PASTE2: Partial Alignment of Multi-slice Spatially Resolved Transcriptomics Data

Xinhao Liu[1], Ron Zeira[2], and Benjamin J. Raphael[*1]

[1]Department of Computer Science, Princeton University, 35 Olden St, Princeton, NJ 08540
[2]Verily Life Sciences, Tel Aviv, Israel

Abstract

Spatially resolved transcriptomics (SRT) technologies measure mRNA expression at thousands of locations in a tissue slice. However, nearly all SRT technologies measure expression in two dimensional slices extracted from a three-dimensional tissue, thus losing information that is shared across multiple slices from the same tissue. Integrating SRT data across multiple slices can help recover this information and improve downstream expression analyses, but multi-slice alignment and integration remains a challenging task. Existing methods for integrating SRT data either do not use spatial information or assume that the morphology of the tissue is largely preserved across slices, an assumption that is often violated due to biological or technical reasons. We introduce PASTE2, a method for *partial* alignment and 3D reconstruction of multi-slice SRT datasets, allowing only partial overlap between aligned slices and/or slice-specific cell types. PASTE2 formulates a novel *partial* Fused Gromov-Wasserstein Optimal Transport problem, which we solve using a conditional gradient algorithm. PASTE2 includes a model selection procedure to estimate the fraction of overlap between slices, and optionally uses information from histological images that accompany some SRT experiments. We show on both simulated and real data that PASTE2 obtains more accurate alignments than existing methods. We further use PASTE2 to reconstruct a 3D map of gene expression in a Drosophila embryo from a 16 slice Stereo-seq dataset. PASTE2 produces accurate alignments of multi-slice datasets from multiple SRT technologies, enabling detailed studies of spatial gene expression across a wide range of biological applications.

Code availability: Software is available at `https://github.com/raphael-group/paste2`

[*]Correspondence: `braphael@princeton.edu`

1 Introduction

Spatially resolved transcriptomics (SRT) technologies measure mRNA expression simultaneously at thousands of locations within a tissue. These technologies include both sequencing based approaches, such as 10X Genomics Visium [1] and Slide-seq [37, 40], as well as hybridization and florescent approaches such as MERFISH [13] and seqFISH [30]. Nearly all of these technologies measure expression at 2D locations within a thin tissue slice ($\approx 10\mu m$), and we will use the term spatial transcriptomics (ST) as a generic term to refer to any of these technologies. ST provides spatial context that is missing from single-cell RNA-sequencing (scRNA-seq) measurement of mRNAs from disassociated cells, and has been widely used to study both normal [2, 45] and diseased tissues, such as cancer [23, 39, 42] and Alzheimer's disease [14]. However, similar to scRNA-seq, ST data suffers from high rates of sparsity. Moreover, recording only the x, y coordinates on a 2D tissue slice loses information along the z (orthogonal) direction of the 3D tissue, hindering a comprehensive analysis of the whole tissue (Fig. 1).

Spatial transcriptomics is often applied to multiple sequential 2D slices from the same tissue (Fig. 1), thus opening the possibility of performing integrative analysis of all slices. Such joint analysis of multiple slices not only helps with the data sparsity problem in individual slices, but also enables innovative downstream tasks such as 3D spatial expression analysis, 3D cell-cell communication, and 3D clustering [29, 48]. However, aligning multiple slices from the same tissue along the orthogonal direction to recover spot-spot correspondence across slices is a challenging task due to morphological differences across slices as well as technical variability in mRNA capture between experiments.

Several approaches have been used for alignment of multiple ST slices. One approach is to apply methods developed for scRNA-seq and multi-omics data integration, such as Seurat [41], SCOT [15, 16], or Pamona [10]. Another approach is to use methods that align an scRNA-seq dataset onto an ST dataset, such as Tangram [5] or RCTD [8]. However, these methods are designed for different alignment tasks and ignore the spatial information within or across slices. Another method, STUtility [4] is designed to align a pair of ST slices, but aligns only the histology images, ignoring both gene expression and spatial information. Morover, this method can only be applied to 10X Genomics Visium data. Another recent method, GPSA [24] integrates multiple ST slices into a common coordinate system, but does not output a mapping between spots that can be used for downstream analysis, and the common coordinate system it produces is different from the 3D coordinates of the tissue.

Another possible solution is to use histological and medical image registration [31] toolkits such as ITK [33] and SimpleITK [3]. However, many image registration methods are supervised and often require manually selected landmarks, creating an extra burden on the user. The spatial alignment problem has also been studied in the context of functional magnetic resonance imaging (fMRI) data registration [7, 26, 27], but these methods are not easily extensible to the spatial genomics setting [24]. Finally, many ST technologies do not have matching histological images.

Recently, PASTE, a method that performs probabilistic alignment of ST slices using both spatial and transcriptional similarities, was introduced [48]. However, PASTE assumes that the slices overlap over the full 2D assayed region, with similar field of view and similar number and proportion of cell types. Essentially, PASTE assumes the two slices are biological/technical replicates of a 2D tissue, an assumption that is often violated in real ST experiments due to technical difficulties in tissue dissection and array placement, or differences in tissue morphology between nearby slices. For example, two slices may only partially overlap along the z axis due to different placements of the array on the tissue, and hence only a part of both slices should be aligned (Fig. 1). Furthermore, two slices may have different compositions of cell types, leading to slice-specific cell types and spots that should not be aligned.

Here, we introduce PASTE2, a method to align multiple adjacent ST slices from the same tissue with several substantial improvements over existing methods. First, PASTE2 performs *partial* pairwise alignment, selecting and aligning only a subset of spots. PASTE2 thus addresses the important case where adjacent

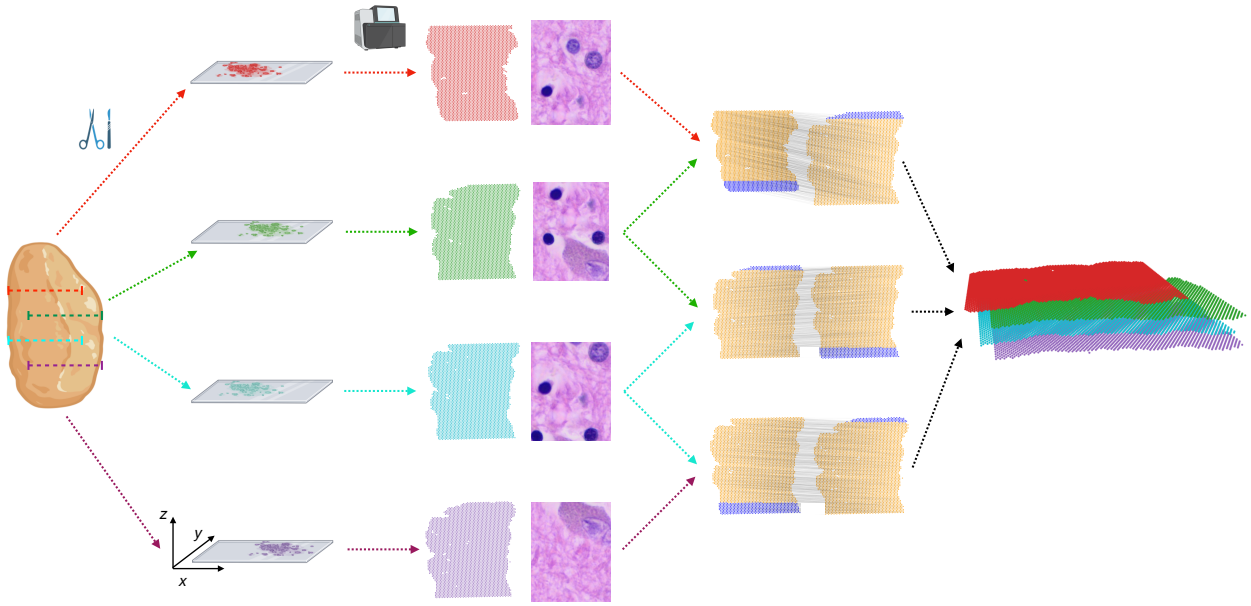

Figure 1: PASTE2 partial alignment of overlapping slices. Four thin slices (red, green, blue, purple) are dissected from the same tissue and placed on an ST array. However, these slices only partially overlap in the z-coordinate direction. The inputs to PASTE2 are the four ST slices, including gene expression, spot locations, and optionally, histology images. PASTE2 computes a *partial* alignment of each pair of adjacent slices by selecting subsets of spots from each slice that preserve transcriptional, spatial and image similarity. PASTE2 uses the partial alignment to create a 3D spatial reconstruction of the tissue.

slices do not fully overlap in space or have different cell type compositions. To solve the partial alignment problem, we introduce the *partial Fused Gromov-Wasserstein* (*partial-FGW*) *optimal transport* framework. Partial-FGW is the partial extension [11] of the Fused Gromov-Wasserstein optimal transport [43] and allows only a fraction of the total mass to be transported between the two distributions. To the best of our knowledge, PASTE2 is the first to formulate the partial-FGW problem and provide an optimization procedure to solve this problem. Second, PASTE2 includes a model selection procedure to estimate the fraction of overlap between two slices to align, which is in general a very difficult problem. Third, PASTE2 optionally uses the histological images. Some ST technologies, such as the 10X Genomics Visium platform, can also produce a Hematoxylin and Eosin (H&E) stained image of the same tissue slice where gene expression is measured. The information in this image can aid in alignment of slices by identifying spots with similar histology. Finally, we provide a generalized Procrustes analysis [46] method for 3D spatial reconstruction of the tissue from partially aligned 2D slices.

We demonstrate PASTE2's advantages on both simulated and real ST datasets. We show on simulated data that PASTE2 achieves accurate alignment and outperforms PASTE when slices do not fully overlap. On ST dataset from the human dorsolateral prefrontal cortex (DLPFC) [32], we show that PASTE2 computes more accurate alignments than competing methods, and the use of histological images can further improve the alignment. Finally, we demonstrate PASTE2's applicability to larger datasets using different SRT technologies by aligning 16 Stereo-Seq slices from a Drosophila embryo [47].

2 Methods

A spatial transcriptomics (ST) experiment on a 2D tissue slice yields a pair (X, Z), where $X \in \mathbb{N}^{n \times p}$ is the gene expression matrix of the tissue slice, and $Z \in \mathbb{R}^{2 \times n}$ is the spatial location matrix of each spot on the slice, where the j-th column $\mathbf{z}_{\cdot j}$ is the x-y coordinate of spot j on the 2D array[1] used by the ST experiment. Here, n is the number of spots on the slice and p is the number of genes measured. $x_{ij} \in \mathbb{N}$ is the transcript

[1]We refer here to array based technologies, but the formulation is the same for other technologies.

count of gene j in spot i. Each row vector $\mathbf{x}_{i\cdot}$ of X is the *expression profile* of spot i. Following [48], we encode the spatial location of each spot in a pairwise distance matrix $D = [d_{ij}] \in \mathbb{R}_+^{n\times n}$, where d_{ij} is the Euclidean distance between spot i and spot j on the slice, calculated from the 2D coordinates $\mathbf{z}_{\cdot i}$ and $\mathbf{z}_{\cdot j}$. Thus, we represent an ST slice of n spots and p genes by a tuple (X, D).

2.1 Partial pairwise slice alignment problem

Given a pair, (X, D) and (X', D') of ST slices, our goal is to compute a *partial pairwise slice alignment*, i.e. to find a probabilistic spot-spot correspondence between spots in the two slices while accounting for the fact that some spots should not be mapped (Fig. 1). The probabilistic mapping is a matrix $\pi = [\pi_{ij}] \in \mathbb{R}^{n\times n'}$ between the n spots in one slice and n' spots in the other slice, where π_{ij} describes the probability (or relative fraction) that a spot i in the first slice is aligned to a spot j in the second slice.

We begin by describing the solution given in [48] to the pairwise slice alignment problem, implemented in the PASTE algorithm. PASTE uses a formulation based on optimal transport to compute the mapping π. Specifically, given probability distributions g and g' over the spots in slice X and X', respectively, PASTE finds the map π (also known as the transport matrix) that minimizes the following transport cost:

$$F(\pi; X, D, X', D', c, \alpha) = (1-\alpha)\sum_{i,j} c(\mathbf{x}_{i\cdot}, \mathbf{x}'_{j\cdot})\pi_{ij} + \alpha \sum_{i,j,k,l} (d_{ik} - d'_{jl})^2 \pi_{ij}\pi_{kl} \tag{1}$$

subject to the regularity constraint that π has to be a probabilistic coupling between g and g':

$$\pi \in \mathcal{F}(g, g') = \{\pi \in \mathbb{R}^{n\times n'} | \pi \geq \mathbf{0}, \pi \mathbf{1}_{n'} = g, \pi^T \mathbf{1}_n = g'\}. \tag{2}$$

Here, $c : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ is an *expression cost function* that gives a non-negative dissimilarity score between the expression profiles of two spots over the same genes. $\mathbf{1}_n$ is an all-one vector of length n. Typically, g and g' are chosen to be uniform distributions over spots in each slice, although other distributions can be used [48].

The PASTE objective function F is composed of an expression similarity term (first summand) and a spatial similarity term (second summand) weighted by a parameter α. The first term, also called the Wasserstein distance in the OT literature [35], represents the cost of moving one unit of probability mass from each spot i to each spot j, with the cost being the gene expression dissimilarity between spots. The second term, also called the Gromov-Wasserstein distance [34, 36], approximately preserves the intra-slice spatial distances between spots. Together, the convex combination of the two terms in F is known as the Fused Gromov-Wasserstein (FGW) optimal transport objective [43].

The regularity condition (2) forces a rigid structure on π such that all spots from both slices must be aligned. However, such constraints may not be appropriate for ST slices with considerable differences in field of view or cell type composition due to both biological variation across tissue sections as well as differences caused by the manual nature of tissue dissection. Therefore, spots containing cell types or tissue regions that are unique to only one slice will be forced to be mapped to somehow arbitrary spots on the other slice.

Thus, in PASTE2, we propose to solve the *partial pairwise slice alignment problem* by minimizing the same objective function as PASTE (Equation (1)), but with a different set of constraints that allow for unmapped spots. Specifically, given a parameter $s \in [0, 1]$ describing the fraction of mass to transport between g and g', we define a set $\mathcal{P}(g, g', s)$ of *s-partial* couplings between distributions g and g' as

$$\mathcal{P}(g, g', s) = \{\pi \in \mathbb{R}^{n\times n'} | \pi \geq \mathbf{0}, \pi \mathbf{1}_{n'} \leq g, \pi^T \mathbf{1}_n \leq g', \mathbf{1}_n^T \pi \mathbf{1}_{n'} = s\}. \tag{3}$$

The parameter $s \in [0, 1]$ is interpreted as the overlap percentage between the two slices to align. The constraint $\mathbf{1}_n^T \pi \mathbf{1}_{n'} = s$ ensures that only the fraction of s probability mass is transported. Equivalently, if $g_i = \frac{1}{n}$ is a point mass for each spot, then roughly s fraction of the spots in each slice are aligned. The feasibility constraints $\pi \geq 0$, $\sum_j \pi_{ij} \leq g_i$ for all spots i in the first slice, and $\sum_i \pi_{ij} \leq g'_j$ for all spots j

in the second slice make sure that each spot only transport probability mass that it already has according to g and g', hence ensures π to be a valid transport plan. In PASTE2 we require that the map π belong to $\mathcal{P}(g, g', s)$, thus replacing the set $\mathcal{F}$ (defined in (2)) by the set $\mathcal{P}(g, g', s)$ (defined in (3)). In analogy to sequence alignment, PASTE calculates a global alignment, while PASTE2 calculates a local alignment.

The general concept of partial optimal transport [9] extends optimal transport theory to allow the transportation of only a specified fraction of mass between distributions. Here, we adapt the idea of partial optimal transport to the Fused Gromov-Wasserstein objective, hence the PASTE2 optimization problem is a novel partial Fused Gromov-Wasserstein (partial-FGW) optimal transport problem. While there are existing solutions to the partial Wasserstein and partial Gromov-Wasserstein problem, to the best of our knowledge, PASTE2 is the first to state and formulate the partial-Fused Gromov-Wasserstein problem and provide an optimization procedure to sovle this problem.

PASTE2 has two parameters: α, the balance between the gene expression dissimilarity and the spatial dissimilarity, and the overlap percentage parameter s indicating the fraction of mass to transport. Unless otherwise specified, we set $\alpha = 0.1$ following [48]. We choose the value of s using a model selection procedure described in Supplement §1. The choice of c, the expression dissimilarity function is described in Supplement §2.

2.2 An iterative conditional gradient algorithm for optimization

We derive an optimization algorithm to minimize the objective (1) subject to the constraint (3). This problem is a large scale (each slice contains thousands of spots) non-convex quadratic program with a convex and compact feasible region. Our algorithm is based on the Frank-Wolfe optimization algorithm [20], also known as the conditional gradient [28] algorithm. This algorithm has been widely adopted in the optimal transport community [11, 17, 19, 43] to compute transport plans because of its ability to handle large-scale quadratic programs [22]. The optimization problem in PASTE2 is thus particularly suitable for the conditional gradient algorithm.

The conditional gradient algorithm is an iterative first-order algorithm for constrained optimization. To fit in the conditional gradient scheme, we first write (1) in matrix form, following [36]

$$F(\pi) = (1-\alpha)\langle \mathbf{C}, \pi \rangle_F + \alpha \langle \mathbf{L}(D, D') \otimes \pi, \pi \rangle_F, \tag{4}$$

where $\mathbf{C} \in \mathbb{R}^{n \times n'}$ encodes the gene expression dissimilarity $c_{ij} = c(\mathbf{x}_{i\cdot}, \mathbf{x}'_{j\cdot})$ between each spot i in the first slice and each spot j in the second slice, and $\mathbf{L}(D, D') \in \mathbb{R}^{n \times n' \times n \times n'}$ is a 4-dimensional tensor defined by $\mathbf{L}_{i,j,k,l}(D, D') = (D_{ik} - D'_{jl})^2$. $\otimes$ is the tensor-matrix multiplication operator, *i.e.* $\mathbf{L} \otimes \pi$ is an $n \times n'$ matrix whose (i, j)-th element is $(\sum_{k,l} \mathbf{L}_{i,j,k,l} \cdot \pi_{k,l})$. $\langle \cdot, \cdot \rangle$ denotes the Frobenius dot product of matrices.

In each iteration, the algorithm moves in the direction that minimizes a linear approximation of the objective function while remaining in the feasible region. The mathematical details of the derivation of each step, as well as the pseudocode, is provided in Supplement §3.

2.3 Using histological image data in alignment

We further extend the PASTE2 partial-FGW framework to incorporate image information. Specifically, we replace the gene expression dissimilarity matrix $\mathbf{C} \in \mathbb{R}^{n \times n'}$ in Equation (4) by a sum of two $n \times n'$ dissimilarity matrices $\frac{1}{2}\mathbf{C}_{gene} + \frac{1}{2}\mathbf{C}_{image}$, where $\mathbf{C}_{gene}$ is the gene expression dissimilarity matrix as defined above and $\mathbf{C}_{image}$ encodes the dissimilarity between the image information at each spot. Thus, we seek a map π that minimizes the following objective function

$$F(\pi) = (1-\alpha)\langle \frac{1}{2}\mathbf{C}_{gene} + \frac{1}{2}\mathbf{C}_{image}, \pi \rangle_F + \alpha \langle \mathbf{L}(D, D') \otimes \pi, \pi \rangle_F \tag{5}$$

Note that to avoid an extra parameter we give equal weight $\frac{1}{2}$ to both gene expression and image information, although substituting other weights is straightforward. Also, since $\mathbf{C}_{gene}$ and $\mathbf{C}_{image}$ may not be on the same

scale, we scale $\mathbf{C}_{image}$ such that the maximum entry of $\mathbf{C}_{image}$ equals the maximum entry of $\mathbf{C}_{gene}$. In our implementation, we define $[\mathbf{C}_{image}]_{ij}$ to be the Euclidean distance between the mean RGB values of the spots. See Supplement §4 for further details.

2.4 3D reconstruction based on the partial alignment matrix

Given a series of consecutive, (partially) overlapping slices from the same tissue, we aim to reconstruct the spatial expression of the tissue in 3D by transforming PASTE2 partial pairwise alignments into a common coordinate system. Specifically, given a series of consecutive slices we first find partial alignments between adjacent slices by solving the partial pairwise slice alignment problem as above. To project all slices onto a common coordinate system, we extend the generalized weighted Procrustes analysis [25, 46] approach in [48] to sequentially project each pair of adjacent slices. While [48] projects a pair of slices onto the same coordinate system by centering both slices followed by calculating a rotation matrix, we derive the centering step for each slice separately to address the case where the alignment matrix π is partial and the aligned regions of the two slices have unique barycenters. The details of the projection are in Supplement §5.

3 Results

3.1 Evaluation on simulated ST data

We first compared PASTE2 and PASTE on a simulated ST dataset based on a human dorsolateral prefrontal cortex (DLPFC) tissue slice from [32]. Specifically, we extracted two partially overlapping subslices from a single DLPFC slice (sample 151674, corresponding to Slice 3B in §3.2) with varying overlap percentages $90\%, 70\%, 50\%, 30\%$ (Fig. 2a). To perturb the gene expression, we resample the gene expression profile of each spot in one of the subslices by sampling from a multinomial distribution with added pseudocount δ, which controls the noise level (Supplement §6). We vary the pseudocount δ in the range[2] from 0.1 to 2 with an increment of 0.1. In total, we generated $4 \times 30 = 120$ pairs of subslices with different overlap percentages and noise levels δ. For each pair of subslices, we ran PASTE2 with $\alpha = 0, 0.1, 1$ and using the ground truth value for the overlap fraction s (we evaluated model selection separately in Supplement §9), as well as full PASTE with default parameters ($\alpha = 0.1$). We evaluated the alignment using *Label Transfer Adjusted Rand Index* (*LTARI*). Given a labeling of cell type/spatial region of spots, LTARI measures how well the alignment preserves the label between the aligned spots. LTARI first defines a new spot labeling for the second slice by assigning to each aligned spot the label of the most likely corresponding spot in the first slice, then calculates the ARI between the induced spot labeling of the second slice and the ground truth labeling (Supplement §7). We used the manual cortical layer annotation from [32] as ground truth spot labeling (Fig. 2a).

We found that PASTE2 with the default parameter setting of $\alpha = 0.1$, which uses both gene expression information and spatial information, outperforms PASTE across most values of the added noise δ for every overlap percentage (Fig. 2b, Fig. S6). Specifically, for all four overlap percentages, PASTE2 ($\alpha = 0.1$) achieves the highest LTARI when $\delta < 2.0$, and achieves almost perfect LTARI when δ is small. Note that PASTE obtains constant accuracy because it aligns overlapping regions well but non-overlapping regions arbitrarily. The gap in accuracy between PASTE2 and PASTE is larger when the overlap is smaller. This indicates that PASTE, which finds an alignment between all pairs of spots, is not suitable for the partial alignment task. In contrast, PASTE2 has high accuracy in partial alignment across a wide range of overlap percentages and gene expression noises. Moreover, PASTE2 achieves near perfect LTARI when the added pseudocount is in the range of variability in read counts ($\approx 0.1 - 0.2$) observed in real data [48].

To investigate the effect of the misspecification of the value of the overlap percentage parameter s on the result of PASTE2, we ran PASTE2 with s ranging from 0.1 to 1, with a step size of 0.1, on a simulated pair where the ground truth overlap percentage is 50% and the added pseudocount is 0.1. We found that PASTE2

[2]With the typical sequence coverage and data sparsity in ST data, $\delta > 2.0$ (adding > 2 counts to each transcript) is a strong perturbation of the data that essentially destroys the signal present in the original data.

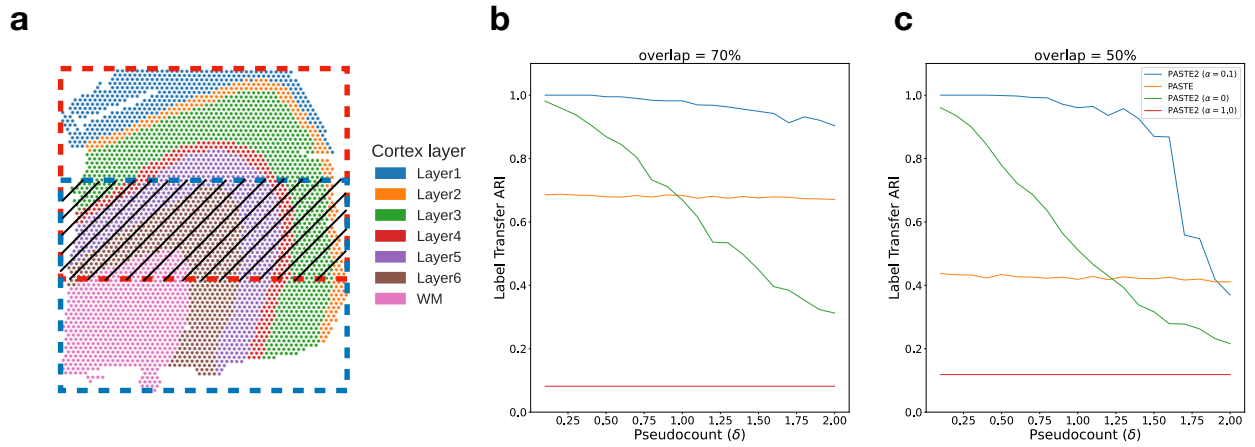

Figure 2: Comparison of PASTE2 and PASTE on simulated partially overlapping subslices. a, DLPFC slice 151674 with spots colored according to the manual annotations of cortical layers from [32]. Red box and blue box indicate two partially overlapping subslices, with central overlapping region containing some fraction of spots from each slice. **b**, Label Transfer ARI of the alignments produced by PASTE2 with $\alpha = 0$ (gene expression information only), PASTE2 $\alpha = 1$ (spatial information only), PASTE2 $\alpha = 0.1$ (both), and PASTE (full alignment, $\alpha = 0.1$) as a function of the pseudocount (δ) for overlap percentage70%. **c**, Label Transfer ARI for overlap percentage 50%.

aligns correctly when s is lower than the ground truth, while the performance degrades for larger values of s (Fig. S7). This is expected because with an overestimation of the overlap percentage, the PASTE2 alignment becomes more similar to the PASTE alignment which includes all the spots. Thus, in selecting a value for s, it is preferable to use a model selection procedure that slightly underestimates s rather than overestimate s. We propose a heuristic for selecting s in Supplement §1.

Finally, we emphasize the importance of using both gene expression and spatial information in computing accurate partial alignments. PASTE2 ($\alpha = 1.0$) has consistently low LTARI, indicating that using spatial coordinates alone cannot recover alignment across slices. The performance of PASTE2 ($\alpha = 0$), which only uses gene expression information of each spot for alignment, drops more quickly than PASTE2 ($\alpha = 0.1$) with increasing pseudocount δ indicating more noise in gene expression. Using both gene expression and spatial information, PASTE2 is able to accurately align two partially overlapping ST slices. The effect of different intermediate values of α on the alignment performance is thoroughly discussed in [48].

3.2 Human dorsolateral prefrontal cortex (DLPFC) slices

We next compared PASTE2 to PASTE [48] and two other transcriptomics alignment methods – Pamona [10] and Tangram [5] – on the full human dorsolateral prefrontal cortex (DLPFC) dataset containing 10X Genomics Visium ST data from three individuals (labeled sample 1, 2, 3) with four slices (labeled slice A, B, C, D) per individual [32]. For each individual, slices A and B and slices C and D are 10 μm apart. However, slices B and C are further apart at a distance of 300 μm. Hence, slice pairs AB and CD are more similar to each other than slice pair BC. Note that Pamona [10], a manifold alignment algorithm for multi-omics datasets, is also based on partial optimal transport, while Tangram [5] is a deep-learning based method that aligns scRNA-seq data onto ST data. We create partial ST alignment problems by generating two partially overlapping DLPFC datasets as follows. For each individual, we extracted the left portions of slice A and C, and the right portions of slice B and D such that the extracted pairs AB, BC, and CD have $\approx 70\%$ overlap in area (Fig. 3a). We also created another set of partially overlapping dataset using horizontal slices (Fig. S8a). We ran each of the methods as described in Supplement §8. We evaluate the accuracy of each method by computing the Label Transfer ARI (LTARI) as previously described (Section 3.1).

We find that PASTE2 achieves the highest LTARI on all adjacent subslices of all individuals, for both vertical partial slices and horizontal partial slices, with the exception of one pair (Fig. 3b, Fig. S8b). About

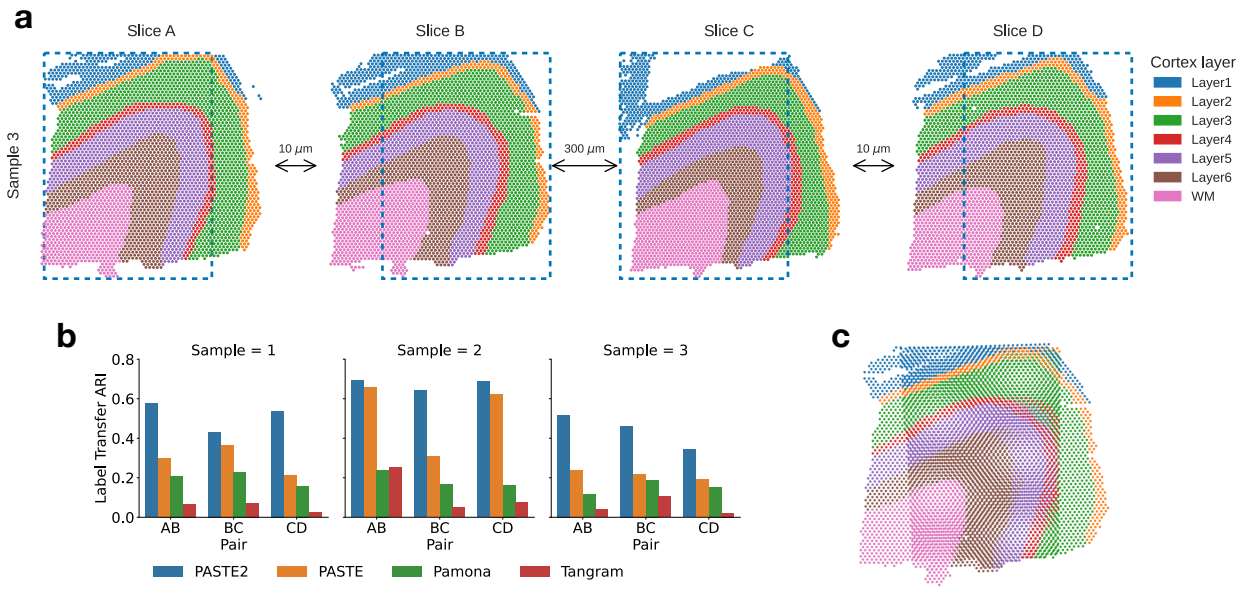

Figure 3: Comparison of alignment methods on partially overlapping DLPFC slices. a, Vertical subslices were obtained by cropping subslices (blue dotted boxes) from four adjacent slices from DLPFC sample 3, with indicated distances between adjacent slices. Each pair of adjacent subslices overlaps in 70% of their areas. **b**, LTARI of pairwise alignments computed by PASTE2, PASTE, Pamona, and Tangram for each pair of adjacent vertical subslices from three samples. **c**, Optimal projection of vertical subslices from slice AB of sample 3 onto the same 2D coordinate system using the PASTE2 partial alignment.

70% - 75% spots from each subslice is aligned in each pair indicating the parameter s corresponds well with slice overlap. For most pairs, PASTE2 has more than twice the LTARI than all other methods, demonstrating PASTE2's ability to identify the overlap region of the two ST slices and align the overlap region reliably. On one pair Pamona has slightly higher LTARI than PASTE2 (Fig. S8b), but all methods have very low LTARI (< 0.1), suggesting that this pair has low spatial coherence. PASTE is the second-highest performing method on most pairs, indicating that even though PASTE does not model partially overlapping slices, it is still more suitable for aligning spatial transcriptomics data than methods designed for different purposes. While Pamona is designed to align datasets with both shared and dataset-specific cells [10] – the analog of the partial pairwise slice alignment problem for single cell datasets – Pamona does not model spatial constraints, perhaps explaining its lower performance. Tangram assumes the single-cell gene expression dataset and the spatial dataset come from the same anatomical region [5], hence the partial slice alignment task violates the Tangram assumption, leading to a low alignment accuracy.

For a more intuitive demonstration of PASTE2's advantage and accuracy, we projected the vertical subslices of sample 3 pair AB onto the same coordinate system, computed as described in §2.4 based on the alignment matrix computed by PASTE2 (Fig. 3c), as well as the optimal projection of the same pair based on the alignment computed by PASTE (Fig. S9a). Qualitatively, the projection of PASTE2 correctly stacks the overlap area of the two slices, with spots from the same cortical layer stacking on top of each other, while PASTE fails to find the corresponding layers in the two slices. Additionally, PASTE2 correctly identifies and aligns the overlap area of all four partial slices of an individual while leaving the rest unaligned, leading to a visually correct 3D reconstruction of the tissue from partial slices (Fig. S9b).

We also ran STUtility [4], a method to align H&E stained images that are generated as part of the 10X Genomics Visium ST workflow. STUtility outputs new coordinates of the aligned slices and does not produce a mapping between pairs of spots; thus, we visualized the alignment results by plotting each pair of partial subslices according to the new coordinates output by STUtility (Fig. S10). The image masking function utilized by STUtility failed for the partial slices of sample 3, so we only visualized the results for sample 1

and 2. STUtility correctly identifies that each pair of input slices are partially overlapping, but it does not align the correct overlapping region, and the output alignment seems quite arbitrary. These results might be because STUtlitiy aligns images by identifying and finding correspondences between edges of the two input tissues, but when two tissues are partially overlapping, the edges do not provide information about spot correspondences. On the other hand, PASTE2 correctly aligns the overlapping region (Fig. 3c). This demonstrates that using transcriptomic similarity, spatial similarity, and image information yields better partial alignments than H&E images alone.

Finally, we compared PASTE2's running time with other methods on the vertical subslices of sample 3. PASTE2 finished in under 10 minutes for all subslice pairs on a Macbook Pro with 2.4GHz Intel Core i5 CPU, with most of the running time spent on the GLM-PCA subroutine (Fig. S11). The conditional gradient optimizer in PASTE2 runs in less than half of the time of Pamona and Tangram, and only runs slightly slower than PASTE. We also used the DLPFC datasets to evaluate the accuracy of PASTE2's model selection procedure for estimating the overlap percentage s, and found that PASTE2 correctly estimates the overlap percentage in many scenarios (Supplement §9).

3.3 Incorporating histology information improves alignment

We compared PASTE2's alignment performance when using both gene expression and histological image (Equation 5) versus using only gene expression data (Equation 4). Note that spatial information is included in both analyses. We ran the two modes of PASTE2 on pairs of horizontal and vertical subslices from DLPFC sample 3. We found that using the histological image substantially improved the alignment performance for pair CD (Fig. 4ab), increasing the LTARI from 0.34 to 0.46. Examining the alignment obtained on this pair using only gene expression information (Fig. 4c) to the alignment obtained with both gene expression and the histology image (Fig. 4d), we observe that the alignment obtained using the images is more spatially contiguous. In particular, there is a curve of unaligned spots (blue spots in Figure 4c) in subslice D inside the yellow region. This curve corresponds to spots that are manually annotated as Layer 6 of the DLPFC [32]. Interestingly, the spots in this layer have lower total UMI counts than other layers: the mean total UMI counts for spots from Layer 6 is 2915 compared to total UMI counts ≈ 4500 in the other layers. This suggests that the gene expression signal is weaker in these spots. In contrast, the PASTE2 alignment obtained using both gene expression and image information (Figure 4d) does not have the same curve of unaligned spots, demonstrating the advantages of using the histological image for spots with a weak gene expression signal.

In the horizontal slices, we see cases where using the image information reduces the alignment performance. For example, the LTARI drops from 0.56 (expression) to 0.50 (expression and image) for horizontal subslices of pair AB (Fig. S14a,b). The aligned part of subslice B in Fig. S14d shows that many spots are left unaligned in the actual overlap region, and there is a clear stripe of unaligned blue spots towards the left part of the subslice. Looking at the H&E image of subslice B in Fig. S14b, we see a clear dark stain on the left of the subslice that is missing from the image of subslice A, at exactly the same location of the unaligned stripe. This indicates that the stain on the H&E image is the cause for the worse alignment performance.

For the other pairs, the LTARI for PASTE2 alignments with and without images is approximately the same. This is not too surprising since the H&E images of DLPFC slices do not display strong heterogeneity across different layers (Fig. 4b, Fig. S14b). However, the fact that utilizing image information corrects the alignment of low UMI spots demonstrates the potential for histological images to guide PASTE2 alignment. The image information can help overcome the sparsity of gene expression, and when the histological images have greater variation across spots, using the images should further improve the alignment quality by complementing the gene expression signal.

3.4 Spatial transcriptomics of Drosophila embryo

We applied PASTE2 to analyze a Stereo-seq dataset from a Drosophila embryo [12]. Stereo-seq is a new SRT technology with $\approx$ 500nm resolution, two orders of magnitude smaller than the 10X Visium platform,

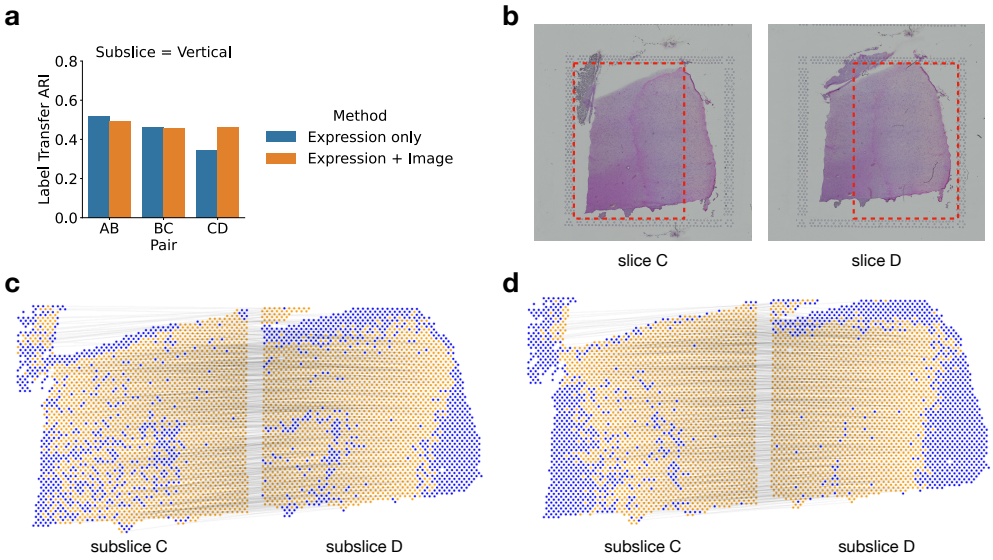

Figure 4: Evaluating the benefit of using histological image information in PASTE2 alignment. **a**, The Label Transfer ARI (LTARI) of PASTE2 partial alignments of pairs of vertical subslices extracted from DLPFC sample 3 using only gene expression (blue) and using both gene expression and image information (orange). **b**, Histological images of sample 3 slice C and slice D. The red boxes bound the vertical subslices extracted for partial alignment. The right part of subslice C should be aligned to the left part of subslice D. **c**, Visualization of PASTE2 alignment of the subslice pair CD using gene expression and spatial information. Yellow spots are aligned by PASTE2, while blue spots are unaligned. Thin black lines connect pairs of spots that are aligned by PASTE2 with high weight. **d**, Visualization of PASTE2 alignment of the same subslice pair when gene expression, histological image, and spatial information are all used.

but with lower UMIs per spot. [47] applied Stereo-seq to two late-stage Drosophila embryos 14-16 hours and 16-18 hours after egg laying (labeled E14-16 and E16-18) and three stages of larvae (labeled L1-L3). Each slice has ≈ 1000 spots with median UMI per spot of ≈ 2000, compared to ≈ 4000 spots and ≈ 5000 median UMI per spot in the 10X Visium DLPFC dataset. In the published analysis, the cell type of each spot was derived by unsupervised clustering of gene expression followed by annotation based on marker genes. The publication used PASTE to align all slices from the same stage and obtain a 3D map of spatial expression of each stage. However, slices from the same stage vary in size and cell type compositions and do not fully overlap in space. For example, inspection of annotated cell types shows that adjacent slices from the E14-16 sample do not fully overlap (Fig. S15). Therefore, it is appropriate to use PASTE2 to realign the adjacent slices respecting the different composition of cell types across slices, and to obtain a more accurate 3D reconstruction of the Drosophila embryo.

We applied PASTE2 to compute a partial alignment for each pair of 16 adjacent slices from the E14-16 sample, estimating the overlap percentage using the PASTE2 model-selection heuristic (Supplement §1). Slices 7 and 8 have clear differences in the composition of cell types annotated by [47], with the carcass cells showing the largest difference in proportion (Fig. 5a). PASTE2 addresses this imbalance by aligning a similar proportion of carcass cells across slices, leaving the excess cells in slice 8 unaligned. The spots from the two slices included in the PASTE2 partial alignment show similar spatial organization (Figure 5b) and cell type composition (Fig. S16b). For example, the proportions of carcass cells in slices 7 and 8 differ by 10% before alignment (Fig. S16a), but after alignment the difference is less than 3% (Fig. S16b). The differences in proportions shrinks for salivary gland cells as well, indicating PASTE2 correctly identifies and aligns the overlapping parts. PASTE2 optimal projection of the two slices to the same coordinate system puts slice 8 slightly higher in y coordinates than slice 7, consistent with the observation that slice 8 has unaligned carcass cells at the top (Figure 5c). The LTARI obtained by PASTE2 for this pair is 0.49, compared to a LTARI of

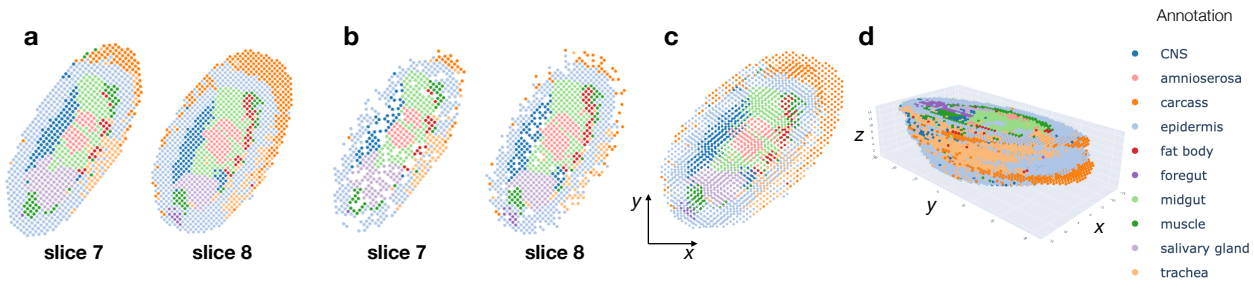

Figure 5: PASTE2 alignment of Stereo-seq data from E14-16 Drosophila embryo from [47]. **a**, Stereo-seq slices 7 and 8 with spots labeled by cell types annotated in [47]. **b**, Spots from slices 7 and 8 that are included in the partial alignment computed by PASTE2. Spots selected by PASTE2 have similar proportions of cell types and spatial locations. **c**, Optimal projection of slices 7 and slice 8 onto the same 2D coordinate system using the PASTE2 partial alignment. **d**, PASTE2 3D reconstruction using all 16 slices from the Drosophila embryo.

0.39 for PASTE, again showing the advantages of partial alignment. Examination of pair of adjacent slices 14 and 15 shows a similar advantage of partial alignment. Slice 15 has a stripe of carcass cells that is absent in slice 14 (Fig. S17a). PASTE2 leaves the stripe unaligned across slices (Fig. S17bc), increasing the LTARI from 0.29 for PASTE to 0.52 for PASTE2. Since PASTE computes an alignment for all spots, the extra carcass cells in slice 15 are mapped somewhere on slice 14, creating false correspondences between spots (Fig. S18).

We compared the LTARI of the PASTE2 alignment with the LTARI scores of PASTE, Pamona, and Tangram on every pair of adjacent slices. PASTE2 achieves the highest LTARI for most pairs, with the largest gain in pairs where the two slices have different compositions of cell types, such as slice 14 and 15 (Fig. S19). Pairs where PASTE2 does not obtain the highest LTARI, such as slice 2 and 3, have relative similar sizes and cell types, and PASTE2 still achieves comparable LTARI with the highest performing method. This indicates that PASTE2 not only aligns partially overlapping slices correctly, but also performs well on pairs of similar slices.

We used PASTE2 to generate a 3D reconstruction of all 16 slices of the E14-16 Drosophila embryo, where adjacent slices have on average 70% of overlapping spots (Fig. 5d). The PASTE2 3D reconstruction will be useful for refining the analyses presented in [47] who demonstrated that the PASTE-generated 3D expression helped detect functional subregions and uncover the dynamics of cell state changes and tissue-specific gene regulation.

4 Discussion

We present PASTE2, a method to perform pairwise alignment and 3D reconstruction of multi-slice spatial transcriptomics data. PASTE2 addresses the important situation where slices partially overlap in space or have different cell type compositions, which is the case for most real datasets. We formulate the ST partial pairwise alignment problem using a partial Fused Gromov-Wasserstein optimal transport framework and derive an optimization algorithm to solve this problem. We further design a model selection procedure to determine the overlap between slices, and extend the framework to incorporate both gene expression and imaging information.

We found that PASTE2 outperforms multiple other methods for alignment of spatial transcriptomics or single cell data including PASTE, Pamona, Tangram, and STUtility. We show that PASTE2's use of histology images can further improve alignments, although the results are variable depending on the quality of the images. We expect that PASTE2 will achieve much higher accuracy incorporating image information in datasets where histological images display stronger signal across spots – in preliminary results on unpublished cancer datasets with high-quality H&E images we observed even larger gains. Finally, we demonstrate PASTE2's capabilities on larger datasets from another SRT technology by generating a 3D reconstruction of

a Drosophila embryo from 16 slices of Stereo-seq data.

There are multiple directions for future work. First, is to extend the partial alignment framework to integrate multiple slices into a single consensus slice to address the data sparsity issue by pooling counts from corresponding spots [48]. Second, one could stitch together multiple partially overlapping slices into a larger 2D slice. This stitching would be helpful in cases where adjacent tissue slices are close in the z-coordinate which is often the case with thin tissue slices ($\approx 10\mu m$). In addition, one could incorporate additional spatial regularization terms to enforce more contiguous overlapping regions. Third, it would be interesting to apply PASTE2 to integrated spatial transcriptomics and imaging data from other platforms such as Slide-seq [37, 40], or combined Stero-seq and imaging data which [47] noted as a future technology development. Finally, it would be interesting to examine the effectiveness of other optimal transport frameworks such as unbalanced OT [38] that impose soft constraints rather than hard constraints on partial alignments.

We anticipate that PASTE2 will be a useful tool for integrating transcriptomic information across multi-slice ST datasets and for building 3D tissue atlas across both normal and diseased tissues, such as in the Human Tumor Atlas Network and related projects.

Acknowledgements

This work is supported by grants U24CA211000 and U24CA248453 from the US National Cancer Institute (NCI).

References

[1] Spatial gene expression - 10x genomics. https://www.10xgenomics.com/products/spatial-gene-expression. Accessed: October 2022.

[2] Chiara Baccin, Jude Al-Sabah, Lars Velten, Patrick M Helbling, Florian Grünschläger, Pablo Hernández-Malmierca, César Nombela-Arrieta, Lars M Steinmetz, Andreas Trumpp, and Simon Haas. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nature cell biology*, 22(1):38–48, 2020.

[3] Richard Beare, Bradley Lowekamp, and Ziv Yaniv. Image segmentation, registration and characterization in r with simpleitk. *Journal of statistical software*, 86, 2018.

[4] Joseph Bergenstråhle, Ludvig Larsson, and Joakim Lundeberg. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC genomics*, 21(1):1–7, 2020.

[5] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Åsa Segerstolpe, Meng Zhang, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature methods*, 18(11):1352–1362, 2021.

[6] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.

[7] Matthew Brett, Kalina Christoff, Rhodri Cusack, Jack Lancaster, et al. Using the talairach atlas with the mni template. *Neuroimage*, 13(6):85–85, 2001.

[8] Dylan M Cable, Evan Murray, Luli S Zou, Aleksandrina Goeva, Evan Z Macosko, Fei Chen, and Rafael A Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526, 2022.

[9] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, pages 673–730, 2010.

[10] Kai Cao, Yiguang Hong, and Lin Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics*, 38(1):211–219, 2022.

[11] Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2903–2913. Curran Associates, Inc., 2020.

[12] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.

[13] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

[14] Wei-Ting Chen, Ashley Lu, Katleen Craessaerts, Benjamin Pavie, Carlo Sala Frigerio, Nikky Corthout, Xiaoyan Qian, Jana Laláková, Malte Kühnemund, Iryna Voytyuk, et al. Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182(4):976–991, 2020.

[15] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Single-cell multiomics integration by scot. *Journal of Computational Biology*, 29(1):19–22, 2022.

[16] Pınar Demetçi, Rebecca Santorella, Björn Sandstede, and Ritambhara Singh. Unsupervised integration of single-cell multi-omics datasets with disproportionate cell-type representation. In *International Conference on Research in Computational Molecular Biology*, pages 3–19. Springer, 2022.

[17] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[18] 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.

[19] Rémi Flamary, Nicolas Courty, Alain Rakotomamonjy, and Devis Tuia. Optimal transport with laplacian regularization. In *NIPS 2014, Workshop on Optimal Transport and Machine Learning*, 2014.

[20] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[21] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):1–15, 2019.

[22] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

[23] Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstråhle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514, 2020.

[24] Andrew Jones, F William Townes, Didong Li, and Barbara E Engelhardt. Alignment of spatial genomics and histology data using deep gaussian processes. *bioRxiv*, 2022.

[25] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[26] Arno Klein, Jesper Andersson, Babak A Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E Christensen, D Louis Collins, James Gee, Pierre Hellier, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *Neuroimage*, 46(3):786–802, 2009.

[27] Jack L Lancaster, Marty G Woldorff, Lawrence M Parsons, Mario Liotti, Catarina S Freitas, Lacy Rainey, Peter V Kochunov, Dan Nickerson, Shawn A Mikiten, and Peter T Fox. Automated talairach atlas labels for functional brain mapping. *Human brain mapping*, 10(3):120–131, 2000.

[28] Evgeny S Levitin and Boris T Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.

[29] Yingxin Lin and Jean YH Yang. 3d reconstruction of spatial expression. *Nature Methods*, 19(5):526–527, 2022.

[30] Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ rna profiling by sequential hybridization. *Nature methods*, 11(4):360–361, 2014.

[31] Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Medical image registration using mutual information. *Proceedings of the IEEE*, 91(10):1699–1722, 2003.

[32] Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.

[33] Matthew McCormick, Xiaoxiao Liu, Julien Jomier, Charles Marion, and Luis Ibanez. Itk: enabling reproducible research and open science. *Frontiers in neuroinformatics*, 8:13, 2014.

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

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

[36] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

[37] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[38] Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

[39] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[40] Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319, 2021.

[41] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[42] Kim Thrane, Hanna Eriksson, Jonas Maaskola, Johan Hansson, and Joakim Lundeberg. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage iii cutaneous malignant melanoma. *Cancer research*, 78(20):5970–5979, 2018.

[43] Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR, 2019.

[44] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20(1):1–16, 2019.

[45] Susanne C van den Brink, Anna Alemany, Vincent van Batenburg, Naomi Moris, Marloes Blotenburg, Judith Vivie, Peter Baillie-Johnson, Jennifer Nichols, Katharina F Sonnen, Alfonso Martinez Arias, et al. Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids. *Nature*, 582(7812):405–409, 2020.

[46] Grace Wahba. A least squares estimate of satellite attitude. *SIAM review*, 7(3):409–409, 1965.

[47] Mingyue Wang, Qinan Hu, Tianhang Lv, Yuhang Wang, Qing Lan, Rong Xiang, Zhencheng Tu, Yanrong Wei, Kai Han, Chang Shi, et al. High-resolution 3d spatiotemporal transcriptomic maps of developing drosophila embryos and larvae. *Developmental Cell*, 57(10):1271–1283, 2022.

[48] Ron Zeira, Max Land, Alexander Strzalkowski, and Benjamin J Raphael. Alignment and integration of spatial transcriptomics data. *Nature Methods*, 19(5):567–575, 2022.