Unsupervised Integration of Single-Cell Multi-omics Datasets with Disproportionate Cell-Type Representation

Pınar Demetçi[1,2], Rebecca Santorella[3], Björn Sandstede[3], and Ritambhara Singh[1,2](✉)

[1] Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA
{pinar_demetci,ritambhara}@brown.edu
[2] Department of Computer Science, Brown University, Providence, RI 02912, USA
[3] Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
{rebecca_santorella,bjorn_sandstede}@brown.edu

Abstract. Integrated analysis of multi-omics data allows the study of how different molecular views in the genome interact to regulate cellular processes; however, with a few exceptions, applying multiple sequencing assays on the same single cell is not possible. While recent unsupervised algorithms align single-cell multi-omic datasets, these methods have been primarily benchmarked on co-assay experiments rather than the more common single-cell experiments taken from separately sampled cell populations. Therefore, most existing methods perform subpar alignments on such datasets. Here, we improve our previous work Single Cell alignment using Optimal Transport (SCOT) by using unbalanced optimal transport to handle disproportionate cell-type representation and differing sample sizes across single-cell measurements. We show that our proposed method, SCOTv2, consistently yields quality alignments on five real-world single-cell datasets with varying cell-type proportions and is computationally tractable. Additionally, we extend SCOTv2 to integrate multiple ($M \geq 2$) single-cell measurements and present a self-tuning heuristic process to select hyperparameters in the absence of any orthogonal correspondence information.

Available at: http://rsinghlab.github.io/SCOT.

Keywords: Single-cell sequencing · Multi-omics · Data integration · Unsupervised learning · Optimal transport · Unbalanced alignment

1 Introduction

The ability to measure multiple aspects of the single-cell offers the opportunity to gain critical biological insights about cell development and diseases. However, many existing single-cell sequencing technologies cannot be simultaneously

I. Pe'er (Ed.): RECOMB 2022, LNBI 13278, pp. 3–19, 2022.
https://doi.org/10.1007/978-3-031-04749-7_1

applied to the same cell, resulting in multi-omics datasets sampled from distinct cell populations. While these measurements can be analyzed separately, integrating them prior to analysis can help explain how different molecular views interact and regulate cellular functions. Unfortunately, single-cell assays that measure different molecular aspects in separately sampled cell populations lack direct sample–sample and feature–feature correspondences across these measurements. This lack of correspondences makes it hard to use integration methods that require some shared information to perform single-cell alignment [4]. Therefore, *unsupervised* single-cell multi-omics data alignment methods are crucial for integrative single-cell data analysis.

Several unsupervised methods [4,10,12,15], including our previous work, SCOT [9], have shown state-of-the-art performance for integrating different single-cell measurement domains. Since these methods were mainly evaluated on real-world co-assay datasets (with 1–1 correspondence between cells across domains), our understanding of their performance on datasets obtained from experiments that are not co-assays is limited. Such experiments perform separate sampling to measure distinct genomic features, like gene expression and 3D chromatin conformation. Therefore, their datasets can consist of varying proportions of cell-types across different measurements, creating cell-type imbalance and lacking 1–1 cell correspondences. We hypothesize that alignment methods that perform well on co-assay datasets may not effectively handle the differences in cell-type proportions of the commonly available non-co-assay datasets. Indeed, a recent method, Pamona [5], extended our SCOT framework and used partial Gromov-Wasserstein (GW) optimal transport to allow for missing or underrepresented cell-types in one domain when performing alignment. It showed that current integration methods [4,9,12,15] tend to perform worse under such settings.

We present SCOTv2, a novel extension of SCOT that can effectively align both co-assay and non-co-assay datasets using a single framework. It uses *unbalanced* GW optimal transport to align datasets with disproportionate cell-types while only introducing one additional hyperparameter. This unbalanced framework relaxes the constraint that each point must be mapped with its original mass during the optimal transport. Specifically, an underrepresented cell-type in one domain can be transported with more mass to match the proportion of that cell-type in the other domain and vice-versa. The SCOTv2 framework is summarized in Fig. 1. We demonstrate that SCOTv2 aligns datasets with imbalance in cell-type representations better than state-of-the-art baselines and computationally scales as well as the fastest methods. Furthermore, we extend SCOTv2 to integrate single-cell datasets with more than two measurements, making it a multi-omics alignment tool. We perform alignments of five real-world single-cell datasets, with both simulated and natural cell-type imbalance as well as two and more than two domains ($M \geq 2$), demonstrating SCOTv2's applicability across a wide range of scenarios. Finally, similar to the previous version, we present a self-tuning heuristic process to select hyperparameters for SCOTv2 without any corresponding information like cell-type annotations or matching cells or features in truly unsupervised settings.

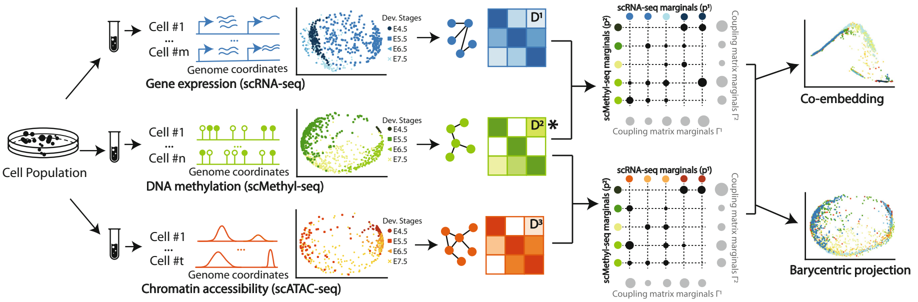
Cell Population
Cell #1
...
Cell #m
Genome coordinates
Gene expression (scRNA-seq)
Dev. Stages
● E4.5
■ E5.5
◄ E6.5
× E7.5
D^1
Cell #1
...
Cell #n
Genome coordinates
DNA methylation (scMethyl-seq)
Dev. Stages
● E4.5
■ E5.5
◄ E6.5
× E7.5
D^2 *
Cell #1
...
Cell #t
Genome coordinates
Chromatin accessibility (scATAC-seq)
Dev. Stages
● E4.5
■ E5.5
◄ E6.5
× E7.5
D^3
scRNA-seq marginals (p^1)
scMethyl-seq marginals (p^2)
Coupling matrix marginals Γ^2
Coupling matrix marginals Γ^1
scRNA-seq marginals (p^1)
scMethyl-seq marginals (p^2)
Coupling matrix marginals Γ^2
Coupling matrix marginals Γ^1
Co-embedding
Barycentric projection

Fig. 1. Overview of SCOTv2 on scNMT-seq dataset [8], which contains unbalanced cell-type representation across three domains - RNA expression, chromatin accessibility, and DNA methylation. SCOTv2 selects an anchor domain (denoted with *) and aligns other measurements to it. First, it computes intra-domain distances matrices D^m for $m = 1, 2, 3$, which are used to solve for correspondence matrices between the anchor and other domains. The circle sizes in the matrices depict the magnitude of the correspondence probabilities or how much mass to transport. Unbalanced GW relaxes the mass conservation constraint, so the transport map does not need to move each point with its original mass. Finally, it either co-embeds the domains into a common space or uses barycentric projections to project them onto the anchor domain.

2 Method

Optimal transport finds the most cost-effective way to move data points from one domain to another. One can imagine it as the problem of moving a pile of sand to fill in a hole through the least amount of work. Our previous framework SCOT [9] uses Gromov-Wasserstein optimal transport, which preserves local geometry when moving data points from one domain to another. The output of SCOT is a matrix of probabilities that represent how likely it is that data points from one modality correspond to data points in the other.

Here, we reintroduce the SCOT formulation to integrate M domains (or single-cell measurements) $X^m = (x_1^m, x_2^m, \dots x_{n_m}^m) \in \mathbb{R}^{d_m}$ for $m = 1, \dots M$ with n_m data points (or cells) each. For each dataset, we define a marginal distribution p^m, which can be written as an empirical distribution over the data points:

$$p^m = \sum_{i=1}^{n_m} p_i^m \delta_{x_i}. \tag{1}$$

Here, δ_{x_i} is the Dirac measure. For SCOT, we choose these distributions to be uniform over the data.

Gromov-Wasserstein optimal transport performs the transport operation by comparing distances between samples rather than directly comparing the samples themselves [2]. Therefore, for each dataset, we compute the intra-domain distance matrix D^m. Next, we construct k-NN graphs based on correlations between data points and use Dijkstra's algorithm to compute the shortest path distance on the graph between each pair of nodes. Finally, we connect all unconnected nodes by the maximum finite distance in the graph and set D^m to be the matrix resulting from normalizing the distances by this maximum.

For two datasets and a cost function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we compute the fourth-order tensor $\mathbf{L} \in \mathbb{R}^{n_x \times n_x \times n_y \times n_y}$, where $\mathbf{L}_{ijkl} = L(D^1_{ik}, D^2_{jl})$. Intuitively, L quantifies how transporting a pair of points x^1_i, x^1_k onto another pair across domains, x^2_j, x^2_l, distorts the original intra-domain distances and helps to preserve local geometry. Then, the discrete Gromov-Wasserstein problem is,

$$GW(p^1, p^2) = \min_{\Gamma \in \Pi(p^1,p^2)} \sum_{i,j,k,l} \mathbf{L}_{ijkl} \Gamma_{ij} \Gamma_{kl}, \tag{2}$$

where Γ is a coupling matrix from the set:

$$\Pi(p^1, p^2) = \{\Gamma \in \mathbb{R}^{n_1 \times n_2}_+ : \Gamma \mathbb{1}_{n_2} = p_1,\ \Gamma^T \mathbb{1}_{n_1} = p_2\}. \tag{3}$$

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. Each entry Γ_{ij} describes how much of the mass of x_i should be mapped to y_j.

To make this problem more computationally tractable, we solve the entropically regularized version:

$$GW_\epsilon(p^1, p^2) = \min_{\Gamma \in \Pi(p^1,p^2)} \langle \mathbf{L}(D^1, D^2) \otimes \Gamma, \Gamma \rangle - \epsilon H(\Gamma). \tag{4}$$

where $\epsilon > 0$ and $H(\Gamma)$ is the Shannon entropy defined as $H(\Gamma) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \Gamma_{ij} \log \Gamma_{ij}$. Larger values of ϵ make the problem more convex but also lead to a denser coupling matrix, meaning there are more correspondences between samples. In SCOT, we use the cost function $L = L_2$.

2.1 Unbalanced Optimal Transport of SCOTv2

Our proposed solution to align datasets with different numbers of samples or proportions of cell-types is to use unbalanced Gromov-Wasserstein optimal transport, which adds divergence terms to allow for mass variations in the marginals [11,16]. We follow Séjourné *et al.* [16], and use the Kullback-Leibler divergence,

$$\mathrm{KL}(p||q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right), \tag{5}$$

to measure the difference between the marginals of the coupling Γ and the input marginals p^1 and p^2. Thus, we solve the unbalanced GW problem:

$$GW_{\epsilon,\rho}(p^1, p^2) = \min_{\Gamma \geq 0} \langle \mathbf{L}(D^1, D^2) \otimes \Gamma, \Gamma \rangle - \epsilon H(\Gamma) + \rho \mathrm{KL}(\Gamma \mathbb{1}_{n_2} || p^1) + \rho \mathrm{KL}(\Gamma^T \mathbb{1}_{n_1} || p^2), \tag{6}$$

where $\rho > 0$ is a hyperparameter that controls the marginal relaxation. When ρ is large, the marginals of Γ should be close to p^1 and p^2, and when ρ is small, the marginals of Γ may differ more, allowing each point to transport with more or less mass than it originally had. We detail the optimization procedure for unbalanced Gromov-Wasserstein optimal transport (UGWOT) in Algorithm 3.

2.2 Extending SCOTv2 for Multi-domain Alignment

We provide the details of SCOTv2 in Algorithm 1. To align more than two datasets ($M > 2$), we use one domain as an anchor to align the other domains. The anchor should be the domain with the clearest biological structures, for example, a dataset with the best-defined cell-type clusters. We propose selecting the anchor via the kNN graph computed. For every node x_i^m in the graph, we calculate the average of the k neighboring node values $\mathcal{N}_k(x_i^m)$. We measure the difference between this average and the true value of the node. This difference reflects how well the averaged neighborhood represents the given node. We then average these differences across the graph and select the domain with the lowest averaged difference as the anchor. Intuitively, we select the anchor whose kNN graph best reflects its dataset. Suppose X^1 is the anchor dataset. Then, for $m = 2, 3, \ldots, N$, we compute the coupling matrix Γ^m according to Eq. 4.

To have all of the datasets aligned in the same domain, we can either use barycentric projection to project each X^m for $m = 2, 3, \ldots, M$ onto X^1 or find a shared embedding space as described in Sect. 2.3. In the first iteration of SCOT, we used a barycentric projection to align and project one dataset onto the other. Due to the marginal relaxation, we now search for a non-negative $n_1 \times n_m$ dimensional matrix Γ instead of $\Gamma \in \Pi(p^1, p^m)$. Because of this change, the adjusted barycentric projection is:

$$x_i^m \mapsto \frac{\sum_{j=1}^{n_1} \Gamma_{ij}^m x_j^1}{\sum_{j=1}^{n_1} \Gamma_{ij}^m}. \tag{7}$$

2.3 Embedding with the Coupling Matrix

Other methods such as MMD-MA and UnionCom align datasets by embedding them into a common latent space of dimension $p \leq \min_{m=1,\ldots,M} d_m$. Here d_m represents the original dimension size of measurement (or domain) m. Embedding the datasets in a new space often leads to a better alignment as it introduces the additional benefits of dimension reduction, allowing more meaningful structures in the datasets such as cell-types to be more prevalent. Due to these benefits, we also enable the embedding option through a modification of the t-SNE method proposed by UnionCom [4]. For each domain m, we compute P^m, an $n_m \times n_m$ cell-to-cell transition matrix; each entry $P_{j|i}^m$ is the conditional probability that a data point x_i^m would pick x_j^m as its neighbor when chosen according a Gaussian distribution centered at x_i^m. Similarly, for the lower-dimensional embeddings, we compute a cell-to-cell probability matrix $Q^{m'}$ through a Student-t distribution. The full descriptions of P^m and $Q^{m'}$ are given in Appendix.

Then, to jointly embed all domains through the anchor domain X^1, the optimization problem is:

$$\min_{X^{1'},\ldots,X^{M'}} \sum_{m=1}^{M} \mathrm{KL}(P^m || Q^{m'}) + \beta \sum_{m=2}^{M} ||X^{1'} - X^{m'}(\Gamma^m)^T||_F^2, \tag{8}$$

Algorithm 1: Pseudocode for SCOTv2 Algorithm

```
Input: Datasets $X^1, \ldots, X^M$, number of graph neighbors $k$, entropic
  regularization coefficient $\epsilon$, mass relaxation coefficient $\rho$.
for $m = 1, \ldots, M$ do
    // Initialize marginal probabilities: $p^m \leftarrow \text{Uniform}(X^m)$;
    //Construct $G^m$, a $k-$NN graph based on pairwise correlations
    // Compute intra-domain distances $D^m$ with Dijsktra's algorithm.
    $c^m = \frac{1}{n_m} \sum_{i=1}^{n_m} \frac{1}{k} \sum_{x_j^m \in \mathcal{N}_k(x_i^m)} \text{corr}(x_j^m, x_i^m)$ //“neighborbood corr.
end
// Select an anchor domain $X^{m*}$: $m^* = \arg\max_{m=1,\ldots M} c^m$
for $m = 1, \ldots, M$ $(m \neq m^*)$ do
    $\Gamma^m \leftarrow UGWOT_{\epsilon,\rho}(p^m, p^{m*})$ // Compute pairwise couplings w/ $X^{m*}$:
    if Barycentric projection then
        $x_i^{m'} \leftarrow \frac{\sum_{j=1}^{n_1} \Gamma_{ij}^m x_j^{m*}}{\sum_{j=1}^{n_1} \Gamma_{ij}^m}$
    end
    else
        $X^{1'} \ldots X^{M'} \leftarrow$
          $\min_{X^{m'},\ldots,X^{M'}} \sum_{m=1}^{M} \text{KL}(P^m||Q^{m'}) + \beta \sum_{m \neq m*} ||X^{m^{*'}} - X^{m'}(\Gamma^m)^T||_F^2$
          // Find shared embedding
    end
end
Return: Aligned datasets, $X^{1'} \ldots X^{M'}$.
```

where $X^{m'}$ is the lower dimensional embedding of X^m, and Γ^m is the coupling matrix from solving Eq. 6 for $m = 2, \ldots, M$. These two terms seek to find an embedding that both preserves the local geometry in the original domain and aligns the domains according to the correspondence found by GW. The intuition behind the term $\text{KL}(P^m||Q^{m'})$ is very similar to that of GW; if two points have a high transition probability in the original space, then they should also have a high transition probability in the latent space. The term $||X^{1'} - X^{m'}(\Gamma^m)^T||_F^2$ measures how well aligned the new embeddings $X^{1'}$ and $X^{m'}$ are according to the prescribed coupling matrix Γ^m. Finally, $\beta > 0$ controls the trade-off between preserving the original geometry with the KL term and enforcing the alignment found with GW. We solve this optimization problem using gradient descent from UnionCom with a default latent space dimension size $p = 3$ [4].

2.4 Heuristic Process for Self-tuning Hyperparameters

SCOTv2 has three hyperparameters: (1) k for the number of neighbors to consider in nearest neighbor graphs, (2) the weight of the entropic regularization term, ϵ, and (3) the coefficient of the mass relaxation constraint, ρ. The barycentric projection of one domain onto another does not require any hyperparameters. However, jointly embedding the domains in a latent space requires selecting the dimension p.

Ideally, orthogonal correspondence information such as 1–1 correspondences and cell-type labels can guide hyperparameter tuning as validation. However, such information is hard to obtain in most cases. First, no validation data on cell-to-cell correspondences exists for non-co-assay datasets. Second, it is challenging to infer cell-types for certain sequencing domains such as 3D chromatin conformation. Lastly, the cell-type annotations may not always agree across single-cell domains.

We provide a heuristic to self-tune hyperparameters in the completely unsupervised setting. We first choose a k for the neighborhood graphs that yields a high average correlation value between the neighborhood predicted values and measured genomic values of the graph nodes. This step is the same as the one used to select the anchor domain for multi-omics alignment in Sect. 2.2. Next, we choose ϵ and ρ values that minimize the Gromov-Wasserstein distance between the aligned datasets. Algorithm 2 gives the details of this procedure.

Algorithm 2: Unsupervised hyperparameter search procedure

```
Input: Datasets $X^1, \dots, X^M$.
for $m = 1, \dots, M$ do
    $k^m = \underset{k \in \{10,20,\dots,150\}}{\text{argmax}} \frac{1}{n_m} \sum_{i=1}^{n_m} \frac{1}{k} \sum_{x_j^m \in \mathcal{N}_k(x_i^m)} \text{corr}(x_j^m, x_i^m)$ // Find $k_m$'s
    // Use $k^m$ to compute $D^m$
end
for $m = 2, \dots, M$ do
    $\epsilon^m, \rho^m = \arg\min_{\epsilon,\rho} GW_{\epsilon,\rho}(\mathbb{1}_{n_1}, \mathbb{1}_{n_m})$ // Use GW distance to pick $\rho$, $\epsilon$
end
Return: $k^m, \epsilon^m, \rho^m$.
```

3 Experimental Setup

3.1 Datasets

We evaluate SCOTv2 on single-cell datasets with disproportionate cell-types using two schemes. (1) We subsample different cell-types in co-assay datasets to simulate cell-type representation disparities between sequencing modalities. (2) We select real-world separately sequenced single-cell multi-omics datasets, which lack 1–1 cell correspondences and have different cell-type proportions across modalities due to the sampling procedure. Additionally, we present results on the original co-assay datasets with 1–1 cell correspondence to demonstrate the flexibility of SCOTv2 across balanced and unbalanced single-cell datasets.

Co-assay Single-Cell Datasets with 1–1 Cell Correspondence. We use three co-assay datasets to validate our model, sequenced by SNARE-seq, scGEM, and scNMT technologies. SNARE-seq is a two-modality sequencing technology that simultaneously captures the chromatin accessibility and transcriptional profiles of cells [6]. This dataset contains a total of 1047 cells from four cell lines: BJ (human fibroblast cells), H1 (human embryonic cells), K562 (human erythroleukemia cells), and GM12878 (human lymphoblastoid cells) (Gene Expression Omnibus access code: GSE126074). We follow the same data preprocessing steps outlined by Chen *et al.* [6]. The scGEM technology is a three-modality sequencing technology that profiles the genetic sequence, gene expression, and DNA methylation states in the same cell [7]. The dataset we use is derived from human somatic cell samples undergoing conversion to induced pluripotent stem cells (Sequence Read Archive accession code SRP077853) [7]. We access the preprocessed data provided by Welch *et al.* [17], which only contains the gene expression and DNA methylation modalities.[1] The dataset sequenced by scNMT-seq method [3] contains three modalities of genomic data: gene expression, DNA methylation, and chromatin accessibility, from mouse gastrulation samples, going through the Carnegie stages of vertebrate development (Gene Expression Omnibus access code: GSE109262). We access the preprocessed data through GitHub[2]. While the SNARE-seq and scGEM datasets contain the same number of cells across measurements, scNMT-seq modalities contain different cell-type proportions after preprocessing due to varying noise levels in measurements (Table 1).

Single-Cell Datasets with Simulated Cell-Type Imbalance. To test alignment performance sensitivity to different levels and types of cell-type proportion disparities across modalities, we generate simulation datasets by subsampling SNARE-seq and scGEM co-sequencing datasets in two ways. (1) We remove a cell-type from one modality. (2) We reduce the proportion of a cell-type in one modality by subsampling it at 50% and another cell-type in the other modality by subsampling it at 75%. We simulate this setting to test how the alignment methods will behave when multiple cell-types have disproportionate representation at different levels (for example, half or quarter percentage of cell-types missing) across modalities. For these cases, we uniformly pick at random which cell-type to subsample or remove. Specifically, for scGEM in simulation case (1), we remove "d16T+" cells in the DNA methylation domain while retaining the original gene expression domain, and remove the "d24T+" cells in the gene expression domain while retaining the original DNA methylation domain. For the SNARE-seq dataset, we remove "GM" cells in the gene expression domain and "K562" in the chromatin accessibility domain. In simulation case (2), we subsample the "d8" cluster of the scGEM dataset at 75% in the gene expression modality and the "d16T+" cluster at 50% in the DNA methylation modality.

[1] Preprocessed data for the scGEM dataset accessed here: https://github.com/jw156605/MATCHER.

[2] Dimensionality reduced data, used by Pamona and us, here: https://github.com/caokai1073/Pamona/tree/master/scNMT. Preprocessing scripts for the raw data provided by the authors here: https://github.com/PMBio/scNMT-seq/.

Table 1. Number of cells in (and percentages of) each cell-type across different modalities in the scNMT-seq co-assayed dataset after quality control procedures and the non-coassay datasets.

	Modality #1 (Gene Expression)	**Modality #2** (Chromatin Accessibility)	**Modality #3** (DNA Methylation or 3D chrom. conform.)
scNMT dataset	**(n = 579)** **E4.5:** 76 (12.73%) **E5.5:** 104 (17.42%) **Day6.5:** 146 (24.46%) **E7.5:** 271 (45.39%)	**(n = 647)** **E4.5:** 63 (9.73%) **E5.5:** 89 (13.76%) **E6.5:** 220 (34.00%) **E7.5:** 175 (42.50%)	**(n =725)** **E4.5:** 65 (8.96%) **E5.5:** 91 (12.55%) **E6.5:** 278 (38.34 %) **E7.5:** 291 (40.14%)
sciOmics dataset	**(n = 1,058)** **Day0:** 489 (46.22%) **Day3:** 127 (12.00%) **Day7:** 78 (7.37%) **Day11:** 145 (13.71%) **NPC:** 219 (20.70%)	**(n = 1,296)** **Day0:** 164 (12.65%) **Day3:** 702 (54.17%) **Day7:** 77 (5.94%) **Day11:** 175 (13.50%) **NPC:** 178 (13.73%)	**(n =2,154)** **Day0:** 987 (45.82 %) **Day3:** 435 (20.19 %) **Day7:** 243 (11.28 %) **Day11:** 164 (7.61 %) **NPC:** 325 (15.09 %)
MEC dataset	**(n = 26,273)** **Basal:** 11,138 (42.39 %) **L-Sec (Prog):** 7,683 (29.24 %) **L-HR:** 3,439 (13.09 %) **L-Sec (Mat):** 2,869 (10.92 %) **L-Sec (Prolif):** 758 (2.89 %) **Stroma:** 386 (1.47 %)	**(n = 21,262)** **Basal:** 13,353 (62.80 %) **L-Sec (Prog):** 3,343 (15.72 %) **L-HR:** 2,624 (12.34 %) **L-Sec (Mat):** 1,165 (5.48 %) **L-Sec (Prolif):** 7 (0.033 %) **Stroma:** 770 (3.62 %)	N/A

For SNARE-seq, we subsample the "H1" cluster at 75% and the "K562" cluster at 50% in the gene expression and chromatin accessibility domains, respectively.

Single-Cell Datasets Without 1–1 Correspondences. We also align non-co-assay datasets, containing separately sequenced single-cell -omic measurements. Bonora *et al.* generated the first dataset we use, "sciOmics" [1]. This dataset consists of sciRNA-seq, sciATAC-seq, and sciHiC measurements, capturing gene expression, chromatin accessibility, and 3D chromosomal conformation profiles of mouse embryonic stem cells undergoing differentiation. The measurements were taken at five stages: days 0, 3, 7, 11, and as fully differentiated neural progenitor cells (NPCs). The second non-co-assay dataset, "MEC," contains gene expression and chromatin accessibility measurements taken using the 10X Chromium scRNA-seq and scATAC-seq technologies on mouse mammary epithelial cells (MEC). Since each modality consists of separately sampled cell populations, these contain disparate cell-type proportions across modalities (Table 1).

3.2 Evaluation Metrics and Baseline Methods

Although most of the datasets lack 1–1 cell correspondences, we can evaluate alignment using cell-type labels through label transfer accuracy (LTA) as in [4,5,9]. This metric assesses the clustering of cell-types after alignment by training a kNN classifier on a training set (50% of the aligned data) and then evaluates its predictive accuracy on a test dataset (the other 50% of the aligned data). Higher values correspond to better alignments, indicating that cells that belong to the same cell-type are aligned close together after integration. We benchmark our method against the current unsupervised single-cell multi-omic alignment methods, Pamona [5], UnionCom [4], MMD-MA [14], bindSC [10], Seuratv4 [15], and the previous version of SCOT, which performs alignment without the KL term [9]. Pamona [5], as previously discussed, uses partial Gromov-Wasserstein (GW) optimal transport to align single-cell datasets. UnionCom [4] performs unsupervised topological alignment through a two-step procedure that first finds a correspondence between the domains, considering both global and local geometries with a hyperparameter to control the trade-off between them, and then embeds them in a new shared space. MMD-MA [14] uses the maximum mean discrepancy (MMD) measure to align and embed two datasets in a new space. BindSC [10] requires the users to bring input datasets to the gene expression feature space by constructing a gene activity score matrix for the epigenomic domains, then finds a correspondence matrix between samples through bi-order canonical correspondence analysis (bi-CCA), and jointly embeds the domains into a new space. Finally, Seuratv4 [15] also requires gene activity score matrices for epigenomic domains and then identifies correspondence anchors via CCA. Based on these anchors, it imputes one genomic domain based from the other domain and co-embeds them into a shared space using UMAP.

Since bindSC and Seurat v4 require the creation of gene activity score matrices for epigenomic datasets, they might be more difficult to use with certain sequencing domains. For instance, gene activity scoring is challenging for 3D chromosomal conformation. Of all the selected baselines, only Pamona and UnionCom can align more than two domains, so we only use them as baselines for experiments with multiple domains ($M > 2$). For each benchmark, we define a hyperparameter grid of similar granularity and perform extensive tuning (see Appendix). We report the alignment results with the best performing hyperparameter combinations in Sect. 4.1.

4 Results

4.1 SCOTv2 Gives High-Quality Alignments Consistently Across All Single-Datasets

We first present the alignment results for real-world co-assay datasets with simulated cell-type imbalance. The results we present are obtained by the best performing hyperparameter combinations for all methods compared in this study. Figure 2(A) visualizes the barycentric projection alignments performed by SCOTv2 plotted as 2D PCA for SNARE-seq and scGEM datasets, respectively. We use barycentric projection for visualizations because we set this to be the default projection method of our method since it does not require additional hyperparameters. Here, we integrate datasets under three different settings described in the previous section: (1) Balanced datasets (or “full datasets” with no subsampling), (2) Missing cell-type in the epigenomic domains, and (3) Subsampled cells

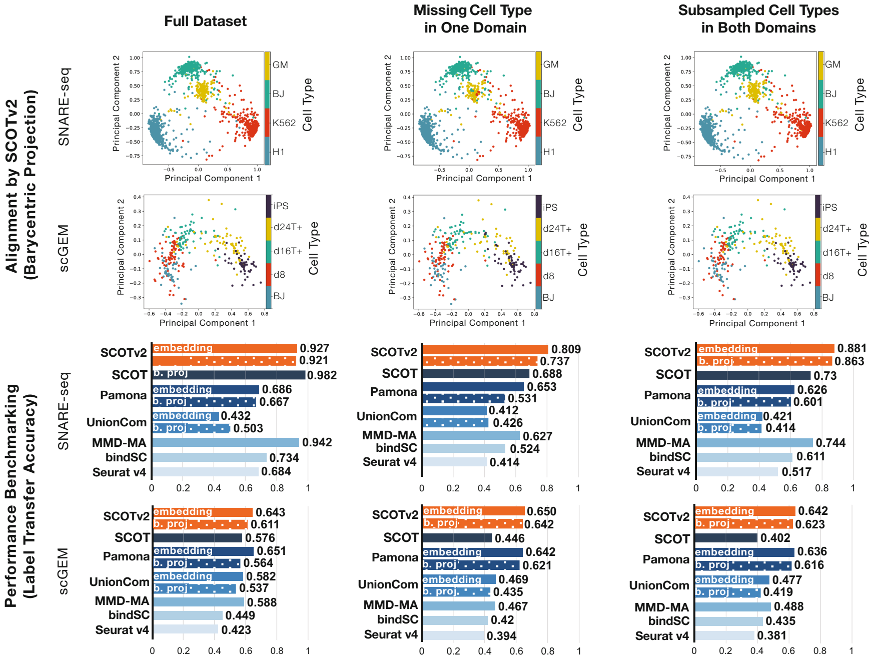

Fig. 2. Alignment results for simulations and balanced co-assay datasets. **A** visualizes the barycentric projection alignment on SNARE-seq and scGEM for the full co-assay datasets, simulations with a missing cell-type in the epigenomic domain, and subsampled cell-types in both domains. **B** compares the alignment performance of SCOTv2 to the benchmarks through LTA. For SCOTvs, Pamona, and UnionCom, we report results on both embedding into a shared space (solid bars) and the barycentric projection (dotted bars).

in both domains (one cell-type at 50% in the epigenomic domains and another cell-type at 75% in the gene expression domains). We include alignment results on the full datasets with 1–1 sample correspondences to ensure that SCOTv2 performs well for balanced cases as well.

Qualitatively, we see that SCOTv2 preserves the cell-type annotations after alignment for all three settings. In Fig. 2(B), we report the quantitative performance of SCOTv2 and all the other state-of-the-art baselines using the Label Transfer Accuracy (LTA) scores. MMD-MA, UnionCom, Seurat, and bindSC fail to reliably align datasets with disproportionate cell-type representation across modalities. While Pamona tends to yield high-quality alignments for cases with cell-type disproportion, it fails to perform well on the SNARE-seq balanced dataset as well as its subsampling simulation.

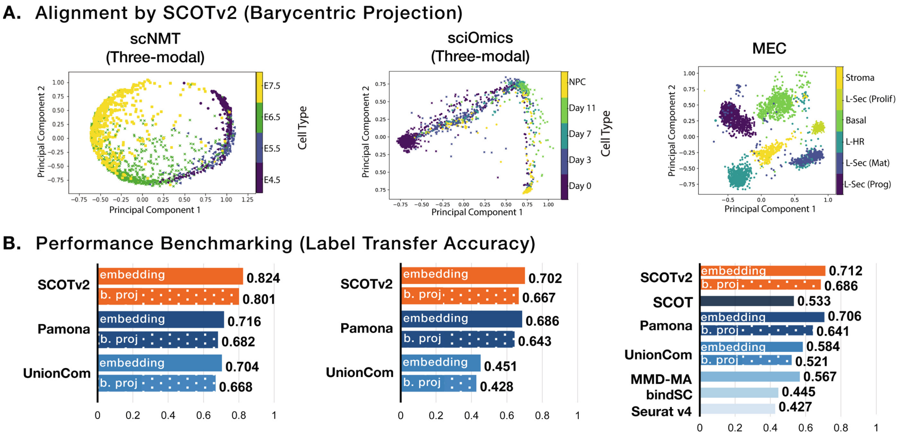

Fig. 3. Alignment results for multi-modal ($M > 2$) and separately sequenced datasets. A visualizes the alignment of scNMT-seq, sciOmics, and MEC. All datasets have unequal sample sizes and cell-type proportions across domains. **B** benchmarks alignment performance through LTA. As in Fig. 2, we report results both by embedding (solid bars) and barycentric projection (dotted bars) for the methods that allow for both. For scNMT-seq and sciOmics, which are three-modal datasets, we only demonstrate results for SCOTv2, Pamona, and UnionCom, which can handle more than two modalities.

Among all methods tested, SCOTv2 consistently gives more high-quality alignments across different scenarios of cell-type representation. It also demonstrates a ∼22% average increase in LTA over the previous version of the algorithm (SCOT) when comparing the barycentric projection results and ∼27% for the embedding results. UnionCom, Pamona, and SCOTv2 allow us to perform both barycentric projections and embed the single-cell domains in a lower-dimensional space. Overall, we observe that embedding yields higher LTA values than barycentric projection. Since the barycentric projection projects one domain onto another, the separation of the domain being projected onto (or anchor domain) limits the clustering separation after alignment. In contrast, the embedding utilizes t-SNE to enhance cell-type separation, allowing for better-separated clusters after alignment.

Next, we report the alignment performance of SCOTv2 on single-cell datasets with disparities in cell-type representation due to sampling during experiments. We include scNMT, a co-assay with varying levels of cells across domains due to quality control procedures, along with sciOmics and MEC for this experiment. Note that scNMT and sciOmics have three different modalities, and hence, we can only report the baselines for methods that can align datasets with $M > 2$. Figure 3(A) presents the qualitative alignment results for SCOTv2 with PCA. SCOTv2 performs well on all three datasets, including the ones with three modalities. The LTA scores in Fig. 3(B) demonstrate that SCOTv2 consistently yields the best alignments on the three real-world datasets. These results highlight its ability to reliably integrate separately sampled with disproportionate cell-type representation and multiple ($M > 2$) modalities simultaneously.

4.2 Hyperparameter Self-tuning Aligns Well Without Depending on Orthogonal Correspondence Information

The benchmarking results above present the alignment performance of each algorithm at its best hyperparameter setting; however, users may not have 1–1 correspondences to validate alignments, for the purpose of hyperparameter selection, in real-world applications. While users may have access to cell-type labels, inferring cell-types is highly difficult in specific modalities of single-cell sequencing, such as 3D chromatin conformation. Additionally, different sequencing modalities might disagree on cell-type clustering (as is often the case with scRNA-seq and scATAC-seq datasets). In these situations, users might not have sufficient validation data for tuning hyperparameters.

Table 2. Alignment performance benchmarking in the fully unsupervised setting. We run SCOTv2 and SCOT using their heuristics to approximately self-tune hyperparameters. We use default parameters for other methods due to a lack of similar procedures for unsupervised self-tuning.

	SNARE (full dataset)	**SNARE** (missing cell-type)	**SNARE** (subsam. dataset)	**scGEM** (full dataset)	**scGEM** (missing cell-type)	**scGEM** (subsam. dataset)	**scNMT**	**sciOmics**	**MEC**
SCOTv2	0.826	**0.653**	**0.751**	**0.509**	**0.521**	**0.415**	**0.727**	**0.537**	**0.584**
SCOT	**0.852**	0.572	0.588	0.423	0.323	0.314	N/A	N/A	0.466
Pamona	0.554	0.423	0.419	0.385	0.414	0.308	0.588	0.329	0.417
MMD-MA	0.523	0.407	0.431	0.360	0.296	0.287	N/A	N/A	0.233
UnionCom	0.411	0.406	0.422	0.332	0.315	0.276	0.474	0.306	0.349
bindSC	0.713	0.584	0.475	0.387	0.254	0.262	N/A	N/A	0.412
Seurat	0.428	0.517	0.503	0.408	0.377	0.329	N/A	N/A	0.387

We design a heuristic process (described in Sect. 2.4), as done previously for SCOT, that allows SCOTv2 to select hyperparameters in a completely unsupervised manner. Other alignment methods do not provide an unsupervised hyperparameter tuning procedure. Therefore, without validation data, a user would have to use the default parameters. In Table 2, we compare alignment performance for our heuristic against the default parameters of other methods. While our heuristic does not always yield the optimal hyperparameter combination, it does give more favorable results over the default settings of the other methods. Thus, we recommend using it in cases that lack orthogonal information for hyperparameter tuning.

4.3 SCOTv2 Scales Well with Increasing Number of Samples

We compare the runtime of SCOTv2 with the top performing methods: Pamona, MMD-MA, UnionCom, and the previous version of SCOT by subsampling various numbers of cells from the MEC dataset. MMD-MA, UnionCom, and SCOTv2 have GPU versions, while Pamona and SCOT only have CPU versions. We run MMD-MA and UnionCom on a single NVIDIA GTX 1080ti GPU with VRAM of 11 GB and Pamona and SCOT on Intel Xeon e5-2670 CPU with 16 GB memory. We also run SCOTv2 on the same CPU to give comparable results to Pamona's runtimes. Figure 4 depicts that SCOT, MMD-MA, Pamona, and SCOTv2 show similar computational scaling.

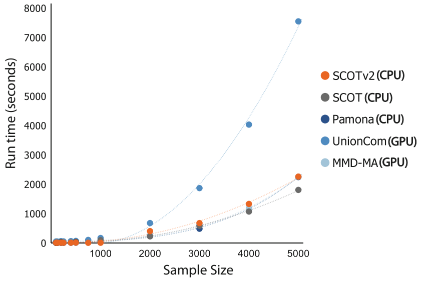

Fig. 4. Runtimes for SCOTv2, SCOT, Pamona, UnionCom, and MMD-MA as the number of samples increases.

5 Discussion

We present SCOTv2, an improved unsupervised alignment algorithm for multi-omics single-cell alignment. It extends the alignment capabilities of SCOT to datasets with cell-type representation disproportions across different sequencing measurements. It also performs alignment for single-cell datasets with more than two measurements ($M > 2$). Experiments on real-world subsampled co-assay datasets and separately sampled and sequenced single-cell datasets demonstrate that SCOTv2 reliably yields high-quality alignments for a wide range of cell-type disproportions without compromising its computational scalability. Furthermore, SCOTv2's flexible marginal constraints enable it to consistently give good alignments results for both balanced and unbalanced single-cell datasets. In addition to effectively handling cell-type imbalances and multi-omics alignment, SCOTv2 can self-tune its hyperparameters making it applicable in complete unsupervised settings. Therefore, SCOTv2 offers a convenient way to align multiple single-cell measurements without requiring any orthogonal correspondence information.

In this second iteration of SCOT, we have utilized the coupling matrix in a new way to find a latent embedding space. While this dimension reduction improves cell-type separation, using the coupling matrix directly may offer even more insights into interactions between the aligned domains. Future work will consider how to use the probabilities in the coupling matrix directly for downstream analysis like improved clustering and pseudo-time inference. Though SCOTv2 has runtimes that scale with other methods, it requires $O(n^2)$ memory storage for the distance matrices, which may be an issue for especially large datasets. One way to address this limitation would be to develop a procedure to align a representative subset of each domain that can be extended to the entire dataset. Therefore, we will explore this direction to further improve the scalability of SCOTv2.

Appendix

Embedding Method Details

The full details of t-SNE can be found in [13]. For each domain m, we compute P^m, an $n_m \times n_m$ cell-to-cell transition matrix; each entry $P^m_{j|i}$ is the conditional probability that a data point x^m_i would pick x^m_j as its neighbor when chosen according a Gaussian distribution centered at x^m_i:

$$P^m_{j|i} = \frac{\exp(-||x^m_i - x^m_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||x^m_i - x^m_k||^2/2\sigma_i^2)}. \tag{9}$$

The bandwidth σ_i is chosen according to the density of the data points through a binary search for the value of σ_i that achieves the user-supplied perplexity value. P^m is computed by averaging $P^m_{i|j}$ and $P^m_{j|i}$ to give more weight to outlier points:

$$P^m_{ij} = \frac{P^m_{i|j} + P^m_{j|i}}{2n_m} \tag{10}$$

Then, to jointly embed all domains through the anchor domain X^1, the optimization problem is:

$$\min_{X^{1'},\ldots,X^{M'}} \sum_{m=1}^{M} \mathrm{KL}(P^m||Q^{m'}) + \beta \sum_{m=2}^{M} ||X^{1'} - X^{m'}(\Gamma^m)^T||_F^2, \tag{11}$$

where $X^{m'}$ is the lower dimensional embedding of X^m, P^m is defined as in Eq. 9, and Γ^m is the coupling matrix from solving Eq. 6 for $m = 1, 2, \ldots, M$, $X^{m'}$. The probability matrix Q^m is computed through a Student-t distribution with one degree of freedom:

$$Q^{m'}_{ij} = \frac{(1 + ||x^{m'}_i - x^{m'}_j||)^{-1}}{\sum_{k\neq l} 1 + (||x^{m'}_k - x^{m'}_l||)^{-1}}. \tag{12}$$

The intuition behind the cost $\mathrm{KL}(P^m||Q^{m'})$ is very similar to that of GW; if two points have a high transition probability in the original space, then they should also have a high transition probability in the latent space.

Hyperparameter Tuning Procedure Details

For each alignment method, we define a grid of hyperparameters and choose the best performing combination for each experiment. If methods share similar hyperparameters in their formulation, we keep the range defined for these consistent across all algorithms. We refer to the publication and the code repository for each method to choose a hyperparameter ranges whenever possible.

For Pamona, we search the number of neighbors in the cell neighborhood graphs, $k \in \{20, 30, \ldots, 150\}$, the entropic regularization coefficient, $\epsilon \in \{5e{-}4, 3e{-}4, 1e{-}4, 7e{-}3, 5e{-}3, \ldots, 1e{-}2\}$, geometry preservation trade-off coefficient, $\lambda \in \{0.1, 0.5, 1, 5, 10\}$, and lastly, embedding dimensionality, $p \in \{3, 4, 5, 10, 30, 32\}$, the output dimension for embedding. For UnionCom, we search the trade-off parameter $\beta \in \{0.1, 1, 5, 10, 15, 20\}$, the regularization coefficient $\rho \in \{0, 0.1, 1, 5, 10, 15, 20\}$, the maximum neighborhood size permitted in the neighborhood graphs, $k_{max} \in \{40, 100, 150\}$, and embedding dimensionality $p \in \{3, 4, 5, 10, 30, 32\}$. For MMD-MA:, we tune the weights λ_1 and $\lambda_2 \in \{1e{-}2, 5e{-}3, 1e{-}3, 5e{-}4, \ldots, 1e{-}9\}$, and the embedding dimensionality, $p \in \{3, 4, 5, 10, 30, 32\}$. For bindSC, we choose the coefficient that assigns weight to the initial gene activity matrix $\alpha \in \{0, 0.1, 0.2, \ldots 0.9\}$, the coefficient that assigns weight factor to multi-objective function $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$, and the number of canonical vectors for the embdedding space $K \in \{3, 4, 5, 10, 30, 32\}$. Lastly, for Seuratv4, we tune the number of neighbors to consider when finding anchors, $k \in \{5, 10, 15, 20\}$, co-embedding dimensionality, $p \in \{3, 4, 5, 10, 30, 32\}$ and the choice of the reference and anchor domains when finding anchors.

Algorithm 3: Pseudocode for Unbalanced GW Optimal Transport (UGWOT)

```
Input: Marginal probabilities $p^1$ and $p^2$, intra-domain distance matrices $D^1$
  and $D^2$, relaxation coefficient $\rho$, regularization coefficient $\epsilon$
Initialize the coupling matrix: $\Gamma = \pi = p^1 \otimes p^2$
while $\Gamma$ not converged do
    $\Gamma_{(mass)} \leftarrow \sum_{i,j} \Gamma_{i,j}$  $\tilde{\epsilon} \leftarrow \Gamma_{(mass)}\epsilon$,  $\tilde{\rho} \leftarrow \Gamma_{(mass)}\rho$
    // Compute cost $C$:
    $\Gamma^1 \leftarrow \Gamma \mathbb{1}_{n_2}$, $\Gamma^2 \leftarrow \Gamma^T \mathbb{1}_{n_1}$
    $A \leftarrow (D^1)^{\circ 2}\Gamma^1$, $B \leftarrow (D^2)^{\circ 2}\Gamma^2$
    $D \leftarrow D^1 \Gamma D^2$
    $E \leftarrow \epsilon \sum_{ij} \log\left(\frac{\Gamma_{i,j}}{p_i^1 p_j^2}\right)\Gamma_{i,j} + \rho\left(\sum_i \log\left(\frac{\Gamma_i^1}{p_i^1}\right)\Gamma_i^1 + \sum_j \log\left(\frac{\Gamma_j^2}{p_j^2}\right)\Gamma_j^2\right)$
    $C \leftarrow A + B - 2D + E$
    while $(u, v)$ not converged do
        $u \leftarrow -\frac{\tilde{\epsilon\rho}}{\tilde{\epsilon}+\tilde{\rho}} \log\left[\sum_{i,j} \exp(v_j - C_{ij})/\tilde{\epsilon} + \log p^2\right]$
        $v \leftarrow -\frac{\tilde{\epsilon\rho}}{\tilde{\epsilon}+\tilde{\rho}} \log\left[\sum_{i,j} \exp(u_i - C_{ij})/\tilde{\epsilon} + \log p^1\right]$
    // Update: $\pi_{ij} \leftarrow \exp\left[u_i + v_j - C_{ij}\right] p_i^1 p_j^2$
    // Rescale: $\pi \leftarrow \sqrt{\Gamma_{(mass)}/\pi_{(mass)}}\pi$ and set $\Gamma \leftarrow \pi$
Return: $\Gamma$
```

References

1. Bonora, G., et al.: Single-cell landscape of nuclear configuration and gene expression during stem cell differentiation and x inactivation. Genome Biol. **22**(1), 279 (2021). https://doi.org/10.1186/s13059-021-02432-w

2. Alvarez-Melis, D., Jaakkola, T.S.: Gromov-wasserstein alignment of word embedding spaces. arXiv preprint arXiv:1809.00013 (2018)
3. Argelaguet, R., Clark, S.J., Mohammed, H., Stapel, L.C., Krueger, C., Kapourani, C.A., et al.: Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature **576**(7787), 487–491 (2019). https://doi.org/10.1038/s41586-019-1825-8
4. Cao, K., Bai, X., Hong, Y., Wan, L.: Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics **36**(Suppl._1), i48–i56 (2020)
5. Cao, K., Hong, Y., Wan, L.: Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. Bioinformatics **38**(1), 211–219 (2021). https://doi.org/10.1093/bioinformatics/btab594
6. Chen, S., Lake, B.B., Zhang, K.: High-throughput sequencing of transcriptome and chromatin accessibility in the same cell. Nat. Biotechnol. **37**(12), 1452–1457 (2019)
7. Cheow, L.F., Courtois, E.T., Tan, Y., Viswanathan, R., Xing, Q., Tan, R.Z., et al.: Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. Nat. Methods **13**(10), 833–836 (2016)
8. Clark, S.J., Argelaguet, R., Kapourani, C.A., Stubbs, T.M., Lee, H.J., et al.: scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat. Commun. **9**(1), 1–9 (2018)
9. Demetci, P., Santorella, R., Sandstede, B., Noble, W.S., Singh, R.: Gromov-wasserstein optimal transport to align single-cell multi-omics data. BioRxiv (2020)
10. Dou, J., Liang, S., Mohanty, V., Cheng, X., Kim, S., Choi, J., et al.: Unbiased integration of single cell multi-omics data. bioRxiv (2020). https://doi.org/10.1101/2020.12.11.422014. https://www.biorxiv.org/content/early/2020/12/11/2020.12.11.422014
11. Liero, M., Mielke, A., Savaré, G.: Optimal entropy-transport problems and a new hellinger-kantorovich distance between positive measures. Invent. Math. **211**(3), 969–1117 (2018)
12. Liu, J., Huang, Y., Singh, R., Vert, J.P., Noble, W.S.: Jointly embedding multiple single-cell omics measurements. BioRxiv, p. 644310 (2019)
13. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
14. Singh, R., Demetci, P., Bonora, G., Ramani, V., Lee, C., Fang, H., et al.: Unsupervised manifold alignment for single-cell multi-omics data. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 1–10 (2020)
15. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., III, W.M.M., et al.: Comprehensive integration of single-cell data. Cell **77**(7), 1888–1902 (2019)
16. Séjourné, T., Vialard, F.X., Peyré, G.: The unbalanced gromov wasserstein distance: Conic formulation and relaxation. arXiv (2021)
17. Welch, J.D., Hartemink, A.J., Prins, J.F.: Matcher: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol. **18**(1), 138 (2017)