Breaking Isometric Ties and Introducing Priors in Gromov-Wasserstein Distances

Pinar Demetci
Eric and Wendy Schmidt Center [1]
Broad Institute of MIT and Harvard

Quang Huy Tran
Université Bretagne-Sud, IRISA
CMAP, Ecole Polytechnique, IP Paris

Ievgen Redko [2]
Paris Noah's Ark Lab
Huawei Technologies

Ritambhara Singh [2]
Department of Computer Science
Center for Computational Molecular Biology
Brown University

Abstract

Gromov-Wasserstein distance has many applications in machine learning due to its ability to compare measures across metric spaces and its invariance to isometric transformations. However, in certain applications, this invariant property can be too flexible, thus undesirable. Moreover, the Gromov-Wasserstein distance solely considers pairwise sample similarities in input datasets, disregarding the raw feature representations. We propose a new optimal transport formulation, called Augmented Gromov-Wasserstein (AGW), that allows for some control over the level of rigidity to transformations. It also incorporates feature alignments, enabling us to better leverage prior knowledge on the input data for improved performance. We first present theoretical insights into the proposed method. We then demonstrate its usefulness for single-cell multi-omic alignment tasks and heterogeneous domain adaptation in machine learning.

1 INTRODUCTION

Optimal transport (OT) theory provides a fundamental tool for comparing and aligning probability measures omnipresent in machine learning (ML) tasks. Following the least effort principle, OT and its associated metrics offer many attractive properties that other divergences, such as the popular Kullback-Leibler or Jensen-Shannon divergences, lack. For instance, OT borrows key geometric properties of the underlying "ground" space on which the distributions are defined (Villani, 2008) and enjoys non-vanishing gradients when measures have disjoint support (Arjovsky et al., 2017). OT theory has also been extended to the challenging case of comparing probability measures supported on different metric-measure spaces. In this scenario, the Gromov-Wasserstein (GW) distance seeks an optimal matching between points in the supports of the considered distributions that will minimize the distortion of intra-domain distances upon such matching.

Since its proposal by Memoli (2011) and further extensions by Peyré et al. (2016), GW distance has been successfully used in a wide range of applications, including domain adaptation (Yan et al., 2018), computational biology (Nitzan et al., 2019; Cao et al., 2021; Cang and Nie, 2020; Demetci et al., 2020, 2022a), generative modeling (Bunne et al., 2019), and reinforcement learning (Nakagawa et al., 2022).

1.1 Limitations of Prior Work

Successful applications of GW distance are often attributed to its invariance to distance-preserving transformations (also called "isometries") of the input domains. Since GW formulation considers only intra-domain distances, it is naturally invariant to any transformation that does not alter them. While this is a blessing in many applications, for example, comparing graphs with the unknown ordering of nodes, it may become a curse when one has to choose the "right" isometry among many that yield the same GW distance. How could one break such ties while keeping the at-

[1]Work partially completed while at Brown University, Department of Computer Science, Center for Computational Molecular Biology.

[2]Co-corresponding authors: ievgen.redko@gmail.com, ritambhara@brown.edu

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

tractive properties of the GW distance? This question remains to be addressed in the field.

Additionally, GW distances are often used in tasks where one may have some *a priori* knowledge about the mapping between the two considered spaces. For example, in single-cell applications, mapping a group of cells in similar tissues across species helps understand the evolutionarily conserved and diverging cell types and functions (Kriebel and Welch, 2022). When performed using OT, this cross-species cell mapping may benefit from the knowledge about an overlapping set of orthologous genes [3]. GW formulation does not offer any straightforward way to incorporate this knowledge, which may lead to suboptimal performance.

1.2 Our Contributions

In this paper, we introduce a new OT formulation that addresses the drawbacks of the GW distance mentioned above. We summarize our contributions as follows:

1. We propose Augmented Gromov-Wasserstein (AGW), a new formulation that leverages both pairwise sample similarities in input datasets and their raw data representations;
2. We demonstrate that AGW allows for tighter control over the isometric transformations of the GW distance and helps break isometric ties;
3. We show that AGW can incorporate prior knowledge to guide how the two metric spaces should be compared, which improves object comparisons;
4. We provide a theoretical analysis of the properties of the proposed formulation and examples that concretely illustrate its unique features;
5. Our empirical results show that AGW outperforms previously proposed cross-domain OT methods in several downstream tasks and tends to converge in fewer iterations than GW distance. We first focus on real-world applications in computational biology, namely the single-cell data integration tasks. Then, we also illustrate its generalizability to the heterogeneous domain adaptation in ML.

The paper is organized as follows. Section 2 presents key notions from the OT theory utilized in the rest of the paper. Section 3 presents our proposed AGW formulation and analyzes its theoretical properties. In Section 4, we present several empirical studies for the single-cell alignment task and demonstrate the applicability of our method to the heterogeneous domain adaptation task. We conclude our paper in Section 5 with a discussion of potential future work.

[3] Genes in two different species that originated from a common ancestor and largely maintained their function and sequence during speciation.

2 TECHNICAL BACKGROUND

This section briefly presents some background knowledge, including the Kantorovich's formulation of the OT problem and two relevant OT-based distances proposed to match samples across incomparable spaces.

In what follows, we denote by $\Delta_n = \{w \in (\mathbb{R}_+)^n : \sum_{i=1}^n w_i = 1\}$ the simplex histogram with n bins. We use $\otimes$ for tensor-matrix multiplication, *i.e.*, $L \otimes B$ is the matrix $(\sum_{k,l} L_{i,j,k,l} B_{k,l})_{i,j}$ for a tensor $L = (L_{i,j,k,l})_{i,j,k,l}$ and a matrix $B = (B_{i,j})_{i,j}$. We use $\langle \cdot, \cdot \rangle$ for the matrix scalar product associated with the Frobenius norm $\|\cdot\|_F$. We write $\mathbf{1}_d \in \mathbb{R}^d$ for a d-dimensional vector of ones. We use the terms "coupling matrix", "transport plan" and "correspondence matrix" interchangeably. A point in the space can also be called "an example" or "a sample". Given an integer $n \geq 1$, denote $[n] := \{1, ..., n\}$.

2.1 Kantorovich's Problem and Wasserstein Distance

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^{m \times d}$ be two input matrices, $\mathbf{C} \in \mathbb{R}^{n \times m}$ be a cost (or ground) matrix. Given two discrete probability measures $\mu \in \Delta_n$ and $\nu \in \Delta_m$, Kantorovich's formulation of OT seeks a coupling γ minimizing the following quantity:

$$W_{\mathbf{C}}(\mu, \nu) = \min_{\boldsymbol{\gamma} \in \Pi(\mu, \nu)} \langle \mathbf{C}, \boldsymbol{\gamma} \rangle, \tag{1}$$

where $\Pi(\mu, \nu)$ is the set of probability distributions on $\Delta_{n \times m}$ with marginals μ and ν. When $\mathbf{C}_{ij} = ||x_i - y_j||^p$, for $p \geq 1$, such an optimization problem defines a proper metric on the space of probability distributions called the Wasserstein distance.

2.2 Gromov-Wasserstein Distance

Samples of input matrices in different spaces, *i.e.*, $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^{m \times d'}$ with $d \neq d'$, are incomparable since it is not possible to define a cost function as the distance between points across the input spaces. To circumvent this limitation of the Wasserstein distance, Memoli (2011) proposed the Gromov-Wasserstein (GW) distance, defined as follows:

$$\mathrm{GW}(\mathbf{X}, \mathbf{Y}, \mu, \nu, d_X, d_Y) := \min_{\boldsymbol{\gamma} \in \Pi(\mu, \nu)} \mathcal{L}_{GW}(\boldsymbol{\gamma}),$$

where

$$\mathcal{L}_{\mathrm{GW}}(\boldsymbol{\gamma}) := \sum_{i,j,k,l} \big(d_X(x_i, x_k) - d_Y(y_j, y_l)\big)^2 \gamma_{i,j}\gamma_{k,l}$$
$$= \langle L(\mathbf{D}_X, \mathbf{D}_Y) \otimes \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle.$$

Here, $\big(L(\mathbf{D}_X, \mathbf{D}_Y)\big)_{i,j,k,l} = \big(d_X(x_i, x_k) - d_Y(y_j, y_l)\big)^2$, where $(x_i, x_k) \in \mathbb{R}^d \times \mathbb{R}^d$ and $(y_j, y_k) \in \mathbb{R}^{d'} \times \mathbb{R}^{d'}$ are tuples of samples in $\mathbf{X}$ and $\mathbf{Y}$, respectively. d_X and d_Y are distances on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, respectively, so that $(\mathbf{D}_X)_{i,k} = d_X(x_i, x_k)$ and $(\mathbf{D}_Y)_{j,l} = d_Y(y_j, y_l)$.

CO-Optimal transport Redko et al. (2020) introduced an alternative to GW distance, termed CO-Optimal transport (COOT). Instead of relying on the intra-domain distance matrices $\mathbf{D}_X$ and $\mathbf{D}_Y$, COOT uses the raw feature information (*i.e.*, the coordinates of the samples) and jointly learns two couplings, one corresponding to sample alignments (denoted by $\boldsymbol{\gamma}^s$ below), and the other corresponding to feature alignments ($\boldsymbol{\gamma}^v$ below). More precisely, COOT assigns two histograms $\mu' \in \Delta_d$ and $\nu' \in \Delta_{d'}$ to the features (columns) of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and defines the distance between two matrices $\mathbf{X}$ and $\mathbf{Y}$ as

$$\mathrm{COOT}(\mathbf{X}, \mathbf{Y}, \mu, \nu, \mu', \nu') := \min_{\substack{\boldsymbol{\gamma}^s \in \Pi(\mu,\nu) \\ \boldsymbol{\gamma}^v \in \Pi(\mu',\nu')}} \mathcal{L}_{\mathrm{COOT}}(\boldsymbol{\gamma}^s, \boldsymbol{\gamma}^v),$$

where

$$\mathcal{L}_{\mathrm{COOT}}(\boldsymbol{\gamma}^s, \boldsymbol{\gamma}^v) := \sum_{i,j}\sum_{a,b} L(x_{i,k}, y_{j,l})\gamma^s_{i,j}\gamma^v_{a,b}$$
$$= \langle L(\mathbf{X}, \mathbf{Y}) \otimes \boldsymbol{\gamma}^v, \boldsymbol{\gamma}^s \rangle.$$

In what follows, we consider $L(x_{i,a}, y_{j,b}) = (x_{i,a} - y_{j,b})^2$ and write simply $\mathrm{GW}(\mathbf{X}, \mathbf{Y})$ and $\mathrm{COOT}(\mathbf{X}, \mathbf{Y})$ when μ, ν, μ', ν' are uniform and when the choice of d_X and d_Y is of no importance.

3 AUGMENTED GROMOV-WASSERSTEIN (AGW)

Here, we start by outlining the motivation for our proposed formulation, highlighting the different invariance properties of GW distance and COOT. Then, we detail our AGW method that interpolates between the two, followed by a theoretical study of its properties. Our implementation is available at https://github.com/pinardemetci/AGW-AISTATS24, along with examples and demonstrations.

3.1 Motivation

Our motivation for this work comes from leveraging different invariance properties of the GW distance and COOT in order to have a tighter control over isometric transformations when comparing objects across different metric spaces.

3.1.1 Invariants of GW distance

GW distance remains unchanged under isometric transformations of the input data as it compares intradomain pairwise distances. This property has contributed much to the popularity of GW distance, as isometries naturally appear in many applications. However, not all isometries are equally desirable. For instance, a rotation of the handwritten digit 6 seen as a discrete measure can lead to its slight variation for small angles or to a digit 9 when the angle is close to 180 degrees. In both cases, however, the GW distance remains unchanged, making it insufficient to distinguish the two digits apart, unable to break such isometric ties

3.1.2 Invariants of COOT

Unlike GW distance, COOT has fewer degrees of freedom in terms of invariance to global isometric transformations as it is limited to permutations of rows and columns of the two matrices, and not all isometric transformations can be achieved via such permutations. For example, Figure S1 shows the effect of the sign change and image rotation in a handwritten digit matching task, to which GW distance is invariant while COOT is not. Additionally, COOT is strictly positive for any two datasets of different sizes either in terms of features or samples, making it much more restrictive than GW distance. It thus provides a finer-grained control when comparing complex objects, yet it lacks the robustness of GW distance to frequently encountered transformations between the two datasets. Further, unlike GW distance, it is invariant to local isometries that can be achieved via permutations of a subset of features.

3.2 AGW Formulation

Given the above discussion on the invariants of COOT and GW distance, interpolating between them will restrict each other's invariants. Additionally, interpolating with COOT is a natural way to introduce raw feature alignments in GW formulation, which allows for leveraging priors on them. We call this interpolation **Augmented GW (AGW)** and define it as follows:

$$\mathrm{AGW}_\alpha(\mathbf{X}, \mathbf{Y}) := \min_{\substack{\boldsymbol{\gamma}^s \in \Pi(\mu,\nu) \\ \boldsymbol{\gamma}^v \in \Pi(\mu',\nu')}} \mathcal{L}_\alpha(\boldsymbol{\gamma}^s, \boldsymbol{\gamma}^v), \qquad (2)$$

where

$$\mathcal{L}_\alpha(\boldsymbol{\gamma}^s, \boldsymbol{\gamma}^v) = \alpha\, \mathcal{L}_{\mathrm{GW}}(\boldsymbol{\gamma}^s) + (1-\alpha)\, \mathcal{L}_{\mathrm{COOT}}(\boldsymbol{\gamma}^s, \boldsymbol{\gamma}^v)$$
$$= \alpha\, \langle L(\mathbf{D}_X, \mathbf{D}_Y) \otimes \boldsymbol{\gamma}^s, \boldsymbol{\gamma}^s \rangle$$
$$+ (1-\alpha)\, \langle L(\mathbf{X}, \mathbf{Y}) \otimes \boldsymbol{\gamma}^v, \boldsymbol{\gamma}^s \rangle,$$

for $0 \leq \alpha \leq 1$. The AGW problem always admits a solution. Indeed, as the objective function is continuous

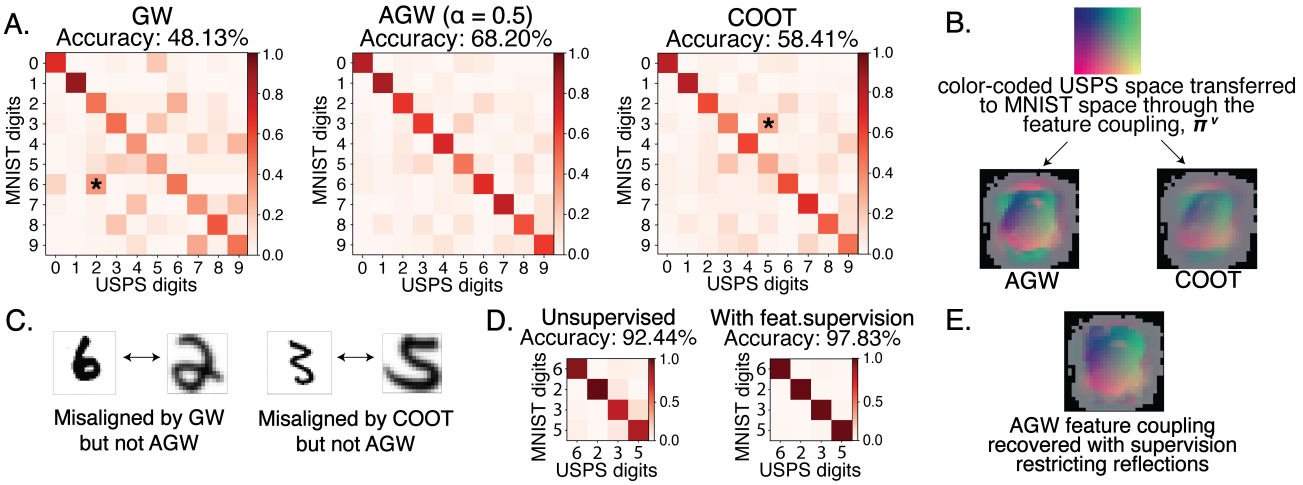

Figure 1: Aligning Digits from MNIST and USPS Datasets. **(A)** Confusion matrices of GW, AGW with $\alpha = 0.5$ and COOT.(*) denote pair alignments ; **(B)** Feature coupling $\boldsymbol{\gamma}^v$ of AGW compared to COOT; **(C)** Illustration of a case from where GW's and COOT's invariants are detrimental for obtaining a meaningful comparison, while AGW remains informative. **(D)** Example showing improved digit alignment with feature-level supervision that restricts reflections **(E)** Feature coupling recovered by AGW ($\alpha = 0.5$) in the supervised setting of (D).

and the sets of admissible couplings are compact, the existence of minimum and minimizer is guaranteed.

Our interpolation offers several important benefits. First, COOT term ensures that AGW will take different values for any two isometries whenever $d \neq d'$. Intuitively, AGW's value will then depend on how "far" a given isometry is from a permutation of rows and columns of the inputs. Thus, we restrict a broad class of (infinitely many) transformations that GW distance cannot distinguish and we tell them apart by assessing whether they can be approximately obtained by simply swapping 1D elements in input matrices.

Second, combining the objective functions of COOT and GW distance allows us to effectively influence the optimization of $\boldsymbol{\gamma}^s$ by introducing priors on feature matchings through $\boldsymbol{\gamma}^v$ and vice versa. This can be achieved by penalizing the costs of matching certain features in the COOT term to influence the optimization of $\boldsymbol{\gamma}^v$. This prior knowledge guides how the two metric spaces should be compared and improves empirical performance. These key properties explain our choice of calling it "augmented": we equip GW distance with an ability to provide finer-grained object comparisons by breaking isometric ties and/or guiding the matching using available prior knowledge.

3.3 Illustrations

We illustrate AGW's properties on a task of aligning handwritten digits from MNIST (LeCun et al., 2010) (28×28 pixels) and USPS datasets (16×16 pixels) (Hull, 1994) in Figure 1, where AGW with $\alpha = 0.5$ outperforms both GW distance and COOT in alignment accuracy (Panel A). The black asterisks show some digit pairs that significantly benefit from AGW interpolation, which are $6 - 2$ for GW distance and $3 - 5$ for COOT. Panel C visualizes examples from these digit pairs that are misaligned by GW distance and COOT but not by AGW [4]. Here, we observe that 6-2 misalignment by GW optimal transport is likely because one is a close reflection of the other across the y-axis. Similarly, COOT mismatches 3 and 5 as one can obtain 3 from 5 by a local permutation of the upper half of the pixels. Panel B visualizes the feature couplings obtained by AGW (on the left) and COOT (on the right). The feature coupling by COOT confirms that COOT allows for a reflection across the y-axis on the upper half of the image but not on the lower half. With AGW, both of these misalignments partially improve, likely because (1) the correct feature alignments in the lower half of the images prevent 6 and 2 from being matched and (2) GW distance is non-zero for 5-3 matches since the transformation is not applied to the whole image. In Panels D and E, we also show that providing supervision on feature alignments to restrict local reflections further improves AGW's performance.

Similar improvement can be seen for aligning cells (samples) for two different single-cell measurements (*i.e.*, measurements generating different types of features) (Chen et al., 2019) in Figure S2: Panel A shows that AGW consistently maps the 4 cell types in the data better than GW alignment (a popular method for this

[4] Here, we define "aligned pairs" as pairs of digits with the highest coupling probabilities.

Algorithm 1 BCD Algorithm to Solve AGW

Initialize $\boldsymbol{\gamma}^s$ and $\boldsymbol{\gamma}^v$
repeat
  Calculate $L_v = L(\mathbf{X}, \mathbf{Y}) \otimes \boldsymbol{\gamma}^s$.
  For fixed $\boldsymbol{\gamma}^s$, solve the OT problem: $\boldsymbol{\gamma}^v \in \arg\min_{\boldsymbol{\gamma} \in \Pi(\mu', \nu')} \langle L_v, \boldsymbol{\gamma} \rangle$.
  Calculate $L_s = L(\mathbf{X}, \mathbf{Y}) \otimes \boldsymbol{\gamma}^v$.
  For fixed $\boldsymbol{\gamma}^v$, solve the fused GW problem: $\boldsymbol{\gamma}^s \in \arg\min_{\boldsymbol{\gamma} \in \Pi(\mu, \nu)} \alpha\, \mathcal{L}_{\mathrm{GW}}(\boldsymbol{\gamma}^s) + (1-\alpha)\langle L_s, \boldsymbol{\gamma}^s \rangle$.
until convergence

task (Cao et al., 2021; Demetci et al., 2020, 2022a; Cao et al., 2022)) over 50 random subsampling of cells. The 2D projection of alignments in Panel B shows that GW distance sometimes completely swaps the cell type clusters when they have a similar number of cells, whereas AGW is more robust to this phenomenon.

3.4 Optimization

For simplicity, let $n = m$ and $d = d'$. With the squared loss in both GW and COOT terms, the computational trick by Peyré et al. (2016) can be applied, which reduces the complexity of AGW from $O(n^4 + n^2 d^2)$ to $O(n^3 + dn^2 + nd^2)$. For optimization, we use the block coordinate descent (BCD) algorithm, where we alternatively fix one coupling and minimize AGW with respect to the other (Algorithm 1). Each iteration then consists of solving two OT problems. To further accelerate the optimization, entropic regularization (Cuturi, 2013) can be used on either $\boldsymbol{\gamma}^s$, $\boldsymbol{\gamma}^v$, or both. In practice, we rely on the built-in functions of the Python Optimal Transport package (Flamary et al., 2021).

3.5 Theoretical Analysis

Intuitively, we expect that AGW interpolates between GW distance and COOT, and satisfies a relaxed triangular inequality since COOT and GW distance are both metrics, similarly to Fused Gromov-Wasserstein (FGW) distance (Vayer et al., 2019) . The following result summarizes these observations, with proofs presented in Appendix A.

Proposition 1. *For every $\alpha \in [0, 1]$, given two input matrices $\mathbf{X}$ and $\mathbf{Y}$,*

1. *When $\alpha \to 0$ (or 1), one has $AGW_\alpha(\mathbf{X}, \mathbf{Y}) \to COOT(\mathbf{X}, \mathbf{Y})$ (or $GW(\mathbf{X}, \mathbf{Y})$).*
2. *AGW satisfies the relaxed triangle inequality: for any input matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, one has $AGW_\alpha(\mathbf{X}, \mathbf{Y}) \leq 2\big(AGW_\alpha(\mathbf{X}, \mathbf{Z}) + AGW_\alpha(\mathbf{Z}, \mathbf{Y})\big)$.*

A more intriguing question is about the invariants that AGW exhibits. $\mathcal{O}_d$ and $\mathcal{P}_d$ denote the sets of orthogonal and permutation matrices of size d, respectively. Given a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, we assume that

Assumption 1. *$\mathbf{X}$ is full-rank and has exactly $\min(n, d)$ distinct singular values.*

The full-rank assumption is not uncommon in the machine learning literature (Kawaguchi, 2016) and can be easily met in practice. Additionally, not only the Hermitian matrices with repeated eigenvalues are rare (see page 56 in (Tao, 2012)), but we can also show that

Corollary 1. *The set of Hermitian matrices with repeated eigenvalues has zero Lebesgue measure.*

Since the singular values of $\mathbf{X}$ are determined by the symmetric matrix $\mathbf{X}\mathbf{X}^T$, Corollary 1 assures that it is reasonable to exclude all symmetric matrices with repeated eigenvalues. With these, we present:

Theorem 1.

1. *Given matrices $\mathbf{X}$ and $\mathbf{Y}$, if $\mu = \nu$ and $\mathbf{Y}$ is obtained by permuting columns of $\mathbf{X}$ via the permutation σ_c (so $\nu' = (\sigma_c)_{\#}\mu'$), then $AGW_\alpha(\mathbf{X}, \mathbf{Y}) = 0$.*
2. *Suppose $\mathbf{X} \in \mathbb{R}^{n \times d}$ satisfies Assumption 1. For any $0 < \alpha < 1$, if $n \geq d$ and $AGW_\alpha(\mathbf{X}, \mathbf{Y}) = 0$, then there exist a symmetric orthogonal matrix $O \in \mathcal{O}_d$ and a permutation matrix $P \in \mathcal{P}_d$ such that $\mathbf{Y} = \mathbf{X}OP$.*

Despite the simplicity of the interpolation structure, the invariants induced by AGW present novel and nontrivial challenges for theoretical analysis. While sharing basic invariants, such as feature swaps, AGW covers much fewer isometries than GW distance. Unlike GW distance, AGW is not fully invariant under translation. However, only its minimum is shifted by a constant under translation, while the feature and sample alignments remain unchanged (discussed in Appendix A.4). Similar to COOT, AGW only has at most finitely many, whereas GW has infinitely many isometries. Under mild conditions, when AGW vanishes, only transformations with a particular structure (compositions of a permutation and a symmetric orthogonal transformation) are eligible. Given the superior empirical performance of AGW over GW and COOT, such isometries appear meaningful and relevant in real-world tasks.

3.6 Related Work

Most related work to our interpolation structure is the Fused Gromov-Wasserstein (FGW) divergence (Vayer et al., 2019) that compares structured objects. Its objective function is a convex combination of the GW term defined based on the pairwise intra-domain distances and the Wasserstein term defined over auxiliary

features that live in the same space for both input matrices. Despite the structural resemblance to FGW; however, AGW and FGW fundamentally differ in several crucial ways:

1. **Structure of input data:** FGW needs two different views of data: relational and auxiliary information for the GW and Wasserstein terms, respectively. To illustrate this, assume that the node embeddings of two graphs are $X \in \mathbb{R}^{n\times d}$ and $X' \in \mathbb{R}^{n'\times d'}$. FGW requires not only intra-graph structure (*e.g.*, adjacency matrices computed based on X and X') in the GW term, but also auxiliary data (independent of X and X', *e.g.*node colors) for the Wasserstein term. For many datasets, including in the applications presented in this paper, such auxilary data is not available. AGW, on the other hand, can directly operate on the embeddings X and X' without any auxiliary information, even when comparing objects across metric spaces. Moreover, the "features" in AGW refer to the columns of X and X', while FGW refers to the auxiliary information as "features". As such, the notion of "features" also differs between AGW and FGW.

2. **Control over isometries:** The above difference implies that the invariants induced by FGW are guided by both structural and auxiliary information. As demonstrated by Vayer et al. (2020), for example, FGW is invariant to rotations. By contrast, since AGW operates on the raw feature space, its invariants are only controlled by (and their corresponding structural information or "intradomain distance matrices", e.g. adjacency matrix computed on node embeddings from the previous example). Section 3.3 demonstrates that AGW leverages interpolation with COOT to provide some explicit control over the invariants of GW distance, for example under rotations (Figure S1) and leads to more meaningful cross-domain matchings. As such, Theorem 1 is the first result of its kind aiming at characterizing the invariances resulting from such interpolation.

3. **Use of transport plans:** FGW only uses one common sample coupling for computing (sample-level) alingments based on both structural and auxiliary feature spaces. By contrast, AGW learns two different couplings, one for sample-level and one for raw feature-level correspondences/alignments.

4 EXPERIMENTAL EVALUATIONS

To test its empirical performance, we apply AGW to the single-cell multi-omics alignment and heterogeneous domain adaptation (HDA) tasks. Overall, we aim to empirically answer: **(1)** Does tightening the invariances improve upon GW's performance in tasks where it was previously used? and **(2)** Does prior knowledge introduced in AGW help in obtaining better cross-domain matchings?

We pick other cross-domain OT methods as baselines, namely COOT, GW, and their unbalanced counterparts, UCOOT (Tran et al., 2023) and UGW (Sejourne et al., 2021). Note that we leave extending AGW to unbalanced scenarios for future work.

In semi-supervised HDA tasks (Table 2), we also consider the KPG-RL and KPG-GW-RL methods Gu et al. (2022), which were developed specifically for leveraging prior information on keypoints in HDA applications. We consider entropic regularization for all methods on both sample and (when applicable) feature couplings. We keep the hyperparameter values considered for all regularization coefficients consistent across all methods. We report the results of the best-performing hyperparameter combination after tuning on a validation set for each method in each experiment. We report empirical runtimes in Appendix E and detail our experimental setup in Appendix F.

4.1 Integrating Single-Cell Multi-omics Datasets

Integrating data from different single-cell sequencing experiments is an important biological task for which OT has proven useful (Cao et al., 2021, 2022; Demetci et al., 2020). Single-cell experiments measure various genomic features at the individual cell resolution. Jointly studying these can give scientists insight into the mechanisms regulating cells. However, experimentally combining multiple measurement types for the same cell is challenging. For a limited combination of measurement types (or "measurement modalities") , there are experimental protocols (termed "co-assays") that jointly profile them on the same cells. For others, scientists rely on the computational integration of multimodal data taken on different but related cells (*e.g.*, by cell type or tissue) to study the relationships and interactions between different aspects of the genome.

We particularly focus on this task for two reasons. First, GW distance was previously used as a state-of-the-art method for this task (Cao et al., 2021; Demetci et al., 2022a; Cao et al., 2022), so it is important to see if AGW improves upon it. Second, several single-cell benchmark datasets provide ground-truth matchings on the feature- and the sample-level alignments. This information allows us to assess the effect of guiding cross-domain matching with partial or full prior knowledge of these relationships.

Table 1: Single-Cell Alignment Error, as Quantified by the Average 'Fraction of Samples Closer Than True Match' (FOSCTTM) Metric. Lower values are better. For each dataset, the size of the two domains they contain are expression in the format (number of samples x number of features) in the second row. Note that for GW, we use the SCOT implementation by Demetci et al. (2020) and similarly for UGW, we use the SCOTv2 implementation by Demetci et al. (2022a).

	Simulation 1 (300x1000, 300x2000)	**Simulation 2** (300x1000, 300x2000)	**Simulation 3** (300x1000, 300x2000)	**Simulated RNA-seq** (5000x50, 5000x500)	**scGEM** (177x28, 177x34)	**SNARE-seq** (1047x1000, 1047x3000)	**CITE-seq** (1000x25, 1000x24)
AGW	**0.0730**	**0.0041**	**0.0082**	**0.0**	0.183	**0.132**	**0.091**
GW	0.0866	0.0216	0.0084	7.1e-5	0.198	0.150	0.121
COOT	0.0752	**0.0041**	0.0088	**0.0**	0.206	0.153	0.132
UGW	0.0838	0.0522	0.0105	0.096	**0.175**	0.160	0.116
UCOOT	0.0850	0.0081	0.0122	0.115	0.181	0.188	0.127
bindSC	N/A	N/A	N/A	3.8e-4	N/A	0.242	0.144

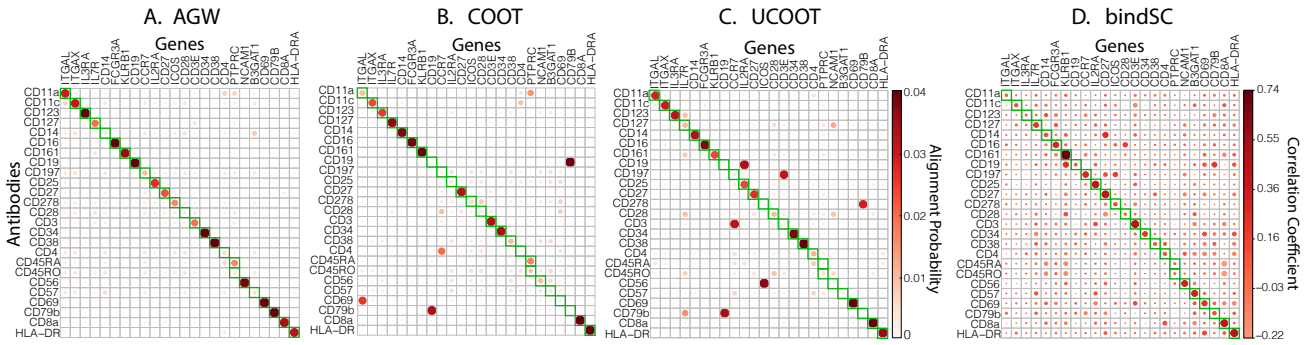

Figure 2: Feature Alignments for the CITE-Seq Dataset. Green boxes indicate where we expect matches (a notion of "ground-truth") based on domain knowledge.

4.1.1 Single-Cell Alignment

We follow the first GW application in this domain (Demetci et al., 2020, 2022b) and align samples (*i.e.*, cells) of simulated and real-world datasets from different measurement types. Note that this work adapted the GW optimal transport framework for single-cell data integration tasks with their implementation (named SCOT), which uses shortest path distances on nearest neighbor graphs constructed with low dimensional embeddings of genomic data. SCOT shows performance improvement over classic choices of intradomain distance metrics, such as Euclidean distances or correlations (Demetci et al., 2022b). Therefore, only in the single-cell multi-omic integration tasks, we use the SCOT implementation when benchmarking GW optimal transport, and the subsequent SCOTv2 implementation Demetci et al. (2022a) when benchmarking UGW optimal transport.

Since we use simulations and real-world co-assayed datasets (datasets where mutliple measurements are jointly profiled on the same set of cells), we have ground-truth information on cell-cell alignments for all datasets. However, we perform unsupervised alignment, and only use this information for benchmarking alignments and hyperparameter tuning. We demonstrate in Table 1 that AGW consistently yields higher quality cell alignments (with lower alignment error) compared to the state-of-the-art baselines, including GW, COOT, and their unbalanced counterparts.

We include bindSC as an additional baseline, which performs bi-order canonical correlation analysis to align single-cell datasets. Unlike other single-cell alignment methods, it internally computes a feature correlation matrix that users can extract. So, we include it as a baseline to compare its feature alignment performance against AGW in the next section. However, bindSC usage is limited to a few measurement types as it requires an input matrix that relates features across domains to bring the datasets into the same space at initialization. We do not have this information for most datasets, thus the "N/A" entries in Table 1.

4.1.2 Aligning Genomic Features

AGW augments GW formulation with a feature coupling matrix. Therefore, we jointly align features and see whether AGW reveals relevant biological relation-

ships. All current single-cell alignment methods only align samples (*i.e.*, cells), except for bindSC as discussed above. We emphasize that the main goal of feature alignment for our single-cell multi-omic integration applications is to leverage priors to yield higher quality sample alignments, as demonstrated later. Cross-modal genomic relationships are complex and our algorithm has not been extensively benchmarked on recovering such relationships. Nevertheless, improving upon feature matching capabilities of approaches such as COOT and bindSC will also help improve sample alignment quality. These experiments also provide a straightforward way to evaluate whether feature couplings can generate meaningful hypotheses.

Among the real-world datasets in Table 1, CITE-seq (Stoeckius et al., 2017) is the only one with ground-truth information on feature correspondences. This dataset has paired single-cell measurements on the abundance levels of 25 antibodies and activity (*i.e.*, "expression") levels of genes, including the genes that encode these 25 antibodies. So, we first present unsupervised feature alignment results on the CITE-seq dataset. For completion, we also report the biological relevance of our feature alignments on SNARE-seq (Chen et al., 2019) and scGEM (Cheow et al., 2016b) datasets in Appendix C. However, note that these datasets (unlike CITE-seq) do not have clear ground-truth feature correspondences. We compare our feature alignments with bindSC, COOT, and UCOOT in Figure 2. Note that due to the size of the CITE-seq dataset ($\sim 60,000$ human and mouse cells in total), we first subsample it by randomly selecting 1000 human cells. Then, we transform the data by computing the `log2`-fold change compared to dataset median. We find that this improves the feature alignment performance of all methods. The entries in Figure 2 matrices are arranged such that the "ground-truth" correspondences lie in the diagonal, marked by green squares. While AGW correctly assigns 19 out of 25 antibodies to their encoding genes with the highest alignment probability, this number is 16 for UCOOT, 15 for COOT and 13 for bindSC (which yields correlation coefficients instead of alignment probabilities). Additionally, the OT methods yield more sparse alignments than bindSC thanks to the "least effort" requirement in their formulation.

4.1.3 Leveraging Prior Knowledge

Finally, we show the advantage of providing priors by aligning a multi-species gene expression dataset containing measurements from the adult mouse prefrontal cortex (Bhattacherjee et al., 2019) and pallium of bearded lizard (Tosches et al., 2018). Since measurements come from two different species, the feature space (*i.e.*, genes) differs, and there is no 1-1 correspondence between the samples (*i.e.*, cells). However, there is a shared subset within the features, *i.e.*, orthologous genes that descend from a common ancestor and maintain similar biological functions in both species. We also have domain knowledge on cells that belong to similar cell types across the two species. Thus, we expect AGW to recover these relationships.

Figure 3A visualizes the cell-type alignment probabilities yielded by AGW when full supervision is provided on the $10,816$ orthologous genes. The green boxes indicate alignment between similar types of cells. This matrix is obtained by averaging the sample alignment matrix (*i.e.*, cell-cell alignments) into cell-type groups. We observe that AGW yields biologically plausible alignments, as all the six cell types that have a natural match across the two species are correctly matched. We also show in Figure 3B that providing supervision on one alignment level (*e.g.*, features) improves the quality on the other alignment level (*e.g.*, samples). The supervision scheme is detailed in Appendix F.2.

4.2 Heterogeneous Domain Adaptation

Finally, we demonstrate the generalizability of our approach on a popular ML task, heterogeneous domain adaptation, where COOT and GW distance were previously successfully used. Domain adaptation (DA) refers to the problem in which a classifier learned on one domain (called *source*) can generalize to the other (called *target*). Here, we apply AGW to unsupervised and semi-supervised heterogeneous DA (HDA) tasks, where the source and target samples live in different spaces, and we have as few as zero labeled target samples.

We mostly follow the experimental setup by Redko et al. (2020) and use source-target pairs from the Caltech-Office dataset (Saenko et al., 2010). We consider all possible pairs between three domains: Amazon (A), Caltech-256 (C), and Webcam (W), whose image embeddings are obtained from the penultimate layers in the GoogleNet (Szegedy et al., 2015) (vectors in $\mathbb{R}^{4096}$) and CaffeNet (Jia et al., 2014) (vectors in $\mathbb{R}^{1024}$) neural network architectures. We randomly choose 20 samples per class and perform adaptation from CaffeNet to GoogleNet and repeat it 10 times. Differently than Redko et al. (2020), we (1) unit normalize the dataset prior to alignment as we empirically found it to boost all methods' average performance compared to using unnormalized datasets, (2) use cosine distances when defining intra-domain distance matrices for GW and AGW, as we found them to perform better than Euclidean distances, and (3) report results after hyperparameter tuning all methods for each pair of datasets based on a validation set (see Section F.3).

As in Redko et al. (2020), for semi-supervised settings,

Table 2: Heterogeneous Domain Adaptation Results (Unsupervised). Best results are bolded, and second-bests are underlined. For AGW, the α values used are respectively $0.6, 0.9, 0.7, 0.9, 0.3, 0.8, 0.7, 0.2, 0.6$.

	A → A	**A → C**	**A → W**	**C → A**	**C → C**	**C → W**	**W → A**	**W → C**	**W → W**
AGW	**93.1±1.6**	**68.3±14.1**	**79.8±3.5**	55.4±7.1	76.4±5.6	**57.7±14.3**	60.1±9.1	**60.9±13.3**	**97.3±0.9**
GW	86.2±2.3	64.1±6.2	77.6±11.1	53.0±13.2	**81.9±10.5**	53.5±15.9	50.4±22.1	54.3±14.7	92.5±2.6
COOT	50.3±15.9	35.0±6.4	39.8±14.5	40.8±15.8	33.5±10.7	37.5±10.4	44.3±14.0	27.4±10.2	57.9±13.4
UGW	90.6±6.5	67.2±12.7	75.4±3.1	**56.3±14.6**	69.2±8.7	51.2±13.1	**66.7±9.9**	58.4±4.7	94.7±1.5
UCOOT	65.4±2.1	44.6±3.8	36.4±1.2	55.1±8.6	52.1±3.8	41.8±14.9	63.2±4.0	59.7±6.3	80.3±2.1

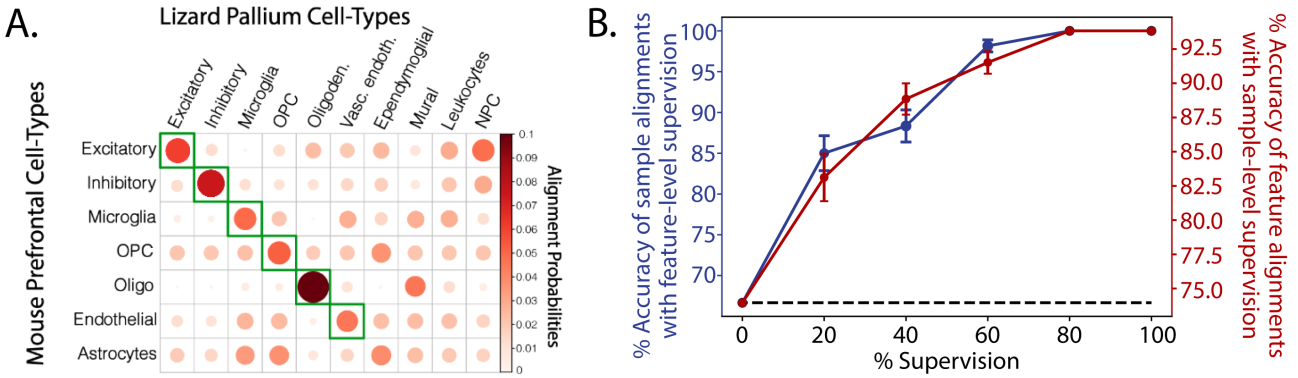

Figure 3: **Aligning cross-species dataset.** A. AGW's cell-type alignments. B. Providing supervision on one level of alignment (*e.g.*, features) boosts alignments on the other. Standard errors computed over 10 random runs. Dashed line indicates the sample alignment performance of GW and bindSC (orthologous gene used in input).

we incorporate prior knowledge on a few target labels by adding an extra cost matrix to the training of sample coupling, so that a source sample will be penalized if it transfers mass to the target samples from different classes . Once the sample coupling $\boldsymbol{\gamma}^s$ is learned, we obtain the final prediction using label propagation: $\widehat{y}_t = \arg\max_k L_{k\cdot}$, where $L = D_s\boldsymbol{\gamma}^s$ and D_s denotes one-hot encodings of the source labels y_s. .

Table 2 presents the performance of each method averaged across ten runs in the unsupervised setting, where AGW yields favorable results in 6 out of 9 cases. In two cases, UGW, and in one case, UCOOT, outperform AGW despite the lower performance of their balanced counterparts. In these cases, unbalanced formulations prove beneficial, and support extending AGW to unbalanced scenarios as future work. Appendix D presents the semi-supervised experiments, which show the same trend where AGW tends to outperform other baselines.

4.3 Empirical Runtime

As described in Section 3.2, the theoretical complexity of AGW is $O(n^3 + dn^2 + nd^2)$. When $d < n$, the dominating term of n^3 is due to the computational burden of computing the GW distance. However, in practice, we observe that AGW converges in much fewer iterations than GW distance (about 1/5 of the number of iterations on average) thanks to the further refinement of the sample coupling per iteration as facilitated by the interpolation with COOT, thus having a shorter runtime (see Appendix E). To further speed up optimization, one can consider low-rank coupling and cost matrix (Scetbon et al., 2022) or use the divide and conquer strategy (Chowdhury et al., 2021), which allows one to scale the GW distance up to a million points.

5 CONCLUSION AND DISCUSSION

We present Augmented Gromov-Wasserstein (AGW), a new OT-based divergence for incomparable spaces. It interpolates between GW distance and CO-Optimal transport and allows to narrow down the choices of isometries induced by GW distance, while efficiently exploiting the prior knowledge on the input data. We study its basic properties and empirically show that such restrictions result in better performance for single-cell multi-omic alignment tasks and transfer learning. Future work will focus on refining the theoretical analysis of the AGW invariants to better understand their performance in practice. We will also extend AGW to the unbalanced and/or continuous setting, and other tasks where feature supervision by domain experts may be incorporated in OT framework.

Acknowledgements

We thank Professor Will Sawin from Princeton University, Department of Mathematics for helpful discussions on Corollary 1. Pinar Demetci's and Ritambhara Singh's contributions were funded by the NIH award 1R35HG011939-01. Pinar Demetci's contributions were also supported in part by funding from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard. Quang Huy Tran's contribution was funded by the projects OTTOPIA ANR-20-CHIA-0030, the 3rd Programme d'Investissements d'Avenir ANR-18-EUR-0006-02, the Chair "Challenging Technology for Responsible Energy" led by l'X – Ecole Polytechnique and the Chair "Business Analytics for Future Banking" sponsored by NATIXIS.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

Aritra Bhattacherjee, Mohamed Nadhir Djekidel, Renchao Chen, Wenqiang Chen, Luis M. Tuesta, and Yi Zhang. Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. *Nature Communications*, 10(1): 4169, Sep 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-12054-3. URL https://doi.org/10.1038/s41467-019-12054-3.

Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning Generative Models across Incomparable Spaces. In *ICML*, pages 851–861, 2019.

Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications*, 11(1): 1–13, 2020.

Kai Cao, Xiangqi Bai, Yiguang Hong, and Lin Wan. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, 36 (Supplement_1):i48–i56, 2020.

Kai Cao, Yiguang Hong, and Lin Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics*, 08 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab594. URL https://doi.org/10.1093/bioinformatics/btab594. btab594.

Kai Cao, Qiyu Gong, Yiguang Hong, and Lin Wan. A unified computational framework for single-cell data integration with optimal transport. *Nature Communications*, 13(1):7419, 2022.

Song Chen, Blue B. Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457, 2019. doi: 10.1038/s41587-019-0290-0. URL https://doi.org/10.1038/s41587-019-0290-0.

Lih Feng Cheow, Elise T Courtois, Yuliana Tan, Ramya Viswanathan, Qiaorui Xing, Rui Zhen Tan, Daniel S Q Tan, Paul Robson, Loh Yuin-Han, Stephen R Quake, and William F Burkholder. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature Methods*, 13(10):833–836, 2016a.

Lih Feng Cheow, Elise T Courtois, Yuliana Tan, Ramya Viswanathan, Qiaorui Xing, Rui Zhen Tan, Daniel S Q Tan, Paul Robson, Loh Yuin-Han, Stephen R Quake, and William F Burkholder. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature Methods*, 13(10):833–836, 2016b.

Samir Chowdhury, David Miller, and Tom Needham. Quantized Gromov-Wasserstein. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 811–827, 2021.

Keith Conrad. Isometries of $\mathbb{R}^n$. https://kconrad.math.uconn.edu/blurbs/grouptheory/isometryRn.pdf, 2019. Accessed: 2023-05-01.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pages 2292–2300, 2013.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-Wasserstein optimal transport to align single-cell multi-omics data. *bioRxiv*, 2020.

Pinar Demetci, Rebecca Santorella, Manav Chakravarthy, Bjorn Sandstede, and Ritambhara Singh. Scotv2: Single-cell multiomic alignment with disproportionate cell-type representation. *Journal of Computational Biology*, 29 (11):1213–1228, 2022a. doi: 10.1089/cmb.2022.0270. URL https://doi.org/10.1089/cmb.2022.0270. PMID: 36251763.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022b. doi: 10.1089/cmb.2021.0446. URL https://doi.org/10.1089/cmb.2021.0446. PMID: 35050714.

Jinzhuang Dou, Shaoheng Liang, Vakul Mohanty, Qi Miao, Yuefan Huang, Qingnan Liang, Xuesen Cheng, Sangbae Kim, Jongsu Choi, Yumei Li, Li Li, May Daher, Rafet Basar, Katayoun Rezvani, Rui Chen, and Ken Chen. Bi-order multimodal integration of single-cell data. *Genome Biology*, 23(1): 112, 2022. doi: 10.1186/s13059-022-02679-x. URL https://doi.org/10.1186/s13059-022-02679-x.

Li Fang, Yunjin Li, Lu Ma, Qiyue Xu, Fei Tan, and Geng Chen. GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic Acids Research*, 49(D1):D97–D103, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa995. URL https://doi.org/10.1093/nar/gkaa995.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Xiang Gu, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu. Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. In *Advances in Neural Information Processing Systems*, volume 35, pages 14972–14985, 2022.

Fayrouz Hammal, Pierre de Langen, Aurélie Bergon, Fabrice Lopez, and Benoit Ballester. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Research*, 50(D1):D316–D325, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab996. URL https://doi.org/10.1093/nar/gkab996.

J.J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678, 2014.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

April R Kriebel and Joshua D Welch. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nature communications*, 13(1):780, 2022.

Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly Embedding Multiple Single-Cell Omics Measurements. In *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, volume 143, pages 10:1–10:13, 2019.

Facundo Memoli. Gromov Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, pages 1–71, 2011.

Nao Nakagawa, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Gromov-Wasserstein Autoencoders. *arXiv preprint arXiv:2209.07007*, 2022.

Jairo Navarro Gonzalez, Ann S Zweig, Matthew L Speir, Daniel Schmelter, Kate R Rosenbloom, Brian J Raney, Conner C Powell, Luis R Nassar, Nathan D Maulding, Christopher M Lee, Brian T Lee, Angie S Hinrichs, Alastair C Fyfe, Jason D Fernandes, Mark Diekhans, Hiram Clawson, Jonathan Casper, Anna Benet-Pagès, Galt P Barber, David Haussler, Robert M Kuhn, Maximilian Haeussler, and W James Kent. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research*, 49(D1):D1046–D1057, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1070. URL https://doi.org/10.1093/nar/gkaa1070.

Mor Nitzan, Nikos Karaiskos, Nir Friedman, and Nikolaus Rajewsky. Gene expression cartography. *Nature*, 576(7785):132–137, Dec 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1773-3. URL https://doi.org/10.1038/s41586-019-1773-3.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *ICML*, pages 2664–2672, 2016.

Ievgen Redko, Titouan Vayer, Rémi Flamary, and Nicolas Courty. CO-Optimal Transport. *34th Conference on Neural Information Processing Systems*, 2020.

K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, LNCS, pages 213–226, 2010.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-Time Gromov Wasserstein Distances using Low Rank Couplings and Costs . In *International Conference on Machine Learning*, 2022.

Thibault Sejourne, Francois-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8766–8779. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4990974d150d0de5e6e15a1454fe6b0f-Paper.pdf.

Ritambhara Singh, Pinar Demetci, Giancarlo Bonora, Vijay Ramani, Choli Lee, He Fang, Zhijun Duan, Xinxian Deng, Jay Shendure, Christine Disteche, and William Stafford Noble. Unsupervised manifold alignment for single-cell multi-omics data. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379649. doi: 10.1145/3388440.3412410. URL https://doi.org/10.1145/3388440.3412410.

Elias M. Stein and Rami Shakarchi. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton Lectures in Analysis. Princeton University Press, 2005.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14 (9):865–868, 2017. doi: 10.1038/nmeth.4380. URL https://doi.org/10.1038/nmeth.4380.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

Terence Tao. *Topics in random matrix theory*. Graduate Studies in Mathematics. American Mathematical Society, 2012.

Maria Antonietta Tosches, Tracy M. Yamawaki, Robert K. Naumann, Ariel A. Jacobi, Georgi Tushev, and Gilles Laurent. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science*, 360(6391):881–888, 2018. doi: 10.1126/science.aar4237. URL https://www.science.org/doi/abs/10.1126/science.aar4237.

Quang Huy Tran, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci, and Ritambhara Singh. Unbalanced co-optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):10006–10016, Jun. 2023. doi: 10.1609/aaai.v37i8.26193. URL https://ojs.aaai.org/index.php/AAAI/article/view/26193.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal Transport for structured data with application on graphs. In *ICML*, pages 6275–6284, 2019.

Titouan Vayer, Laetitia Chapel, Remi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9), 2020. ISSN 1999-4893. doi: 10.3390/a13090212. URL https://www.mdpi.com/1999-4893/13/9/212.

Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, 2009 edition, September 2008.

Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2969–2975, 2018.

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

A Proofs

A.1 Proposition 1

Proof. The proof of this proposition can be adapted directly from (Vayer et al., 2019). For self-contained purpose, we give the proof here. Denote

- $(\boldsymbol{\gamma}_\alpha^s, \boldsymbol{\gamma}_\alpha^v)$ the optimal sample and feature couplings for $\text{AGW}_\alpha(\mathbf{X}, \mathbf{Y})$.
- $(\boldsymbol{\gamma}_0^s, \boldsymbol{\gamma}_0^v)$ the optimal sample and feature couplings for $\text{COOT}(\mathbf{X}, \mathbf{Y})$.
- $\boldsymbol{\gamma}_1^s$ the optimal sample coupling for $\text{GW}(\mathbf{X}, \mathbf{Y})$.

Due to the suboptimality of $\boldsymbol{\gamma}_\alpha^s$ for GW and $(\boldsymbol{\gamma}_1^s, \boldsymbol{\gamma}_0^v)$ for AGW, we have

$$\alpha\langle L(\mathbf{D}_X, \mathbf{D}_Y) \otimes \boldsymbol{\gamma}_1^s, \boldsymbol{\gamma}_1^s\rangle \leq \alpha\langle L(\mathbf{D}_X, \mathbf{D}_Y) \otimes \boldsymbol{\gamma}_\alpha^s, \boldsymbol{\gamma}_\alpha^s\rangle + (1-\alpha)\langle L(\mathbf{X}, \mathbf{Y}) \otimes \boldsymbol{\gamma}_\alpha^v, \boldsymbol{\gamma}_\alpha^s\rangle \tag{3}$$

$$\leq \alpha\langle L(\mathbf{D}_X, \mathbf{D}_Y) \otimes \boldsymbol{\gamma}_1^s, \boldsymbol{\gamma}_1^s\rangle + (1-\alpha)\langle L(\mathbf{X}, \mathbf{Y}) \otimes \boldsymbol{\gamma}_0^v, \boldsymbol{\gamma}_1^s\rangle, \tag{4}$$

or equivalently

$$\alpha\text{GW}(\mathbf{X}, \mathbf{Y}) \leq \text{AGW}_\alpha(\mathbf{X}, \mathbf{Y}) \leq \alpha\text{GW}(\mathbf{X}, \mathbf{Y}) + (1-\alpha)\langle L(\mathbf{X}, \mathbf{Y}) \otimes \boldsymbol{\gamma}_0^v, \boldsymbol{\gamma}_1^s\rangle. \tag{5}$$

Similarly, we have

$$(1-\alpha)\text{COOT}(\mathbf{X}, \mathbf{Y}) \leq \text{AGW}_\alpha(\mathbf{X}, \mathbf{Y}) \leq (1-\alpha)\text{COOT}(\mathbf{X}, \mathbf{Y}) + \alpha\langle L(\mathbf{D}_X, \mathbf{D}_Y) \otimes \boldsymbol{\gamma}_0^s, \boldsymbol{\gamma}_0^s\rangle. \tag{6}$$

The interpolation property then follows by the sandwich theorem.

Regarding the relaxed triangle inequality, given three triplets $(\mathbf{X}, \mu_{sx}, \mu_{fx}), (\mathbf{Y}, \mu_{sy}, \mu_{fy})$ and $(\mathbf{Z}, \mu_{sz}, \mu_{fz})$, let $(\pi^{XY}, \gamma^{XY}), (\pi^{YZ}, \gamma^{YZ})$ and (π^{XZ}, γ^{XZ}) be solutions of the problems $\text{AGW}_\alpha(\mathbf{X}, \mathbf{Y}), \text{AGW}_\alpha(\mathbf{Y}, \mathbf{Z})$ and $\text{AGW}_\alpha(\mathbf{X}, \mathbf{Z})$, respectively. Denote $P = \pi^{XY}\text{diag}\left(\frac{1}{\mu_{sy}}\right)\pi^{YZ}$ and $Q = \gamma^{XY}\text{diag}\left(\frac{1}{\mu_{fy}}\right)\gamma^{YZ}$. Then, it is not difficult to see that $P \in \Pi(\mu_{sx}, \mu_{sz})$ and $Q \in \Pi(\mu_{fx}, \mu_{fz})$. The suboptimality of (P, Q) implies that

$$\frac{\text{AGW}_\alpha(\mathbf{X}, \mathbf{Z})}{2} \tag{7}$$

$$\leq \alpha \sum_{i,j,k,l} \frac{|\mathbf{D}_X(i,j) - \mathbf{D}_Z(k,l)|^2}{2} P_{i,k}P_{j,l} + (1-\alpha)\sum_{i,j,k,l} \frac{|\mathbf{X}_{i,j} - \mathbf{Z}_{k,l}|^2}{2} P_{i,k}Q_{j,l} \tag{8}$$

$$= \alpha \sum_{i,j,k,l} \frac{|\mathbf{D}_X(i,j) - \mathbf{D}_Z(k,l)|^2}{2} \left(\sum_e \frac{\pi_{i,e}^{XY}\pi_{e,k}^{YZ}}{(\mu_{sy})_e}\right)\left(\sum_o \frac{\pi_{j,o}^{XY}\pi_{o,l}^{YZ}}{(\mu_{sy})_o}\right) \tag{9}$$

$$+ (1-\alpha) \sum_{i,j,k,l} \frac{|\mathbf{X}_{i,j} - \mathbf{Z}_{k,l}|^2}{2} \left(\sum_e \frac{\pi_{i,e}^{XY}\pi_{e,k}^{YZ}}{(\mu_{sy})_e}\right)\left(\sum_o \frac{\gamma_{j,o}^{XY}\gamma_{o,l}^{YZ}}{(\mu_{fy})_o}\right) \tag{10}$$

$$\leq \alpha \sum_{i,j,k,l,e,o} |\mathbf{D}_X(i,j) - \mathbf{D}_Y(e,o)|^2 \frac{\pi_{i,e}^{XY}\pi_{e,k}^{YZ}}{(\mu_{sy})_e}\frac{\pi_{j,o}^{XY}\pi_{o,l}^{YZ}}{(\mu_{sy})_o} + (1-\alpha)\sum_{i,j,k,l,e,o} |\mathbf{X}_{i,j} - \mathbf{Y}_{e,o}|^2 \frac{\pi_{i,e}^{XY}\pi_{e,k}^{YZ}}{(\mu_{sy})_e}\frac{\gamma_{j,o}^{XY}\gamma_{o,l}^{YZ}}{(\mu_{fy})_o} \tag{11}$$

$$+ \alpha \sum_{i,j,k,l,e,o} |\mathbf{D}_Y(e,o) - \mathbf{D}_Z(k,l)|^2 \frac{\pi_{i,e}^{XY}\pi_{e,k}^{YZ}}{(\mu_{sy})_e}\frac{\pi_{j,o}^{XY}\pi_{o,l}^{YZ}}{(\mu_{sy})_o} + (1-\alpha)\sum_{i,j,k,l,e,o} |\mathbf{Y}_{e,o} - \mathbf{Z}_{k,l}|^2 \frac{\pi_{i,e}^{XY}\pi_{e,k}^{YZ}}{(\mu_{sy})_e}\frac{\gamma_{j,o}^{XY}\gamma_{o,l}^{YZ}}{(\mu_{fy})_o} \tag{12}$$

$$= \alpha \sum_{i,j,e,o} |\mathbf{D}_X(i,j) - \mathbf{D}_Y(e,o)|^2 \pi_{i,e}^{XY}\pi_{j,o}^{XY} + (1-\alpha)\sum_{i,j,e,o} |\mathbf{X}_{i,j} - \mathbf{Y}_{e,o}|^2 \pi_{i,e}^{XY}\gamma_{j,o}^{XY} \tag{13}$$

$$+ \alpha \sum_{k,l,e,o} |\mathbf{D}_Y(e,o) - \mathbf{D}_Z(k,l)|^2 \pi_{e,k}^{YZ}\pi_{o,l}^{YZ} + (1-\alpha)\sum_{k,l,e,o} |\mathbf{Y}_{e,o} - \mathbf{Z}_{k,l}|^2 \pi_{e,k}^{YZ}\gamma_{o,l}^{YZ} \tag{14}$$

$$= \text{AGW}_\alpha(\mathbf{X}, \mathbf{Y}) + \text{AGW}_\alpha(\mathbf{Y}, \mathbf{Z}). \tag{15}$$

where the second inequality follows from the inequality: $(x+y)^2 \leq 2(x^2+y^2)$. ■

A.2 Corollary 1

Proof. First, let us recall the Schwartz-Zippel lemma. Denote $F(x_1, ..., x_n)$ a multivariate polynomial. Its total degree is the maximum of the sums of the powers of the variables in any monomial. The Schwartz-Zippel lemma states that: let $F(x_1, ..., x_n)$ be a nonzero multivariate polynomial of total degree d and S be a finite subset of $\mathbb{R}$. Denote $Z_S := \{(x_1, ..., x_n) \in S^n : F(x_1, ..., x_n) = 0\}$ the zero set of F on S^n. Then $|Z_S| \leq d|S|^{n-1}$.

Note that, the set of Hermitian matrices of size n forms a finite-dimensional real vector space. In particular, it is isomorphic to the Euclidean space $\mathbb{R}^{n^2}$. Denote I set of Hermitian matrices of size n with repeated eigenvalues. It is enough to show that I has measure zero. We have $I \simeq E$, for some $E \subset \mathbb{R}^{n^2}$. Since I is closed (see page 56 in (Tao, 2012)), it is measurable, by Proposition 4 in (Stein and Shakarchi, 2005). If I does not have zero measure, then the intersection $E \cap [0, 1]^{n^2}$ has positive measure $p > 0$. If, for each $i \in [n^2]$, we sample m i.i.d coordinates uniformly in $[0, 1]$, then we have m^{n^2} points uniformly distributed in $[0, 1]^{n^2}$. So, the expected number of points lying in E is pm^{n^2}.

On the other hand, recall that a (Hermitian) matrix has repeated eigenvalues if and only if the discriminant of its characteristic polynomial is zero. Moreover, the discriminant of the characteristic polynomial is a polynomial in n^2 entries of the matrix. Thus, the measure of I (or, equivalently E) is the measure of the set of values of these n^2 variables which make a certain polynomial of total degree d vanish. By Schwartz-Zippel lemma, on average, there are at most dm^{n^2-1} points in E. By choosing $m > d/p$, we obtain a contradiction. Thus E (or equivalently I) must have zero measure. ■

A.3 Theorem 1

Proof. Regarding the first claim, note that $\mathbf{Y} = \mathbf{X}Q$, where Q is a permutation matrix corresponding to the permutation σ_c. Since $\mathbf{Y}$ is obtained by swapping columns of $\mathbf{X}$, it is easy to see that $\mathrm{GW}(\mathbf{X}, \mathbf{Y}) = 0$ and the optimal plan between $\mathbf{X}$ and $\mathbf{Y}$ is $\boldsymbol{\gamma}^s = \frac{1}{n^2}\mathrm{Id}_n$. Similarly, $\mathrm{COOT}(\mathbf{X}, \mathbf{Y}) = 0$ and $\boldsymbol{\gamma}^s, \boldsymbol{\gamma}^v = \frac{1}{n}Q$ are the optimal sample, feature couplings, respectively. In other words, $\langle L(\mathbf{D}_X, \mathbf{D}_Y) \otimes \boldsymbol{\gamma}^s, \boldsymbol{\gamma}^s \rangle = 0$ and $\langle L(\mathbf{X}, \mathbf{Y}) \otimes \boldsymbol{\gamma}^v, \boldsymbol{\gamma}^s \rangle = 0$. We deduce that $\mathrm{AGW}_\alpha(\mathbf{X}, \mathbf{Y}) = 0$.

Now, for $0 < \alpha < 1$, if $\mathrm{AGW}(\mathbf{X}, \mathbf{Y}) = 0$, then $\mathrm{GW}(\mathbf{X}, \mathbf{Y}) = \mathrm{COOT}(\mathbf{X}, \mathbf{Y}) = 0$. In particular, $\mathbf{X}$ and $\mathbf{Y}$ must have the same shape, so $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$. As $\mathrm{GW}(\mathbf{X}, \mathbf{Y}) = 0$, there exists an isometry from $\mathbf{X}$ to $\mathbf{Y}$. Note that every isometry from $\mathbb{R}^d$ to $\mathbb{R}^d$ is a composition of at most $d + 1$ reflections (see, for example, Corollary A.7 in (Conrad, 2019)). So, $\mathbf{Y} = \mathbf{X}O$, for some $O \in \mathcal{O}_d$. As $\mathrm{COOT}(\mathbf{X}, \mathbf{Y}) = 0$, there exist two permutations σ_r and σ_c such that $\mathbf{X}_{i,j} = \mathbf{Y}_{\sigma_r(i),\sigma_c(j)}$, or equivalently two permutation matrices $P \in \mathcal{P}_n, Q_1 \in \mathcal{P}_d$ such that $\mathbf{Y} = P\mathbf{X}Q_1$. We deduce that $\mathbf{X}O = P\mathbf{X}Q_1$, or equivalently $\mathbf{X} = P\mathbf{X}Q$, for $Q = Q_1O^T \in \mathcal{O}_d$. We will show that Q is symmetric.

Indeed, consider the singular value decomposition of $\mathbf{X}$, *i.e.* $\mathbf{X} = U\Sigma V^T$, where $U \in \mathbb{R}^{n \times d}$ such that $U^TU = I_d$, $V \in \mathcal{O}_d$ and $\Sigma \in \mathbb{R}^{d \times d}$ is a diagonal matrix whose diagonal contains d strictly decreasing singular values (since $n \geq d$). As $\mathbf{X} = P\mathbf{X}Q$, we have $U\Sigma V^T = (PU)\Sigma(V^TQ)$. For $i \in [d]$, let $u_i \in \mathbb{R}^n$ and $v_i \in \mathbb{R}^d$ be columns of U and V, respectively. As the singular values are positive and distinct, the columns are unique up to the sign change of both columns in U and V. This means $u_i = \pm Pu_i$ and $v_i = \pm Q^Tv_i$. In other words, ± 1 are eigenvalues of P and Q^T, and u_i, v_i are their corresponding eigenvectors, respectively. Denote $D \in \mathbb{R}^{d \times d}$ any diagonal matrix whose diagonal values are in $\{\pm 1\}$, then $Q^T = VDV^{-1} = VDV^T = Q$. So, Q is symmetric. Theorem 1 then follows by observing that $O = Q^TQ_1$. ■

A.4 Weak invariance to translation

While enjoying the interpolation and metric properties, AGW does not inherit the invariance to the translation of the GW distance. However, we find that it satisfies a relaxed version of this invariant defined as follows.

Definition 1. *We call* $D = \inf_{\pi \in \Pi} F(\pi, \mathbf{X}, \mathbf{Y})$*, where* $\mathbf{X}, \mathbf{Y}$ *are input data and* Π *is a set of feasible couplings and* F *is a real-valued functional, an OT-based divergence. Then* D *is weakly invariant to translation if for every* $a, b \in \mathbb{R}$*, we have* $\inf_{\pi \in \Pi} F(\pi, \mathbf{X}, \mathbf{Y}) = C + \inf_{\pi \in \Pi} F(\pi, \mathbf{X} + a, \mathbf{Y} + b)$*, for some constant* C *depending on* $a, b, \mathbf{X}, \mathbf{Y}$ *and* Π*.*

Here, we denote the translation of $\mathbf{X}$ as $\mathbf{X} + a$, whose elements are of the form $\mathbf{X}_{ij} + a$. Intuitively, an OT-based divergence is weakly invariant to translation if only the optimal transport plan is preserved under translation, but not necessarily the divergence itself. In practice, we would argue that the ability to preserve the optimal plan

under translation is much more important than preserving the distance itself. In other words, the translation only shifts the minimum but has no impact on the optimization procedure, meaning that the minimizer remains unchanged. Now, we can show that

Corollary 2. *COOT is weakly invariant to translation.*

Proof of Corollary 2. Given $\boldsymbol{\gamma}^s \in \Pi(\mu, \nu), \boldsymbol{\gamma}^v \in \Pi(\mu', \nu')$, for any $c \in \mathbb{R}$, we have

$$\sum_{ijkl}(\mathbf{X}_{ik} - \mathbf{Y}_{jl} - c)^2 \boldsymbol{\gamma}^s_{ij}\boldsymbol{\gamma}^v_{kl} = \sum_{ijkl}(\mathbf{X}_{ik} - \mathbf{Y}_{jl})^2 \boldsymbol{\gamma}^s_{ij}\boldsymbol{\gamma}^v_{kl} - 2c\sum_{ijkl}(\mathbf{X}_{ik} - \mathbf{Y}_{jl})\boldsymbol{\gamma}^s_{ij}\boldsymbol{\gamma}^v_{kl} + c^2 \tag{16}$$

Now,

$$\begin{aligned}
\sum_{ijkl}(\mathbf{X}_{ik} - \mathbf{Y}_{jl})\boldsymbol{\gamma}^s_{ij}\boldsymbol{\gamma}^v_{kl} &= \sum_{ijkl}\mathbf{X}_{ik}\boldsymbol{\gamma}^s_{ij}\boldsymbol{\gamma}^v_{kl} - \sum_{ijkl}\mathbf{Y}_{jl}\boldsymbol{\gamma}^s_{ij}\boldsymbol{\gamma}^v_{kl} && (17)\\
&= \sum_{ik}\mathbf{X}_{ik}\left(\sum_j \boldsymbol{\gamma}^s_{ij}\right)\left(\sum_l \boldsymbol{\gamma}^v_{kl}\right) - \sum_{jl}\mathbf{Y}_{jl}\left(\sum_i \boldsymbol{\gamma}^s_{ij}\right)\left(\sum_k \boldsymbol{\gamma}^v_{kl}\right) && (18)\\
&= \sum_{ik}\mathbf{X}_{ik}\mu_i\mu'_k - \sum_{jl}\mathbf{Y}_{jl}\nu_j\nu'_l && (19)\\
&= \mu^T\mathbf{X}\mu' - \nu^T\mathbf{Y}\nu'. && (20)
\end{aligned}$$

So, $\mathrm{COOT}(\mathbf{X}, \mathbf{Y} + c) = \mathrm{COOT}(\mathbf{X}, \mathbf{Y}) - 2c\left(\mu^T\mathbf{X}\mu' - \nu^T\mathbf{Y}\nu'\right) + c^2$. This implies that COOT is weakly invariant to translation. ■

AGW inherits the weak invariant to translation from COOT.

Proposition 2. *For any $\alpha \in [0, 1]$, AGW is weakly invariant to translation.*

Proof of Proposition 2. Note that the GW term in AGW remains unchanged by translation. By adapting the proof of Corollary 2, we obtain

$$\mathrm{AGW}_\alpha(\mathbf{X}, \mathbf{Y} + c) = \mathrm{AGW}_\alpha(\mathbf{X}, \mathbf{Y}) + (1 - \alpha)\left[c^2 - 2c\left(\mu^T\mathbf{X}\mu' - \nu^T\mathbf{Y}\nu'\right)\right].$$

The result then follows. ■

B Supplementary Illustrations

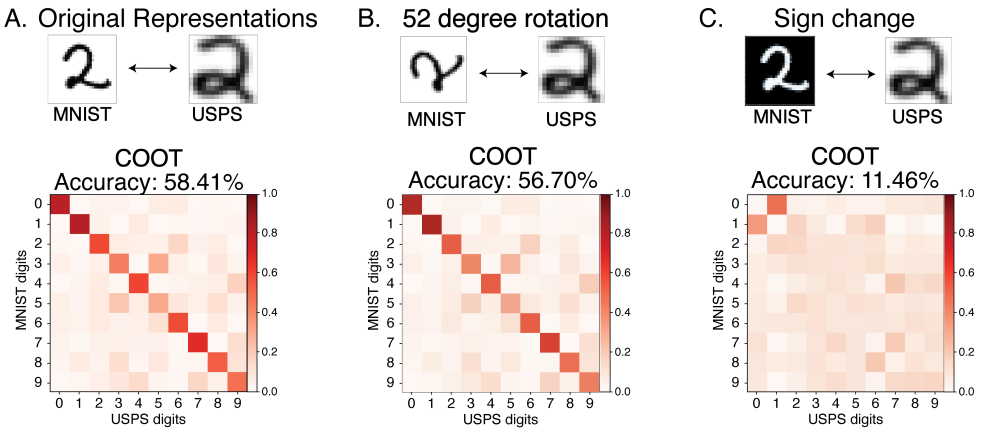

Figure S1: Examples of Isometric Transformations to which COOT is not invariant, unlike GW distance, which yields a consistent alignment accuracy of 48.13%.

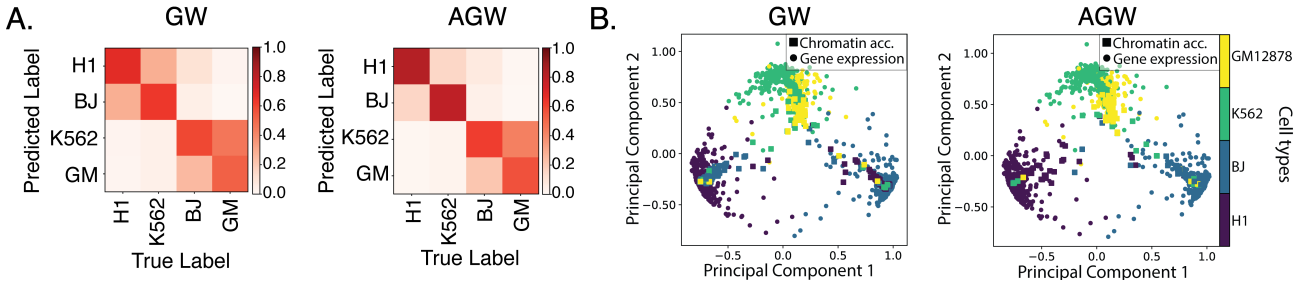

Figure S2: **(A)** Confusion matrix showing GW distance tends to mismatch H1 and BJ clusters more often than GW over 50 random subsampling of cells in the SNARE-seq dataset. **(B)** An example where GW distance mismatches these clusters (H1 cells being matched to the BJ cluster) when their local geometries look similar upon downsampling while AGW yields more accurate results.

C Supplementary results on single-cell multi-omic integration

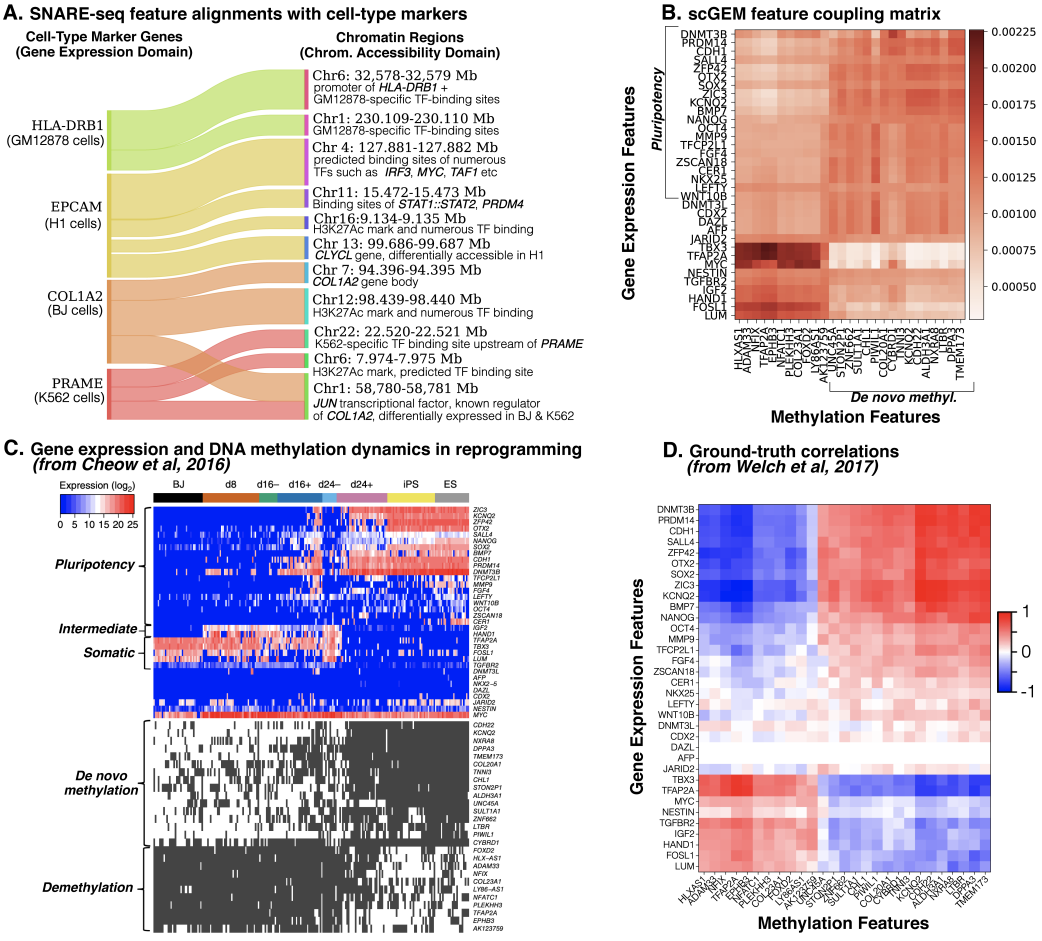

Figure S3: AGW's Feature Alignments for (A) SNARE-Seq, and (B) scGEM Datasets. **A.** The Sankey plot presents the four cell-type marker genes and their top correspondences in the open chromatin regions. **B.** visualizes the feature coupling matrix for the scGEM dataset. **C** is borrowed from the publication for scGEM dataset (Cheow et al., 2016a), which shows how the genomic features in the gene expression and DNA methylation domains vary during cellular differentiation. **D** is a heatmap borrowed from Welch *et al.* that shows empirical correlations between the features of the two measurement domains, which we use for comparison with Panel B.

For completeness of results, we visualize the feature alignments AGW obtains for the SNARE-seq and scGEM datasets in Figure S3 despite the lack of ground-truth information on these. Due to the high number of features

in the SNARE-seq dataset (1000 genes and 3000 chromatin regions), we only show the four cell-type marker genes and their top correspondences from the accessible chromatin regions in Panel A. For the scGEM dataset, we visualize the full feature coupling matrix in Panel B, which can be interpreted in light of the information presented in Panels C and D. We detail the significance of the inferred correspondences below.

SNARE-seq feature correspondences To validate the correspondences in Figure S3 Panel A, we consult the biological annotations on the UCSC Genome Browser (Navarro Gonzalez et al., 2020), as well as gene regulatory information databases, such as GRNdb (Fang et al., 2020) and RepMap Atlas of Regulatory Regions (Hammal et al., 2021). Most of these correspondences agree with either experimentally validated or computationally predicted regulatory relationships.

Three of the alignments are between marker genes and their own chromatin regions. The first is the *PRAME* and Chr22: 22.520-22.521 Mb region alignment, which is a region upstream of the *PRAME* gene body that is rich with predicted transcription factor (TF) binding sites based on the RepMap Atlas of Regulatory Regions (Hammal et al., 2021) annotations on UCSC Genome Browser (Human hg38 annotations) (Navarro Gonzalez et al., 2020). Among the predicted TF bindings, some are K562-specific predictions, and are previously reported regulators of *PRAME*, such as *E2F6*, *HDAC2*, *CTCF* (based on GRNdb database (Fang et al., 2020) of TF-gene relationships). Then *COL1A2* and *HLA-DRB1* also have recovered correspondences with their own chromosomal region, "Chr7:94.395-94.396 Mb" and "Chr6:32,578-32,579 Mb", respectively. Both these genes are also additionally aligned with "Chr1: 58,780 - 58,781 Mb" region, which corresponds to the gene body of *JUN* transcription factor. Indeed, Chen et al. (2019) identify *JUN* as a TF differentially expressed in the K562 and BJ cells, but more strongly in the latter. GRNdb also lists *JUN* as a regulator of *COL1A2* gene.

PRAME has another region abundant in predicted TF binding sites among its top correspondences: "Chr6: 7.974-7.975 Mb". This region is annotated with an H3K27Ac mark on the UCSC Genome Browser, which is an acetylation mark that is often found near gene regulatory elements on the genome (Navarro Gonzalez et al., 2020). Furthermore, this region contains multiple predicted binding sites of TFs GRNdb identifies as regulators of *PRAME*, such as *IRF1, HDAC2, HOXC6* and POU2AF1. The *HLA-DRB1* gene is also aligned with a chromosomal region rich in GM12878-specific predictions of TF bindings, such as *IRF4*, *IRF8*, *ETV6*, and *CREM*, which GRNdb lists as potential regulators of *HLA-DRB1*. Lastly, although we could not find a biological relationship reported between the *EPCAM* gene (marker gene for the H1 cell-line) and the chromatin region for the *CLYBL* gene, this region indeed appears to be differentially accessible in H1 cells in the SNARE-seq dataset.

scGEM feature correspondences In the absence of ground-truth data, we consult the publication that introduced the scGEM dataset (Cheow et al., 2016a) to interpret the feature coupling yielded by AGW in Panel B. Figure S3 Panel C presents a plot from this paper, which shows how the expression of genes that drive pluripotency during cell differentiation correlate or anti-correlate with the methylation of the genes in the "DNA methylation" domain. We observe B that AGW's coupling also recovers an alignment between the expression profiles of pluripotency-driving genes and the methylation levels of associated genes. Panel D visualizes the underlying correlations between the gene expression and DNA methylation domains computed by Welch *et al.* AGW's feature coupling reflects the structure in this ground-truth correlation matrix. Note that the rows and columns are ordered identically in Panel B to aid with the visual comparison.

D Supplementary results on heterogeneous domain adaptation (HDA)

We present the results for the semi-supervised heterogeneous domain adaptation (HDA) experiments below. Differently than the unsupervised HDA experiments, we consider two additional baselines that were specifically developed for semi-supervised HDA, namely the "KeyPoint-Guided model by ReLational preservation" (KPG-RL) and the "KPG-RL with Gromov-Wasserstein" (KPG-RL-GW) (Gu et al., 2022). For these baselines, we use the implementation provided at https://github.com/XJTU-XGU/KPG-RL. Both KPG-RL and KPG-RL-GW create a "guidance matrix" to supervise the coupling such that the coupling probabilities between keypoints and their match (points selected for supervision) are set to 1 while all other possible matches for these keypoints are assigned 0. Therefore, their supervision scheme assumes there is $1-1$ correspondence between keypoints. We note that, however, the datasets in our HDA experiments contain images from the same classes but they don't necessarily contain the same image (*e.g.* in the Caltech10 vs Amazon datasets). As such, an image here realistically has multiple possible correspondences.

Table S1: Heterogeneous Domain Adaptation Results in the Semi-Supervised Case with $t = 1$ Sample Used for Supervision. Best results are bolded and second best results are underlined. For AGW, the α values used are: $0.6, 0.2, 0.2, 0.9, 0.4, 0.5, 0.9, 0.3, 0.1$ from left to right.

	A → A	**A → C**	**A → W**	**C → A**	**C → C**	**C → W**	**W → A**	**W → C**	**W → W**
AGW	**93.1±1.6**	90.2±5.1	**90.3±3.5**	**78.8±7.7**	**84.2±2.3**	**77.3±4.2**	**92.8±3.8**	**90.3±3.**	98.5±0.8
COOT	87.1±4.9	44.4±4.9	54.5±9.8	74.5±3.1	45.4±12.2	40.8±12.2	86.9±2.3	39.7±3.6	76.1±13.5
GW	92.4±1.8	90.6±5.5	85.7±3.9	77.6±8.6	82.8±3.0	73.3±6.9	91.9±7.9	81.9±9.6	97.2±1.4
UCOOT	85.3±5.2	49.0±5.0	63.5±1.5	76.6±3.5	62.8±4.6	68.2±3.0	80.3±5.6	66.2±4.5	85.3±2.1
UGW	91.9±3.7	**90.9±4.6**	84.3±2.1	78.7±8.2	77.5±5.4	72.8±4.9	88.3±6.0	80.9±5.2	**98.7±1.5**
KPG-RL	85.0±7.1	85.5±5.0	85.4±5.2	73.5±5.6	70.3±4.2	73.4±4.8	86.7±4.3	84.0±4.6	88.1±3.6
KPG-GW-RL	86.1±2.6	79.2±4.3	84.0±3.8	71.9±6.6	69.4±6.0	71.1±3.1	83.8±5.2	78.8±4.7	87.5±3.4

Table S2: Domain Adaptation Results in the Semi-Supervised Case with $t = 3$ Samples Used for Supervision. Best results are bolded and second best results are underlined. For the AGW, the α values used are $0.2, 0.1, 0.2, 0.7, 0.2, 0.9, 0.8, 0.9, 0.4$ from left to right.

	A → A	**A → C**	**A → W**	**C → A**	**C → C**	**C → W**	**W → A**	**W → C**	**W → W**
AGW	**96.0±0.8**	**93.5±1.8**	**93.8±0.7**	**85.6±1.2**	**86.5±2.0**	**83.2±2.4**	**97.1±0.8**	**94.7±1.1**	98.7±0.5
COOT	91.1±2.0	59.7±3.6	72.6±4.4	83.1±5.1	59.3±8.4	64.6±6.2	94.3±2.2	55.0±7.1	87.4±4.4
GW	93.2±0.9	92.8±2.1	91.6±1.8	81.2±1.2	85.3±2.8	79.7±2.5	93.4±5.2	90.9±3.5	97.4±2.6
UCOOT	90.0±3.2	67.3±2.9	80.3±1.8	84.6±2.3	64.9±4.1	68.7±3.4	83.6±4.1	69.9±6.2	89.9±1.8
UGW	92.6±2.8	93.2±3.0	88.5±2.9	82.5±6.2	81.8±5.6	77.3±7.0	91.2±3.8	88.9±8.2	**98.8±3.2**
KPG-RL	89.9±3.8	89.7±2.1	90.3±1.8	82.2±2.9	78.4±2.2	79.6±3.4	93.1±2.3	91.5±3.2	95.1±2.1
KPG-GW-RL	86.6±4.3	83.0±3.6	87.9±2.8	78.7±3.4	75.2±5.0	78.8±5.8	90.1±2.7	84.9±5.1	93.6±2.3

Table S3: Heterogeneous Domain Adaptation Results in the Semi-Supervised Case with $t = 5$ Samples Used for Supervision. Best results are bolded and second best results are underlined. For the AGW, the α values used are $0.3, 0.1, 0.7, 0.1, 0.5, 0.8, 0.9, 0.2, 0.9$ from left to right.

	A → A	**A → C**	**A → W**	**C → A**	**C → C**	**C → W**	**W → A**	**W → C**	**W → W**
AGW	**96.4±1.3**	**96.4±1.7**	93.1±2.9	**86.2±2.3**	**86.6±1.9**	83.9±1.4	**96.7±1.1**	**95.7±2.1**	98.7±0.9
COOT	93.6±1.5	66.1±3.8	75.7±3.5	85.2±2.1	64.7±7.2	67.0±7.5	96.3±1.9	60.5±5.3	90.3±1.9
GW	93.8±2.1	91.9±2.2	**93.3±1.2**	85.2±3.6	84.5±2.7	81.9±3.6	96.6±1.1	94.5±3.1	98.4±0.9
UCOOT	91.3±4.3	70.1±5.5	88.8±1.6	85.1±4.2	70.8±3.4	76.3±1.7	91.5±3.6	73.6±6.4	92.4±2.0
UGW	93.2±1.7	93.4± 3.0	90.8±1.9	85.8±2.6	82.7±2.2	80.6±3.2	95.1±0.8	92.9±2.4	**98.9±0.7**
KPG-RL	92.6±1.9	90.5±2.9	91.9±2.1	83.9±2.7	82.0±2.8	**84.4±4.2**	94.7±3.8	93.3±2.5	97.0±2.3
KPG-GW-RL	89.9±1.9	85.0±1.8	91.9±2.3	81.0±4.4	79.9±3.1	82.1±2.4	91.9±2.9	88.5±3.9	97.1±1.5

E Empirical Runtime Analysis

For runtime comparisons, we present AGW, GW and COOT runtimes in Table S4 on HDA experiments. We ran all algorithms on an Intel Xeon e5-2670 CPU with 16GB memory. To ensure consistency in comparisons, we kept the regularization coefficients the same across all runs, with the coefficient of entropic regularization over sample coupling as $5e-4$ and the coefficient of entropic regularization over the feature couplings (in COOT and AGW) as $1e-3$. These were the values most often picked by hyperparameter tuning.

Table S4: Runtime per Iteration and Number of Iterations Before Convergence of AGW, GW and COOT Algorithms on HDA Experiments. The same convergence criteria are used across all methods.

	Number of Iterations			**Runtime per Iteration**		
	COOT	**AGW (α=0.5)**	**GW**	**COOT**	**AGW (α=0.5)**	**GW**
A $\rightarrow$ **A**	2.1 $\pm$ 0.5	13.9 $\pm$ 1.8	42.7 $\pm$ 17.3	0.18 $\pm$ 0.02	0.22 $\pm$ 0.05	0.22 $\pm$ 0.04
A $\rightarrow$ **C**	1.9 $\pm$ 0.7	13.7 $\pm$ 2.5	56.4 $\pm$ 21.4	0.16 $\pm$ 0.01	0.21 $\pm$ 0.07	0.26 $\pm$ 0.02
A $\rightarrow$ **W**	2.3 $\pm$ 0.8	13.4 $\pm$ 1.0	41.3 $\pm$ 14.8	0.18 $\pm$ 0.03	0.20 $\pm$ 0.01	0.23 $\pm$ 0.06
C $\rightarrow$ **A**	2.0 $\pm$ 0.0	15.7 $\pm$ 3.5	62.4 $\pm$ 14.5	0.23 $\pm$ 0.05	0.26 $\pm$ 0.09	0.28 $\pm$ 0.03
C $\rightarrow$ **C**	2.0 $\pm$ 0.4	12.0 $\pm$ 1.9	54.1 $\pm$ 12.0	0.21 $\pm$ 0.04	0.24 $\pm$ 0.07	0.22 $\pm$ 0.04
C $\rightarrow$ **W**	2.4 $\pm$ 0.8	11.6 $\pm$ 2.2	72.5 $\pm$ 19.7	0.20 $\pm$ 0.02	0.21 $\pm$ 0.01	0.24 $\pm$ 0.05
W $\rightarrow$ **A**	2.0 $\pm$ 0.0	14.5 $\pm$ 1.4	32.7 $\pm$ 17.8	0.20 $\pm$ 0.06	0.22 $\pm$ 0.01	0.27 $\pm$ 0.04
W $\rightarrow$ **C**	2.2 $\pm$ 0.6	11.8 $\pm$ 2.3	48.3 $\pm$ 8.2	0.17 $\pm$ 0.02	0.19 $\pm$ 0.01	0.20 $\pm$ 0.09
W $\rightarrow$ **W**	2.0 $\pm$ 0.0	13.6 $\pm$ 1.1	52.7 $\pm$ 7.2	0.20 $\pm$ 0.03	0.21 $\pm$ 0.02	0.23 $\pm$ 0.02

Table S4 shows that AGW tends to converge in fewer iterations than GW. We observe that this is thanks to the further refinement of the sample coupling matrix and quicker drop in GW cost as influenced by the feature coupling (after feature optimization step of the COOT term in the same iteration, as all else remains the same between the two algorithms).

F Experimental Set-up Details

F.1 MNIST Illustrations

We align 1000 images of hand-written digits from the MNIST dataset with 1000 images from the USPS dataset. Each dataset is subsampled to contain 100 instances of each of the 10 possible digits (0 through 9), using the random seed of 1976. We set all marginal distributions to uniform, and use cosine distances for GW and AGW. We consider both the entropically regularized and non-regularized versions for all methods. For entropic regularization, we sweep a grid of ϵ_1, ϵ_2(=if applicable) $\in [5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 1e-1, 5e-1]$. For AGW, we consider $[0.1, 0.2, 0.3, ..., 0.9]$, and present results with the best-performing hyperparameter combination of each method, as measured by the percent accuracy of matching images from the same digit across the two datasets.

F.2 Single-cell multi-omic alignment experiments

As a real-world application of AGW, we align single-cell data from different measurement domains. Optimal transport has recently been applied to this problem in computational biology by multiple groups (Demetci et al., 2020; Cao et al., 2021, 2022). To briefly introduce the problem: Biologists are interested in jointly studying multiple genomic (*i.e.*, "multi-omic") aspects of cells to determine biologically relevant patterns in their co-variation. Such studies reveal how the different molecular aspects of a cell's genome (*e.g.*, its 3D structure, chemical modifications it undergoes, activity levels of its genes, etc) interact to regulate the cell's response to its environment. These studies are of interest for both fundamental biology research, as well as drug discovery applications. However, as Liu et al. (2019) describe, combining multiple measurements on the same cells is experimentally difficult. Consequently, computational approaches are developed to integrate data from different measurement modalities using biologically relevant cell populations. In this paper, we apply AGW to jointly align both cells and genomic features of single-cell datasets. This is a novel direction in the application of optimal

transport (OT) to single-cell multi-omic alignment tasks, as the existing OT-based algorithms only align cells.

Datasets We largely follow the first paper that applied OT to single-cell multi-omic alignment task (Demetci et al., 2020) in our experimental set-up and use four simulated datasets and three real-world single-cell multi-omic datasets to benchmark our cell alignment performance.

Three of the simulated datasets have been generated by Liu et al. (2019) by non-linearly projecting 600 samples from a common 2-dimensional space onto different 1000- and 2000-dimensional spaces with 300 samples in each. In the first simulation, the data points in each domain form a bifurcating tree structure that is commonly seen in cell populations undergoing differentiation. The second simulation forms a three-dimensional Swiss roll. Lastly, the third simulation forms a circular frustum that resembles what is commonly observed when investigating the cell cycle. These datasets have been previously used for benchmarking by other cell-cell alignment methods (Liu et al., 2019; Singh et al., 2020; Cao et al., 2020, 2021; Demetci et al., 2020). We refer to these datasets as "Sim 1", "Sim 2", and "Sim 3", respectively.

We include a fourth simulated dataset generated by Demetci et al. (2020) using a single-cell RNA-seq data simulation package in R, called Splatter (Zappia et al., 2017). We refer to this dataset as "Synthetic RNA-seq". This dataset includes a simulated gene expression domain with 50 genes and 5000 cells divided across three cell types and another domain created by non-linearly projecting these cells onto a 500-dimensional space. As a result of their generation schemes, all simulated datasets have ground-truth 1-1 cell correspondence information. We use this information solely for benchmarking. We do not have access to ground-truth feature relationships in these datasets, so we exclude them from feature alignment experiments.

Additionally, we include three real-world single-cell sequencing datasets in our experiments. To have ground-truth information on cell correspondences for evaluation, we choose three co-assay datasets which have paired measurements on the same individual cells: a scGEM dataset (Cheow et al., 2016a), a SNARE-seq dataset (Chen et al., 2019), and a CITE-seq dataset (Stoeckius et al., 2017) (these are exceptions to the experimental challenge described above). These first two datasets have been used by existing OT-based single-cell alignment methods (Cao et al., 2020; Singh et al., 2020; Demetci et al., 2020; Cao et al., 2021; Demetci et al., 2022a), while the last one was included in the evaluations of a non-OT-based alignment method, bindSC (Dou et al., 2022). The scGEM dataset contains measurements on gene expression and DNA methylation states of 177 individual cells from the human somatic cell population undergoing conversion to induced pluripotent stem cells (iPSCs) (Cheow et al., 2016a). We accessed the pre-processed count matrices for this dataset through the following GitHub repository: `https://github.com/caokai1073/UnionCom`. The SNARE-seq dataset contains gene expression and chromatin accessibility profiles of 1047 individual cells from a mixed population of four cell lines: H1(human embryonic stem cells), BJ (a fibroblast cell line), K562 (a lymphoblast cell line), and GM12878 (lymphoblastoid cells derived from blood) (Chen et al., 2019). We access their count matrices online from the Gene Expression Omnibus platform with the accession code GSE126074. Finally, the CITE-seq dataset has gene expression profiles and epitope abundance measurements on 25 antibodies from 30,672 cells from human bone marrow tissue, which we randomly downsample to 1000 cells (Stoeckius et al., 2017). The count matrices for this dataset were downloaded from the Seurat website [5]. We use these three real-world single-cell datasets for both cell-cell (*i.e.*, sample-sample) and feature-feature alignment benchmarking.

In addition to these three datasets, we include a fourth single-cell dataset, which contains data from the same measurement modality (*i.e.*, gene expression) but from two different species: mouse (Bhattacherjee et al., 2019) and bearded lizard (Tosches et al., 2018). Our motivation behind including this dataset is to demonstrate the effects of both sample-level (*i.e.*, cell-level) and feature-level (*i.e.*, gene-level) supervision on alignment qualities. We refer to this dataset as the "cross-species dataset", which contains 4,187 cells from lizard pallium (a brain region) and 6,296 cells from the mouse prefrontal cortex. The two species share a subset of their features: 10,816 paralogous genes. Each also has species-specific genes: 10,184 in the mouse dataset and 1,563 in the lizard dataset. The data comes from different species, so there is no 1–1 correspondence between cells. However, the two species contain cells from similar cell types. Unlike the other single-cell dataset, there is a subset of the features (the paralogous genes) that have 1–1 correspondences across the two domains (domains are defined by species in this dataset).

[5] `https://satijalab.org/seurat/v4.0/weighted_nearest_neighbor_analysis.html`

Baselines and hyperparameter tuning We benchmark AGW's performance on single-cell alignment tasks against three algorithms: (1) COOT (Redko et al., 2020), (2) SCOT (Demetci et al., 2020), which is a Gromov-Wasserstein OT-based algorithm that uses k-nearest neighbor (kNN) graph distances on dimensionality reduced datasets (top 30 principal components for gene expression domains and simulated domains, 15-25 topics with latent Dirichlet allocation for other measurement domains) as intra-domain distance matrices. This choice of distances has been shown to perform better than Euclidean distances, cosine distances by Demetci et al. (2020)), and bindSC (Dou et al., 2022). For consistency, we keep the intra-domain distance computations the same for AGW and UGW, too. Among all baselines, bindSC is not an OT-based algorithm: It employs bi-order canonical correlation analysis to perform alignment. We include it as a benchmark as it is the only existing single-cell alignment algorithm that can perform feature alignments (in addition to cell alignments) for a few limited types of measurement modalities.

When methods share similar hyperparameters in their formulation (*e.g.*, entropic regularization constant, ϵ for methods that employ OT), we use the same hyperparameter grid to perform their tuning. Otherwise, we refer to the publication and the code repository for each method to choose a hyperparameter range. For SCOT, we tune four hyperparameters: $k \in \{20, 30, \ldots, 150\}$, the number of neighbors in the cell neighborhood graphs, $\epsilon \in \{5e-4, 3e-4, 1e-4, 7e-3, 5e-3, \ldots, 1e-2\}$, the entropic regularization coefficient for the optimal transport formulation. Similarly, for both COOT, UCOOT, UGW (SCOTv2 Demetci et al. (2022a)) and AGW, we sweep $\epsilon_1, \epsilon_2 \in \{5e-4, 3e-4, 1e-4, 7e-3, 5e-3, \ldots, 1e-2\}$ for the coefficients of entropic regularization over the sample and feature alignments. We use the same intra-domain distance matrices in AGW as in SCOT (based on kNN graphs). For AGW, we consider the interpolation coefficient of $\alpha \in \{0.1, 0.2, ..., 0.9\}$.For the unbalanced formulations, namely UGW and UCOOT, we sweep $lambda_1 \in \{1e-3, 5e-3, ..., 100\}$, corresponding to the mass conservation relaxation coefficient over samples (*i.e.*cells). In UCOOT, we sweep the same interval for λ_2, the same relaxation parameter over features. For all OT-based methods, we perform barycentric projection to complete the alignment.

For bindSC, we choose the coupling coefficient that assigns weight to the initial gene activity matrix $\alpha \in \{0, 0.1, 0.2, \ldots 0.9\}$ and the coupling coefficient that assigns a weight factor to multi-objective function $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$. Additionally, we choose the number of canonical vectors for the embedding space $K \in \{3, 4, 5, 10, 30, 32\}$. For all methods, we report results with the best-performing hyperparameter combinations.

Evaluation Metrics When evaluating cell alignments, we use a metric previously used by other single-cell multi-omic integration tools (Liu et al., 2019; Singh et al., 2020; Cao et al., 2020; Demetci et al., 2020; Cao et al., 2021; Demetci et al., 2022a; Dou et al., 2022) called "fraction of samples closer than the true match" (FOSCTTM). For this metric, we compute the Euclidean distances between a fixed sample point and all the data points in the other domain. Then, we use these distances to compute the fraction of samples that are closer to the fixed sample than its true match and then average these values for all the samples in both domains. This metric measures alignment error, so the lower values correspond to higher-quality alignments.

We investigate the accuracy of feature correspondences recovered to assess feature alignment performance. We mainly use two real-world datasets for this task - CITE-seq, and the cross-species scRNA-seq datasets (results on SNARE-seq and scGEM datasets are qualitatively evaluated due to the lack of ground-truth information). For the CITE-seq dataset, we expect the feature correspondences to recover the relationship between the 25 antibodies and the genes that encode them. To investigate this, we simultaneously align the cells and features of the two modalities using the 25 antibodies and 25 genes in an unsupervised manner. We compute the percentage of 25 antibodies whose strongest correspondence is their encoding gene.

For the cross-species RNA-seq dataset, we expect alignments between (1) the cell-type annotations common to the mouse and lizard datasets, namely excitatory neurons, inhibitory neurons, microglia, OPC (Oligodendrocyte precursor cells), oligodendrocytes, and endothelial cells and (2) between the paralogous genes. For this dataset, we generate cell-label matches by averaging the rows and columns of the cell-cell alignment matrix yielded by AGW based on these cell annotation labels. We compute the percentage of these six cell-type groups that match as their strongest correspondence. For feature alignments, we compute the percentage of the 10,816 shared genes that are assigned to their corresponding paralogous gene with their highest alignment probability. For this dataset, we consider providing supervision at increasing levels on both sample and feature alignments. For feature-level supervision, 20% supervision means setting the alignment cost of $\sim 20\%$ of the genes with their paralogous pairs to 0. For sample-level supervision, 20% supervision corresponds to downscaling the alignment cost of $\sim 20\%$ of

the mouse cells from the aforementioned seven cell types with the $\sim 20\%$ of lizard cells from their corresponding cell-type by $\frac{1}{\#\text{ lizard cells in the same cell-type}}$.

F.3 Heterogeneous domain adaptation experiments

For each pair of domains (A)-(C), (A)-(W), (C)-(C) etc, we sweep a hyperparameter grid over 5 runs of random selecting 10 samples in each class to form a validation dataset, and choose the hyperparameter combination that best performed on average. We then randomly select 20 other samples per class and perform adaptation from CaffeNet to GoogleNet, report average performance along with standard deviations over 10 repetitions. For all methods that allow for entropic regularization, we consider their version with no entropic regularization (either on the sample-level alignments, feature-level alignments, or both), along with various levels of regularization. Similarly to other experiments, for entropic regularization over sample alignments, we consider $\epsilon_1 \in [5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 0.1]$ in the hyperparameter grid. For entropic regularization over feature alignments in COOT and AGW, we also consider $\epsilon_2 \in [5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 0.1]$. For unbalanced formulations, we consider the mass relaxation constants of λ_1, (and if applicable)$\lambda_2 \in \{1e-3, 5e-3, 1e-2, ..., 100\}$. We consider $\alpha \in [0.1, 0.2, ..., 0.9]$ for interpolation coefficient of AGW.