Gromov–Wasserstein based optimal transport to align single-cell multi-omics data

Pinar Demetci*[1,2], Rebecca Santorella*[3], Björn Sandstede[3], William Stafford Noble[4,5], and Ritambhara Singh[1,2]

[1]*Department of Computer Science, Brown University*
[2]*Center for Computational Molecular Biology, Brown University*
[3]*Division of Applied Mathematics, Brown University*
[4]*Department of Genome Sciences, University of Washington*
[5]*Paul G. Allen School of Computer Science and Engineering, University of Washington*
[*]*Equal Contribution*

1 Introduction

Single-cell measurements provide a fine-grained view of the heterogeneous landscape of cells in a sample, revealing distinct subpopulations and their developmental and regulatory trajectories. The availability of measurements capturing various genomic properties, such as gene expression, chromatin accessibility, and histone modifications, has increased the need for data integration methods for disparate data types. Due to technical limitations, it is hard to obtain multiple types of measurements from the same cell, so datasets lack sample correspondence. Furthermore, we cannot *a priori* identify correspondences between features in different domains. Accordingly, integrating two or more single-cell data modalities requires methods that do not rely on either cell-wise or feature-wise correspondences. [1–4].

Two unsupervised manifold alignment algorithms address this challenge in single-cell sequencing: (1) MMD-MA [5], which is based on the maximum mean discrepancy (MMD) measure, and (2) UnionCom [6], which performs topological alignment while emphasizing both local and global alignment. Although neither MMD-MA nor UnionCom requires any correspondence information, they require tuning three and four hyperparameters, respectively. Though hyperparameter values significantly affect the quality of the alignment for both methods, selecting the best hyperparameters is challenging in the completely unsupervised setting. One usually requires some correspondence information to pick the settings that provide the most accurate alignment.

We present **S**ingle-**C**ell alignment using **O**ptimal **T**ransport (SCOT), an unsupervised learning algorithm that employs Gromov-Wasserstein optimal transport to align single-cell multi-omics datasets while preserving local geometry. We compare the alignment performance of SCOT with MMD-MA and UnionCom on three simulated and two real-world datasets. We show that SCOT performs on par with state-of-the-art methods and converges $\sim$10 and $\sim$28 times faster than GPU implementations of MMD-MA and UnionCom, respectively. Unlike MMD-MA and UnionCom, our algorithm requires tuning only two hyperparameters and is robust to the choice of one. Moreover, we demonstrate that the Gromov-Wasserstein distance can guide SCOT's hyperparameter tuning to align the datasets effectively. As a result, SCOT is the first algorithm to perform single-cell alignment in a completely unsupervised manner. It does not require any correspondence information to align datasets or select hyperparameters. The source code and scripts for replicating results are available at `https://github.com/rsinghlab/SCOT`.

2 Method

SCOT uses the Gromov-Wasserstein based optimal transport, which preserves local neighborhood geometry when moving data points between domains. This transport problem yields a matrix of probabilities representing how likely it is that data points from one domain correspond to data points in another domain. These probabilities can then be used to project the data into the same space for alignment. While optimal transport has been used for other biological applications [7–10], SCOT is the first algorithm to apply it for single-cell

sequencing data integration. In this section, we first introduce optimal transport followed by its extension to Gromov-Wasserstein distance. Finally, we present the details of our SCOT algorithm.

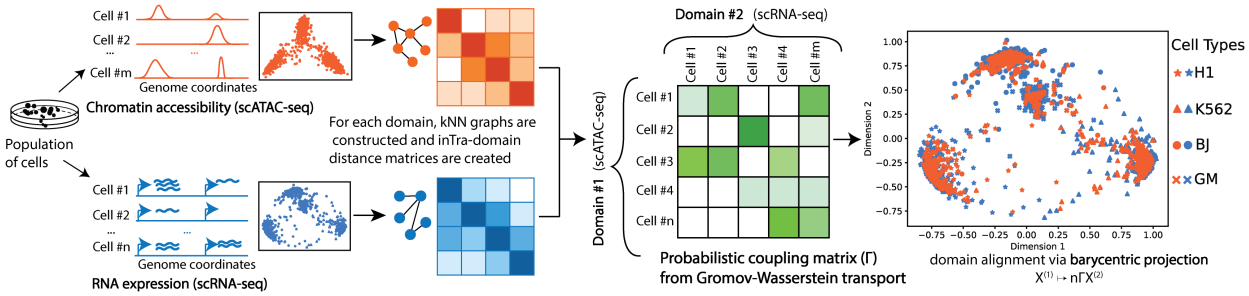

Figure 1: Schematic of SCOT applied to single-cell multi-omics data alignment (alignment of real-world SNARE-seq dataset is shown here).

We present these methods for two sets of points: $X = (x_1, x_2, \ldots, x_{n_x})$ and $Y = (y_1, y_2, \ldots, y_{n_y})$, from the measure spaces $(\mathcal{X}, p)$ and $(\mathcal{Y}, q)$, respectively. We define discrete measures p and q over our data points, representing marginal distributions of X and Y respectively, which we can write as $p = \sum_{i=1}^{n_x} p_i \delta_{x_i}$ and $q = \sum_{j=1}^{n_y} q_j \delta_{y_j}$, where δ_{x_i} is the Dirac measure. We do not require any correspondence information across the datasets but assume that there is some underlying shared manifold structure.

Optimal Transport The Kantorovich formulation of optimal transport problem seeks a minimal cost coupling between two probability distributions to tie them in a meaningful way [11]. For discrete measures, the set of possible couplings are the matrices $\Pi(p, q) = \{\Gamma \in \mathbb{R}_+^{n_x \times n_y} : \Gamma \mathbb{1}_{n_y} = p,\ \Gamma^T \mathbb{1}_{n_x} = q\}$. Each row Γ_i of a coupling Γ tells us how to split the correspondence probabilities of data point x_i onto the points y_j for $j = 1, \ldots, n_y$, and the condition $\Gamma \mathbb{1}_{n_y} = p$ requires that the sum of each row Γ_i is equal to p_i, the probability of sample x_i.

Given discrete measures p and q and a cost matrix $C \in \mathbb{R}^{n_x \times n_y}$ where C_{ij} is the cost of transporting point x_i to point y_j, the discrete optimal transport problem learns the coupling that optimizes:

$$\min_{\Gamma \in \Pi(p,q)} \langle \Gamma, C \rangle. \tag{1}$$

Intuitively, the cost function represents how many resources it will take to move x_i to y_j, and the coupling Γ assigns a probability Γ_{ij} that x_i should be moved to y_j for each x_i and y_j in the two spaces. Although this problem can be solved with minimum cost flow solvers, it is usually regularized with entropy for more efficient optimization and empirically better results [12]. Thus, the optimal transport problem that is solved numerically is

$$\min_{\Gamma \in \Pi(p,q)} \langle \Gamma, C \rangle - \epsilon H(\Gamma), \tag{2}$$

where $\epsilon > 0$ and $H(\Gamma) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \Gamma_{ij} \log \Gamma_{ij}$ is Shannon entropy. Adding entropy diffuses the optimal coupling, meaning that correspondence probabilities will be split over more data points. Equation 2 is a strictly convex optimization problem, and the solution can be obtained efficiently via Sinkhorn's algorithm [11].

Gromov-Wasserstein Optimal Transport Classic optimal transport requires defining a cost function directly on the samples themselves, which can be difficult for data in different dimensions and metric spaces. Gromov-Wasserstein distance allows for comparing distributions in different metric spaces by comparing pairwise distances between the samples across these domains, instead of looking at the samples themselves. For this extension, we need to assume we have metric measure spaces $(\mathcal{X}, D^x, p)$ and $(\mathcal{Y}, D^y, q)$, where

D^x and D^y are distance matrices on the two datasets with $D^x_{ij} = d_x(x_i, x_j)$ and $D^y_{ij} = d_y(y_i, y_j)$ for some distances d_x and d_y [13].

Given a cost function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the discrete Gromov-Wasserstein distance between p and q is

$$GW(p, q) = \min_{\Gamma \in \Pi(p,q)} \sum_{i,j,k,l} \mathbf{L}_{ijkl} \Gamma_{ij} \Gamma_{kl}, \tag{3}$$

where $\mathbf{L} \in \mathbb{R}^{n_x \times n_x \times n_y \times n_y}$ is the fourth-order tensor defined by $\mathbf{L}_{ijkl} = L(D^x_{ik}, D^y_{jl})$. Intuitively, $L(D^x_{ik}, D^y_{jl})$ captures how transporting x_i onto y_j and x_k onto y_l would distort the original distances between x_i and x_k and between y_j and y_l. This change ensures that the optimal transport plan Γ will preserve some local geometry. For our problem, we use square distance: $L(x, y) = \frac{1}{2}(x - y)^2$. As in the case of classic optimal transport, this problem can be solved efficiently through entropic regularization [14]:

$$\min_{\Gamma \in \Pi(p,q)} \sum_{i,j,k,l} \mathbf{L}_{ijkl} \Gamma_{ij} \Gamma_{kl} - \epsilon H(\Gamma). \tag{4}$$

Smaller values of ϵ lead to sparser solutions, meaning that the coupling matrix is more likely to find the correct one-to-one correspondences when they exist. However, it also means a harder (more non-convex) optimization problem [15].

Single-Cell alignment using Optimal Transport (SCOT) Our method, SCOT, first computes the pairwise distances on our data using graph distances as in [10]. To do this, we construct a $k-$nearest neighbor connectivity graph based on the correlation distances within each data set. Then, we calculate the shortest path distance on the graph between each pair of nodes using Dijkstra's algorithm and set the distance of any unconnected nodes to be the maximum (finite) distance in the graph. Our approach is robust to the choice of k. Following [15], we set p and q to be the uniform distributions on the data points. Then we solve for the optimal coupling Γ, which minimizes Equation 4.

One of this approach's major advantages is that we end up with a coupling matrix Γ with a probabilistic interpretation. The entries of the normalized row $p_i \Gamma_i$ are the probabilities that the fixed data point x_i corresponds to each y_j. To use the established correspondence metrics to evaluate the alignment, we need to project the two datasets into the same space. We use the barycentric projection $x_i \mapsto \frac{1}{p_i} \sum_{j=1}^{n_y} \Gamma_{ij} y_j$. This barycentric projection of point x_i is a weighted average of the y_j's, where the weight Γ_{ij} is the probability of correspondence between x_i and y_j. Figure 1 presents the schematic of the SCOT algorithm.

Experimental Setup We benchmark our method on three different simulation schemes. They consist of points with underlying structures of a bifurcated tree (Sim. 1), a Swiss roll (Sim. 2), and a circular frustum (Sim. 3) that were generated by Liu *et al.* [5]) and projected to higher dimensions. Next, we align two real-world single-cell co-assay datasets: (1) sc-GEM [16], which simultaneously profiles gene expression and DNA methylation at multiple loci on human somatic cell samples undergoing conversion to induced pluripotent stem cells (iPSCs) and (2) SNARE-seq [17], which links chromatin accessibility with gene expression data on a mixture of BJ, H1, K562, GM12878 cells. These datasets were pre-processed according to descriptions in [16] and [17], respectively. Our datasets have varying number of samples, Sim. 1,2,3 consist of 300 samples, sc-GEM has 177 cells, and SNARE-seq contains 1047 cells. All datasets have 1–1 correspondence information, which we use only to evaluate the alignments through the average "fraction of samples closer than the true match (FOSCTTM)" metric from [5]. We report the average FOSCTTM score across all the samples for each dataset. We compare the performance of SCOT with UnionCom and MMD-MA. For SCOT, a grid of hyperparameters is defined over the regularization weight (ϵ) and the number of neighbors (k) in the $k-$NN graph. For baseline methods, we define the grid based on recommendations in the original papers and their source code and select the set that yields the best performance (minimal average FOSCTTM).

3 Results

SCOT provides state-of-the-art alignment results We report the alignment results for SCOT for all the datasets and compare them with MMD-MA and UnionCom, in Table 1. The qualitative alignment for SNARE-seq data is shown in Figure 1 as an example. SCOT (first row) gives comparable performance to MMD-MA and UnionCom for both simulated and real-world datasets. We also observe that the Gromov-Wasserstein (GW) distance (Equation 3) serves as a proxy for measuring alignment since lower values of GW distance correspond to lower values of the average FOSCTTM. SCOT (GW), in the second row of Table 1, reports the alignment scores achieved when we select the hyperparameters corresponding to the lowest GW distance. While this selection does not always provide the best alignment, it is consistently close to it. Selecting hyperparameters is particularly challenging in an unsupervised setting. In contrast to other methods that require some corresponding information to select optimal hyperparameters, SCOT can use GW distance to pick them effectively without any other information.

Table 1: Alignment performances using average FOSCTTM scores (lower is better).

	Sim. 1	**Sim. 2**	**Sim. 3**	**sc-GEM**	**SNARE-seq**
SCOT	0.087	0.021	**0.009**	**0.198**	0.150
SCOT (GW)	0.098	0.025	0.010	0.223	0.218
MMD-MA	0.124	0.032	0.012	0.201	**0.149**
UnionCom	**0.083**	**0.016**	0.152	0.210	0.265

SCOT is faster than the state-of-the-art alignment methods We compare the running times of SCOT with the baseline methods for the best performing hyperparameters. We ran the CPU versions of the algorithms on an Intel i5-8259U CPU (base frequency 2.30GHz) with 16GB memory. For GPU versions, we used a single NVIDIA GTX 1080ti with VRAM of 11GB. We observe that SCOT converges $\sim$10, and $\sim$28, times faster than the GPU versions of MMD-MA, and UnionCom, respectively, for the largest SNARE-Seq dataset (Table 2). Unlike MMD-MA and UnionCom, which require the tuning of three or four parameters, SCOT requires the tuning of only two and is robust to the choice of one. Therefore, it drastically reduces the parameter search space making the algorithm a fast tool for unsupervised single-cell alignment tasks.

Table 2: Running times (in seconds) of all the methods averaged over ten runs.

		Sim. 1	**Sim. 2**	**Sim. 3**	**sc-GEM**	**SNARE-seq**
CPU	**SCOT**	**3.51**	**3.47**	**4.95**	**3.72**	**12.22**
	MMD-MA	30.06	29.69	28.84	16.12	547.71
	UnionCom	525.85	442.19	302.69	143.60	2169.74
GPU	**MMD-MA**	79.01	84.13	76.43	90.17	119.28
	UnionCom	117.72	112.41	109.73	70.21	345.37

4 Conclusion

SCOT uses Gromov Wasserstein-based optimal transport to perform unsupervised integration of single-cell multi-omics data. It performs on par with two state-of-the-art methods but in less time and with fewer hyperparameters, which can be selected in an unsupervised manner. Future work will focus on developing effective ways to utilize the coupling matrix, closely investigating the cases where SCOT and MMD-MA outperform each other, and extending our framework to handle more than two alignments at a time.

References

[1] M. Amodio and S. Krishnaswamy. MAGAN: Aligning biological manifolds. 2018.

[2] Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.

[3] Joshua D Welch, Alexander J Hartemink, and Jan F Prins. Matcher: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome biology*, 18(1):138, 2017.

[4] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papelexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 77(7):1888–1902, 2019.

[5] Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly embedding multiple single-cell omics measurements. *BioRxiv*, page 644310, 2019.

[6] Kai Cao, Xiangqi Bai, Yiguang Hong, and Lin Wan. Unsupervised topological alignment for single-cell multi-omics integration. *bioRxiv*, 2020.

[7] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[8] Karren D Yang, Karthik Damodaran, Saradha Venkatchalapathy, Ali C Soylemezoglu, GV Shivashankar, and Caroline Uhler. Autoencoder and optimal transport to infer single-cell trajectories of biological processes. *bioRxiv*, page 455469, 2018.

[9] Karren D Yang and Caroline Uhler. Multi-domain translation by learning uncoupled autoencoders. *arXiv preprint arXiv:1902.03515*, 2019.

[10] Mor Nitzan, Nikos Karaiskos, Nir Friedman, and Nikolaus Rajewsky. Gene expression cartography. *Nature*, 576(7785):132–137, 2019.

[11] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[13] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

[14] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

[15] David Alvarez-Melis and Tommi S Jaakkola. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

[16] Lih Feng Cheow, Elise T Courtois, Yuliana Tan, Ramya Viswanathan, Qiaorui Xing, Rui Zhen Tan, Daniel S Q Tan, Paul Robson, Loh Yuin-Han, Stephen R Quake, and William F Burkholder. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature Methods*, 13(10):833–836, 2016.

[17] Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457, 2019.