Unsupervised Manifold Alignment with TopoGAN

Aligning multi-modal biological data without correspondence information available across modalities

Akash Singh
5156416
Thesis Project for M.Sc. Computer Science, Data Science Track
(Specialisation in Bioinformatics)

Thesis Committee

Prof. dr. ir. M.J.T. Reinders,
Pattern Recognition and Bioinformatics, TU Delft

Dr. C. Lofi,
Web Information Systems, TU Delft

Dr. A. Mahfouz,
Pattern Recognition and Bioinformatics and Dept. of Human Genetics, LUMC

T.Abdelaal,
Dept. of Radiology, LUMC

Electrical Engineering, Mathematics, and Computer Science
Delft University of Technology
Netherlands

Contents

1 Preface

This report discusses my masters thesis where I explored the problem of aligning biological datasets of different modalities in an unsupervised fashion. The project was carried out from December 2020 till August 2021. It has been challenging, fascinating, and eventually, humbling, to employ modern computer science towards tackling problems from the realm of biology.

For readers unfamiliar with the jargon, different modalities in biological data imply measurement of different aspects of a cell. A living cell is a complex machine with myriad different functions and processes co-occuring to achieve fundamental objectives such as cell preservation, cell replication, and cell interaction with the external environment. This begins with a central genetic code, or DNA, of the cell which dictates the machinery and execution of almost everything happening within that cell, including its own replication, thus perpetuating the most fundamental objective of biology - life. In this process, many subtle, but crucial events happen within the cell which leave molecular signatures. The past decade has seen emergence of technologies which can measure one or many of these signatures. It is the measurement of any of these molecular signatures which has been referred to as a "modality" or an "omics-layer" throughout this report. Additionally, this project focuses exclusively on single-cell data. Therefore, "sample" in this report always means a cell.

The motivation behind aligning different modalities is to be able to leverage, simultaneously, all the biological insight present in them. Each modality, in isolation, only captures a small part of the picture. It is when we bring these parts together through integration that we begin to see this picture in entirety, which is the underlying biology driving both life and disease. This realisation has been a profound source of meaning and motivation for me during this project.

I would like to thank Dr. Ahmed Mahfouz, my supervisor, for helping me address the "bigger-picture" questions of the project and Tamim Abdelaal, my daily supervisor, for helping me navigate the day-to-day challenges of the thesis. I would also like to thank Prof. Reinders and Dr. Lofi for being part of my thesis defense committee. I look forward to their feedback on this work. Finally, I feel gratitude towards the scientific community including but also extending beyond the bibilography of this report. It is the framework of ideas developed by the community which allowed me to come up with scientific contributions of my own.

Akash Singh,
August, 2021

2 Abstract

Single-cell multi-modal omics promises to open new doors in bioinformatics by measuring different aspects of cells, thus offering multiple perspectives on the underlying biological phenomenon. Although simultaneous multi-modal measurement protocols do exist, their inherent technical limitations necessitate focus on single modality measurements. These single modality measurements, however, destroy the cell in question, thus making simultaneous measurements impossible. This gives rise to a great availability of multi-modal biological data with no inter-data set sample/feature correspondence. This work proposes a novel approach to align multi-modal data sets in an unsupervised fashion using an Autoencoder to obtain latent embeddings of the modalities and a Generative Adversarial Network to align these latent representations. Minimising the topological error between the original and latent representations of a data set is central to this approach which enables not just the superposition but also alignment of different modalities. Two recently published methods, UnionCom and MMD-MA, have been used for comparison and benchmarking. The approach, termed *TopoGAN*, has been demonstrated to give consistently stable alignments, give better quantitative performance in realistic unsupervised settings, and scale much better in terms of memory requirements as compared to these state-of-the-art methods.

3 Introduction

The technology of Single-cell sequencing has come a long way since being selected as Method of The Year by Nature in 2013 [1]. Initial biological insights offered by single-cell sequencing paved the way for researchers seeking to pair different bio-molecular measurements at single-cell resolution. Measurements of different aspects of a cell, for example scRNA-seq and protein profiling, enables resolution of different cell-types with greater precision [2]. It is almost unsurprising then that Nature selected "Single-cell Multi Modal Omics" as Method of the Year in 2019 [2]. The field has grown tremendously ever since, with new measurement technologies generating a great amount of multi-modal data sets at single-cell resolution. However, there exists a trade-off between noise and scalability for these technologies [3]. Technologies which measure millions of cells in a high-throughput fashion tend to produce noisy and sparse data sets while prioritising high-quality measurements usually means a lower-throughput data set, limited sometimes to only a few hundred cells [3].

Multi-modal measurements on the same cell (figure 1.a), with its practical limitations, is only a part of the overall landscape of singe-cell sequencing. Other technologies obtain multi-modal measurements on distinct cells from the same cellular population. Sometimes, these modalities share some common features (figure 1.b), like spatial transcriptomics and scRNA-seq on the same tissue [4, 5]. The sources of variation common to all modalities can be used as anchors to integrate these measured cells in a common space. Methods like SEURAT, Harmony, ComBat and LIGER use this idea to integrate such data sets [6]. For example, Seurat projects all modalities into a lower dimensional space. Post this projection, it uses Canonical Correlation Analysis (CCA) to identify basis vectors such that the variation along these vectors is the most correlated across modalities [7]. These canonical vectors are then used as "anchors" to align all the modalities. However, because the method relies on a partial/complete overlap in the original feature space across modalities, it is a feasible approach only for datasets depicted by figure 1.b.

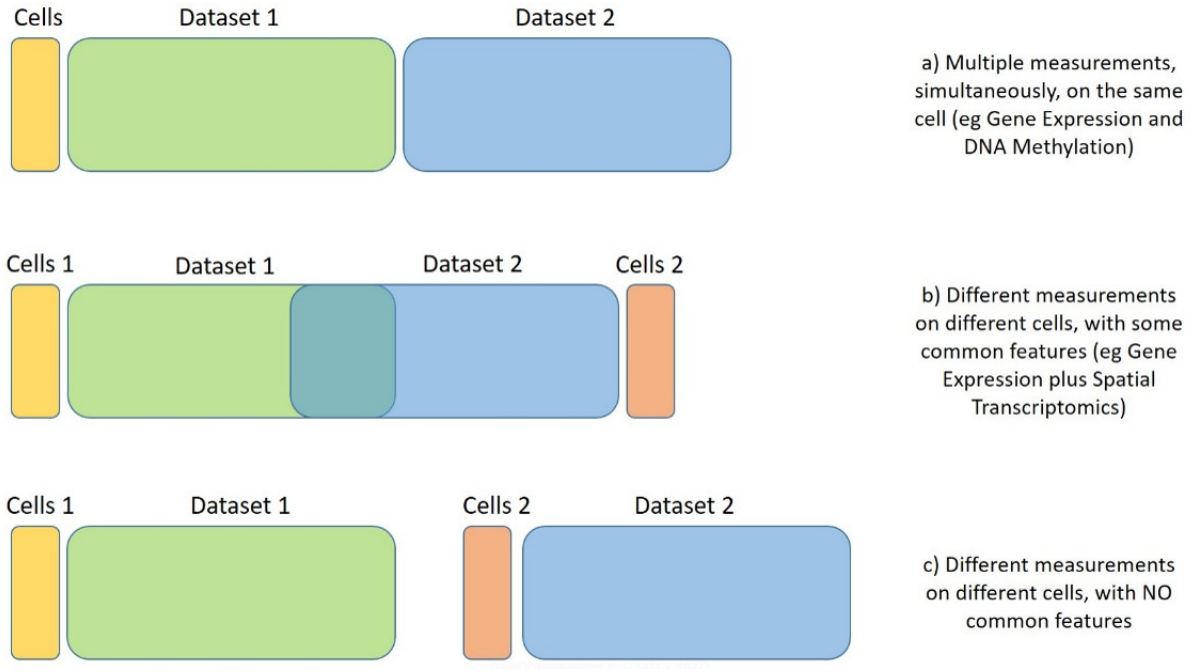

Figure 1: Classification of data availability in single-cell multimodal omics landscape

The most general case of multi-modal measurements is when completely distinct modalities are being measured on distinct cells from the same tissue, hence same cellular population. These measurements destroy the cell in question, but generate data with relatively high-fidelity [8], leaving us with a significant amount of multi-modal data sets measured on the same cellular population without any overlap in their samples or features (figure 1.c). This makes the problem of integrating these multi-modal data sets an important one, despite its many challenges. The most general description of this problem assumes the lack of cell labels or inter-modality correspondence information, making it a completely unsupervised

setting. Methods developed to align data sets with partial correspondence among samples or features have been shown to not work well in the completely unsupervised setting [9].

However, there are approaches developed specifically to tackle the unsupervised alignment problem. UnionCom [9] tries to achieve local and global alignment among different modalities using a geodesic-distance based graph. Although the UnionCom paper demonstrates state-of-the-art performance on both simulated and real data sets, all the data sets reported upon contained perfect cell-cell correspondence. In other words, although the method is developed to align two data sets without using any inter-modality correspondence information (because it does not exist, figure 1.c), it has been tuned and tested on data sets where perfect cell-cell correspondence across modalities does exist (figure 1.a). However, when applied on data sets where this correspondence is removed, performance of UnionCom drops and also becomes inconsistent with multiple runs with different initialisation seeds (experimental results in section 6.5). Additionally, the computational memory requirements of UnionCom scale up in an impractical fashion with number of samples. Combining this with the need for extensive hyperparameter tuning makes UnionCom an expensive method. MMD-MA, like UnionCom, has been developed to not require any correspondence information to perform the actual alignment. However, MMD-MA also requires tuning of 3 hyperparameters, which does require labels or correspondence information. The authors of MMD-MA acknowledge the inability to tune hyperparameters in a truly unsupervised setting [8]. In cases where one-to-one cell correspondence has been partially/completely destroyed, the performance of MMD-MA drops as well, even when implemented with the recommended hyperparameter values as indicated by the authors. On the other hand, SCIM uses a Variational Autoncoder followed by correspondence discovery among cells across modalities using the Network Simplex Algorithm [10]. However, SCIM demands partial correspondence information across modalities in order to align the rest of cells. As a result, this excludes data sets which have absolutely no available correspondence across modalities. There is a need for a method which performs well in the truly unsupervised setting without incurring a prohibitive amount of computational cost. This work proposes a strategy, termed *TopoGAN* to align multi-modal data sets catering to these requirements. Formally, the scientific contributions are:

1. A computationally inexpensive approach which scales well for large number of samples.
2. The method performs well not only for data sets where one-to-one correspondence already exists, but also for cases where the correspondence has been partially/completely destroyed.
3. A loss computation which does not require any correspondence information or labels to be computed. Tuning the method is therefore possible in the truly unsupervised setting as well.

4 Unsupervised Manifold Alignment

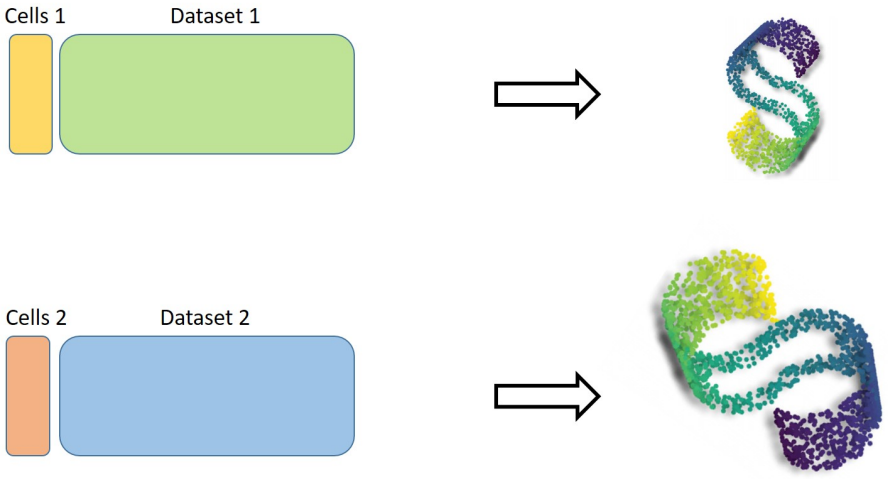

Figure 2: Multi-modal biological data sets can be assumed to lie on similar latent manifolds

Let us talk about the fundamental assumptions which will help us solve this problem. Single-cell data sets, being high-dimensional in nature (gene expression data sets contain 20K features while methylation data sets may contain 150K upto 850K features per sample) are considered to assume the shape of a manifold [11]. This assumption stems from the idea that there is a lot of redundancy in biological data sets. For example, a lot of genes work together within gene-expression pathways, thus constraining the cells to a relatively lower number of states as indicated by the dimensionality of gene-expression data. In traditional machine-learning settings, high-dimensional data sets with correlated features are assumed to have a lower-dimensional manifold structure [12]. A manifold is a low-dimensional shape arranged in a way that it resides in a higher number of dimensions. We now have our first assumption - *high dimensional biological data sets have a latent manifold structure.* Because these single-cell multi-modal data sets are generated from the same tissue sample, cells in all modalities come from the same biological phenomenon/generative distribution even though the individual, inter-modality cells do not match [13]. This idea gives us our second assumption – *different modalities of the same cellular population lie on the same underlying manifold* (figure 2). Thus, if we align these manifolds, we align the data sets. It is important to note the distinction between aligning two manifolds and merely superimposing them. While a superimposition will simply ensure global matching of manifolds, alignment will ensure local regions on one manifold are aligned to the corresponding ones from the other (figure 3). In terms of a pair of multi-modal data sets, alignment of the data sets ensures cells of a particular type are aligned to the same cell-type in the other data set (figure 4).

Because in many cases, obtaining metadata, like cell-type information can be difficult, unavailability of metadata is an additional constraint. As a result, this manifold alignment needs to be completely unsupervised, thus defining the problem statement as *Unsupervised Manifold Alignment.*

4.1 Aligning Multi-modal Biological data sets

Guided by the two central assumptions as discussed in the previous section, the problem of *Unsupervised Manifold Alignment* can be broken down into two sub-problems:

1. **Manifold Projection**: projecting high-dimensional data to the underlying latent manifold, while preserving relevant *intra-modality structure.*

2. **Manifold Alignment**: projecting individual manifolds in a shared latent space such that relevant *inter-modality correspondence* is preserved.

What is meant by preserving intra-modality structure? Effectively, it means that we are reducing the dimensionality of the data in such a way that local and global correspondence among the samples

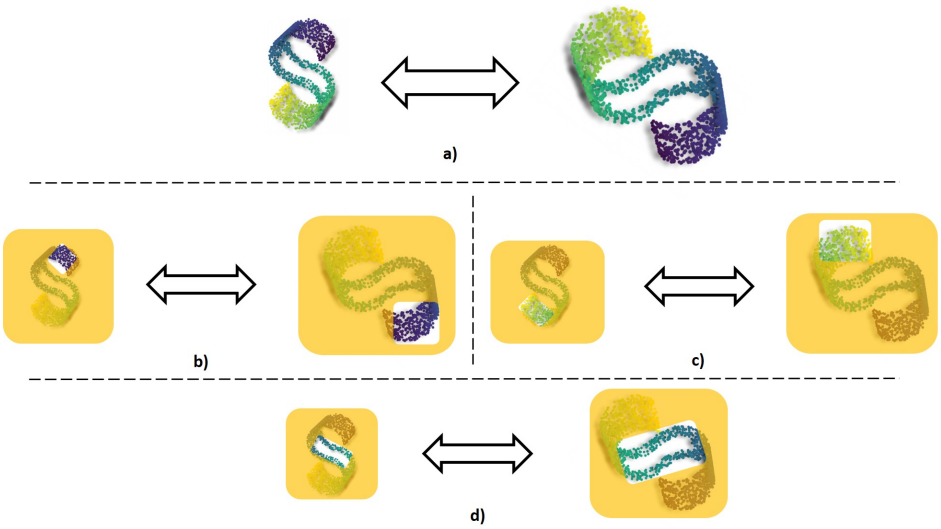

Figure 3: Manifold Alignment: a - The manifolds to be aligned; b, c, d - different corresponding regions in both manifolds

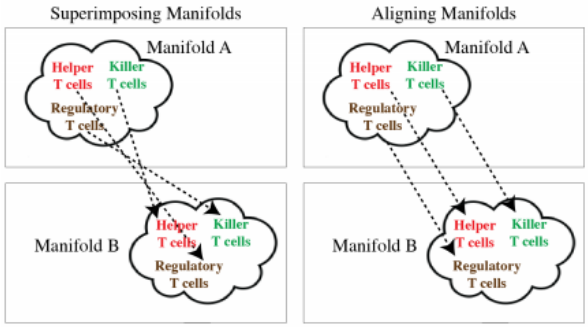

Figure 4: Manifold Superposition vs Manifold Alignment *(Image Courtesy: Amodio, M. and Smita Krishnaswamy. "MAGAN: Aligning Biological Manifolds." ICML (2018) [14])*

are preserved. In other words, this can be understood as obtaining a lower dimensional representation of the data set such that the underlying manifold structure is represented with fidelity (figure 2). Inter-modality correspondence, on the other hand, refers to the way a sample relates to samples from the other modality. For instance, cells from a modality would have maximum correspondence with regions in the other modality such that the local neighbourhoods (in terms of distribution of cell-types) would be similar in both cases (figure 4).

Encapsulating the topology of a data set through pair-wise distance matrices is an idea widely used in manifold alignment methods. Unioncom [9] and MMD-MA [8] use different strategies to preserve the structure represented by these distance matrices by minimising a distance/dissimilarity score computed between the pairwise distance matrices of the original and latent representations. The size of these matrices increases by the square of the number of samples in a data set. For data sets with tens of thousands of cells, this demands a significant amount of compute memory. However, the point-pairs crucial to the topology is usually a subset of all possible point-pairs [15]. A simple presentation of this idea is that in a group of points arranged in the form of a 3-dimensional pyramid, the point pairs defining the edges and faces of the pyramid are much more significant than any point-pair inside the body of the pyramid.

As a result, the approach of Topological Autoencoders [16] is particularly interesting, because it tries to preserve only a subset of all pair-wise distances. Topological Autoencoder focuses only on those point-pairs which are most crucial in determining the topology of the manifold instead of trying to optimise all possible point-pairs. Determination of these point-pairs relies upon assessing the most crucial topological features in a given data modality. Topological Autoencoders have been shown to

give reliable topology approximations with mini-batches as well [16], meaning the problem of topology assessment can be approached in a piece-wise fashion. Therefore, this approach offers the potential to significantly reduce memory requirements. Additionally, Topological Autoencoders have never been tried on biological data sets, to the best of our knowledge. The core idea of Topological Autoencoders is minimising the *topological error* between the original and latent representations. As will be discussed ahead, computing the *topological error* between representations of the same data set in different spaces is central not only to our first sub-problem (Manifold Projection) but also to the more challenging sub-problem of Manifold Alignment.

5 Method

5.1 Minimising topological error between original and latent spaces

As discussed in the previous section, preserving the inherent structure of a data set is an important objective in manifold alignment methods. Topological Autoencoder, in particular, uses Persistence Homology to infer the topology of a high-dimensional data set [16]. Persistence Homology selectively considers edges connecting point-pairs below a certain distance threshold. These edges are used to construct local neighbourhoods which in turn constitute large-scale topological features. By repeating this procedure for a large range of distance thresholds, topological features most *persistent* over this range are revealed [15]. It is only these topological features and the point-pairs constituting them which are considered significant to the topology of the data set. As a result, preserving the distances between these point-pairs across projections of a data set in different spaces will preserve the topology of the data set. These topologically relevant point-pairs/edges, formally known as *Persistence Pairings* are denoted as π^X for the original space and π^Z for the latent space in equation 2.

Formally, the loss function to preserve topology of the original data when projected in a latent space can be formulated as (equation 2):

$$L = L_r + \lambda L_t \tag{1}$$

where,
L is the overall loss for Topological Autoencoders [16]
L_r is the reconstruction loss
λ is the weight of topological loss in the overall loss
L_t is the topological loss

$$\begin{aligned} L_t &= L_{XZ} + L_{ZX} \\ L_{XZ} &= \frac{1}{2}\|A^X[\pi^X] - A^Z[\pi^X]\|^2 \\ L_{ZX} &= \frac{1}{2}\|A^Z[\pi^Z] - A^X[\pi^Z]\|^2 \end{aligned} \tag{2}$$

A^X / A^Z: Distance matrix in original/latent space
π^X / π^Z: Topologically relevant point-pairs (persistence pairings) in original/latent space
$A^X[\pi^Z]$: Subset of distances in *original* space defined by topologically relevant edge indices in *latent* space
$A^Z[\pi^X]$: Subset of distances in *latent* space defined by topologically relevant edge indices in *original* space
L_{XZ}: Ensures point pairs relevant to the *original* manifold are equidistant in both spaces (original and latent)
L_{ZX}: Ensures point pairs relevant to the *latent* manifold are equidistant in both spaces

5.2 Manifold projection with Topological Autoencoder

Topological Autoencoder (TopoAE) [16] uses equation 1 to obtain a lower dimensional representation of a given input data set. The framework of Topological Autoencoder was used to project all modalities of a multi-modal data set independently into lower-dimensional spaces. The original implementation was slightly modified to accommodate the initial dimensionalities of different data sets. The encoder network uses 2 hidden layers and applies Batch Normalisation and Relu activation after each hidden layer. Having a hidden layer between the input and output layers enables dimensionality reduction of the input data in a step-wise fashion. The decoder network simply mirrors the encoder network in terms of input-output sizes of each layer (figure 5). The Topological Error term (λL_t) in the loss function (equation 1) is expanded in equation 2. It is this term which ensures the topology prevalent in the input space is preserved in the latent embedding. The authors of Topological Autoencoders recommend a

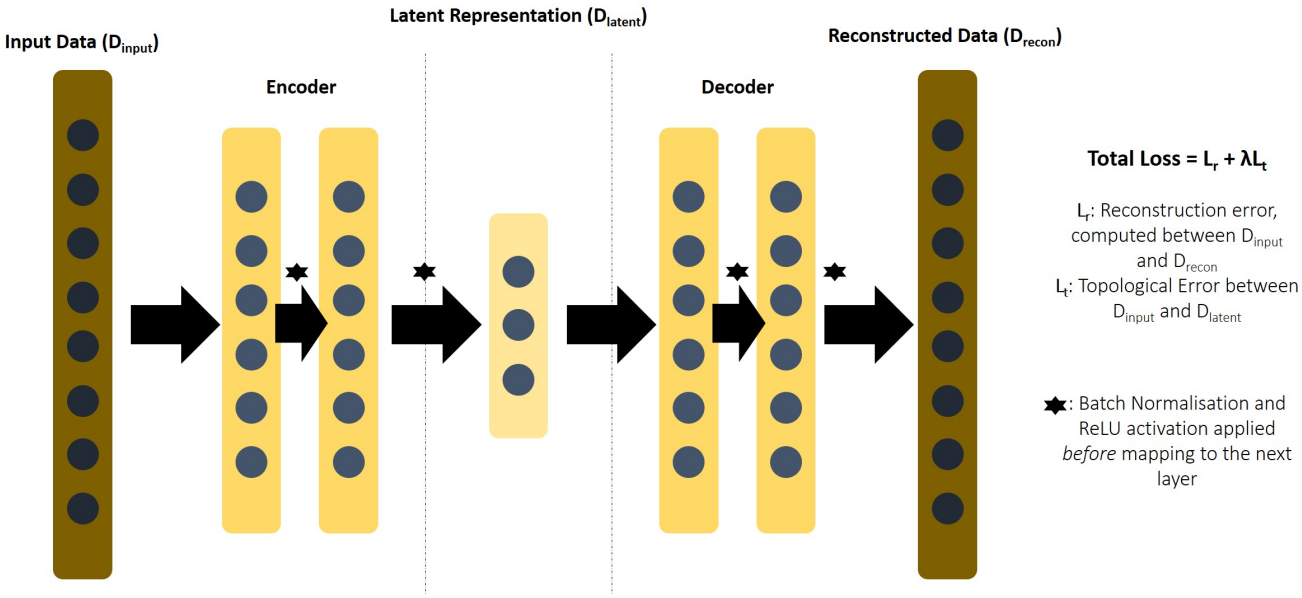

Figure 5: Schematic diagram of Topological Autoencoder, as implemented in this work. The loss function defined in equations 1 and 2 are indicated in the figure

range of values for the hyperparameter λ (0.5 - 3.0) [16]. At this point, the performance of Topological Autoncoder was compared with a standard Variational Autoencoder, which is similar to the approach used by SCIM [10] for its manifold projection step. A Variational Autoencoder (VAE) projects the original data into a latent space, just like a regular autoencoder with the difference being that instead of learning the point representation of data, the VAE tries to learn the generative distribution of the data [17], thus making it more generalizable (schematic displayed in figure 6). The VAE loss function is described below (equation 3):

$$L = L_r + \alpha KLD \tag{3}$$

where,
L is the overall loss for the Variational Autoencoder
L_r is the reconstruction loss between the original data and its reconstruction by the decoder
KLD is the KL-Divergence between the original data and its latent representation
α is a hyperparameter, controlling the influence of KL-Divergence in the total loss

Both models were trained for 100 epochs, with early stopping if the loss did not improve for 10 consecutive epochs. Another manifold learning technique, UMAP was used as a baseline against these two methods. UMAP uses the theoretical framework of Riemannian geometry to learn the underlying manifold, thus projecting a given data set into a lower dimensional space [18]. The standard UMAP implementation (available as *umap-learn* in python) was used to benchmark these two methods. Topological Autoencoder performed decisively better than the VAE and UMAP on both qualitative and quantitative grounds. Additionally, the quantitative assessment helped determine the optimum dimensionality for the lower dimensional projection (discussed in results section). Ultimately, it was decided to project all modalities into an 8-dimensional space.

5.3 Manifold alignment with TopoGAN

Having obtained a latent embedding of each data set through Topological Autoencoder, the next step is to obtain representations of all modalities in a common space. This is because independently reducing all modalities to a lower-dimensional embedding does not necessarily project them in same feature space. Methods like UnionCom [9] ultimately project all modalities in a new, common space, thus achieving the alignment of manifolds. However, it is equally valid to use one of the modalities as a target space and simply project the remaining modalities in that feature space. Recalling our second assumption

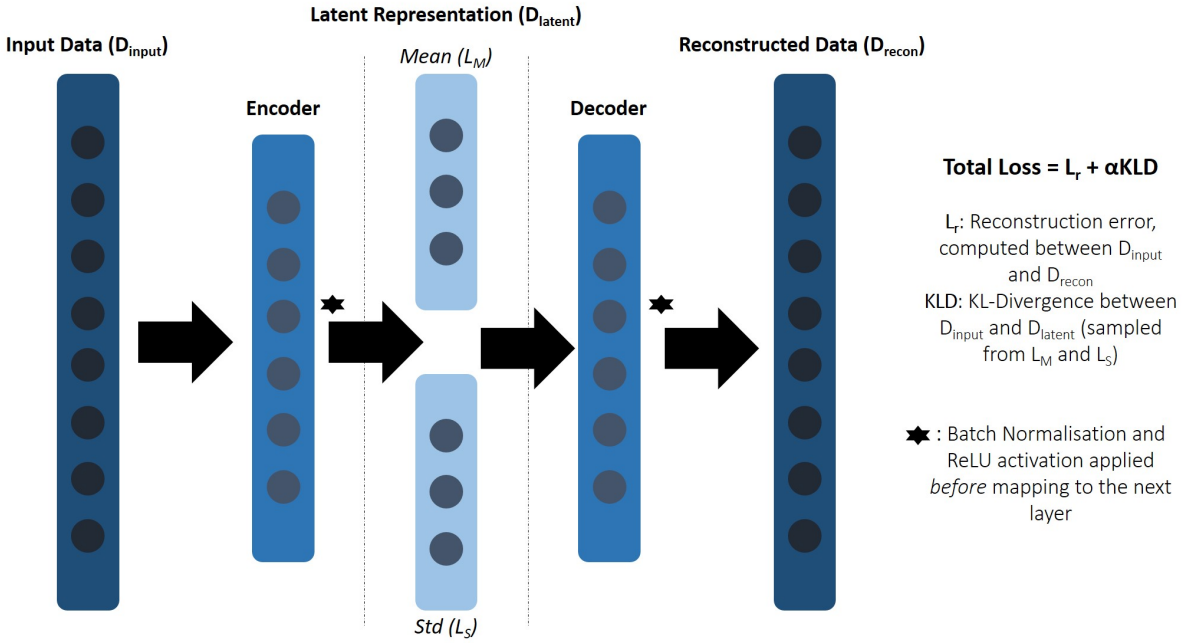

Figure 6: Schematic diagram of Variational Autoencoder, as implemented in this work. The loss function defined in equation 3 is indicated in the figure

from section 4, these sets of unmatched cells present in different modalities are assumed to lie on the same latent manifold. In order to align the latent embeddings of different modalities, it was decided to use Generative Adversarial Networks (GANs). An existing method, MAGAN [14], uses a modified GAN setting to project biological data sets onto each other. However, MAGAN requires correspondence information between points across different modalities as an anchor. In the unsupervised setting being focused upon in this work, we do not have that luxury. Regardless, the potential of GANs to convincingly replicate data generative distributions has been well researched by the Computer Vision community [19]. Because the target modality is untouched during this process and the Generator tries to make all other modalities indistinguishable from the target, it is fair to say the Generator will try to project these modalities onto the manifold of the target modality.

In a standard GAN setting, a Generator Network, which is a deep neural network, tries to generate, from random noise, data which mimics a target distribution. A Discriminator Network, another deep neural network, tries to distinguish between samples belonging to the actual target distribution and those projected by the Generator Network. As these networks are trained in an alternating fashion, the Discriminator gets better at telling the fake samples apart from the true ones. This, in turn, forces the Generator to mimic the generative distribution process of the target in a much more authentic fashion. Once the training process is stabilised, the Generator Network can be used to synthesise fake samples. This is possible because the Generator has learned to generate a distribution which resembles the target distribution.

After extensive experimentation with various GAN architectures, it was decided to use a single hidden-layer Generator against a double hidden-layer Discriminator. Ideas from existing research in GAN training were borrowed to achieve stable GAN training - the discriminator weights and biases were sampled from a normal distribution (with mean 0 and standard deviation 0.02), as it helps stabilise the training process [20]. Additionally, Leaky Relu was used as the activation function for the Discriminator with an activation value of 0.2, again, to stabilise the training [20].
As discussed in the MAGAN paper as well, conventional GANs will do a good job at manifold *superposition* but good performance on manifold *alignment* remains elusive in the absence of inter-modality correspondence information [14]. As will be demonstrated in detail in the results section, this phenomenon was observed while training GANs for this work as well - performance of a given GAN architecture fluctuated beyond acceptable limits with different initialisation seeds. This motivated the computation of topological error between the source data and its projection in target space. An in-depth discussion

of this idea and its validity is presented in results section.

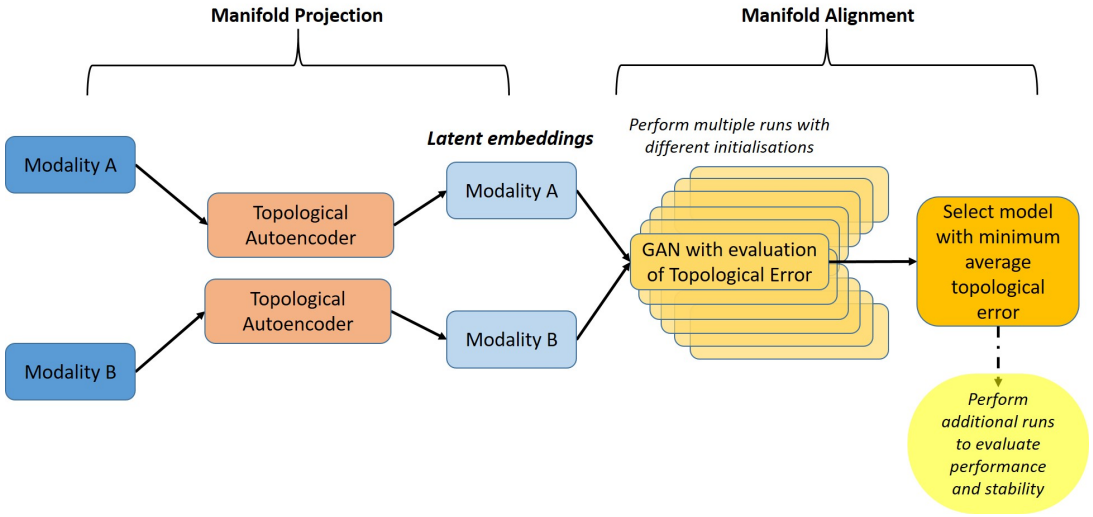

Figure 7: Workflow for Unsupervised Manifold Alignment

These topological error computations have then been used in the following manner to perform *Unsupervised Manifold Alignment*:

1. Use Topological Autoencoder to obtain independent lower-dimensional embeddings of the input modalities. In this work, all data sets were projected into an 8-dimensional space, as discussed.

2. Fix one of the modalities as the target, and the other one as source. For both PBMC and CITE-Seq data, RNA has been chosen as the target modality. A modified GAN setup is then initialised with a Generator and Discriminator network where the Generator trains to make the source modality indistinguishable from the target while the Discriminator trains to distinguish the two modalities. This setup is trained for 1000 epochs and the trained Generator model is saved every 100 epochs. The topological error between the source data and its representation in the target space (projected by the Generator) is computed every 100 epochs according to equation 2.

3. The above experiment is repeated 20 times for the exact same GAN architecture, with a different initialisation seed each time. For each of the 20 expriments, the average topological error for the last 6 computations (epochs 500 to 1000) is calculated. The experiment with the lowest average topological error is then selected for the next step. A lower topological error averaged across 600 epochs (epoch 500 to 1000) is a much stronger indicator of appropriate alignment of manifolds as compared to low topological error for just one of the epochs, which could simply be the result of a local minimum.

4. The Generator model thus obtained is loaded as the Generator model in a new GAN with a new discriminator as its adversary. This second generation GAN is then trained for 1000 epochs with a number of different initialisation settings. The evaluation metrics are computed to evaluate performance of the GAN. Please note that multiple GAN trainings in this step is only meant to evaluate the method. In order to obtain aligned manifolds, training the second generation GAN once is enough.

The performance of the second generation GAN thus obtained has been discussed and compared against UnionCom in the results section.

5.4 Data

5.4.1 PBMC Data - Full and Partial

The PBMC data set consists of human peripheral blood mononuclear cells obtained from a healthy female homo sapiens donor (aged 25). The cells were obtained by the 10x multiome RNA+ATAC kit and the data set is publicly available on the 10xGenomics website. The data set consists of two assays,

RNA and ATAC. Both modalities have been measured on the same 10,412 cells, thus giving us a multi-modal data set where one-to-one cell correspondence is actually available. The Seurat Vignette (available at *satijalab.org*) was used to pre-process the data set using standard procedures for RNA and ATAC data. The data is processed by first loading the raw counts data in R. This data is then log-normalized, post which the most variable features are extracted. These variable features are then projected in a 50-dimensional space using Latent Semantic Indexing (LSI) which essentially obtains a given number of "latent" dimensions to faithfully represent the data in. The LSI output is our starting point, i.e. the full PBMC data set displayed in figure 8.

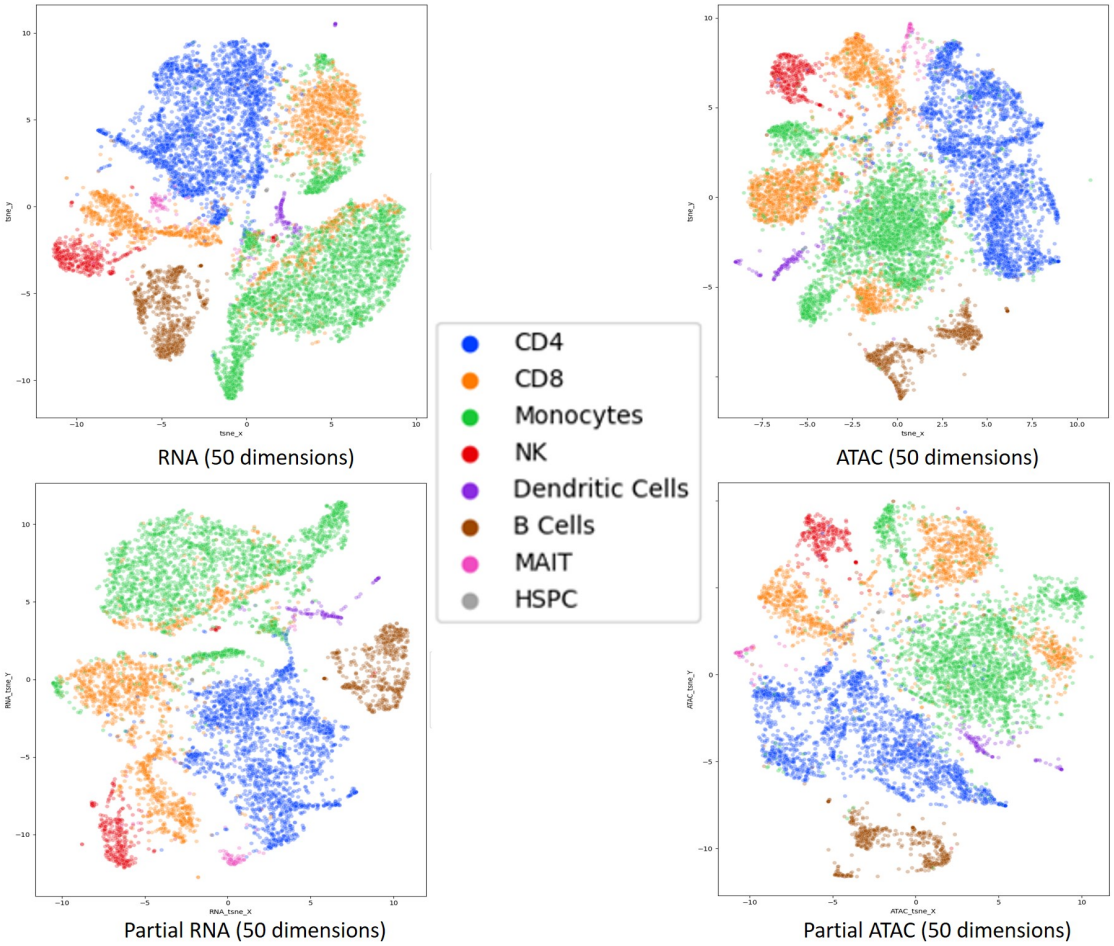

Figure 8: tSNE plots of 50-dimensional PBMC Data for two modalities - RNA (left column) and ATAC (right column); top row - full PBMC data with 10,412 cells per modality, bottom row - partial PBMC data with 7,329 cells per modality

This pre-processing stage delivered two data sets (RNA and ATAC), each with 10,412 cells and 50 features. These 10,412 cells were sampled from 19 different cell-types which were then arranged into 8 broad categories - CD4, CD8, Monocytes, NK, Dendritic Cells, B Cells, MAIT, and HSPC. The data sets depicted in figure 8 were used as the starting point, i.e., as inputs for the Manifold Projection step (left most step in figure 7).

However, it is not prudent to develop a method for *Unsupervised Manifold Alignment* and test it only on a data set where perfect cell-to-cell correspondence already exists across modalities. Therefore, 30 percent cells are randomly and independently removed (in a stratified fashion) from both RNA and ATAC data sets. This results in a modified PBMC data set with 7,329 cells in each modality. Out of these, 5,187 cells are common in both modalities. The remaining 2,142 cells in RNA are completely different from the remaining 2,142 cells of ATAC. However, as can be seen in figure 8, the overall structure of the data sets has not been drastically distorted by this random removal of cells. Therefore, trying to align the partial data sets should not be extraordinarily more difficult than the original ones.

5.4.2 CITE-Seq - Full and Partial

The CITE-Seq data set [21] consists of 30,672 scRNA-seq profiles measured alongside a panel of 25 antibodies from bone marrow. The data is available in two assays, RNA and antibody-derived tags (ADT). Like PBMC, CITE-Seq data has also been pre-processed by the Seurat Vignette in a similar fashion - most variable features are extracted post normalisation of raw data. After scaling these features, data is projected into a lower dimensional space using Principal Component Analysis. The final processed RNA data set contains 30,672 cells and 50 features while the ADT data set contains the same 30,672 cells and 24 features (figure 9) . Therefore, one-to-one cell correspondence is available across modalities for CITE-Seq data as well. There were 27 different cell-types available in the annotations data for these 30,672 cells which were then arranged into 5 broad categories - Progenitor cells, T Cell, Mono/DC, NK, and B Cell.

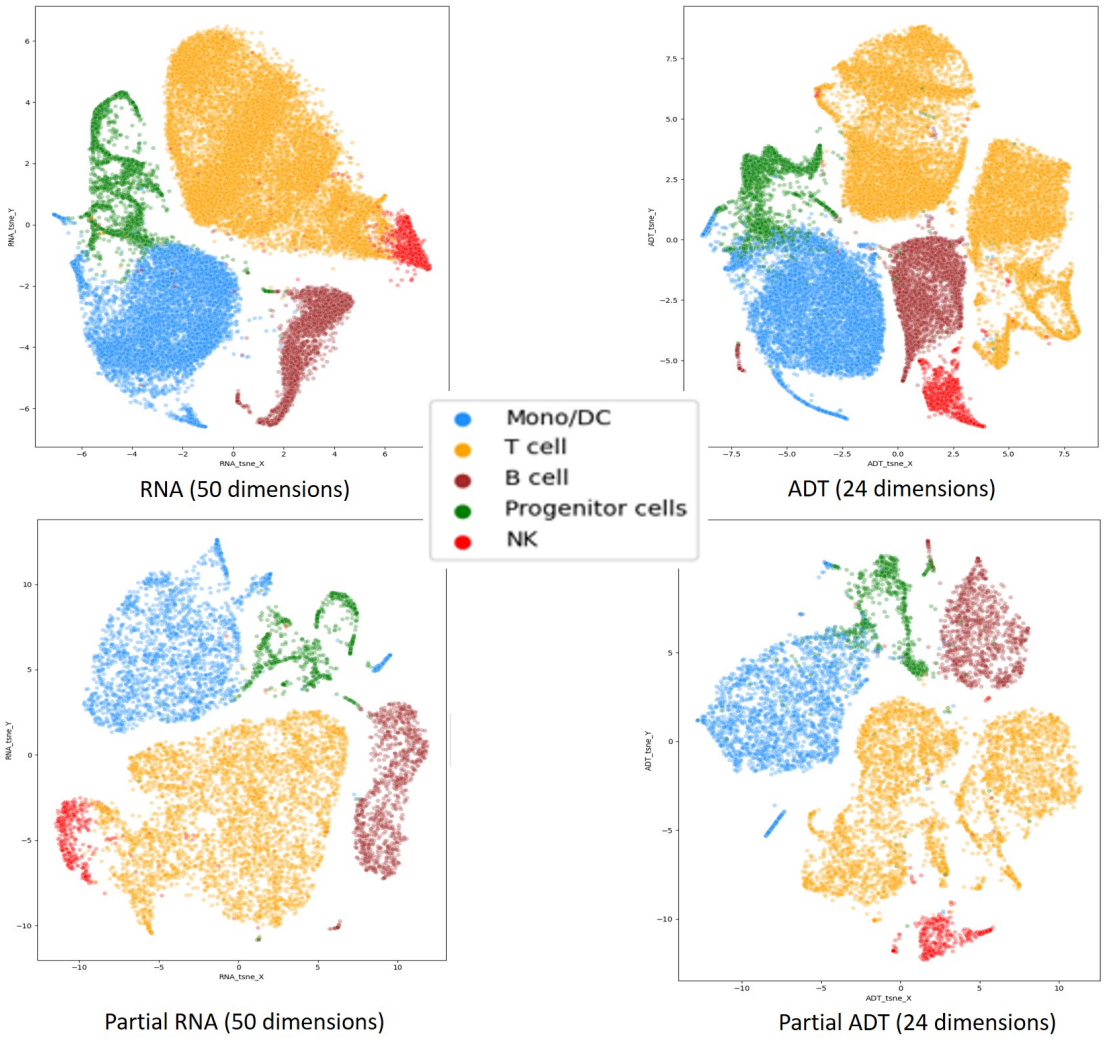

Figure 9: tSNE plots of CiteSeq Data for two modalities - RNA (left column) and ADT (right column); top row - full CiteSeq data with 30,672 cells per modality, bottom row - partial CiteSeq data with 9,053 cells per modality

A smaller data set has been derived from the original CITE-Seq data [21] introduced above. 9,053 cells were randomly selected from the RNA data to generate a partial RNA data set. Then, 9,053 cells were randomly selected from the ADT data, ensuring that there is *no one-to-one cell correspondence* whatsoever of the resulting data with the partial RNA data. As a result, a partial CITE-Seq data set was generated with 2 modalities and 9,053 cells per modality and no overlap between cells across modalities (figure 9). It might occur to the reader that complete destruction of one-to-one cell correspondence across modalities in the original CITE-Seq data can be already achieved with 15,000 cells per modality,

which is correct. But in order to enable running MMD-MA and UnionCom on CITE-Seq data on a reasonable budget of resources, only 9,053 cells per modality were included for CITE-Seq data. Because cell correspondence is completely destroyed in this case, it is an instance of the truly unsupervised setting.

5.5 Evaluation Metrics

5.5.1 Assessing Manifold Projection

SCIM uses a Variational Autoencoder [10] to project the input data sets into a lower-dimensional space. Although SCIM uses partial cell-cell correspondence information across modalities for alignment, its manifold projection step is completely unsupervised. As a result, a Variational Autoencoder was used as a baseline, in addition to UMAP, to benchmark the performance of Topological Autoencoder. The first quantitative metric used for assessing the performance of different methods is Silhouette score. The silhouette score is a measure of similarity of a point to other points from its own cluster (cohesion) and distance/dissimilarity of this point to points from other clusters (separation) [22] (equation 4).

$$SilhouetteScore = (b - a)/max(a, b) \tag{4}$$

where,
a is the average intra-cluster distance i.e the average of distances between all point-pairs within a cluster
b is the average inter-cluster distance i.e the average distance between all clusters

Additionally, the Kullback–Leibler divergence (KL_σ) between the density estimates of the input and latent spaces is computed. KL_σ quantifies the dissimilarity between two distributions - representation of the data in the input and latent spaces in this case. The Topological Autoencoder paper also relies on this metric for evaluation of the obtained latent embeddings [16]. In this case, the value of σ, which represents the length scale of the Gaussian kernel, was chosen to be 0.01.

5.5.2 Assessing Manifold Alignment

Because all the data sets used to perform the Manifold Alignment step were accompanied by cell annotation information, it became possible to assess the quality of alignment achieved in quantitative terms in addition to qualitative evaluations like 2-dimensional tSNE plots of aligned data. The primary metrics for quantitative evaluation are *sub cell-type matches* and *cell-type matches*. As discussed in section 5.4, the data sets used contain cell annotation information, in the form of sub cell-type and cell-type. In a perfect alignment of a multi-modal data set, cells of a certain type from one modality will be projected in the local neighbourhood of cells of the same type. Therefore, for each cell in the projected data, its k-neighbours from the other modality are determined. The cell-label with the highest frequency among these neighbours is considered the dominant label. If this matches with the label of the original cell from the projected modality, it is considered to be a match, otherwise not. After computing this binary metric for all cells in the projected modality, the percentage of cells which report a match is computed. This percentage is reported as the final metric - sub cell-type/cell-type matches. For UnionCom and MMD-MA, there is no projected modality as all modalities are projected into one common space. In these cases, one modality has arbitrarily been selected as the projected modality as the selection criteria bears no effect on the final conclusions drawn from the results.

The authors of UnionCom use *Label Transfer Accuracy* as the evaluation criteria in cases where perfect one-to-one correspondence is not available [9]. After the manifolds have been appropriately aligned, a k-nn classifier is first trained on one of the modalities to predict the cell-type label for any given cell. Once trained, this classifier is used to predict the labels for all cells from the other modality. These predictions are then compared against the actual labels of these cells to compute the final Label Transfer Accuracy score. Upon closer inspection of the implementation of UnionCom (https://github.com/caokai1073/UnionCom), the two evaluation metrics, Label Transfer Accuracy and Cell-type Matches were found to eventually rely upon the same nearest-neighbour computation routines provided in *scikit-learn*. Not surprisingly, the two scores tend to agree with each other. Nevertheless, to make a fair comparison between this work and UnionCom, performance has been reported on both of these metrics. Additionally, the procedure described above has been used to create another metric, *Sub Label Transfer Accuracy*, where the k-nn classifier predicts the sub cell-type labels instead of the cell-type labels.

6 Results

6.1 Manifold Projection using Topological Autoencoder

The full PBMC data set was primarily used to compare the performances of Topological Autoencoder and Variational Autoencoder for Manifold Projection. Qualitatively, the Topological Autoencoder results did not vary too much with changes in the topological error weight, λ (equation 1). The Variational Autoencoder demands tuning of its hyperparameters - the coefficients of KL-Divergence and L_1-loss (α and β, equation 3). Figures 10 and 11 compare the outputs of both approaches, the 8-dimensional latent embeddings, for RNA and ATAC data. It may be a little difficult to make a qualitative evaluation in the case of RNA data (figure 10), but it can still be observed that the VAE mixes the CD4 (blue) and CD8 (orange) populations while Topological Autoencoder maintains them as separate clusters. Similarly, VAE mixes the CD8 (orange) and NK (red) populations which are projected separately by the Topological Autoencoder. For ATAC data (figure 11), it is abundantly clear that Topological Autoencoder does a much better job of preserving the inherent topology of the data as compared to the Variational Autoencoder. UMAP, on the other hand, displays a much more rigid and sharp approximation of the manifold. Please note that the outputs shown in figures 10 and 11 are the best results obtained for both methods post hyperparameter tuning.

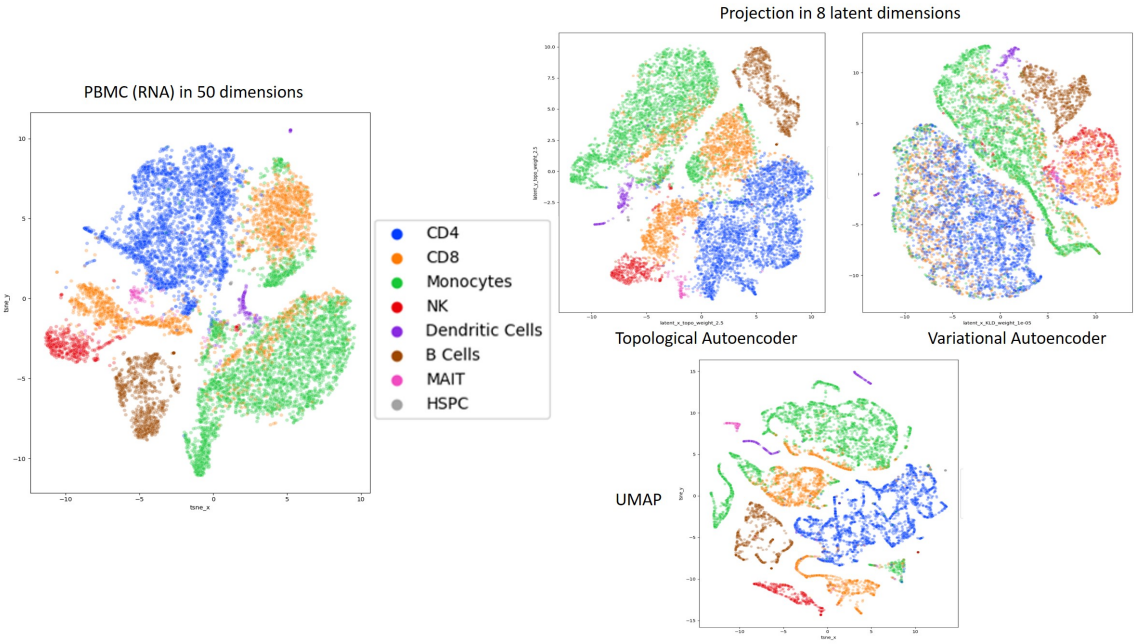

Figure 10: 2-D tSNE Outputs of Topological Autoencoders, Variational Autoencoders, and UMAP plotted against the original RNA data - Topological Autoencoder displays the most faithful preservation of structure inherent in the original 50-dimensional data (left)

Tables 1 and 2 provide empirical evidence that TopoAE performs better than a VAE and the baseline - both tables have been sorted in decreasing order of $KL_{0.01}$ score. A lower value of $KL_{0.01}$ indicates less divergence between the original and latent projections, therefore better preservation of the manifold structure. To evaluate silhouette score values, it is helpful to recall that higher silhouette scores imply tightly arranged and well-separated clusters. Topological Autoencoder achieves a much higher silhouette score as compared to the Variational Autoencoder. At the same time, it achieves a much lower $KL_{0.01}$ as compared to the VAE. For the ATAC data, the Topological Autoencoder is shown to clearly outperform the VAE on both metrics (table 2). Although UMAP does achieve high values for silhouette score in both modalities, it also scores relatively higher on the $KL_{0.01}$ metric, indicating a less faithful preservation of the manifold structure. This is evident in the qualitative results as well (figures 10 and 11), where UMAP learns a much more rigid approximation of the manifolds. Please note that the input data for all these methods was the same, the 50-dimensional RNA/ATAC datasets obtained after pre-processing the raw data, as discussed in section 5.4. Additionally, these tables indicate 8 to be a slightly more favourable number of latent dimensions as compared to 6, based on the values for silhouette score and

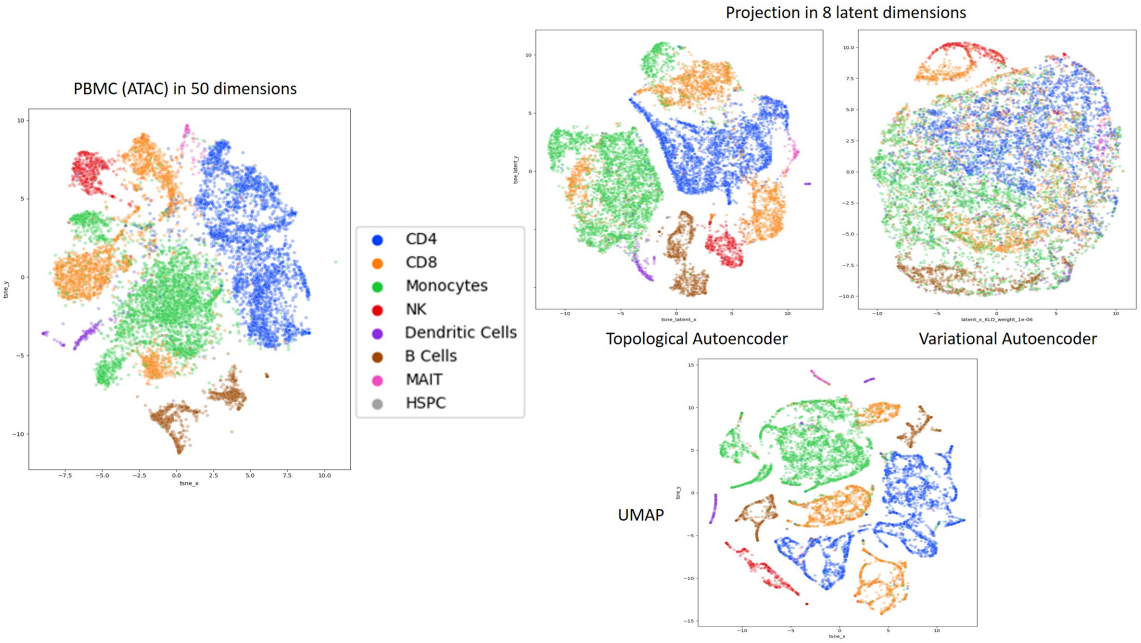

Figure 11: 2-D tSNE Outputs of Topological Autoencoders, Variational Autoencoders, and UMAP plotted against the original ATAC data - Topological Autoencoder displays the most faithful preservation of structure inherent in the original 50-dimensional data (left)

$KL_{0.01}$ value.

Method	**Latent Dimensions**	**Hyperparameter**	**Hyperparameter Value**	**Silhouette Score: Latent**	$KL_{0.01}$
Topological Autoencoder	8	Topology Weight	2.0	0.175	0.007
Topological Autoencoder	8	Topology Weight	3.0	0.061	0.007
Topological Autoencoder	6	Topology Weight	1.5	0.096	0.009
Variational Autoencoder	8	KLD Weight	1e-5	-0.123	0.020
Variational Autoencoder	8	KLD Weight	1e-4	-0.03	0.020
Variational Autoencoder	8	KLD Weight	1e-6	-0.2	0.026
UMAP	8	Default	-	0.229	0.33

Table 1: Quantitative comparison of top experiments of Topological Autoencoder and Variational Autoencoder, with UMAP as baseline for Manifold Projection on PBMC RNA data. The *Silhouette Score* for original RNA data is 0.135. Length scale for the Gaussian kernel, σ, is selected as 0.01 ($KL_{0.01}$)

6.2 Inconsistent Alignments with Generative Adversarial Network (GAN)

Once the samples in each modality have been projected into a lower-dimensional space, they need to be projected into one common space in a manner that they are appropriately aligned. As discussed earlier,

Method	**Latent Dimensions**	**Hyperparameter**	**Hyperparameter Value**	**Silhouette Score: Latent**	$KL_{0.01}$
Topological Autoencoder	8	Topology Weight	1.0	0.091	0.001
Topological Autoencoder	8	Topology Weight	0.5	0.061	0.001
Topological Autoencoder	6	Topology Weight	3.0	0.038	0.001
Variational Autoencoder	8	KLD Weight	1e-6	-0.099	0.02
Variational Autoencoder	8	KLD Weight	1e-5	-0.160	0.012
Variational Autoencoder	8	KLD Weight	5e-5	-0.124	0.021
UMAP	8	Default	-	0.275	0.2

Table 2: Quantitative comparison of top experiments of Topological Autoencoder and Variational Autoencoder, with UMAP as baseline for Manifold Projection on PBMC ATAC data. The *Silhouette Score* for original RNA data is -0.011. Length scale for the Gaussian kernel, σ, is selected as 0.01 ($KL_{0.01}$)

Epoch	**Source**	**Target**	**Run 01**	**Run 02**	**Run 03**	**Run 04**	**Run 05**	**Mean**	**Standard Deviation**
100	ATAC	RNA	54.1	30.0	15.8	36.8	57.7	34.5	17.3
200	ATAC	RNA	50.7	18.3	20.2	37.0	59.3	37.1	18.3
300	ATAC	RNA	63.3	30.2	14.2	51.5	70.3	39.0	19.5
400	ATAC	RNA	64.7	29.7	11.9	46.6	71.1	39.2	18.6
500	ATAC	RNA	65.9	28.1	9.6	45.6	71.6	40.5	20.1
600	ATAC	RNA	58.4	29.1	8.8	37.1	75.4	40.7	20.3
700	ATAC	RNA	65.9	26.4	10.5	37.3	74.0	41.2	20.6
800	ATAC	RNA	62.9	30.1	8.6	42.0	76.8	41.0	20.7
900	ATAC	RNA	60.6	30.5	8.8	39.8	74.8	41.5	20.8
1000	ATAC	RNA	61.8	26.8	6.1	41.1	75.8	41.4	20.6

Table 3: GAN performance on 5 different runs (PBMC data); all displayed results are for percentage *cell-type matches* for k-neighbourhood=5; Mean and Standard Deviation are computed from the results of 40 experiments

we selected Generative Adversarial Networks (GAN) for this task. Figure 12 displays the results for 2 different experiments in the top panel. Cases I and II in the figure differ only by the initialisation of model weights. Everything else - the data, model architecture, hyperparameters, is the same. In both cases, the ATAC modality (from PBMC data set) has been displayed *after* it has been projected into the RNA space by the Generator network of GAN. As will be argued now, the Generator has so far only tried to fool the discriminator by *superposing* them without any explicit attempt at *aligning* them. By deliberately displaying the top panel of figure 12 in greyscale, all cell-label information has been completely removed. By simply looking at the greyscale images in figure 12, one may judge both cases to have achieved a similar degree of success in terms of manifold *superposition*. The bottom panel of Figure 12 reveals the same images in color, including cell-label information. Please note that both modalities contain the exact same set of cells (the full PBMC data set). It is immediately revealed that both cases vastly differ when it comes to success in *aligning* the two manifolds. While case II exhibits significant fidelity in aligning the manifolds by successfully projecting the vast majority of CD4 and Monocytes onto the correct cell-type in RNA space, these two cell-types have been almost completely mis-aligned in case I.

This inconsistency of GAN in aligning cells with the correct cell-types from the target modality (in other words, manifold *alignment*) persists despite extensive experimentation with different GAN architectures, hyperparameter combinations, and loss functions. Additionally, this behaviour has been observed to be independent of the Generator/Discriminator losses during the corresponding training runs. Figure 13 displays the performance (percentage cell-type matches) for 10 experiments with the

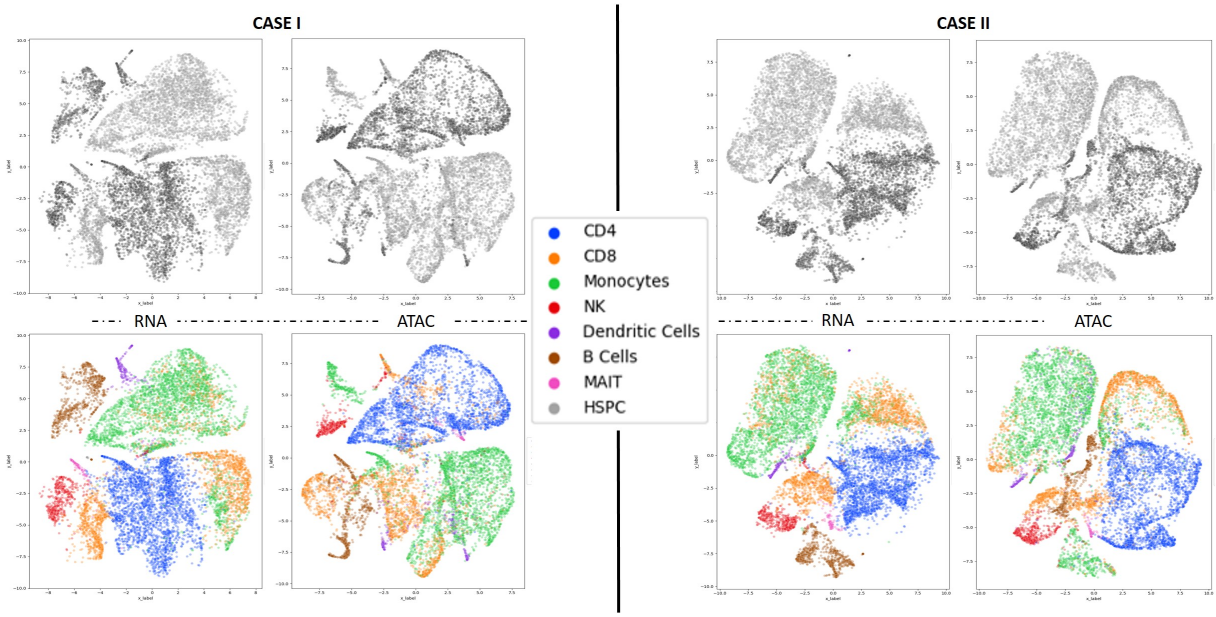

Figure 12: Manifold Alignment with GAN with 2 different weight initialisations - ATAC data is shown after being projected into RNA space; the top (greyscale) and bottom (color) panels show the same plots; both cases demonstrate comparable performance in *superposing* ATAC data onto RNA (see top panel), but vastly different performance in *aligning* the same (bottom panel) - case I reported 6% cell-type matches while case II reported 72% cell-type matches

same architecture and different initialisation seeds post training for 1000 epochs. In the same graph, the corresponding Generator and Discriminator losses have been depicted for these experiments. As can be seen, the losses remain approximately constant at the end of 1000 epochs while the GAN performance displays wild fluctuations. These observations, along with the results shown in table 3 implies that the same GAN with the same architecture and source/target modality pair cannot be relied upon to consistently produce the same results in different runs. In other words, the GAN architecture as discussed so far can be relied upon to only achieve manifold *superposition*, as is evident from the discussion presented here about figure 12. The performance of these experiments in terms of manifold *alignment* varies with every experiment, covering a large spectrum from very good to very bad. Another important observation to be made from table 3 is that performance of any given run does not improve dramatically over the course of 1000 epochs. Therefore, it is reasonable to assume that simply training a GAN for longer epochs will not improve its performance. Therefore, to achieve reasonable and more importantly, consistent performance in manifold *alignment*, this problem needs another breakthrough idea.

6.3 Using topological similarity to guide manifold alignment

As discussed in the last section, extensive experimentation with different model architectures and hyperparameter settings makes it apparent that without giving the Generator Network any metadata information about the modalities, it is unreasonable to expect it to align a cell to its appropriate sub cell/cell-type in the target modality. The Generator is simply being trained to fool the Discriminator. When initialised and trained with this objective in multiple independent runs, learning to superimpose the source data manifold onto the target data manifold is enough to achieve this objective. It is only sometimes that it learns to align the two manifolds as well. Looking at figure 12 makes it clear that when the GAN merely superposes the source data onto target, the *source topology* is destroyed while in cases where the GAN also learns to align the source data, this topology is preserved. For example, the original topology of ATAC data (figure 11) displays three distinct groups of CD8 populations, one of which lies in close vicinity to the Monocytes while the others can be observed as distinct cell-islands. In case I of figure 12, 4 distinct gropus of CD8 cells can be observed - 2 lying side-by-side (bottom left corner), 1 overlapping with CD4 cells (top half center), and another near the Monocytes (bottom half center). By comparison, for case II, which reported 72% cell-type matches (as compared to 6% for case I), we can again observe 3 CD8 groups - one in close vicinity of Monocytes and the others as distinct

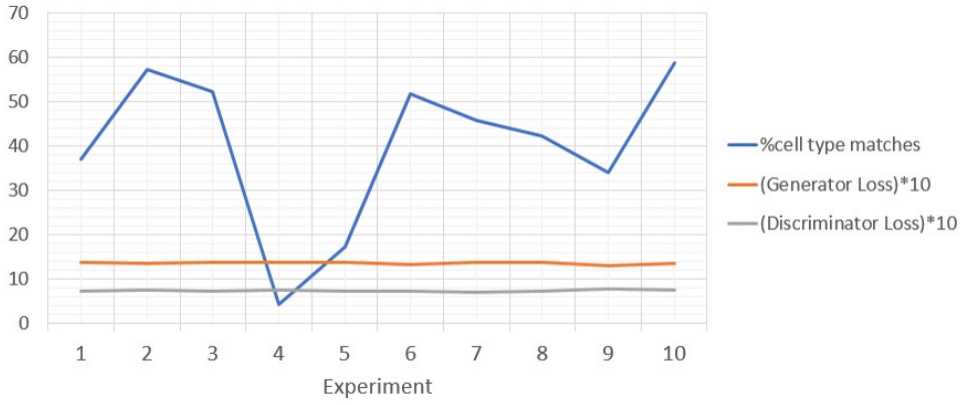

Figure 13: Performance (percent cell-type matches) of GAN compared against Generator and Discriminator losses after training for 1000 epochs. Results for 10 experiments with same architecture and different initialisation seeds have been shown. The losses have been scaled up by a factor of 10 for ease of visualisation.

cell-islands. This makes sense in the context of our second fundamental assumption - *different modalities sampled from the same cellular population have the same underlying manifold.* Because the manifolds of different modalities are already same/similar, a faithful alignment will not distort the existing manifold of the source data, but merely align it on top of the target manifold.

	All Epochs	**Epochs 400 - 1000**
Cell-type matches - Pearson	-0.69 ($1.84 \cdot 10^{-58}$)	-0.73 ($3.23 \cdot 10^{-48}$)
Sub cell-type matches - Pearson	-0.65 ($5.01 \cdot 10^{-49}$)	-0.67 ($1.62 \cdot 10^{-38}$)
Cell-type matches - Spearman	-0.72 ($5.19 \cdot 10^{-64}$)	-0.73 ($2.92 \cdot 10^{-47}$)
Sub cell-type matches - Spearman	-0.71 ($6.3 \cdot 10^{-62}$)	-0.73 ($2.13 \cdot 10^{-47}$)

Table 4: Correlation Analysis of Topological Error with cell-type/sub cell-type matches; for each correlation value, its p-value (displayed in parentheses) indicates the probability of similarly correlated values being generated by a random system

In a manner similar to the Topological Autoencoder, the topological distance between the Persistence Homology calculations of a modality in its actual (obtained as the output of Topological Autoencoder) and projected (obtained as projection by the GAN into source modality space) spaces can be computed. This computation can be used to assess if the topology of original data is destroyed in the process of projecting it into a lower dimensional space. Table 4 demonstrates a negative correlation of the evaluation metrics with this topological error. For case I of figure 12, the topological error was computed to be 1187 while, by comparison, topological error for case II was only 786. To evaluate this correlation, a GAN setup on PBMC data where ATAC was projected into RNA space was used. We performed 50 experiments using identical GAN architectures but different random seeds, training the GAN for 1000 epochs each time. Every 100 epochs, the evaluation metrics (% cell-type matches and % sub cell-type matches) and topological error between the input (ATAC) and output (ATAC projected into RNA space) of Generator was computed according to equation 2. The topological error was calculated for a batch-size of 1000 to strike a balance between coverage of global structure in each batch and compute memory requirements. Finally, the Pearson's and Spearman's correlation coefficients were calculated to evaluate the relationship between the evaluation metrics and the topological error. As is apparent in table 4, the absolute value of Spearman's coefficient is slightly higher than Pearson's, indicating the relationship is monotonic, even if not strictly linear. The p-values indicate the probability of a random system, with no inherent correlation of topological error with the reported metric, generating these values. The extremely low probabilities coupled with the fact that the analysis has been performed on 500 samples indicates that there indeed exists a correlation between the two variables and is not merely a coincidence. Additionally, the absolute values of these correlation coefficients are higher when we only consider epochs 400 and above. This is

not unexpected because a typical loss curve of one of these GAN experiments usually stabilises between epochs 300 and 400 (figure 14). Similar values of these coefficients were obtained with batch-size of 5000, indicating that computing this error for batch-size 1000 is enough. This saves additional computational resources when computing the topological error of alignment.

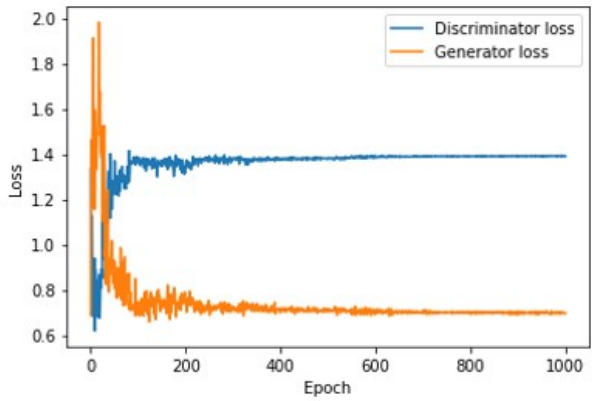

Figure 14: Typical Training Curve of a GAN while aligning PBMC data. Both losses take about 300 epochs to stabilise

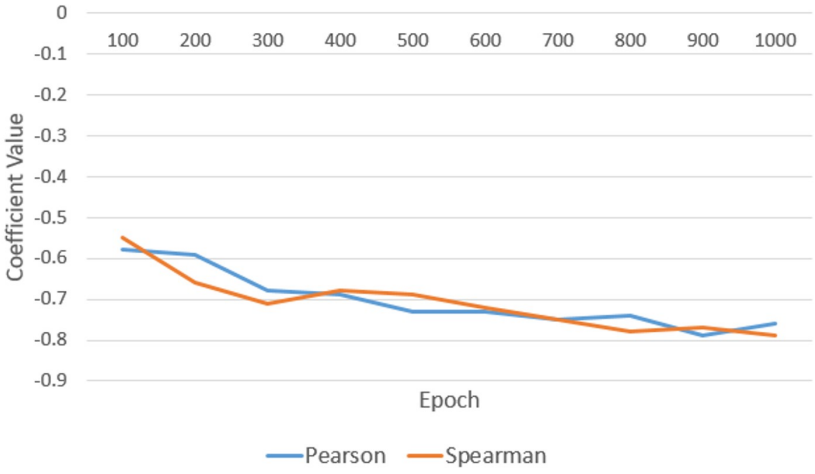

Figure 15: Correlation of Topological Error with cell-type matches, plotted per epoch. The correlation gets progressively stronger until epoch 900 (approximately)

Figure 15 demonstrates correlation calculations for the same set of experiments as in table 4 for each epoch separately. There is a definite increase in the correlation scores with increasing number of epochs. This can be understood in terms of the training curve of a GAN (figure 14) - a Generative Adversarial Network in this setting takes about 400 epochs to stabilise the losses after which it only makes slight improvements in its loss function. As a results, cases where the network also achieves a good alignment of modalities (in addition to minimising the loss function) will take about 400 epochs to do so. These two tables help in arriving at the conclusion that performance of the GAN on the chosen evaluation metrics is correlated with the topological error. Therefore, computing the topological error can be used to guide a GAN setting to align the manifolds instead of just superposing them. Finally, calculation of this topological distance does not require any metadata or cell-label information, which makes it possible to guide manifold alignment using the topological error in the truly unsupervised setting as well.

Epoch	%Cell-type matches (Mean; Std)	%Sub cell-type matches (Mean; Std)
100	66.5; 3.5	45.4; 5.1
200	67.7; 4.0	47.4; 6.3
300	66.9; 3.9	51.0; 4.1
400	66.6; 2.5	52.1; 3.0
500	67.0; 3.4	53.0; 1.9
600	66.7; 1.4	53.9; 1.7
700	67.6; 2.0	54.7; 2.0
800	67.0; 2.4	53.7; 2.1
900	67.0; 2.9	53.9; 3.0
1000	67.5; 2.0	54.5; 2.3

Table 5: Aggregate TopoGAN performance for 1000 epochs (PBMC data). For both metrics, mean and standard deviation for 10 different experiments have been reported. Size of k-neighbourhood is 5

6.4 TopoGAN gives consistent alignments

As previously discussed, without any cell-annotation information, the success with which a GAN *aligns* the manifolds varies with different initial conditions. It was also demonstrated that in a run of 20 experiments, a GAN covers a large spectrum of alignment solutions, ranging from very good to very bad (table 3). As a result, computing the topological error between the original and projected representations of the Generator data (termed *source data*) and using this metric to pick the most successful model should give a second generation GAN which consistently aligns the two manifolds. Table 5 demonstrates the aggregate results from 10 different runs for this strategy. As is immediately clear, the performance is much better and extremely stable against different initialisation seeds now. Additionally, it can also be seen that the performance on either of the metrics (cell-type/sub cell-type matches) does not show any significant improvement beyond 700 epochs, which means the training for the second generation GAN can be stopped earlier, saving additional computational resources. Similar consistency of performance is observed on the partial PBMC data (table 6), demonstrating that TopoGAN is robust to the lack of one-to-one cell correspondences across modalities. The same observation can be made on comparison of performance on CITE-Seq (table 10, Supplementary Material) and partial CITE-Seq (table 11, Supplementary Material) data sets.

Epoch	%Cell-type matches (Mean; Std)	%Sub cell-type matches (Mean; Std)
100	54.3; 6.8	38.9; 4.6
200	57.7; 4.4	41.6; 4.6
300	56; 4.7	42.1; 3.8
400	58.8; 4.2	44.5; 4.5
500	58.8; 3.2	44.1; 2.9
600	59.4; 2.0	45.1; 1.7
700	59.2; 1.7	45.5; 2.5
800	59.4; 1.7	46.3; 2.1
900	59.2; 2.3	46.6; 0.9
1000	59.3; 3.0	47.4; 1.4

Table 6: Aggregate TopoGAN performance for 1000 epochs (Partial PBMC data). For both metrics, mean and standard deviation for 10 different experiments have been reported. Size of k-neighbourhood is 5

6.5 TopoGAN gives better alignments in truly unsupervised settings

Now that it has been demonstrated that the TopoGAN architecture solves the problem of performance fluctuation with a standard GAN, it is time to establish a fair comparison of TopoGAN with state-of-the art methods - UnionCom and MMD-MA. To do so, all three methods were implemented on the data sets discussed in this work. For UnionCom and MMD-MA, the standard implementations were obtained

from the original papers. MMD-MA requires three hyperparameters to be specified by the user. The values recommended by the authors for these hyperparameters was used. For the standard UnionCom implementation, only the epochs for which training must be carried out needs to be specified. This was set at 1000 epochs to ensure the alignment happens within a reasonable compute time. Each method was run several times on each data set to observe consistency in performance. Table 7 compares aggregate results for 10 different runs for TopoGAN, UnionCom, and MMD-MA. While TopoGAN performs better than MMD-MA, UnionCom performs much better on all the reported metrics, with great consistency, on the entire PBMC data. Performance of TopoGAN, in comparison, is consistent, but slightly less spectacular. However, there is an important caveat which must be addressed - the data set used in this particular comparison already contains *perfect one-to-one cell correspondence* across modalities. The kind of data sets we should expect to encounter in real-life, however, will be like the one demonstrated in figure 1.c. It is important to compare any set of methods on these data sets, where one-to-one cell correspondence is completely, or at least partially, missing. All the real-world data sets the UnionCom paper performed its validation on contained perfect one-to-one cell correspondence [9].
As can be seen here, UnionCom performs extremely well on the entire PBMC data set. However, when UnionCom is implemented, with the same hyperparameters, on the partial PBMC data, where one-to-one cell correspondence is imperfect, a sharp drop in performance is observed (table 8) on all reported metrics. Performance of TopoGAN, by comparison does not deteriorate so drastically. Comparing the two methods on the partial CiteSeq data, where one-to-one cell correspondences do not exist *at all* further illustrates this point (table 9) - TopoGAN achieves reasonable alignment performance on the partial CiteSeq data set, reporting scores far better than UnionCom. As discussed in section 5.5.2, in order to make a fair comparison between TopoGAN and benchmark methods, performance has been compared on the metric used in UnionCom as well, namely *label transfer accuracy* [9]. A more granular version of this metric, *sub label transfer accuracy*, has been computed based on the sub cell-type annotations. TopoGAN demonstrates robust performance on all reported metrics across PBMC, Partial PBMC, and Partial CITE-Seq data sets.

Metric	**TopoGAN**	**UnionCom**	**MMD-MA**
Cell-type matches	67.5; 2.1	95.7; 0.5	23.1; 9.5
Sub cell-type matches	54.5; 2.4	90.4; 0.9	12.4; 5.7
Label transfer accuracy	67.5; 2.1	95.7; 0.5	22.8; 9.3
Sub label transfer accuracy	54.6; 2.4	90.4; 0.8	13.9; 6.2

Table 7: TopoGAN comparison with UnionCom and MMD-MA on Full PBMC data; *Mean* and *Standard Deviation*, computed over 10 experiments, are reported; size of k-neighbourhood is 5

Metric	**TopoGAN**	**UnionCom**	**MMD-MA**
Cell-type matches	56.9; 2.5	23.1; 6.6	31.7; 10.9
Sub cell-type matches	36.7; 1.3	13.4; 5.0	14.5; 8.6
Label transfer accuracy	56.9; 2.6	23.4; 7.1	30.2; 9.7
Sub label transfer accuracy	36.8; 1.3	14.3; 5.4	15.3; 9.2

Table 8: TopoGAN comparison with UnionCom and MMD-MA on Partial PBMC data; *Mean* and *Standard Deviation*, computed over 10 experiments, are reported; size of k-neighbourhood is 5

This phenomenon can be seen qualitatively by considering figure 16 - while UnionCom manages to preserve and align the overall structures of RNA and ATAC data in the full PBMC data, it fails to do so in the partial PBMC data, where the one-to-one cell correspondences is partially absent. On the other hand, TopoGAN achieves stable alignments on both the full PBMC data and partial PBMC data. Figure 17 displays ATAC and RNA modalities (after alignment) side-by-side for UnionCom and TopoGAN. Despite the almost perfect alignment achieved on the full PBMC data (figure 17 - a, b),

Metric	TopoGAN	UnionCom	MMD-MA
Cell-type matches	56.6; 0.5	29.1; 11.8	45.9; 10.1
Sub cell-type matches	32.3; 7.7	9.9; 5.2	15.4; 6.68
Label transfer accuracy	56.6; 0.5	28.8; 11.6	44.9; 10.8
Sub label transfer accuracy	32.0; 7.7	10.1; 5.2	16.3; 7.15

Table 9: TopoGAN comparison with UnionCom and MMD-MA on Partial CiteSeq data; *Mean* and *Standard Deviation*, computed over 10 experiments, are reported; size of k-neighbourhood is 5

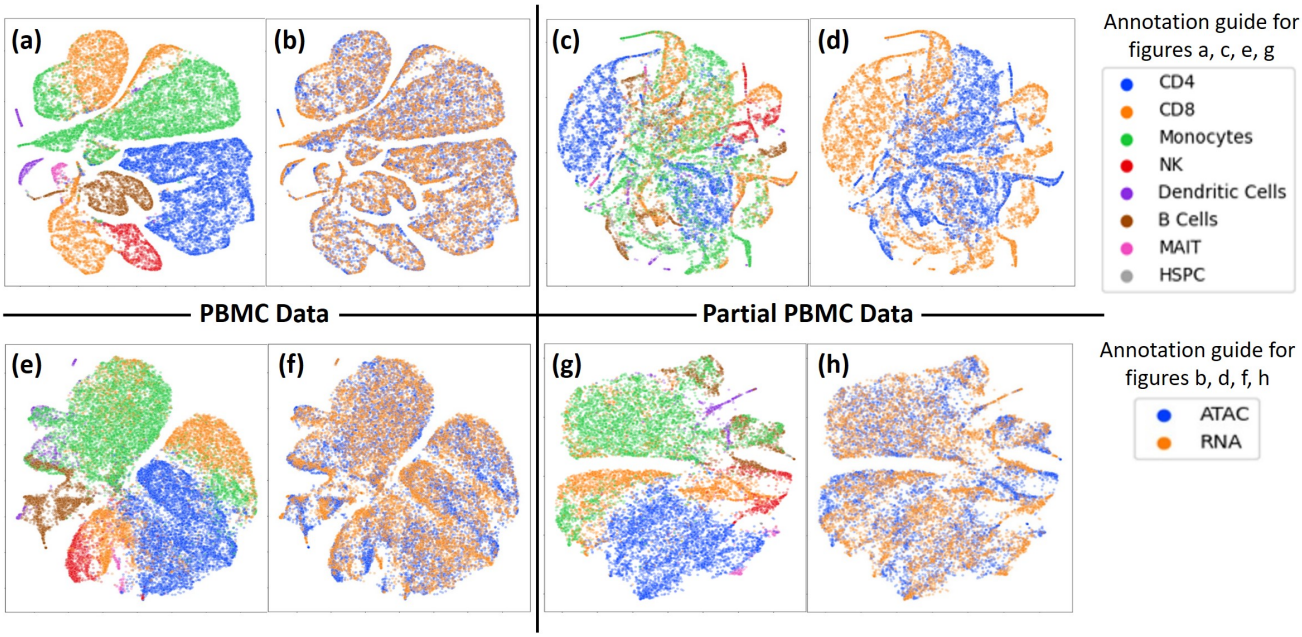

Figure 16: Qualitative comparison of UnionCom and TopoGAN. In each figure, both modalities have been plotted together after being projected in a common space. Top row (figures a, b, c, d) depicts UnionCom results and bottom row (figures e, f, g, h) depict TopoGAN results. Figures a, b, e, f (left half of the figure) are results on the full PBMC data while figures c, d, g, h (right half) are on the partial PBMC data. Figures a, c, e, g are annotated according to cell-type while figures b, d, f, h are annotated according to modality.

UnionCom cannot replicate the alignment quality on partial PBMC ((figure 17 - c, d)). TopoGAN, on the other hand, maintains stable alignment performance on both the full and partial PBMC datasets, despite making certain local errors in alignment ((figure 17 - e, f, g, h)). These arguments are further illustrated on the partial CiteSeq data (figures 19 and 20, Supplementary Material) where UnionCom merely arranges both modalities in the same space without really aligning them while TopoGAN superposes both modalities, and aligns the major cell-types (T cell and Mono/DC with each other) despite a complete lack of one-to-one cell correspondence across modalities.
MMD-MA, by comparison, gives lower performance on all metrics than TopoGAN, for all datasets. Although we used the recommended hyperparameter settings for MMD-MA, it could be possible that retuning of its 3 hyperparameters results in a better performance. However, given the unsupervised nature of our problem statement, such extensive hyperparameter tuning is impractical - cell-label information has not been used to improve the alignment performance of TopoGAN. As a result, tuning the MMD-MA hyperparameters has not been explored in this work.

6.6 TopoGAN is memory frugal

Both UnionCom and MMD-MA require the computation of a cell-cell similarity matrix for each modality for manifold alignment [9, 8]. Because these similarity matrices need to be computed for the *entire* data set, including all the modalities being aligned, these methods scale up very rapidly in terms of memory requirements. TopoGAN, by comparison, relies on the similarity matrix of a *batch* of cells. Additionally,

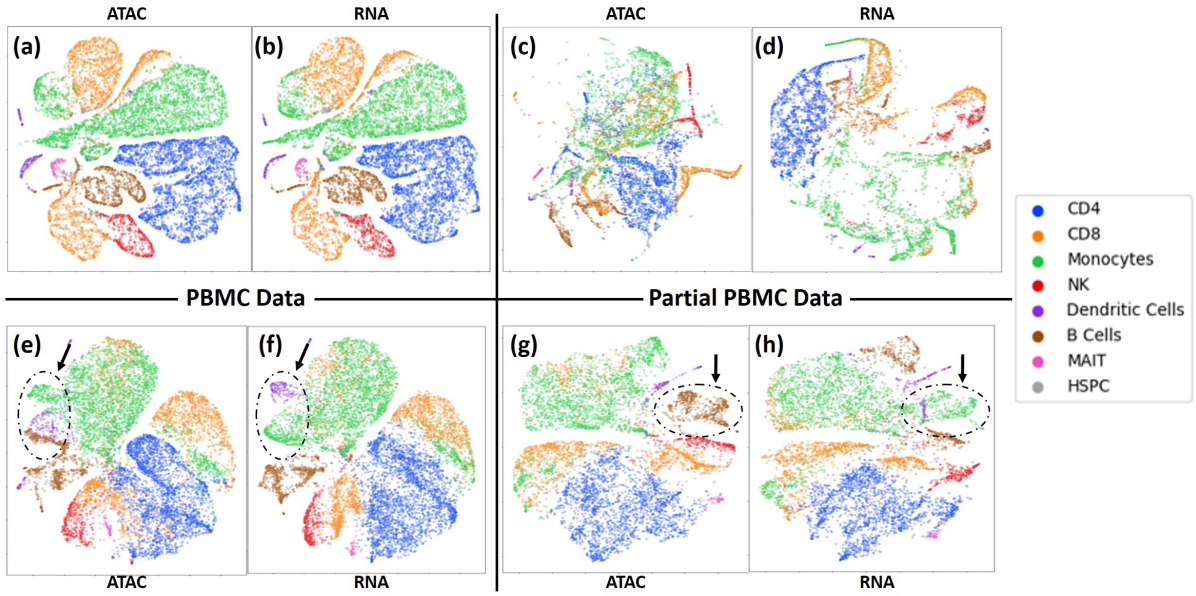

Figure 17: Qualitative comparison of UnionCom and TopoGAN. Each plot represents the indicated modality in the same feature space as the other one. Top row (figures a, b, c, d) depicts UnionCom results and bottom row (figures e, f, g, h) depicts TopoGAN results. Figures a, b, e, f (left half of the figure) are results on the full PBMC data while figures c, d, g, h (right half) are on the partial PBMC data. All figures are annotated according to cell-type. The regions outlined and indicated by an arrow in figures e and f compare local alignment errors by TopoGAN on PBMC data. Local alignment errors by TopoGAN on Partial PBMC data have been similarly pointed out in figures g and h. Doing this for UnionCom on partial PBMC data (figures c and d) is pointless.

this batch is only from the query modality which is being projected into the space of source modality. When this batch size is 1000 (as is the case in this work), it caps the upper limit of memory requirements to a reasonable amount. Figure 18 compares the increase in memory demands of TopoGAN, UnionCom, and MMD-MA with $log_{10}(\#cells)$. All experiments needed to obtain figure 18 were run on the same GPU node on the compute cluster of TU Delft. The reported memory requirements are explicitly for the entire workflow on PBMC data. We start from data in its original, 50-dimensional space. UnionCom and MMD-MA directly use this data to perform the alignment while for TopoGAN, the modalities were first independently projected into 8-dimensions using Topological Autoencoder (the *manifold projection* step). The final objective is to obtain the RNA and ATAC modalities in the same feature space. Once this is achieved, the maximum amount of memory required at any point in this process is noted and reported in figure 18. As can be seen, for data sets with 100 cells, all 3 methods demand less than 1000 MB. However, for data sets with 1000 cells, UnionCom blows up in terms of memory demands. At 10,000 cells, MMD-MA becomes too expensive as well. TopoGAN, however, maintains a stable memory load, owing to its batch-wise approach to the problem - at any given point, TopoGAN considers only a fixed batch of samples to be projected into the space of source modality. The most memory-expensive step in TopoGAN is computation of the topological error. As discussed earlier, computing this error on batches of 1000 samples correlates well with alignment performance. As a result, for any dataset with more than 1000 samples, the memory load of TopoGAN is constant.

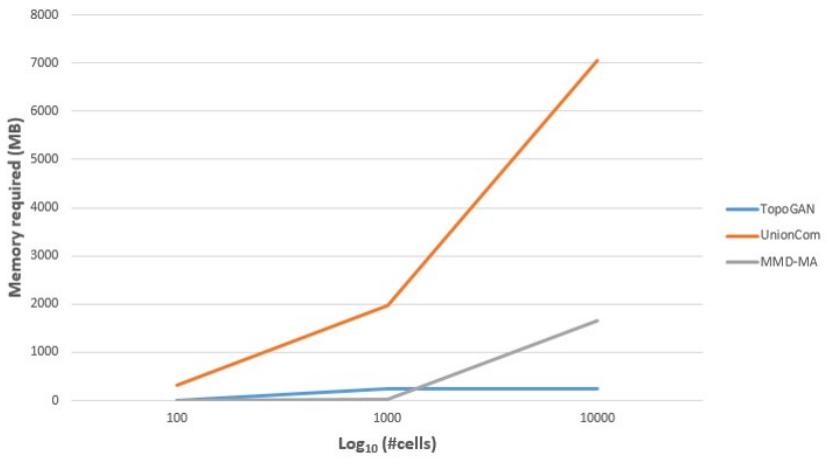

Figure 18: Comparison of memory requirements of TopoGAN, UnionCom, and MMD-MA

7 Discussion

7.1 Why is a multi-omics approach to disease desirable?

The field of single-cell sequencing is in the midst of a data explosion at the moment [23], with measurements of different omic-layers becoming more and more feasible. While any single one of these omic layers provides clues to disease, it tends to be correlation-based - for example, uncharacteristic deviation from the standard methylation levels is often correlated with cancer diagnosis. However, within an intgrated multi-omic paradigm, phenomenon casually associated with disease can become visible, thus motivating targeted treatment strategies [24]. As new assays sample ever larger data sets across modalities, the ability to faithfully align different modalities of a common tissue sample/cellular population will pave the path for cutting-edge scientific research in the field. As has been discussed in the results section, state-of-the-art methods like UnionCom and MMD-MA can not give reliably good performance in cases where perfect one-to-one cell correspondence across modalities does not exist. TopoGAN proposes an alignment strategy which aligns different modalities on a common manifold even when above mentioned correspondence is not available. Additionally, TopoGAN does not require cell-label information during hyperparameter tuning - it only relies on the topological error between the original and projected versions of a given modality. MMD-MA, by contrast relies on the tuning of its hyperparameters which in turn is influenced by computing an error metric using available cell-label information. In the absence of such information, it is not possible to tune MMD-MA [8]. Therefore, TopoGAN can be used in the truly unsupervised settings, where we have multiple modalities with no sample or feature correspondence and no cell-label information either.

7.2 Topological Autoencoder as a tool for dimensionality reduction

As we saw earlier, using a Topological Autoencoder to project a high-dimensional biological data set in a low-dimensional space (the *manifold projection* step of our problem) is a promising new avenue in the field of bioinformatics. For any analysis, the prime objective of dimensionality reduction is preserving the information inherent in the data. In this work, we have referred to this idea as preserving the manifold structure of the data. We discussed in the results section that Topological Autoencoder does a better job of preserving the manifold structure as compared to Variational Autoencoder and UMAP, both qualitatively and quantitatively. Therefore, the isolated idea to use Topological Autoencoder for dimensionality reduction may be adapted in many bioinformatics projects. For instance, DNA Methylation data offers promising epigenetic information to develop models for cancer diagnosis [25]. Leveraging methylation data to develop cancer diagnostic tests is a problem currently chased by both academia and industry. However, methylation data tends to be extremely high dimensional - modern methylation sequencing technologies tend to produce 450,000 - 850,000 features per cell. Needless to say, despite the amount of epigenetic information condensed in these datasets, the number of features per sample needs to be drastically reduced in order to implement any machine learning approach on them. Topological Autoencoder, in combination with preliminary feature selection techniques may offer a promising way to condense relevant information from methylation data to facilitate downstream analyses.

The original Topological Autoencoder paper argues that their method goes further from merely segregating different classes to arrange them in a spatially meaningful manner [16]. They further argue that Topological Autoencoder generates a much more faithful representation of the original manifold structure as compared to methods like PCA, UMAP, or a regular Autoencoder. In the context of our first fundamental assumption (high-dimensional biological data lies on a low-dimensional manifold), this adds weight to the argument that Topological Autoencoder could be a promising dimension reduction method in its own right. One avenue where this idea offers promising results is Trajectory Inference. Trajectory Inference is a technique to arrange cells in an order representative of dynamic cellular processes. This allows one to go beyond a static snapshot of cellular biology, towards the evolution of the biological phenomenon being studied. However, the high-dimensional nature of biological data necessitates dimensionality reduction before the inference of any biological trajectory [26]. Logically, the preservation of inter-cellular relationships in this step is critically important if they indicate an underlying dynamic phenomenon which we wish to infer. These inter-cellular relationships can be represented as the topology of the dataset in question in an n-dimensional space, where each cell represents a point. Topological Autoencoder explicitly tries to preserve this topology in the generated embeddings, and does a better job at it as compared to current approaches like PCA, VAE or UMAP, as argued both in this work and the original Topological Autoencoder paper [16]. Current approaches to infer biological trajectory use PCA

or UMAP as the dimension reduction step [26]. Therefore, it can be expected that using Topological Autoencoder can offer better solutions to trajectory inference problems, especially for datasets with a multitude of different cell-types which usually hint at a complicated underlying manifold.

7.3 Inconsistent Alignments with Generative Adversarial Networks (GAN)

The idea of using Generative Adversarial Networks to perform unsupervised learning tasks has been explored extensively in the research field of Computer Vision. While developing TopoGAN, we frequently sought inspiration from the field of Computer Vision. As discussed in section 6.2, multiple runs of the same GAN settings gives vastly different performance on metrics like percentage cell-type matches. Extensive experimentation with model architectures (varying the number and composition of the hidden layers) and loss functions (Wasserstein Loss, Binary Cross-Entropy) during the initial phases of this thesis revealed that the problem of inconsistent alignments persists despite these changes. This observation indicated that the problem originates not from a flaw in the model setup but instead from the lack of relevant information to *guide* the alignment. Conventional GANs are used for two kinds of task - data generation and data translation. The starting point of data generative tasks tends to be random noise which is then converted into a meaningful pattern, such as an image. It gradually became clear that the ideas mentioned above failed to work because they aim to improve the quality of the training process. In our case, this means that these ideas will help achieve a better *superposition* of modalities which is directly reflected in the final loss of the Generator. As we have already seen, the Generator/Discriminator losses are independent of the metrics we wish to optimise on (e.g. % cell-type matches, figure 13).
Instead, our problem statement aligns more with the data translation direction - we are not trying to generate a modality but instead align an existing one onto a target modality. A popular idea in the unsupervised data translation sub-domain is Conditional GANs [27, 28], where the core idea is to provide some kind of information which gives the Generator Network a *sense of direction* while training. Image-to-image translation, for instance, can be achieved by, among other ideas, training on *(source, target)* image pairs [27] or by explicitly pre-training a Generator network on *(class label, image)* pairs [28]. The message being that GANs usually require additional information while training (hence the name Conditional GAN). Because of the lack of cell-label information or inter-modality correspondence information in our case, implementing Conditional GANs becomes infeasible. The reason for this is that in our case, we have to provide information about the relationship *between* different modalities, either in the form of cell-type labels for samples (which may tell us which local neighbourhood from the target modality to aim for during alignment) or cell-to-cell correspondence across modalities. However, one might argue that by minimising the topological error between the initial and post-GAN representations of a modality, information from the target is being indirectly provided, because of our second assumption (all modalities of a multi-modal dataset lie on the same underlying manifold, if sampled from the same cellular population).

7.4 Limitations and Future Work

The core idea of this work, put in simple words, is this - Unsupervised Manifold Alignment can be achieved by evaluating the topology distortion of a modality when it is projected on a common manifold. Using this idea to select the best network out of a series of GAN experiments and then evaluating the chosen network empirically proves that topological error is a trustworthy guide for a GAN architecture to go beyond manifold superposition towards alignment. However, it is also evident that the quality of alignment eventually achieved is limited by the best alignment any GAN training run achieves in a series of training runs. The current work, presented in this report relies on performing these runs 20 times, which makes it computationally expensive. Additionally, while using topological error computation guarantees that one of the best Generators will be selected for the second generation GAN, there is no guarantee that the GAN will find the best possible solution in those 20 runs of the first generation. Hence, future work in this direction should attempt to encapsulate this idea in a regularisation term which can be applied while training the GAN for the first time itself. This will guide the GAN to minimise the topological error every time it is training, thus eliminating the need to train the GAN multiple times (for the best model to be picked from these runs). Preliminary experiments in this direction suggest that the Generator needs to strike a balance between minimising the topology loss and fooling the Discriminator. Various variations of a GAN setting were implemented which dynamically balanced the focus between these two sub-objectives. However, these experiments were performed with a smaller batch of 50. The correlation between topology error and TopoGAN performance, as explained in section 6.3, breaks down

at such small batch sizes, perhaps because there is not enough coverage of the global topology at this batch scale (topological error is computed only across 50 cells at a time now). Therefore, experimenting with larger batch sizes (for example 1000) might help with developing a topologically regularised GAN for Unsupervised Manifold Alignment.

The datasets which we worked on were a result of simultaneous measurement of multiple-modalities on a cell. Although we specifically removed samples from these datasets to eliminate cell-cell correspondence across modalities, these modalities are still derived from the same experiment. It will be interesting to test TopoGAN in a setting where different modalities were measured from completely different/independent experiments. However, in such cases, we need to consider an important question before proceeding - can the different modalities still be assumed to lie on the same underlying manifold? This becomes an important consideration because TopoGAN works with the assumption that different modalities do lie on the same manifold. Pragmatically speaking, some deviation from this assumption may be expected when it comes to modalities measured independently, simply because different experiments might focus on the same tissue type but not the exact same cellular population (e.g. measuring cells from the same tissue but different donors). It is simply a result of trying to capture real world biology, which tends to be messy. However, the ultimate objective of TopoGAN is to align biological modalities being measured exactly in such situations. Therefore, despite the possible practical challenges, this research direction is an important one in further improvement of TopoGAN.

8 Supplementary Material

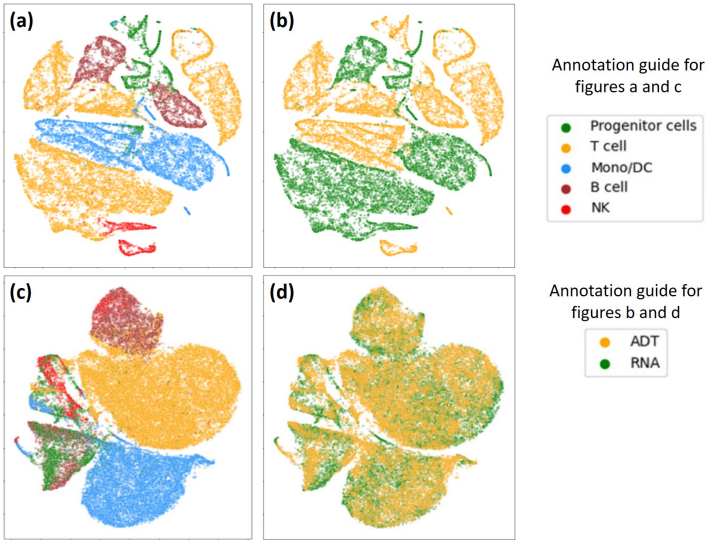

Figure 19: Qualitative comparison of UnionCom and TopoGAN on Partial CiteSeq data. In each figure, both modalities have been plotted together after being projected in a common space. Top row (figures a and b) depicts UnionCom results and bottom row (figures e and f) depict TopoGAN results. Figures a and c are annotated according to cell-type while figures b and d are annotated according to modality.

Epoch	**% Cell-type matches (Mean; Std)**	**% Sub cell-type matches (Mean; Std)**
100	81.0; 1.1	44.2; 3.1
200	80.1; 1.5	42.8; 3.0
300	79.7; 0.7	45.2; 1.7
400	78.7; 1.4	44.8; 2.4
500	78.6; 1.4	45.0; 2.0
600	78.2; 1.4	44.3; 1.7
700	78.1; 1.5	44.2; 1.5
800	77.9; 1.4	44.9; 2.5
900	78.1; 1.4	44.2; 3.4
1000	77.8; 2.0	43.2; 3.4

Table 10: Aggregate TopoGAN performance for 1000 epochs (CITE-Seq data). Reported statistics are mean and standard deviation for both metrics (%cell-type matches and %sub cell-type matches). Size of k-neighbourhood is 5

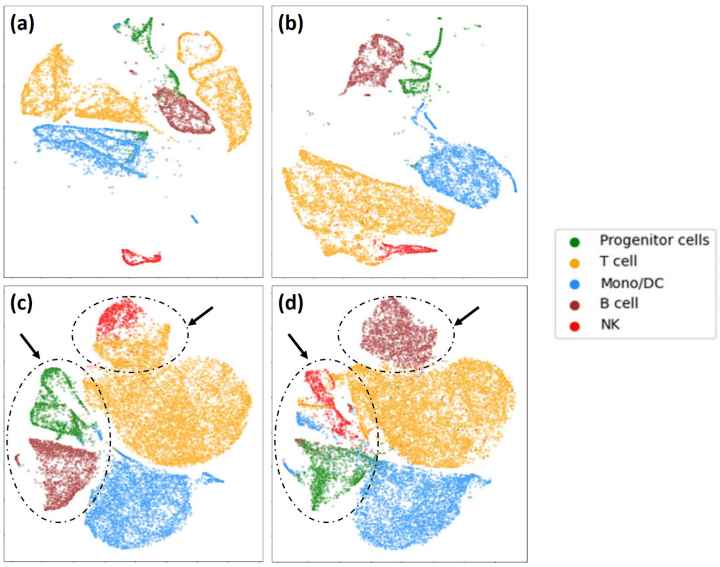

Figure 20: Qualitative comparison of UnionCom and TopoGAN. Each plot represents the indicated modality in the same feature space as the other one. Top row (figures a, b) depicts UnionCom results and bottom row (figures c, d) depicts TopoGAN results. Figures a and c (left half of the figure) are results on the full PBMC data while figures b and d (right half) are on the partial PBMC data. All figures are annotated according to cell-type. The regions outlined and indicated by an arrow in figures c and d compare local alignment errors by TopoGAN. Doing this for UnionCom (figures a and b), however, is pointless.

Epoch	**% Cell-type matches (Mean; Std)**	**% Sub cell-type matches (Mean; Std)**
100	57.5; 4.1	28.4; 4.7
200	58.1; 1.6	31.2; 3.6
300	57.1; 1.4	29.5; 8.8
400	57.8; 1.1	30.7; 6.2
500	57.5; 0.9	32.3; 5.9
600	57.6; 0.9	31.4; 4.5
700	56.8; 0.8	34.8; 6.9
800	56.8; 0.8	34.5; 8.8
900	56.9; 0.6	34.3; 8.2
1000	56.6; 0.4	32.3; 7.3

Table 11: Aggregate TopoGAN performance for 1000 epochs (Partial CITE-Seq data). Reported statistics are mean and standard deviation for both metrics (%cell-type matches and %sub cell-type matches). Size of k-neighbourhood is 5

References

[1] Editorial. "Method of the Year 2013: Single-cell multimodal omics". In: *Nat Methods* 11.30 (Dec. 2013). DOI: doi.org/10.1038/nmeth.2801. URL: https://doi.org/10.1038/nmeth.2801.

[2] Editorial. "Method of the Year 2019: Single-cell multimodal omics". In: *Nat Methods* 17.06 (Jan. 2020). DOI: 10.1038/s41592-019-0703-5. URL: https://doi.org/10.1038/s41592-019-0703-5.

[3] Chenxu Zhu, Sebastian Preissl, and Bing Ren. "Single-cell multimodal omics: the power of many". In: *Nat Methods* 17.01 (Jan. 2020), pp. 11–14. DOI: 10.1038/s41592-019-0691-5. URL: https://doi.org/10.1038/s41592-019-0691-5.

[4] Darren J. Burgess. "Spatial transcriptomics coming of age". In: *Nat Methods* 20.06 (Jan. 2019), pp. 317–317. DOI: 10.1038/s41576-019-0129-z. URL: https://doi.org/10.1038/s41576-019-0129-z.

[5] Alexander F. Schier. "Single-cell biology: beyond the sum of its parts". In: *Nat Methods* 17.01 (Jan. 2020), pp. 17–20. DOI: 10.1038/s41592-019-0693-3. URL: https://doi.org/10.1038/s41592-019-0693-3.

[6] MD Luecken et al. "Benchmarking atlas-level data integration in single-cell genomics". In: *bioRxiv* (2020). DOI: 10.1101/2020.05.22.111161. eprint: https://www.biorxiv.org/content/early/2020/05/23/2020.05.22.111161.full.pdf. URL: https://www.biorxiv.org/content/early/2020/05/23/2020.05.22.111161.

[7] Andrew Butler et al. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". In: *Nature biotechnology* 36 (May 2018). DOI: 10.1038/nbt.4096.

[8] Ritambhara Singh et al. "Unsupervised manifold alignment for single-cell multi-omics data". In: *bioRxiv* (2020). DOI: 10.1101/2020.06.13.149195. eprint: https://www.biorxiv.org/content/early/2020/06/15/2020.06.13.149195.full.pdf. URL: https://www.biorxiv.org/content/early/2020/06/15/2020.06.13.149195.

[9] Kai Cao et al. "Unsupervised topological alignment for single-cell multi-omics integration". In: *Bioinformatics* 36.Supplement$_1$ (July 2020), pp. i48–i56. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa443. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/Supplement_1/i48/33558199/btaa443.pdf. URL: https://doi.org/10.1093/bioinformatics/btaa443.

[10] Stefan G. Stark et al. "SCIM: Universal Single-Cell Matching with Unpaired Feature Sets". In: *bioRxiv* (2020). DOI: 10.1101/2020.06.11.146845. eprint: https://www.biorxiv.org/content/early/2020/06/12/2020.06.11.146845.full.pdf. URL: https://www.biorxiv.org/content/early/2020/06/12/2020.06.11.146845.

[11] Chang Wang and Sridhar Mahadevan. "Manifold Alignment without Correspondence." In: (Jan. 2009), pp. 1273–1278.

[12] Jonathan Bac and Andrei Zinovyev. "Lizard Brain: Tackling Locally Low-Dimensional Yet Globally Complex Organization of Multi-Dimensional Datasets". In: *Frontiers in Neurorobotics* 13 (2020), p. 110. ISSN: 1662-5218. DOI: 10.3389/fnbot.2019.00110. URL: https://www.frontiersin.org/article/10.3389/fnbot.2019.00110.

[13] Xiang-Jun Sun et al. "An integrated analysis of genome-wide DNA methylation and gene expression data in hepatocellular carcinoma". In: *FEBS open bio* 8.7 (July 2018), pp. 1093–1103. ISSN: 2211-5463. DOI: 10.1002/2211-5463.12433. URL: https://europepmc.org/articles/PMC6026698.

[14] Matthew Amodio and Smita Krishnaswamy. "MAGAN: Aligning Biological Manifolds". In: (2018). arXiv: 1803.00385 [cs.CV].

[15] Herbert Edelsbrunner and John Harer. "Persistent homology—a survey". In: *Discrete and Computational Geometry - DCG* 453 (Jan. 2008). DOI: 10.1090/conm/453/08802.

[16] Michael Moor et al. "Topological Autoencoders". In: *CoRR* abs/1906.00722 (2019). arXiv: 1906.00722. URL: http://arxiv.org/abs/1906.00722.

[17] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (2014). arXiv: 1312.6114 [stat.ML].

[18] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML].

[19] Jie Gui et al. “A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications”. In: (2020). arXiv: 2001.06937 [cs.LG].

[20] Alec Radford, Luke Metz, and Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. arXiv: 1511.06434 [cs.LG].

[21] Tim Stuart et al. “Comprehensive integration of single cell data”. In: *bioRxiv* (2018). DOI: 10.1101/460147. eprint: https://www.biorxiv.org/content/early/2018/11/02/460147.full.pdf. URL: https://www.biorxiv.org/content/early/2018/11/02/460147.

[22] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257.

[23] David Laehnemann et al. “Eleven Grand challenges in single-cell data science”. In: *Genome Biology* (2020). DOI: doi.org/10.1186/s13059-020-1926-6.

[24] Yehudit Hasin-Brumshtein, Marcus Seldin, and Aldons Lusis. “Multi-omics Approaches to Disease”. In: *Genome Biology* 18 (May 2017). DOI: 10.1186/s13059-017-1215-1.

[25] Chunlei Zheng and Rong Xu. “Predicting cancer origins with a DNA methylation-based deep neural network model”. In: *bioRxiv* (2019). DOI: 10.1101/860171.

[26] Wouter Saelens et al. “A comparison of single-cell trajectory inference methods: towards more accurate and robust tools”. In: *bioRxiv* (2018). DOI: 10.1101/276907. eprint: https://www.biorxiv.org/content/early/2018/03/05/276907.full.pdf. URL: https://www.biorxiv.org/content/early/2018/03/05/276907.

[27] Phillip Isola et al. *Image-to-Image Translation with Conditional Adversarial Networks*. 2018. arXiv: 1611.07004 [cs.CV].

[28] Hao Dong et al. “Unsupervised Image-to-Image Translation with Generative Adversarial Networks”. In: *CoRR* abs/1701.02676 (2017). arXiv: 1701.02676. URL: http://arxiv.org/abs/1701.02676.