

Genome analysis

SCOT+: a comprehensive software suite for single-cell alignment using optimal transport

Colin Baker^{1,*}, Tuan Pham¹, Pinar Demetci², Quang Huy Tran^{3,4}, Ievgen Redko⁵, Bjorn Sandstede⁶, Ritambhara Singh¹

¹Center for Computational Molecular Biology, Brown University, Providence, RI 02906, United States

²Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard, Cambridge, MA 02412, United States

³Université Bretagne Sud, CNRS, IRISA, Vannes, 56000, France

⁴CMAF, Ecole Polytechnique, Palaiseau, 91120, France

⁵Noah's Ark Lab, Huawei Technologies, Paris, 92100, France

⁶Division of Applied Mathematics, Brown University, Providence, RI, 02906, United States

*Corresponding author. Center for Computational Molecular Biology, Brown University, Providence, RI 02906, United States. E-mail: colin_baker@brown.edu.

Associate Editor: Kieran Campbell

Abstract

Summary: New advances in single-cell multi-omics experiments have allowed biologists to examine how various biological factors regulate processes in concert on the cellular level. However, measuring multiple cellular features for a single cell can be quite resource-intensive or impossible with the current technology. By using optimal transport (OT) to align cells and features across disparate datasets produced by separate assays, Single Cell alignment using Optimal Transport+ (SCOT+), our unsupervised single-cell alignment software suite, allows biologists to align their data without the need for any correspondence. SCOT+ implements a generic optimal transport solution that can be reduced to multiple different previously studied OT optimization procedures including SCOT, SCOTv2, SCOOTR, and AGW for single cell, each of which provides state-of-the-art single-cell alignment performance. Outside of giving a unified framework to interact with prior formulations, the generality of SCOT+ optimization naturally gives rise to a new OT loss, Unbalanced Augmented Gromov-Wasserstein (UAGW), and a corresponding optimizer. With our user-friendly website and tutorials, this new package will help improve biological analyses by allowing for more accurate downstream analyses on multi-omics single-cell measurements.

Availability and implementation: Our algorithm is implemented in Pytorch and available on PyPI and GitHub (<https://github.com/scotplus/scotplus>). Additionally, we have many tutorials available in a separate GitHub repository (https://github.com/scotplus/book_source) and on our website (<https://scotplus.github.io/>).

1 Introduction

Single-cell multi-omic study integrates various modalities of omics, such as genomic, transcriptomic, proteomic, or epigenomic data, at the single-cell resolution (Stoeckius *et al.* 2017, Ma *et al.* 2020, Genomics 2021). Such an integration is crucial for biologists to gain a deeper understanding of cell regulation and development. However, simultaneously measuring multi-modal data for single cells is either difficult or expensive due to technical limitations (Ma *et al.* 2020). Thus researchers rely on separately sequenced single-cell measurements that have to be computationally integrated.

As a result, several computational methods have been developed to integrate multiple single-cell datasets that draw from similar cell populations (Stuart *et al.* 2019, Korsunsky *et al.* 2019, Liu *et al.* 2019, Dou *et al.* 2022). A single-cell dataset is usually represented as a count matrix where the “sample” (row) is a specific cell and the “feature” (column) is a genomic measurement on this cell. To integrate two single-cell datasets, existing methods reconstruct a unified matrix that combines samples and feature sets from both. This

requires the assumption that there exist some meaningful relationships between each set of features. However, this is a difficult problem because such relationships are not observable across disparate sample sets without 1–1 correspondence of cells. In order to combat such issues, supervised methods like Seurat (Stuart *et al.* 2019) and Harmony (Korsunsky *et al.* 2019) utilize explicit, ground-truth sample or feature correspondences between the two datasets to extrapolate the rest of the alignment. On the other hand, unsupervised methods like MMD-MA (Liu *et al.* 2019) and BindSC (Dou *et al.* 2022) attempt to estimate a multi-model dataset without any prior sample or feature correspondence information.

Among the existing unsupervised single-cell multi-omic alignment methods, optimal transport (OT) approaches (Cao *et al.* 2020, 2021, Demetci *et al.* 2022a,b, 2023, 2024) have shown state-of-the-art performance for aligning separately profiled multi-omic datasets. Optimal transport constructs a probabilistic correspondence matrix between the data points of two input domains in order to match them in the most cost effective way possible (Peyré and Cuturi 2019). A popular

view of the problem is to imagine moving a pile of sand to fill in a hole through the least amount of work.

In the single-cell context, OT-based approaches allow for both distribution-free alignments and additional probabilistic interpretation of the coupling matrix prior to integration. Prior methods like UnionCom (Cao *et al.* 2020), Pamona (Cao *et al.* 2021), and uniPort (Cao *et al.* 2022) also leverage these properties to compute single-cell alignments. These methods have proven useful to, for example, co-embed and compare single-cell datasets across different conditions (Zeppilli *et al.* 2023). However, many of these different OT methods for alignment are scattered throughout different packages. The lack of a single, easy to use package that unifies such methods has made OT less accessible to the single-cell community.

Given the advantages of OT-based methods, we introduce Single Cell alignment using Optimal Transport+ (or SCOT+), a software suite that leverages three different OT formulations to integrate single-cell datasets. (i) Gromov-Wasserstein (GW) OT, applied for the single-cell integration task as SCOT (Demetci *et al.* 2022a), is best used in contexts where the geometric structure of the two datasets is important, while potential feature relationships between the two domains to be aligned could be nonlinear but are not as relevant to alignment (Mémoli 2011, Cuturi 2013, Vayer *et al.* 2019). (ii) Co-Optimal Transport (COOT), applied as SCOOTR (Demetci *et al.* 2023), is best used in contexts where potential feature relationships are close to linear and well-known so that linear supervision might aid alignment more meaningfully (Redko *et al.* 2020). Finally, (iii) Augmented Gromov-Wasserstein (AGW) is a convex combination of GW and COOT distance that allows for feature supervision without any restriction to linearity and therefore brings together the benefits of both formulations at the cost of an extra hyperparameter (Demetci *et al.* 2024). Additionally, each of these formulations is applicable in cases where there is a disproportionate distribution of cell types in either dataset. Unbalancing a formulation, in this case, improves alignment, as shown in the SCOTv2 and UCOOT papers (Sejourné *et al.* 2021, Demetci *et al.* 2022b, Thual *et al.* 2022, Tran *et al.* 2022). The unbalanced version of AGW, which we term as UAGW, is a novel construct that arose from our generalization of the single-cell OT pipelines for SCOT+. It is the first OT formulation that allows for unbalancing in the feature space while retaining GW-like structure.

SCOT+ serves as a unified wrapper to access the best solutions to each of these prior formulations (GW, UGW, COOT, UCOOT, AGW), so that users have many options available for integrating different types of single-cell datasets. Rather than having to handle significant hyperparameter tuning to appropriately align their data, with SCOT+, users can choose the formulation that more intuitively applies to their data for the best results. Additionally, SCOT+ utilizes a single, generic optimizer to solve all of its formulations, making it more user-friendly and accessible. We leverage our insight that all other OT formulations mentioned above are partial functions of UAGW, which we introduce for the first time in this paper and SCOT+ solves on the backend.

We show that SCOT+ performs well on real world single-cell datasets, both with respect to replicating co-assay data and correctly mapping cell type clusters across domains for separately sequenced single-cell measurements. We also discuss the potential for downstream analysis of the estimated multi-omic mappings for the community.

2 Methods

2.1 SCOT+ optimal transport setup

Given X and Y , the count matrices—or other datasets of shape n samples, d features with continuous or discrete numerical values—from two single-cell measurements, SCOT+ seeks to learn a probabilistic pairing between their cells (and features) that integrates the two datasets. We assume below that X has n_x cells and d_x features, while Y has n_y cells and d_y features. Each of the GW, COOT, and AGW formulations in SCOT+ treat their input domains X and Y as probability distributions. By minimizing the cost of transporting cells from one domain to another, these methods produce a *coupling matrix* that is the learned pairing between them. We first examine the standard optimal transport problem (Villani 2009, Cuturi 2013) that is defined as:

$$\min_{P \in \Pi(\mu_1, \mu_2)} \langle C, P \rangle \quad (1)$$

where C is some cost matrix such that C_{ij} is the cost of transporting cell i to cell j , and μ_1, μ_2 are the marginals for the solution to the problem. A coupling matrix P , which we force to have marginals μ_1 and μ_2 by inclusion in the set $\Pi(\mu_1, \mu_2) = \{P \in \mathbb{R}_+^{n_x \times n_y} : P_{\#1} = \mu_1, P_{\#2} = \mu_2\}$, maps the cells across disparate sample spaces. Here, $\langle C, P \rangle := \sum_{ij} C_{ij} P_{ij}$. Note that we use $P_{\#1}$ to represent the marginal distribution of P in the first domain and $P_{\#2}$ to represent the marginal distribution of P in the second domain. In the single-cell context and throughout this document, we initialize the marginals to be uniform; in principle, they can also be established by cell type to mitigate non-uniform stratification in a sample. However, this problem can also be generalized (Chizat *et al.* 2018) as follows:

$$\min_P (\langle C, P \rangle + \epsilon \text{KL}(P | \mu_1 \otimes \mu_2) + \rho_1 \text{KL}(P_{\#1} | \mu_1) + \rho_2 \text{KL}(P_{\#2} | \mu_2)) \quad (2)$$

Here, the first Kullback–Leibler (KL) divergence term $\text{KL}(P | \mu_1 \otimes \mu_2)$ is for regularization. It controls the sparsity in P , making the problem more computationally tractable. Note that $\mu_1 \otimes \mu_2$ denotes the outer product here. For example, if $\mu_1, \mu_2 \in \mathbb{R}^{n \times m}$, $\mu_1 \otimes \mu_2 \in \mathbb{R}^{n \times m \times n \times m}$. In this case, $\mu_1 \otimes \mu_2 \in \mathbb{R}^{n \times n \times m \times m}$.

Additionally, the $\rho_1 \text{KL}(P_{\#1} | \mu_1)$ and $\rho_2 \text{KL}(P_{\#2} | \mu_2)$ terms relax the constraints on P 's marginals that were previously quite restrictive and initialization-dependent. They allow flexibility of cell-cell mappings when there is a disproportionate cell type representation across the datasets. For example, $P_{\#1} = \sum_j P_{ij}$.

Another useful feature is that the user can provide supervision when learning the coupling matrix by adding the term $\beta \langle D, P \rangle$ to Equation (2). Here, D is some user-constructed matrix according to known cell-cell or feature-feature mappings and β is a coefficient for the magnitude of supervision (Vayer *et al.* 2019).

Each of the SCOT+ formulations break down into solving the general problem presented in Equation (2) at each iteration of block coordinate descent (see Section 1, available as [supplementary data](#) at *Bioinformatics Advances* online). At a high level, the SCOT+ optimizer breaks down a general loss function into smaller problems by only optimizing over one coupling matrix at a time. Each of these smaller problems is considered an iteration of block coordinate descent, where each iteration holds two coupling matrices constant and

optimizes the third. As a result, SCOT+ currently utilizes Sinkhorn's algorithm (Chizat *et al.* 2018) as its main engine to solve these smaller problems. To retrieve an alignment from a coupling matrix P mapping X to Y , we take the weighted average of the genomic measurements in Y according to the weights in P (Ferradans *et al.* 2014, Courty *et al.* 2017). This produces a new matrix $\hat{Y}$ with the d_y features of Y on the n_x cells of X :

$$\hat{Y}_i = \sum_{j=1}^{n_y} \frac{P_{ij}}{P_{\#1i}} Y_j \quad (3)$$

where the index i refers to cells in X and j refers to the cells in Y . More information on this projection is presented in Section 5, available as [supplementary data](#) at *Bioinformatics Advances* online. In the next section, we take the general formulation designed for SCOT+ in Equation (2) and show how it can be easily extended to yield six different OT formulations relevant for different single-cell integration tasks.

2.2 Different SCOT+ formulations

First, the GW OT formulation that SCOT+ utilizes to align count matrices X and Y is the following minimization problem (Mémoli 2011, Peyré *et al.* 2016):

$$\begin{aligned} \text{GW}(D_X, D_Y) := \min_{P, P'} (\langle \mathcal{L}(D_X, D_Y), P \otimes P' \rangle \\ + \epsilon \text{KL}(P \otimes P', \mu_1 \otimes \mu'_1 \otimes \mu_2 \otimes \mu'_2)) \end{aligned} \quad (4)$$

In this case, D_X and D_Y are intra-domain distance matrices that capture the local geometry of each domain and $\mathcal{L}(D_X, D_Y)_{ijkl} = |D_{X_{ij}} - D_{Y_{kl}}|^2$. We calculate D_X and D_Y by constructing a graph of cells in Euclidean space for each domain and computing the shortest path distances between two cells in the graph. SCOT+ aligns these two graphs by mapping their internal local geometries onto each other. This formulation encourages the minimization of the sum implied by the first term in Equation (4):

$$\sum_{i,j,k,l} |D_{X_{ij}} - D_{Y_{kl}}|^2 P_{ik} P'_{jl} \quad (5)$$

so that SCOT+ outputs an alignment of samples encoded by P and P' which minimizes the distortions between the local geometry in the domain X and the one in Y . Note that this formulation can also be marginally relaxed with the additional terms $\rho_1 \text{KL}(P_{\#1} \otimes P'_{\#1} | \mu_1 \otimes \mu'_1) + \rho_2 \text{KL}(P_{\#2} \otimes P'_{\#2} | \mu_2 \otimes \mu'_2)$ (Sejourne *et al.* 2021). While GW and its unbalanced counterpart (UGW) are very effective for aligning cells, they do not explicitly produce any feature mappings. Such feature mappings, e.g. gene to chromatin regions, could be useful for biological hypothesis generation. Therefore, our next formulation addresses this gap.

SCOT+ also includes the COOT formulation (Redko *et al.* 2020) that learns two coupling matrices, one for cells and the other for features:

$$\begin{aligned} \text{COOT}(X, Y) := \min_{P, Q} (\langle \mathcal{L}(X, Y), P \otimes Q \rangle + \\ \epsilon \text{KL}(P \otimes Q | \mu_1^s \otimes \mu'_1 \otimes \mu_2^f \otimes \mu'_2)) \end{aligned} \quad (6)$$

In this case, Q is a coupling matrix for the features of X and Y so that we minimize the joint transport of cells and

features from one domain to another. In other words, some cell located at i, j in Q can be interpreted as some metric of correspondence for feature i in X and feature j in Y . Additionally, we introduce the superscripts s and f to denote that, for example, μ_1^s is the desired (often uniform) first marginal for P , supported on the distribution of samples (s) of X (1). We can also relax this formulation by adding $\rho_1 \text{KL}(P_{\#1} \otimes Q_{\#1} | \mu_1^s \otimes \mu'_1) + \rho_2 \text{KL}(P_{\#2} \otimes Q_{\#2} | \mu_2^f \otimes \mu'_2)$ (Tran *et al.* 2022) resulting in its unbalanced form (or UCOOT). Unfortunately, the COOT formulation can be sensitive to outliers and the magnitude of features in each domain (Tran *et al.* 2022).

Finally, the newest formulation utilized by SCOT+, AGW OT, merges these two loss functions (GW and COOT) as follows (Demetci *et al.* 2024):

$$\begin{aligned} \text{AGW}(X, Y) := \min_{P, P', Q} (\alpha \text{GW}(D_X, D_Y) + \\ (1 - \alpha) \text{COOT}(X, Y)) \end{aligned} \quad (7)$$

This formulation is more robust to outliers as the GW term preserves the local geometry and it also produces feature coupling matrices (Demetci *et al.* 2024). We introduce a novel relaxed AGW formulation by individually relaxing the GW and COOT terms as in the above descriptions. Our general optimization approach designed for SCOT+ lends itself to this novel unbalanced version of AGW (or UAGW). More details are provided in Section 1, available as [supplementary data](#) at *Bioinformatics Advances* online.

The availability of each of these formulations through a single SCOT+ package allows users to tailor their integration details to the specifics of their dataset. For example, one user with disproportionate cell types and a prior on the feature relationships might use UCOOT, but another with uniform cell types and no prior on the features might use GW.

2.3 Data preprocessing

2.3.1 GW and UGW settings

To evaluate performance in the GW and UGW settings, we considered aligning scRNA-seq with scATAC-seq data. We aimed to recover correct matchings between the two modalities. For the scRNA-seq data, raw counts were normalized using scanpy (Wolf *et al.* 2018), log-transformed, and subjected to principal component analysis (PCA) to capture variation in gene expression. For the scATAC-seq data, after standard quality control, we applied pycisTopic (González-Blas *et al.* 2023) to perform latent Dirichlet allocation (LDA)-based topic modeling, extracting latent factors corresponding to co-accessible chromatin regions. These preprocessing strategies are widely used to obtain lower-dimensional embeddings while preserving biologically meaningful information (Wolf *et al.* 2018, González-Blas *et al.* 2023) as done in (Demetci *et al.* 2024). These lower-dimensional embeddings were then used as input for the matching.

2.3.2 AGW and UAGW settings

For AGW scenarios, we used CITE-seq dataset of (Stoeckius *et al.* 2017), which jointly profiles RNA and surface protein abundances. Our goal was to assess whether AGW can align both features and samples simultaneously. We first selected human cells with at least 500 unique molecular identifiers (UMIs) mapped to human genes and subsampled 1000 cells for computational efficiency. For proteins, we used the pre-processed antibody-derived tag (ADT) counts provided in the

dataset. For RNA, we log-normalized counts using Seurat (Stuart *et al.* 2019) and then identified the top 50 most variable genes and proteins per modality. These features were used as input for the alignment procedure.

2.3.3 Supervised UAGW settings

To demonstrate the usefulness of UAGW, we evaluated its performance on a cross-species alignment task. We analyzed gene expression datasets from Tosches *et al.* (2018) and Bhattacharjee *et al.* (2019), which profile the lizard and mouse brain, respectively. We aimed to test whether the method could recover biologically meaningful correspondences across species despite evolutionary divergence. Both datasets were L2-normalized to ensure comparability and then used as input for alignment. To provide supervision, we curated a set of homologous genes from the literature and incorporated them as supervised features, allowing the method to supervise known cross-species relationships while aligning the full datasets.

2.4 Hyperparameter tuning

For a generic SCOT+ alignment, we generally recommend tuning by first running a grid search over ϵ_{gw} and ϵ_{coot} until a reasonable level of density in the sample and feature coupling matrices is achieved. A coupling matrix might be too dense if it maps all cells in one domain to the same feature values in the other domain; it might be too sparse if each cell is mapped to some unique cell in the other domain. From here, we recommend running another grid search over ρ_{gw} and ρ_{coot} , where in each case we set $\rho_x^1 = \rho_x^2$, depending on whether the two cell populations are unbalanced in some way (i.e. disproportionate cell type representation, disproportionate feature contributions/relevance). Finally, when tuning with AGW, we recommend following the same steps, but additionally searching over a narrow range for α (i.e. $0.1 \leq \alpha \leq 0.5$) in order to provide further resolution on feature relationships.

For example, in our UGW application to the PBMC dataset, we searched over $\rho_{\text{gw}}^1 = \rho_{\text{gw}}^2 \in \{0.01, 0.1, 1\}$ jointly with $\epsilon_{\text{gw}} \in \{10^{-3}, 10^{-2.9}, \dots, 10^{-1}\}$. The specific values we used for each experiment in this document are listed in Section 3, available as [supplementary data](#) at *Bioinformatics Advances* online. More details for such tuning can be found on the SCOT+ website.

Some methods for calculating pairwise distance matrices (in GW settings) may also require a small search. For example, we typically construct a k -nearest-neighbor graph for each dataset in Euclidean space and compute the shortest path between each pair of samples, as in Demetci *et al.* (2023, 2024). Here, one may want to tune k ; we chose 110 for both PBMC applications and 100 for both CITE-seq applications.

Finally, applying supervision to any SCOT+ formulation calls for the construction of a supervision matrix D and the input of a corresponding scalar β . Typically, we construct some matrix D where $D_{ij} = 0$ when samples (or features, when supervising Q) are known to be related and $D_{ij} = 1$ otherwise. If these correspondences are well known, then β can be taken arbitrarily large; in this sense, β should somewhat reflect the confidence of the user in how they choose to bias the alignment procedure. For example, in Fig. 2, we construct the supervision matrix in this way and use $\beta = 1$. Our website contains examples of how scaling β affects performance when supervision is well validated.

2.5 Memory and runtime

Using the SCOT+ default settings for UGW, GW, AGW, and UAGW, we benchmarked memory utilization and runtime across simulated datasets with the number of samples varying from 50 to 10 000 against Pamona and UnionCom. Overall, SCOT+ methods scale comparably to, or better than, these baseline methods (Section 4, available as [supplementary data](#) at *Bioinformatics Advances* online).

3 Results

We now demonstrate the versatility of SCOT+ for the integration of real-world single-cell datasets. We specifically focus on balanced and unbalanced scenarios where GW and AGW are applicable.

3.1 Cell-cell alignment on balanced datasets

As formalized above, SCOT+ GW OT aligns disparate single-cell measurements in a way that maintains the cell-cell similarities upon integration. In many cases where we do not need feature-feature relationships, GW can provide meaningful cell-cell alignments.

Specifically, balanced GW is applicable in cases where we have similar cell type proportions across domains. To showcase this, we use a PBMC co-assay dataset (Genomics 2021) that simultaneously measures scRNA-seq and scATAC-seq data. This dataset has a 1–1 correspondence between single cells across these two domains, which we use only to quantify the alignment performance.

To execute an alignment between these domains, we split a subset of the PBMC dataset into two separate count matrices with the same set of 2407 cells. Then, we applied PCA to the scRNA-seq data to select the 50 most contributing components and topic modeling (González-Blas *et al.* 2023) to the scATAC-seq data to embed the domain into a smaller space with 50 topics. This reduces the dimension of the data to remove noise.

Next, we run SCOT+'s GW formulation and obtained a sample-level alignment across our two separate domains. A barycentric projection using the optimized coupling matrix returns a new single-cell dataset with the same set of original samples from the scATAC-seq data, but the features of the scRNA-seq data. We call this data the projected scATAC-seq data, which is visualized in Fig. 1B.

To quantify the quality of this projected scATAC-seq data, we calculated the fraction of samples closer than the true match (FOSCTTM) using the ground-truth sample correspondence of the PBMC data. FOSCTTM measures the fraction of samples in the projected scATAC-seq data that are closer to a given sample in the original scRNA-seq data than its true match (Demetci *et al.* 2022a). When this metric is lower, it means that each sample in the respective domains is relatively close to its correct match. In this example, we attained a FOSCTTM of 0.12, meaning that the projected scATAC-seq data had a similar geometry and shape to the original scRNA-seq data. This FOSCTTM value is comparable to the previous results in (Demetci *et al.* 2024).

3.2 Cell-cell alignment on unbalanced datasets

The balanced formulation of GW OT fails to account for disproportionate cell type representation that usually exists in real-world datasets (Demetci *et al.* 2022b). The unbalanced case of UGW (implemented in SCOT+) closely models what

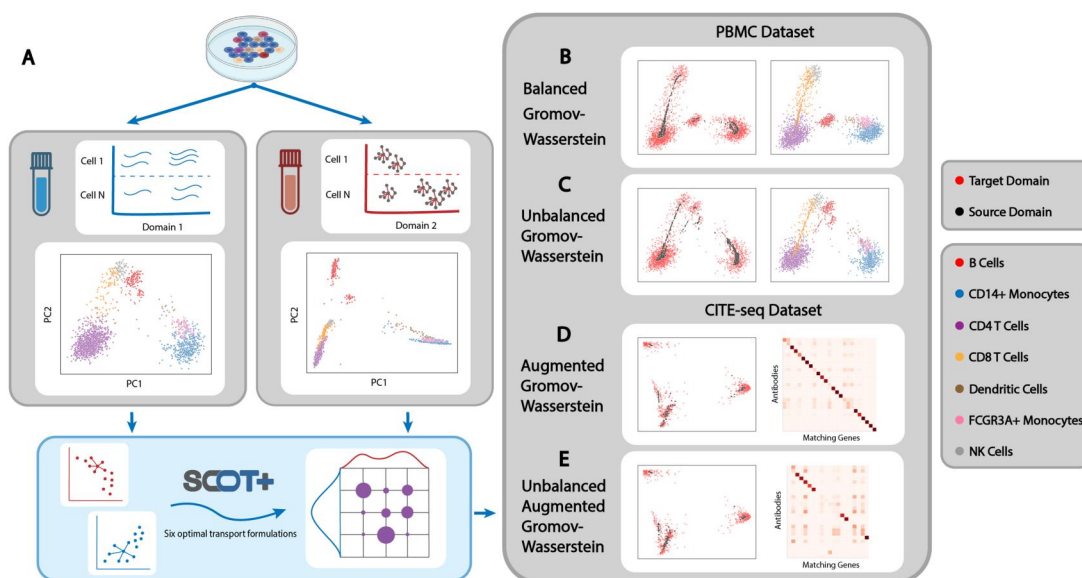


Figure 1 An example of the SCOT+ workflow on a PBMC dataset (Genomics 2021) and a CITE-seq dataset (Stoeckius et al. 2017). All plotted results highlight the projection of the data onto the top two principal components. (A) SCOT+ workflow. (B) Results of running GW on the PBMC data (scRNA-seq and scATAC-seq) in a balanced fashion, followed by a barycentric projection of the accessibility domain onto the gene expression domain. Note that the rightmost GW figure highlights the retained clustering of each cell type when we examine the full aligned dataset. GW reconstructs the gene expression domain from accessibility with low alignment error. (C) Results of running UGW on the same two domains after downsampling each cell type by a random amount. UGW reconstructs each cell type cluster with high accuracy. (D) Results of balanced AGW on the two CITE-seq domains (scRNA-seq and ADT), followed by a barycentric projection of the gene expression domain onto the ADT domain. AGW reconstructs the ADT domain from accessibility low alignment error and 73% of the feature matrix mass assigned to ground-truth feature relationships. (E) Results of running UAGW on the same two domains when we remove five ADT features. UAGW reconstructs the ADT domain with low alignment error and 52% feature mass assigned to ground-truth relationships.

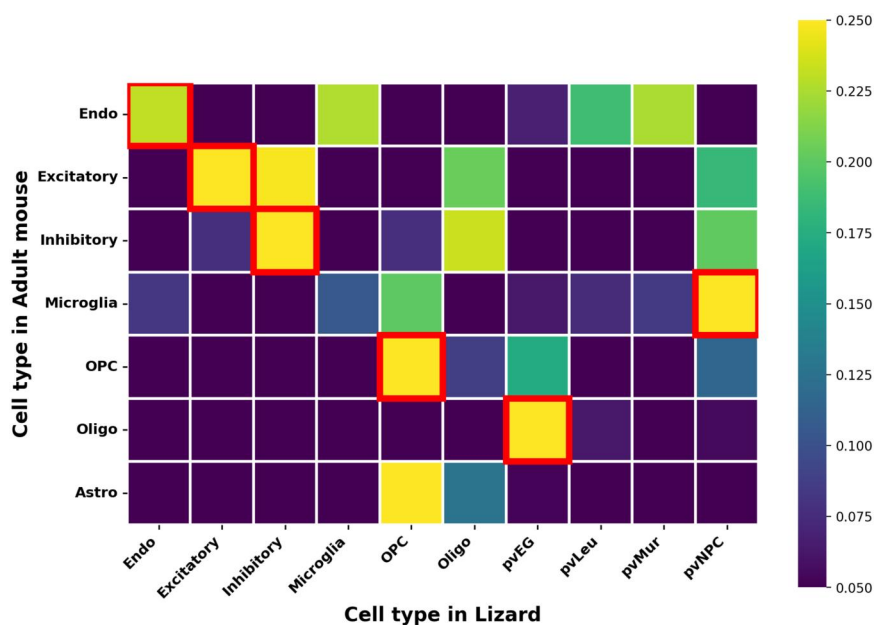


Figure 2 Heatmap showing cross-species cell type alignment between adult mouse (rows) and lizard (columns) under full feature supervision. Outlined boxes denote the strongest correspondence for each adult mouse cell type, highlighting the most likely lizard counterpart. Notably, astrocytes do not have corresponding cell types in lizards.

we expect biologists will observe under real-world conditions in separately sequenced single-cell datasets.

To simulate the unbalanced setting, we systematically downsampled both the scRNA-seq and scATAC-seq data from our previous example by cell type. In particular, we drew a random fraction between 0 and 0.5 for each cell type and discarded the corresponding number of samples in the

scRNA-seq data. We then repeated the same process with different fractions for the scATAC-seq data and used UGW to compute the alignment and barycentric projections shown in Fig. 1C.

In this case, we used a metric called label transfer accuracy (LTA) (Cao et al. 2020) to score our alignment as we do not have exact 1–1 correspondences anymore. To compute LTA,

we trained a k -nn classifier on the original domain to be projected, where k is the actual number of cell types. Next, the score is determined by the accuracy of this classifier on the projected domain. In this example, SCOT+ achieved an LTA of 92.5%, meaning that it succeeds in transporting cell types from one domain to another in a more realistic single-cell integration setting.

3.3 Feature-based alignment on balanced datasets

While GW and UGW work well in cases where we mostly care about obtaining a cell-cell alignment, AGW works better in cases where there are well-defined feature relationships (Demetci et al. 2024). In particular, AGW learns both cell-cell and feature-feature mappings, which is relevant for hypothesis generation in biology. Furthermore, this joint formulation allows for correspondence on either level to inform the other.

To show the usability of SCOT+’s AGW, we used a CITE-seq dataset (Stoeckius et al. 2017), which has co-assayed scRNA-seq and ADT (antibody) data. We downsampled the features of both domains to 25 gene-antibody pairs that are known to be biologically related. After running AGW on these two datasets, we got a FOSCTTM of 0.08, but more importantly, we recovered the accurate feature map displayed in Fig. 1D, which has 73% of its mass assigned to biologically validated antibody-gene pairs (Demetci et al. 2024). This example illustrates the capacity for SCOT+ to learn meaningful feature relationships whilst also integrating single domain datasets. It also demonstrates the additional cell-level alignment power gained from integrating feature information.

3.4 Feature-based alignment on unbalanced datasets

AGW’s unbalanced counterpart, named UAGW in SCOT+, is a new construct that arose from the generalization of our optimization framework. With the fully generic version of our solver, we can get impressive cell-cell alignments in the unbalanced feature setting.

To display this functionality, we downsampled the features of Y from our previous experiment to contain only 20 antibodies. After running UAGW on this data, where we balanced the GW term and left the COOT term unbalanced, we recovered a cell-level alignment as seen in Fig. 1E. This alignment had FOSCTTM 0.09 and a reasonable feature coupling matrix, which had 52% of its mass assigned to known antibody-gene pairs (Demetci et al. 2024). If we had instead generated the coupling matrix randomly, the expected mass assigned to these known pairs would instead be 4%, indicating that UAGW assigned significant additional mass to these known relationships. In some settings without a known correspondence, cells with exaggerated mass relative to random could help generate hypotheses about feature relationships. As a result, we found that UAGW can successfully recover unbalanced feature relationships. UAGW additionally allows for supervision of the feature matrix in an unbalanced fashion, such that we can improve cell-level alignments.

3.5 Feature supervision on sample alignment with UAGW

In addition to the more transparent use case of applying cell type labels directly in the supervision matrix to encourage cell type-preserving alignments, supervision can also be applied to the feature coupling matrix Q . To demonstrate this

Table 1 SCOT+ UAGW’s sample (i.e. cell) alignment performance with increasing supervision on feature (i.e. gene) alignments.^a

Level of supervision	Cell type matching accuracy
No supervision	57.14%
With supervision	71.43%

^a In this case, accuracy is computed as the fraction of cell types in the mouse dataset that were aligned most often to the matching cell type in the lizard dataset.

feature of SCOT+, we performed an experiment aligning transcriptomic data between the pallium of the bearded lizard and the mouse prefrontal cortex, as in (Demetci et al. 2024), but under unbalanced feature settings. The lizard dataset consists of 4187 cells and 12 379 genes, while the mouse dataset contains 6296 cells and 21 000 genes, with 10 816 homologous genes shared between them. We used these homologous pairs to construct a feature-level cost matrix and evaluated alignment with and without this matrix as supervision. After obtaining the alignment matrix, we aggregated cell-cell alignments by cell type to form an alignment matrix at the cell type level. For each cell type, we then identified the counterpart cell type to which it was most strongly aligned, and determined the accuracy of cell type matching by comparing predicted alignments to known biological correspondences. As highlighted in Table 1, biologically related cell types (arranged along the first diagonal in Fig. 2) were reasonably well aligned (71.43%) under feature supervision, although they were already reasonably well aligned prior to supervision. This result suggests that the alignment preserves some true biological structure across datasets, providing insights into conserved cellular processes during evolution.

3.6 Comparison with existing baselines

To put these results in context, we additionally evaluated UnionCom and Pamona on the same balanced and unbalanced PBMC datasets, and applied the same procedure to a SNARE-seq dataset (Chen et al. 2019). As shown in Section 2, available as supplementary data at Bioinformatics Advances online, SCOT+ outperforms UnionCom and Pamona, achieving lower FOSCTTM in the balanced setting and higher LTA in the unbalanced setting for both datasets.

4 Discussion

With SCOT+, we believe that the single-cell community will be better able to understand and utilize the end-to-end alignment procedure of OT methods. Since it uses only one generic solution in the background, SCOT+ unifies single-cell optimal transport into the most digestible framework since the topic’s inception for the single-cell domain.

SCOT+ will enable biologists to make additional progress toward inferring gene regulatory networks as well as producing more robust cross-modal genomic models. Specifically, these downstream analyses require co-assay datasets that are available through the use of SCOT+, but are otherwise hard to generate experimentally. Additionally, biologists will be able to better infer cell trajectories during development or cellular evolution across species.

In the future, we are looking into the theoretical implications of unbalancing the AGW formulation as we have in this package. Additionally, we are continuing to expand SCOT+

to additional real-world scenarios to aid users in tuning SCOT+ for their specific needs.

Supplementary material

Supplementary material is available at *Bioinformatics Advances* online.

Conflicts of interest

All authors declare that they have no conflicts of interest.

Funding

This work was supported by the National Institutes of Health (NIH) (grant numbers: 1 R01 DC020478-01A1, T32 MH115895). It was also supported by Brown University's Undergraduate Teaching and Research Award (UTRA) and the Brown Center for Computation and Visualization.

Data availability

All of these results, including thorough tutorials on the theory and application of each formulation as well as practical examples on real-world data, are available in downloadable Jupyter notebooks on our website, hosted at <https://scotplus.github.io/>. The website additionally contains documentation that will help users use this tool on their own data. The PBMC dataset can be found at <https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-re-moved-through-cell-sorting-3-k-1-standard-2-0-0>. The CITE-seq dataset can be found at <https://www.nature.com/articles/nmeth.4380>.

References

- Bhattacharjee A, Djekidel MN, Chen R *et al*. Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. *Nat Commun* 2019;10:4169.
- Cao K, Bai X, Hong Y *et al*. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* 2020;36:i48–56.
- Cao K, Hong Y, Wan L. Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics* 2021;38:211–9.
- Cao K, Gong Q, Hong Y *et al*. A unified computational framework for single-cell data integration with optimal transport. *Nat Commun* 2022;13:7419.
- Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 2019;37:1452–7.
- Chizat L, Peyré G, Schmitzer B *et al*. Scaling algorithms for unbalanced transport problems. *Math Comp* 2018;87:2563–609.
- Courty N, Flamary R, Tuia D *et al*. Optimal transport for domain adaptation. *IEEE Trans Pattern Anal Mach Intell* 2017;39:1853–65.
- Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport. In: *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc., 2013, 26.
- Demetci P, Santorella R, Sandstede B *et al*. SCOT: single-cell multi-omics alignment with optimal transport. *J Comput Biol* 2022a;29:3–18.
- Demetci P, Santorella R, Chakravarthy M *et al*. SCOTv2: single-cell multiomic alignment with disproportionate cell-type representation. *J Comput Biol* 2022b;29:1213–28.
- Demetci P, Tran QH, Redko I *et al*. Jointly aligning cells and genomic features of single-cell multi-omics data with co-optimal transport. *BioRxiv*, <https://doi.org/2022.11.09.515883>, 2023, preprint: not peer reviewed.
- Demetci P, Tran QH, Redko I *et al*. Breaking isometric ties and introducing priors in Gromov-Wasserstein distances. In: *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research, Vol. 238. PMLR, 2–4 May 2024, 298–306.
- Dou J, Liang S, Mohanty V *et al*. Bi-order multimodal integration of single-cell data. *Genome Biol* 2022;23:112.
- Ferradans S, Papadakis N, Peyré G *et al*. Regularized discrete optimal transport. *SIAM J Imaging Sci* 2014;7:1853–82.
- 10X Genomics. PBMC from a healthy donor. In: *No Cell Sorting (3k) Epi Multiome ATAC + Gene Expression Dataset Analyzed Using Cell Ranger ARC 2.0.0*. Pleasanton, CA: 10x Genomics, 2021.
- González-Blas C, De Winter S, Hulselmanns G *et al*. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 2023;20:1355–67.
- Harris CR, Millman KJ, van der Walt SJ *et al*. Array programming with NumPy. *Nature* 2020;585:357–62.
- Korsunsky I, Millard N, Fan J *et al*. Fast, sensitive, and accurate integration of single-cell data with harmony. *Nat Methods* 2019;16:1289–96.
- Liu J, Huang Y, Singh R *et al*. Jointly embedding multiple single-cell omics measurements. In: *Leibniz International Proceedings in Informatics (LIPIcs)*, Vol. 143. Dagstuhl: Schloss Dagstuhl—Leibniz-Zentrum für Informatik, 2019, 10:1–10:13.
- Luo J, Yang D, Wei K. Improved complexity analysis of the sinkhorn and greenkhorn algorithms for optimal transport. *arXiv*, 2023, preprint: not peer reviewed.
- Ma A, McDermaid A, Xu J *et al*. Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol* 2020;38:1007–22.
- Mémoli F. Gromov-Wasserstein distances and the metric approach to object matching. *Found Comput Math* 2011;11:417–87.
- Paszke A, Gross S, Massa F *et al*. Pytorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc., 2019, 8024–35.
- Peyré G, Cuturi M. Computational optimal transport. *Foundations Trends Mach Learn* 2019;11:355–607.
- Peyré G, Cuturi M, Solomon J. Gromov-Wasserstein averaging of kernel and distance matrices. In: *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48. New York, NY: PMLR, 2016, 2664–72.
- Redko I, Vayer T, Flamary R *et al*. Co-optimal transport. In: *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., 2020, 33.
- Sejourne T, Vialard F, Peyré G. The unbalanced Gromov-Wasserstein distance: conic formulation and relaxation. In: *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., 2021, 34.
- Stoeckius M, Hafemeister C, Stephenson W *et al*. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;14:865–8.
- Stuart T, Butler A, Hoffman P *et al*. Comprehensive integration of single-cell data. *Cell* 2019;177:1888–902.e21.
- Thual A, Tran QH, Zemskova T *et al*. Aligning individual brains with fused unbalanced Gromov-Wasserstein. In: *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., 2022, 35.
- Tosches MA, Yamawaki TM, Naumann RK *et al*. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* 2018;360:881–8.
- Tran QH, Janati H, Courty N *et al*. Unbalanced co-optimal transport. In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. Washington, DC: AAAI Press, 2023, 37.
- Vayer T, Chapel L, Flamary R *et al*. Optimal transport for structured data with application on graphs. In: *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, CA: PMLR, 2019, 97.

- Villani C. *Optimal Transport: Old and New*, Volume 338 of *Grundlehren Der Mathematischen Wissenschaften*. Berlin Heidelberg: Springer-Verlag, 2009.
- Virtanen P, Gommers R, Oliphant TE *et al.*; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods* 2020;17:261–72.
- Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;19:15.
- Zeppilli S, Gurrola AO, Demetci P *et al.* Single-cell genomics of the mouse olfactory cortex reveals contrasts with neocortex and ancestral signatures of cell type evolution. In: *Nature Neuroscience*, Vol. 28. London: Springer Nature, 2025, 937–48.