# Scalable unsupervised alignment of general metric and non-metric structures

**Anonymous Authors**[1]

## Abstract

Aligning data from different domains is a fundamental problem in machine learning with broad applications across very different areas, most notably aligning experimental readouts in single-cell multiomics. Mathematically, this problem can be formulated as the minimization of disagreement of pair-wise quantities such as distances and is related to the Gromov-Hausdorff and Gromov-Wasserstein distances. Computationally, it is a quadratic assignment problem (QAP) that is known to be NP-hard. Prior works attempted to solve the QAP directly with entropic or low-rank regularization on the permutation, which is computationally tractable only for modestly-sized inputs, and encode only limited inductive bias related to the domains being aligned. We consider the alignment of metric structures formulated as a discrete Gromov-Wasserstein problem and instead of solving the QAP directly, we propose to *learn* a related well-scalable linear assignment problem (LAP) whose solution is also a minimizer of the QAP. We also show a flexible extension of the proposed framework to general non-metric dissimilarities through differentiable ranks. We extensively evaluate our approach on synthetic and real datasets from single-cell multiomics and neural latent spaces, achieving state-of-the-art performance while being conceptually and computationally simple.

## 1. Introduction

Unsupervised alignment of data that are related, yet not directly comparable, is a fundamental problem in machine learning. This problem is ubiquitous across a multitude of tasks such as *non-rigid shape correspondence* in computer vision (Bronstein et al., 2006; Halimi et al., 2019), *unlabeled sensing* in signal processing (Unnikrishnan et al., 2018; Emiya et al., 2014), and *latent space communication* in representation learning (Moschella et al., 2022; Maiorca et al., 2024). From an application perspective, we are particularly interested in single-cell biology. In fact, the development of single-cell sequencing technologies has led to the profiling of different molecular aspects within the cell at an unparalleled resolution. Profiling techniques have been developed to assay gene expression (Kukurba & Montgomery, 2015), chromatin accessibility and 3D conformation (Grandi et al., 2022; Deshpande et al., 2022), DNA methylation (Gouil & Keniry, 2019), and histone modifications (O'Geen et al., 2011). The analysis of genome (Navin et al., 2011; Zong et al., 2012), transcriptome (Tang et al., 2010; Guo et al., 2013), and DNA methylation (Smallwood et al., 2014; Guo et al., 2013) profiles has led to enhanced understanding of the heterogeneity across cell populations. The development of high-throughput sequencing (Macosko et al., 2015; Klein et al., 2015; Zheng et al., 2017), and spatial transcriptomics (Rao et al., 2021) technologies further enabled molecular profiling of cells at a high temporal and spatial resolution. One of the central problems within single-cell multiomics is integrating data from different molecular profiles, which is crucial in understanding joint regulatory mechanisms within the cell. Most single-cell sequencing techniques are invasive; thus, carrying out multiple assays on the same cell is rarely possible. While experimental co-assaying techniques are an active area of research (Cheow et al., 2016; Lee et al., 2020), they currently lack the high throughput of their single-assay counterparts. Computationally integrating data from different experimental modalities is, therefore, an important problem, and is the focus of the current paper.

Using the formalism of Gromov-Hausdorff (GH) (Gromov et al., 1999) and Gromov-Wasserstein (GW) (Mémoli, 2011) distances, unsupervised alignment can be formulated as the minimization of disagreement in pair-wise distances. Given two point clouds, both the GH and GW problems aim to find an *assignment* that is invariant to distance-preserving transformations (isometries) of the point clouds. GH seeks an *exact* point-wise assignment and can be shown to be a quadratic assignment problem (QAP) that is known to be an NP-hard (Burkard et al., 1998) and, thus, computationally intractable. GW relaxes the GH problem to find a *soft assignment* and it is more tractable in practice. The most common approach to solving QAP relaxations like GW is by solving

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

a sequence of *linear assignment problems* (LAPs) (Gold & Rangarajan, 1995) or *entropy-regularized optimal transport* ($\epsilon$-OT) (Cuturi, 2013) problems. This approach, coupled with the idea of kernel matching and, specifically, simulated annealing of kernel matrices, has been demonstrated very successful in shape analysis, practically rendering non-rigid shape correspondence a solved problem (Vestner et al., 2017; Melzi et al., 2019). For more general, less structured and higher-dimensional data, recent works have aimed to accelerate the GW solver by (i) reducing the problem size by applying *recursive clustering* (Blumberg et al., 2020) or through the *quantization* of the input dissimilarities (Chowdhury et al., 2021); and (ii) imposing *low-rank constraints* on the pairwise distance matrices and the assignment matrix within the internal $\epsilon$-OT solver (Scetbon et al., 2022).

Specifically on the problem of unsupervised alignment of single-cell multiomic data, GW solvers have already shown promise. Nitzan *et al.* (Nitzan et al., 2019) showed that they could map spatial coordinates in 2D tissues that were obtained with fluorescence in situ hybridization (FISH) to gene expression data. More recently, Demetci *et al.* (Demetci et al., 2022) demonstrated that GW solvers outperform other unsupervised alignment approaches on real data generated by the SNAREseq assay (Chen et al., 2019a), which links chromatin expression to gene expression. Unfortunately, existing solvers have several limitations, including poor scalability to very large ($N \sim 10^4$) datasets, convergence to local minima, and lack of inductivity in the sense that the solver has to be run anew once new data are obtained. This paper proposes remedies to these shortcomings.

**Contributions.** In this work, we introduce a new *framework* for solving GW-like problems. The core idea of our approach is to *learn the cost* of an OT problem (essentially, a LAP) whose solution is also the minimizer of the GW problem (essentially, a QAP). Instead of *explicitly* learning the cost matrix for the given set of samples, we propose to *implicitly* parametrize the cost as a ground-cost measured on neural network embeddings of the points that are being aligned. In order to learn the the neural networks parametrizing the cost, we render the entropy-regularized OT problem as an implicitly differentiable layer using the methodology proposed in (Eisenberger et al., 2022), and demand that the soft assignment produced by $\epsilon$-OT minimizes the GW objective.

This framework offers unique advantages over the standard approach of solving GW as a sequence of LAPs. Firstly, our method is *inductive*. Since we implicitly parametrize the cost with neural networks, when we encounter new pairs of unaligned samples at inference, we simply need to solve an $\epsilon$-OT problem on the embeddings produced by our trained network. This is in contrast to all the other GW solvers that, to the best of our knowledge, are *transductive* and would need to solve the GW problem anew by augmenting the test points. Secondly, our framework is *scalable* requiring to only solve a point-wise $\epsilon$-OT problem at inference. Compared to GW, $\epsilon$-OT is far simpler, and efficient solvers can be employed to solve this problem at scale (Cuturi, 2013; Genevay et al., 2016). Thirdly, our framework is gradient descent-based and is, therefore, *more expressive and general*, as it is straightforward to induce additional domain knowledge into the problem or impose additional regularization on the minimizer. Furthermore, it is straightforward to extend our method to the semi-supervised setting where a partial correspondence is known, and to the *fused GW* (Vayer et al., 2020) setting where a shared attribute is provided in both domains.

Leveraging the advantages of the proposed framework, we propose several novel extensions. Firstly, we demonstrate, for the first time, solving *arbitrary non-metric* assignment problems. To this end, we propose a new objective that matches distance ranks instead of the absolute distances themselves and demonstrate that it is more effective in single-cell multiomic alignment. The standard GW solvers rely on the linearization of QAPs, and it is unclear how they can be extended to handle more complex objectives such as those involving ranking. Secondly, inspired by techniques in geometric matrix completion (Kalofolias et al., 2014; Boyarski et al., 2022), by interpreting the learned cost as a signal on the product manifold of both domains, we impose a regularization that demands that the cost is smooth on its domain. This is intuitive because similar samples in one domain incur a similar cost with respect to the samples from the other domain, and vice-versa. Thirdly, in order to robustify training through $\epsilon$-OT solvers, we propose a simulated-annealing–based approach allowing tuning of the regularization coefficient in the Sinkhorn algorithm during the training process.

We evaluate our method both in inductive and transductive settings, on synthetic and real data. We demonstrated in an inductive setting our solver generalizes and scales to large sample sizes. We demonstrate that it outperforms the entropic GW solver on the SNARE-seq data from (Chen et al., 2019b) and on human bonemarrow scATAC vs. scRNA mapping task proposed in Luecken et al. (2021).

## 2. Background and closely related works in Optimal Transport

Given two sets of points $\{\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_N\} \in \mathcal{X}$ and $\{\mathbf{y}_1, \dots \mathbf{y}_N\} \in \mathcal{Y}$, the goal of unpaired alignment is to find a *point-wise correspondence* $\mathbf{P} \in \mathcal{P}^N$ such that each point in $\mathcal{X}$ is mapped to a point in $\mathcal{Y}$, and vice-versa, where $\mathcal{P}^N$ is the space of permutations. The central theme of metric-based alignment approaches (GH and GW) is to compare the sets of points as *metric spaces*. $\mathcal{X}$ and $\mathcal{Y}$ are considered similar if the metrics between corresponding points, as defined by $\mathbf{P}$, are similar as measured in $\mathcal{X}$ and in $\mathcal{Y}$. Denote by $d_\mathcal{X}$ and $d_\mathcal{Y}$ the metrics associated to $\mathcal{X}$ and $\mathcal{Y}$, and by $\mathbf{D}_\mathcal{X} \in \mathbb{R}^{N\times N}$ and $\mathbf{D}_\mathcal{Y} \in \mathbb{R}^{N\times N}$ the corresponding pairwise distance matrices computed over the points from $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let further $\mu$ and $\nu$ be the associated discrete probability measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Depending on what the spaces represent, these can be uniform measures or incorporate discrete volume elements.

**Gromov-Hausdorff distance.** The *distortion* induced by a correspondence $\mathbf{P}$ between $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ is defined as $\text{dis}(\mathbf{P}) = \|\mathbf{D}_\mathcal{X} - \mathbf{P}\mathbf{D}_\mathcal{Y}\mathbf{P}^\top\|_\infty$. This measures how well the distances between the matched points are preserved. The Gromov-Hausdorff (GH) distance (Gromov et al., 1999) is then defined as

$$d_{\text{GH}}((\mathcal{X}, d_\mathcal{X}), (\mathcal{Y}, d_\mathcal{Y})) = \min_{\mathbf{P}\in\mathcal{P}^N} \text{dis}(\mathbf{P}). \tag{1}$$

The optimization problem in Eq. 1 results in an integer linear program and is an NP-hard problem (Burkard et al., 1998). Therefore, it is computationally intractable.

**Gromov-Wasserstein distance.** Mémoli (2011) proposed relaxing the constraint on $\mathbf{P}$ from an exact assignment defined over $\mathcal{P}^N$ to a *probabilistic* (soft) assignment, i.e., to the space of couplings with marginals $\mu$ and $\nu$ denoted by $U(\mu,\nu) := \{\mathbf{\Pi} \in \mathbb{R}_+^{N\times N} \,|\, \mathbf{\Pi}\mathbf{1}_N = \boldsymbol{\mu}, \mathbf{\Pi}^\top\mathbf{1}_N = \boldsymbol{\nu}\}$. Using this relaxation, the squared Gromov-Wasserstein distance between discrete metric spaces is defined as

$$\begin{aligned} d^2_{\text{GW}} &= \min_{\mathbf{\Pi}\in U(\mu,\nu)} \sum_{i,j,i',j'} \left(d_\mathcal{X}(\mathbf{x}_i, \mathbf{x}_{i'}) - d_\mathcal{Y}(\mathbf{y}_i, \mathbf{y}_{i'})\right)^2 \pi_{ij}\pi_{i'j'} \\ &= \min_{\mathbf{\Pi}\in U(\mu,\nu)} \|\mathbf{D}_\mathcal{X} - \mathbf{\Pi}\mathbf{D}_\mathcal{Y}\mathbf{\Pi}^\top\|^2_{\text{F}}. \end{aligned} \tag{2}$$

To avoid confusion, we reserve the notation $\mathbf{P}$ to the true permutation matrix, while denoting the "soft" assignment by $\mathbf{\Pi}$. Notice that the definition of the GW distance results in a quadratic function in $\mathbf{\Pi}$; thus, it is referred to as the *quadratic assignment problem.* Alternative relaxations to the GH problem exist based on semi-definite programming (SDP) (Villar et al., 2016), but due to the poor scalability of SDP problems, they do not apply to the scales discussed in this paper.

**Optimal transport.** Aligning data that lie *within the same space* is a *linear* optimal transport (OT) problem (Peyré et al., 2019). Given two sets of points $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{x}'_j\}_{i=1}^N$ in the same space $\mathcal{X}$ with two discrete measures $\mu$ and $\nu$, respectively, the OT problem is defined as the minimization of $\sum_{i,j}\pi_{i,j}c(\mathbf{x}_i, \mathbf{x}'_j)$, such that $\mathbf{\Pi}$ satisfies marginal constraints $U(\mu,\nu)$ and $c$ defines transport cost (often, $c(\boldsymbol{x}, \boldsymbol{x}') = d_\mathcal{X}(\boldsymbol{x}, \boldsymbol{x}')$). Note that the objective is *linear* in $\mathbf{\Pi}$, in contrast to GW (Eq. 2), where it is quadratic. Entropy-regularized OT ($\epsilon$-OT) introduces an entropic regularization term, $\epsilon\langle\mathbf{\Pi}, \log\mathbf{\Pi}\rangle$, that can be very efficiently solved using the Sinkhorn algorithm (Cuturi, 2013) (see Appendix for details). More recently, Eisenberger et al. (2022) introduced *differentiable Sinkhorn layers* that uses implicit-differentiation (Amos & Kolter, 2017) to cast the Sinkhorn algorithm as a differentiable block within larger auto-differentiation pipelines. They calculate the Jacobian of the resulting assignment matrix with respect to both the primal and dual variables of the entropic-regularized OT problem. While $\epsilon$-OT solvers (minimizing a point-wise loss) cannot directly solve the GW problem with its pairwise loss, it is a crucial building block in the most efficient GW solver existing today, which is described below.

**Entropic Gromov-Wasserstein.** In a similar spirit to $\epsilon$-OT, Solomon et al. (2016) proposed to solve an entropy-regularized version of GW problem (Eq. 2). Peyré et al. (2019) introduced a mirror-descent-based algorithm that iteratively linearizes the objective in Eq. 2 and then performs a projection onto $U(\mu,\nu)$ by solving an $\epsilon$-OT problem to obtain an assignment (see Appendix for details). This procedure is repeated for a number of iterations. Since each outer iteration involves solving an OT problem in the projection step, this quickly becomes expensive and intractable even in moderate sample sizes. In our experiments, we observed that entropic GW solvers result in out-of-memory for $N > 25000$ even when running on optimized implementation from `ott-jax` (Cuturi et al., 2022) on a high-end GPU, whereas the implementation in `POT` (Flamary et al., 2021), since it is CPU-based, is intractable already for $N > 8000$. Scetbon et al. (2022) proposed *low-rank GW* that imposes low-rank constraints both on the cost and assignment matrices as an alternative to entropic GW and demonstrated that it could provide speed-up compared to entropic counterpart. We observed that if the data violates the low-rank assumptions, as is generally true for distance matrices and was specifically the case in our real data experiments, the benefits from this approach become void. Explicitly imposing low-rank constraints led to a severe degradation in the quality of the estimated assignment.

## 3. Our approach to the GW problem

In order to scale GW solvers to large sample sizes, we start with the following question: can we find an entropic OT problem whose solution coincides with that of the entropic GW problem (Eq. 2)? The rationale is that, given an unpaired set of samples, if we determine an equivalent entropic OT problem, we can employ fast entropic OT solvers to calculate the assignment. One obvious problem that fits this criterion, by construction, is the entropic OT problem that is solved in the last iterate of the entropic GW solver. However, computing this problem would require iterating through

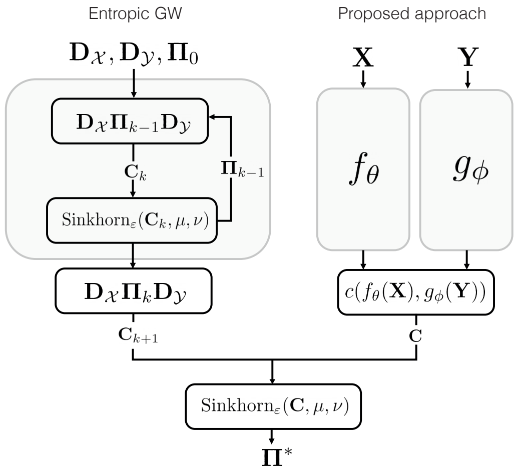
Entropic GW
Proposed approach
$\mathbf{D}_\mathcal{X}, \mathbf{D}_\mathcal{Y}, \mathbf{\Pi}_0$
$\mathbf{D}_\mathcal{X}\mathbf{\Pi}_{k-1}\mathbf{D}_\mathcal{Y}$
$\mathbf{C}_k$
$\mathbf{\Pi}_{k-1}$
$\text{Sinkhorn}_\varepsilon(\mathbf{C}_k, \mu, \nu)$
$\mathbf{D}_\mathcal{X}\mathbf{\Pi}_k\mathbf{D}_\mathcal{Y}$
$\mathbf{C}_{k+1}$
$\mathbf{X}$
$\mathbf{Y}$
$f_\theta$
$g_\phi$
$c(f_\theta(\mathbf{X}), g_\phi(\mathbf{Y}))$
$\mathbf{C}$
$\text{Sinkhorn}_\varepsilon(\mathbf{C}, \mu, \nu)$
$\mathbf{\Pi}^*$

Figure 1: Entropic Gromov-Wasserstein solver (left) solves a sequence of regularized optimal transport ($\epsilon$-OT) problems using the Sinkhorn algorithm. In contrast, the proposed approach learns, via a pair of embeddings, $f_\theta$ and $g_\phi$, the transport cost that directly produces the sought alignment $\mathbf{\Pi}^*$ by solving a single $\epsilon$-OT problem. While the learning of the embeddings still requires multiple calls to the $\epsilon$-OT solver, their cost is amortized at inference time.

the GW solver, and it is thus impractical. By phrasing this question as an optimization problem, we get the following,

$$\begin{aligned} \mathbf{\Pi}^* = &\arg\min_{\mathbf{C}} \left\| \mathbf{D}_\mathcal{X} - \mathbf{\Pi}(\mathbf{C})\mathbf{D}_\mathcal{Y}\mathbf{\Pi}^\top(\mathbf{C}) \right\|_F^2 \\ &\text{s.t. } \mathbf{\Pi}(\mathbf{C}) = \arg\min_{\mathbf{\Pi} \in U(\mu,\nu)} \langle \mathbf{\Pi}, \mathbf{C} \rangle. \end{aligned} \tag{3}$$

It is a bilevel optimization problem: the inner problem is linear OT and it produces an assignment that is optimal with respect to the cost $\mathbf{C}$, and the outer problem demands that the resulting $\mathbf{\Pi}(\mathbf{C})$ is GW-optimal, i.e., it aligns the metrics $\mathbf{D}_\mathcal{X}$ and $\mathbf{D}_\mathcal{Y}$. While seemingly elegant, Equation (3) has two major problems: (i) because $\mathbf{C}$ is unbounded, this objective is very unstable and difficult to optimize; (ii) more practically, Eq. 3 results in a *transductive* approach; given a new set of unpaired samples, this problem needs to be solved anew, which is not scalable.

To mitigate this, instead of optimizing the cost matrix $\mathbf{C}$ (in Eq 3), we propose to *implicitly* parametrize it as a pairwise cost measured on the *learned embeddings* of pointwise features $\mathbf{X}$ and $\mathbf{Y}$. This leads us to the following modified objective,

$$\begin{aligned} \mathbf{\Pi}^* = &\arg\min_{\theta,\phi} \left\| \mathbf{D}_\mathcal{X} - \mathbf{\Pi}(\theta,\phi)\mathbf{D}_\mathcal{Y}\mathbf{\Pi}^\top(\theta,\phi) \right\|_F^2 \\ &\text{s.t. } \mathbf{\Pi}(\theta,\phi) = \arg\min_{\mathbf{\Pi} \in U(\mu,\nu)} \langle \mathbf{\Pi}, c(f_\theta(\mathbf{X}), g_\phi(\mathbf{Y})) \rangle, \end{aligned} \tag{4}$$

where $f, g$ are learnable functions, modeled via neural networks, embedding $\mathbf{X}$ and $\mathbf{Y}$, respectively. It is important to emphasize that the cost is realized through the embedding, while the function $c$ is fixed to the simple Euclidean ($c(\boldsymbol{z}, \boldsymbol{z}') = \|\boldsymbol{z} - \boldsymbol{z}'\|^2$) or cosine ($c(\boldsymbol{z}, \boldsymbol{z}') = \boldsymbol{z}^\top\boldsymbol{z}'$) form. We solve the above problem via gradient descent. In order to backpropagate gradients to $f$ and $g$, we first relax the inner problem to be an $\epsilon$-OT problem, and then employ implicit differentiation (Amos & Kolter, 2017) to calculate $\frac{\partial \mathbf{\Pi}}{\partial c}$ (Eisenberger et al., 2022) which is backpropagated to update the weights of $f$ and $g$.

From a geometric perspective, we are embedding the samples from $\mathcal{X}$ and $\mathcal{Y}$ into a common domain $\mathcal{Z}$, where the samples are *OT-aligned* with the same assignment that makes the metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ *GW-aligned*. From a *computational* point of view, our framework can be viewed as an *amortized entropic GW solver*. Figure 1 presents the parallels between our solver and the entropic GW solver (Solomon et al., 2016). The ground cost of measured on the embeddings $c(f_\theta(\mathbf{X}), g_\phi(\mathbf{Y}))$ can be interpreted as the cost matrix $\mathbf{C}_{k+1} = \mathbf{D}_\mathcal{X}\mathbf{\Pi}_k\mathbf{D}_\mathcal{Y}$ (as depicted in the Fig. 1) produced by running the entropic GW solver for $k$ iterations. Post training, the neural networks can be viewed to be amortizing the GW iterations, in similar spirit to recent amortized optimization techniques proposed for fast calculation of convex conjugates (Amos et al., 2023; Amos, 2022).

From a practical standpoint, this results in an *inductive* GW solver. At inference, when a new set of unpaired samples from $\mathcal{X}$ and $\mathcal{Y}$ are encountered, we simply need to solve an entropic OT problem that is highly scalable. Moreover, since our solver is gradient-descent-based, it allows the flexibility to induce domain knowledge, additional regularization, and inductive biases on the assignment, on the cost, and in the neural networks $f, g$, respectively. We will discuss a few such examples in the sequel. Finally, while our approach may resemble the *inverse OT* (iOT) problem (Dupuy et al., 2016; Chiu et al., 2022) in the sense that it involves the learning of the transport cost, it greatly differs in the minimized objective. While iOT targets finding a cost realizing a given assignment (hence, requiring coupled data), our learning problem does not assume a known target permutation; instead, it tries to find one minimizing the pairwise distance disagreement on unaligned data. From this perspective, the proposed approach can be seen as a variational analog of the iOT problem.

## 4. Extensions

While the objective function in Eq. 4 is easy to *evaluate*, the resulting optimization problem is still an NP-hard QAP. In practice, it is challenging to reach good local minima consistently without imposing further inductive biases. This is also true for the entropic GW solver (Peyré et al., 2019) – while sequential linearization and projection via Sinkhorn algorithm works reasonably in practice, there exist no guar-

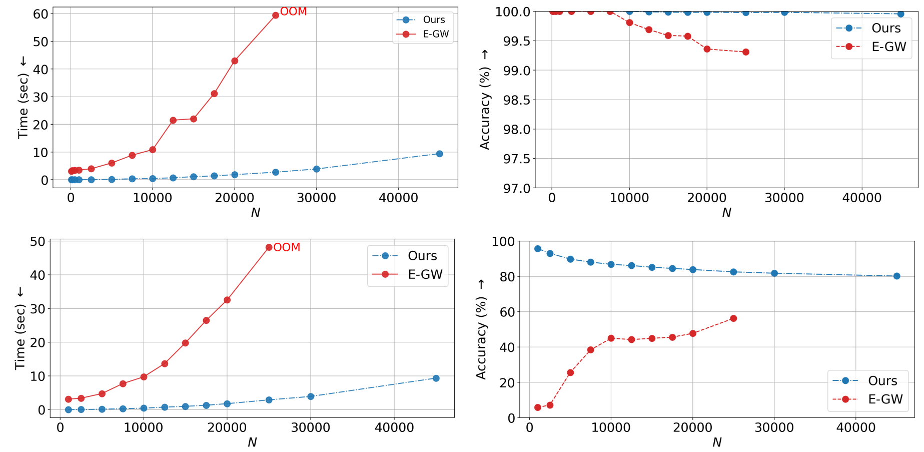

Figure 2: **The proposed solver generalizes to unseen samples and scales to large-sample sizes post-training.** In both top and bottom experiments, $\mathcal{X}$ and $\mathcal{Y}$ are ViT embeddings. The entropic GW solver can only operate in the transductive regime and runs out of memory for $N > 25000$.

antees on its global convergence. In fact, there exist many scenarios where it fails to recover a meaningful local minimum. Here we introduce several regularization techniques in the problem in Eq. 4 to remediate poor convergence: (i) *simulated annealing* of the entropic regularization strength $\epsilon$, and (ii) *spectral-geometric regularization of the OT cost*. We also propose a new objective that matches distance ranks instead of distances themselves that can be employed as an alternative Eq. 4.

**Simulated annealing of $\epsilon$.** While evaluating our solver on the scSNARE-seq data (Chen et al., 2019a), where the goal is to align transcriptomic readouts against those of chromatin accessibility and the ground-truth is available thanks to a co-assaying technique developed by (Chen et al., 2019a), we observed that our solver, while it is accurate on average, it is sensitive to the initialization of the neural networks $f$ and $g$. As a result of symmetries in the metric spaces of these data, we observed that the assignment sometimes consistently mismapped the cell line of GM12878 to H1, and vice-versa. The right panel of Fig. 3 depicts the distribution of alignment errors (lower is better) obtained by solving Eq. 4 with multiple random initializations of the embedding parameters $\theta$ and $\phi$. In the right column, the largest mode corresponds indeed to accurate assignment, whereas the two other modes with larger errors represent the aforementioned symmetry-induced cell-line mismappings.

We mitigate this problem by performing *simulated annealing* on $\epsilon$ of the entropic OT problem within the Sinkhorn layer. We propose a schedule for $\epsilon$ that starts high and is gradually decayed (see Fig. 3, left). Our rationale is that this results in a coarse-to-fine refinement of the learned cost (implicitly parametrized via $f$ and $g$) during training, and it is similar in spirit to the idea of a multi-scale version of kernel matching in shape correspondence problems (Vestner et al., 2017; Melzi et al., 2019; Holzschuh et al., 2020). When $\epsilon$ is high, the entropic regularization is strong, and the resulting assignment is "softer". By scheduling $\epsilon$ from a large value to a small one, we demand that the learned cost matrix, and as a consequence, the resulting assignment, gets refined during training. In practice, we observe that the proposed $\epsilon$-scheduling works remarkably well; it practically reduces the variance across seeds to zero and is effective in breaking symmetries in the metric spaces that lead to bad local minima and making the solver more reliable (Fig. 3, middle).

**Spectral representation on graphs.** Before introducing our proposed spectral-geometric regularization, we provide a brief background on graphs. A familiar reader may skip to the following paragraph. Let $\mathcal{G} = (V, E, \mathbf{\Omega})$ be a weighted graph with the vertex set $V$, edge set $E$, and adjacency matrix $\mathbf{\Omega}$. The *combinatorial graph Laplacian* is defined as $\mathbf{L} = \mathbf{D} - \mathbf{\Omega}$, where $\mathbf{D} = \mathrm{diag}(\mathbf{\Omega 1})$ is the *degree matrix*. Given a scalar-valued *signal* $\mathbf{z} \in \mathbb{R}^{|V|}$ on the graph $\mathcal{G}$, the *Dirichlet energy* is defined to be $\mathbf{z}^{\top}\mathbf{Lz}$, and it measures the *smoothness* of $\mathbf{z}$ on $\mathcal{G}$ (Spielman, 2012). Given two graphs $\mathcal{G}_1 = (V_1, E_1, \mathbf{\Omega}_1)$ and $\mathcal{G}_2 = (V_2, E_2, \mathbf{\Omega}_2)$, the Cartesian product of $\mathcal{G}_1$ and $\mathcal{G}_2$, denoted by $\mathcal{G}_1 \square \mathcal{G}_2$, is defined as a graph with the vertex set $|V_1| \times |V_2|$, on which two nodes $(u, v), (u', v')$ are adjacent if either $u = u'$ and $(v, v') \in E_2$ or $v = v'$ and $(u, u') \in E_1$. The Laplacian of $\mathcal{G}_1 \square \mathcal{G}_2$ is defined as a tensor sum of $\mathbf{L}_1$ and $\mathbf{L}_2$, i.e., $\mathbf{L}_{\mathcal{G}_1 \square \mathcal{G}_2} = \mathbf{L}_1 \oplus \mathbf{L}_2 = \mathbf{L}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{L}_2$. Denote the spectral decompositions of the Laplacians by $\mathbf{L}_1 = \mathbf{\Phi \Lambda}_1 \mathbf{\Phi}^{\top}$ and $\mathbf{L}_2 = \mathbf{\Psi \Lambda}_2 \mathbf{\Psi}^{\top}$. A signal $\mathbf{Z}$ on the product graph $\mathcal{G}_1 \square \mathcal{G}_2$ can be represented using the bases of the individual Laplacians as $\mathbf{Z} = \mathbf{\Phi}^{\top}\mathbf{F}\mathbf{\Psi}$, with the coefficients $\mathbf{F}$.

**Spectral-geometric regularization of the OT cost.** We propose a spectral-geometric regularization on the learned

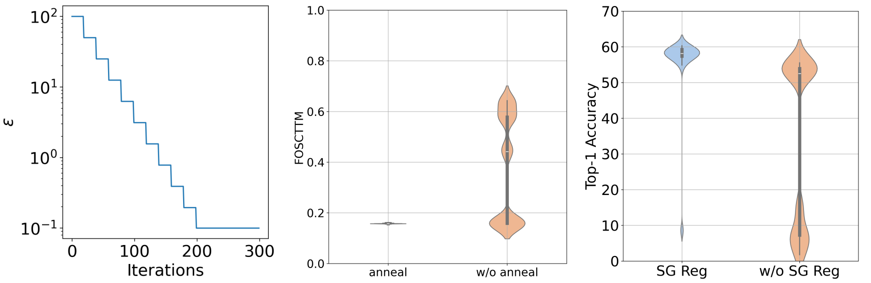

Figure 3: **Simulated annealing of $\epsilon$ and spectral geometric regularization are effective in stabilizing the solver and improving the accuracy of the assignment.** Left: simulated annealing schedule used. Middle: distribution of the alignment error (measured as FOSCTTM) over 20 runs with and without $\epsilon$-annealing. Right: distribution of the alignment error with and without the spectral geometric regularization of the transport cost.

OT cost that demands "similar" items in $\mathcal{X}$ to incur "similar" cost with respect to all items in $\mathcal{Y}$, and vice-versa. To formally represent this notion, let $\mathcal{G}_{\mathcal{X}} = (\mathcal{X}, \mathcal{E}_{\mathcal{X}}, \mathbf{\Omega}_{\mathcal{X}})$ and $\mathcal{G}_{\mathcal{Y}} = (\mathcal{Y}, \mathcal{E}_{\mathcal{Y}}, \mathbf{\Omega}_{\mathcal{Y}})$, be two graphs *inferred* on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and let $\mathbf{L}_{\mathcal{X}}$ and $\mathbf{L}_{\mathcal{Y}}$ be their corresponding graph Laplacians. We interpret the learned OT cost $\mathbf{C} = c(f_\theta(\mathbf{X}), g_\phi(\mathbf{Y}))$ from Eq. 4 as a signal on the product graph $\mathcal{G}_{\mathcal{X}} \,\Box\, \mathcal{G}_{\mathcal{Y}}$, and demand that $\mathbf{C}$ is *smooth* on $\mathcal{G}_{\mathcal{X}} \,\Box\, \mathcal{G}_{\mathcal{Y}}$. The latter smoothness can be expressed as the Dirichlet energy of $\mathbf{C}$ measured on $\mathcal{G}_{\mathcal{X}} \,\Box\, \mathcal{G}_{\mathcal{Y}}$,

$$\begin{aligned}\mathcal{E}_{\text{sm}} &= \operatorname{trace}\left(\mathbf{C}^\top \left(\mathbf{L}_{\mathcal{X}} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{L}_{\mathcal{Y}}\right) \mathbf{C}\right) \\ &= \operatorname{trace}\left(\mathbf{C}^\top \mathbf{L}_{\mathcal{X}} \mathbf{C} + \mathbf{C} \mathbf{L}_{\mathcal{Y}} \mathbf{C}^\top\right),\end{aligned} \tag{5}$$

and added to Eq. 4 as an additional regularization. Figure 3 (right) demonstrates the effectiveness of the proposed spectral regularization on the task of aligning embeddings from neural latent spaces.

From the spectral perspective, interpreting the OT cost $\mathbf{C}$ as a signal on $\mathcal{G}_{\mathcal{X}} \,\Box\, \mathcal{G}_{\mathcal{Y}}$, learning $\mathbf{C}$ given the pointwise features from domains $\mathcal{X}$ and $\mathcal{Y}$ is equivalent to directly learning the functional map of $\mathbf{C}$, this makes our work intimately related to the works of that learn functional maps (Litany et al., 2017; Halimi et al., 2019; Vestner et al., 2017; Boyarski et al., 2022; Kalofolias et al., 2014) from shape correspondence and geometric matrix completion literature.

**Matching ranks instead of distances.** The choice of the comparison criterion for the pairwise distances crucially influences the usability of the GW problem for real applications. Consider, for example, two point clouds that differ only by a scale factor; since distances are not scale-invariant, solving Eq. 4 to match distances would produce meaningless results. As a remedy, we propose to match the *ranks* of the pairwise distances instead of their absolute values. Ranks preserve the order and are insensitive to scale or, more generally, monotone transformations. This departs from the standard framework of GH and GW, which align metric spaces, and generalizes it to a more general problem of performing unpaired alignment by matching non-metric quantities. In order to be able to differentiate the objective with respect to ranks, which is an inherently non-differentiable function, we use the differentiable soft ranking operators introduced by Blondel et al. (2020). We optimize the following modified objective:

$$\begin{aligned}\mathbf{\Pi}^* = &\operatorname*{arg\,min}_{\theta,\phi} \left\| \mathcal{R}_\delta\left(\mathbf{D}_{\mathcal{X}}\right) - \mathcal{R}_\delta\left(\mathbf{\Pi}(\theta,\phi)\mathbf{D}_{\mathcal{Y}}\mathbf{\Pi}^\top(\theta,\phi)\right)\right\|_{\text{F}}^2 \\ &\text{s.t. } \mathbf{\Pi}(\theta,\phi) = \operatorname*{arg\,min}_{\mathbf{\Pi}\in U(\mu,\nu)} \langle \mathbf{\Pi}, c(f_\theta(\mathbf{X}), g_\phi(\mathbf{Y}))\rangle,\end{aligned} \tag{6}$$

where $\mathcal{R}_\delta$ is a soft-ranking operator applied separately to each row of the matrix, and $\delta$ controls the level of "softness" of the rank. Because ranking is a nonlinear operation, this results in a problem that is no longer quadratic in $\mathbf{\Pi}$, it is unclear how standard GW solvers can be adapted to such settings, and also highlights the benefit of having a gradient-descent–based solver. Applying ranking to other groups of distances effectively results in a different GW-like distance. We defer the systematic exploration of this new family of distances to future work.

**Further extensions.** Although we do not explore it within this work, it is easy to see that (i) the proposed framework can be extended to a fused GW (Vayer et al., 2020) setting by adding a linear objective to Eqs. 4 and 6; (ii) the rank of the OT cost can be controlled by modifying the dimension of the embeddings' output by $f$ and $g$; and (iii) when partial supervision is available on the assignment ("semi-supervised" alignment), it can be incorporated into the loss as a data term.

## 5. Experiments

We split this experiment section into three parts. Firstly, we demonstrate that our solver works in the inductive setting and that is much more scalable to large sample sizes in this setting. Secondly, we showcase experiments that demonstrate the effects of (i) *simulated annealing of* $\epsilon$, (ii) *spectral geometric regularization*, and (iii) the *ranking-based* formu-

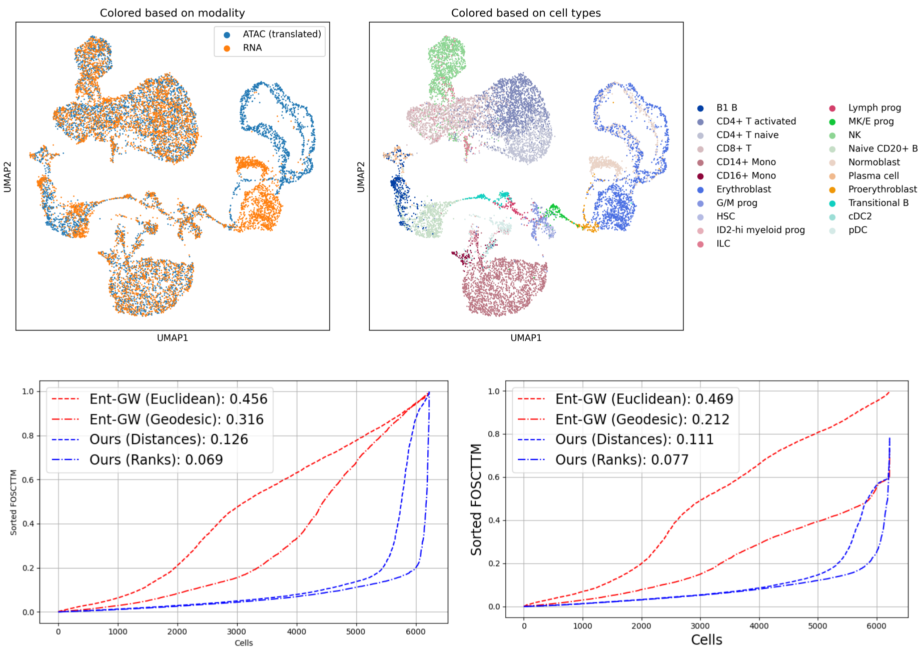

Figure 4: **Qualitative and quantitative results on the human bone marrow single-cell dataset.** Top plots depict the UMAP of the translated cells colored by domain (left) and by the cell type (right).Bottom plots report the FOCSTTM metrics for $\mathcal{Y}$ projected onto $\mathcal{X}$ (left) and $\mathcal{X}$ projected onto $\mathcal{Y}$ (right).

lation. Thirdly, we demonstrate that the proposed solver, in the transductive setting, outperforms the entropic GW solver on two single-cell multiomics benchmarks. We use both real and synthetic data wherever appropriate.

**Inductivity and scale.** In order to evaluate the inductivity of the method and to benchmark it against the entropic GW solver, we consider two experiments (i) when $\mathcal{X}$ and $\mathcal{Y}$ are *isometric*, and (ii) when $\mathcal{X}$ and $\mathcal{Y}$ are not exactly isometric. For the **first experiment**, we consider $\mathcal{X}$ to be CIFAR100 encodings obtained from a vision transformer (Dosovitskiy et al., 2020). We apply an orthogonal transformation to each element of $\mathcal{X}$ to generate $\mathcal{Y}$. We parametrize our encoders $f$ and $g$ to be 3-layer multi-layer perceptrons (MLPs), and optimize the Eq. 4 with respect to their parameters on 200 unaligned samples for 500 iterations (12 seconds). Then, we evaluate our method in an *inductive setting* with an increasing number of unaligned samples available at inference up to $N = 45000$. We benchmark it against the GPU-accelerated entropic GW solver available from `ott-jax` (Cuturi et al., 2022). The results are presented in the top panels of Figure 2. We measure accuracy as whether the predicted correspondence is correct *in terms of the class label*. We observe that both solvers recover the orthogonal transformation perfectly. Further, we can observe that an inductive solver, because it solves only a $\epsilon$-OT problem at inference, is much faster and more memory efficient. Employing an entropic GW solver, on the other hand, goes out of memory for $N > 25000$. Note that the times we reported *do not include* the time required to compute a geodesic distance matrix for both $\mathcal{X}$ and $\mathcal{Y}$, which is significantly time-consuming at large sample sizes (>10 mins for $N = 20000$). In contrast, using our solver would not require computing $\mathbf{D}_\mathcal{X}$ and $\mathbf{D}_\mathcal{Y}$ at inference. For the **second experiment**, we use the data from (Maiorca et al., 2024) and choose $\mathcal{X}$ to be ViT embeddings as in the previous experiment, while $\mathcal{Y}$ is set to be ViT embeddings generated from *rescaled* images. We train the $f$ and $g$ for 1000 iterations ($\sim$ 2 minutes), using 1000 unpaired samples during the training time. The results are presented in the bottom two panels of Fig.2. The results suggest, again, that our solver both generalizes well and scales gracefully with sample sizes, whereas the entropic-GW solver produced inferior results in this setting. These experiments corroborate our claim that our solver both attains high-quality solutions and scales well in the inductive regime.

**Spectral geometric regularization.** For this experiment, we consider the above setting where $\mathcal{X}$ and $\mathcal{Y}$ are two unaligned sets of embeddings obtained from a pre-trained vision transformer (Dosovitskiy et al., 2020). We set $f$ and $g$ to be 3-layer MLPs solve Eq. 4 with and without $\mathcal{E}_{\text{sm}}$ regularization (Eq. 5). We solve this problem on 20 unaligned datasets drawn from $\mathcal{X}$ and $\mathcal{Y}$, each of size $N = 1000$. Figure 3 (right panel) presents the accuracy of the assignment by measuring if the predicted corresponding point belongs to the same class as the groundtruth correspondence. Notice that geometric regularization improves the accuracy of the assignment (+20% in terms of mean accuracy over trials). Moreover, it also reduces the variance thereby inducing meaningful inductive bias into the solver.

**Simulated annealing of $\epsilon$.** As discussed in Section 4, we consider the scSNARE-seq data (Chen et al., 2019b), which is a co-assay of transcriptome and chromatin accessibility measurements performed on $N = 1047$ cells. We run our experiment with and without the proposed simulated annealing of $\epsilon$ for 20 random initializations of $f$ and $g$, the results are presented in Figure 3. We observe that using this seed stabilizes the training process significantly. We used this as a default choice across all real data experiments.

**Ranking-based GW.** Figure 5 (right panel) depicts the assignment produced by the ranking-based GW solver in an inductive setting. We observe that ranking-based GW outperforms the distance-based counterpart in the setting of single-cell multiomic alignment. Consequently, the results that we present in the sequel (Figures 4 and 6) use ranking-based loss and they outperform both the entropic GW solver and the distance-based variant of our solver.

**Single-cell multiomic alignment.** We consider two real-world datasets: (i) **scSNARE-seq** data which contains gene expression (RNA) and chromatin accessibility (ATAC) profiles form 1047 individual cells from four cell lines: H1, BJ, K562, nd GM12878, with known groundtruth thanks to a co-assaying technique developed by (Chen et al., 2019b). We obtained the processed data of RNA and ATAC features from the Demetci et al. (2022), whose method uses entropic GW to align these two modalities and serves as the baseline we evaluate against. (ii) **human bone marrow** single-cell dataset that contains *paired* measurements of single-cell RNA-seq and ATAC-seq measurements released by Luecken et al. (2021). We obtained the processed data from `moscot` (Klein et al., 2023a). In the RNA space, we used PCA embedding of 50 dimensions, and in the ATAC space, we used an embedding given by LSI (latent semantic indexing) embedding, followed by $L_2$ normalization.

In the scSNARE-seq experiment, we used the entropic GW solver with the same hyperparameters used by (Demetci et al., 2022) as the baseline. It was shown by (Demetci et al., 2022) to outperform the other baselines for unpaired alignment on this data. In the bone marrow single-cell experiment, we compared to the entropic GW solver with Euclidean metric and the geodesic distance metric. To establish a fair baseline, following the methodology of Demetci et al. (2022), for both baselines, we perform a grid search on the $\epsilon$ used in Sinkhorn iterations of the solver, and $k$ corresponding to the $k$-NN graph constructed for geodesic computation (for the latter setting), and pick the hyperparameters with the least GW loss. In the case of both bone marrow data and the scSNARE-seq data, we observe that our ranking-based solver produces the best FOCSTTM score (see Appendix). In the case of bone marrow data, especially, our solver produces a significant margin over the entropic GW solvers. The results of scSNARE-seq alignment are presented in Figure 6 in the Appendix. In scSNARE-seq, the margin of our improvement is lower, this could be attributed to limited diversity in cell-lines and small sample-size in scSNARE-seq compared to the bone-marrow data.

## 6. Conclusion

In this paper, we presented a new scalable approach to the Gromov-Wasserstein problem. The GW loss is pair-wise and thus is hard to minimize directly yet simple to evaluate. On the other hand, the OT loss is point-wise and is thus simple to minimize efficiently. We showed practical approaches to learning data embeddings such that the solution of the corresponding OT problem minimizes the GW loss. Unlike existing GW solvers that optimize the assignment matrix or the corresponding dual variables directly, our optimization variables are the parameters of the embedding functions. In addition to better scalability in the transductive regime, the proposed approach is also inductive, as the computed embeddings can be applied to new data previously unseen in training. We further proposed regularization techniques demonstrating consistently better convergence. We emphasize that GW is an NP-hard problem, and no existing polynomial-time algorithms (including ours) are guaranteed to find its global minimum. However, we showed in many synthetic and real data experiments that the proposed solver is significantly more accurate and scalable.

We also introduced a new distance between metric-measure spaces in which distance ranks are matched instead of the distances themselves, which is more appropriate for metric structures coming from distinct modalities that do not necessarily agree quantitatively. Being oblivious to any monotone transformation of the metric structure, this new distance can be applied to general non-metric dissimilarities in the spirit of non-metric multidimensional scaling (MDS) (Cox & Cox, 2000). We defer to future studies the exploration of its geometric and topological properties.

**Limitations.** Our current approach focuses on the discrete GW problem in which the correspondence is found explicitly. Future work should study the continuous setting, with the correspondence represented, e.g., in the form of a functional map (Ovsjanikov et al., 2012) – an operator mapping functions on to $\mathcal{X}$ to functions on $\mathcal{Y}$ which can be represented efficiently using truncated bases of the product graph constructed on $\mathcal{X} \times \mathcal{Y}$. Another limitation is the use of full batches for the minimization of the GW loss, which restricts scalability in the transductive regime. Future studies should consider extending the proposed approach to the mini-batch setting, in the spirit of mini-batch optimal flow-matching (Tong et al., 2023; Klein et al., 2023b).

## References

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

Amos, B. On amortizing convex conjugates for optimal transport. *arXiv preprint arXiv:2210.12153*, 2022.

Amos, B. and Kolter, J. Z. OptNet: Differentiable optimization as a layer in neural networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/amos17a.html.

Amos, B. et al. Tutorial on amortized optimization. *Foundations and Trends® in Machine Learning*, 16(5):592–732, 2023.

Blondel, M., Teboul, O., Berthet, Q., and Djolonga, J. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pp. 950–959. PMLR, 2020.

Blumberg, A. J., Carriere, M., Mandell, M. A., Rabadan, R., and Villar, S. Mrec: a fast and versatile framework for aligning and matching point clouds with applications to single cell molecular data. *arXiv preprint arXiv:2001.01666*, 2020.

Boyarski, A., Vedula, S., and Bronstein, A. Spectral geometric matrix completion. In *Mathematical and Scientific Machine Learning*, pp. 172–196. PMLR, 2022.

Bronstein, A. M., Bronstein, M. M., and Kimmel, R. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006.

Burkard, R. E., Cela, E., Pardalos, P. M., and Pitsoulis, L. S. *The quadratic assignment problem*. Springer, 1998.

Chen, S., Lake, B. B., and Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12): 1452–1457, 2019a.

Chen, S., Lake, B. B., and Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12): 1452–1457, 2019b.

Cheow, L. F., Courtois, E. T., Tan, Y., Viswanathan, R., Xing, Q., Tan, R. Z., Tan, D. S., Robson, P., Loh, Y.-H., Quake, S. R., et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature methods*, 13(10):833–836, 2016.

Chiu, W.-T., Wang, P., and Shafto, P. Discrete probabilistic inverse optimal transport. In *International Conference on Machine Learning*, pp. 3925–3946. PMLR, 2022.

Chowdhury, S., Miller, D., and Needham, T. Quantized gromov-wasserstein. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pp. 811–827. Springer, 2021.

Cox, T. F. and Cox, M. *Multidimensional scaling*. CRC Press, 2000.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. Scot: single-cell multi-omics alignment with optimal transport. *Journal of computational biology*, 29 (1):3–18, 2022.

Deshpande, A. S., Ulahannan, N., Pendleton, M., Dai, X., Ly, L., Behr, J. M., Schwenk, S., Liao, W., Augello, M. A., Tyer, C., et al. Identifying synergistic high-order 3d chromatin conformations from genome-scale nanopore concatemer sequencing. *Nature Biotechnology*, 40(10): 1488–1499, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dupuy, A., Galichon, A., and Sun, Y. Estimating matching affinity matrix under low-rank constraints. *arXiv preprint arXiv:1612.09585*, 2016.

Eisenberger, M., Toker, A., Leal-Taixé, L., Bernard, F., and Cremers, D. A unified framework for implicit sinkhorn differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 509–518, 2022.

Emiya, V., Bonnefoy, A., Daudet, L., and Gribonval, R. Compressed sensing with unknown sensor permutation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1040–1044. IEEE, 2014.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

Gold, S. and Rangarajan, A. Softassign versus softmax: Benchmarks in combinatorial optimization. *Advances in neural information processing systems*, 8, 1995.

Gold, S. and Rangarajan, A. A graduated assignment algorithm for graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(4):377—388, 1996.

Gouil, Q. and Keniry, A. Latest techniques to study dna methylation. *Essays in biochemistry*, 63(6):639–648, 2019.

Grandi, F. C., Modi, H., Kampman, L., and Corces, M. R. Chromatin accessibility profiling by atac-seq. *Nature protocols*, 17(6):1518–1552, 2022.

Gromov, M., Katz, M., Pansu, P., and Semmes, S. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152. Springer, 1999.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome research*, 23 (12):2126–2135, 2013.

Halimi, O., Litany, O., Rodola, E., Bronstein, A. M., and Kimmel, R. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4370–4379, 2019.

Holzschuh, B., Lähner, Z., and Cremers, D. Simulated annealing for 3d shape correspondence. In *2020 International Conference on 3D Vision (3DV)*, pp. 252–260. IEEE, 2020.

Kalofolias, V., Bresson, X., Bronstein, M., and Vandergheynst, P. Matrix completion on graphs. *arXiv preprint arXiv:1408.1717*, 2014.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

Klein, D., Palla, G., Lange, M., Klein, M., Piran, Z., Gander, M., Meng-Papaxanthos, L., Sterr, M., Bastidas-Ponce, A., Tarquis-Medina, M., Lickert, H., Bakhti, M., Nitzan, M., Cuturi, M., and Theis, F. J. Mapping cells through time and space with moscot. *bioRxiv*, 2023a. doi: 10.1101/2023.05.11.540374. URL https://www.biorxiv.org/content/early/2023/05/11/2023.05.11.540374.

Klein, D., Uscidda, T., Theis, F., and Cuturi, M. Generative entropic neural optimal transport to map within and across spaces. *arXiv preprint arXiv:2310.09254*, 2023b.

Kukurba, K. R. and Montgomery, S. B. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11): pdb–top084970, 2015.

Lee, J., Hyeon, D. Y., and Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine*, 52(9):1428–1442, 2020.

Litany, O., Remez, T., Rodola, E., Bronstein, A., and Bronstein, M. Deep functional maps: Structured prediction for dense shape correspondence. In *Proceedings of the IEEE international conference on computer vision*, pp. 5659–5667, 2017.

Luecken, M., Burkhardt, D., Cannoodt, R., Lance, C., Agrawal, A., Aliee, H., Chen, A., Deconinck, L., Detweiler, A., Granados, A., Huynh, S., Isacco, L., Kim, Y., Klein, D., DE KUMAR, B., Kuppasani, S., Lickert, H., McGeever, A., Melgarejo, J., Mekonen, H., Morri, M., Müller, M., Neff, N., Paul, S., Rieck, B., Schneider, K., Steelman, S., Sterr, M., Treacy, D., Tong, A., Villani, A.-C., Wang, G., Yan, J., Zhang, C., Pisco, A., Krishnaswamy, S., Theis, F., and Bloom, J. M. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Maiorca, V., Moschella, L., Norelli, A., Fumero, M., Locatello, F., and Rodolà, E. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Melzi, S., Ren, J., Rodola, E., Sharma, A., Wonka, P., and Ovsjanikov, M. Zoomout: Spectral upsampling for efficient shape correspondence. *arXiv preprint arXiv:1904.07865*, 2019.

Mémoli, F. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., and Rodolà, E. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.

Nitzan, M., Karaiskos, N., Friedman, N., and Rajewsky, N. Gene expression cartography. *Nature*, 576(7785): 132–137, 2019.

Ovsjanikov, M., Ben-Chen, M., Solomon, J., Butscher, A., and Guibas, L. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (ToG)*, 31(4):1–11, 2012.

O'Geen, H., Echipare, L., and Farnham, P. J. Using chip-seq technology to generate high-resolution profiles of histone modifications. *Epigenetics Protocols*, pp. 265–286, 2011.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Rao, A., Barkley, D., França, G. S., and Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

Scetbon, M., Peyré, G., and Cuturi, M. Linear-time gromov wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pp. 19347–19365. PMLR, 2022.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817–820, 2014.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Spielman, D. Spectral graph theory. *Combinatorial scientific computing*, 18:18, 2012.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M. A. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell rna-seq analysis. *Cell stem cell*, 6(5):468–478, 2010.

Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Unnikrishnan, J., Haghighatshoar, S., and Vetterli, M. Unlabeled sensing with random linear measurements. *IEEE Transactions on Information Theory*, 64(5):3237–3253, 2018.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

Vestner, M., Lähner, Z., Boyarski, A., Litany, O., Slossberg, R., Remez, T., Rodola, E., Bronstein, A., Bronstein, M., Kimmel, R., et al. Efficient deformable shape correspondence via kernel matching. In *2017 international conference on 3D vision (3DV)*, pp. 517–526. IEEE, 2017.

Villar, S., Bandeira, A. S., Blumberg, A. J., and Ward, R. A polynomial-time relaxation of the gromov-hausdorff distance. *arXiv preprint arXiv:1610.05214*, 2016.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

Zong, C., Lu, S., Chapman, A. R., and Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*, 338(6114): 1622–1626, 2012.

# A. Appendix.

**Sinkhorn algorithm.** The Sinkhorn algorithm allows efficient solution of the entropy-regularized linear OT problem of the form

$$\min_{\mathbf{\Pi} \in U(\mu,\nu)} \langle \mathbf{C}, \mathbf{\Pi} \rangle + \epsilon \langle \mathbf{\Pi}, \log \mathbf{\Pi} \rangle.$$

Defining the kernel matrix $\mathbf{K} = e^{-\mathbf{C}/\epsilon}$ and initializing $\mathbf{u}_1 = \mathbf{v}_1 = \mathbf{1}$, the algorithm proceeds with iterating

$$\mathbf{u}_{k+1} = \frac{\mu}{\mathbf{K}\mathbf{v}_k}; \qquad \mathbf{v}_{k+1} = \frac{\nu}{\mathbf{K}^\top \mathbf{u}_{k+1}},$$

from which the assignment matrix $\mathbf{\Pi}_{k+1} = \mathrm{diag}(\mathbf{u}_{k+1})\,\mathbf{K}\,\mathrm{diag}(\mathbf{v}_{k+1})$. Here $\mathrm{diag}(\mathbf{u})$ denotes a diagonal matrix with the entries of the vector $\mathbf{u}$ on the diagonal, and exponentiation and division are performed element-wise. The iterations are usually stopped when the change $\|\mathbf{\Pi}_{k+1} - \mathbf{\Pi}_k\|$ falls below a pre-defined threshold.

**Entropic GW solver.** The entropic GW solver aims at solving the entropy-regularized GW problem

$$\min_{\mathbf{\Pi} \in U(\mu,\nu)} \|\mathbf{D}_{\mathcal{X}} - \mathbf{\Pi}\mathbf{D}_{\mathcal{Y}}\mathbf{\Pi}^\top\|_{\mathrm{F}}^2 + \epsilon \langle \mathbf{\Pi}, \log \mathbf{\Pi} \rangle.$$

Without the entropy term, the problem is a linearly constrained quadratic program, which Gold and Rangarajan (Gold & Rangarajan, 1996) proposed to solve as a sequence of linear programs. Applied here, this idea leads to a sequence of entropy-regularized linear OT problems of the form

$$\mathbf{\Pi}_{k+1} = \arg \min_{\mathbf{\Pi} \in U(\mu,\nu)} \langle \mathbf{C}_{k+1}, \mathbf{\Pi} \rangle + \epsilon \langle \mathbf{\Pi}, \log \mathbf{\Pi} \rangle,$$

with the cost $\mathbf{C}_{k+1} = \mathbf{D}_{\mathcal{X}} \mathbf{\Pi}_k \mathbf{D}_{\mathcal{Y}}$ defined using the previous iteration. Each such problem is solved using Sinkhorn inner iterations.

**Barycentric projection.** For visualization and comparison purposes, it is often convenient to represent the points from $\mathcal{X}$ and $\mathcal{Y}$ in the same space. Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ and $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ denote the coordinates of the points in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Given the “soft" assignment $\mathbf{\Pi}$ and using $\mathcal{Y}$ as the representation space, we can represent $\mathbf{X}$ in the form of the weighted sum, $\hat{\mathbf{X}} = \mathbf{Y}\mathbf{\Pi}$, so that the representation of a point $\mathbf{x}_i$ in $\mathbf{Y}$ becomes (Alvarez-Melis & Jaakkola, 2018)

$$\hat{\mathbf{x}}_i = \sum_j \pi_{ij} \mathbf{y}_j,$$

We remind that $\mathbf{\Pi}$ is by definition a stochastic matrix, implying that the weights in the above sum are non-negative and sum to 1.

**FOSCTTM score.** The *fraction of samples closer than the true match* (FOSCTTM) measures the alignment quality of two equally-sized sets with known ground-truth correspondence. Let $U = \{\mathbf{u}_i\}$ and $V = \{\mathbf{v}_i\}$ be two sets of points in a common metric space $\mathcal{Z}$ ordered, without loss of generality, in trivial correspondence order (i.e., every $\mathbf{u}_i$ corresponds to $\mathbf{v}_i$). Given a point $\mathbf{u}_i$, we define the fraction of points in $V$ that are closer to it than the true match $\mathbf{v}_i$,

$$p_i = \frac{1}{N} \left| \{ j : d_{\mathcal{Z}}(\mathbf{u}_i, \mathbf{v}_j) < d_{\mathcal{Z}}(\mathbf{u}_i, \mathbf{v}_i) \} \right|.$$

Similarly, we define the fraction of points in $U$ that are closer to $\mathbf{v}_i$ and the true match $\mathbf{u}_i$,

$$q_i = \frac{1}{N} \left| \{ j : d_{\mathcal{Z}}(\mathbf{v}_i, \mathbf{u}_j) < d_{\mathcal{Z}}(\mathbf{v}_i, \mathbf{u}_i) \} \right|.$$

The FOCSTTM score is defined as

$$\mathrm{FOCSTTM} = \frac{1}{2N} \sum_{i=1}^{N} (p_i + q_i).$$

The score is normalized in the range of $[0,1]$ with perfect alignment having $\mathrm{FOCSTTM} = 0$.

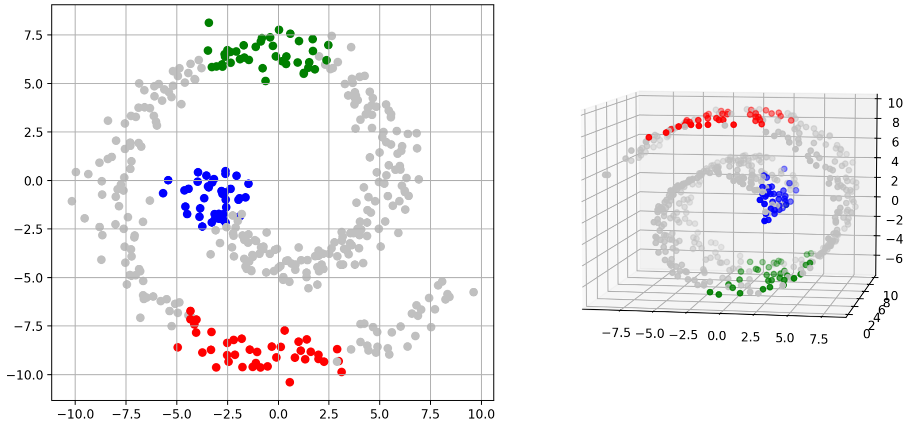

Figure 5: **Qualitative evaluation of the proposed GW solver in inductive setting.** The plot depicts the assignment produced by our distance-based GW solver (Eq. 4) on a new set of samples.

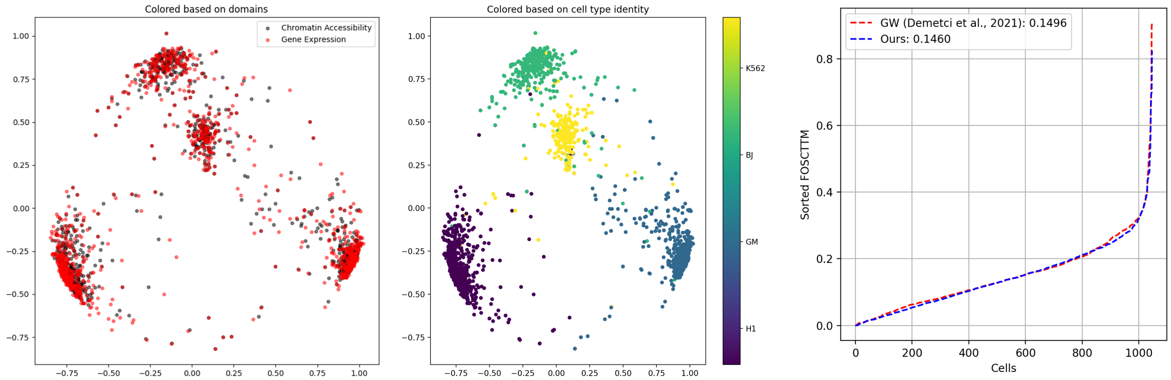

Figure 6: **Qualitative and quantitative results on the scSNARE-seq dataset.** Left and middle: Aligned samples from ATAC and RNAl, colored by the domains (ATAC: black, RNA: red) and cell types, respectively. Right: the sorted FOCSTTM plot, a quantitative metric measuring the quality of the assignment.