Gromov-Wasserstein 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*

Abstract

Data integration of single-cell measurements is critical for understanding cell development and disease, but the lack of correspondence between different types of measurements makes such efforts challenging. Several unsupervised algorithms can align heterogeneous single-cell measurements in a shared space, enabling the creation of mappings between single cells in different data domains. However, these algorithms require hyperparameter tuning for high-quality alignments, which is difficult in an unsupervised setting without correspondence information for validation. We present Single-Cell alignment using Optimal Transport (SCOT), an unsupervised learning algorithm that uses Gromov Wasserstein-based optimal transport to align single-cell multi-omics datasets. We compare the alignment performance of SCOT with state-of-the-art algorithms on four simulated and two real-world datasets. SCOT performs on par with state-of-the-art methods but is faster and requires tuning fewer hyperparameters. Furthermore, we provide an algorithm for SCOT to use Gromov Wasserstein distance to guide the parameter selection. Thus, unlike previous methods, SCOT aligns well without using any orthogonal correspondence information to pick the hyperparameters. Our source code and scripts for replicating the results are available at `https://github.com/rsinghlab/SCOT`.

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 across time. The availability of single-cell measurements that capture various properties of the genome, such as gene expression, chromatin accessibility, DNA methylation, histone modifications, and chromatin 3D conformation, has increased the need for data integration methods capable of combining disparate data types.

Despite the importance of this task, the heterogeneity among single cells presents unique challenges. For example, due to technical limitations, it is hard to obtain multiple types of measurements from the same individual cell. Furthermore, 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 do not rely on either common cells or 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, either among the cells or the features [1–4]. For example, Seurat [4] requires correspondence information in the form of cells from

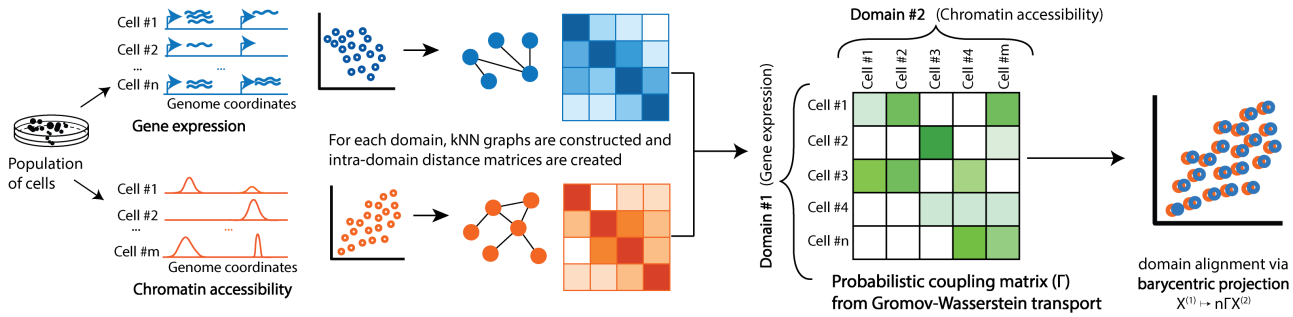


Figure 1: **Schematic of SCOT alignment of single-cell multi-omics data.** A population of cells is aliquoted for different single-cell sequencing assays to capture complementary aspects (e.g. gene expression and chromatin accessibility) of the molecular dynamics. SCOT constructs k-NN graphs based on sample-wise correlations, where vertices represent cells and finds a probabilistic coupling between the samples of each domain which minimizes the distance between the two intra-domain graph distance matrices. Barycentric projection uses this coupling matrix to project one domain onto another.

similar biological state that are shared across the two datasets (known as *anchor points*). Cao *et al.* [5] have shown that such methods cannot perform good alignments under fully unsupervised settings.

Some approaches have tried to align datasets in an entirely unsupervised fashion. One of the earliest attempts, the joint Laplacian manifold alignment (JLMA) algorithm, constructs eigenvector projections based on local k-nearest neighbor graph Laplacians of the data [6]. The generalized unsupervised manifold alignment (GUMA) [7] algorithm seeks a 1–1 correspondence between two datasets based on a local geometry matching term. Liu *et al.* [8] showed that these methods do not perform well on the single-cell alignment task.

Liu *et al.* [8] proposed a manifold alignment algorithm based on the maximum mean discrepancy (MMD) measure, called MMD-MA, which can integrate different types of single-cell measurements. Another method, UnionCom [5], extends GUMA to perform unsupervised topological alignment. MMD-MA aims to match the global distributions of the datasets in a shared latent space, whereas UnionCom emphasizes learning both local and global alignments between the two distributions. Neither method requires any correspondence information either among samples or the features. The respective papers demonstrate state-of-the-art performance on simulated and real datasets. Although these results are encouraging, MMD-MA and UnionCom require that the user specify three and four hyperparameters, respectively. Selecting these hyperparameter values can be difficult and time-consuming in an unsupervised setting.

An emerging number of applications, including several in biology, are using optimal transport to learn a mapping between data distributions [9, 10]. 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. Schiebinger *et al.* [11] use optimal transport to study how gene expression changes over time; they use regularized unbalanced optimal transport to compute differences in gene expression from one time point to the next. ImageAEOT [12] 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 use autoencoders and optimal transport to learn transport maps among multiple domains [13]. However, the application of their method to single-cell datasets requires some form of supervision, like class labels, to preserve the underlying structure during transport.

The classic optimal transport problem requires datasets in the same metric space. Mémoli *et al.*

[14] generalized optimal transport to the Gromov-Wasserstein distance, which compares metric spaces directly instead of comparing samples across spaces. In the natural language processing community, Alvarez *et al.* [10] used this approach to measure similarities between pairs of words across languages to compute the distances between languages. As far as we are aware, the only biological application of Gromov-Wasserstein optimal transport comes from [15], which uses it to reconstruct the spatial organization of cells from transcriptional profiles.

In this paper, we present Single-Cell alignment using Optimal Transport (SCOT), an unsupervised learning algorithm that uses Gromov-Wasserstein-based optimal transport to align single-cell multi-omics datasets (presented schematically in Figure 1). Like UnionCom, SCOT aims to preserve local geometry when aligning single-cell data. SCOT achieves this by constructing a k-nearest neighbor ($k-$NN) graph for each dataset. SCOT then finds a probabilistic coupling between the samples of each dataset that minimizes the distance between the graph distance matrices produced by the k-NN graph. Finally, it uses the coupling matrix to project one single-cell dataset 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 datasets. SCOT aligns all datasets 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 SCOT's hyperparameter tuning in a fully unsupervised setting, when no orthogonal alignment information is available.

2 Methods

SCOT uses Gromov-Wasserstein optimal transport, which preserves local geometry when moving data points form one domain to another. The output of this transport problem is a matrix of probabilities that represent how likely it is that data points from one space correspond to data points in the other space. In this section, we introduce optimal transport followed by its extension to the Gromov-Wasserstein distance. Finally, we present the details of our SCOT algorithm.

We present the case for two datasets: $X = (x_1, x_2, \ldots, x_{n_x})$ from $\mathcal{X}$ and $Y = (y_1, y_2, \ldots, y_{n_y})$ from $\mathcal{Y}$. The datasets have n_x and n_y points, respectively. We do not require any correspondence information but assume there is some underlying shared structure so that the datasets can be meaningfully aligned.

Optimal transport The Kantorovich optimal transport problem seeks to find a minimal cost mapping between two probability distributions [16]. Referring back to the problem of moving a sand pile to fill in a hole, 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. For probability measures μ and ν defined on $\mathcal{X}$ and $\mathcal{Y}$, respectively, this optimal transport problem finds a minimal coupling π that attains

$$\min_{\pi \in \Pi(\nu,\mu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y), \tag{1}$$

where $c(x,y)$ is a cost function and $\Pi(\mu,\nu)$ is the set of couplings of μ and ν given by

$$\Pi(\mu,\nu) = \{\pi \in P(\mathcal{X}\times\mathcal{Y}) : \pi(A\times\mathcal{Y}) = \mu(A) \text{ for } A \subset \mathcal{X},\ \pi(\mathcal{X}\times B) = \nu(B) \text{ for } B \subset \mathcal{Y}\}. \tag{2}$$

Intuitively, the cost function says how many resources it will take to move x to y, and the coupling π assigns a probability $\pi(x,y)$ that x should be moved to y. When the spaces of interest are the same metric space with set $\mathcal{M}$, distance d, and cost function $c(x,y) = d(x,y)^p$, the optimal transport distance

(Equation 1) is equivalent to the $p-$th Wasserstein distance:

$$W^p(\mu,\nu) = \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} d(x,y)^p d\pi(x,y)\right)^{\frac{1}{p}}. \tag{3}$$

The Wasserstein distance measures the distances between probability distributions on a metric space and is commonly used in machine learning applications.

Since we align observed data points, we define the marginals as discrete empirical distributions:

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

where δ_{x_i} is the Dirac measure. Then, the cost function is given as a matrix $C \in \mathbb{R}^{n_x \times n_y}$, e.g. $C_{ij} = \|x_i - y_j\|$, and the set of couplings are the matrices $\Pi(p,q) = \{\Gamma \in \mathbb{R}_+^{n_x \times n_y} : \Gamma \mathbb{1}_{n_y} = p,\ \Gamma^T \mathbb{1}_{n_x} = q\}$. A discrete coupling Γ relates two measures p and q: each row Γ_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 $\Gamma\mathbb{1}_{n_y} = p$ requires that the sum of each row Γ_i is equal to p_i, the probability of sample x_i. The discrete optimal transport problem finds a coupling that minimizes the cost of moving samples through the linear program:

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

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 [17]. Entropy diffuses the optimal coupling, meaning that more masses will be split. Thus, the numerical optimal transport problem is

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

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

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 $K = \exp\left(-\frac{C}{\epsilon}\right)$, element-wise. This solution can be obtained efficiently via Sinkhorn's algorithm, which iteratively computes

$$u \leftarrow p \oslash Kv \text{ and } v \leftarrow q \oslash K^T u, \tag{6}$$

where $\oslash$ denotes element-wise division. This derivation immediately follows from solving the corresponding dual problem for Equation 5 [16].

Gromov-Wasserstein distance Classic optimal transport requires defining a cost function across domains, which can be difficult to implement when the domains are in different metric spaces. Gromov-Wasserstein distance extends optimal transport by comparing distances between samples rather than directly comparing the samples themselves [10]. We assume that we have metric measure spaces $(\mathcal{X}, d_x, \mu)$ and $(\mathcal{Y}, d_y, \nu)$, where d_x and d_y are distances on $\mathcal{X}$ and $\mathcal{Y}$, respectively [14]. Instead of defining a cost function between spaces, Gromov-Wasserstein uses the difference between pairwise distances. Given a cost function $L : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, the Gromov-Wasserstein distance between μ and ν is defined by

$$GW(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} L(d_x(x_1,x_2), d_y(y_1,y_2)) d\pi(x_1,y_1) d\pi(x_2,y_2). \tag{7}$$

The main change from basic optimal transport (Equation 1) to Gromov-Wasserstein (Equation 7) is that we consider the effect of transporting pairs of points rather than single points. Intuitively, $L(d_x(x_1, x_2), d_y(y_1, y_2))$ captures how transporting x_1 to y_1 and x_2 to y_2 would distort the original distances between x_1 and x_2 and between y_1 and y_2. This change ensures that the optimal transport plan π will preserve some local geometry. In the case of $L(x, y) = L_2(x, y) = \frac{1}{2}(x - y)^2$, Gromov-Wasserstein is a distance on the space of metric measure spaces [14].

For the discrete case, we compute pairwise distance matrices D^x and D^y and the fourth order tensor $\mathbf{L} \in \mathbb{R}^{n_x \times n_x \times n_y \times n_y}$, where $\mathbf{L}_{ijkl} = L(D^x_{ik}, D^y_{jl})$. The discrete Gromov-Wasserstein problem is

$$GW(p, q) = \min_{\Gamma \in \Pi(p,q)} \sum_{i,j,k,l} \mathbf{L}_{ijkl} \Gamma_{ij} \Gamma_{kl}. \tag{8}$$

The summation can also be expressed as the inner product $\langle \mathbf{L}(D^x, D^y) \otimes \Gamma, \Gamma \rangle$. Equation 8 is now both non-linear and non-convex 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.* show 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 [18]. In particular, for the case $L = L_2$, the inner product can be computed by

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

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

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

Larger values of ϵ lead to 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 datasets where there are one-to-one correspondences. However, it also yields a harder (more non-convex) optimization problem [10].

Peyré *et al.* [18] propose using a projected gradient descent approach for optimization, where both the projection and the gradient are taken with respect to Kullback-Leibler divergence. These projections are computed via Sinkhorn iterations. Algorithm 1 in the supplement presents the algorithm for $L = L_2$.

Single-Cell alignment using Optimal Transport (SCOT) Our method, SCOT, works as follows. First, we compute the pairwise distances on our data by using a geodesic distance as in [15]. To do this, we use the correlations between data points within each dataset to construct k-NN connectivity graphs. Then we compute the shortest path distance on the graph between each pair of nodes. We set the distance of any unconnected nodes to be the maximum (finite) distance in the graph and rescale the resulting distance matrix by dividing by the 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 with zeros on the diagonal. 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. Our approach is robust to the choice of k (Supplementary Section 1.5)

Since we do not know the true distribution of the original datasets, we follow [10] and set p and q 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/`) [19].

One of the advantages of using optimal transport is that we end up with a coupling matrix Γ with a probabilistic interpretation. The entries of the normalized row $\frac{1}{p_i}\Gamma_i$ are the probabilitie-s that the fixed data point x_i corresponds to each y_j. However, to use the correspondence metrics previously used in the field to evaluate the alignment, we need to project the two datasets into the same space. The Procrustes approach proposed in [10] does not generalize to datasets with different feature and sample dimensions, so we use a barycentric projection:

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

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 [5, 8] or by performing cross-validation [20]. This selection strategy is problematic for truly unsupervised setting, where no correspondence information is available a priori. As a solution, we provide an alternative procedure to learn reasonable alignments based on tracking the Gromov-Wasserstein distance (Equation 8). This procedure is based on our observation that the Gromov-Wasserstein distance serves as a proxy for measuring alignment quality (see Supplementary Figure S5). In this procedure, we alternate between optimizing ϵ and k to minimize the Gromov-Wasserstein distance between the domains (detailed in Algorithm 2 in Supplementary Materials). 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

Simulated datasets We follow Liu *et al.* [8] and benchmark SCOT on three different simulations[1]. All three simulations contain two domains with 300 samples that have been non-linearly projected to 1000- and 2000-dimensional feature spaces, respectively. The three simulations are a bifurcation, a Swiss roll, and a circular frustum (Supplementary Figure S1) with points belonging to three different groups. In addition to these three previously existing simulations, we use Splatter [21] to create simulated single-cell RNA sequencing count data, which we call synthetic RNA-seq. We generate 5000 cells with 1000 genes from three cell groups and reduce the count matrix to the five genes with the highest variances. This count matrix is randomly 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 datasets 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 [8].

Single-cell multi-omics datasets We use two sets of single-cell multi-omics data to demonstrate the applicability of our model to real datasets. Both datasets are generated by co-assays; thus, we have known cell-level correspondence information for benchmarking. The first dataset is generated using the scGEM assay [22], which simultaneously profiles gene expression and DNA methylation. The dataset (Sequence Read Archive accession SRP077853) is derived from human somatic cell samples undergoing conversion to induced pluripotent stem cells (iPSCs). This dataset was also used by Cao *et al.* [5] to demonstrate the

[1] `https://noble.gs.washington.edu/proj/mmd-ma/`

performance of their UnionCom algorithm. The data dimensions are 177×34 for the gene expression data and 177×27 for the chromatin accessibility data.

The second dataset is generated by the SNAREseq assay [23], 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. We pre-process the datasets following Chen *et al.* [23], as follows. We reduce data sparsity and noise in the ATAC-seq data by performing dimensionality reduction using the topic modeling framework cisTopic [24]. The dimensions of the RNA-seq data were reduced using PCA. The resulting input matrices for the SNARE-seq data were of size 1047×19 and 1047×10 for ATAC-seq and RNA-seq, respectively. We unit normalize all real datasets as done in [20].

Evaluation metrics We compare SCOT with the two state-of-the-art unsupervised single-cell alignment methods MMD-MA [8] and UnionCom [5]. None of these methods use any correspondence information for aligning the datasets. However, all datasets 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.* [8]. 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 datasets have group-specific (simulations) or cell-type-specific (real experiments) labels, we also adopt the metric used by Cao *et al.* [5] called "label transfer accuracy" to assess the quality of the cell label assignment. It measures the ability to correctly transfer sample labels from one domain to another based on their neighborhood in the aligned domain. As described in [5], we train a k-nearest neighbor classifier on one of the domains and predict the sample labels in the other domain. The label transfer accuracy 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.

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 $\epsilon \in \{0.0001, 0.0005, 0.001, 0.005, ..., 0.1\}$ and number of neighbors $k \in \{10, 20, 30, 40, \ldots 100, \frac{1}{6}n_x\}$. 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 dataset, 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 Supplementary Materials) to align all the datasets. Since MMD-MA and UnionCom do

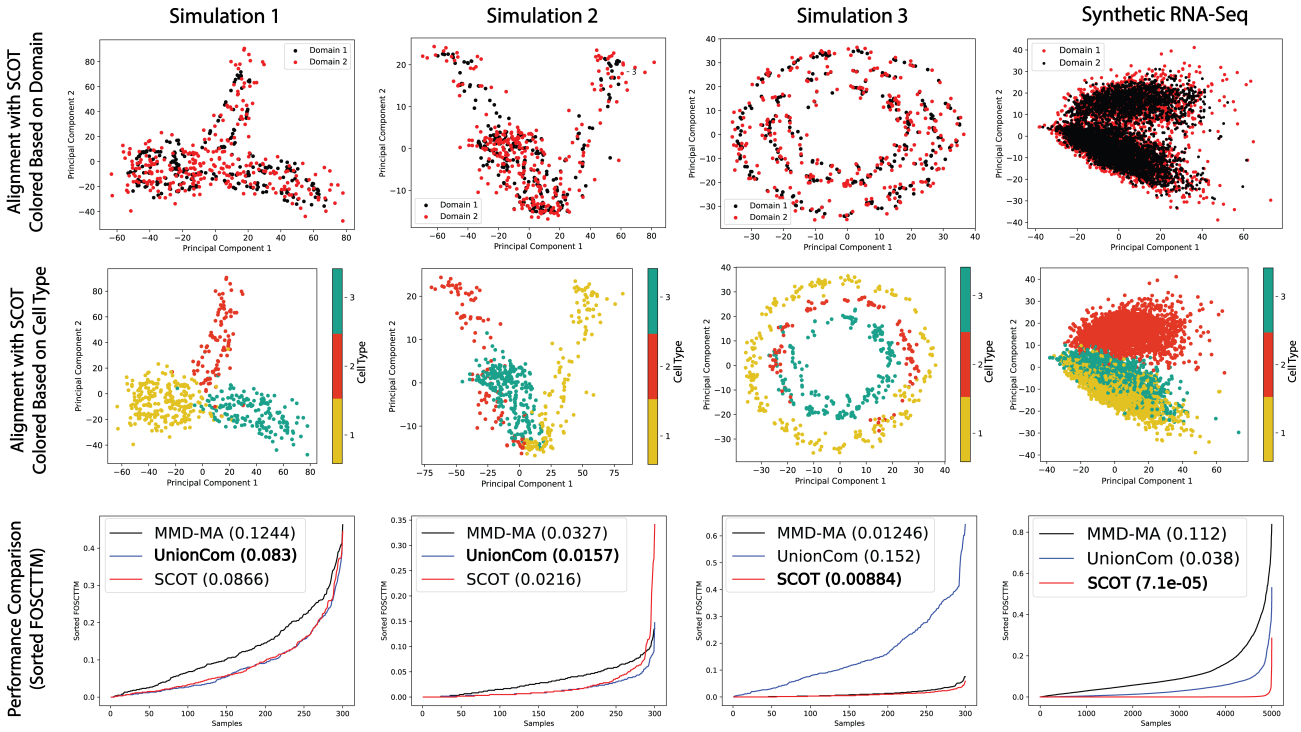


Figure 2: **Aligning simulated datasets.** Each column presents a different simulation. **Top:** our alignment colored by domain (plotted in 2D using PCA). **Middle:** our alignment colored by group. **Bottom:** sorted "fraction of samples closer than the true match" (FOSCTTM) for MMD-MA, UnionCom, and SCOT to visualize the distribution across samples with the average FOSCTTM values in the legend.

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

SCOT successfully aligns the simulated datasets We first compare SCOT's performance with MMD-MA and UnionCom for the four simulation datasets. In this experiment, we select the best performing hyperparameters for each method using the tuning process described in the previous section. In Figure 2, we sort and plot the FOSCTTM score for each sample for the simulations from [8], as well as the synthetic RNA-seq count data from Splatter [21]. Overall, we observe that SCOT consistently achieves one of the lowest average FOSCTTM scores, thereby demonstrating its ability to recover the correct correspondences. We also report the label transfer accuracy results (Table 4) when the first domain is used to train a classifier to predict the labels in the second domain. We observe that SCOT consistently yields high label transfer accuracy scores, indicating that samples are correctly mapped to their assigned groups.

SCOT gives state-of-the-art performance for single-cell multi-omics alignment Next, we apply our method to real single-cell sequencing data. Overall, SCOT gives the lowest average FOSCTTM measure in comparison to MMD-MA and UnionCom (Figure 3, last column) and recovers accurate 1–1 corre-

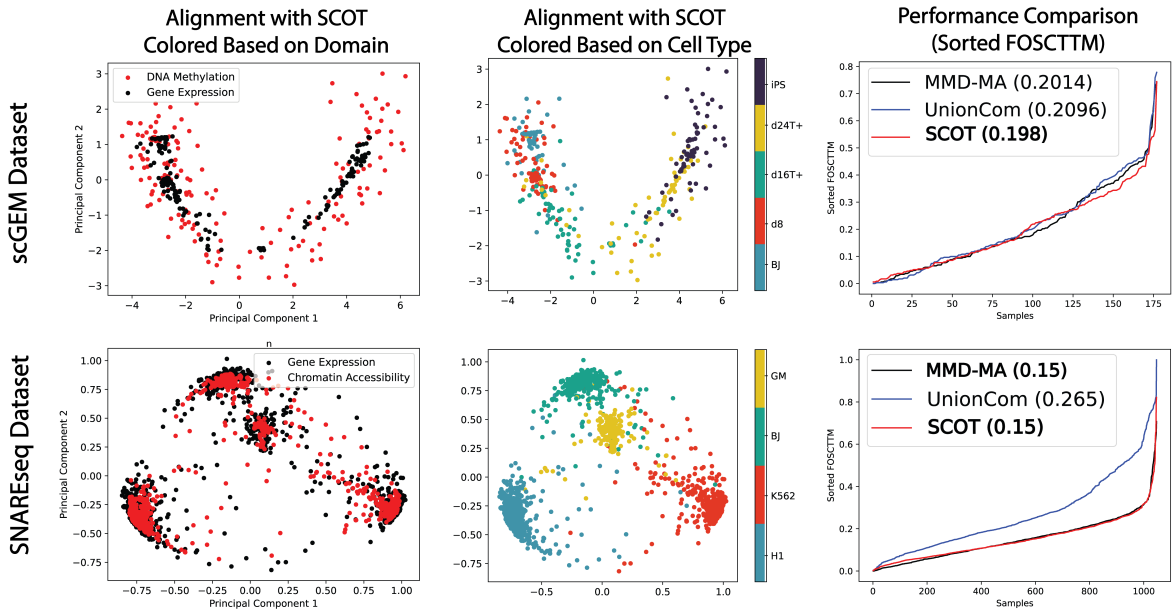


Figure 3: **Aligning real world single-cell sequencing dataset.** Each row presents a different real-word single-cell sequencing dataset. **Left:** our alignment colored based by domain (plotted in 2D using PCA). **Middle:** our alignment colored by cell-type. **Right:** sorted "fraction of samples closer than the true match" (FOSCTTM) for MMD-MA, UnionCom, and SCOT order to visualize the distribution across samples with the average FOSCTTM values in the legend.

spondences in single-cell datasets. For the scGEM data, we report label transfer accuracy using the DNA methylation domain for predicting the cell-type labels in the gene expression domain. For the SNARE-seq dataset, we use the gene expression domain for predicting cell labels in the chromatin accessibility domain. SCOT yields the best label transfer accuracy result on SNAREseq dataset and performs comparably to the other methods for scGEM (Table 4.)

While MMD-MA and UnionCom project both datasets to a shared low-dimensional space, SCOT projects one dataset onto the other. We project SCOT in both directions for all datasets, but we do not observe a significant difference in performance between the two directions (Supplementary Materials Table 3).

SCOT's alternative unsupervised hyperparameter tuning procedure achieves good alignments We compare the alignment performances in Table 2 when given by SCOT's alternative tuning procedure guided by the Gromov-Wasserstein distance and MMA-MA's and UnionCom's default parameters. SCOT returns nearly the same alignments for simulated data and only marginally worse alignments for real data. In contrast, MMD-MA and UnionCom fail to align some of the simulated and all real datasets with the

	Sim. 1	Sim. 2	Sim. 3	Syn. 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

Table 1: Alignment performance by label transfer accuracy ($k = 5$).

	Sim. 1	Sim. 2	Sim. 3	Syn. 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

Table 2: Alignment performance by FOSCTTM scores for SCOT chosen by lowest Gromov-Wasserstein distance, default MMD-MA, and default UnionCom for simulated and real datasets.

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.

SCOT's computation speed scales well with the number of samples We compare SCOT's running times with the baseline methods for the best performing hyperparameters on the synthetic RNA-seq dataset by varying the number of cells. We run CPU computations on an Intel Xeon e5-2670 with 16GB memory and GPU computations on a single NVIDIA GTX 1080ti with VRAM of 11GB. SCOT's running time scales similarly to that of MMD-MA, even though SCOT runs on a CPU and MMD-MA runs on a GPU (Figure 4). Both methods scale better than the GPU-based UnionCom implementation.

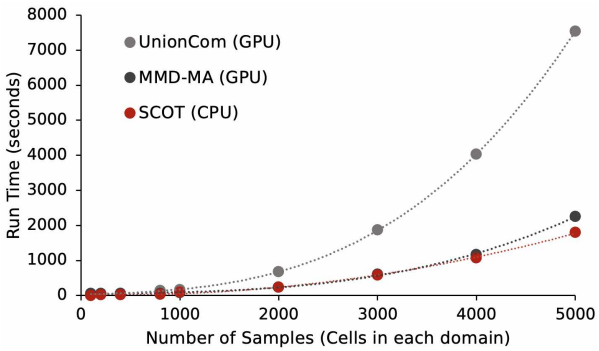


Figure 4: **Runtime comparisons with growing sample size** Dotted lines are polynomial trend lines.

5 Discussion

We have demonstrated that SCOT, which uses Gromov Wasserstein optimal transport for unsupervised single-cell multi-omics data integration, performs on par with UnionCom and MMD-MA. Our formulation of a coupling matrix based on matching graph distances is somewhat similar to UnionCom's initial step; however, UnionCom only matches sample-to-sample distances, while 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 SCOT's state-of-the-art performance. Furthermore, SCOT's 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 fully unsupervised setting.

To visualize and measure alignment, we project data into the same space through barycentric projection, but there are other ways to use the coupling matrix to infer alignment. For example, the coupling matrix could also be used 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.

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

Funding William S. Noble's contribution to this work was funded by NIH award U54 DK107979. Bjorn Sandstede was partially supported by NSF awards 1714429 and 1740741. Rebecca Santorella is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1644760.

References

[1] Matthew Amodio and Smita 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] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 77(7):1888–1902, 2019.

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

[6] Chang Wang and Sridhar Mahadevan. Manifold alignment without correspondence. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[7] Zhen Cui, Hong Chang, Shiguang Shan, and Xilin Chen. Generalized unsupervised manifold alignment. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2014.

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

[9] Alfred Galichon. A survey of some recent applications of optimal transport methods to econometrics. *Econometrics Journal*, 20(2), 2017.

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

[11] 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.

[12] 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.

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

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

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

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

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

[18] 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.

[19] Rémi Flamary and Nicolas Courty. Pot python optimal transport library, 2017.

[20] Ritambhara Singh, Pinar Demetci, Giancarlo Bonora, Vijay Ramani, Choli Lee, He Fang, Zhijun Duan, Xinxian Deng, Jay Shendure, Christine Disteche, et al. Unsupervised manifold alignment for single-cell multi-omics data. *BioRxiv*, 2020.

[21] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):1–15, 2017.

[22] 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.

[23] 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.

[24] Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. cisTopic: cis-regulatory topic modelling on single-cell ATAC-seq data. 16(5):397–400, 2018.

Supplementary Materials for "Gromov-Wasserstein optimal transport to align single-cell multi-omics data"

1 SCOT algorithm

As described in Section 2, SCOT takes in two datasets X and Y and constructs $k-$NN graphs on each dataset to create the distance matrices D_x and D_y. Then, it finds the coupling Γ that minimizes the Gromov-Wasserstein distance. Finally, the coupling matrix is used to project one domain onto the other. In Algorithm 1, we present the full SCOT algorithm, including the Gromov-Wasserstein calculation, which uses the projections proposed in [18].

Algorithm 1: Gromov-Wasserstein Alignment

Input: Datasets $X,\ Y$. Regularization ϵ. Number of neighbors k.
// Compute graph distances $D_x,\ D_y$;
p = Uniform(X), q = Uniform(Y);
$D_{xy} \leftarrow D_x^2 \mathbb{1}_{n_y}^T + \mathbb{1}_{n_x} q (D_x^2)^T$;
while *not converged* **do**
  // Compute cost matrix
  $\hat{D}_\Gamma \leftarrow D_{xy} - 2 D_x \Gamma D_y^T$;
  // Perform Sinkhorn iterations
  $u \leftarrow \mathbb{1}\,,\ K \leftarrow \exp\{-\hat{D}_\Gamma / \epsilon\}$;
  while *not converged* **do**
    $u \leftarrow p \oslash K v,\ v \leftarrow q^T \oslash K^T u$;
  end
  $\Gamma \leftarrow \text{diag}(u) K \text{diag}(v)$;
end
Return: $n_x \Gamma Y$

1.1 Unsupervised Hyperparameter Selection Procedure for SCOT

As detailed in Section 2, one way to select SCOT hyperparameters in the absence of correspondence information or validation dataset, is to use the Gromov-Wasserstein distance as a proxy for alignment quality. Here, we present the procedure for carrying this out, where we alternate between the hyperparameters k and ϵ, and fix one to tune the other:

Algorithm 2: Unsupervised hyperparameter search algorithm for SCOT.

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

1.2 Visualization of Original Data Sets

In the main text, we display the alignment results performed by SCOT. Here, we visualize the original datasets:

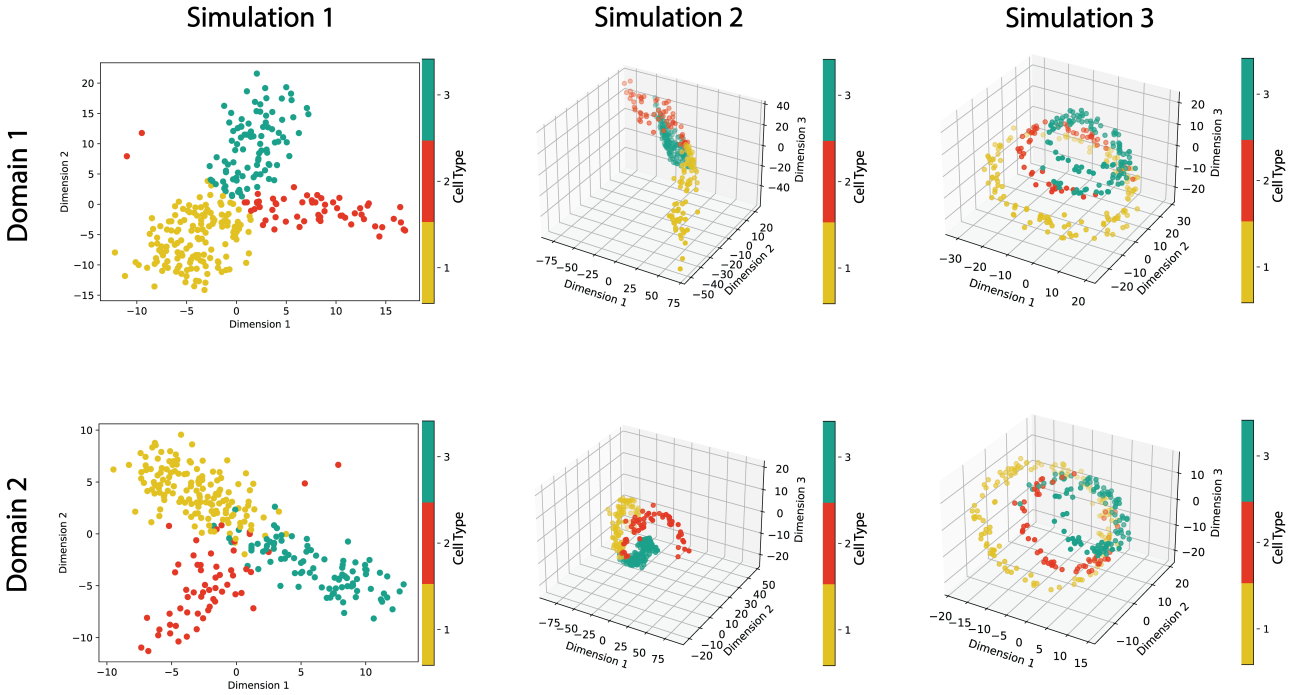


Figure S1: **Original simulation data visualized before alignment.** Data was generated by Liu et al [8] and retrieved from `https://noble.gs.washington.edu/proj/mmd-ma/`. Each simulation set has two domains. Their MDS projections in two dimensional and three dimensional space are visualized here. The first set of simulations form a branched tree in two dimensional space (first column); the second set of simulations form Swiss roll in three dimensional space (second column); and lastly, the third set of simulations form a circular frustum.

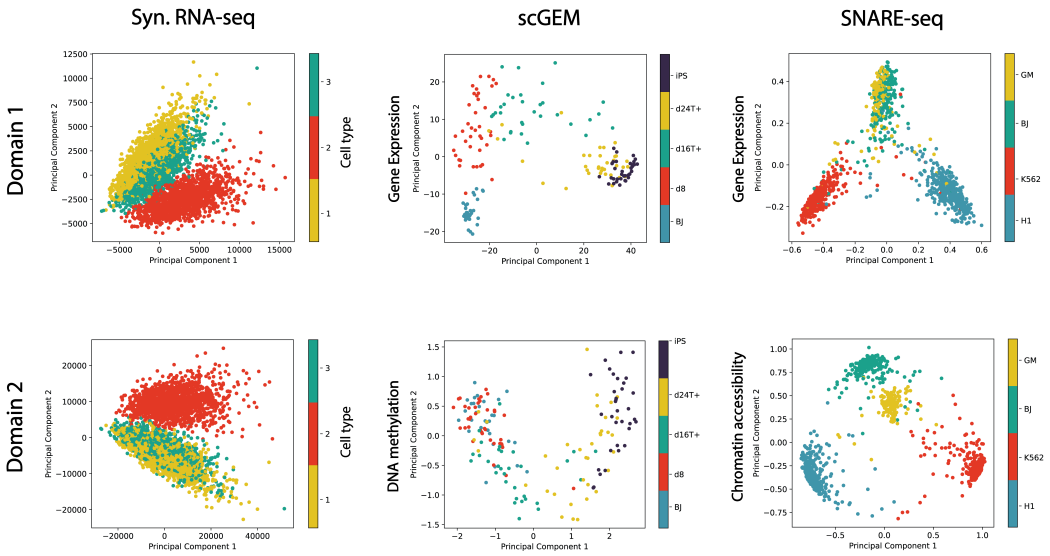


Figure S2: **Original synthetic RNA-seq and real world single-cell data visualized before alignment.** We use Splatter [21] to generate a count matrix with 5000 cells and 1000 genes from three cell groups. We reduce the dataset to the 5 genes with the highest variances, and then use random matrices to project the data to new dimensions $p_1 = 50$ and $p_2 = 500$. Here we visualize the two domains with PCA projections for this dataset as well as the real world single-cell sequnecing datasets

1.3 Barycentric Projections in Both Directions

While MMD-MA and UnionCom project both datasets to a shared low-dimensional space, SCOT projects one dataset onto the other. We project SCOT in both directions for all datasets, but we do not observe a significant difference in performance between the two directions. In Table 3, we present the averaged FOSCTTM values for barycentric projection in both directions (domain 1 projected onto domain 2, as well as domain 2 projected onto domain 1).

	Domain 1 onto Domain 2	**Domain 2 onto Domain 1**
Sim. 1	0.0872	0.0866
Sim. 2	0.0216	0.0230
Sim. 3	0.0088	0.0091
Syn. RNA-Seq	7.12×10^{-5}	7.68×10^{-5}
scGEM	0.2118	0.1978
SNARE-seq	0.1496	0.1514

Table 3: **Best mean FOSCTTM for each direction of the barycentric projection for all datasets.** The method is robust to the direction of the projection.

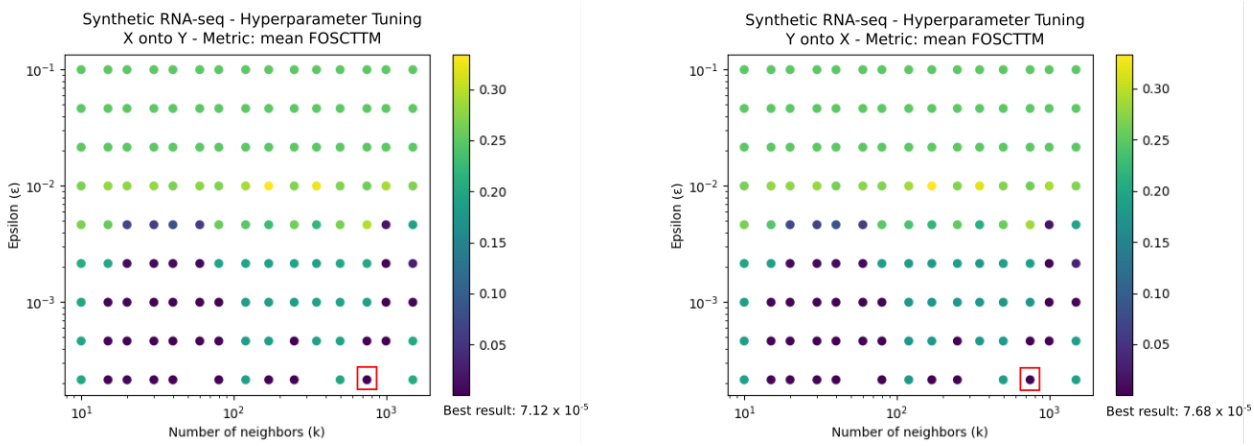
1.4 Label Transfer Accuracy with the Second Domain used in Training

In Table 4, we present the label transfer accuracies when the first domain is used as the training set. Here we report the opposite direction.

	Sim. 1	Sim. 2	Sim. 3	Syn. RNA-Seq	scGEM	SNAREseq
SCOT	**0.953**	**0.987**	**0.957**	**0.998**	0.435	0.936
MMD-MA	0.893	0.806	0.933	0.899	**0.638**	**0.967**
UnionCom	0.912	0.97	0.62	0.97	0.508	0.717

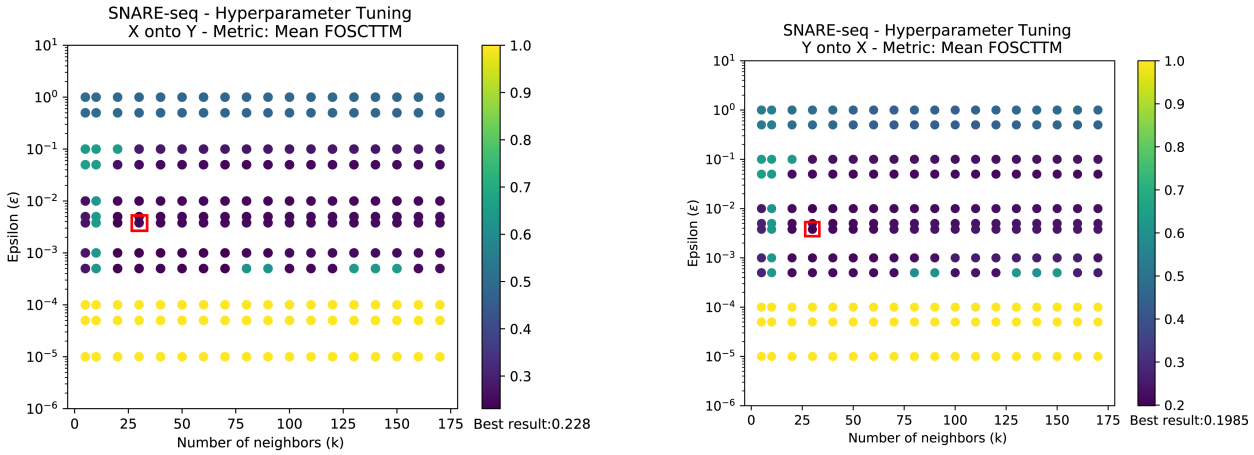
Table 4: Alignment performance by label transfer accuracy ($k = 5$) for SCOT, MMD-MA, and UnionCom for simulated and real datasets when the second domain is used for training.

1.5 Hyperparameter Tuning for SCOT



(a) (b)

Figure S3: **Hyperparameter optimization results for synthetic RNA-seq dataset.** Mean FOSCTTM metric was used to assess performance (indicated by color). **(a)** Results when first domain (X) is projected onto second domain (y). **(b)** Results when second domain (y) is projected onto first domain (X). The algorithm is largely robust to the choice of k. For both projections, the best performing hyperparameter setting was $\epsilon = 0.000215,\ k = 750$. The hyperparameter combination that yielded the best performance is highlighted with red square. For ease of visualization, a subset of the values are plotted.



(a) (b)

Figure S4: **Hyperparameter optimization results for SNARE-seq dataset.** Mean FOSCTTM metric was used to assess performance (indicated by color). **(a)** Results when chromatin accessibility domain (X) is projected onto gene expression domain (y). **(b)** Results when expression domain (y) is projected onto chromatin accessibility domain (X). The algorithm is largely robust to the choice of k. For both projections, the best performing hyperparameter setting was $\epsilon = 0.0038, k = 30$. The hyperparameter combination that yielded the best performance is highlighted with red square. For ease of visualization, a subset of the ϵ values are plotted.

1.6 Visualizing the Empirical Relationship between Gromov-Wasserstein Distance and Correspondence in Alignment as Measured by Average FOSCTTM

We observe that lower values of the Gromov-Wasserstein distance tend to correspond to lower average FOSCTTM values. Below, we have plotted the Gromov-Wasserstein values against average FOSCTTM for each dataset over a range of parameter values.

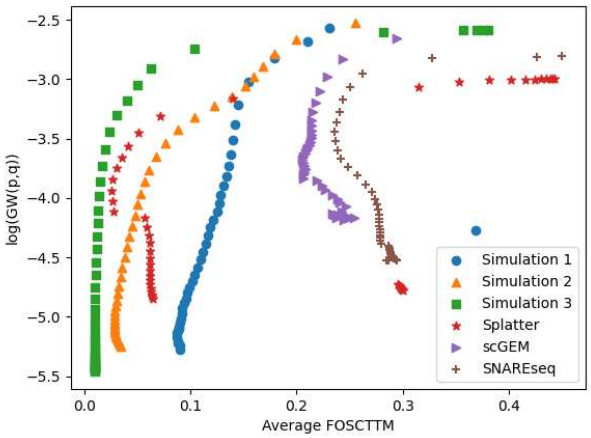


Figure S5: **Gromov-Wasserstein distance vs average FOSCTTM values** for all datasets with a range of ϵ parameter values (k fixed at $\min(50, 0.2n_x)$).

1.7 Label Transfer Accuracy for Automatic Alignment

In Table 2, we report the average FOSCTTM values for SCOT when chosen by lowest Gromov-Wasserstein distance and default parameters for MMD-MA and UnionCom. In the tables below, we also report the label transfer accuracy scores.

	Sim. 1	**Sim. 2**	**Sim. 3**	**Syn. 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

Table 5: Alignment performance by label transfer accuracy ($k = 5$) when the first domain is used for training for SCOT, MMD-MA, and UnionCom for simulated and real datasets.

	Sim. 1	**Sim. 2**	**Sim. 3**	**Syn. RNA-Seq**	**scGEM**	**SNAREseq**
SCOT	**0.93**	**0.98**	**0.957**	**0.998**	**0.571**	**0.736**
MMD-MA	0.893	0.9	0.757	0.299	0.225	0.557
UnionCom	0.91	0.943	0.143	0.971	0.113	0.292

Table 6: Alignment performance by label transfer accuracy ($k = 5$) when the second domain is used for training for SCOT, MMD-MA, and UnionCom for simulated and real datasets.