scTopoGAN

unsupervised manifold alignment of single-cell data

Singh, Akash; Biharie, Kirti; Reinders, Marcel J.T.; Mahfouz, Ahmed; Abdelaal, Tamim

DOI
10.1093/bioadv/vbad171

Publication date
2023

Document Version
Final published version

Published in
Bioinformatics Advances

Bioinformatics Advances, 2023, **00**, vbad171
https://doi.org/10.1093/bioadv/vbad171
Advance Access Publication Date: 24 November 2023
Original Article

OXFORD

Gene regulation

scTopoGAN: unsupervised manifold alignment of single-cell data

Akash Singh[1], **Kirti Biharie**[1,2,3], **Marcel J.T. Reinders**[1,2,3], **Ahmed Mahfouz**[1,2,3], **Tamim Abdelaal**[1,2,4,*]

[1]Delft Bioinformatics Lab, Delft University of Technology, 2628 XE Delft, The Netherlands
[2]Leiden Computational Biology Center, Leiden University Medical Center, 2333 ZC Leiden, The Netherlands
[3]Department of Human Genetics, Leiden University Medical Center, 2333 ZC Leiden, The Netherlands
[4]Department of Radiology, Leiden University Medical Center, 2333 ZC Leiden, The Netherlands
*Corresponding author. Department of Radiology, Leiden University Medical Center, 2333 ZC Leiden, The Netherlands. E-mail: t.r.m.abdelaal@lumc.nl

Associate Editor: Magnus Rattray

Abstract

Motivation: Single-cell technologies allow deep characterization of different molecular aspects of cells. Integrating these modalities provides a comprehensive view of cellular identity. Current integration methods rely on overlapping features or cells to link datasets measuring different modalities, limiting their application to experiments where different molecular layers are profiled in different subsets of cells.

Results: We present scTopoGAN, a method for unsupervised manifold alignment of single-cell datasets with non-overlapping cells or features. We use topological autoencoders (topoAE) to obtain latent representations of each modality separately. A topology-guided Generative Adversarial Network then aligns these latent representations into a common space. We show that scTopoGAN outperforms state-of-the-art manifold alignment methods in complete unsupervised settings. Interestingly, the topoAE for individual modalities also showed better performance in preserving the original structure of the data in the low-dimensional representations when compared to other manifold projection methods. Taken together, we show that the concept of topology preservation might be a powerful tool to align multiple single modality datasets, unleashing the potential of multi-omic interpretations of cells.

Availability and implementation: Implementation available on GitHub (https://github.com/AkashCiel/scTopoGAN). All datasets used in this study are publicly available.

1 Introduction

A growing number of single-cell technologies allow the characterization of distinct molecular features of cells, such as single-cell RNA-sequencing (scRNA-seq) or measuring chromatin accessibility at single-cell resolution (scATAC-seq). Despite advances in multi-modal technologies (Zhu *et al.* 2020), these molecular features are mostly measured from different subsets of cells. Sometimes the measured modalities share common features, for example when spatial transcriptomics and scRNA-seq are applied on the same tissue. Because the datasets are not measured from the same cells, they have to be aligned into a common space using data integration methods (Argelaguet *et al.* 2021).

Multi-omic data integration methods aim to find a joint latent space representing information from multiple modalities. These methods include MOFA+ (Argelaguet *et al.* 2020), Seurat WNN (weighted nearest neighbor) (Hao *et al.* 2021), totalVI (Gayoso *et al.* 2021) and Mixture of Experts (Shi *et al.* 2019). These methods require the cell–cell correspondence between the different omics modalities and cannot be applied when multiple unimodal assays are used to profile different cells from the same biological sample. This problem is referred to as diagonal integration, and represents the most challenging case of single-cell multi-omics data integration (Argelaguet *et al.* 2021).

Previous methods, such as MATCHER (Welch *et al.* 2017), SCIM (Stark *et al.* 2020), UnionCom (Cao *et al.* 2020), MMD-MA (Singh *et al.* 2020), SCOT (Demetci *et al.* 2022a), Pamona (Cao *et al.* 2022b), and uniPort (Cao *et al.* 2022a), have addressed this challenging integration task by assuming a similar cellular composition between unimodal datasets collected from the same tissue. MATCHER uses Gaussian processes to embed cells from multiple modalities onto a 1D trajectory. SCIM uses variational autoencoders with an adversarial objective function to learn a modality-invariant latent representation. UnionCom first defines geometrical matches between cells across different modalities and projects the different features onto a common latent representation which is comparable for the matched cells. MMD-MA minimizes the maximum mean discrepancy between the different modalities in the learned latent space. While SCOT, Pamona and uniPort use different variations of an optimal transport formulation to perform the integration. Many of these multi-modal alignment methods, however, suffer from several limitations. MATCHER and SCIM are not fully unsupervised as they require (partial) cell type annotations for each of the different modalities in order to

Received: April 19, 2023; **Revised:** October 30, 2023; **Editorial Decision:** November 19, 2023; **Accepted:** November 23, 2023

align the data. Additionally, MATCHER can only align 1D trajectory structures and cannot deal with more complex structures. SCOT showed promising integration results; however, SCOT was originally tested on small single-cell datasets, which makes its scalability to realistically large single-cell datasets questionable. UnionCom and MMD-MA represent fully unsupervised manifold alignment methods; however, both methods were tested on single-cell multi-omics data, in which the multiple modalities were measured from the same cell, with perfect cell-to-cell correspondences. Although they did not exploit this correspondence in their methods, the integration performance drops significantly (as we show later) when, more realistic, datasets lacking this correspondence are used.

Recently, a new category of diagonal integration methods emerged, including bridge integration (Hao *et al.* 2022), UINMF (Kriebel and Welch 2022), and StabMap (Ghazanfar *et al.* 2023). These methods require an additional single-cell multi-modal dataset containing cell-cell correspondence and measuring the same modalities of the unimodal datasets to be diagonally integrated. This additional multi-modal dataset can act like a bridge and translate information between the distinct unimodal datasets guiding the data integration process. Moreover, GLUE (Cao and Gao 2022) is using a graph-based modeling of the regulatory interactions in order to integrate gene expression and chromatin accessibility data. These methods are beyond the scope of our study since they require an additional representative multi-modal layer.

Considering different single-cell modalities measured from the same biological sample, the main assumption in integration is that the different modalities lie on the same underlying manifold (Sun *et al.* 2018). Preserving the topology of the datasets is crucial when constructing and integrating the different manifolds. Since the different modalities are measuring distinct features, it is necessary to first find a low-dimensional representation of each modality separately. Topological autoencoders (topoAE) have been recently introduced to project high-dimensional data into a low-dimensional latent space while preserving the data topology (Moor *et al.* 2020). Next, these low-dimensional manifolds have to be aligned into a common space with minimal distortion to the original topology of each data modality. Generative Adversarial Networks (GANs) were successfully used in the computer vision field (Gui *et al.* 2021). GANs were previously used to project biological datasets onto each other (Amodio and Krishnaswamy 2018); however, based on correspondence information between the datasets, and not in a fully unsupervised setting.

We propose scTopoGAN, a topology-preserving multi-modal alignment of two single-cell modalities with non-overlapping cells or features. scTopoGAN first finds topology-preserving latent representations of the different modalities, which are then aligned in an unsupervised way using a topology-guided GAN. scTopoGAN is fully unsupervised with no requirement for cell type annotations. Our results show that scTopoGAN outperforms state-of-the-art methods, producing joint representation of distinct datasets with better matching between cellular populations.

2 Methods

2.1 scTopoGAN overview

scTopoGAN is designed to align two datasets measuring two different single-cell modalities, each measured on different non-matching cells. scTopoGAN consists of two steps: (i) manifold projection and (ii) manifold alignment (Fig. 1). Assuming a lower-dimensional manifold structure for single-cell datasets (Bac and Zinovyev 2019), scTopoGAN first finds the manifold for each modality separately, with explicit preservation of the data topology. Then, the latent space representation of the two modalities is aligned in a topology-preserving manner, exploiting the assumption that the topology of the cells in the two modalities is the same. This alignment step should preserve relevant inter-modality correspondence such that similar cell types should be aligned between different modalities.

2.1.1 Manifold projection

To project each modality to a lower-dimensional latent space, scTopoGAN uses a topoAE (Moor *et al.* 2020), which chooses point-pairs that are crucial in defining the topology of the manifold instead of trying to optimize all possible point-pairs. A topoAE is based on the concept of persistent homology (Edelsbrunner and Harer 2008) which selectively considers edges connecting point-pairs below a certain distance threshold. These edges are used to construct local neighborhoods together constituting large-scale topological features. Similar to the Mapper method (Singh *et al.* 2007) used in scTDA (Rizvi *et al.* 2017), persistent homology can identify simple topological features, such as connected components/tree-like structures, as well as higher-order structures such as cycles and holes/voids. Repeating the above procedure by increasing the distance threshold, persistent topological features are defined as the topological structures which are preserved and observed in the data over a wide range of distance thresholds. The point-pairs constituting these persistent features are known as persistent pairings. Preserving the distances between these pairings in a lower-dimensional projection of the data preserves the data topology. The loss function of the topoAE is defined as:

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

where L_r is the reconstruction loss between the input and reconstructed output of the autoencoder across all cells, and L_t represents the topological loss, while λ is the weight of the topological loss. The topological loss is defined as:

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

where X is the original input data and Z is the encoded latent representation, A^X and A^Z are the distance matrices in the original and latent spaces respectively, π^X and π^Z are the persistent pairings in the original and latent spaces, respectively. $A^{\cdot}[\pi^{\cdot}]$ represent subset of distances in the space $A^{\cdot}$ defined by the topologically relevant edges in that space $\pi^{\cdot}$. The term L_{XZ} ensures that persistent pairings relevant to the original manifold are equidistant in both the original and the latent spaces, while L_{ZX} ensures that persistent pairings relevant to the latent manifold are equidistant in both spaces.

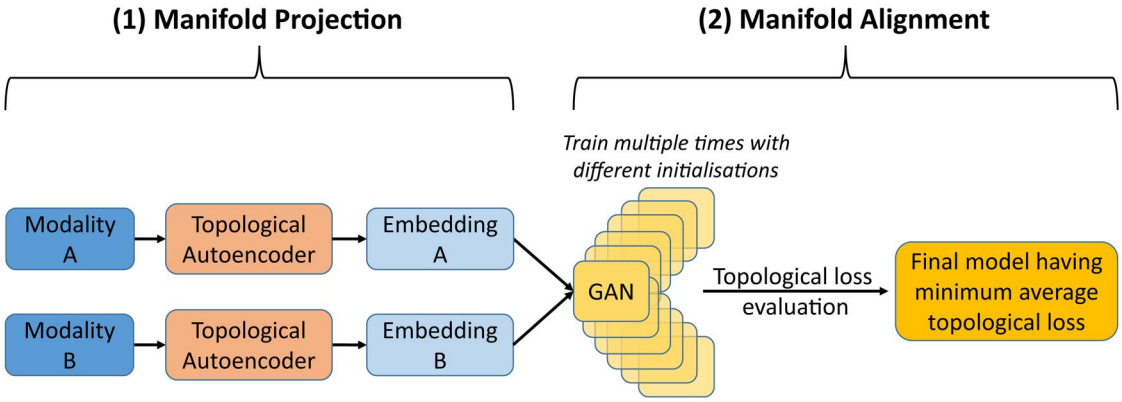

Figure 1. scTopoGAN overview. scTopoGAN consists of two stages: (1) Manifold projection using topological autoencoders to obtain a low-dimensional embedding (manifold) for each modality independently. (2) Manifold alignment using a GAN. Hereto, 20 different GAN models are trained with random initializations. Then that model is selected which has the minimum average topological loss. The selected model is further trained for 1000 additional epochs to produce the final alignment.

2.1.2 Manifold alignment

We used a GAN (Goodfellow *et al.* 2014) to align one modality (source) to the other modality (target). The generator part of the GAN aims to project the source modality onto the target modality, resulting in a combined dataset. We use a single hidden layer generator network against a double hidden-layer discriminator network. The GAN was trained for 1000 epochs.

To ensure a topology-preserving alignment of the two modalities, we trained 20 different GANs and selected the GAN which best preserves the topological loss (Equation (2)) between the source data and its projection in the target data space. To do so, for each GAN, the topological loss is calculated from epoch 500 every 100th epoch until epoch 1000 (six values) and then averaged. The topological loss was calculated for a batch size of 1000 to balance between coverage of global structure in each batch and compute memory requirements. The generator network of the selected GAN is then loaded into a new GAN model with a new discriminator network as its adversary. This final model is then trained for an additional 1000 epochs to obtain the final aligned manifolds.

2.2 Datasets

2.2.1 Peripheral blood mononuclear cells (PBMC) dataset

The peripheral blood mononuclear cells (PBMC) dataset consists of healthy human PBMCs, simultaneously profiling gene expression (RNA) and chromatin accessibility (ATAC) from the same cells using the 10x multiome protocol. The dataset was downloaded from the 10x Genomics website (https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k). The "Full PBMC" dataset contained 11 910 cells, profiling 36 601 genes and 108 377 peaks, having one-to-one correspondence between the two modalities, and including 7 major cell classes (CD4 T cells, CD8 T cells, Monocytes, NK cells, Dendritic cells, B cells, and HSPC) which are further divided into 20 cell subclasses.

To simulate a realistic data in which the cell–cell correspondences do not exist between the two modalities, we generated the "Partial PBMC" dataset where we randomly removed 30% of the cells (after preprocessing) from both RNA and ATAC independently stratified across the different cell classes. This results in a total of 7349 cells for each modality, including 2100 cells which have no corresponding cells in the other modality. In this case, the "Partial PBMC" dataset represents an example with partial correspondence.

2.2.2 Bone marrow (BM) dataset

The original bone marrow (BM) dataset consists of human bone marrow cells, simultaneously profiling gene expression (RNA) and protein expression (antibody-derived tags, ADT) from the same cells using the CITE-seq protocol (Stoeckius *et al.* 2017). The dataset contained 30 672 cells, profiling 17 009 genes and 25 ADT, with cell-cell correspondence between both modalities, and including 5 major cell classes (T cells, B cells, Mono/DC cells, NK cells, and Progenitor cells), further categorized into 27 different cell subclasses. For the purpose of our study, we randomly selected 10 235 cells from each modality independently in a stratified manner across the cell classes. In this case, the "BM" dataset does not contain any cell–cell correspondences between the two modalities.

2.2.3 scGEM dataset

The scGEM dataset measures the continuous differentiation trajectory of human fibroblast reprogramming to induced pluripotent stem cells (iPS), simultaneously profiling gene expression (RNA) and DNA methylation (MET) from the same cells (Cheow *et al.* 2016). The scGEM dataset was also used for evaluation by UnionCom and SCOT. The dataset contained 177 cells, measuring 34 genes and 27 methylation states, and is composed of five developing cell states starting from fibroblasts (BJ), going through three intermediate states (d8, d16T+, d24T+), and finally differentiating into iPS.

2.3 Data preprocessing

We performed all data preprocessing using the Seurat v4.0 R package (Hao *et al.* 2021). For the PBMC dataset, we filtered out cells with RNA count below 1000 or above 25 000, cells with ATAC count below 5000 or above 70 000, and cells with mitochondrial percentage above 20%, resulting in a total of 10 412 cells. Further, the RNA modality is normalized using `SCTransform` (Hafemeister and Satija 2019), selecting the top 3000 variable genes. The ATAC modality was normalized using the `RunTFIDF` function using a scaling factor of 10 000, followed by finding the top peaks using the `FindTopFeatures` function with min.cutoff = q0. Next, we reduced the dimensionality of the RNA and ATAC data to 50 dimensions using principal component analysis (PCA) and latent semantic indexing

(LSI), respectively. These 50-dimensional datasets are used as input to the scTopoGAN workflow.

For the BM dataset, the RNA modality was normalized using a scaling factor of 10 000 followed by log-transformation. The top 2000 variable genes were selected, next the data was scaled and centered. The ADT modality was centered log-ratio (CLR) normalized, scaled and centered. The RNA data was reduced to 50 dimensions using PCA, while dimensionality reduction was not necessary for the ADT modality which only had 25 features. The scGEM dataset was obtained from (Demetci *et al.* 2022a) in a preprocessed form, no further preprocessing was applied.

2.4 Benchmarking methods

For the manifold projection step, we compared the performance of the topoAE with a standard variational autoencoder (VAE) (Kingma and Welling 2014), which is used for manifold projection in SCIM (Stark *et al.* 2020), and a regular autoencoder (AE). Further, we used uniform manifold approximation and projection (UMAP) (McInnes *et al.* 2018) as a base-line for the manifold projection evaluation. Next, we compared the alignment performance of scTopoGAN with the state-of-the-art methods UnionCom, MMD-MA, and SCOT.

2.5 Downstream analysis

We tested downstream analysis tasks performed on the integrated data. First, we performed unsupervised clustering to reconstruct the original cell classes from each dataset. We used the Leiden graph-based clustering (Traag *et al.* 2019) and adjusted the resolution to match the number of clusters to the number of cell classes per dataset. Second, we performed cross-modality prediction using the "Full PBMC" and the "BM" datasets. When predicting ATAC from RNA ("Full PBMC" dataset), for each RNA cell we define the 50-nearest neighboring ATAC cells and perform a weighted nearest-neighbor regression, previously described in (Abdelaal *et al.* 2020), to predict the corresponding ATAC expression. Similarly, we predicted RNA from ATAC ("Full PBMC" dataset), ADT from RNA and RNA from ADT ("BM" dataset).

2.6 Evaluation metrics

To evaluated the manifold projection, we used the Silhouette score (Rousseeuw 1987) which assesses the separation between the cell classes. The Silhouette score ranges from −1 to 1, where a higher value indicates better separated classes. Additionally, we calculated the Kullback–Leibler divergence KL_σ between the density estimates of the input data and its latent space representation (Moor *et al.* 2020). The KL_σ calculation requires the pairwise distance matrices of the original input data and its latent representation. Gaussian kernel of size σ (we used $\sigma = [0.01, 0.1, 1]$) is applied for each point to estimate its density based on the distances to other points. The KL_σ value quantifies the dissimilarity between the density estimates in both spaces (input and latent), thus lower values (≈ 0) indicate better manifold projection performance.

To evaluate the manifold alignment, for each cell in one modality, we determine its k-neighboring cells from the other modality in the final aligned common space ($k = 5$, Euclidean distance). Next, we compare the class/subclass annotation of that cell with the majority vote of its neighbors and check whether it is a match or not. We report the percentage of cells with matching cell class/subclass denoted as the Celltype matching and the Subcelltype matching scores, respectively. Additionally, to evaluate the ability of the methods to mix the different modalities, we calculated the Local Inverse Simpson's Index (LISI) using the batch (dataset) identifiers (Korsunsky *et al.* 2019), and report the average batch LISI score calculated over 10 different runs.

Finally, to evaluate the downstream analysis tasks, we used the Adjusted Rand Index (ARI) to evaluate the clustering assignment obtained in comparison with the cell class labels. For the cross-modality prediction task, we calculated the Spearman correlation between the original measured and predicted features.

2.7 Implementation details

To train the topoAE, we used a learning rate of 1e−03, batch size of 50, latent size of 8 dimensions, and an architecture of two hidden layers followed by batch normalization and rectified linear unit (ReLU) activation. For the hyperparameter λ, we tested values ranging from 0.5 to 3.0 as recommended by (Moor *et al.* 2020). The hidden layers are both of size 32, except for the "BM" ADT data (input dimensions = 25), the size of the hidden layers is 16. We used the same architecture for VAE and AE. For all autoencoder models, we split the data into 80% training and 20% validation. We trained the models for a minimum of 50 epochs and a maximum of 200 epochs, with an early stop if the validation loss did not improve for 10 consecutive epochs (after the initial 50 epochs).

For the GAN model, we used a generator hidden layer of size 30 and a discriminator hidden layers of sizes 60 and 30, batch size of 512 and learning rates of 1e−03 and 1e−02 for the generator and discriminator, respectively. We followed previous work using GANs to stabilize the training process (Radford *et al.* 2016) by sampling initialization weights from a normal distribution N(0,0.02), and using a Leaky ReLU as the activation function for the discriminator with an activation value of 0.2, while using ReLU activation for the generator. UnionCom and MMD-MA were trained for 1000 epochs using default hyperparameters settings. We tested the following learning rate values [1e−2, 1e−3, 1e−4, 1e−5] for UnionCom, and [1e−3, 1e−4, 1e−5, 1e−6, 1e−7] for MMD-MA. We reported the best result obtained for each dataset. For SCOT, we used the recent version SCOTv2 (Demetci *et al.* 2022b), and tested the following hyperparameters settings and reported the best result obtained for each dataset, $\epsilon = [0.001, 0.01, 0.1]$ and $\rho = [0.01, 0.1, 1.0]$. For all models, input datasets were randomly shuffled to ensure cell–cell correspondence is not implicitly provided to the model. The PyTorch version 1.7 was used in all experiments.

3 Results

3.1 Topological autoencoder produced better manifold projections compared to other methods

Before integrating different data modalities, it is crucial to acquire a proper low-dimensional embedding of each modality separately. For this manifold projection task, we used a topoAE which has been shown to produce reliable topology approximations (Moor *et al.* 2020). To the best of our knowledge, topoAEs have not been applied on biological datasets which, compared to classical datasets used in machine learning, contain continuous topological structures (Rizvi *et al.* 2017). Using both RNA and ATAC modalities of the "Full PBMC" dataset, and both RNA and ADT of the

Table 1. Manifold projection evaluation results.[a]

Dataset	Method	Silhouette score	$KL_{0.01}$
PBMC RNA	topoAE ($\lambda = 0.5$)	0.272 ± 0.014	0.059 ± 0.007
	topoAE ($\lambda = 1.0$)	0.267 ± 0.017	0.051 ± 0.005
	topoAE ($\lambda = 2.0$)	0.277 ± 0.013	0.045 ± 0.007
	topoAE ($\lambda = 3.0$)	0.274 ± 0.006	**0.044 ± 0.004**
	VAE	0.211 ± 0.008	0.074 ± 0.014
	AE	0.263 ± 0.018	0.081 ± 0.012
	UMAP (8 dimensions)	**0.356**	0.290
	UMAP (2 dimensions)	0.319	0.284
PBMC ATAC	topoAE ($\lambda = 0.5$)	0.095 ± 0.015	0.075 ± 0.042
	topoAE ($\lambda = 1.0$)	0.071 ± 0.010	0.090 ± 0.045
	topoAE ($\lambda = 2.0$)	0.053 ± 0.004	0.123 ± 0.043
	topoAE ($\lambda = 3.0$)	0.047 ± 0.013	0.063 ± 0.031
	VAE	0.150 ± 0.034	**0.031 ± 0.014**
	AE	0.228 ± 0.019	0.080 ± 0.026
	UMAP (8 dimensions)	0.318	0.182
	UMAP (2 dimensions)	**0.336**	0.164
BM RNA	topoAE ($\lambda = 0.5$)	0.431 ± 0.024	0.036 ± 0.006
	topoAE ($\lambda = 1.0$)	0.396 ± 0.011	0.024 ± 0.003
	topoAE ($\lambda = 2.0$)	0.391 ± 0.012	0.018 ± 0.003
	topoAE ($\lambda = 3.0$)	0.375 ± 0.009	**0.017 ± 0.003**
	VAE	0.301 ± 0.023	0.034 ± 0.006
	AE	0.455 ± 0.008	0.055 ± 0.005
	UMAP (8 dimensions)	**0.541**	0.064
	UMAP (2 dimensions)	0.510	0.063
BM ADT	topoAE ($\lambda = 0.5$)	0.360 ± 0.012	0.047 ± 0.003
	topoAE ($\lambda = 1.0$)	0.357 ± 0.006	0.044 ± 0.007
	topoAE ($\lambda = 2.0$)	0.347 ± 0.005	0.044 ± 0.004
	topoAE ($\lambda = 3.0$)	0.351 ± 0.003	**0.040 ± 0.004**
	VAE	0.272 ± 0.022	0.124 ± 0.009
	AE	0.405 ± 0.026	0.146 ± 0.037
	UMAP (8 dimensions)	**0.427**	0.178
	UMAP (2 dimensions)	0.291	0.213

[a] Reported results for different AE models are computed over 10 different runs (mean ± std). Bold values indicate the best method for each dataset.

"BM" dataset, we compared the manifold projection performance of the topoAE with three other methods (Table 1). All methods were used to reduce the 50-dimensional (PCA or LSI) data or the 25-dimensional ADT data to 8 dimensions, additionally UMAP was used to produce 2-dimensional embedding for visualization purpose. Furthermore, we tested different settings for the topological loss weight λ of the topoAE. Results show that topoAE is the best method in preserving the original data density estimates having overall the lowest $KL_{0.01}$ value, except for the PBMC ATAC data, where VAE is performing better in terms of density preservation. Similar conclusion can be obtained when using larger σ values of 0.1 and 1. However, UMAP obtained the highest Silhouette score producing better separation between different cell classes across all datasets.

Further, to qualitatively compare the low-dimensional manifolds produced by each method, we generated two-dimensional UMAP embeddings of the 8-dimensional manifolds of the topoAE, VAE and AE, in comparison with the 2-dimensional UMAP embeddings (Supplementary Fig. S1). Overall, all methods obtained similar maps with good separation between the cell types, however, VAE could not group the CD8 populations for the PBMC data separate from the CD4 T cells. Taken together, topoAE showed better performance in producing low-dimensional manifolds preserving the original density of the data, with comparable performance in terms of cell type separation compared to other autoencoder models.

3.2 Minimum topological loss ensured manifold alignment instead of superposition

After obtaining the lower-dimensional manifold of each modality using topoAEs, these manifolds are integrated into one common space. We applied the scTopoGAN manifold alignment on the "Full PBMC" dataset, aligning the ATAC modality (source) to the RNA space (target). We observed an inconsistency in the alignment performance when training multiple GANs initialized with different weights. Although their different losses were more or less equal, the Celltype matching score (see Section 2.6) of 40 different GANs was 37.1 ± 16.5% (mean ± standard deviation). We visualized the resulting alignments for the best and the worst models (Fig. 2). Both GANs achieve a good superposition of the ATAC manifold onto the RNA manifold, aligning the ATAC data to match the shape of the RNA data. However, the worst GAN produced a poor alignment of cell classes, e.g. projecting T cells to Monocytes (Fig. 2A and B). Whereas, the best GAN correctly aligns most cell classes (Fig. 2C and D).

To quantify how distorted the source manifold is after projection to the target space, we inspected the topological loss between the source data and the projected source. Using the "Full PBMC" dataset, we performed 40 experiments using identical GAN architectures but different random initializations, trained for 1000 epochs. GAN training showed that the generator and discriminator losses stabilized around 400 epochs. Therefore, we calculated the topological loss from epoch 500 to 1000 every 100 epochs, between the input ATAC (source) manifold and the output ATAC manifold projected onto the RNA space. To interpret the topological losses, we correlated them with the Celltype and the Subcelltype matching scores at the same epochs, resulting in a negative Pearson correlation of −0.73 and −0.67, respectively. This negative correlation indicates that GANs with low topological loss (i.e. preserving the topology of the source data after alignment) tend to produce better manifold alignment. This observation promoted us to train 20 different GANs and select the model with the minimum average topological loss (see Section 2.1.2) as the final scTopoGAN model. We chose to train 20 base GANs as that showed to cover a wide range of alignment scores.

3.3 scTopoGAN outperforms state-of-the-art methods

We benchmarked scTopoGAN against UnionCom, MMD-MA and SCOT. First, we tested the four methods using all three datasets ("Full PBMC," "Partial PBMC," and "BM"), and evaluated the results using the Celltype and Subcelltype matching scores (Table 2). Shuffling the input data and random sampling of the batches in each iteration (in the case of scTopoGAN) ensured that the cell–cell correspondence is not implicitly captured by any of the methods. scTopoGAN outperformed all other methods based on both scores for the "Full PBMC" and "Partial PBMC" datasets, producing joint embeddings with better cell type separation. However, SCOT produced the highest Celltype and Subcelltype matching scores for the "BM" dataset, followed by comparable performance between scTopoGAN and UnionCom, while MMD-MA ranked last.

Further, we qualitatively compared the performance of scTopoGAN against other methods in order to interpret their performances. We visualized the final alignment results for

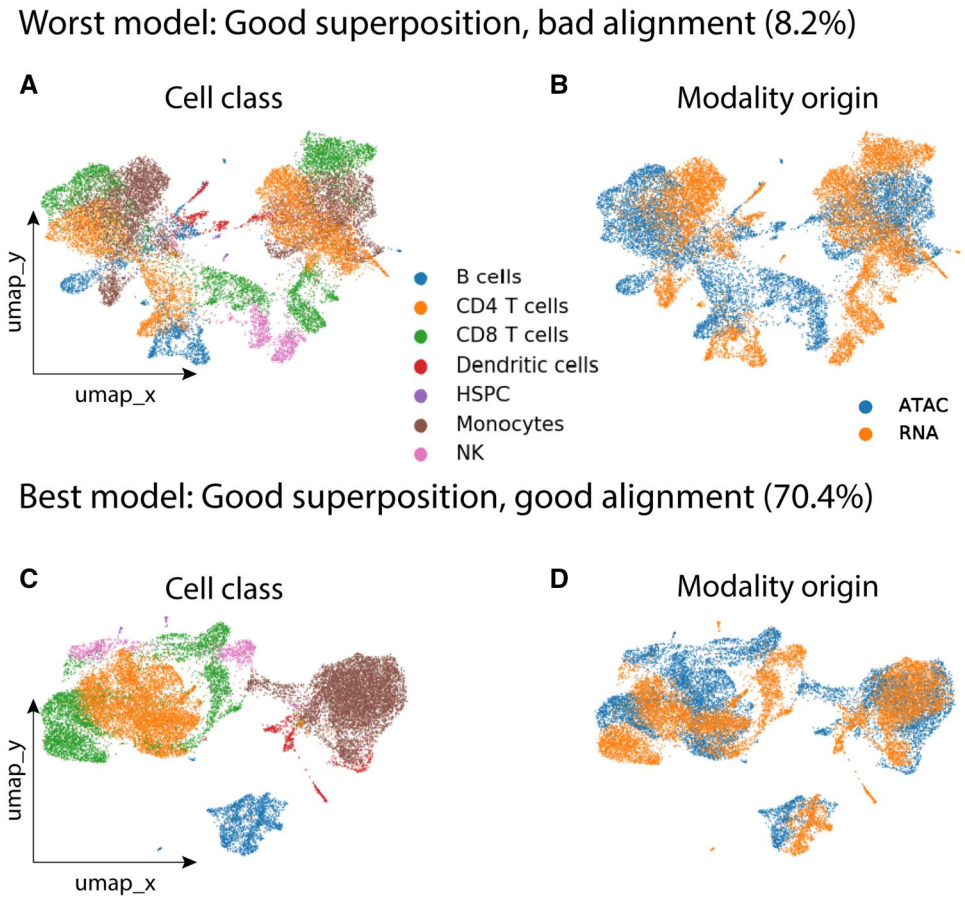

Figure 2. GAN alignment versus superposition. Plots show UMAP embeddings of the final alignment of two GAN models, showing (A, B) worst GAN model performing good superposition of the two manifolds, but bad alignment in terms of cell classes (8.2% Celltype matching score), and (C, D) best GAN model with good alignment projecting the correct cell classes across the two modalities (70.4% Celltype matching score). Each UMAP is plotted twice, once colored with the cell classes showing how well different cell classes are separated, and once colored with the modality of origin showing how well different modalities are mixed.

Table 2. Benchmarking scTopoGAN against UnionCom, MMD-MA and SCOT.[a]

Dataset	Method	Celltype matching (mean ± std) %	Subcelltype matching (mean ± std) %	Batch LISI score
Full PBMC	scTopoGAN	**61.7 ± 8.6**	**41.3 ± 6.5**	**1.49**
	UnionCom	34.8 ± 10.9	22.9 ± 7.2	1.10
	MMD-MA	28.3 ± 6.4	10.5 ± 4.8	1.01
	SCOT	18.9 ± 1.1	2.4 ± 0.2	1.01
Partial PBMC	scTopoGAN	**72.5 ± 5.1**	**45.1 ± 1.8**	**1.49**
	UnionCom	30.2 ± 7.7	15.5 ± 6.4	1.08
	MMD-MA	30.1 ± 7.4	8.6 ± 7.7	1.00
	SCOT	13.0 ± 0.8	2.6 ± 0.2	1.01
BM	scTopoGAN	50.9 ± 14.7	22.5 ± 5.4	**1.41**
	UnionCom	51.8 ± 3.7	20.9 ± 2.6	1.07
	MMD-MA	38.8 ± 17.9	10.4 ± 8.4	1.01
	SCOT	**90.5 ± 0.0**	**31.6 ± 0.0**	1.02
scGEM	scTopoGAN	58.8 ± 0.0		1.4
	UnionCom	51.1 ± 0.9		**1.8**
	MMD-MA	32.2 ± 19.3		1.4
	SCOT	**59.6 ± 1.3**		1.7

[a] Reported results are computed over 10 different runs. Bold values indicate the best method for each dataset.

all three datasets (Fig. 3). For the "Full PBMC" and "Partial PBMC" datasets, scTopoGAN showed better mixing of the RNA and ATAC modalities compared to UnionCom, MMD-MA, and SCOT, while keeping the cell classes separable (Fig. 3A and B). The "BM" dataset is more challenging to correctly match the smaller cell classes; however, scTopoGAN produced better alignment and mixing of the RNA and ADT modalities compared to other methods (Fig. 3C). Next, we quantitatively evaluate the mixing of the modalities using the batch LISI score. In line with the qualitative assessment observed in the UMAP (Fig. 3A–C), results show that scTopoGAN had the highest batch LISI score

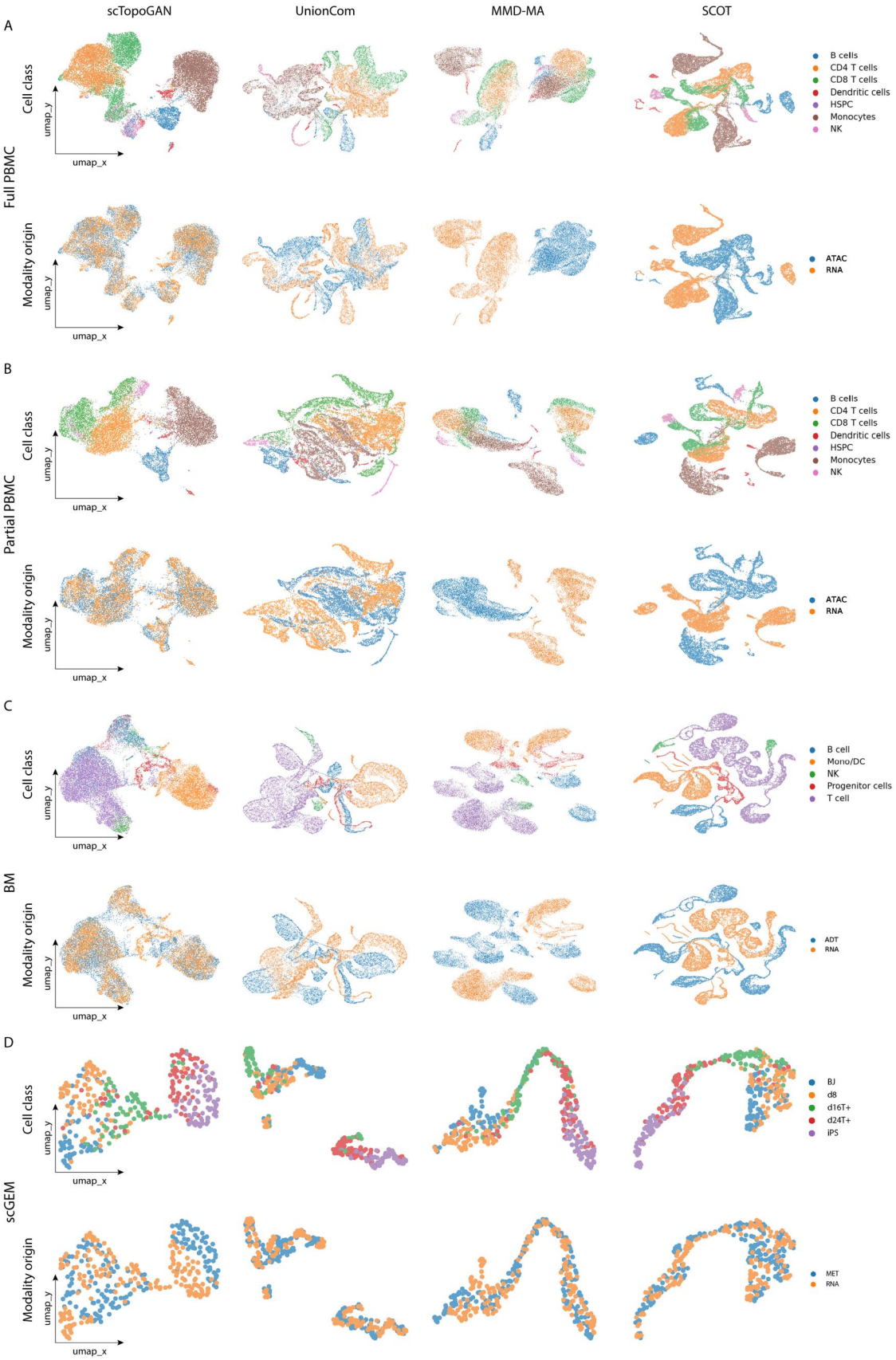

Figure 3. Qualitative comparison of scTopoGAN, UnionCom, MMD-MA, and SCOT. Plots show UMAP embeddings of the final alignment produced by each method when applied on (A) Full PBMC dataset, (B) Partial PBMC dataset, (C) BM dataset, and (D) scGEM dataset. Each UMAP is plotted twice, once colored with the cell classes showing how well different cell classes are separated, and once colored with the modality of origin showing how well different modalities are mixed.

(optimal value is 2) across all three datasets compared to other methods (Table 2), which shows overall superior integration performance for scTopoGAN.

Additionally, we quantified the computational complexity of the different methods. In terms of memory requirements, scTopoGAN memory requirement is almost constant with the number of cells, due to the fixed batch size. For the three datasets ("Full PBMC," "Partial PBMC," and "BM"), scTopoGAN is the least memory-demanding method (Supplementary Table S1), while SCOT was the fastest method in terms of computation time.

In order to test the robustness of scTopoGAN when presented with more realistic integration scenarios, we performed the following downsampling experiments using the "Full PBMC" dataset: (i) First, in a more realistic setup, the two modalities will have different number of cells. We evaluated the effect of changing the total number of cells across modalities, where we kept the RNA modality fixed and tested the integration performance of scTopoGAN when aligning 90%, 80%, …, 10% of the ATAC modality, downsampled in stratified manner across cell types, thus keeping the cell type proportions similar across modalities. The results show a robust alignment performance until the ATAC data has only 30% of the original number of cells, at which we observed a drop in the average Celltype matching score (blue solid line, Supplementary Fig. S2). (ii) Second, in reality the cell type proportions across modalities will vary. To test such effect, we kept the RNA modality fixed and downsampled 90%, 80%, …, 10% of cells from one specific cell type from the ATAC modality. Here, we used the three largest cell populations (Monocytes, CD4 T cells, and CD8 T cells) as varying the proportions of these populations should have the largest effect on the ATAC data manifold. We performed this experiment for each of the three populations separately. Initially, we observed a robust alignment performance when 50% or less of one cell population is removed (dashed lines, Supplementary Fig. S2); however, the alignment performance slightly decreased compared with the downsampling performed in a stratified manner across all cell types. This suggests that matching of the cell type proportions affects the overall data manifold to some degree. Furthermore, results show more variability in the alignment performance when 60% or more of one cell population is removed from the ATAC side (dashed lines, Supplementary Fig. S2).

3.4 scTopoGAN performs well on continuous differentiation data

The previously tested datasets are mainly composed of discontinuous structures having discrete cell classes. However, it is interesting to test the integration performance on developmental data, which is mainly composed of continuous structures. Thus, we tested scTopoGAN on the "scGEM" dataset, which consists of a continuous differentiation trajectory from fibroblasts to iPS. In comparison to other methods, SCOT and scTopoGAN were the top performing methods with comparable Celltype matching scores (Table 2), while UnionCom ranked third and MMD-MA ranked last.

Investigating the UMAP embeddings of the integration showed that scTopoGAN, MMD-MA, and SCOT were able to capture the continuous differentiation trajectory along the five states in the correct order (Fig. 3D). However, UnionCom splits the data into two groups. Additionally, all methods showed good mixing between the two modalities.

Table 3. Clustering evaluation of the aligned data using scTopoGAN, UnionCom, MMD-MA and SCOT.[a]

Dataset	Method	Cell class ARI (mean ± std)
Full PBMC	scTopoGAN	**0.63 ± 0.08**
	UnionCom	0.27 ± 0.03
	MMD-MA	0.19 ± 0.05
	SCOT	0.38 ± 0.01
BM	scTopoGAN	0.41 ± 0.12
	UnionCom	0.23 ± 0.02
	MMD-MA	0.38 ± 0.08
	SCOT	**0.47 ± 0.00**

[a] Reported results are computed over 10 different runs. Bold values indicate the best method for each dataset.

When evaluated quantitatively using the batch LISI score (optimal value is 2), UnionCom ranked first, followed by SCOT, while scTopoGAN and MMD-MA ranked third with similar scores (Table 2). Overall, these results show the ability of scTopoGAN to align data containing continuous cellular structures.

3.5 Post alignment downstream analysis

After a successful integration, we assessed the usability of the integrated/aligned data in downstream analysis. We tested two downstream tasks using the "Full PBMC" and "BM" datasets: (i) unsupervised clustering of the aligned data, aiming to reconstruct the known cell classes, and (ii) cross-modality prediction using nearest-neighbor regression applied on the aligned data (see Section 2.5). For the "Full PBMC" dataset, clustering the aligned data using scTopoGAN yielded the highest ARI when compared to the original cell classes, followed in order by SCOT, UnionCom, and MMD-MA (Table 3). For the "BM" dataset, scTopoGAN ranked second behind SCOT, followed by MMD-MA and lastly UnionCom.

For the cross-modality prediction, we found that this task is quite challenging given the relatively low correlations obtained in general across all methods. Results show that scTopoGAN outperformed all other methods when predicting ATAC from RNA using the "Full PBMC" dataset, with a median Spearman correlation of 0.044 (Supplementary Fig. S3), in comparison to 0.005, 0.001, and −0.024 for UnionCom, MMD-MA, and SCOT, respectively. Similarly, scTopoGAN performed best when predicting RNA from ATAC (0.011), compared to UnionCom (0.002), MMD-MA (0.001), and SCOT (−0.005). Using the "BM" dataset, when predicting ADT from RNA, scTopoGAN ranked third (0.170) after SCOT (0.335) and UnionCom (0.332), while MMD-MA (0.077) ranked last (Supplementary Fig. S3). Finally, SCOT was the best method for predicting RNA from ADT (0.036), followed in order by scTopoGAN (0.005), UnionCom (0.003), and MMD-MA (0.001). Overall, the aligned data using scTopoGAN had the best downstream analysis performance for the "Full PBMC" dataset and was the runner-up for the "BM" dataset, showing the potential usage of the aligned data by scTopoGAN for further downstream tasks.

4 Discussion

We present scTopoGAN, a method to integrate multi-modal single-cell data with non-overlapping cells or features. scTopoGAN is fully unsupervised and relies on the

assumption that different single-cell modalities measured from the same tissue have the same underlying manifold, hence the topological structure of these modalities should be similar. To perform manifold alignment, scTopoGAN uses a GAN in combination with a topological loss guiding the selection of the best performing GAN. We would like to stress that although the topological loss idea was inspired based on the evaluation using the cell type annotations, these annotations are not at all used by scTopoGAN (the topological loss is fully unsupervised).

We showed that scTopoGAN outperforms current state-of-the-art methods; however, the best Celltype matching obtained was ∼70% which shows how difficult and challenging the task of diagonal integration is. Although the topological loss calculation is capable, in most cases, to select the generator model with relatively good alignment performance, it is important to note that training the generator is quite a difficult task since the generator has no ground to link cells between the source and the target data.

For manifold projection, we used topoAEs and showed their ability to preserve the structure of the data in the low-dimensional embedding. topoAEs showed better results compared to VAE, AE, and UMAP. However, UMAP still produced the best separation between the cell classes. Nevertheless, this evaluation is biased toward UMAP, as these cell classes were defined using clustering analysis, where UMAP is used to interpret the clusters and annotate them.

Further, it was previously shown that topoAEs have superior performance to PCA and regular autoencoder (Moor *et al.* 2020). Therefore, it might be interesting to explore the applicability of topoAEs in other single-cell analysis tasks. One example is trajectory inference studying the differentiation trajectory of cells using scRNA-seq datasets (Saelens *et al.* 2019). Most trajectory inference methods rely on a lower-dimensional representation of the data, where topoAEs can be applied to produce low-dimensional space preserving the topology of the inter-cellular relationships in the data.

When tested on the continuous structured "scGEM" dataset, scTopoGAN was able to align and reconstruct the differentiation trajectory across different cell states, with comparable performance to other methods. However, it is worth noting that this extremely small sized dataset represents quite a challenge to scTopoGAN, which is a deep-learning based model usually requiring large quantity of data point to be properly trained.

The main assumption of scTopoGAN, different modalities measured from the same tissue have the same underlying manifold structure, is not completely true. Although this assumption is based on the fact that the cell pool where different modalities sample from is the same, hence similar cellular structure, different modalities are measuring different molecular features capturing different views of this cellular structure. As a result, the underlying manifolds of each modality are not identical; however, for the tested datasets, we showed that there is enough similarity between these manifolds that can be used to perform the data integration.

A major limitation in the current scTopoGAN workflow is the requirement of training multiple GAN networks in order to choose the best model based on the topological loss. It is evident that the quality of the alignment achieved is limited by the best alignment obtained in this set of GAN models. Here, we trained 20 different GAN models which is computationally expensive and there is no guarantee that the selected GAN model is the best possible solution for the tested dataset. Future improvement in this direction can incorporate the topological loss as a regularization term in the overall loss function of the GAN. This will guide the GAN to minimize the topological loss during training, thus eliminating the need to train multiple GAN models.

In all our experiments, we used the RNA modality as the target modality to which other source modalities (ATAC or ADT) were aligned. The choice of the target modality has an impact on the final alignment performance. Furthermore, we did not fine-tune the hyperparameters used for each dataset. Optimizing these hyperparameters specifically for each dataset may improve the overall results.

In conclusion, scTopoGAN opens new opportunities in studying complex tissues as it represents a step toward better integration of multiple molecular views without the restriction that these are measured from the same cell.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This project was supported by the NWO Gravitation project: BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (NWO: 024.004.012) and the NWO TTW project 3DOMICS (NWO: 17126).

References

Abdelaal T, Mourragui S, Mahfouz A *et al.* SpaGE: spatial gene enhancement using scRNA-seq. *Nucleic Acids Res* 2020;**48**:e107.

Amodio M, Krishnaswamy S. MAGAN: aligning biological manifolds. In: *Proceedings of the 35th International Conference on Machine Learning*, *ICML 2018*, Stockholmsmässan, Stockholm, Sweden. Proceedings of Machine Learning Research 80, PMLR, 2018.

Argelaguet R, Arnol D, Bredikhin D *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:111.

Argelaguet R, Cuomo ASE, Stegle O *et al.* Computational principles and challenges in single-cell data integration. *Nat Biotechnol* 2021; **39**:1202–15.

Bac J, Zinovyev A. Lizard brain: tackling locally low-dimensional yet globally complex organization of multi-dimensional datasets. *Front Neurorobot* 2019;**13**:110.

Cao K, Bai X, Hong Y *et al.* Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* 2020; **36**:i48–i56.

Cao K, Gong Q, Hong Y *et al.* A unified computational framework for single-cell data integration with optimal transport. *Nat Commun* 2022a;**13**:7419.

Cao K, Hong Y, Wan L *et al.* Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics* 2022b;**38**:211–9.

Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022; **40**:1458–66.

Cheow LF, Courtois ET, Tan Y *et al.* Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat Methods* 2016; **13**:833–6.

Demetci P, Santorella R, Sandstede B *et al.* SCOT: single-cell multi-omics alignment with optimal transport. *J Comput Biol* 2022a;**29**:3–18.

Demetci P, Santorella R, Sandstede B *et al.* Unsupervised integration of single-cell multi-omics datasets with disparities in cell-type representation. In: Pe'er, I. (ed.), *Research in Computational Molecular Biology. RECOMB 2022*. Lecture Notes in Computer Science, Vol. 13278. Springer, Cham, 2022b.

Edelsbrunner H, Harer J. Persistent homology—a survey. In: J. E. Goodman, J. Pach, & R. Pollack (eds.), *Contemporary mathematics* (Vol. 453, pp. 257–282). American Mathematical Society, 2008.

Gayoso A, Steier Z, Lopez R *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* 2021;**18**:272–82.

Ghazanfar S, Guibentif C, Marioni JC. Stabilized mosaic single-cell data integration using unshared features. *Nat Biotechnol* 2023.

Goodfellow IJ, Pouget-Abadie J, Mirza M *et al.* Generative adversarial nets. In *Advances in Neural Information Processing Systems*, Montréal, Canada, 2014.

Gui J, Sun Z, Wen Y *et al.* A review on Generative Adversarial Networks: algorithms, theory, and applications. *IEEE Trans Knowl Data Eng* 2021;**35**:3313–32.

Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**:296.

Hao Y, Stuart T, Kowalski MH *et al.* Dictionary learning for integrative, multimodal, and scalable single-cell analysis. *Nat Biotechnol* 2022. https://doi.org/10.1038/s41587-023-01767-y.

Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87.e29.

Kingma DP, Welling M. Auto-encoding variational Bayes. In: *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, Banff, Canada, 2014.

Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019; **16**:1289–96.

Kriebel AR, Welch JD. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nat Commun* 2022;**13**:780.

McInnes L, Healy J, Saul N *et al.* UMAP: uniform manifold approximation and projection. *JOSS* 2018;**3**:861.

Moor M, Horn M, Rieck B *et al.* Topological autoencoders. In: *37th International Conference on Machine Learning, ICML 2020*, Proceedings of Machine Learning Research 119, PMLR, 2020.

Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, San Juan, Puerto Rico, 2016.

Rizvi AH, Camara PG, Kandror EK *et al.* Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat Biotechnol* 2017;**35**:551–60.

Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65.

Saelens W, Cannoodt R, Todorov H *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 2019;**37**:547–54.

Shi Y, Siddharth N, Paige B *et al.* Variational mixture-of-experts autoencoders for multi-modal deep generative models. In: *Advances in Neural Information Processing Systems*, Vancouver, Canada. 2019.

Singh G, Mémoli F, Carlsson G. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In: Botsch M (ed.), *Eurographics Symposium on Point-Based Graphics*. Prague, Czech Republic: The Eurographics Association, 2007.

Singh R, Demetci P, Bonora G *et al.* Unsupervised manifold alignment for single-cell multi-omics data. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* 2020.

Stark SG, Ficek J, Locatello F *et al.*; Tumor Profiler Consortium. SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* 2020;**36**:i919–27.

Stoeckius M, Hafemeister C, Stephenson W *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017; **14**:865–8.

Sun X-J, Wang M-C, Zhang F-H *et al.* An integrated analysis of genome-wide DNA methylation and gene expression data in hepatocellular carcinoma. *FEBS Open Bio* 2018;**8**:1093–103.

Traag VA, Waltman L, van Eck NJ *et al.* From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**:5233.

Welch JD, Hartemink AJ, Prins JF *et al.* MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol* 2017;**18**:138.

Zhu C, Preissl S, Ren B *et al.* Single-cell multimodal omics: the power of many. *Nat Methods* 2020;**17**:11–4.