JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 1, 2022

Pp. 3–18
DOI: 10.1089/cmb.2021.0446

Research Articles

SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport

PINAR DEMETCI,[1,2,*,i] REBECCA SANTORELLA,[3] BJÖRN SANDSTEDE,[3,*] WILLIAM STAFFORD NOBLE,[4,5] and RITAMBHARA SINGH[1,2,ii]

ABSTRACT

Recent advances in sequencing technologies have allowed us to capture various aspects of the genome at single-cell resolution. However, with the exception of a few of co-assaying technologies, it is not possible to simultaneously apply different sequencing assays on the same single cell. In this scenario, computational integration of multi-omic measurements is crucial to enable joint analyses. This integration task is particularly challenging due to the lack of sample-wise or feature-wise correspondences. We present single-cell alignment with optimal transport (SCOT), an unsupervised algorithm that uses the Gromov–Wasserstein optimal transport to align single-cell multi-omics data sets. SCOT performs on par with the current state-of-the-art unsupervised alignment methods, is faster, and requires tuning of fewer hyperparameters. More importantly, SCOT uses a self-tuning heuristic to guide hyperparameter selection based on the Gromov–Wasserstein distance. Thus, in the fully unsupervised setting, SCOT aligns single-cell data sets better than the existing methods without requiring any orthogonal correspondence information.

Keywords: data integration, manifold alignment, multi-omics, optimal transport, single-cell genomics.

1. INTRODUCTION

THE GROWING VARIETY OF SINGLE-CELL ASSAYS allows us to measure the heterogeneous landscape of cell state in a sample, revealing distinct subpopulations and their developmental and regulatory trajectories across time. Different technologies can interrogate different molecular aspects of the cell, such as gene expression, protein synthesis, chromatin accessibility, DNA methylation, histone modifications, and chromatin three-dimensional (3D) confirmation. Combining data generated by these single-cell assays can

[1]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, USA.
[2]Department of Computer Science, Brown University, Providence, Rhode Island, USA.
[3]Division of Applied Mathematics, Brown University, Providence, Rhode Island, USA.
[4]Department of Genome Sciences, University of Washington, Seattle, Washington, USA.
[5]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, Washington, USA.
*These authors contributed equally to this work.
[i]ORCID ID (https://orcid.org/0000-0002-5644-0326).
[ii]ORCID ID (https://orcid.org/0000-0002-7523-160X).

provide novel insights into the interactions between these molecular views and their joint regulatory mechanisms. Hence, learning this combined information is critical to our understanding of complex biological processes and heterogeneous diseases.

Despite its importance, combining single-cell multi-omics data is a challenging task. Aside from a few recent co-assay procedures that simultaneously isolate separate molecular material for each measurement, applying multiple assays on the same single cell is impossible. Sometimes, sequencing assays need access to the same molecular material, such as with chromatin accessibility and 3D chromatin conformation capture assays. In such cases, the measurements are taken by dividing a cell population into subpopulations and assaying them separately, losing the potential for 1–1 correspondence of cells that is required for easy data integration.

Moreover, in cases where we can take measurements in the same cell and preserve the 1–1 correspondences, the choice of the experimental method for processing cells and isolating molecular materials of interest can introduce additional challenges and noise in the co-assayed data (Hu et al., 2018). For example, for simultaneous isolation of DNA and RNA, there are two general approaches: physical separation of DNA and RNA followed by separate amplification, or simultaneous preamplification followed by physical separation of the two materials. For the first approach, separation techniques such as centrifugation and micropipetting are not high-throughput; however, high-throughput approaches (Macaulay et al., 2015; Angermueller et al., 2016) have been found to introduce variability in coverage and sequencing depth of various genomic regions in the isolated DNA (Hu et al., 2018).

In recent years, computational methods have been developed to solve the single-cell data integration problem. Many of these methods combine different experiments from a single modality such as RNA sequencing for correcting batch effects (Welch et al., 2017, 2019; Amodio and Krishnaswamy, 2018; Barkas et al., 2019; Stuart et al., 2019). However, integrating data from multiple modalities such as gene expression and DNA methylation presents unique challenges. For example, when we measure different properties of a cell, we cannot a priori identify correspondences between features in the two domains.

Accordingly, integrating two or more single-cell data modalities requires methods that rely on neither common cells nor features across the data types. This aspect prevents the application of some existing single-cell alignment methods to unsupervised settings because they require some correspondence information to perform alignment (Welch et al., 2017, 2019; Amodio and Krishnaswamy, 2018; Barkas et al., 2019; Stuart et al., 2019). Earlier versions of the popular batch integration method Seurat required correspondence information in the form of cells from a similar biological state that are shared across the two data sets (known as ''anchor points'').

While a more recent version automatically selects these anchor points, it still requires features from one domain to be mapped to the other domain to perform the single-cell alignment (Stuart et al., 2019). This mapping might be possible for experiments such as gene expression and chromatin accessibility, where one can map the chromatin region read counts to the corresponding gene regions. However, it can be difficult to perform for other sequencing assay combinations. Furthermore, Cao et al. (2020) have shown that such methods do not yield quality alignments in unsupervised settings.

Multiple approaches have tried to align data sets in an entirely unsupervised manner. One of the earliest attempts, the joint Laplacian manifold alignment algorithm, constructs eigenvector projections based on *k*-nearest neighbor (*k*-NN) graph Laplacians of the data (Wang and Mahadevan, 2009). The generalized unsupervised manifold alignment (GUMA) (Cui et al., 2014) algorithm seeks a 1–1 correspondence between two data sets based on optimization of a local geometry matching term. Liu et al. (2019) showed that these methods do not perform well on the single-cell alignment task and proposed a manifold alignment (MA) algorithm based on the maximum mean discrepancy (MMD) measure, called MMD-MA. Another method, UnionCom (Cao et al., 2020), extends GUMA to perform unsupervised topological alignment and makes it more suitable for single-cell multi-omics integration.

While MMD-MA aims to match the global distributions of the data sets in a shared latent space, UnionCom emphasizes learning both local and global alignments between the two distributions. Neither method requires any correspondence information, either among samples or features, to perform an alignment. The respective articles demonstrate state-of-the-art performance on simulated and real data sets. Although these results are encouraging, MMD-MA and UnionCom require that the user specify three and four hyperparameters, respectively. Hyperparameter selection can significantly affect the quality of alignments. Therefore, in an unsupervised real-world setting with no validation data on correspondences, hyperparameter tuning can be difficult to perform and can lead to subpar alignments.

In this article, we propose an unsupervised alignment method based on optimal transport theory. Optimal transport finds the most cost-effective way to move data points from one domain to another. One way to think about it is as the problem of moving a pile of sand to fill in a hole through the least amount of work.

Traditionally, optimal transport problems have been difficult to compute, especially for large-scale data sets. However, subsequent relaxations (Kantorovich, 1942; Peyré et al., 2019) modify the original optimal transport problem, making it more applicable and easier to compute. Recently, several regularization procedures (Peyré et al., 2016) have further improved the computational scalability of optimal transport.

In biology, an emerging number of applications are using optimal transport to learn a mapping between data distributions (Alvarez-Melis and Jaakkola, 2018; Yang et al., 2018; Schiebinger et al., 2019; Yang and Uhler, 2019; Cang and Nie, 2020). Schiebinger et al. (2019) used it to study temporal changes in gene expression by using regularized unbalanced optimal transport to compute expression differences between time points. SpaOTsc (Cang and Nie, 2020) maps cells with high ligand expression onto cells with high receptor expression to recover cell signaling relationships in spatially resolved single-cell RNA-seq data sets. ImageAEOT (Yang et al., 2018) maps single-cell images to a common latent space through an autoencoder and then uses optimal transport to track cell trajectories. In related work, the same authors used autoencoders and optimal transport to learn transport maps among multiple domains (Yang and Uhler, 2019). However, the application of their method to single-cell data sets requires some form of supervision, such as class labels, to be used during transport.

The classic optimal transport problem requires data sets from the same metric space. Mémoli (2011) generalized optimal transport to the Gromov–Wasserstein distance, which compares metric spaces directly instead of comparing samples across spaces, making optimal transport suitable for multimodal alignment. In natural language processing, Alvarez-Melis and Jaakkola (2018) used this approach to measure similarities between pairs of words across languages to compute the similarity between languages. As far as we are aware, the only biological application of the Gromov–Wasserstein optimal transport comes from the study by Nitzan et al. (2019), which uses it to reconstruct the spatial organization of cells from transcriptional profiles.

We present single-cell alignment with optimal transport (SCOT), an unsupervised algorithm that uses the Gromov–Wasserstein-based optimal transport to align single-cell multi-omics data sets (presented schematically in Fig. 1). Like UnionCom, SCOT aims to preserve local geometry when aligning single-cell data. SCOT achieves this by constructing a k-NN graph for each data set (or domain) and then computing graph distance matrices for each k-NN graph to capture the intra-domain distances. SCOT then finds a probabilistic coupling matrix that minimizes the discrepancy between the intra-domain distance matrices. Finally, it uses the coupling matrix to project one single-cell data set onto another through barycentric projection, thus aligning them.

Unlike MMD-MA and UnionCom, SCOT requires tuning only two hyperparameters and is robust to the choice of one. We compare the alignment performance of SCOT with MMD-MA and UnionCom on four simulated and two real-world data sets. SCOT aligns data sets as well as the state-of-the-art methods and scales well with increasing numbers of samples. Moreover, we demonstrate that the Gromov–Wasserstein distance can guide SCOTs hyperparameter tuning in a fully unsupervised setting when no orthogonal alignment information is available. Thus, unlike other methods, SCOT provides a heuristic for hyperparameter selection without validation data. The source code for SCOT is publicly available at http://rsinghlab.github.io/SCOT.

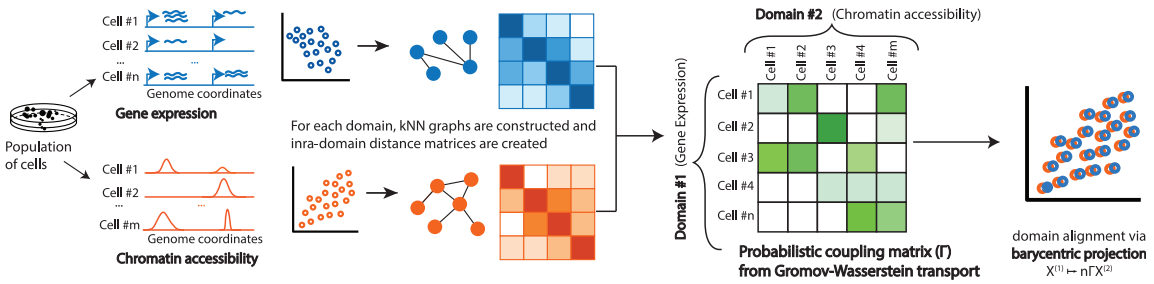

FIG. 1. Schematic of SCOT alignment of single-cell multi-omics data. A population of cells is aliquoted for different single-cell sequencing assays. SCOT constructs k-NN graphs based on sample-wise correlations and finds a probabilistic coupling between the samples of each domain that minimizes the distance between the two intra-domain graph distance matrices. Barycentric projection projects one domain onto another based on this coupling matrix. SCOT, single-cell alignment with optimal transport.

2. METHODS

SCOT relies on the Gromov–Wasserstein optimal transport to move data points from one domain to another while preserving the original local geometry. The goal of the transport problem at the core of SCOT is to find an ideal ''coupling'' (also called ''correspondence'') matrix that describes the probability of alignment between each point across domains. In this section, we first introduce optimal transport theory, followed by its extension to the Gromov–Wasserstein distance. Then, we present the details of our algorithm.

We have two data sets representing two domains, $X=(x_1, x_2, \ldots, x_{n_x})$ from $\mathcal{X}$ and $Y=(y_1, y_2, \ldots, y_{n_y})$ from $\mathcal{Y}$. The data sets have n_x and n_y points, respectively. We do not require any correspondence information or assume that there is any ground truth for 1–1 correspondence between samples or features, but we do assume that there is some underlying shared biology (e.g., cells across the data sets sharing a lineage or belonging to shared cell types), so that the data sets can be meaningfully aligned.

2.1. Optimal transport

The Kantorovich optimal transport problem seeks to find a minimal cost mapping between two probability distributions or discrete measures (Peyré et al., 2019). Referring back to the problem of moving a sand pile to fill in a hole, the Kantorovich optimal transport allows us to split the mass of a grain of sand instead of moving the whole grain; therefore, the mappings need not be 1–1. Consider discrete measures μ and ν as such

$$\mu = \sum_{i=1}^{n_x} p_i \delta_{x_i} \text{ and } \nu = \sum_{j=1}^{n_y} q_j \delta_{y_j}, \tag{1}$$

where $\sum_{i=1}^{n_x} p_i = 1 = \sum_{j=1}^{n_y} q_j,\ p_i \geq 0,\ q_j \geq 0$ and δ_{x_i} is the Dirac measure. This optimal transport problem finds a minimal coupling π that attains

$$\min_{\pi \in \Pi(\nu,\ \mu)} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} c(i,\ j)\pi(i,\ j) \tag{2}$$

$$\text{subject to : } \pi(i,\ j) \ \geq \ 0,\ \sum_{i=1}^{n_x} \pi(i,j) \ = \ q_j,\ \sum_{j=1}^{n_y} \pi(i,j) \ = \ p_i$$

where $c(i, j)$ is a cost function defined over the samples from the two data sets and $\Pi(\mu, \nu)$ is the set of couplings of μ and ν given by

$$\Pi(\mu, \nu) = \left\{\pi \in \mathbb{R}_+^{n_x \times n_y} : \pi 1_{n_y} = \mu,\ \pi^T 1_{n_x} = \nu\right\}. \tag{3}$$

Intuitively, the cost function says how many resources it will take to move point x_i in the first data set to point y_j in the second data set, and the coupling π relates the two discrete measures μ and ν by correspondence probabilities. Each row π_i tells us how to split the mass of data point x_i onto the points y_j for $j=1, \ldots, n_y$, and the condition $\pi 1_{n_y} = p$ requires that the sum of each row π_i is equal to p_i, the probability of sample x_i. The discrete optimal transport problem finds a coupling matrix, Γ, that minimizes the cost of moving samples through the linear program:

$$\min_{\Gamma \in \Pi(\mu,\ \nu)} \langle \Gamma, C \rangle. \tag{4}$$

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 (Cuturi, 2013). Entropy diffuses the optimal coupling, meaning that more masses will be split. Thus, the numerical optimal transport problem is

$$\min_{\Gamma \in \Pi(\mu,\ \nu)} \langle \Gamma, C \rangle - \epsilon H(\Gamma), \tag{5}$$

where $\epsilon > 0$ and $H(\Gamma)$ is the Shannon entropy ($\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \Gamma_{ij} \log \Gamma_{ij}$).

Equation (5) is a strictly convex optimization problem, and for some unknown vectors $u \in \mathbb{R}^{n_x}$ and $v \in \mathbb{R}^{n_y}$, the solution has the form $\Gamma^* = \text{diag}(u)K\,\text{diag}(v)$, with $K = \exp\left(-\frac{C}{\varepsilon}\right)$, element-wise. This solution can be obtained efficiently via Sinkhorn's algorithm, which iteratively computes

$$u \leftarrow \mu \% Kv \text{ and } v \leftarrow \nu \% K^T u, \tag{6}$$

where % denotes element-wise division. This derivation immediately follows from solving the corresponding dual problem for Equation (5) (Peyré et al., 2019).

2.2. The Gromov–Wasserstein optimal transport

While the classic optimal transport formulation requires us to define a cost function across domains [Eq. (2)], this is difficult to do when working with data from different metric spaces. This is because we cannot directly compare data points with different modalities, such as in the case of multi-omic alignment. The Gromov–Wasserstein distance extends optimal transport by comparing distances between data points rather than directly comparing the data points themselves (Alvarez-Melis and Jaakkola, 2018) and allows us to work with data from different modalities. Consider the same discrete measures μ and ν as above, the cost function in the formulation of the optimal transport problem will now be defined over sample-wise pairwise distances $d_x(i,\ k)$ and $d_y(j,\ l)$ in the X and Y data sets, respectively:

$$GW(\mu,\ \nu) : = \min_{\pi \in \Pi(\mu,\ \nu)} \sum_{i,\ k}^{n_x} \sum_{j,\ l}^{n_y} L(d_x(i,\ k), d_y(j,\ l))\, \pi\,(i, j)\, \pi(k,\ l). \tag{7}$$

where L indicates the cost function. The main change from basic optimal transport [Eq. (2)] to the Gromov–Wasserstein optimal transport [Eq. (7)] is that we consider the effect of transporting pairs of samples rather than single samples. Intuitively, $L(d_x(i,\ k),\ d_y(j,\ l))$ captures how transporting x_i to y_j and x_k to y_l would distort the original distances between i and k and between x_j and x_l. This change ensures that the optimal transport plan π will preserve some local geometry.

For solving the Gromov–Wasserstein optimal transport formulation, we compute pairwise distance matrices D^x and D^y for the two domains separately, as well as the fourth-order tensor $\mathbf{L} \in \mathbb{R}^{n_x \times n_x \times n_y \times n_y}$, where $\mathbf{L}_{ijkl} = L(D^x_{ik},\ D^y_{jl})$. Then, the discrete Gromov–Wasserstein problem can also be expressed as the inner product

$$GW(\mu,\ \nu) = \min_{\Gamma \in \Pi(\mu,\ \nu)} \langle \mathbf{L}(D^x,\ D^y) \otimes \Gamma,\ \Gamma \rangle \tag{8}$$

Equation (8) is now both nonlinear and nonconvex and involves operations on a fourth-order tensor, including the $\mathcal{O}(n_x^2 n_y^2)$ operation tensor product $L(D^x,\ D^y) \otimes \Gamma$ for a naive implementation. Peyré et al. (2016) showed that for some choices of loss function this product can be computed in $\mathcal{O}(n_x^2 n_y + n_x n_y^2)$ cost. In particular, for the case $L = L_2$, the inner product can be computed by

$$\mathbf{L}(D^x,\ D^y) \otimes \Gamma = (D^x)^2 \mu 1_{n_y}^T + 1_{n_x} \nu^T ((D^y)^2)^T - D^x \Gamma (D^y)^T. \tag{9}$$

As in the classic optimal transport case, the coupling matrix can be efficiently computed for an entropically regularized optimization problem:

$$GW(\mu,\ \nu) = \min_{\Gamma \in \Pi(\mu,\ \nu)} \langle \mathbf{L}(D^x,\ D^y) \otimes \Gamma,\ \Gamma \rangle - \epsilon H(\Gamma). \tag{10}$$

Larger values of ϵ lead to not only an easier optimization problem but also a denser coupling matrix, meaning that solutions will indicate significant correspondences between more data points. Smaller values of ϵ lead to sparser solutions, meaning that the coupling matrix is more likely to find the correct one-to-one correspondences for data sets where there are one-to-one correspondences. However, it also yields a harder (more nonconvex) optimization problem (Alvarez-Melis and Jaakkola, 2018).

Peyré et al. (2016) proposed using a projected gradient descent approach for optimization, where both the projection and the gradient are taken with respect to the Kullback–Leibler divergence. These projections are computed via the Sinkhorn iterations. Algorithm 1 in the Supplementary Materials presents the algorithm for $L = L_2$.

Algorithm 1: Unsupervised hyperparameter search procedure

Input: Data sets $X,\ Y$.
$n \leftarrow \min(n_x, n_y)$, $k_1 \leftarrow \min(0.2n, 50)$
$\epsilon_1 \leftarrow \arg\ \min_{\epsilon \in [10^{-3},\ 10^{-2}]} \mathrm{SCOT}(X,\ Y,\ k_1,\ \epsilon)$ // Fix k_1 and vary ϵ
// Fix ϵ_1 and vary k
if $n > 250$ **then**
  $k_2 \leftarrow \arg\ \min_{k \in [20,\ 100]}\ \mathrm{SCOT}(X,\ Y,\ k,\ \epsilon_1)$
end
else
  $k_2 \leftarrow \arg\ \min_{k \in [0.05n,\ 0.2n]}\ \mathrm{SCOT}(X,\ Y,\ k,\ \epsilon_1)$
end
// Do a more refined search around k_2 and ϵ_1
$k_{\text{best}},\ \epsilon_{\text{best}} \leftarrow \arg\ \min_{k \in [k_2 - 5,\ k_2 + 5],\ \epsilon \in [10^{-0.25}\epsilon_1,\ 10^{0.25}\epsilon_1]}\ \mathrm{SCOT}(X,\ Y,\ k,\ \epsilon)$
Return: $k_{\text{best}},\ \epsilon_{\text{best}}$

2.3. Single-cell alignment with optimal transport

Our method, SCOT, works as follows. First, we compute the pairwise distances on our data in a way similar to Nitzan et al. (2019). To do this, we use the correlations between data points within each data set to construct k-NN connectivity graphs. We find that connectivity graphs, which connect nodes with binary edges, empirically work better than weighted edges. This could be because connectivity graphs potentially denoise the data. Next, we compute the shortest path distance on the graph between each pair of nodes via Dijkstra's algorithm.

We set the distance of any unconnected nodes to be the maximum finite distance in the graph and normalize the matrix by dividing the elements by this maximum distance. If k is the number of samples, then the k-NN graph is the complete graph, so the corresponding distance matrix is a matrix of all ones. In this case, the distance matrix does not provide information about the local geometry, so we recommend keeping k small relative to the number of samples to avoid this scenario. We find that our approach is robust to the choice of k (Fig. 5).

Since we do not know the true distribution of the original data sets, we follow the study by Alvarez-Melis and Jaakkola (2018) and empirically set μ and ν to be the uniform distributions on the data points. Then, we solve for the optimal coupling Γ, which minimizes Equation (10). To implement this method, we use the Python Optimal Transport toolbox (https://pot.readthedocs.io/en/stable) (Flamary and Courty, 2017).

One of the advantages of using optimal transport is the probabilistic interpretation of the resulting coupling matrix Γ, where the entries of the normalized row $\frac{1}{p_i}\Gamma_i$ are the probabilities that the fixed data point x_i corresponds to each y_j. However, to use the evaluation metrics previously used in the field and to visualize alignment, we need to project the two data sets into the same space. The Procrustes approach proposed in the study by Alvarez-Melis and Jaakkola (2018) does not generalize to data sets with different feature and sample dimensions, so we use a barycentric projection:

$$x_i \mapsto \frac{1}{p_i} \sum_{j=1}^{n_y} \Gamma_{ij} y_j. \tag{11}$$

2.4. Alternative unsupervised alignment procedure

In the description of SCOT, the number k for nearest neighbors and the entropy weight ϵ are hyperparameters. One way to set these hyperparameters for optimal alignment is to use some orthogonal correspondence information to select the best alignment either directly (Liu et al., 2019; Cao et al., 2020) or by performing cross-validation (Singh et al., 2020). This selection strategy is problematic for truly unsupervised setting, where no correspondence information is available a priori upon sequencing separate cell cultures.

As a solution, we provide an alternative procedure to learn reasonable alignments based on tracking the Gromov–Wasserstein distance [Eq. (8)]. This procedure is based on our observation that the Gromov–Wasserstein distance serves as a proxy for measuring alignment quality (Fig. 4A). In this procedure, we

alternate between optimizing ϵ and k to minimize the Gromov–Wasserstein distance between the domains (detailed in Algorithm 1). Although the lowest Gromov–Wasserstein distance is not always the best alignment, it consistently appears to be one of the better alignments.

3. EXPERIMENTAL SETUP

3.1. Simulated data sets

We follow the study by Liu et al. (2019) and benchmark SCOT on three different simulations (https://noble.gs.washington.edu/proj/mmd-ma). All three simulations contain two domains with 300 samples that have been nonlinearly projected to 1000- and 2000-dimensional feature spaces, respectively. The three simulations are a bifurcation, a Swiss roll, and a circular frustum (Fig. 2) with points belonging to three different groups. In addition to these three previously existing simulations, we use Splatter (Zappia et al., 2017) to create simulated single-cell RNA sequencing count data, which we call synthetic RNA-seq. We generate 5000 cells with 1000 genes from 3 cell groups and reduce the count matrix to the 5 genes with the highest variances. This count matrix is mapped into two new domains with dimensions $p_1 = 50$ and $p_2 = 500$ by multiplying it with two randomly generated matrices, resulting in data with dimensions 5000×50 and 5000×500.

All four data sets were simulated with 1–1 sample-wise correspondences, which are solely used for evaluating model performance. Each domain is projected to a different dimension, so there is no feature-wise correspondence either. In all simulations, we Z-score normalize the features before running the alignment algorithms as in the study by Liu et al. (2019).

3.2. Single-cell multi-omics data sets

We use two sets of single-cell multi-omics data to demonstrate the applicability of our model to real data sets. Both data sets are generated by co-assays; thus, we have known cell-level correspondence information for benchmarking. The first data set is generated using the scGEM assay (Cheow et al., 2016), which simultaneously profiles gene expression and DNA methylation. The data set (Sequence Read Archive accession SRP077853) is derived from human somatic cell samples undergoing conversion to induced pluripotent stem cells and shows a continuous trajectory. This data set was also used by Cao et al. (2020) to demonstrate the performance of their UnionCom algorithm. We preprocessed the data as described in the original publications (Cheow et al., 2016; Cao et al., 2020) and ended up with dimensions 177×34 for the gene expression data and 177×27 for the chromatin accessibility data.

The second data set is generated by the SNAREseq assay (Chen et al., 2019), which links chromatin accessibility with gene expression. The data (Gene Expression Omnibus accession GSE126074) is derived from a mixture of human cell lines: BJ, H1, K562, and GM12878 and show distinct cell type clusters. We preprocess the data sets following the study by Chen et al. (2019).The resulting data matrices for the SNARE-seq data set were of size 1047×19 and 1047×10 for ATAC-seq and RNA-seq, respectively. We unit normalize all real data sets as done in the study by Singh et al. (2020). Both data sets that we work with have been previously published and made publicly available by the original authors. Therefore, they do not require IRB approval.

3.3. Evaluation metrics

We compare SCOT with the two state-of-the-art unsupervised single-cell alignment methods MMD-MA (Liu et al., 2019) and UnionCom (Cao et al., 2020). None of these methods use any correspondence information for aligning the data sets. However, all data sets have 1–1 sample-level correspondence information, which we use to quantify the alignment performance through the ''fraction of samples closer than the true match'' (FOSCTTM) metric introduced by Liu et al. (2019). For each domain, we compute the Euclidean distances between a fixed sample point and all the data points in the other domain. Next, we use these distances to compute the fraction of samples that are closer to the fixed sample than its true match. Finally, we average these values for all the samples in both domains. For perfect alignment, all samples would be closest to their true match, yielding an average FOSCTTM of zero. Therefore, a lower average FOSCTTM corresponds to better alignment performance.

Since all the data sets have group-specific (simulations) or cell-type-specific (real experiments) labels, we also adopt the metric used by Cao et al. (2020) called ''label transfer accuracy'' (LTA) to assess the quality of the cell label assignment and to allow for a more direct comparison with their results. This metric

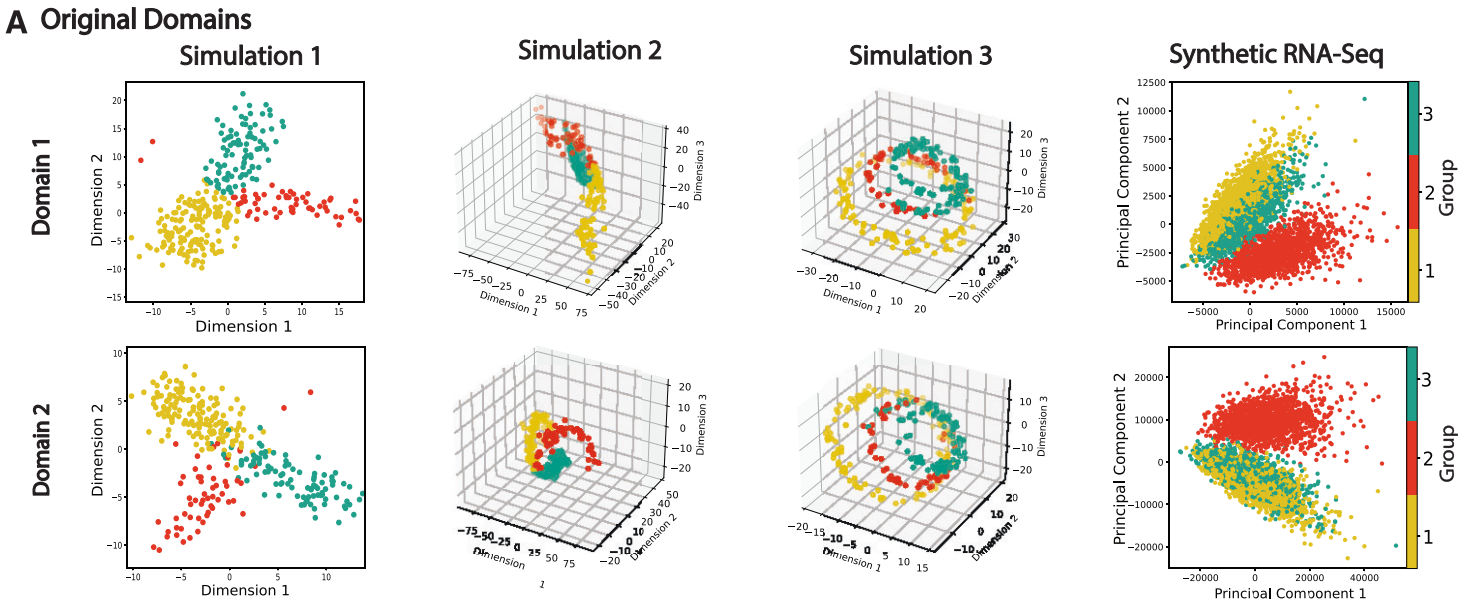

FIG. 2. Alignment results for simulated data sets. We present the alignment result on four simulations (left to right)—a bifurcation, a Swiss roll, a circular frustum, and synthetic RNA-seq data generated from Splatter (Zappia et al., 2017). **(A)** Visualization of the data set before alignment. Each data set has two domains to be aligned. **(B)** Visualization of data sets after alignment by SCOT. The upper row plots samples colored by domain they come from, whereas the bottom row shows samples colored by their group (or cell type) identity. **(C)** Performance benchmarking. We plot sorted FOSCTTM measures for alignments performed by SCOT, MMD-MA, and UnionCom for benchmarking. The mean FOSCTTM measures for each alignment and data set are included in figure legends. Best performing results are bolded. FOSCTTM, fraction of samples closer than the true match; MMD-MA, maximum mean discrepancy-manifold alignment.

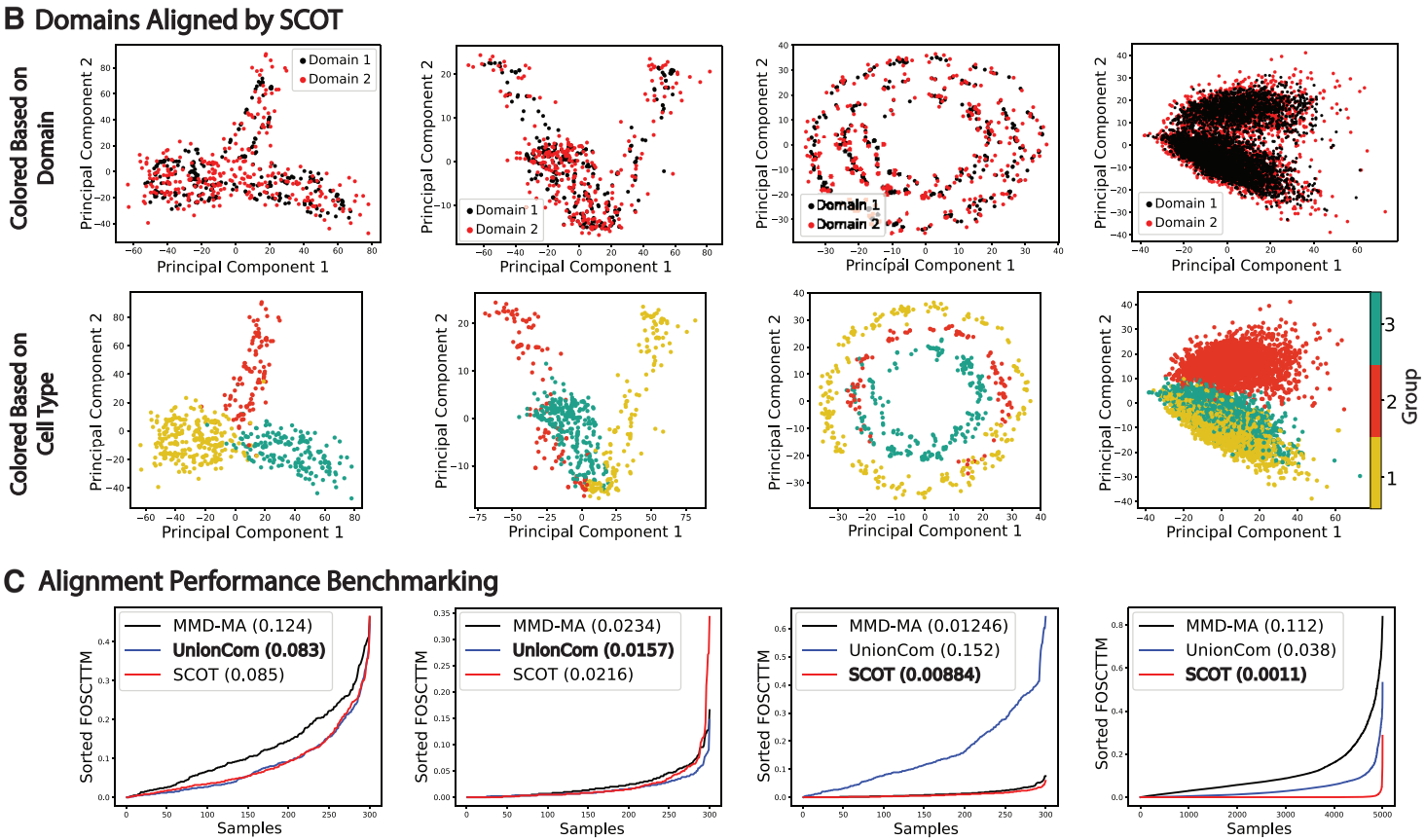

FIG. 2. (*Continued*)

measures the ability to correctly transfer sample labels from one domain to another based on their neighborhood in the aligned domain. As described in the study by Cao et al. (2020), we train a k-NN classifier on one of the domains and predict the sample labels in the other domain. The LTA is the proportion of correctly predicted labels, so it ranges from 0 to 1, and higher values indicate good performance. We apply this metric to alignments selected by the FOSCTTM measure.

3.4. Hyperparameter tuning

We run each method over a grid of hyperparameters and select the setting that yields the lowest average FOSCTTM. For SCOT, the grid covers the regularization weight $\epsilon \in \{0.0001, 0.0005, 0.001, 0.005, \ldots, 0.1\}$ and number of neighbors $k \in \{10, 15, 20, 25, 30, 35, \ldots 100, \frac{1}{6}n_x\}$. We observe empirically that going above $\frac{1}{6}n$ for k does not yield any improvement in alignment.

We pick the hyperparameters for MMD-MA and UnionCom based on the default values and recommended ranges. MMD-MA has three hyperparameters: weights $\lambda_1, \lambda_2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ for the terms in the optimization problem and the dimensionality $p \in \{4, 5, 6, 16, 32, 64\}$ of the embedding space. UnionCom requires the user to specify four hyperparameters: the number $kmax \in \{40, 100\}$ of maximum number of neighbors in the graph, the dimensionality $p \in \{4, 5, 6, 16, 32, 64\}$ of the embedding space, the trade-off parameter $\beta \in \{0.1, 1, 10, 15, 20\}$ for the embedding, and a regularization coefficient $\rho \in \{0, 5, 10, 15, 20\}$. We select the embedding dimension $p \in \{16, 32, 64\}$ around the default value of 32 set by UnionCom but also add $p \in \{4, 5, 6\}$ to match the recommended values for MMD-MA. We keep the hyperparameter search space size approximately consistent across the three methods. For each data set, we present alignment and runtime results for the best performing hyperparameters.

Furthermore, we consider the scenario where correspondence information is unavailable to pick the optimal hyperparameters. For SCOT, we apply the alternative unsupervised alignment algorithm (Algorithm 2 in the Supplementary Materials S1) to align all the data sets. Since MMD-MA and UnionCom do not provide a hyperparameter selection strategy, we rely on the default hyperparameters; we use UnionCom's provided default parameters of $kmax = 40$, $p = 32$, $\rho = 10$, and $\beta = 1$, and the center values of MMD-MA's recommended range: $p = 5$, $\lambda_1 = 10^{-5}$, and $\lambda_2 = 10^{-5}$. We also present the alignment results for all three methods in this fully unsupervised setting.

4. RESULTS

We use four simulation data sets and two real-world single-cell sequencing data sets to assess the alignment performance of SCOT. We benchmark it against the two state-of-the-art unsupervised single-cell multi-omics alignment algorithms, MMD-MA and UnionCom, using FOSCTTM and LTA metrics. The former assesses cell-to-cell alignment error and the latter assesses the cell-type grouping accuracy upon alignment.

4.1. SCOT successfully aligns the simulated data sets

In this experiment, we align the three simulation data sets from the study by Liu et al. (2019), as well as the synthetic single-cell RNA-seq count data generated with Splatter (Zappia et al., 2017). Before alignment, we first select the best performing hyperparameters for each method using the ground-truth correspondence information, as described in Section 3.4.

In Figure 2, we visualize the original domains, as well as the alignment performed by SCOT. We color the samples by their domain and cell-type identity. We observe that the global structure is matched, and cells cluster correctly based on cell-type identity. We then sort and plot the FOSCTTM score for each sample in Figure 2C. The mean FOSCTTM values are summarized in Table 1. We also report the LTA values in Table 2 when the first domain is used to train a classifier to predict the labels in the second domain. Overall, we observe that SCOT consistently achieves one of the lowest average FOSCTTM scores, thereby demonstrating its ability to recover the correct correspondences. SCOT also consistently yields high LTA scores indicating that samples are correctly mapped to their assigned groups.

Table 1. Alignment Performance by Average FOSCTTM Measure When the First Domain Is Projected onto the Second Domain

	Simulation 1	*Simulation 2*	*Simulation 3*	*Synthetic RNA-seq*	*scGEM*	*SNAREseq*
SCOT	0.085	0.022	**0.009**	**0.001**	**0.192**	**0.150**
MMD-MA	0.124	0.023	0.012	0.112	0.201	**0.150**
UnionCom	**0.083**	**0.016**	0.152	0.038	0.209	0.265

For real-world data sets, we picked gene expression domain in scGEM and chromatin accessibility domain in SNAREseq to be projected.

Bold values indicate the best performing alignment for each data set.

FOSCTTM, fraction of samples closer than the true match; MMD-MA, maximum mean discrepancy-manifold alignment; SCOT, single-cell alignment with optimal transport.

4.2. SCOT gives state-of-the-art performance for single-cell multi-omics alignment

Next, we apply our method to real single-cell sequencing data and visualize the alignments in Figure 3. To have ground-truth information on cell–cell correspondences solely for benchmarking purposes, we use data sets generated by co-assaying technology. Overall, SCOT gives the lowest average FOSCTTM measure in comparison to MMD-MA and UnionCom (Table 1) and recovers accurate 1–1 correspondences in single-cell data sets. For the scGEM data, we report LTA using the DNA methylation domain for predicting the cell-type labels in the gene expression domain.

For the SNARE-seq data set, we use the gene expression domain for predicting cell labels in the chromatin accessibility domain (Table 2). SCOT yields the best LTA result on SNAREseq data set and performs comparably to the other methods for scGEM. All methods have higher LTA performance on SNAREseq data set compared with scGEM data set because SNAREseq data set contains a mixture of different cell types that cluster separately, whereas scGEM data set contains cells going through a continuous differentiation.

While MMD-MA and UnionCom project both data sets to a shared low-dimensional space, SCOT projects one data set onto the other. We find that the direction of projection makes no significant difference in performance (Supplementary Table S1).

4.3. SCOTs alternative unsupervised hyperparameter tuning procedure achieves quality alignments

We compare the alignment performances in fully unsupervised settings, when we have no validation data on correspondences to use for hyperparameter tuning, as described in Section 3.4. We present the alignment performances, measured by average FOSCTTM measures, in Table 3, and by LTA in Table 4, when using SCOTs alternative self-tuning procedure. In this procedure, hyperparameter choice is guided by the Gromov–Wasserstein distance, as we have observed a correlation between the Gromov–Wasserstein distances between the aligned data sets and alignment quality (Fig. 4A).

In this unsupervised setting, we use MMD-MA's and UnionCom's default parameters since they lack self-tuning capability. SCOT returns nearly the same alignments for simulated data and only marginally worse alignments for real data. In contrast, MMD-MA and UnionCom show inconsistent alignment performance and fail to align some of the simulated and all real data sets with the default parameter values. Therefore, the proposed procedure could guide a user to an alignment close to the optimal result when no orthogonal information is available.

Table 2. Alignment Performance by Label Transfer Accuracy ($k=5$) when the First Domain is Used in Training (Gene Expression Domains for Real World Data Sets)

	Simulation 1	*Simulation 2*	*Simulation 3*	*Synthetic RNA-seq*	*scGEM*	*SNAREseq*
SCOT	0.937	**0.977**	**0.957**	**0.998**	0.576	**0.982**
MMD-MA	0.89	0.783	0.947	0.706	**0.588**	0.942
UnionCom	**0.96**	0.62	0.613	0.997	0.582	0.423

Bold values indicate the best performing alignment for each data set.

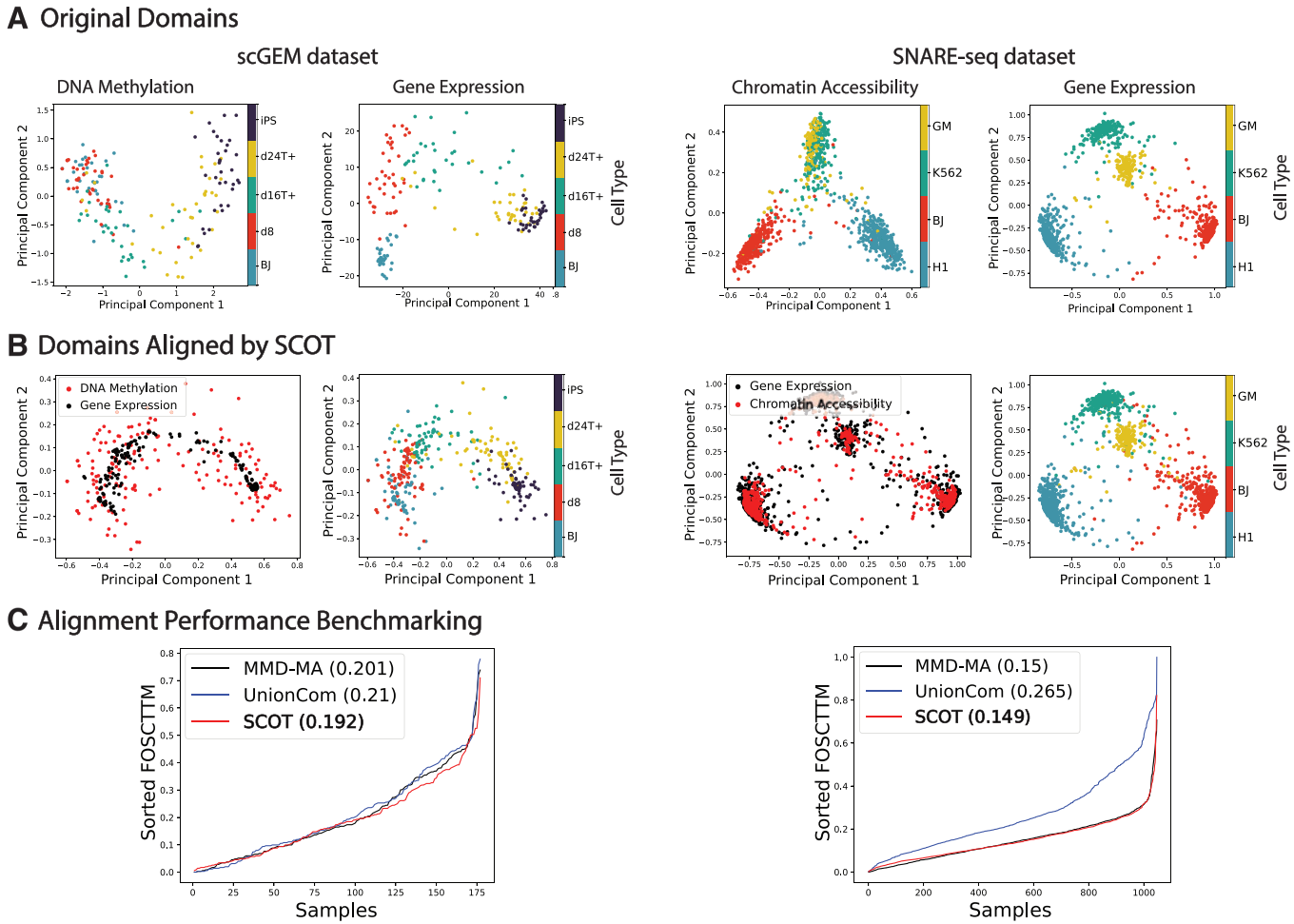

FIG. 3. Aligning real-world single-cell sequencing data set. **(A)** We first visualize the original data sets before alignment. Each data set has two domains with different sequencing modalities. Left: our alignment colored based by domain (plotted in 2D using PCA). **(B)** We visualize the aligned data sets after running SCOT. For each data set, we plot alignments both by coloring data points by domain and by cell-type identity. **(C)** We benchmark SCOT against MMD-MA and UnionCom algorithms by comparing FOSCTTM values we get. Graphs here plot sorted FOSCTTM measures, and the legend contains average FOSCTTM measures for each alignment. 2D, two-dimensional; PCA, principal component analysis.

Table 3. Alignment Performance by the Mean FOSCTTM Scores in Fully Unsupervised Setting

	Simulation 1	*Simulation 2*	*Simulation 3*	*Synthetic RNA-seq*	*scGEM*	*SNAREseq*
SCOT (GW)	**0.088**	0.025	**0.009**	**0.001**	**0.209**	**0.218**
MMD-MA	0.125	**0.012**	0.739	0.384	0.437	0.473
UnionCom	0.091	0.028	0.684	0.028	0.691	0.510

The hyperparameters for SCOT are chosen by the lowest Gromov–Wasserstein distance and the default hyperparameters are used for MMD-MA and UnionCom.

Bold values indicate the best performing alignment for each data set.

4.4. SCOTs computation speed scales well with the sample size

We compare SCOTs running times with the baseline methods for the best performing hyperparameters on the synthetic RNA-seq data set by varying the number of cells to demonstrate how each algorithm scales to larger data sets. While SCOT is implemented for CPU, both MMD-MA and UnionCom algorithms provide GPU versions, which run faster. Therefore, we use them for benchmarking. We run CPU computations on an Intel Xeon e5-2670 with 16 GB memory and GPU computations on a single NVIDIA GTX 1080ti with VRAM of 11 GB. SCOTs running time scales similar to that of MMD-MA, even though SCOT runs on a CPU and MMD-MA runs on a GPU (Fig. 4B). Both methods scale better than the GPU-based UnionCom implementation.

4.5. Investigating algorithmic choices and hyperparameters of SCOT

To better understand our method, we investigated the effects of different algorithmic choices and hyperparameter combinations on the alignment performance of the real-world data sets. Figure 5 shows the range of average FOSCTTM values we receive for alignments with different combinations of k (number of neighbors in k-NN graphs) and ϵ (entropic regularization coefficient) values for the two real-world sequencing data sets. Overall, we observe that the choice of ϵ tends to make a larger impact on the alignment performance than k. Next, we consider the effect of different algorithmic choices on the alignment performance of SCOT.

We compare the final SCOT model with (1) no entropic regularization, (2) using Euclidean distances for intra-domain distance matrices, and (3) using correlation-based intra-domain distance matrices in lieu of graph distances. For each of these settings, we run alignments for the same combinations of hyperparameters as described in Section 3.4 and record the average FOSCTTM measure we receive for each alignment. In Figure 6, we compare these in violin plots for scGEM and SNARE-seq data sets. This experiment shows that both entropic regularization and modeling the single-cell data sets as graphs for intra-domain distance computations yield lower FOSCTTM measures, corresponding to higher quality alignments.

5. DISCUSSION

We have demonstrated that SCOT, which uses the Gromov–Wasserstein optimal transport for unsupervised single-cell multi-omics data integration, performs on par with UnionCom and MMD-MA when sample correspondence information is available for hyperparameter tuning and shows advantages in other scenarios and aspects. Our formulation of a coupling matrix based on matching graph distances is

Table 4. Alignment Performance by Label Transfer Accuracy ($k=5$) in the Fully Unsupervised Setting When the First Domain Is Used for Training

	Simulation 1	*Simulation 2*	*Simulation 3*	*Synthetic RNA-seq*	*scGEM*	*SNAREseq*
SCOT	**0.977**	**0.977**	**0.95**	**0.996**	**0.582**	**0.701**
MMD-MA	0.897	0.957	0.7	0.506	0.237	0.412
UnionCom	0.947	0.947	0.133	0.948	0.107	0.288

The hyperparameters for SCOT are chosen by the lowest Gromov–Wasserstein distance and the default hyperparameters are used for MMD-MA and UnionCom.

Bold values indicate the best performing alignment for each data set.

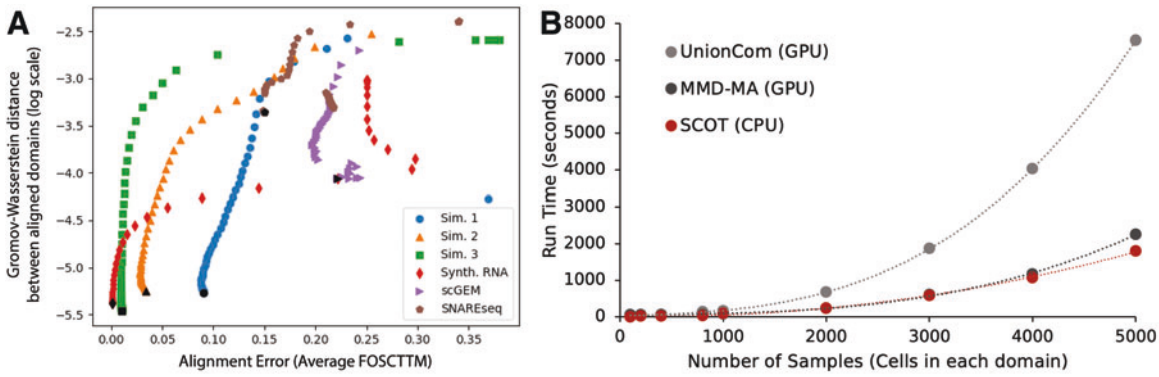

FIG. 4. **(A)** Runtime comparisons with growing sample size. Dotted lines are polynomial trend lines. **(B)** Relationship between the Gromov–Wasserstein distance between the aligned data sets and alignment quality. Lower Gromov–Wasserstein values tend to correspond to better alignments (lower FOSCTTM measures).

somewhat similar to UnionCom's initial step; however, UnionCom only matches sample-to-sample distances, whereas the Gromov–Wasserstein distance considers the cost of moving pairs of points, enabling our method to better preserve local geometry.

Additionally, SCOT performs global alignment of the marginal distributions, which is similar to how MMD-MA uses the MMD term to ensure that the two distributions agree globally in the latent space. We hypothesize that these properties result in SCOTs state-of-the-art performance. Furthermore, SCOTs optimization runs in less time and with fewer hyperparameters, and the Gromov–Wasserstein distance can guide the user to choose an alignment when no validation information exists. Therefore, unlike other methods, SCOT easily yields high-quality alignments in the realistic fully unsupervised setting.

While barycentric projection provides a way to visualize the alignment, it assumes that cells in one data set should be mapped to the convex hull of the other data set. Future work will develop unbalanced optimal transport, which would take care of outliers as well as under- or overrepresented groups. There are also other ways to use the coupling matrix to infer alignment such as using it with other dimension reduction methods such as t-SNE (as in UnionCom) to align the manifolds while embedding them both into a new space. Alternatively, depending on the application, a projection may not be required; it may be sufficient to have probabilities relating the samples to one another. Future work will develop effective ways to utilize the coupling matrix and extend our framework to handle more than two alignments at a time.

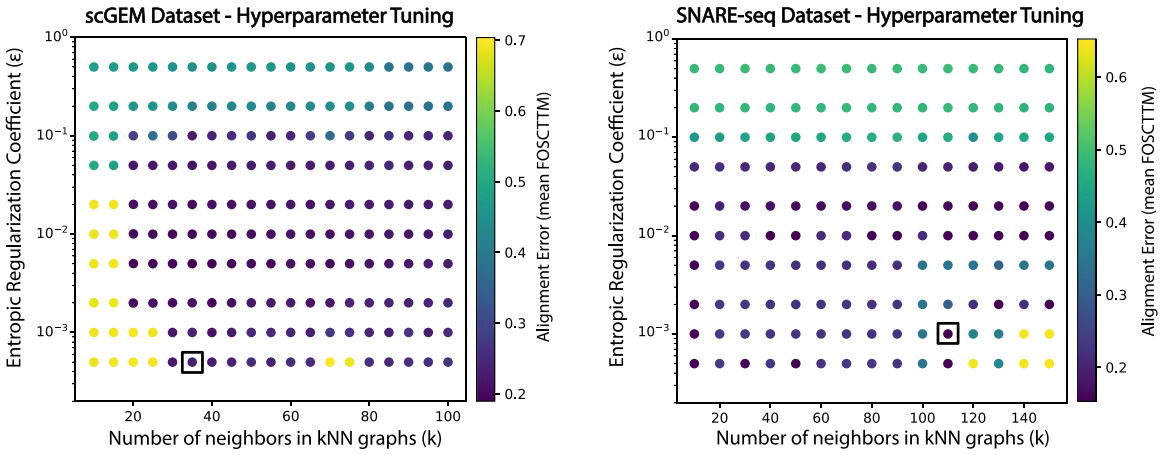

FIG. 5. Hyperparameter tuning results for scGEM (left) and SNARE-seq (right) data sets. We sweep a range of values for the two hyperparameters in our model: number of neighbors in k-NN graphs, k (on the x-axis), and the entropic regularization coefficient, ϵ (on the y-axis). The color of the scattered dots corresponds to the average FOSCTTM values we receive for each alignment, with lower values corresponding to better alignments. The hyperparameter combinations that yielded the best FOSCTTM values are in black squares.

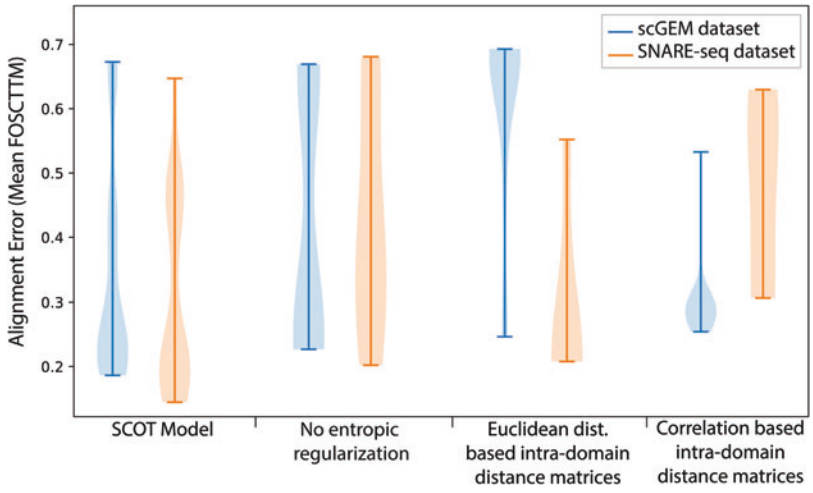

FIG. 6. Ablation test results. We considered several modifications to algorithmic choices in SCOT and investigated the range of average FOSCTTM values we received in our alignments for scGEM (blue) and SNARE-seq (orange) data sets. The modifications considered are: (1) removing the entropic regularization term from the Gromov–Wasserstein optimal transport objective function, (2) using Euclidean distances for intra-domain distance, and (3) using correlation-based distances instead of graph distances for the intra-domain distance matrices.

ACKNOWLEDGMENTS

We are grateful to Yang Lu, Jean-Philippe Vert, and Marco Cuturi for helpful discussion of the Gromov–Wasserstein optimal transport.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

FUNDING INFORMATION

W.S.N.'s contribution to this work was funded by the NIH award U54 DK107979. B.S. was partially supported by the National Science Foundation (NSF) awards 1714429 and 1740741. R.S. is supported by the NSF Graduate Research Fellowship under Grant No. 1644760.

SUPPLEMENTARY MATERIAL

Supplementary Materials S1

REFERENCES

Alvarez-Melis, D., and Jaakkola, T.S. 2018. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*.

Amodio, M., and Krishnaswamy, S. 2018. MAGAN: Aligning biological manifolds. *In International Conference on Machine Learning*, *PMLR* 80, 215–223. Stockholm, Sweden.

Angermueller, C., Clark, S.J., Lee, H.J., et al. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* 13, 229–232.

Barkas, N., Petukhov, V., Nikolaeva, D., et al. 2019. Joint analysis of heterogeneous single-cell rna-seq dataset collections. *Nat. Methods* 16, 695–698.

Cang, Z., and Nie, Q. 2020. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun*. 11, 1–13.

Cao, K., Bai, X., Hong, Y., et al. 2020. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* 36 (Supplement 1), i48–i56.

Chen, S., Lake, B.B., and Zhang, K. 2019. High-throughput sequencing of transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol*. 37, 1452–1457.

Cheow, L.F., Courtois, E.T., Tan, Y., et al. 2016. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat. Methods* 13, 833–836.

Cui, Z., Chang, H., Shan, S., et al. 2014. Generalized unsupervised manifold alignment, 2429–2437. *In Advances in Neural Information Processing Systems*, 27, 2429–2437. Montreal, Canada.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport, 2292–2300. *In Advances in Neural Information Processing Systems*, 26, 2294–2300. Lake Tahoe, Nevada, USA.

Flamary, R., and Courty, N. 2017. POT: Python Optimal Transport. *Journal of Machine Learning Research* (*JMLR*) 22, 1–8.

Hu, Y., An, Q., Sheu, K., et al. 2018. Single cell multi-omics technology: Methodology and application. *Front. Cell Dev. Biol.* 6, 28.

Kantorovich, L.V. 1942. On the translocation of masses. *Dokl. Akad. Nauk. USSR (N.S.)*. 37, 199–201.

Liu, J., Huang, Y., Singh, R., et al. 2019. Jointly embedding multiple single-cell omics measurements. *In 19th International Workshop on Algorithms in Bioinformatics (WABI); LIPICS* 143, 10:1–10:13.

Macaulay, I.C., Haerty, W., Kumar, P., et al. 2015. G&t-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* 12, 519–522.

Mémoli, F. 2011. Gromov–wasserstein distances and the metric approach to object matching. *Found. Comput. Math*. 11, 417–487.

Nitzan, M., Karaiskos, N., Friedman, N., et al. 2019. Gene expression cartography. *Nature* 576, 132–137.

Peyré, G., and Cuturi, M. 2019. Computational optimal transport. *Found. Trends Mach. Learn*. 11, 355–607.

Peyré, G., Cuturi, M., and Solomon, J. 2016. Gromov-wasserstein averaging of kernel and distance matrices, 2664–2672. *In International Conference on Machine Learning*. PMLR 48, 2664–2672. New York, NY, USA.

Schiebinger, G., Shu, J., Tabaka, M., et al. 2019. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* 176, 928–943.

Singh, R., Demetci, P., Bonora, G., et al. 2020. Unsupervised manifold alignment for single-cell multi-omics data. *In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, 1–10. Association for Computing Machinery (ACM): New York, NY, USA.

Stuart, T., Butler, A., Hoffman, P., et al. 2019. Comprehensive integration of single-cell data. *Cell*. 77, 1888–1902.

Wang, C., and Mahadevan, S. 2009. Manifold alignment without correspondence. *In Twenty-First International Joint Conference on Artificial Intelligence*, 21, 1273–1278. AAAI Press: Pasadena, California, USA.

Welch, J.D., Hartemink, A.J., and Prins, J.F. 2017. Matcher: Manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol*. 18, 138.

Welch, J.D., Kozareva, V., Ferreira, A., et al. 2019. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 177, 1873–1887.

Yang, K.D., Damodaran, K., Venkatchalapathy, S., et al. 2020. Predicting cell lineages using autoencoders and optimal transport. *PLOS Computational Biology* 16, e1007828.

Yang, K.D., and Uhler, C. 2019. Multi-domain translation by learning uncoupled autoencoders. *arXiv preprint arXiv:1902.03515*.

Zappia, L., Phipson, B., and Oshlack, A. 2017. Splatter: Simulation of single-cell rna sequencing data. *Genome Biol*. 18, 1–15.

Address correspondence to:
Dr. Ritambhara Singh
Center for Computational Molecular Biology
Brown University
164 Angell Street, 3rd Floor
Providence, RI 02912
USA

E-mail: ritambhara@brown.edu