

Optimal Transport Methods for Analyzing Highly Multiplexed Spatial Proteomic Data



Joseph Boen
The Queen's College
University of Oxford

A thesis for the degree of
Master of Science
Trinity 2023

Acknowledgements

I would like to thank Professors Helen Byrne, Simon Leedham, and Eric O'Neill for their mentorship, and my colleagues Joshua Bull, Eoghan Mulholland, and George Atkinson for their help and collaboration. I am grateful for the support provided by the Oxford University Clarendon Fund and The Queen's College Oxford Graduate Scholarship. Most of all, thank you to Mama, Papa, Nathaniel, and Hannah. None of this would have been possible without your love and support.

Abstract

Cell microenvironments are complex niches of interacting cells that drive tissue function and disease response. In recent years, *in situ* multiplexed imaging has enabled the multi-molecular profiling of single cells while preserving their spatial context. Although these new technologies capture cell phenotypes and spatial organization at unprecedented resolution, the technology has many trade-offs, and the analysis of the rich spatial information they provide remains a challenge. In this thesis, we develop novel computational methods tailored for multiplex imaging data leveraging principles from Optimal Transport (OT), a branch of statistics for comparing probability distributions.

In the first part, we address the challenge of integrating spatial profiles from different tissues, a task essential to enhancing the information obtained from multiplex experiments. We introduce a novel algorithm for optimizing tissue alignment derived from partial OT, and additionally propose an unbiased statistical estimator for OT distances that scales to massive tissues at a fraction of the cost. We benchmark our method against state-of-the-art algorithms, and demonstrate our method’s improved robustness to rips, tears, and other challenging spatial distortions induced by the multiplexing process.

In the second part, we consider the task of quantitatively characterizing cell microenvironments. We introduce a novel framework that captures both microenvironment cell composition and spatial organization using graph OT, and demonstrate its applications in identifying cell-cell interactions, spatial domains, and disease relevant microenvironments on simulated and real datasets. On pathologist annotated multiplexed lung tissues, we demonstrate that our methods recover microenvironments directly linked to known cell interactions and functions. On multiplexed head neck tumors, we identify microenvironments that are highly predictive of patient outcome. We benchmark our method against two state-of-the-art competitors and various baselines, and show that our method outperforms all comparisons on each dataset.

Lastly, we present a comprehensive case study that showcases the practical application of both our methods for spatial integration and microenvironment characterization on a novel dataset of multiplexed mouse colorectal cancer tissues. We align massive tissues with hundreds of thousands of cells, and identify distinct, condition-specific microenvironment subtypes that correspond to unique cell enrichment profiles and spatial domains. Together, our methods provide a powerful approach to mapping the complex landscape of the tumor microenvironment in colorectal cancer.

Contents

1	Introduction	1
1.1	<i>In situ</i> spatial profiling of tissues	2
1.2	Computational Challenges and Opportunities	5
1.2.1	Spatial Integration	6
1.2.2	Microenvironment Characterization	7
1.3	Optimal Transport for Multiplexed Imaging	8
1.3.1	Motivation	8
1.3.2	Optimal Transport for Spatial Data	11
1.4	Mathematical Foundations	13
1.4.1	Assignment and The Monge Problem	13
1.4.2	Coupling and the Kantorovich Relaxation	15
1.4.3	Computation	17
1.4.4	Theoretical Guarantees	18
1.5	Contributions of this Thesis	20
2	Alignment and Integration of Spatially Resolved Proteomics Data with Partial Optimal Transport	23
2.1	Introduction	23
2.2	Point Set Registration	25
2.2.1	Mathematical Formulation	25
2.2.2	Geometric Transformations	26
2.2.3	Solving the Point Set Registration Problem	27
2.3	Related Work	28
2.4	Optimal Transport for Point Set Registration	30
2.4.1	Partial Optimal Transport	31
2.4.2	Optimizing POT for Point Set Registration	33
2.4.3	Improving Computational Efficiency	36
2.5	Results	38

2.5.1	The POT Loss Quantifies Alignment Quality	38
2.5.2	The Quantized Estimator Enables Fast and Accurate Alignment	41
2.5.3	POT Alignment Outperforms ICP and CPD	45
2.6	Conclusion	46
3	Characterizing Cell Microenvironments with Graph Optimal Transport	48
3.1	Introduction	48
3.2	Related Work	50
3.2.1	Spatial LDA	50
3.2.1.1	Summary	50
3.2.1.2	Method	51
3.2.1.3	Analysis	52
3.2.2	UTAG	53
3.2.2.1	Summary	53
3.2.2.2	Method	53
3.2.2.3	Analysis	55
3.3	Characterizing Microenvironments with Graph Optimal Transport . .	55
3.3.1	Microenvironments as Spatial Graphs	56
3.3.2	Comparing Microenvironments with the Fused Gromov-Wasserstein Distance	59
3.3.3	Analyzing Microenvironments with Graph-OT Distances . . .	63
3.4	Results	67
3.4.1	Graph-OT quantifies changes in microenvironment organization and composition	67
3.4.2	Graph-OT Microenvironments Reveal Tissue Architectures in the Healthy Human Lung	72
3.4.3	Graph-OT microenvironments are generalizable, and predictive of patient outcome in HNSCC	78
3.5	Conclusion and Discussion	86
4	Case Study: Microenvironment Remodeling of Colorectal Cancer through TGF-β Inhibition	88
4.1	Introduction	88
4.2	Methods	91
4.3	Results	91
4.3.1	Spatial Integration of Immune and Fibroblast Panels	91

4.3.2	Graph Optimal Transport Identifies Highly Stable Microenvironment Clusters	93
4.3.3	TGF- β Remodels Tumor Microenvironments in AKPT CRC tissues	98
4.4	Discussion and Conclusion	102
5	Conclusion and Outlook	104
5.1	Review of Contributions	105
5.2	Limitations	106
5.3	Future Work	107
A	Additional Methods	110
A.1	Software Implementations	110
A.2	Data Availability	111
A.3	Mouse Multiplexing and Image Preprocessing	111
B	Additional Figures	113
	Bibliography	115

List of Figures

1.1	Multiplex Imaging and Analysis Pipeline. In the top panel, tissues are measured for various protein markers. In the second panel, cell segmentation and cell type assignment convert raw, multi-channel images into spatial point clouds of cell types. The images are from a $1\text{ mm} \times 1\text{ mm}$ patch from a multiplex IHC image of murine colorectal cancer. In the third panel, different spatial profiles are integrated for joint analysis. In the fourth, microenvironments and spatial domains are identified and characterized.	4
1.2	Illustrative Example. In the left panel, cell positions on a 1D lattice are shown. In the middle panel, the cell locations are converted into a probability distribution, and the optimal plan for transporting the red to the blue distribution is shown. In the right panel, the optimal transport distance calculation is shown.	11
1.3	Optimal Transport Co-Localization. Kernel Density Estimates were applied to cell locations from three different cell types to generate spatial probability distributions, where brighter colors indicate higher likelihoods of observing a particular cell type at a given spatial location. By visual inspection, we surmise that cell types A and C are more closely co-localized than they are with cell type B. Optimal Transport quantifies this intuition, and confirms that $OT(A, C) < OT(B, C)$. . .	12
1.4	Comparing Statistical Distance Metrics. We compute the statistical distance between two probability measures defined on domain $\mathbb{R}$. While all the distance functions presented are metrics, only the 2-Wasserstein respects the geometry of the domain. The distributions were parameterized as $p(x) \sim \mathcal{N}(\mu, \sigma^2)$, $q(x \tau) \sim \mathcal{N}(\mu - \tau, \sigma^2)$. For each metric, the range of $d(p(x), q(x \tau))$ over $\tau \in [-15, 15]$ was $[0, 1]$ normalized for ease of comparison.	20

2.1	Spatial Integration. Two thin tissue slices (orange, blue) are sectioned from the same tissue. One slice is stained to detect one set of protein markers, and the second slice another. Due to technical variation in the sample collection process, there are rips and holes in the blue slice not present in the orange slice and vice versa. Spatial integration is used to align the blue and orange tissue slices, and allow for joint analysis of both the blue and orange cell populations in an integrated tissue.	24
2.2	Partial Optimal Transport Alignment. a) Given two synthetic tissue slices modeled as point clouds, the integration task is equivalent to transforming Panel 1’s initial position to the ground truth. b) Using Algorithm 1, we learn the transformation parameters that minimize the POT distance between point clouds. The POT distance accurately quantifies alignment quality, and gradient descent converges to the true parameters. c) The POT distance between Panel 1 and (transformed) Panel 2 is shown across different transformation parameters. In comparison, the Mean Squared Error (MSE) between the (normalized) transformed Panel 2 and its (normalized) ground truth position is also shown. The POT distance is highly correlated with the true alignment error (MSE), even though the true correspondence is unknown. . . .	39
2.3	Estimator Comparison. a) The leftmost panel shows the full tissue consisting of over 100,000 cells, with spatial locations colored by relative cell density. The next two panels show the plug in and quantized representations of this distribution with 100 samples, with points colored by weight. b) For the same number of samples, ($k = 100$), we compare the relative error between the analytic POT distance, and the approximate distance estimated by plug in and quantized sampling. While both are reasonably unbiased, the quantized estimator has a greatly reduced variance. c) We compute the absolute relative error as a function of sample size. The quantized estimator is always more accurate, and converges slightly faster than the plug in.	41

2.4	Comparison of Estimator Strategies on Spatial Alignment. In this plot, “Full” refers to Algorithm 1 using a gradient computed between complete point clouds, while “Stochastic” and “Quantized” refer to gradients calculated using the plug in and quantized estimators respectively. (a) Tissue alignment results for different estimators after 500 iterations, compared with the ground truth and original starting positions. POT Loss convergence as a function of iterations (b) and time (c) for different strategies. Since the Full POT loss computation is deterministic, its optimization trajectory is noiseless. In contrast, the plug in and quantized estimators are stochastic, and yield noisy optimization trajectories. The reduced noiseness of the quantized trajectory in comparison to the stochastic trajectory can be explained by the reduced variance of the quantized estimator shown in Figure 2.3 b).	43
2.5	Methods Comparison. a) We compare the alignment quality between ICP, CPD, and our method (OT) under increasingly smaller overlaps. As before, the MSE is computed between the final position of the normalized moving point cloud and its normalized ground truth position. b) Alignment quality for a 60% overlap. The orange cells represent the moving point cloud.	46
3.1	Overview. (a) We first construct a weighted spatial graph from the original cell centers using Voronoi tessellations followed by Delauney triangulations. (b) From this tissue graph, microenvironments are defined as induced subgraphs. A set of subgraphs can then be compared using graph optimal transport to form a distance matrix. (c) This distance matrix can clustered to group microenvironments into canonical subtypes, which can then investigated for various biological applications, such as cell type enrichment, archetypical cell organization, or spatial domain mapping.	57

3.2	Microenvironment Simulation. (a) Microenvironment subgraphs from High and Low Green cell compositions with fixed, exclusionary Interactions. (b) Graph Optimal Transport between microenvironments correctly distinguishes between states, as exhibited in the block anti-diagonal distance matrix structure, and confirmed by the Composition distance matrix. The Average Nearest Neighbor Distance (ANND) statistic is not designed to detect changes in composition, and as such displays a randomized structure. (c) Microenvironment subgraphs with High and Low Green-Grey cell Interactions under fixed, High Green cell Composition. (d) Once again, Graph Optimal Transport between microenvironments correctly distinguishes between states as confirmed by the ANND matrix. Because the composition is fixed, the composition distance matrix is uniform.	71
3.3	Mapping Human Lung Tissue Architecture from multiplex IMC data. (a) comparison between pathologist annotations and spatially mapped, graph-OT microenvironments. (b) Comparing spatial domain predictions from graph-OT and state-of-the-art methods for a single tissue. Cell type enrichment of graph-OT microenvironments over $n = 18$ tissues. Benchmark comparisons over $n = 18$ tissues evaluated by (d) accuracy and (e) homogeneity scores. Significance bars indicate $p < 0.001$ as computed by a Bonferroni corrected two-sided Mann Whitney U Test.	74
3.4	Identifying clinically relevant microenvironments in HNSCC tumors. In biomarker discovery, microenvironments are characterized and associated with NED or non-NED outcomes on labeled patients. Given tissue from an unseen patient (a), similarity scores are computed between biomarkers and microenvironments (b). These scores are aggregated and used as features for a logistic regression classifier, with performance evaluated in (c). In (d), classifier weights are used to isolate the contribution of each microenvironment cluster to the overall patient prognosis. Cell enrichment profiles are provided to summarize microenvironment cluster compositions. The weights associated with the red triangle and blue pentagon biomarkers indicate that their presence strongly contributes to non-NED and NED prognosis respectively.	81

4.1	Experimental Outline. Tumor organoids from an AKPT mouse model of CRC were orthotopically implanted into $n = 10$ mice, half of which were treated with a TFG- β inhibitor. Adjacent tissue slices were resected and stained with IHC markers for fibroblast and immune cell populations, and subsequently imaged and preprocessed (cell segmentation and type assignment). Following data acquisition, we applied our algorithms for spatial integration and microenvironment characterization.	90
4.2	Spatial Integration of Immune and Fibroblast Panels. Tissues from two mice are shown, with tissues stained with immune markers in blue, and fibroblast markers shown in orange. Their positions and orientations pre and post integration are shown. Note that tears and holes are present that are not shared across slices.	92
4.3	Cluster Stability Analysis. (a) By randomly removing microenvironments across and within tissues, we can estimate cluster stability by reclustering on the perturbed datasets. (b) Using an intra-cluster variance curve, we can identify the optimal clustering resolution by selecting the "knee point". The graph shown is for cluster selection on the full dataset. (c) On the left, UMAP visualization of microenvironment clusters. The bottom panel shows the probability of cluster assignment across the resampled datasets, while the top panel shows the cluster assignment on the full dataset. Most microenvironments are almost always uniquely assigned to a single cluster. (d) Cluster agreement between the full and resampled datasets, compared with randomized cluster assignment. Significance bars were calculated using a two-sided Mann-Whitney U Test. (e) Average microenvironment cluster composition across all resampled datasets, with error bars indicating a 95% confidence interval. Microenvironment clusters have small variance and show distinct patterns of cell composition.	94

4.4	TGF-β Remodels the Tumor Microenvironment Landscape.	
	(a) UMAP visualization of microenvironment clustering. (b) Cluster condition specificity, significance bars are calculated using a Bonferroni corrected pairwise χ^2 test. While cluster A is not significantly enriched for either condition, B is significantly enriched in untreated AKPT tissues, and C is significantly enriched in TGF- β inhibited AKPT tissues. (c) Microenvironment cell enrichment profiles. We observe that each cluster has a distinct cell enrichment profile showing unique cell-cell interactions. Using spatial similarity scoring (d), we can construct discrete spatial domains (e) and observe the change in microenvironment locations as a function of TGF- β inhibition.	99
B.1	Graph-OT Spatial Domain Mapping of Healthy Lung Tissues	
	For each of the three panels, the first column indicates the cell type, the second pathologist annotations, and the third spatial domains identified using Graph Optimal Transport.	113
B.2	Spatial Integration of Mouse Fibroblast and Immune Panels	
	Adjacent tissues from all $n = 10$ mice are shown, with tissues stained with immune markers in blue, and fibroblast markers shown in orange. Row labels indicate sample IDs. Positions and orientations pre and post integration are shown.	114
B.3	Spatial Domain Mapping of Mouse CRC Tissues	
	In each tissue, all cell microenvironments were assigned to their most similar microenvironment cluster using the procedure outlined in Section 4.3.3. Detailed cluster descriptions, including cell enrichment and condition specificity, are provided in Figure 4.4.	114

Chapter 1

Introduction

Cells do not exist or function in a vacuum, but in complex microenvironments where their interactions play a fundamental role in the coordination of global tissue function. In cancer, complex interactions between fibroblasts, immune, and cancer cells within the tumor microenvironment are known to be important for immune inhibition, tumor growth, and treatment resistance [57, 77, 119]. Given their importance, there has been increased interest in understanding and characterizing cell microenvironments, and deciphering their role in disease and therapeutic responsiveness [39, 28, 86]. In recent years, spatial biology has been revolutionized by advancements in spatial proteomics technologies, which enable researchers to profile tissues with multiple protein markers at single cell resolution while preserving their spatial context [72, 44, 102, 62, 61]. However, manual analysis is challenging due to the rapidly increasing number of cells and markers profiled, and traditional computational methods struggle to capture and exploit the rich spatial information in these datasets.

In this thesis, we develop novel computational methods tailored for spatial data, leveraging principles from Optimal Transport (OT), a branch of statistics for comparing probability distributions [116]. OT measures the distance between distributions by identifying the cost of the most efficient way of transporting one distribution into

another, offering an intuitive and rigorous metric for comparing diverse data, including points clouds, graphs, and densities [83]. Our methods address two key tasks in the analysis of multiplex information, the integration of spatial profiles from different tissues, and the characterization of microenvironments and cell-cell interactions (Figure 1.1). We evaluate our methods on published and synthetic datasets, and rigorously benchmark them against state-of-the-art algorithms. Finally, we apply them to a new dataset of multiplex immunohistochemistry profiled murine colorectal carcinoma, in order to investigate changes in cell microenvironments and interactions in response to transforming growth factor (TGF) β inhibition, a signalling molecule commonly dysregulated in cancer.

1.1 *In situ* spatial profiling of tissues

Historically, *in situ* proteomic tissue analysis was constrained to provide either molecular or spatial information, but not both. Technologies such as mass spectrometry allow for the profiling of hundreds of proteins in single cells, but without their spatial context, as the sample preparation process involves tissue lysis and enzymatic digestion to release peptide fragments [52]. In contrast, staining methods such as hematoxylin and eosin (H&E) and immunohistochemistry (IHC) capture the spatial organization of entire tissues, but with little to no molecular information, as limited specificity and increased cross-talk restrict the number of antibody markers that can be applied. [107, 24].

In recent years, advanced multiplex imaging technologies such as multiplex immunohistochemistry (multiplex IHC), imaging mass cytometry (IMC), cyclic immunofluorescence (CyCIF), co-detection by indexing (CODEX/Phenocycler), and multiplexed ion beam imaging (MIBI) have bridged this modality gap, allowing researchers to analyze tissues across dozens of proteins at single cell, sub-micrometer

spatial resolution. While these technologies employ diverse methods to acquire spatially resolved measurements of protein expression, they all encode this information as a large 2-dimensional, multichannel pixel image, where each channel for each pixel represents the abundance of a particular marker measured at that location [72, 44, 102, 62, 61]. This standardization allows for scalable visualization and analysis across platforms and modalities.

After *in situ* profiling, most approaches for viewing and analyzing multiplex images employ image analysis pipelines to preprocess the raw, multichannel pixels to highlight cell positions and categorize them into distinct cell types. Cell locations are obtained through image segmentation, typically using machine learning computer vision algorithms. These algorithms are readily available and highly automated, either as standalone software (e.g., DeepCell [41] and CellPose [101]), or integrated into open source platforms (e.g., Qupath [5] and MCMICRO [94]) and commercial platforms (e.g., HALO and Visiopharm). These algorithms are capable of generating high-resolution pixel masks, but are often used to provide simple summary statistics such as cell center location or cell size. Cell type assignment is typically a more iterative and supervised process, due to the diversity of available marker panels. Some approaches categorize cell types with manually tailored thresholds for each marker, for example using CD4 and CD8 expression for immune cell identification and stratification. Other algorithms (e.g., histoCAT for IMC panels and CELESTA for IHC data) offer a semi-supervised solution, with initial unsupervised clustering in marker space, followed by manual cluster annotation [93, 125]. Together, these pipelines transform massive, multichannel images into spatial point clouds of cell types, streamlining downstream visualization and analysis (Figure 1.1).

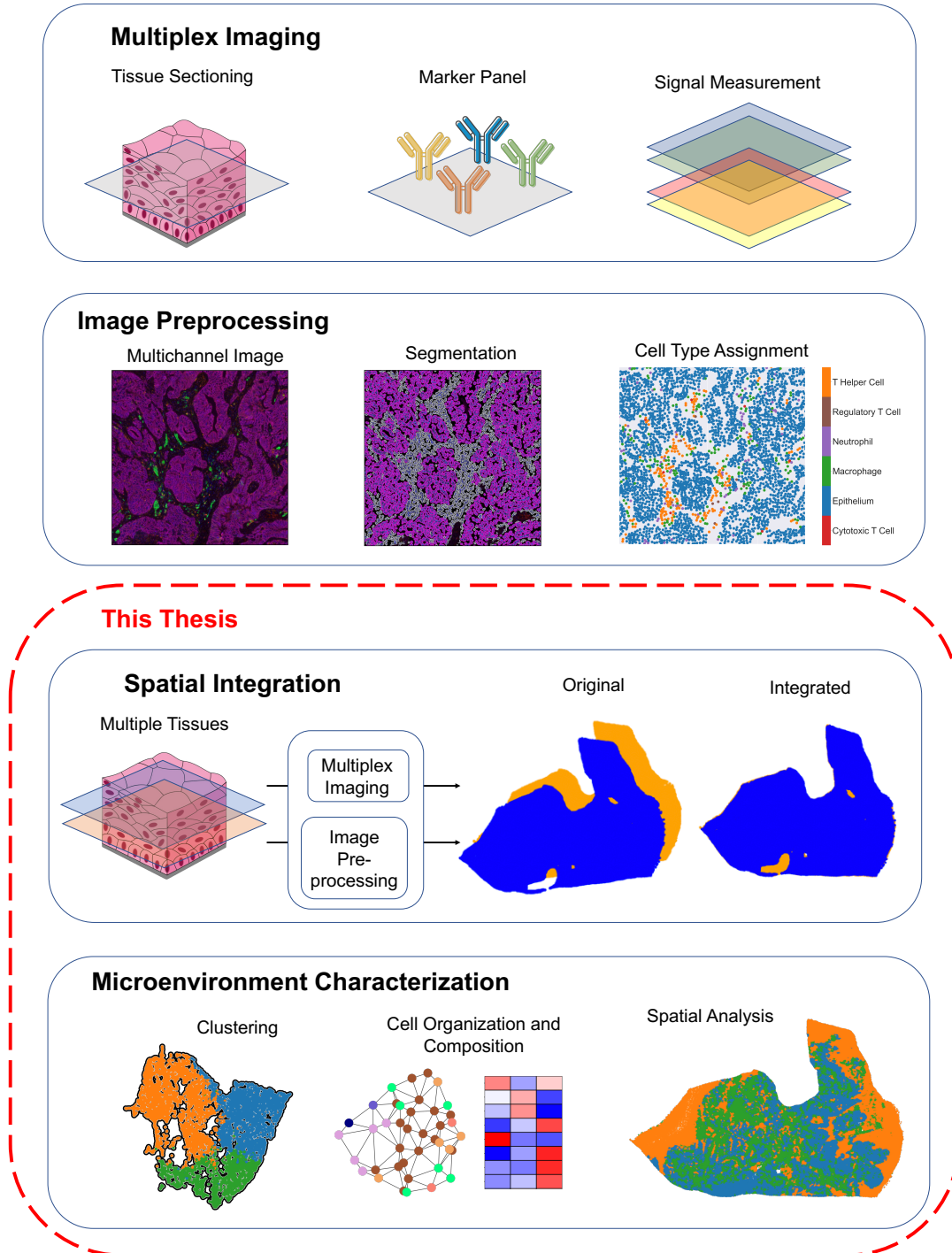


Figure 1.1: **Multiplex Imaging and Analysis Pipeline.** In the top panel, tissues are measured for various protein markers. In the second panel, cell segmentation and cell type assignment convert raw, multi-channel images into spatial point clouds of cell types. The images are from a $1 \text{ mm} \times 1 \text{ mm}$ patch from a multiplex IHC image of murine colorectal cancer. In the third panel, different spatial profiles are integrated for joint analysis. In the fourth, microenvironments and spatial domains are identified and characterized.

1.2 Computational Challenges and Opportunities

The confluence of rapidly advancing multiplexing technologies with the growing convenience and accessibility of preprocessing pipelines, is increasing the availability and abundance of spatial data. This information is highly relevant to many diseases, and there is increased evidence that aberrant spatial organization and cell-cell interactions drives disease progression and impact therapeutic responsiveness. For example, analysis of breast tumors profiled with multiplex IMC has identified distinct tumor microenvironments that are highly predictive of clinical outcome [28]. In SARS-CoV-2, multiplex IMC revealed that excessive immune infiltration of alveolar epithelial cells induce a localized hyperinflammatory cell state that is associated with lung damage [86]. In CyCIF imaged lung adenocarcinoma, networks of interacting lymphocytes have emerged as a distinctive feature of the anti-cancer immune response. Upon immune checkpoint blockade therapy, these lymphocyte networks were observed to recruit and co-localize with cytotoxic CD8+ T cell populations [39]. These recent examples highlight the wealth of biological insights contained in the spatial information generated by multiplex technologies.

In order to exploit and facilitate further biological discoveries from multiplex data, computational methods are needed to characterise the spatial and cell type information contained in these complex and increasingly massive datasets. There are a variety of computational challenges and opportunities at each step of the multiplex workflow, from image acquisition to preprocessing to analysis to biological interpretation. In this thesis we focus on two crucial steps: tissue integration and microenvironment characterization. The first task involves unifying spatial information obtained from different multiplex experiments, and plays a crucial role in enhancing the amount of information obtained from multiplex technologies. The second involves analyzing and quantifying the spatial organization of cell types, and is directly relevant to biological discovery.

1.2.1 Spatial Integration

While revolutionary, multiplex technologies have many limitations and trade-offs. Existing technologies are limited to measuring protein expression in two dimensional slices extracted from three-dimensional tissues, thus losing information that is shared across tissue slices. Furthermore, an important trade-off exists between the size of the tissue profiled and the number of unique proteins measured. Low-plex platforms such as multiplex IHC and MIBI are capable of Whole Slide Imaging (WSI) but with relatively small numbers of markers (e.g., 9-12 markers). On the other hand, high-plex platforms such as CODEX and CyTOF can measure 40-50 markers on small tissue regions ($\approx 10^3$ cells), but are extremely expensive and time consuming to scale to WSIs ($\approx 10^5$ cells) [102, 11].

To circumvent these challenges, low-plex multiplex imaging can be applied to multiple, adjacent WSI slices (e.g. along the z -axis) sectioned from the same tissue with different slices profiled with different marker panels. A joint analysis across all adjacent slices allows for spatial analysis across a larger diversity of cell types, and better captures the overall composition and structure of the tissue. This approach is both cost effective and scalable, and can be performed with most existing multiplex platforms [65].

Unfortunately, analyzing information across slices requires *spatial integration* to ensure that the the local cell locations are globally consistent or “aligned” across slices. This task is tedious to perform manually, and challenging for traditional image registration algorithms. While image registration is well studied and commonly applied in photography or object recognition, a key requirement is that the registered images look similar, or have user annotated anchor points between images. These requirements are not met in multiplex imaging. Despite coming from the same three-dimensional tissue, the two-dimensional slices often exhibit a high degree of spatial variability, due to changes in tissue architecture orthogonal to the slicing direction,

or through rips and tears induced by the sample collection process. Automatic algorithms for image registration struggle with these artefacts, and in practice integration is typically performed manually through image manipulation software [5, 94]. To address these shortcomings, unsupervised computational methods that are robust to spatial artefacts are needed to perform spatial integration and enable the analysis of diverse cell types across whole slide tissues.

1.2.2 Microenvironment Characterization

Traditionally, the analysis of multiplex imaging data has been through qualitative visual inspection. However, with the increasing number of cells and markers profiled by new and emerging technologies, this has become more challenging. While algorithms that focus on cell counts or marker distributions are available from single cell analysis ecosystems, they do not take into account cell locations, and cannot describe spatial organization. Most current computational methods for analyzing spatially multiplexed data tend to draw from two approaches: point process statistics and machine learning.

Point process statistics are mainly derived from ecology and geostatistics, and involve fitting parametric models to cell locations or designing custom spatial statistics to quantify properties of interest [4, 73]. The simplest example is the self-explanatory average nearest neighbor distance between two cell types, but more complex variations exist. For example, Moran’s I and Ripley’s K, measure spatial clustering and exclusion within a cell type, while the Morista-Horn index and J-function measure spatial correlations between pairs of cell types [75, 87, 45, 114]. Designing these statistics is a complex process, and establishing their significance even more so. Most methods are limited to the analysis of pairwise relationships, and rely on simple null models of complete spatial randomness (CSR) or computationally intensive Monte Carlo permutation tests [88, 6]. Because of these challenges, the direct application of

spatial statistics has so far been limited to single or low marker histology data, and current research efforts for multiplex data have focused on combining or extending these statistics [80, 15, 14].

Rather than attempting to “feature engineer” custom statistics, machine learning algorithms attempt to empirically learn functions from spatial data to biological classes, and are usually repurposed or adapted algorithms from computer vision applications. Convolutional Neural Networks (CNNs) have found enormous success in computer vision tasks, and extensions to multiplex data apply CNNs directly to the raw, multichannel images. These methods characterize microenvironments by framing the biological task as a patch resolution classification problem, or a pixel resolution segmentation task [99, 36]. In recent years, the rise of graph neural networks (GNNs) has fueled similar applications. In this framework, cell locations and types are represented as graphs, with edges typically constructed by threshold radius proximity, or through nearest neighbors. GNNs have been successfully applied to predicting disease or stratifying patients from spatial graph motifs [120, 66]. While powerful, machine learning methods employ “black-box” models that are difficult to interpret and heavily reliant on large quantities of labeled training data, and unsupervised or self-supervised training is an active area of research [122, 89]. A summary of these paradigms is provided in Table 1.1.

1.3 Optimal Transport for Multiplexed Imaging

1.3.1 Motivation

A recurring challenge in the computational tasks that we have introduced is the lack of ground truth. When performing spatial integration, this is unavoidable, as the optimal alignment of every multiplex experiment will depend on the shape of the underlying tissue. The same holds true for microenvironment characterization. While

Paradigm	Summary	Advantages	Disadvantages
Molecular Expression	Derived from classical bioinformatics or single-cell transcriptomic analysis. Statistical analysis of molecular expression in each cell independently.	Semi or fully unsupervised, and highly interpretable. Can apply large, biomedical databases (e.g. Gene Ontology, Pathway Interaction Database) for increased power and interpretability.	Does not make use of spatial information. Multiplex proteomics datasets contain far fewer markers per cell ($\sim 10^1$) than transcriptomics ($\sim 10^3$), constraining capacity for comprehensive pathway analysis.
Point Process Statistics	Derived from geostatistics, ecology, public health etc. Constructs custom spatial statistics based on cell locations to identify spatial patterns or trends.	Fully unsupervised, and highly interpretable. Allows for statistical significance testing.	Challenging to define, has difficulty accounting for marker information. Largely constrained to pairwise relationships between categorical populations. Relies on simple null models of complete spatial randomness.
Machine Learning	Derived from network analysis, computer vision, etc. Models tasks as graph or image classification problems.	Fully supervised. Accounts for spatial and marker information, learns directly from data. Many drop-in model architectures and optimizers.	Requires large annotated or labeled datasets. Weak interpretability due to black-box models.

Table 1.1: **Comparison of Spatial Analysis Paradigms.** This summary is not exhaustive and serves as a concise overview.

correlations with various external covariates (disease status, treatment outcome, etc.) can be used as weak supervision or validation, microenvironment characterization for biological discovery is necessarily an unsupervised task, since the ground truth is the

object of discovery. This challenge precludes the application of most data driven machine learning algorithms, which require many labeled examples, and are difficult to interpret.

A second theme is the difficulty of designing computational algorithms that can account for the spatial information contained in multiplex images. In spatial integration, this requires algorithms that are robust to multiple, different spatial deformations, and precludes the use of simple object detection or feature based image registration algorithms. In microenvironment characterization, interactions between diverse cell types define biological niches. Since these interactions are difficult to encode in hand designed point process statistics, existing methods have been largely limited to studying pairwise interactions.

In this thesis, we present an approach that circumvents both of these challenges by relying on statistical distance *metrics*. Distance metrics are analytic functions that satisfy a strict set of mathematical properties, thus guaranteeing principled and consistent behavior. Furthermore, distance metrics do not require supervision and are not reliant on annotated examples or correlated labels; rather, they are defined directly on the mathematical structures they compare.

In this thesis, we employ Optimal Transport (OT), a branch of statistics that describes distance metrics between complex probability distributions, such as point clouds, densities, and graphs. Multiplexed data is naturally described by such distributions. For example, tissues can be represented as joint distributions between spatial and cell type random variables. Additionally, microenvironments can be described as labeled graphs, with nodes encoding cell types and edges describing interactions or proximity. OT calculates the distance between distributions by finding the cost of the most efficient, or optimal, way of transporting one distribution into another. As such, this intuitive and theoretically rigorous metric is ideally suited for quantifying distances in spatial data [83].

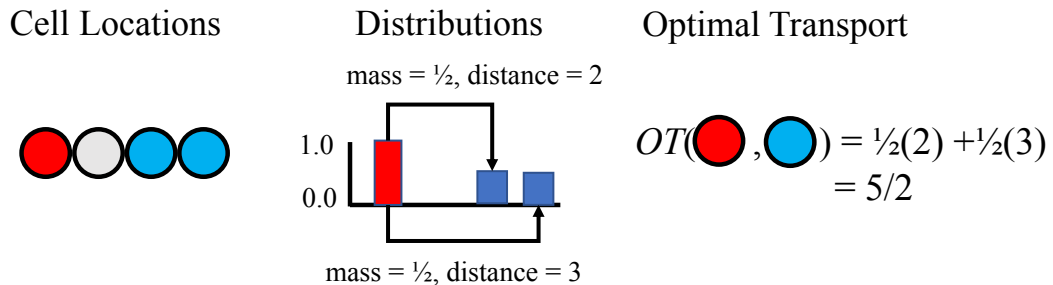


Figure 1.2: **Illustrative Example.** In the left panel, cell positions on a 1D lattice are shown. In the middle panel, the cell locations are converted into a probability distribution, and the optimal plan for transporting the red to the blue distribution is shown. In the right panel, the optimal transport distance calculation is shown.

1.3.2 Optimal Transport for Spatial Data

One popular Optimal Transport (OT) metric is the Wasserstein distance. This distance metric has a remarkably intuitive interpretation. If we consider probability distributions P, Q as mounds of mass or dirt, the Wasserstein distance can be interpreted as the least amount of “work” required to “transport” pile P into pile Q . In this setting, “work” is defined as the amount of mass moved multiplied by the distance it is transported, summed over every trip. A “transport plan” is an assignment that moves mass from P to Q and satisfies two conditions: (i) all the mass from P must be moved to Q , e.g. no mass can be “left over” or otherwise be unaccounted for, (ii) only non-negative quantities of mass may be moved. For this reason, the Wasserstein Metric is often known as the “Earth Mover’s Distance”, and its value can be viewed as the work required under the most efficient, or optimal, transport plan [83].

In Figure 1.2 we present a simple application of the Wasserstein distance which illustrates several of the advantages that OT offers over alternative spatial statistics that are not distance metrics. For example, the OT distance between red and blue

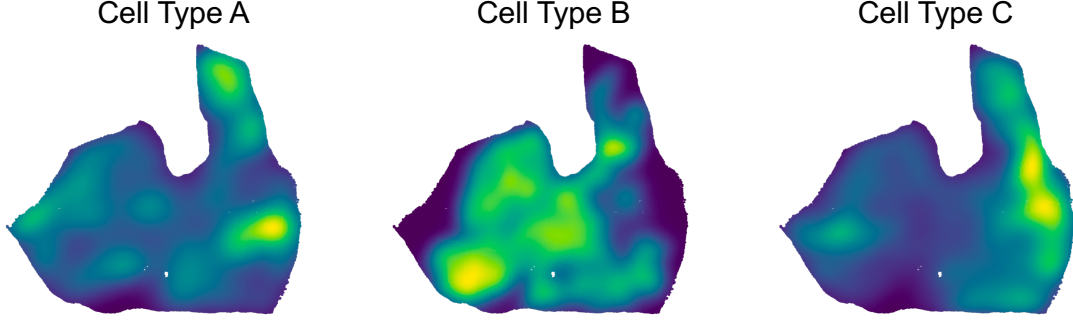


Figure 1.3: **Optimal Transport Co-Localization.** Kernel Density Estimates were applied to cell locations from three different cell types to generate spatial probability distributions, where brighter colors indicate higher likelihoods of observing a particular cell type at a given spatial location. By visual inspection, we surmise that cell types A and C are more closely co-localized than they are with cell type B. Optimal Transport quantifies this intuition, and confirms that $OT(A, C) < OT(B, C)$.

cells is symmetric ($OT(R, B) = OT(B, R)$), and invariant to unequal populations. These properties are not guaranteed for most spatial statistics. Consider for example, the average nearest neighbor distance (ANND), which is the average minimum distance between pairs of cell types.

$$ANND(R, B) = \frac{1}{n} \sum_{i=1}^n \min_j \|R_i - B_j\|_2^2 = \frac{1}{1}(2) = 2$$

$$ANND(B, R) = \frac{1}{m} \sum_{j=1}^m \min_i \|B_j - R_i\|_2^2 = \frac{1}{2}(2 + 3) = 2.5$$

The ANND is not symmetric, and sensitive to the number of samples in each population, leading to difficulties in interpreting the spatial relationships quantified by ANND. In contrast, the Wasserstein distance is guaranteed to be symmetric, and can quantify spatial relationships invariant to population fluctuations. These metric properties allow OT to robustly quantify spatial relationships, as demonstrated in Figure 1.3.

1.4 Mathematical Foundations

Optimal Transport is a diverse field with a rich history. It was originally conceived as an engineering problem, and the analogies of “Earth Moving” are in fact the original historical motivations for its invention, when the French mathematician Gaspard Monge was tasked with constructing earthen military fortifications in the late 18th century [74]. Since then various extensions have been introduced. In 1939 Soviet economist Leonard Kantorovich introduced the Kantorovich Relaxation to the original Optimal Transport problem posed by Monge for the optimization of factory logistics [51]. The Wasserstein distance previously introduced is directly derived from this innovation, and would win him the Nobel Prize in Economics in 1975. In 1949, George Dantzig showed how to efficiently solve Optimal Transport problems with linear programming, giving the field a firm footing in optimization [29]. With the rise of fast computing and machine learning, further advancements have arisen, and many variants now exist for applying Optimal Transport to all kinds of probability distributions with varying applications [27, 3, 100, 71].

While we will exploit several state-of-the-art variants of Optimal Transport in later chapters, in this section, we will describe the basic mathematical principles underlying the general theory. We will concretely define the important notions of couplings and transportation plans, algorithmic details regarding their computation, and theoretical guarantees underlying their performance, starting with Monge’s original optimal transport distance.

1.4.1 Assignment and The Monge Problem

A discrete measure is a probability vector associated with a set of locations on a metric space. Concretely, let $\mathbf{a} \in \mathbb{R}_{\geq 0}^n, \mathbf{b} \in \mathbb{R}_{\geq 0}^m$ be probability vectors such that $\sum_{i=1}^n \mathbf{a}_i = \sum_{j=1}^m \mathbf{b}_j = 1$, and let $(\mathbf{x}_1 \cdots \mathbf{x}_n), (\mathbf{y}_1 \cdots \mathbf{y}_m)$ be locations in domain Ω ,

where Ω is some metric space with measurable cost function $c : \Omega \times \Omega \rightarrow \mathbb{R}$. The discrete measures α, β over Ω are then defined as

$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta(\mathbf{x}_i), \quad \beta = \sum_{j=1}^m \mathbf{b}_j \delta(\mathbf{y}_j)$$

where $\delta(\cdot)$ is a Dirac measure at a specified location in Ω .

A Monge (transport) map T assigns each source point $\mathbf{x}_i$ a target point $\mathbf{y}_j$, and pushes the mass α toward the mass β . That is, $T : \{\mathbf{x}_{i=1}^n\} \rightarrow \{\mathbf{y}_{j=1}^m\}$ and satisfies

$$\forall j, 1 \leq j \leq m, \mathbf{b}_j = \sum_{i: T(\mathbf{x}_i) = \mathbf{y}_j} \mathbf{a}_i \quad (1.1)$$

For compactness, we refer to any T that satisfies these conditions as the “push-forward operator” of α , and denote it as $T_{\#}\alpha = \beta$. The Monge Problem seeks to find the transport map T with minimal cost, subject to the condition that T is a valid push forward operator of α .

$$\min_T \left\{ \sum_i^n c(\mathbf{x}_i, T(\mathbf{x}_i)) : T_{\#}\alpha = \beta \right\} \quad (1.2)$$

When α, β are continuous measures on Ω , the transport map T is defined by a function $T : \Omega \rightarrow \Omega$, and the “push-forward” condition $T_{\#}\alpha = \beta$ is satisfied iff, for any set $B \subset \Omega$,

$$\beta(B) = \alpha(\{x \in \Omega : T(x) \in B\}) = \alpha(T^{-1}(B)) \quad (1.3)$$

In the continuous setting, the Monge Problem can be defined analogously.

$$\inf_T \left\{ \int_{\Omega} c(\mathbf{x}, T(\mathbf{x})) d\alpha(\mathbf{x}) \mid T_{\#}\alpha = \beta \right\} \quad (1.4)$$

The Monge Problem can be viewed as a generalized version of the assignment problem,

as it allows probability vectors to be non-uniform and of different size. Indeed for the particular case when $n = m$ and $\forall i, j, \mathbf{a}_i = \mathbf{b}_j = \frac{1}{m} = \frac{1}{n}$, the Monge Problem reduces to finding the optimal bijection between sets α and β .

This generality is not wholly complete however, and solutions to the Monge Problem do not exist or satisfy uniqueness for all measures α, β . No Monge maps exist when $n < m$, and the push forward operator $T_{\#}\alpha$ cannot be constructed for incompatible probability weight vectors (e.g. when $\mathbf{a} = e_1, \mathbf{b} = \frac{1}{m}\mathbf{1}$). Furthermore, solving the Monge Problem requires combinatorial optimization.

1.4.2 Coupling and the Kantorovich Relaxation

The Kantorovich Problem resolves issues associated with the Monge problem by relaxing the requirement for deterministic matching and introducing probabilistic correspondence. In the Kantorovich Problem the satisfactory conditions are only that the exact mass from source α is allocated to target β . The transport plan that accomplishes this allocation can assign mass from a single source point to multiple target points, in a phenomenon most commonly referred to as “mass-splitting”.

In place of a Monge Map T , the Kantorovich transportation plan π can be expressed as a coupling matrix, where $\pi_{i,j}$ denotes the mass sent from source point $\mathbf{x}_i$ to target point $\mathbf{y}_j$. The satisfactory conditions can then be summarized as requiring that the row and column sums of π equate to $\mathbf{a}$ and $\mathbf{b}$ respectively, and the set of admissible coupling matrices forms a convex polytope denoted as Π .

$$\Pi(\mathbf{a}, \mathbf{b}) = \{\pi \in \mathbb{R}_{\geq 0}^{n \times m} \mid \pi \mathbf{1}_m = \mathbf{a}, \pi^T \mathbf{1}_n = \mathbf{b}\} \quad (1.5)$$

Similar to the Monge Problem, the Kantorovich Problem seeks to find the admissible

plan with minimal cost

$$\min_{\pi \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i=1}^n \sum_{j=1}^m c(\mathbf{x}_i, \mathbf{y}_j) \pi_{ij} = \min_{\pi \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \pi \rangle \quad (1.6)$$

where $\mathbf{C} \in \mathbb{R}^{n \times m}$ is a pairwise cost matrix between all points in α, β , i.e. $\forall i, j, \mathbf{C}_{ij} = c(\mathbf{x}_i, \mathbf{y}_j)$.

When α, β over Ω are continuous, the transportation plan π is a coupling, or joint probability measure on $\Omega \times \Omega$ whose marginals are α and β respectively. The set of admissible couplings is given by

$$\Pi(\alpha, \beta) = \left\{ \pi_{\geq 0} \in \Omega \times \Omega \left| \int_{\Omega} \pi(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \alpha, \int_{\Omega} \pi(\mathbf{x}, \mathbf{y}) d\mathbf{x} = \beta \right. \right\} \quad (1.7)$$

Analogously, the Kantorovich Problem for continuous measures finds the minimal cost associated with a valid coupling.

$$\inf_{\pi \in \Pi(\alpha, \beta)} \int_{\Omega \times \Omega} c(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y}) \quad (1.8)$$

The Kantorovich problem improves upon the Monge problem's original formulation in its generality and practical implementation. Solutions are guaranteed to exist for all α, β , regardless of their size (n, m) and the construction of $\mathbf{a}, \mathbf{b}$ [51]. Furthermore, the space $\Pi(\mathbf{a}, \mathbf{b})$ forms a convex polytope, which reduces the complexity of optimal transport from a combinatorial optimization problem to solving a linear program [83]. For these reasons, the Kantorovich Problem is the most popular formulation of the optimal transport problem, and is applied (and renamed) in various fields [83, 51, 46, 13]. While the Kantorovich formulation is applicable to any cost function c , in practice c is commonly defined as an ℓ_p metric. In such cases, the transportation cost obtained by solving the Kantorovich Problem is called the p -Wasserstein distance.

1.4.3 Computation

As introduced in the previous section, the p -Wasserstein distance between discrete measures α, β is given by the Kantorovich Problem with an ℓ_p distance metric as the cost function.

$$\mathcal{W}_p(\alpha, \beta) = \min_{\pi \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \pi \rangle$$

$$\text{where } \mathbf{C}_{ij} = d_p(\mathbf{x}_i, \mathbf{y}_j) = \left(\sum_k |\mathbf{x}_{ik} - \mathbf{y}_{jk}|^p \right)^{\frac{1}{p}} \quad (1.9)$$

To solve this optimization problem, we can rewrite the objective and constraints as a linear program. This requires encoding the restrictions of admissible coupling matrices $\pi \in \Pi(\mathbf{a}, \mathbf{b})$ into an equality with the product between a matrix and a vector, rewriting the matrix dot product $\langle \mathbf{C}, \pi \rangle$ in the objective function into a vector dot product, and enforcing non-negativity for all variables.

We can convert the matrices in the objective function into the desired vectors through vectorization. Let $\mathbf{p} \in \mathbb{R}_+^{nm}$ denote the column-stacked coupling matrix π , and $\mathbf{c} \in \mathbb{R}_+^{nm}$ column-stacked cost matrix $\mathbf{C}$. That is, $\mathbf{p}_{i+n(j-1)} = \pi_{ij}$, and $\mathbf{c}_{i+n(j-1)} = \mathbf{C}_{ij}$. To encode the row and column-sum restrictions, we define a matrix $\mathbf{A} \in \mathbb{R}^{(n+m) \times nm}$,

$$\mathbf{A} = \begin{bmatrix} \mathbf{1}_n^T \otimes \mathbf{I}_m \\ \mathbf{I}_n \otimes \mathbf{1}_m^T \end{bmatrix}$$

where $\mathbf{I}_n$ denotes an identity matrix of size n , and $\otimes$ the Kronecker Product. The set of admissible coupling matrices $\pi \in \Pi(\mathbf{a}, \mathbf{b})$ is then equivalent to

$$\left\{ \mathbf{p} \in \mathbb{R}_+^{nm} \mid \mathbf{A}\mathbf{p} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \right\}$$

After these modifications, the original optimization problem for the p -Wasserstein can be rewritten as

$$\mathcal{W}_p(\alpha, \beta) = \min_{\mathbf{p} \in \mathbb{R}_+^{nm}} \mathbf{c}^T \mathbf{p} \quad \text{s.t.} \quad \mathbf{A}\mathbf{p} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \quad (1.10)$$

This is a linear program in canonical form, and can be solved using any Linear Programming solver. In practice, most implementations rely on Orlin’s algorithm [79] or interior point methods [64], which have complexity $\mathcal{O}(k^3 \log k)$, where $k = nm$. When α, β are continuous measures, the same procedure can be applied after discretization [83, 29].

1.4.4 Theoretical Guarantees

Defining appropriate measures of “distance” between probability distributions is a fundamental problem in statistics, and many methods exist. The strongest notions of “distance” are characterized as distance *metrics*, and satisfy the three *metric* conditions. Formally, given a domain M and a distance $d : M \times M \rightarrow \mathbb{R}$, d is a metric if and only if d satisfies the following conditions for all distributions $P, Q, R \in M$

1. (Non-negativity) The distance between two distributions is non-negative:

$$d(P, Q) \geq 0$$

2. (Identity of Discernables). The distance between two distributions is 0 if and only if they are identical:

$$d(P, Q) = 0 \iff P = Q$$

3. (Symmetry) The distance between two distributions is independent of their

order in the distance function:

$$d(P, Q) = d(Q, P)$$

4. (Triangle Inequality) Distances satisfy geometric and physical notions of space:

$$d(P, R) \leq d(P, Q) + d(Q, R)$$

Most statistical distances are not metrics; for example, the Kullback-Leibler Divergence is not symmetric [56], while the Bhattacharyya distance does not satisfy the triangle inequality [9].

The Wasserstein distance satisfies all four properties, and as such is a metric [51]. Furthermore, the concept of an optimal transportation plan that underlies Wasserstein metric ensures that the Wasserstein distance respects the geometry of the measure space. This property separates Optimal Transport metrics from other statistical distance metrics. As an illustrative example, consider two 1-dimensional probability distributions $p(x), q(x)$ defined as

$$p(x) \sim \mathcal{N}(\mu, \sigma^2), \quad q(x|\tau) \sim \mathcal{N}(\mu - \tau, \sigma^2)$$

Here $\tau \in \mathbb{R}$ is a free parameter that dictates the “difference” between $p(x)$ and $q(x)$, which we will attempt to quantify using a statistical distance metric d . For ease of notation, we will refer to $d(p(x), q(x|\tau))$ as simply $d(p, q(\tau))$. When $\tau = 0$, $p(x)$ and $q(x|\tau)$ are identical, and we expect that $d(p, q(\tau)) = 0$. Conversely, as $|\tau|$ increases, we expect that $d(p, q(\tau))$ will also increase.

In Figure 1.4 we can see that this is not the case for most statistical distance metrics. For the Jensen-Shannon, L_1, L_2 , and Hellinger metrics, $d(p, q(\tau = 10)) = d(p, q(\tau = 15))$. In fact, $d(p, q(\tau = 10)) = d(p, q(\tau = h))$ for any value of h with

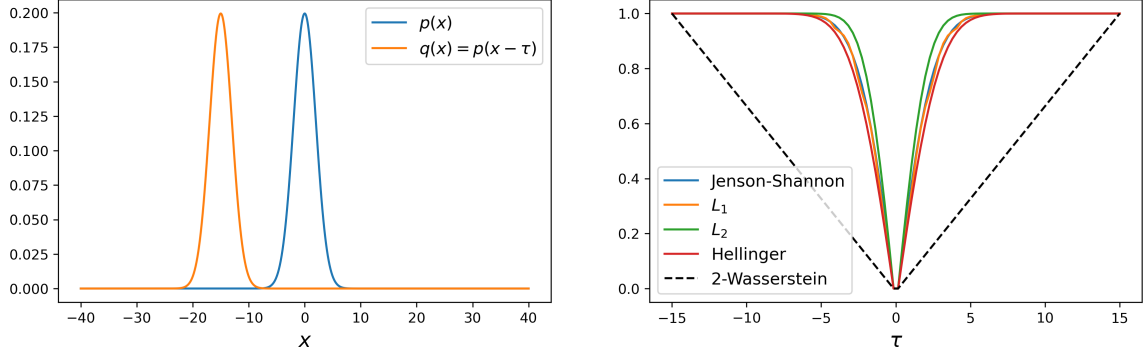


Figure 1.4: **Comparing Statistical Distance Metrics.** We compute the statistical distance between two probability measures defined on domain $\mathbb{R}$. While all the distance functions presented are metrics, only the 2-Wasserstein respects the geometry of the domain. The distributions were parameterized as $p(x) \sim \mathcal{N}(\mu, \sigma^2)$, $q(x|\tau) \sim \mathcal{N}(\mu - \tau, \sigma^2)$. For each metric, the range of $d(p(x), q(x|\tau))$ over $\tau \in [-15, 15]$ was $[0, 1]$ normalized for ease of comparison.

a magnitude greater than 10. Furthermore, we can see that the magnitude of the gradient $\partial d(p, q(\tau))/\partial \tau$ is not constant for these metrics. The change in $|d(p, q(\tau))|$ when τ is close to 0 is extremely large, however as the magnitude of τ increases, $|d(p, q(\tau))|$ appears to “saturate” and becomes constant. While the Jensen-Shannon, L_1 , L_2 , and Hellinger distances satisfy all four metric properties, their behavior is counter intuitive to our notions of “distance”, and does not respect the geometry of the underlying space (i.e. $\mathbb{R}$). In contrast, the Wasserstein distance increases linearly as $|\tau|$ increases and the distributions “move” further apart. This behavior better captures our intuition, and respects the geometry of the domain.

1.5 Contributions of this Thesis

This thesis addresses key challenges in multiplexed imaging by developing new computational methods using techniques from Optimal Transport. We focus on two challenges: spatial integration and microenvironment characterization. We evaluate our methods on published and synthetic data, and compare their performance with state-of-the-art benchmarks. Lastly, we apply our methods to analyze a new multiplexed

IHC dataset of murine colorectal cancer.

In this chapter, we have introduced the computational challenges and opportunities in multiplexed spatial imaging, and reviewed existing methods for the quantitative analysis of spatial data, highlighting their advantages and shortcomings. We have proposed a novel approach that relies on statistical distance metrics derived from Optimal Transport. We provided a mathematical review of the principles, computation, and theoretical guarantees underlying OT, and showcased applications to spatial data.

In Chapter 2, we discuss the problem of integrating spatial profiles from adjacent tissue slices. We introduce a novel learning algorithm for optimizing alignment between tissues based on partial OT. Our method addresses the practical and challenging case where slices do not fully overlap and have different cell type compositions. We evaluate our method on simulated datasets, and demonstrate its superior resilience to spatial distortions and artefacts commonly encountered multiplexed datasets compared to current methods. To address the challenge of scaling alignment to tissues with hundreds of thousands of cells, we introduce a statistical estimator for optimal transport distances that achieves similar performance with a fraction of the computational cost.

In Chapter 3, we consider the task of quantitatively characterizing and identifying cell microenvironments. We introduce a novel approach to microenvironment characterization that captures both cell type and spatial information using graph OT. On synthetic datasets, we demonstrate that our method can accurately describe changes in spatial interactions and cell type compositions. On pathologist annotated multiplexed IMC images of lung tissue, we demonstrate that graph-OT reveals unique microenvironments that are directly linked to known cell type interactions and functional tissues. We benchmark our method against two state-of-the-art algorithms for microenvironment characterization, and demonstrate that it outperforms both with

statistical significance. On a dataset of CODEX imaged head and neck squamous cell carcinoma tumors, we show that graph-OT identifies microenvironments that are highly predictive of patient outcome, once again outperforming baseline benchmarks and state-of-the-art competitors.

Finally, in Chapter 4, we present a case study that showcases the practical application of our spatial integration and microenvironment characterization methods on a novel dataset of mouse colorectal cancer tissues profiled with multiplex IHC. Using partial OT spatial integration, we align tissues with hundreds of thousands of cells, enabling an integrated spatial analysis of 16 different cell types. Using graph-OT, we identify distinct microenvironment subtypes that corresponded to unique cell enrichment profiles and spatial domains. Furthermore, we showed that some of these subtypes were highly sensitive to TGF- β signalling, demonstrating that TGF- β plays an important role in shaping tumor microenvironment landscapes in colorectal cancer.

Chapter 2

Alignment and Integration of Spatially Resolved Proteomics Data with Partial Optimal Transport

2.1 Introduction

Spatial integration is the process by which information from different multiplex samples is integrated into a cohesive framework. As explained in the previous chapter, current multiplex imaging technologies are limited to two-dimensional profiles of thinly sliced tissues, and enforce a strict trade-off between the number of proteins measured and the size of the tissue profiled. Thus, spatial integration of adjacent slices profiled with different marker panels is a necessary step to profiling WSIs across diverse cell types (Figure 2.1). Due to rips and tears introduced in sample collection and measurement, as well as intrinsic biological variation orthogonal to the direction of slicing, this task is difficult for automated image registration algorithms, and tedious

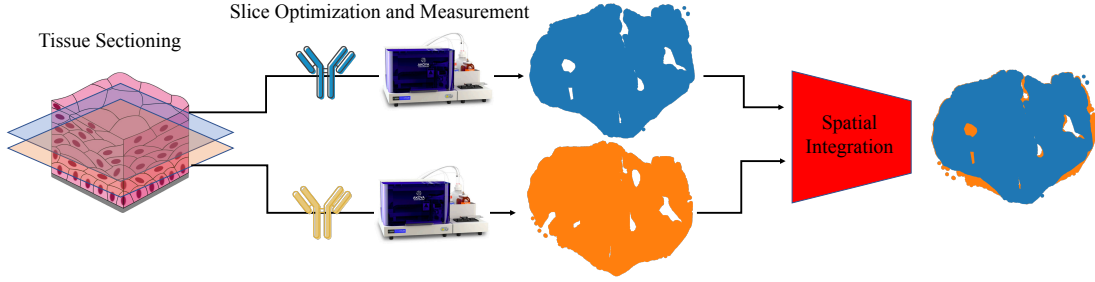


Figure 2.1: **Spatial Integration.** Two thin tissue slices (orange, blue) are sectioned from the same tissue. One slice is stained to detect one set of protein markers, and the second slice another. Due to technical variation in the sample collection process, there are rips and holes in the blue slice not present in the orange slice and vice versa. Spatial integration is used to align the blue and orange tissue slices, and allow for joint analysis of both the blue and orange cell populations in an integrated tissue.

to perform manually.

Several strategies exist for the unsupervised integration of spatial profiles. One possible approach is to employ tools from spatial transcriptomics, such as CytoSPACE [113], STUtility [7] and PASTE [124]. These methods use both spatial information and molecular expression profiles to perform slice alignment. Having been designed for spatial transcriptomics data, they assume a high degree of conserved molecules between slices, which is not guaranteed if the slices have been optimized for different cell phenotypes. An alternative strategy is to apply methods from image registration, such as ITK [67] and elastix [54]. Unfortunately, many of these methods struggle when only partial correspondences are possible, such as when tears or holes occur in the sample collecting process, and require manually selected landmarks. One of the most promising strategies for slice alignment is point set registration, which treats each cell as a point in 2D space. Each tissue slice is then modeled as a point cloud, and the spatial integration task reduces to merging point clouds into a globally consistent model. This approach is uniquely suited to spatial proteomics, as it makes no assumptions on shared molecules beyond cell detection.

In this chapter, we introduce a computational method for integrating multiple spatial measurements with substantial improvements over existing methods. Our

method applies *partial optimal transport* to point set registration, and addresses the important and practical case where slices do not fully overlap in space and have different cell phenotypes. We also introduce an efficient algorithm called *quantized optimization*, that enables rapid alignment of large tissues with hundreds of thousands of cells. To the best of our knowledge, our method is the first to introduce *quantized partial optimal transport* to the spatial proteomics integration problem, as well as the optimization procedures to solve it. We benchmark our method against Iterative Closest Point (ICP) [126] and Coherent Point Drift (CPD) [76], two state of the art algorithms, and demonstrate our method’s robustness to rips, tears, and other challenging spatial distortions induced by morphological variation and sample collection present in spatial proteomics datasets.

2.2 Point Set Registration

2.2.1 Mathematical Formulation

Spatial integration can be formulated as a point set registration problem. In this framework, each cell is treated as a point in 2D space, commonly defined as the center of a bounding box that encloses the cell. Since each cell is represented by a point, a tissue is represented as a point cloud, and the spatial integration task is then equivalent to aligning multiple point clouds into a single globally consistent model, a task known as point set registration. Unlike algorithms for spatial transcriptomics, which perform alignment by matching gene expression profiles between tissues, point set registration only requires locations, and does not require conserved cell types or a large set of overlapping markers between tissues.

Point set registration can be formulated as follows. Let $M \in \mathbb{R}^{n \times d}, S \in \mathbb{R}^{m \times d}$ denote point clouds with n, m points in a real-valued vector space of dimension d . The problem is to find a transformation that can be applied to the “moving” point

set M such that the alignment error between M and the fixed, or “static”, set S is minimized.

$$T^* = \arg \min_{T \in \mathcal{T}} \mathcal{L}(T(M), S) \quad (2.1)$$

Here $\mathcal{T}$ denotes the set of valid geometric transformations, and $\mathcal{L}$ is a cost function that quantifies the alignment error. The alignment quality is heavily dependent on the formulation of $\mathcal{L}$, and different Point Set Registration algorithms are constructed from different variants [22, 76]. When multiple (≥ 2) point sets are to be aligned, this process can be repeated, with each additional point set aligned with the point cloud formed by the union of previously aligned sets.

2.2.2 Geometric Transformations

A geometric transformation T is any bijective function $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$, and there are many possible transformation classes. The choice of the set of geometric transformations $\mathcal{T}$ is problem dependent, and selected for the properties it preserves. For example, *displacements* only allow for translations, which preserve distances and angles, whereas *diffeomorphisms* allow for “warping”, which only preserves differentiability or smoothness.

In spatial integration, we choose to constrain our class of transformations to *rigid* transformations, which preserve Euclidean distances between all pairs of points. This distance-preserving property allows for an unbiased measurement of cell proximity across tissues, which is essential for further analysis of cell interactions and communication. For our 2-dimensional tissues, Rigid Transformations can be parameterized

by the function $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, defined as

$$T(\mathbf{v}) = R\mathbf{v} + \mathbf{t} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \mathbf{v} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \quad (2.2)$$

$\forall \mathbf{v} \in \mathbb{R}^2$. The parameters $\theta \in [0, 2\pi)$ control the angle of rotation, and $t_1, t_2 \in \mathbb{R}$ the horizontal and vertical shifts respectively. For compactness, we define $\Theta := \{\theta, t_1, t_2\}$ as the set of parameters for a rigid transformation, and $T(\mathbf{v} \mid \Theta)$ to be the transformed vector under parameter set Θ . An object acted upon by a rigid transformation T will preserve its size and shape, but not its position or orientation.

2.2.3 Solving the Point Set Registration Problem

Practical point set registration algorithms typically include two components: one component assigns correspondences between S, M , and the other computes the optimal transformation between S, M based on these correspondences. If the correspondence is known, for example using manually defined landmarks or feature matching, then the registration problem reduces to finding the transformation. This problem is known as the Procrustes Problem, and has a closed form solution [95]. The more challenging and common scenario is when correspondences are unknown. In this setting, the correspondence and transformation must be jointly optimized, in a process called “simultaneous pose and correspondence registration”. In this formulation, a correspondence is hypothesized, and based on this correspondence, the pose is computed and the moving point clouds are projected to their destinations. From this “new” moved point cloud, another correspondence is hypothesized, and another pose is computed. This process proceeds iteratively until the distance between consecutive projections is less than a prescribed tolerance [34].

In the context of integrating spatial proteomics profiles, the point set registration task is particularly challenging because perfect correspondences do not exist. Since

each point cloud represents a different tissue slice, it is almost guaranteed that $n \neq m$, and no direct assignment exists since almost no cells are shared between tissues. Furthermore, changes in the tissue architecture parallel to the axis of slice acquisition, as well as rips or holes induced during the sample collection process, mean that even the overall shape of the tissues will be different across slices. If these deformations are severe enough, a complete alignment may not exist, and a partial alignment must be sought, since only overlapping parts of the different tissues can be registered to one another.

2.3 Related Work

Point set registration is a well studied problem in computer vision, with important applications in robotics, navigation, and computer graphics. Two of the most popular and successful algorithms are the Iterative Closest Point (ICP) and Coherent Point Drift (CPD) [126, 76].

In the ICP algorithm, correspondences are defined in terms of the closest (Euclidean) distance criterion. That is, the correspondence between the transformed moving point $T(m_i)$ and the scene point s_j is given by

$$\pi(T(m_i), s_j) = \begin{cases} 1 & \text{if } d(T(m_i), s_j) < d(T(m_i), s_{k \neq j}) \\ 0 & \text{otherwise.} \end{cases}$$

Given these correspondences, the optimal transformation minimizes the Euclidean distance between each point in the transformed moving point set M , and its corresponding point in the scene point set S .

$$\mathcal{L}_{ICP} = \sum_{i=1}^N \sum_{j=1}^M \pi(T(m_i), s_j) \|T(m_i) - s_j\|_2^2 \quad (2.3)$$

ICP is optimized in an alternating, iterative fashion. The correspondence matrix π is estimated by associating each transformed moving point $T(m_i)$ with its nearest neighbor in S , and the transformation T that minimizes the loss conditioned on the correspondence matrix π is obtained by solving the Orthogonal Procrustes Problem. This alternating process is repeated until convergence.

ICP is simple to implement and efficient to compute, and is often used for real-time point registration tasks, such as in autonomous navigation and robotic manipulation. Unfortunately, the performance of ICP is heavily dependent on the quality of its correspondence matrix, which can be difficult to estimate in settings where there are a high proportion of outliers or missing points, or when only partial correspondences may exist. Various extensions have been introduced to increase the robustness of ICP, for example by including a Random Sample Consensus (RANSAC) step to identify and reject outliers [37].

CPD attempts to correct for the brittleness of ICP by introducing probabilistic correspondences. In CPD, the moving point set M is represented by an n -component Gaussian Mixture Model (GMM) with centers $T(m_i)$, while the scene point set S represents observations from the GMM. Subsequently, the correspondence between the transformed center $T(m_i)$ and scene point s_j is simply the posterior probability of observing or sampling s_j from component $T(m_i)$,

$$\begin{aligned}\pi(T(m_i), s_j) &= p(s_j \mid T(m_i)) \\ &= \exp\left(\frac{-1}{2\sigma^2}\|s_j - T(m_i)\|_2^2\right) / \sum_{k=1}^n \exp\left(\frac{-1}{2\sigma^2}\|s_j - T(m_k)\|_2^2\right)\end{aligned}$$

Given this probabilistic correspondence, the optimal registration is given by maximizing the expectation of observing S under the parameters (transformed centers) of

$T(M)$.

$$\mathcal{L}_{CPD} = - \sum_{i=1}^N \log \sum_{j=1}^M p(T(m_i)) p(s_j | T(m_i)) \quad (2.4)$$

The CPD algorithm is optimized through Expectation-Maximization (EM); at each iteration the correspondence is computed (Expectation Step), and the transformation parameters are updated (Maximization).

Due to the computational cost of computing the expectation per iteration, CPD is slower than ICP but remains highly popular due to its improved robustness. While the CPD’s use of probabilistic correspondences confers greater robustness to missing points and outliers, it struggles when only a partial registration is possible. Similarly to ICP, various extensions have been introduced to combat this weakness, for example by introducing density adaptive sampling to account for variations in point cloud densities [58].

2.4 Optimal Transport for Point Set Registration

In this chapter, we introduce a computational method for spatial integration that applies optimal transport to point set registration. As described in the previous chapter, optimal transport quantifies the distance between distributions by computing the cost associated with the most efficient (or optimal) transport plan that transports the mass from one distribution to another. OT is ideally suited to tissue alignment. In optimal transport, each slice can be represented as an empirical distribution sampled from an underlying density (the 3D tissue). Because these empirical samples are drawn from the same distribution, their optimal transport distance should be small, and we can optimize for transformations that reduce the optimal transport distance between them to recover the optimal alignment.

Furthermore, if we represent each point cloud as a discrete probability distribu-

tion, then obtaining correspondences between point clouds is equivalent to finding the optimal transport plan. As the Kantorovich formulation (Equation 1.6) allows for mass splitting, these correspondences do not require a one-to-one matching between point clouds, and allow for probabilistic assignments between source and target points. While the original OT problem (Equation 1.9) requires mass conservation (e.g. the source and target masses must be equivalent), we can relax this constraint and allow for plans that only transport part of the mass in a formulation known as *partial* optimal transport. This formulation allows us to construct *partial*, probabilistic correspondences between distributions, rendering our metric robust to missing tissue, shape changes, and other spatial distortions and artefacts induced by sample collection or tissue variation. Altogether, these mathematical properties provide an effective metric for measuring alignment quality in point set registration.

2.4.1 Partial Optimal Transport

While there are many variants of both the Optimal and partial Optimal Transport frameworks, for simplicity in this chapter we will use OT and POT to denote the 2–Wasserstein and partial 2–Wasserstein problems, unless otherwise stated. Returning to our notation from the previous chapter, let α, β denote two discrete measures over metric space Ω with a measurable cost function $c : \Omega \times \Omega \rightarrow \mathbb{R}$.

$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta(\mathbf{x}_i), \quad \beta = \sum_{j=1}^m \mathbf{b}_j \delta(\mathbf{y}_j)$$

Then as before, the 2-Wasserstein distance between α, β is defined as

$$\begin{aligned} \text{OT}(\alpha, \beta) &:= \mathcal{W}_2(\alpha, \beta) = \min_{\pi \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \pi \rangle \\ \Pi(\mathbf{a}, \mathbf{b}) &= \{ \pi \in \mathbb{R}_{\geq 0}^{n \times m} \mid \pi \mathbf{1}_m = \mathbf{a}, \pi^T \mathbf{1}_n = \mathbf{b} \} \end{aligned}$$

where $\mathbf{C}_{ij} = \|\mathbf{x}_i - \mathbf{y}_j\|_2^2$, and $\mathbf{a}, \mathbf{b}$ denote the probability mass vectors of α, β . The OT distance requires that α, β have the same probability mass, $\|\mathbf{a}\|_1 = \|\mathbf{b}\|_1$ and that all the mass has to be transported. This requirement is not suitable for domains like the point set registration task, where only partial correspondence is needed.

The *partial* Optimal Transport (POT) problem requires that only a fraction $0 \leq s \leq \min(\|\mathbf{a}\|_1, \|\mathbf{b}\|_1)$ of the mass be transported. In this case, the set of admissible couplings is denoted as $\Pi^s(\mathbf{a}, \mathbf{b})$, and analogously to the OT framework, it requires finding the admissible coupling $\pi \in \Pi^s(\mathbf{a}, \mathbf{b})$ that transports this fraction s with minimum cost.

$$\begin{aligned} \text{POT}(\alpha, \beta) &:= \text{partial } \mathcal{W}_2(\alpha, \beta) = \min_{\pi \in \Pi^s(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \pi \rangle \\ \Pi^s(\mathbf{a}, \mathbf{b}) &= \{\pi \in \mathbb{R}_{\geq 0}^{n \times m} \mid \pi \mathbf{1}_m \leq \mathbf{a}, \pi^T \mathbf{1}_n \leq \mathbf{b}, \mathbf{1}_n^T \pi \mathbf{1}_m = s\} \end{aligned} \quad (2.5)$$

Solving the POT problem is equivalent to solving an OT problem on an “augmented” distribution.

$$\text{POT}(\alpha, \beta) = \text{OT}(\bar{\alpha}, \bar{\beta}) = \min_{\pi \in \Pi(\bar{\mathbf{a}}, \bar{\mathbf{b}})} \langle \bar{\mathbf{C}}, \pi \rangle$$

where $\bar{\mathbf{C}} \in \mathbb{R}^{(n+1) \times (m+1)}$, $\bar{\mathbf{a}} \in \mathbb{R}^{n+1}$, $\bar{\mathbf{b}} \in \mathbb{R}^{m+1}$ denote an augmented cost matrix and extended probability mass vectors respectively, such that

$$\bar{\mathbf{C}} = \begin{bmatrix} \mathbf{C} & \mathbf{0}_n \\ \mathbf{0}_m^T & A \end{bmatrix}, \quad \bar{\mathbf{a}} = \begin{bmatrix} \mathbf{a}, & 1-s \end{bmatrix}, \quad \bar{\mathbf{b}} = \begin{bmatrix} \mathbf{b}, & 1-s \end{bmatrix}$$

and $A > 0$ is an arbitrary scalar [18].

2.4.2 Optimizing POT for Point Set Registration

We model the static and transformed moving point clouds as empirical probability distributions $\mathcal{S}$ and $T(\mathcal{M})$ respectively over locations in $\mathbb{R}^2$:

$$\mathcal{S} \triangleq \sum_{i=1}^n \frac{1}{n} \delta(\mathbf{s}_i), \quad T(\mathcal{M} \mid \Theta) \triangleq \sum_{j=1}^m \frac{1}{m} \delta(T(\mathbf{m}_j \mid \Theta)) \quad (2.6)$$

The loss function is defined as the POT Distance between these distributions, where the (unbalanced) amount of mass to be transported is defined as the ratio between the number of cells in the smaller and larger tissues, $s = \min \left\{ \frac{|M|}{|S|}, \frac{|S|}{|M|} \right\}$. The Point Set Registration Problem is then equivalent to finding the set of parameters $\Theta = \{\theta, t_1, t_2\}$ for the rigid transformation T that minimizes the POT distance between $\mathcal{S}$ and $T(\mathcal{M} \mid \Theta)$:

$$\min_{T \in \mathcal{T}} \mathcal{L}(S, T(M)) = \min_{\Theta} \text{POT}(\mathcal{S}, T(\mathcal{M} \mid \Theta)) \quad (2.7)$$

We solve Equation 2.7 to obtain the optimal alignment with gradient descent using Algorithm 1.

Gradient descent is an iterative optimization algorithm. For a convex loss function and small enough step size η , gradient descent is guaranteed to converge to the global minimum under any initialization [26]. We can prove the convergence of gradient descent under these conditions as follows.

$$\begin{aligned} \mathcal{L}(\Theta^{(k+1)}) &= \mathcal{L}(\Theta^{(k)} - \eta \nabla \mathcal{L}(\Theta^{(k)})) && \text{Parameter Update Rule} \\ &\approx \mathcal{L}(\Theta^{(k)}) - \nabla \mathcal{L}(\Theta^{(k)})^T \eta \nabla \mathcal{L}(\Theta^{(k)}) && \text{First-order Taylor series} \\ &= \mathcal{L}(\Theta^{(k)}) - \eta \|\nabla \mathcal{L}(\Theta^{(k)})\|_2^2 && \text{Definition of } \ell_2 \text{ norm} \\ \Rightarrow \mathcal{L}(\Theta^{(k+1)}) &\leq \mathcal{L}(\Theta^{(k)}) && \text{Non-negativity of } \ell_2 \text{ norm} \end{aligned} \quad (2.8)$$

For point set registration using a rigid transformation, our loss function is not globally convex, as an equivalent global minima can be obtained by modifying the optimal transformation with a rotation of $2n\pi$, where n is any non-zero integer. Thus, our loss landscape can only be locally convex over the restricted transformation domain $\Theta_{restricted} = \{\theta \in [n\pi, n\pi + 2\pi), t_1 \in \mathbb{R}, t_2 \in \mathbb{R}\}$ for some fixed n . Depending on the tissues to be registered, pathological point clouds may induce multiple global or local minima even in a restricted domain. For instance, multiple global minima may arise if one tissue is multiple times larger than the other and constructed of repeated sub-structures, allowing for multiple “matching” locations for the smaller tissue in the larger. Alignments between tissues with drastically varying densities, or irregular protrusions of cells can induce local minima. In this setting, a local loss minima can be achieved by greedily aligning the larger and denser tissue regions, at the expense of the sparser protrusions. Modifying the transformation parameters will increase the loss as the dense tissue “body” will need to move out of alignment before settling into the globally optimal configuration. In practice, even for non-convex or locally convex functions, gradient descent is an effective optimization strategy when combined with multiple, random, initializations. These random initializations allow gradient descent to better explore the loss landscape, and avoid consistently converging to the potentially local minima [97]. The parameter set that produces the lowest loss over all initializations is selected as the optimal solution.

In this Chapter, all experiments represent the performance of our method, as well as ICP and CPD, using an initialization at $\Theta = \{0, 0, 0\}$ for consistency and fairness. In Chapter 4, we apply our POT alignment method to a novel dataset of large, mouse tissues profiled with multiplex IHC, and rerun each alignment multiple times with varying initialization parameters sampled over $\Theta_{restricted}$, and select the transformation parameters that produce the lowest POT loss.

While there are many strategies for computing the gradient in a numerically stable

and computationally efficient way, we employ automatic differentiation in PyTorch [81]. To improve convergence, we z -normalize the point clouds S, M :

$$\begin{aligned} \forall \mathbf{s}'_i \in S', \mathbf{s}'_i &= \left(\frac{\mathbf{s}_i - \mathbb{E}[(P_1)]}{\sigma(P_1)}, \frac{\mathbf{s}_i - \mathbb{E}[(P_2)]}{\sigma(P_2)} \right) \\ \forall \mathbf{m}'_i \in M', \mathbf{m}'_i &= \left(\frac{\mathbf{s}_i - \mathbb{E}[(P_1)]}{\sigma(P_1)}, \frac{\mathbf{m}_i - \mathbb{E}[(P_2)]}{\sigma(P_2)} \right) \end{aligned} \quad (2.9)$$

Here S', M' are the normalized point clouds, $P = S \cup M$ is the union of S, M , and P_1, P_2 are the first and second columns (e.g. x, y coordinates) of P respectively. This ensures that individual parameter updates are of the same magnitude, which helps smooth the loss surface and prevent one gradient direction from dominating the others [17].

Algorithm 1 Gradient Descent POT Optimization for Point Set Registration

Given: point clouds S, M , step-size $\eta > 0$, initial transformation parameters $\Theta^{(0)}$, maximum iterations K .

Set s , the amount of mass to be transported, so that $s = \min\{\frac{|M|}{|S|}, \frac{|S|}{|M|}\}$.

Construct z -normalized point clouds S', M' using Eq. 2.9.

Construct a distribution $\mathcal{S}$ for the normalized scene points S' using Eq. 2.6.

for $k = 1 \dots K$ **do**

 Construct a distribution $\mathcal{T}(\mathcal{M} \mid \Theta)$ for the normalized moving points using Eq. 2.6.

 Compute the loss gradient $\nabla \mathcal{L}(\Theta)$ with respect to the current parameters:

$$\forall i, \nabla \mathcal{L}(\Theta)_i = \frac{\partial \mathcal{L}}{\partial \Theta_i} = \lim_{\epsilon \rightarrow 0} \frac{\mathcal{L}(\Theta_i + \epsilon, \Theta_{\forall j \neq i}) - \mathcal{L}(\Theta_i, \Theta_{\forall j \neq i})}{\epsilon}$$

 where $\mathcal{L}(\Theta) = \text{POT}(\mathcal{S}, T(\mathcal{M} \mid \Theta))$

 Update the transformation parameters so that $\Theta^{(k+1)} = \Theta^{(k)} - \eta \nabla \mathcal{L}(\Theta^{(k)})$.

end for

 Inverse z -normalize the aligned point cloud $T(\mathcal{M} \mid \Theta^{(K)})$.

2.4.3 Improving Computational Efficiency

Computing optimal transport distances between large point clouds can be computationally expensive. This is a particularly relevant challenge in our setting: multiplexed WSIs will have hundreds of thousands of cells, and gradient descent is an iterative algorithm that requires POT computations at each iteration. To reduce the computational cost, we avoid computing optimal transport distances between the original (large) point clouds, and instead use stochastic estimations from smaller samples, effectively exchanging computationally cheaper iterations for a potentially slower convergence rate.

The cheapest and simplest approximation of the optimal transport distance is the *plug in* estimator, which is computed by sampling from the original point clouds S, M and computing the POT distance between samples:

$$\begin{aligned} \text{POT}_{\text{plug in}}(\mathcal{S}, \mathcal{M}) &:= \text{POT}(\hat{\mathcal{S}}_k, \hat{\mathcal{M}}_l) \\ \text{where } \hat{\mathcal{S}}_k &\triangleq \sum_{i=1}^k \frac{1}{k} \delta(\hat{s}_i), \quad \forall \hat{s}_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{S} \\ \text{and } \hat{\mathcal{M}}_l &\triangleq \sum_{j=1}^l \frac{1}{l} \delta(\hat{m}_j), \quad \forall \hat{m}_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{M}. \end{aligned} \tag{2.10}$$

Here k, l denote the number of sampled points from S, M respectively, and are typically chosen to preserve the original ratio in the number of points in S, M , i.e. $\frac{k}{l} \approx \frac{n}{m}$. Because the subsampling process is random, the POT distance will be a random variable, and the resulting gradient computations will be stochastic. For this reason, the corresponding iterative optimization algorithm will be analogous to stochastic gradient descent (SGD).

The plug in estimator is cheap to compute and simple to implement. Given enough iterations, SGD will converge to the same solution as gradient descent [97]. However, the plug in estimator is an extremely noisy approximation of the true POT distance

with a weak convergence rate, i.e. it requires exponentially more samples for a linear reduction in error [117]. Consequently, each gradient step will only be approximately in the right “direction” in parameter space, and the optimization trajectory will be very noisy. In order to converge to global optima, SGD must employ a large number of iterations, or k, l must be very large in order to reduce the approximation error. However, the bottleneck in Algorithm 1 is not constructing $\mathcal{S}, \mathcal{M}$, but the computational cost of repeatedly computing optimal transport distances between large numbers of points. Thus, increasing the number of samples or running for more iterations degrades the benefit of using a stochastic estimator. Instead of increasing k, l to reduce the variance of the POT estimate, a better strategy is to construct an approximation of $\mathcal{S}, \mathcal{M}$, that more accurately captures the underlying spatial structure.

To accomplish this, we employ an algorithm introduced in [8] that draws inspiration from the field of signals processing, in particular *quantization*, a form of compression algorithm that maps values from a large (or continuous) set to a much smaller, finite set. Instead of constructing $\hat{\mathcal{S}}_k, \hat{\mathcal{M}}_l$ by drawing k, l locations from S, M and assigning them uniform weights, we *cluster* the original point clouds S, M , and define their *quantized* distributions to be probability measures located at the cluster centers with weights equivalent to the normalized cluster size:

$$\begin{aligned} \text{POT}_{\text{quantized}}(\mathcal{S}, \mathcal{M}) &:= \text{POT}(\hat{\mathcal{S}}_k, \hat{\mathcal{M}}_l) \\ \text{where } \hat{\mathcal{S}}_k &\triangleq \sum_{i=1}^k \frac{|S^i|}{|S|} \delta(c(S^i)), \bigcup_{i=1}^k S^i = S \\ \text{and } \hat{\mathcal{M}}_l &\triangleq \sum_{j=1}^l \frac{|M^j|}{|M|} \delta(c(M^j)), \bigcup_{j=1}^l M^j = M \end{aligned} \quad (2.11)$$

where S^i denotes all the points in the i -th cluster of S with center $c(S^i)$, and analogously for M^j and $c(M^j)$. In principle, any clustering method can be used, but in practice the k -means clustering [42] is preferred.

k -means is computationally inexpensive, and clusters data by assigning each point to the cluster whose center it is closest to while optimizing the location of cluster centers so that the “spread”, or variance, of each cluster is minimized. This process effectively partitions the data space into Voronoi cells, which ensures a widespread dispersion of centers across the point clouds, thus capturing the underlying spatial structure more effectively. Depending on the size of the tissues, k -means can be applied to the original or subsampled point clouds. However, because k -means clustering is computationally cheap, a large number of samples may be drawn from S, M to find the most representative k, l centers at relatively inexpensive cost. Importantly, this process allows us to compute better approximations of the POT distance than the plug in estimator, while maintaining similar computational requirements: while fast k -means clustering can be performed on up to M, N points, the expensive optimal transport computation is still computed on only $k \ll N, l \ll M$ points.

2.5 Results

2.5.1 The POT Loss Quantifies Alignment Quality

We first demonstrate the performance of our method on simulated tissues where the true alignment is known. We randomly sample 2000 cell locations from a single slice of tissue and divide them into two artificial point clouds called Panels 1 and 2. This ensures that no cell exists on both panels, and no markers are conserved between slices beyond cell location. This construction mimics a multiplex experiment where two, 2D slices are sectioned from the same 3D tissue and stained with different protein markers. To simulate spatial deformations and artefacts induced by sample collection, we artificially corrupt Panel 2, and simulate a dramatic “tear” by removing all of the cells below a certain break point, before randomly rotating and shifting. Together, this synthetic experiment ensures that there is no one-to-one correspondence between

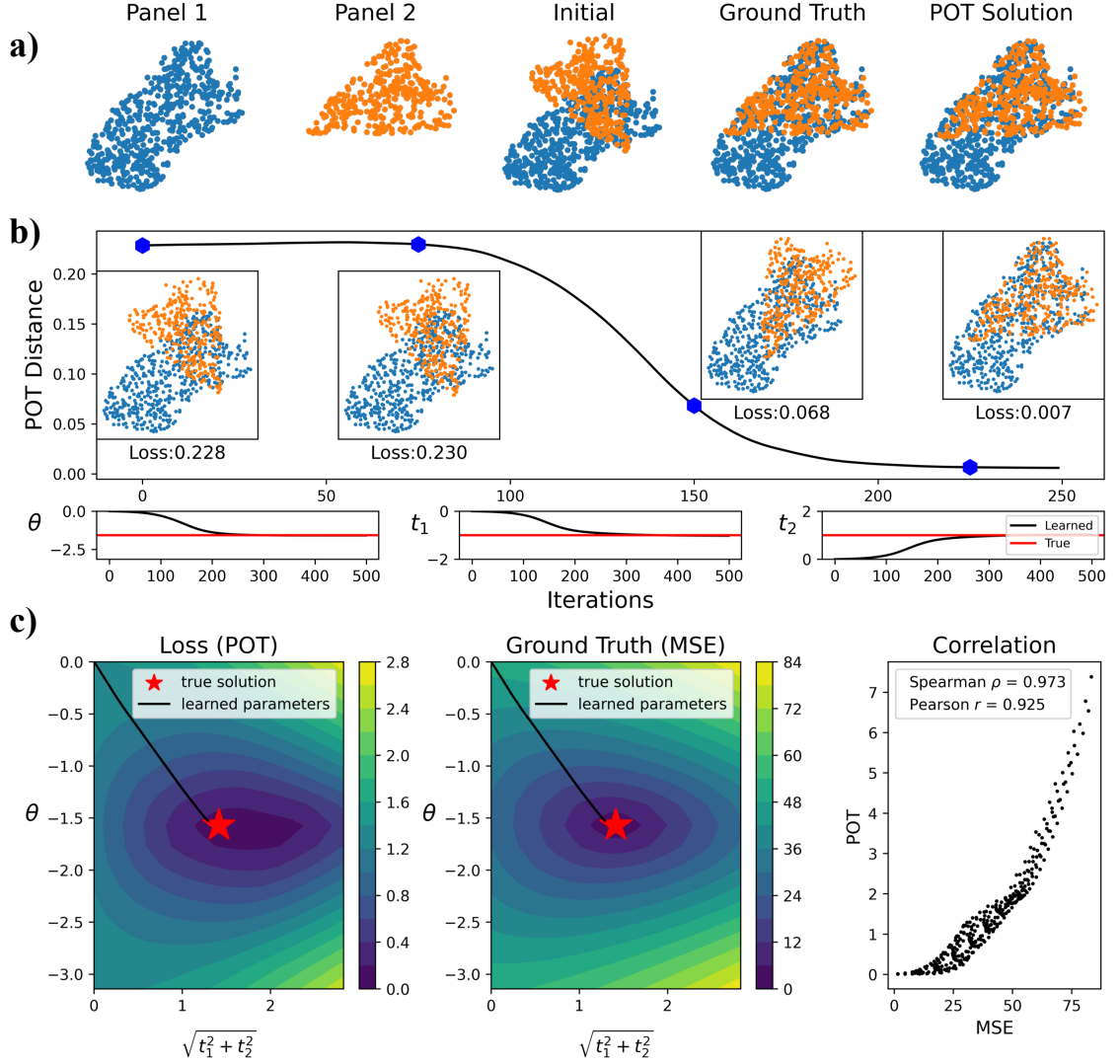


Figure 2.2: **Partial Optimal Transport Alignment.** a) Given two synthetic tissue slices modeled as point clouds, the integration task is equivalent to transforming Panel 1's initial position to the ground truth. b) Using Algorithm 1, we learn the transformation parameters that minimize the POT distance between point clouds. The POT distance accurately quantifies alignment quality, and gradient descent converges to the true parameters. c) The POT distance between Panel 1 and (transformed) Panel 2 is shown across different transformation parameters. In comparison, the Mean Squared Error (MSE) between the (normalized) transformed Panel 2 and its (normalized) ground truth position is also shown. The POT distance is highly correlated with the true alignment error (MSE), even though the true correspondence is unknown.

cells, and simulates an extreme example of partial registration, since only the top half of Panel 1 overlaps with Panel 2. Since we provide the original rotation and shifts, the optimal parameter set Θ^* that returns Panel 2 to its original location is known, allowing us to quantitatively benchmark the performance of Algorithm 1.

This experiment demonstrates that Algorithm 1 recovers the optimal transformation parameters, and that the POT distance can be used to measure alignment quality. In Figure 2.2 (a), we show that the solution to Equation 2.7 is equivalent to the ground truth alignment between slices. In Figure 2.2 (b), we demonstrate that our loss function provides an accurate, quantitative measure of alignment quality, and that the parameters learned by gradient descent converge to their optimal values. In Figure 2.2 (c), we compare our loss function across different transformation parameters with the Mean Squared Error (MSE) between the transformed (normalized) position of Panel 2 and its (normalized) ground truth position. Since our parameter space Θ has three variables, for visualization purposes we combine the x, y shifts t_1, t_2 into a single “distance” variable $\sqrt{t_1^2 + t_2^2}$. As shown in Proof 2.8, the parameter trajectory induced by the gradient step follows the direction of steepest descent, allowing for fast optimization of the loss function. We observe that the global minimum of the POT loss is precisely located at the true transformation parameters Θ^* , providing evidence that the POT loss is an unbiased measure of alignment quality. We further validate this by computing correlation metrics between the POT loss and the ground truth MSE distance. The stricter Pearson correlation yields a correlation of 0.925, while the Spearman correlation, which only requires a monotonic relationship, records 0.973. Both metrics demonstrate that the POT distance is a powerful tool for measuring alignment quality, even when the true correspondence is unknown, and only partial registration is possible.

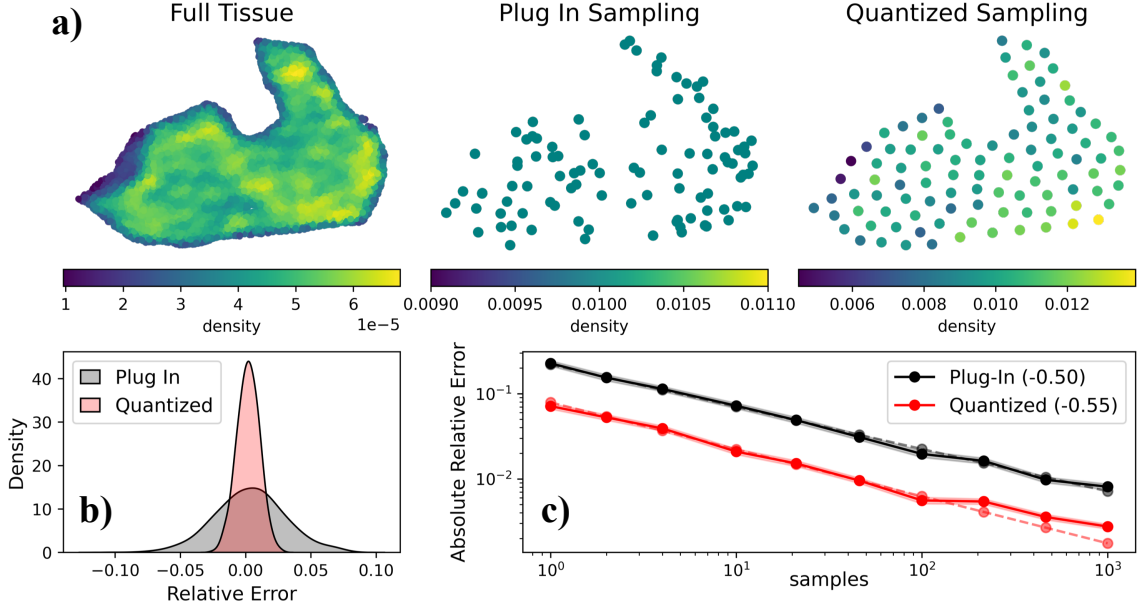


Figure 2.3: **Estimator Comparison.** **a)** The leftmost panel shows the full tissue consisting of over 100,000 cells, with spatial locations colored by relative cell density. The next two panels show the plug in and quantized representations of this distribution with 100 samples, with points colored by weight. **b)** For the same number of samples, ($k = 100$), we compare the relative error between the analytic POT distance, and the approximate distance estimated by plug in and quantized sampling. While both are reasonably unbiased, the quantized estimator has a greatly reduced variance. **c)** We compute the absolute relative error as a function of sample size. The quantized estimator is always more accurate, and converges slightly faster than the plug in.

2.5.2 The Quantized Estimator Enables Fast and Accurate Alignment

We next demonstrate the improved efficiency of the quantized estimator. In this experiment, we compute the POT distance between a tissue X consisting of over 100,000 cells and a shifted version of itself, $T(X \mid \{0, t_1, t_2\})$. While calculating the POT distance between such large datasets is computationally challenging, the analytical POT distance between any distribution and a shifted version of itself is given by the magnitude of the shift [83], in this case $\sqrt{t_1^2 + t_2^2}$. For simplicity, we set the number of samples from each point cloud to be equal so that $k = l$, and the

amount of transported mass to $s = 1$, and compare the accuracy of the quantized and plug in estimators for increasing larger k . For the quantized estimator, we perform k -means clustering on $10k$ samples drawn randomly from X .

This example demonstrates the improved performance of the quantized estimator. In Figure 2.3 a), we show that the plug in sample does not properly account for variations in tissue density, and does not capture the spatial structure of the entire tissue. In contrast, the quantized sample accurately captures the entire tissue structure complete with varying density. Figure 2.3 b) demonstrates that the quantized estimator has a much smaller variance than the plug in estimator at an equivalent number of samples. This variance reduction can be explained by the visualizations shown in the previous panel. Because the plug in estimator uses random samples, the sample distributions $\hat{\mathcal{S}}, \hat{\mathcal{M}}$ struggle to comprehensively capture the overall tissue structure and produce noisy estimates of the POT distance. In contrast, the clustering step in the quantized estimator produces stable cluster centers, reducing the variability in the POT estimate. Figure 2.3 c) demonstrates that the quantized estimator is over $10\times$ more accurate than the plug in estimator for all sample sizes. Furthermore, by computing the slopes of the log-log plot of error versus sample size, we can estimate their convergence speed as a function of sample size. We observe that the quantized estimator converges approximately 11% faster than the plug-in estimator. Again, this behavior can be explained by the sampling strategy: the plug in estimator improves its approximation of spatial structure by gathering more random data points. In contrast, the quantized estimator adds new cluster centers, progressively increasing the “resolution” of the approximated spatial structure.

In Figure 2.4, we demonstrate the impact of the POT estimation strategy on tissue alignment. Similar to the our construction in the previous section, we generate synthetic data with artificial deformations that simulate a spatial proteomics workflow on two 2D slices, before randomly rotating and shifting one point cloud, to determine

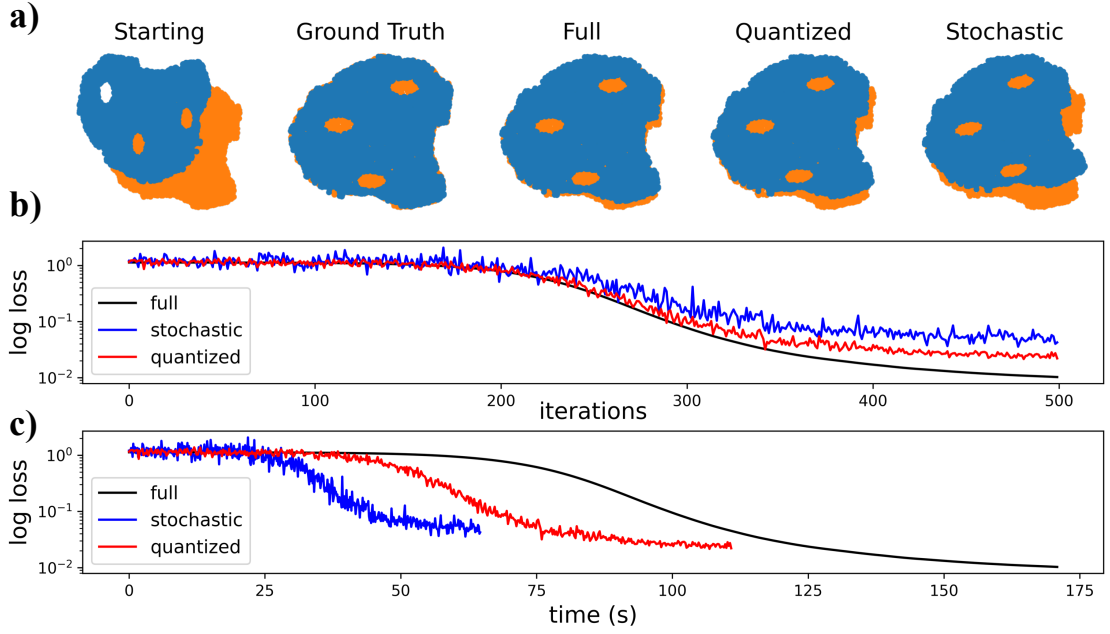


Figure 2.4: **Comparison of Estimator Strategies on Spatial Alignment.** In this plot, “Full” refers to Algorithm 1 using a gradient computed between complete point clouds, while “Stochastic” and “Quantized” refer to gradients calculated using the plug in and quantized estimators respectively. **(a)** Tissue alignment results for different estimators after 500 iterations, compared with the ground truth and original starting positions. POT Loss convergence as a function of iterations **(b)** and time **(c)** for different strategies. Since the Full POT loss computation is deterministic, its optimization trajectory is noiseless. In contrast, the plug in and quantized estimators are stochastic, and yield noisy optimization trajectories. The reduced noiseness of the quantized trajectory in comparison to the stochastic trajectory can be explained by the reduced variance of the quantized estimator shown in Figure 2.3 b).

whether our methods can recover the optimal transformation parameters. In this example, we increase the number of cells sampled to 10,000, limited only by our need for comparison with the baseline POT solver, and introduce artificial holes to simulate accidental punctures during the sample collection process, or cavities in a 3D tissue. We apply Algorithm 1 to the point clouds, and vary the gradient computation strategy between “Full”, “Stochastic”, and “Quantized”. In the “Full” setting, the POT Loss is computed between the complete point clouds, while the “Stochastic” and “Quantized” settings refer to losses computed between sampled point clouds using the plug in and quantized estimators respectively with $k = l = 100$ samples.

This experiment highlights several interesting trends. In Figure 2.4 a), we can observe that for a fixed number (500) of iterations, POT optimization with the Full gradient computed between complete point clouds yields the best alignment, with the Quantized gradient falling slightly behind, and the Stochastic gradient performing worst. In Figure 2.4 b), we observe that the Full gradient converges much faster than the other two methods, achieving a final POT loss $100\times$ smaller than the Stochastic gradient, and $10\times$ smaller than the Quantized gradient after 500 iterations. However, analysis of the CPU time required to compute 500 iterations in Figure 2.4 c) reveals the computational acceleration induced by the different strategies. The Stochastic Gradient is approximately $3\times$ faster than the Full Gradient, and almost $2\times$ faster than the Quantized. Altogether, this experiment demonstrates the tradeoff between a slower convergence rate and faster iterations. For larger tissues, this tradeoff is no longer optional, as they contain too many cells for a Full Gradient to be computed. In this case, the combination of the faster loss convergence by the Quantized gradient as a function of iterations compared to the Stochastic gradient, and the faster convergence of the Quantized Estimator as a function of samples compared to the Plug-In Estimator, makes quantization a powerful strategy for integrating massive tissues.

2.5.3 POT Alignment Outperforms ICP and CPD

Finally, we compare our method to the ICP and CPD algorithms for spatial integration under increasingly challenging spatial distortions. As before, we sample 2000 points from a single tissue slice and divide them between two artificial point clouds. For each alignment task, we “tear” one tissue, and progressively remove larger and larger sections, minimizing the overlap between slices. The small number of points sampled was chosen to amplify the effect of the reduced overlap. We then randomly rotate and shift the torn tissue, and apply ICP, CPD, and our method to perform point set registration and recover the optimal transformation parameters that reverse the random transformation. We compare our algorithm with robust implementations of ICP and CPD. The ICP implementation [22] includes RANSAC for outlier detection, and the CPD implementation uses a Bayesian density adaptive strategy to compensate for severe density variations [58]. While the number of points is small enough for full gradient POT optimization, we employ Quantized gradient descent fixed at 500 iterations to more accurately reflect the speed considerations encountered in real tissue alignment.

The results presented in Figure 2.5 (a) demonstrate that while the performances of ICP and CPD rapidly deteriorate, our method is robust to challenging artefacts. As the degree of overlap decreases, the MSE between the normalized ground truth tissue and the final transformation increases exponentially for both CPD and ICP. In contrast, the MSE for our method increases linearly, and outperforms both CPD and ICP on all overlap conditions. To understand this behavior, Figure 2.5 b) shows the final alignments for tissues with a 60% overlap. While the starting position of the moving point cloud (orange) is not shown, ICP simply moves the moving point cloud from its original, unaligned position to the closest points in the (blue) static point cloud. This behavior is due its assumption of nearest-neighbor correspondence. CPD performs slightly better, but the probabilistic correspondence generated us-

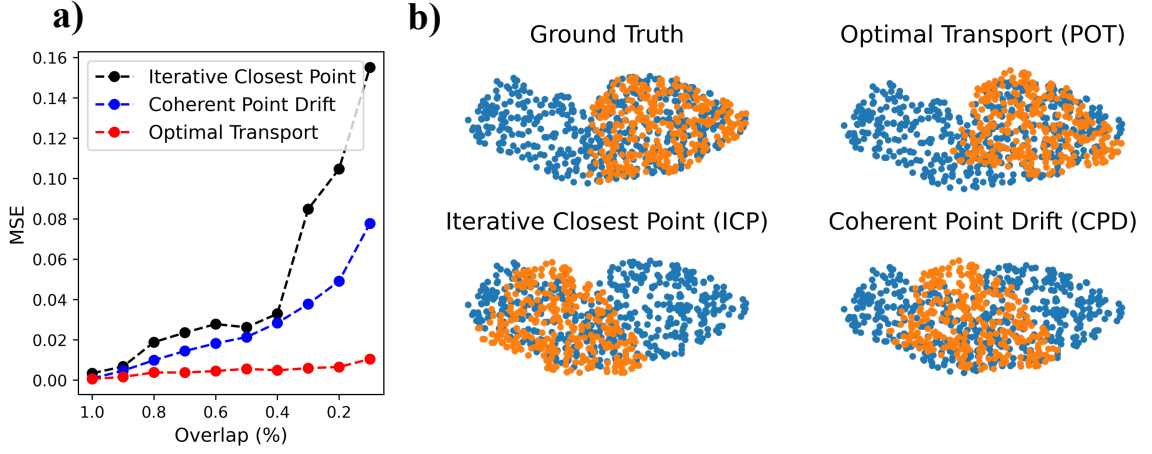


Figure 2.5: **Methods Comparison.** **a)** We compare the alignment quality between ICP, CPD, and our method (OT) under increasingly smaller overlaps. As before, the MSE is computed between the final position of the normalized moving point cloud and its normalized ground truth position. **b)** Alignment quality for a 60% overlap. The orange cells represent the moving point cloud.

ing Expectation-Maximization attempts to maximize the correspondence across all points, and converges to a failed solution that effectively aligns the centers of the point clouds. In contrast, our method of aligning tissues using *partial* Optimal Transport allows for the transportation of *fractional* mass between distributions, thus allowing for partial registration. While our alignment is imperfect, due to the slow optimization induced by the Quantized gradient which has not converged after 500 iterations, it is the only method that successfully identifies the correct overlapping regions.

2.6 Conclusion

In this chapter, we introduced a computational method for integrating multiple spatial measurements, by applying quantized partial optimal transport alignment to point set registration. Our method enables the analysis of diverse cell types across large tissues, by automatically aligning multiplex measurements taken from tissues stained with

different markers and affected by spatial deformations induced by sample collection. To scale our method to WSIs with hundreds of thousands of cells, we introduced a quantized estimator for the POT distance that operates with a fraction of the original cost. We benchmark our spatial integration method against state of the art algorithms, and demonstrate our method’s improved robustness to rips, tears, and other challenging spatial distortions commonly present in multiplex datasets. In this chapter we have focused on evaluation, and as such have mainly demonstrated our method on small, synthetic datasets where ground truth alignments are known. In Chapter 4, we will apply our method to a dataset of mouse tissues profiled with multiplex IHC, and align massive tissues consisting of hundreds of thousands of cells.

Chapter 3

Characterizing Cell Microenvironments with Graph Optimal Transport

3.1 Introduction

Cell microenvironments are complex biological systems comprising diverse, co-localized, and interacting cell types. Increasingly, microenvironments are recognized as crucial components in governing immune response, tumor progression, and tissue function [39, 28, 86]. Advancements in multiplexed imaging technology along with the widespread adoption of standardized, automated image pre-processing technologies, have enabled the observation of cellular microenvironments within their native spatial context [72, 44, 102, 62, 61]. Although these technologies provide rich, highly resolved information, analysis of the resulting spatial data poses new challenges. In particular, how to identify microenvironments from multiplex images and characterize their relevance to tissue architectures and/or diseases are important open questions.

Several strategies exist for the identification of distinct microenvironments. One

method is to apply techniques from single cell transcriptomics. These techniques typically employ unsupervised clustering of high-dimensional molecular expression data to identify enriched marker genes within clusters that can be annotated and queried for biological interpretation using curated biomedical ontologies (GO, KEGG) [63, 23, 50]. Unfortunately, these methods cannot account for spatial information, and are not ideally suited for multiplex proteomics data, which have far fewer markers than transcriptomic technologies. Another promising strategy is machine learning, which frames microenvironment identification as a graph or image classification problem [120, 66, 99, 36]. While effective in supervised settings, these methods require massive quantities of labeled data or otherwise annotated information which may not be present in most domains of interest, where the biological ground truth is unknown and is itself the object of discovery. Feature engineering with spatial statistics is another approach. These methods involve computing statistical descriptors for different spatial features or quantities of interest (e.g. attraction, exclusion, mixing etc.). Unfortunately, these methods are difficult to analytically parameterize for complex microenvironments with diverse cell types and, at present, are limited to correlations between pairs of cell types [80, 15, 14].

In this chapter, we introduce a computational strategy for the *unsupervised* characterization of cell microenvironments relying on statistical distance metrics. Our method accounts for both spatial and cell type information by applying graph optimal transport to labeled spatial graphs. To the best of our knowledge, our method is the first application of graph optimal transport to microenvironment characterization. We evaluate our methods on real and synthetic datasets, and demonstrate their utility in mapping tissue architectures and identifying disease relevant microenvironments. On synthetic data, we demonstrate that our algorithm can discern changes in both cell type composition and spatial organization within cell microenvironments. On healthy lung tissues profiled by multiplex IMC, our method identifies spatially

varying microenvironments that map to distinct functions within complex tissue architectures, and directly correspond with pathologist annotations. On CODEX profiled Head Neck Squamous Cell Carcinoma (HNSCC) tumors, our method identifies disease-relevant microenvironments that are highly predictive of patient outcome. We benchmark our algorithm against informative baselines, as well as state of the art competitive algorithms.

3.2 Related Work

In its simplest form, a microenvironment can be described as a group of cells of one or more cell types, organized in a certain spatial pattern. Together these two dimensions, cell type and spatial organization, provide rich insights into the cell-cell interactions and signalling that drive tissue architectures and disease specific environments. Thus, the computational paradigm by which these two sources of information are integrated and analyzed is a crucial algorithmic design choice in microenvironment characterization. In this section, we introduce two state-of-the-art algorithms with unique approaches: Spatial Latent Dirichlet Allocation (Spatial LDA) [21], and Unsupervised Tissue Architecture Discovery with Graphs (UTAG) [53].

3.2.1 Spatial LDA

3.2.1.1 Summary

Spatial LDA is directly inspired by Latent Dirichlet Allocation (LDA), a popular probabilistic model originally designed for natural language processing [12]. Spatial LDA assumes that cell microenvironments are invariant to the local organization of cells, as any pair of cells within a small neighborhood can directly communicate with each other using diffusive signalling molecules. Using this approximation, cell

microenvironments can be described as a “bag of cells”, which counts the abundances of each cell type within a fixed radius from a center cell, and encodes this information as a fixed length, normalized frequency vector. A tissue is then a collection of different “bags”, and can be described a distribution of microenvironments. An LDA model attempts to parameterize this distribution and learn its latent factors or “topics”. These “topics” are viewed as “canonical” or idealized microenvironments, and each cell microenvironment can be described as a mixture of topics.

3.2.1.2 Method

LDA models tissues as distributions of microenvironments, which are themselves represented by distributions of latent topics, where each topic is a distribution of cell types. More precisely, a tissue $\mathcal{D}$ comprises M microenvironments w , consisting of N cells, each of which is associated with a topic z . There are K unique topics (user defined parameter), and V unique cell types. Each topic is defined by its preference for particular cell types, and is described by a Multinomial distribution. The preferences of all unique topics across all unique cell types is captured by the matrix $\beta \in \mathbb{R}_{\geq 0}^{K \times V}$. The distribution of topics is given by the vector $\theta \in \mathbb{R}^{1 \times K}$, which is drawn from a Dirichlet prior parameterized by $\alpha \in \mathbb{R}^{1 \times K}$. Together, α, β are parameters that describe the distribution and types of microenvironments present in a tissue respectively, and must be learned.

In this model, the joint distribution of all the microenvironments in a tissue can be expressed as follows.

$$p(\mathcal{D}|\alpha, \beta) = \prod_{d=1}^M \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_{dn} \quad (3.1)$$

Microenvironments are characterized by their topic mixture, which can be obtained by computing the posterior likelihood of observing a particular topic given the com-

position observation.

$$p(\theta|w, \alpha, \beta) = \frac{p(\theta, w|\alpha, \beta)}{p(w|\alpha, \beta)} = \frac{p(\theta|\alpha) \prod_{i=1}^N p(z_n|\theta)p(w_n|z_n, \beta)}{\int p(\theta|\alpha) \left(\prod_{i=1}^N \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta) \right) d\theta} \quad (3.2)$$

The task of applying spatial LDA to a tissue involves (i) constructing microenvironments w by calculating the relative abundance of every cell type within radius R of the center cell, over all center cells (ii) fitting an LDA model by optimizing for α, β that maximize the likelihood of Equation 3.1 over all the collected microenvironments, (iii) assigning topic probabilities to individual microenvironments by solving Equation 3.2. Unfortunately, computing steps (ii) and (iii) involves an analytically intractable integral, and thus exact inference is not possible. In practice, both equations can be approximately solved through Variational Inference, which iteratively maximizes the Evidence Based Lower Bound. This optimization problem is highly non-convex and difficult to solve, and scales poorly with increasing numbers of microenvironments [12].

3.2.1.3 Analysis

Spatial LDA is a complex algorithm with many unique characteristics. Since it requires a discrete vocabulary of possible cell types, Spatial LDA cannot account for continuous expression values, and can only be applied to tissues post cell-type assignment. However, Spatial LDA characterizes each microenvironment as a *mixture* of topics, allowing users to visualize environmental *gradients* across tissues rather than arbitrarily distinct spatial domains. Unfortunately, the number of topics must be pre-selected, and adding or removing a topic requires the entire model to be retrained. Since learning a new Spatial LDA model requires solving a non-convex Variational Inference optimization problem, this computational bottleneck prevents users from quickly fitting new models and observing microenvironment subtypes at differing lev-

els of granularity.

3.2.2 UTAG

3.2.2.1 Summary

Instead of assuming that communication within a microenvironment is diffusive, UTAG describes interactions as directed between neighboring cells. In this setting, neighbors are not determined by a fixed radius, but instead by the k -closest cells to a center cell. This structure can be modeled as an S_k star graph, with the center cell as the center node, and its k neighbors forming the leaves. To avoid extraneous neighbors, edges can also be thresholded by absolute physical distance. Within this star graph, nodes representing cells are endowed with feature vectors, which are either “one-hot” vectors denoting cell type, or else continuous vectors representing marker intensity (e.g. protein abundance). Communication between cells is then modeled by passing node information across edges and summing in a process called “message passing”, to create spatially aggregated features.

3.2.2.2 Method

Formally, given a cell x described by d features (either one-hot for cell type, or continuous for marker intensity), $x \in \mathbb{R}^d$, the spatially aggregated feature vector $y \in \mathbb{R}^k$ constructed by message passing is defined as

$$y = x + \sum_{i=1}^k x_i \quad (3.3)$$

where $\{x\}_{i=1}^k \in \mathbb{R}^d$ are the features of the k nearest neighbors of x . The microenvironment centered around cell x is then represented by the vector y , its spatially aggregated feature vector.

Message passing is scalable to large tissues. Consider a tissue with n cells and d

features per cell (either markers or unique cell types). To compute message passing efficiently over all n cells, Equation 3.3 can be rewritten as a graph diffusion operator by vectorizing over all cell positions and features. Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the unweighted adjacency matrix of a (possibly thresholded) k Nearest Neighbors spatial graph constructed from the n cell center positions, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ the cell feature matrix. Then the spatially aggregated feature matrix $\mathbf{Y}$ over all cells is simply given by the inner product

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \in \mathbb{R}^{n \times d} \quad (3.4)$$

This inner product is efficient to compute, and the main computational bottleneck in the UTAG algorithm is the cost of storing the adjacency matrix, which grows quadratically with the number of cells in the tissue, though this can be remedied through matrix partitioning.

From this spatially aggregated feature matrix, individual microenvironments are grouped into canonical or representative niches by Leiden clustering [111]. The Leiden method is a network community detection algorithm that finds groups of nodes, or communities, that have a greater number of edges within the community relative to edges crossing outside, through an iterative optimization process termed modularity maximization. To convert the spatially aggregated feature matrix $\mathbf{Y} \in \mathbb{R}^{n \times d}$ into a network, a distance matrix D is computed by comparing the spatially aggregated features between every pair of microenvironments,

$$D \in \mathbb{R}^{n \times n}, \text{ s.t. } D_{ij} = \|\mathbf{Y}_i - \mathbf{Y}_j\|_2^2 \quad (3.5)$$

From this distance matrix, a k -Nearest Neighbors graph is constructed by representing every microenvironment as a node, and connecting it to its k most similar microenvironments as determined by D . This network is then passed into the Leiden

community detection algorithm, which then assigns a community (cluster) to every node (microenvironment).

3.2.2.3 Analysis

The UTAG algorithm offers an intriguing contrast with Spatial LDA. Unlike Spatial LDA, UTAG’s message passing method for constructing spatially aggregated features is compatible with continuous expression values, and thus applicable to tissues both pre and post cell type assignment. UTAG is also computationally inexpensive, by replacing the naive definition of message passing with a vectorized graph diffusion operator, the spatially aggregated feature matrix (Equation 3.4) can be quickly computed even for massive tissues. Furthermore, the Leiden clustering algorithm does not required a predetermined number of clusters, but does accept a resolution parameter that allows users to control the granularity of the clusters provided. While the Leiden algorithm can be slow for extremely large networks, Equation 3.4 does not need to be recomputed in order to re-cluster microenvironments, allowing users to compute features over large tissues once, and then cluster at different resolutions as needed. However, regardless of resolution, Leiden clustering can only return discrete cluster assignments. While Spatial LDA represents microenvironments as topic mixtures that can be interpreted as environmental gradients, UTAG’s microenvironment clusters only allow for the characterization of discrete spatial domains.

3.3 Characterizing Microenvironments with Graph Optimal Transport

In this section, we introduce a computational strategy for the *unsupervised* characterization of cell microenvironments relying on optimal transport. Our contributions are organized as follows: (i) we first introduce a mathematical paradigm for *describing*

microenvironments using graphs, (ii) we then describe a method for *comparing* microenvironments, and explicitly quantifying their similarity using optimal transport, and (iii) we explain how to *interpret* these comparisons, and relate them to biological applications of interest. Altogether, these methods provide a general framework for the identification of microenvironments from multiplexed images, and the analysis of their role in tissue architectures as well as disease; the framework is summarized in Figure 3.1. In addition, we provide a comprehensive comparison between the paradigms, limitations, and strategies employed by our method, Spatial LDA, and UTAG in Table 3.1.

3.3.1 Microenvironments as Spatial Graphs

Unlike Spatial LDA, we do not assume that all pairs of cells within a microenvironment can communicate with equal ease or likelihood. Unlike UTAG, we do not restrict communication to neighboring cells. Instead, we assume that cells can communicate with their direct neighbors and beyond, potentially across vast distances, but the *strength* of these interactions, or the influence that they have on each others behavior, is inversely proportional to their proximity.

Under this paradigm, we first construct a tissue-wide cell interaction network, and describe microenvironments as local subnetworks of closely interacting cells from this larger graph. The tissue graph is both a labeled and weighted spatial graph, and is constructed as follows (Figure 3.1).

1. from the cell center positions, the 2D space is partitioned using a Voronoi tessellation.
2. Voronoi polygons are connected if they share a facet, forming a Delauney triangulation [59].
3. Each edge is weighted by the physical distance separating cell centers, and

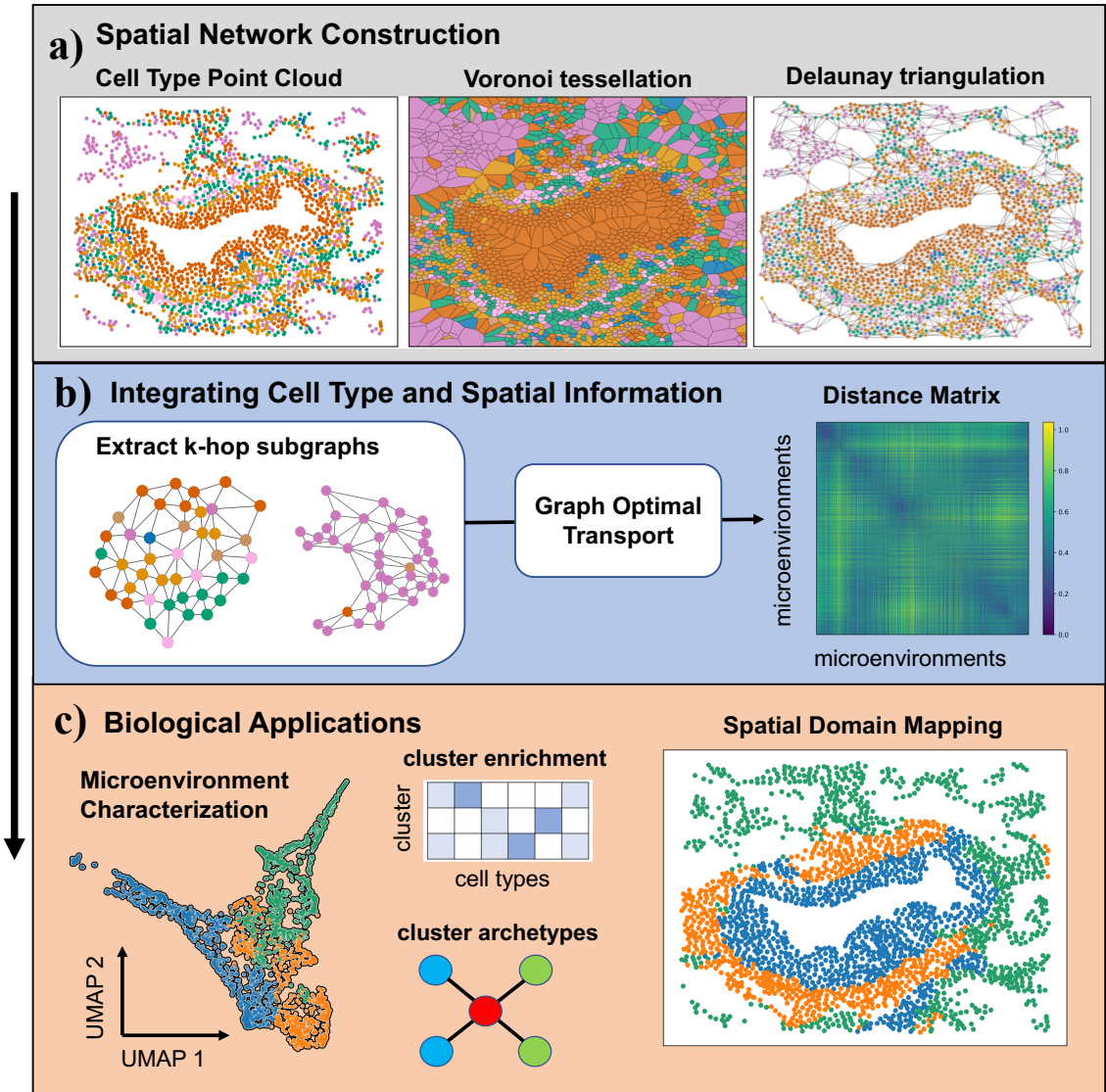


Figure 3.1: **Overview.** (a) We first construct a weighted spatial graph from the original cell centers using Voronoi tessellations followed by Delaunay triangulations. (b) From this tissue graph, microenvironments are defined as induced subgraphs. A set of subgraphs can then be compared using graph optimal transport to form a distance matrix. (c) This distance matrix can be clustered to group microenvironments into canonical subtypes, which can then be investigated for various biological applications, such as cell type enrichment, archetypal cell organization, or spatial domain mapping.

removed if it is larger than some maximum radius.

4. Each node is endowed with a feature vector representing either cell type (one-hot vector) or marker expression (continuous)

The resulting graph can be interpreted as an interaction graph, with edges connecting cells that are likely to be interacting based on proximity and blocking (e.g. no edge will cut "through" another cell), and the weight of the edge inversely encoding the strength of this interaction (e.g. distance). From this tissue-wide interaction network, a microenvironment can be described as a concentrated subnetwork of strong, local interactions, and is constructed as an induced subgraph around a center node.

This approach has the following advantages. Firstly, choosing to model microenvironments with a graph based paradigm allows us to capture relevant biological information without a prohibitively large set of descriptors. That is, the number of equivalent multicellular patterns generated by translation, rotation, or mirroring is vast, but since graphs are invariant to such group actions, we can capture all of these equivalent patterns with a *single microenvironment*. Secondly, describing tissue organization as a weighted and planar graph, allows us to capture varying *degrees of interaction*. In the tissue interaction network, cells are connected to other cells even across vast regions. However, their degree of communication or influence on each other is directly related to the number and weight of the edges in the shortest path connecting them. Lastly, our microenvironments are *flexible in size*, allowing us to capture the density of the underlying tissue. While all UTAG microenvironments have k neighbors, and all Spatial LDA microenvironments are frequency normalized bags-of-cells, a subgraph in a dense region of tissue will have more nodes and edges, whereas a subgraph in a sparser region will contain fewer cells and sparse interactions, or even only a single cell. Together, these advantages offer a powerful way of describing microenvironments with both cell type and spatial information, while respecting their role in overall tissue organization.

3.3.2 Comparing Microenvironments with the Fused Gromov-Wasserstein Distance

Under the model introduced in the previous section, a microenvironment can be described as a labeled graph with weighted edges. Thus, *comparing* two microenvironments is equivalent to quantifying the similarity, or computing the distance between labeled graphs. Quantifying graph similarity is a common problem in network science, and most approaches for this task try to convert a labeled graph into a fixed length vector, termed its embedding, which can then be compared with other graph embeddings using ℓ_p distance metrics or statistical tests [16]. These embeddings can be learned with machine learning algorithms (e.g. Graph Neural Networks), explicitly constructed through network summary statistics (e.g. clustering coefficient, diameter etc), or defined using spectral or topographical descriptors (e.g. graph Laplacian spectrum, random walk dynamics) [19].

Unfortunately, each of these strategies suffer from limitations that restrict their applicability to our setting of microenvironment characterization. In order to learn an embedding, GNNs must optimize some objective function that quantifies how accurately embedding distances reflect similarities in the original graphs. This results in a circular problem, and most GNNs replace a direct notation of distance with some kind of external objective, for example by comparing how well embeddings can be used to separate graphs into different classes defined by human annotation [121]. This supervised requirement cannot be met for microenvironment characterization, since the role or class of a given microenvironment is unknown and the object of discovery. While network summary statistics are selected and defined independently of external labels, this approach is not robust as similar statistics do not guarantee similar network structures [85, 103]. Conversely, spectral methods [127, 40] that compare the spectrum between graph adjacency or Laplacian matrices suffer from abnormal sensitivity, as small changes in network structure can radically alter the

spectrum [103]. Diffusion methods [112] compare the dynamics of random walks along each graph and are more robust, however neither spectral nor diffusion methods can account for node attributes, and instead focus solely on network structure. This is a severe limitation, as these methods will not be able to account for cell type.

In our method, we circumvent these limitations and directly compute distances between *labeled graphs without supervision* using graph optimal transport, and in particular the Fused Gromov-Wasserstein Distance metric [108]. While there may be other applications of optimal transport to graphs, in this work we will use Graph Optimal Transport (Graph-OT), Fused Gromov-Wasserstein (FGW) Distance, and Fused Gromov-Wasserstein Optimal Transport (FGW-OT) interchangeably unless otherwise stated.

In FGW, graphs are considered as probability measures. The probability measure space of a graph $\mathcal{G}$ is described by the triplet $(\mathbf{X}, \mathbf{A}, \mu)$. Here $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^{n \times d}$ represents the set of n nodes, each described by d features. $\mathbf{A} \in \mathbb{R}_{\geq 0}^{n \times n}$ is a square matrix that describes the structure or topology of a graph, and $\mu \in \mathbb{R}^n$ describes a probability vector over the nodes. Let two graphs $\mathcal{G}_1, \mathcal{G}_2$ be differently sized graphs with n, m nodes respectively, described by feature, structure, and probability triplets $(\mathbf{X}, \mathbf{A}, \mu), (\mathbf{Y}, \mathbf{B}, \nu)$. The FGW distance computes the distance between $\mathcal{G}_1, \mathcal{G}_2$ by finding a soft coupling between the nodes of $\mathcal{G}_1$ and $\mathcal{G}_2$ that minimizes the distance between their features, while simultaneously preserving their relative distances to other nodes within their respective graphs. Namely,

$$\begin{aligned} \text{FGW}(\mathcal{G}_1, \mathcal{G}_2) &:= \min_{\pi \in \Pi(\mu, \nu)} \sum_{i,j,k,l} (1 - \alpha) c(\mathbf{x}_i, \mathbf{y}_j) + \alpha \|\mathbf{A}_{ik} - \mathbf{B}_{jl}\|^2 \pi_{ij} \pi_{kl} \\ \Pi(\mu, \nu) &= \{\pi \in \mathbb{R}_{\geq 0}^{n \times m} \mid \pi \mathbf{1}_m = \mu, \pi^T \mathbf{1}_n = \nu\} \end{aligned} \quad (3.6)$$

here $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$ is a metric cost function that quantifies the dissimilarity between node features, and $\alpha \in [0, 1]$ is a scalar parameter that controls the balance

between node and structure information.

The formulation in Equation 3.6 provides an extension to the field of network science by extending network comparisons to *labeled* graphs, i.e. networks with node attributes, by applying a *transportation* or “assignment” based strategy for comparing networks. This strategy tries to find a probabilistic assignment between nodes that minimizes their feature distances, as well as their respective “relationships” within their graphs. When $\alpha = 0$ or when the networks have the exact same topology but different features, the FGW distance is equivalent to a (soft) assignment problem that tries to match nodes with similar features. Conversely, when $\alpha = 1$ or when all node features are identical, FGW tries to match nodes that have similar intra-graph relationships, and reduces to measuring how far two networks are from being isometric. More precisely, given node $a_i \in \mathcal{G}_1$, and node $b_j \in \mathcal{G}_2$, their respective intra-graph “relationships” can be summarized as $\{d(a_i, a_{k \neq i}), \forall a_k \in \mathcal{G}_1\}$ and $\{d(b_j, b_{l \neq j}), \forall b_l \in \mathcal{G}_2\}$ respectively, where $d(., .)$ is defined as either shortest path distance or adjacency. This comparison of the pairwise node distances within graphs bears resemblance to random walk or graph diffusion methods. FGW differs from these methods by including an optimal assignment problem that tries to find the optimal coupling between nodes so that this difference is minimized. Overall, the key characteristics of the FGW distance that differentiate this method from other strategies are (i) the introduction of an assignment or *optimal transportation* problem, and (ii) the combination of feature *and* structure information to enable an extension to labeled graphs.

This formulation enjoys several properties consistent with optimal transport theory. First, if $\mathcal{G}_1, \mathcal{G}_2$ are isometric, then the transportation matrix π is a square permutation matrix representing a bijection between the nodes of $\mathcal{G}_1, \mathcal{G}_2$, and the FGW cost is precisely 0. Secondly, the FGW distance is symmetric. The distance $\text{FGW}(\mathcal{G}_1, \mathcal{G}_2)$ is identical to $\text{FGW}(\mathcal{G}_2, \mathcal{G}_1)$, since the optimal transportation matrix π for one distance is the transpose of the other. Strictly speaking, FGW does not satisfy the

triangle inequality for all choices of c . However, it can be proved that FGW satisfies the relaxed triangle inequality, meaning that the FGW distance between any pair of graphs in a triplet is upper bounded by a constant scaling of the sum of the distances between the other two pairs. Thus, FGW defines a semi-metric on the space of labeled graphs.

The probability measures μ, ν encode the relative importance between nodes within a graph, and the α parameter allows the FGW distance to interpolate between a purely structural comparison and a purely feature based comparison. In practice, most implementations of the FGW distance define the (node) feature distance metric c as a 2-Wasserstein distance, choose μ, ν to be uniform distributions $\mu = \frac{1}{n}\mathbf{1}_n, \nu = \frac{1}{m}\mathbf{1}_m$ over the nodes, and set $\alpha = 0.5$ to be a perfectly balanced trade-off between structure and node features. Unless otherwise stated, our methods also use the above parameters. While more optimal parameter choices are perhaps possible through external knowledge or brute force parameter tuning, in practice, such knowledge is unlikely to be available beforehand, and most likely highly specific to the underlying graphs. As such, we elect to employ these “standard” parameter choices so as to benchmark the performance of our algorithms in their most generalizable configuration.

Solving the the FGW distance (Equation 3.6) involves a quadratic optimization problem with respect to the coupling or transportation matrix π . If the 2-Wasserstein distance is applied as the feature cost function, it can be shown that the complexity of this problem can be reduced from $\mathcal{O}(m^2n^2)$ to $\mathcal{O}(mn^2 + nm^2)$ [84, 83]. In this case, Equation 3.6 is equivalent to finding optimal transportation plan π^*

$$\pi^* = \arg \min_{\pi \in \Pi(\mu, \nu)} \text{vec}(\pi)^T Q(\alpha) \text{vec}(\pi) + \text{vec}(D(\alpha))^T \text{vec}(\pi) \quad (3.7)$$

where $Q = -2\alpha(\mathbf{A} \otimes \mathbf{B})$ and $D(\alpha) = (1 - \alpha)\mathbf{C}(\mathbf{X}, \mathbf{Y})$. For clarity, $\otimes$ denotes the

Kronecker product, vec a column stacking operator, and $\mathbf{C}(\mathbf{X}, \mathbf{Y})$ a pairwise distance matrix of Wasserstein distances between nodes feature pairs in $\mathcal{G}_1, \mathcal{G}_2$. The gradient that arises from Equation 3.7 can be expressed with the following partial derivative with respect to π ,

$$G = (1 - \alpha)\mathbf{C}(\mathbf{X}, \mathbf{Y}) + 2\alpha L(\mathbf{A}, \mathbf{B}) \otimes \pi$$

where $L(\mathbf{A}, \mathbf{B})$ is a 4-dimensional tensor that measures the similarity between all pairwise distances within each graph,

$$L(\mathbf{A}, \mathbf{B})_{i,j,k,l} = \|\mathbf{A}_{ik} - \mathbf{B}_{jl}\|$$

From the gradient in 3.7, the Quadratic Program 3.6 can be solved iteratively using conditional gradient descent. The full algorithm can be found in [108].

3.3.3 Analyzing Microenvironments with Graph-OT Distances

In the previous two sections, we outlined a method for describing microenvironments as spatial graphs, and a method for comparing two microenvironments by computing the similarity between their graphs using the Fused Gromov-Wasserstein distance. Given these two methods, we can analyze microenvironments with a variety of strategies.

Clustering: To identify microenvironment clusters, we first compute a pairwise distance matrix between microenvironments using the FGW distance. Namely, given a set of microenvironments $\{\mathcal{G}\}_{i=1}^n$, we construct a pairwise distance matrix D such that

$$D \in \mathbb{R}_{\geq 0}^{n \times n}, \quad D_{ij} = \text{FGW}(\mathcal{G}_i, \mathcal{G}_j) \quad (3.8)$$

From this distance matrix, microenvironment clusters can be found by partitioning D such that the within cluster distances are minimized, while the between cluster distances are maximized. This procedure can be implemented using any standard clustering algorithm, such as Hierarchical clustering, Spectral Clustering, or even K -means and DBSCAN after dimensionality reduction (see next section) [123]. In general, computing the FGW distance over n^2 possible pairs of microenvironments is computationally intensive. To increase the efficiency of this step, we note that since FGW is a semi-distance metric between labeled graphs, only $\frac{n(n-1)}{2}$ comparisons are required due to symmetry and non-negativity properties. Furthermore, we can also compute D in parallel.

Dimensionality Reduction: While clustering algorithms provide rigid partitions of microenvironments into different sets, appropriate visualizations can highlight gradients and mixtures between different groups of microenvironments. Strictly speaking, the distance matrix can be directly visualized with a simple heatmap. However, this visualization is unlikely to be informative, as the user must permute the rows and columns to find clusters. A more effective strategy is to employ dimensionality reduction to embed D into a 2 or 3-dimensional space, allowing users to visualize separations and groups. Dimensionality reduction algorithms such as UMAP [68] or Multi-Dimensional Scaling [70] try to find a set of n points in euclidean space, $X \in \mathbb{R}^{n \times \{2,3\}}$, such that the Euclidean distance between any pair of points is approximately equivalent (e.g. monotonically correlated) to the original distance found in the distance matrix.

$$\|X_i - X_j\|_2 \approx D_{ij} = \text{FGW}(\mathcal{G}_i, \mathcal{G}_j)$$

The matrix X is termed an embedding, and beyond visualization, these embeddings can also be fed into clustering algorithms for further analysis.

Spatial Analysis: After clustering, microenvironments can be mapped back to their original spatial locations to observe different spatial domains or tissue architectures. This allows researchers to observe changes in cellular organization and composition across large tissues, and to highlight the different functionality of different regions.

Median Microenvironment Identification: To analyze the most common or average structure of microenvironments belonging to a particular cluster, we can use the FGW distance to compute a so-called “median” subgraph. Let $\mathcal{S}_c = \{\mathcal{G}\}_{i=1}^m$ be the set of m microenvironments in cluster c . Then, the *median* of $\mathcal{S}_c$ is the subgraph that minimizes the following.

$$\mathcal{G}_i = \arg \min_i \sum_{j=1}^m \text{FGW}(\mathcal{G}_i, \mathcal{G}_j) \quad (3.9)$$

This *median* subgraph allows researchers to visualize “common” or “standard” microenvironments that epitomize certain niches.

Cell Type Enrichment: To analyze microenvironment specific cell type interactions or abundances, we can calculate cell type enrichment scores. Briefly, an enrichment score measures how “over” or “under” represented a certain cell type is in a microenvironment cluster compared to a uniform distribution. Let $\{\mathcal{G}\}_{i=1}^N$ denote a set of microenvironments partitioned into M disjoint clusters $\{\mathcal{S}\}_{j=1}^M$. The number of cells in microenvironment $\mathcal{G}_i$ is denoted by $|\mathcal{G}_i|$, while the number of microenvironments in cluster $\mathcal{S}_j$ is denoted by $|\mathcal{S}_j|$. We will use $\mathbb{I}(\cdot)$ to denote an indicator function, i.e. $\mathbb{I}(\cdot)$ returns 1 if the argument is true, and 0 otherwise.

The “proportion” of cell type k in microenvironment $\mathcal{G}_i$ is given by

$$\text{prop}(k, \mathcal{G}_i) = \frac{1}{|\mathcal{G}_i|} \sum_{c \in \mathcal{G}_i} \mathbb{I}(\text{cell type}(c) = k) \quad (3.10)$$

The “composition” of cell type k in microenvironment cluster $\mathcal{S}_j$ is defined as the

average proportion of k across all the microenvironments in cluster $\mathcal{S}_j$.

$$\text{comp}(k, \mathcal{S}_j) = \frac{1}{|\mathcal{S}_j|} \sum_{\mathcal{G}_i \in \mathcal{S}_j} \text{prop}(k, \mathcal{G}_i) \quad (3.11)$$

The “global composition” of cell type k is defined as the the average proportion of k across all microenvironments.

$$\text{global comp}(k) = \frac{1}{N} \sum_{i=1}^N \text{prop}(k, \mathcal{G}_i) \quad (3.12)$$

Then, the enrichment of cell type k in cluster $\mathcal{S}_j$ is the logfold change between the microenvironment cluster composition and the global composition.

$$\text{enrichment}(k, \mathcal{S}_j) = \log_2 \left(\frac{\text{comp}(k, \mathcal{S}_j)}{\text{global comp}(k)} \right) \quad (3.13)$$

The proportion metric (Equation 3.10) is a straightforward measure of a particular cell type’s abundance in a microenvironment, and the composition metric (Equation 3.11) is a simple generalization to multiple microenvironments assigned to the same cluster. Unfortunately, both metrics are heavily skewed towards populous cell types, making it difficult to comprehend the interactions and behavior of less abundant cell types. Cell enrichment scores counter these deficiencies. An enrichment score of 0 implies that the proportion of a particular cell type within that cluster is exactly equivalent to the proportion obtained if that cell type was evenly distributed across all clusters. In contrast, large positive and negative enrichment scores imply that clusters are significantly concentrated or depleted for certain cell types. Unlike proportion or composition metrics, an enrichment score is unbiased to cell type population.

Cell Interactions: By combining spatial analysis with cell enrichment scores, we can determine *environment specific, multi-cell* interactions. Within a microenvironment cluster, we can compare enrichment scores across cell types to determine

multi-cell interactions. For example, a group of cell types that are simultaneously enriched indicate that group co-localization, and imply recruitment or activation. In contrast, if some cell types are highly enriched while others are highly depleted, this indicates segregation or exclusion. Importantly, these relationships can be inferred jointly across all cell types, and are not restricted to pairwise interactions. Furthermore, these relationships are specific to a *subset* or cluster of microenvironments, and may change from cluster to cluster. Since microenvironment clusters can be mapped back to their original spatial locations, these cell relationships are specific to distinct spatial domains or regions within a larger tissue architecture.

Condition Specific Microenvironments: The distance matrix D can be constructed from any set of n microenvironment subgraphs. Thus, these microenvironments do not need to come from the same tissue graph, and can be sampled from different tissues corresponding to different treatments or conditions. After clustering D , the microenvironment clusters may contain a mixture of microenvironments from tissues spanning different conditions, or else be entirely concentrated from a single condition. To assess the condition specificity of a microenvironment cluster, one can utilize a pairwise χ^2 test. This statistical test compares the actual proportion of microenvironments within a cluster associated with a specific condition to the proportion that would be expected from a random distribution across all conditions, while taking into account the original condition abundances.

3.4 Results

3.4.1 Graph-OT quantifies changes in microenvironment organization and composition

We hypothesize that the function of a microenvironment is a product of its cell type composition and its spatial organization. As such, our strategy of describing microen-

	Spatial LDA	UTAG	Graph-OT
Micro-environment Paradigm	Diffusive signalling. "Bag-of-cells" model. Frequency normalized cell counts within fixed radius r from center cell.	Directed communication. Node information from k -nearest neighbors of center cell.	Proximity based communication. Node information from k -hops on a weighted spatial graph from center cell.
Statistical Model	Latent Dirichlet Allocation. Probabilistic modeling of cell type distributions.	Message Passing / Graph Diffusion. Spatial feature aggregation between nodes.	Optimal Transport (Fused Gromov Wasserstein distance) between graphs.
Information Accepted	Discrete cell types only. Cannot use markers.	Discrete cell types and continuous markers. Neighbor information only (simple graph).	Discrete cell types, continuous markers, continuous or discrete distances between cells (weighted spatial graph).
Micro-environment Characterization	Continuous topic mixtures from LDA model. Must prespecify number of clusters (topics).	Discrete clusters using Leiden Clustering. Does not accept prespecified number of clusters.	Any clustering method, continuous distance between any pair of micro environments. Accepts, but does not require prespecified number of clusters.
Computational Cost	High. Requires retraining if the number of clusters (topics) is changed.	Low. Fit once, recluster at any resolution.	High initial cost (computing distance matrix). Fit once, recluster at any resolution.

Table 3.1: **Comparison of Microenvironment Characterization Algorithms**

vironments as labeled subnetworks of local interactions, and comparing them using graph optimal transport, is carefully designed to capture and quantify both spatial organization and cell type information. In contrast, Spatial LDA’s ”bag-of-cells” model cannot distinguish between different arrangements of cells, and approximates cell communication as a diffusive mechanism that allows all possible pairs of cells within a fixed radius to interact equally. While UTAG incorporates some spatial information through message passing, all UTAG microenvironments share the same S_k star graph structure. This does not imply that either method is incapable of microenvironment analysis; in real tissues, cell type composition and spatial organization are likely to be correlated, so the approximations used by UTAG and Spatial LDA can still be effective. However, we believe that their approximations of spatial organization are not as effective as our explicit treatment of spatial organization using weighted interaction graphs and graph optimal transport in capturing the rich cell-cell interactions that drive microenvironment function.

In this section, we use simulated microenvironments to demonstrate that our methods can characterize microenvironments based on their cell type composition and their spatial organization, even in cases where these two characteristics are *not* correlated. Our simulated microenvironments contain only two cell types: Gray Cells and Green Cells. Microenvironments can have high proportions of Green cells, or low proportions of Green cells. Within these microenvironments, Grey cells and Green cells can either be highly interacting, or exclusionary. We term these two axes of variation as ”Composition” and ”Interaction”, and together they generate four possible states (High/Low Composition combined with High/Low Interaction). We show that our method of describing microenvironments as subgraphs and comparing them with graph optimal transport can detect changes in spatial organization even under fixed cell type compositions, and vice versa.

Specifically, a simulated microenvironment consists of $N \sim \text{Poisson}(\lambda)$ cells, where

$\lambda = 30$. Cell positions are generated by fixing a center cell around the origin, and sampling $N - 1$ locations from a standard multivariate random normal. For each cell, cell types are drawn from $\text{Binomial}(p)$, with $p = \frac{2}{3}$ denoting a high Green composition, and $p = \frac{1}{3}$ a low Green composition. Highly interacting microenvironments are generated by randomly assigning cells to be either Green or Grey while respecting their proportions, while low interaction (exclusionary) environments are generated by selecting the nearest neighbors of the center cell (always Green) to be Green, up to the total number of Green cells.

To quantify if our methods can accurately characterize microenvironments, we generate 100 microenvironments, with the first 50 sampled from one state, and the last 50 from another. We then compute a 100×100 pairwise distance matrix between every pair of microenvironments. If our method is capable of differentiating between the two states, we would expect that distances between microenvironments sampled from different states will be larger than distances between microenvironments sampled from the same state. These properties can be observed in the distance matrix. Methods that can detect changes between states will exhibit a block anti-diagonal matrix structure, with high values concentrated in blocks crossing different microenvironment states. In contrast, methods that are unable to detect changes will exhibit uniform or randomly distributed distance matrices.

In Figure 3.2, we demonstrate that our method can detect changes in both interaction and composition. In Figure 3.2 (a) and (b), we show the case where cell composition changes while spatial organization is fixed, and in 3.2 (c) and (d), spatial organization changes while cell composition is fixed. In both cases, the Graph Optimal Transport distance matrices exhibit the desired block anti-diagonal structure indicative of accurate state change detection. To avoid confounding results, we omit the correlated case where both cell composition and spatial organization change. Beyond simply detecting changes, we confirm that our characterizations are directly

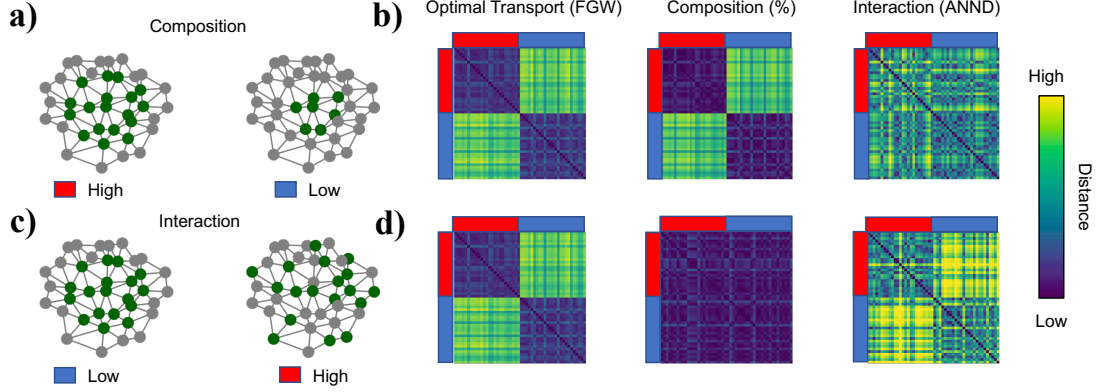


Figure 3.2: **Microenvironment Simulation.** (a) Microenvironment subgraphs from High and Low Green cell compositions with fixed, exclusionary Interactions. (b) Graph Optimal Transport between microenvironments correctly distinguishes between states, as exhibited in the block anti-diagonal distance matrix structure, and confirmed by the Composition distance matrix. The Average Nearest Neighbor Distance (ANND) statistic is not designed to detect changes in composition, and as such displays a randomized structure. (c) Microenvironment subgraphs with High and Low Green-Grey cell Interactions under fixed, High Green cell Composition. (d) Once again, Graph Optimal Transport between microenvironments correctly distinguishes between states as confirmed by the ANND matrix. Because the composition is fixed, the composition distance matrix is uniform.

related to the axes of variation (e.g. composition and interaction) by verifying with statistics specifically designed to detect changes in composition or interactions. In particular, we compare the distances calculated under Graph Optimal Transport with the Euclidean difference between normalized cell frequency vectors, and the Average Nearest Neighbor Distance from Green to Grey Cells. Both statistics corroborate the changes detected under our method, and this experiment demonstrates that our approach is capable of characterizing microenvironments across spatial organization and cell type composition.

3.4.2 Graph-OT Microenvironments Reveal Tissue Architectures in the Healthy Human Lung

We next apply our methods to a publically available dataset of healthy human lung tissues. [91]. 18 highly multiplexed IMC tissues each containing approximately 1000 – 2000 cells were sectioned from three donor patients, and profiled with 28 different markers assigned to 7 different cell types. Importantly, each tissue was annotated by trained pathologists into four distinct anatomical domains: airways, airway walls, alveolar space, and vessels. The human lung is a highly compartmentalized organ, in which different spatial domains perform different functions. Airways carry air in and out of the lungs, and are surrounded by an airway wall, that expands and contracts during inhalation and exhalation. The alveolar space acts as a bridge between airway walls and blood vessels and is responsible for gas exchange, shuttling oxygen from the airway walls to blood vessels, while removing carbon dioxide [118]. In this experiment, we evaluate if our methods can identify distinct microenvironments that map to these spatial domains, and compare our predictions to pathologist annotations.

We construct our microenvironments using the workflow outlined in Figure 3.1. For each tissue, we construct a tissue wide interaction network, and extract a k -hop microenvironment subgraph centered around each cell. We then compare these microenvironments using Graph Optimal Transport, and construct a pairwise distance matrix (see Equations 3.6, 3.8). After experimentation, we chose a value of $k = 3$ for computational efficiency, striking a balance between microenvironment size (roughly 40 cells) and manageable computation time for the distance matrix (10 – 15 minutes). After computing the distance matrix, we cluster our microenvironments and apply spatial domain mapping, and compare with the pathologist’s annotations. In principle, any clustering algorithm can be used, but in order to facilitate comparison with the UTAG algorithm, we employ the Leiden clustering method. This standardization allows us to attribute differences in performance between UTAG and our approach to

the quality of the microenvironment representations, rather than the quality of the clustering algorithm.

In Figure 3.3 (a), we observe that the spatial domains generated by graph optimal transport are similar to pathologist annotations, and in some cases may be more granular. For example, in the tissue shown in the top left corner, our method identifies airway microenvironments that follow the folds of epithelial cells, while the pathologist annotation delineates an idealized straight boundary. In order to further analyze the function of these microenvironments, we compute cell enrichment scores (Equation 3.13) for each microenvironment across all $n = 18$ tissues, as shown in Figure 3.3 (c). We observe that Vessel and Airway Walls have highly enriched smooth muscle populations, while the Alveolar Space and Airway are respectively depleted. This observation is consistent with the physiological requirements of Airway Walls and Blood Vessels, as they must expand and contract during inhalation and exhalation. We also observe that the Alveolar Space includes a balanced mixture of endothelial and epithelial cells. These cell types are responsible for diffusion, secretion, and macro-molecule transportation, and are necessary for gas exchange [118, 91]. Together, these results provide compelling evidence that the microenvironments identified through our method align accurately with the spatial domains identified by pathologist annotations and recover relevant physiological functions. These outcomes demonstrate the utility of our methods in mapping tissue architectures and discerning microenvironment function.

We compare our method’s performance with UTAG and Spatial LDA. As mentioned previously, to compare against UTAG we standardize the clustering step to use the Leiden algorithm. Recall that UTAG constructs a k Nearest Neighbor Graph to pass into Leiden clustering by computing pairwise Euclidean distances between spatially aggregated features. By contrast, in our method, we directly employ the Graph Optimal Transport distance matrix. To select k and the resolution parameter,

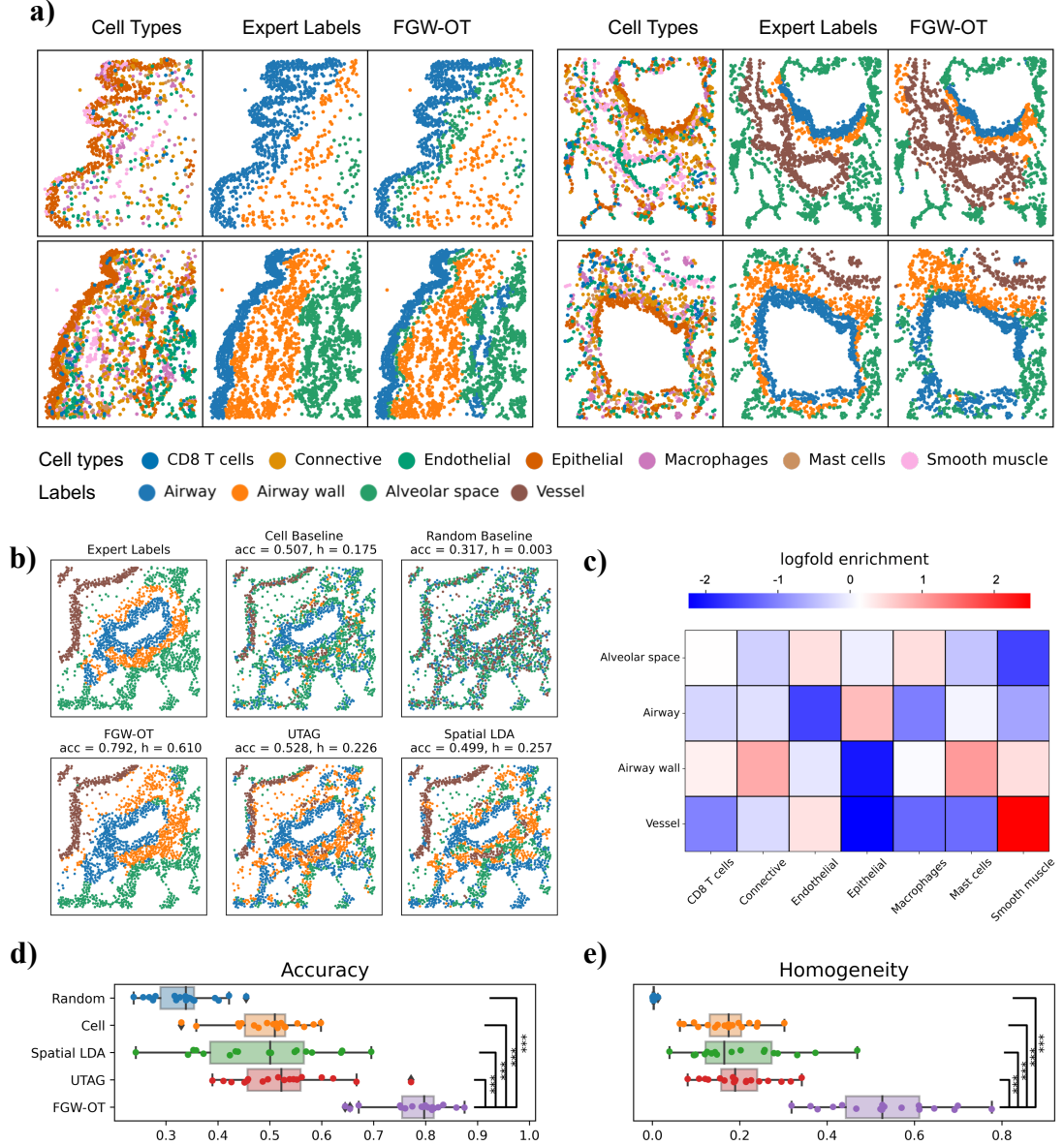


Figure 3.3: Mapping Human Lung Tissue Architecture from multiplex IMC data. (a) comparison between pathologist annotations and spatially mapped, graph-OT microenvironments. (b) Comparing spatial domain predictions from graph-OT and state-of-the-art methods for a single tissue. Cell type enrichment of graph-OT microenvironments over $n = 18$ tissues. Benchmark comparisons over $n = 18$ tissues evaluated by (d) accuracy and (e) homogeneity scores. Significance bars indicate $p < 0.001$ as computed by a Bonferroni corrected two-sided Mann Whitney U Test.

we use the values provided in the original UTAG publication. Spatial LDA does not require a clustering step, and instead defines microenvironments as a mixture of topics. Therefore, we fix the number of LDA topics to be equal to the number of domains in the tissue. While our method and UTAG can be applied directly to the marker information, Spatial LDA requires cell type assignment. To prevent a biased comparison, we elect to use only cell type information for all three methods. For UTAG and our method, this means that node features are one-hot vectors that encode cell types. UTAG and Spatial LDA also have parameters controlling the size of the microenvironments profiled. Standardizing microenvironment size is fairly difficult across all three algorithms, since each uses a different size metric: our method defines microenvironments as three-hop subgraphs in our weighted interaction networks, UTAG uses k -neighbors for its message passing star-graph, and Spatial LDA uses a fixed radius r denoting the maximum diffusion of local cell signalling for a bag-of-cells. We endeavour to achieve a fair comparison by fixing the UTAG k parameter (not to be confused with the clustering k) to be the median size of a three-hop subgraph in our weighted spatial graphs, and the Spatial LDA radius r to be the size for which the distributions of microenvironment sizes (number of cells) are most similar.

To evaluate performance, we compare each method with the pathologist annotations, using Accuracy and Homogeneity [90] scores. The Accuracy score measures the direct correspondence between pathologist and predicted labels, with 0 indicating perfect disagreement and 1 perfect agreement. For Spatial LDA, UTAG, and our method, the predicted labels are generated by constructing a correspondence matrix between clusters/topics and pathologist labels, and assigning each cluster to its most likely label. Formally, given a cluster C and a set of labels L , the predicted label is given by

$$L_{pred} = \arg \max_{L_i} p(L_i|C)$$

The Homogeneity metric measures how *uniquely* a cluster maps to a label, without penalizing the detection of more granular sublabels. As such, it does not require a label mapping step and operates on the original clusters/topics. Formally, the homogeneity score h is given by

$$h = 1 - \frac{H(L|C)}{H(L)}$$

where $H(L|C)$ is the conditional entropy of the labels given the cluster assignment. Intuitively, if multiple clusters correspond to a single label, the homogeneity score is high, since the conditional entropy is low. In contrast, if a single cluster corresponds to multiple labels, the conditional entropy is high and the homogeneity score will be minimized. A homogeneity score of 1 indicates that each cluster maps to exactly 1 label, while a score of 0 means that each cluster maps to every label completely randomly.

As a baseline comparison, we calculate the same metrics using two simple null models: the Random Baseline and the Cell Baseline models:

- The Random Baseline model is a location null model. Every unique cell location is assigned a spatial domain label according to the frequency of that domain in the overall tissue.
- The Cell Baseline model is cell type null model, where each *cell type* is randomly assigned to a spatial domain label according to the label prevalence.

While the Random Baseline can and will map similar cell types to different spatial labels, the Cell Baseline ensures a cohesive labeling within cell types. Together, these baselines evaluate the benefit of characterizing microenvironments as opposed to simply relying on individual cells.

In Figure 3.3 (b), we show the different spatial domains predicted by each method (UTAG, Spatial LDA, and ours) along with the baselines, when they are applied to a

single tissue, and list the accuracy and homogeneity scores. As expected, the random baseline shows no spatial cohesion, since every cell location is randomly assigned to a label. The Cell baseline offers a little more structure, particularly for the Vessel domain. This is likely due to the relative homogeneity of this domain. As shown in Figure 3.3 (c), Vessels are highly enriched for smooth muscle cells, and relatively depleted for most other cell types. This lack of spatial cohesion is reflected in the Spatial LDA and UTAG predictions. While both methods accurately resolve some spatial domains, such as Vessels and, to some degree, the Airway, they fail to accurately model more complex microenvironments, in particular the Alveolar Space. As previously discussed, the Alveolar Space is a highly diverse compartment, with many cell types and a complex structure designed for gas exchange. As such, it is difficult for Spatial LDA and UTAG to model with their simplified approximations of spatial organization. In contrast, our strategy provides a spatially coherent predictions, and most closely matches the Expert labels. These visual observations are confirmed by the homogeneity and accuracy scores.

We compare each of the five methods (UTAG, Spatial LDA, and ours, Random Baseline, Cell Type Baseline) across all $n = 18$ tissues, and compute their accuracy and homogeneity scores (Figure 3.3 (d,e)). Our method significantly outperforms both baseline methods, as well as UTAG and Spatial LDA across both accuracy and homogeneity. We confirm this using a two-sided Mann-Whitney U-Test with a Bonferroni multiple comparison p -value correction. As expected, the Random Baseline achieves an Accuracy roughly equal to the most prevalent spatial domain, which is consistent with the generation of those random labels. Since each “cluster” will map completely randomly to each label, the corresponding homogeneity scores will be 0. This baseline is easily outperformed by all other methods. In contrast, we note that the performances of UTAG and Spatial LDA are similar to the performance of the Cell Type Baseline, and surpassed by our method ($p < 10^{-5}$). This provides compelling

evidence for our previous hypothesis that microenvironment function is derived from cell type composition as well as spatial organization. Since spatial organization and cell type performance are likely to be correlated in highly structured tissues, UTAG and Spatial LDA’s approximations of spatial organization will still provide useful microenvironment characterizations. However, our explicit treatment of spatial organization using weighted interaction graphs and graph optimal transport, allows us to derive a more complete characterization of microenvironment structure and function.

3.4.3 Graph-OT microenvironments are generalizable, and predictive of patient outcome in HNSCC

In this section, we demonstrate that our methods can uncover microenvironments that are highly relevant to disease and patient prognosis on a dataset of CODEX profiled Head Neck Squamous Cell Carcinoma (HNSCC) Tissues [30]. This dataset contains 19 CODEX images of HNSCC tumor tissues containing 1000 – 2000 cells, each from a different patient. Each tissue was profiled with 39 markers assigned to 16 different cell types, and sampled from tumors removed during surgical resection. After a median period of five-years, each patient’s status was updated into one of several categories: ‘NED’ = ‘no evidence of disease’, ‘DOD’ = ‘died of disease’, ‘DOC’ = ‘died of complications’, and ‘AWD’ = ‘alive with disease’. For simplicity, we condensed these categories into two labels: “NED”, indicating complete remission, and “non-NED”, indicating recurrence or death. Of the $n = 19$ patients, 9 had NED outcomes and 10 non-NED. We randomly stratified the dataset into a training and a testing group, $n_{train} = 9, n_{test} = 10$ with 5 non-NED patients in each group. In this experiment, we used the training group to identify microenvironments relevant to disease outcome. We then evaluated whether these microenvironments were predictive of patient outcome (NED vs non-NED) on the testing set.

We construct our microenvironments in largely the same fashion as for the lung

dataset, following the workflow outlined in Figure 3.1, and define a microenvironment to be a $k = 3$ hop subgraph on the tissue interaction network. However, instead of constructing a distance matrix per tissue, we construct a *patient to patient* distance matrix that compares all microenvironments across and within patients. As before, to facilitate comparison with UTAG, we cluster this distance matrix with the Leiden method using the resolution and k parameters outlined in the original UTAG publication. The resulting clusters contain microenvironments originating from different patients. Some clusters are heavily enriched with microenvironments from non-NED patients or vice versa, while others contain a much more balanced mixture. We use pairwise χ^2 tests to determine the statistical significance of this enrichment with a Bonferroni multiple comparison p -value correction. This statistical test compares the proportion of microenvironments within the cluster originating from non-NED patients, to the proportion expected from a random distribution while taking into account the overall abundance of non-NED microenvironments, and vice versa. We view these significantly enriched microenvironments as “biomarkers” for their respective conditions.

Once a biomarker is obtained, it can be easily applied to new patients by observing how similar their microenvironments are to the biomarker. We can quantify this using a similarity score $s \in [0, 1]$, where 1 indicates a perfectly similar microenvironment, and 0 a perfectly dissimilar microenvironment. Given a biomarker $\mathcal{G}^*$ from the training set and a microenvironment $\mathcal{G}_i$ from a new patient, its similarity score s is given by

$$s = \frac{1}{\text{FGW}(\mathcal{G}^*, \mathcal{G}_i) + 1} \quad (3.14)$$

Intuitively, the similarity between biomarkers is the inverse of their Graph Optimal

Transport distance.

$$\lim_{\text{FGW}(\mathcal{G}^*, \mathcal{G}_i) \rightarrow 0} s = \frac{1}{0 + 1} = 1$$

$$\lim_{\text{FGW}(\mathcal{G}^*, \mathcal{G}_i) \rightarrow +\infty} s = \frac{1}{\infty + 1} = 0$$

By applying Equation 3.14 to a set of microenvironments sampled across a tissue, we can analyze the *spatial prevalence* of a given biomarker across any patient’s tissue architecture.

In Figure 3.4 (a) we show a sample CODEX image from an unseen patient in the testing set. In this image, we observe a tumor front located in the top left region of the image, surrounded by an immune-stroma microenvironment in the center and bottom right. We extract the most heavily enriched NED and non-NED clusters from the training set as determined by the Bonferroni corrected pairwise χ^2 tests, and compute the median microenvironment of each cluster using Equation 3.9 to produce the networks depicted in Figure 3.4 (b). We observe that the non-NED biomarker is heavily enriched with tumor cell types, while the NED biomarker is more diverse, being composed of a mix of immune and stromal cells. We also note that the structure of the non-NED biomarker is very densely packed, containing a total of 39 tumor cells, with each tumor cell is strongly interacting with multiple other tumor cells to form a tightly connected network. In contrast, the NED biomarker is far less dense, containing only 20 cells, and shows mixed and sparse interactions. Using the similarity score outlined in Equation 3.14, we “convolve” these biomarkers across all microenvironments from the unseen patient, and compute spatial similarity maps. These similarity maps correspond to the tumor front and immune-stromal microenvironments observed in the original image, and reflect similarities in both cell composition and structure. The compact, tumor rich non-NED biomarker is highly prevalent in the dense tumor front, while the sparse, immune-stromal NED biomarker

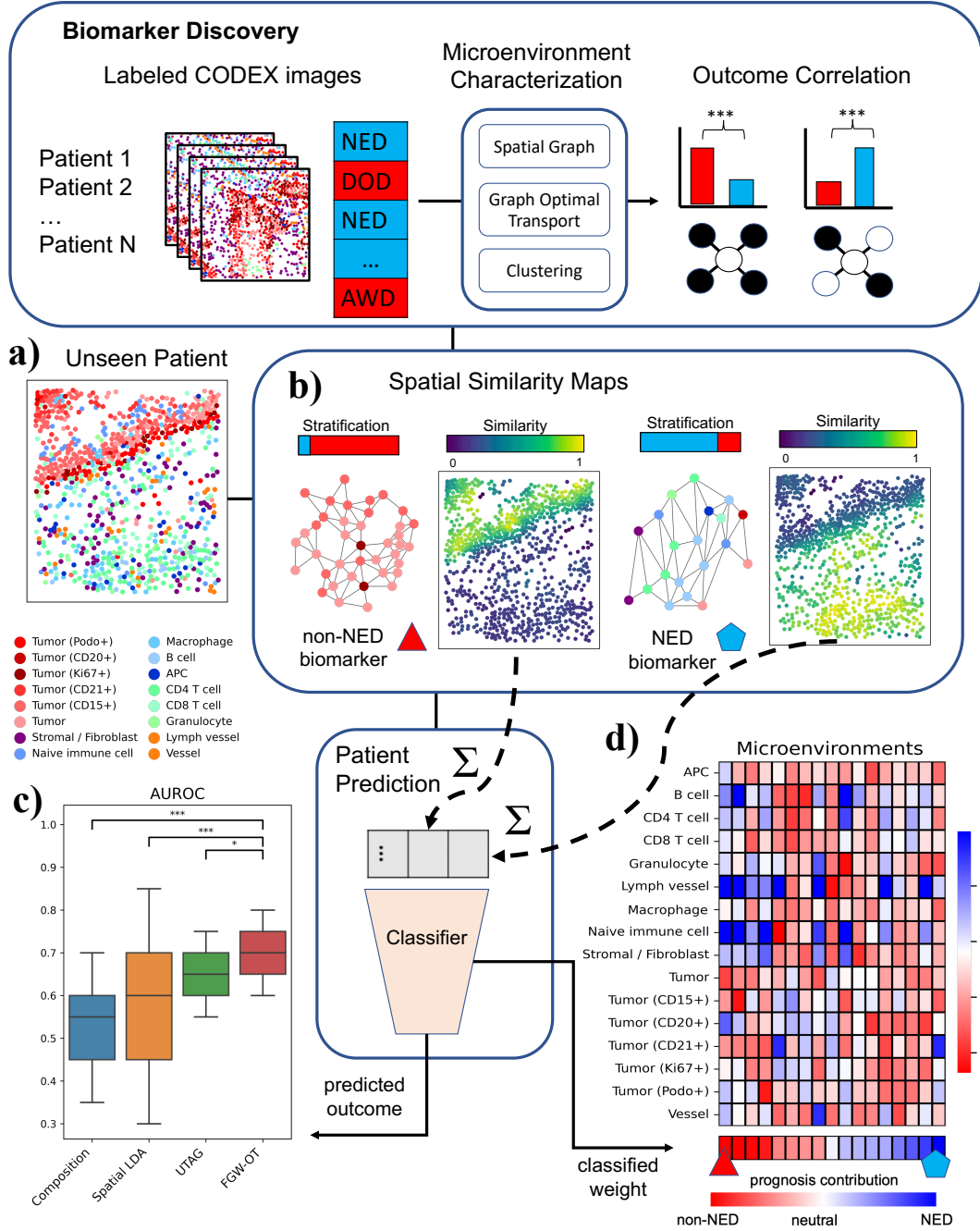


Figure 3.4: Identifying clinically relevant microenvironments in HNSCC tumors. In biomarker discovery, microenvironments are characterized and associated with NED or non-NED outcomes on labeled patients. Given tissue from an unseen patient (a), similarity scores are computed between biomarkers and microenvironments (b). These scores are aggregated and used as features for a logistic regression classifier, with performance evaluated in (c). In (d), classifier weights are used to isolate the contribution of each microenvironment cluster to the overall patient prognosis. Cell enrichment profiles are provided to summarize microenvironment cluster compositions. The weights associated with the red triangle and blue pentagon biomarkers indicate that their presence strongly contributes to non-NED and NED prognosis respectively.

is highly prevalent in the less dense surrounding stromal space.

In order to evaluate the clinical relevance of our biomarkers, we can use them to predict patient outcomes on the test set. Given a CODEX image from a test patient and a training set biomarker, we compute the similarity score between the biomarker and microenvironment across all possible microenvironments extracted from the test tissue as shown in Figure 3.4 (b). From this distribution of similarity scores, we compute the average and use it as a single feature. By repeating this process across all the biomarkers obtained from the training set, we construct a feature vector for the test patient. Each entry in this vector represents the average similarity, or prevalence, of a particular biomarker in the patient. The vector can then be passed into any classification algorithm to produce a predicted risk score, which can then be compared with the patient’s true clinical outcome.

We compare our method’s performance against UTAG and Spatial LDA, and benchmark the clinical relevance of our respective microenvironments by comparing patient prediction performance as measured by the Area Under the Receiver Operating Characteristic (AUROC) score. An AUROC score of 1 indicates perfect classification ability, 0.5 is that of a random classifier, and 0 is of a perfectly incorrect classifier. While any classification algorithm can be used, we elect to use a simple Logistic Regression model for its interpretability. To quantify the added information gained by microenvironment characterization, we use tissue composition as a baseline comparison. In this baseline, the feature vector describing a patient is a vector containing the (normalized) abundance of every cell type in the tissue.

As before, in order to facilitate comparison, we select the radius and number of neighbors parameters in Spatial LDA and UTAG to achieve the most similar distribution of microenvironment sizes, and elect to use only cell type information. For UTAG and our method, we use Leiden clustering with the default parameters from the UTAG publication. For Spatial LDA, we set the number of topics to be the

median number of clusters generated by our method. Since Spatial LDA does not produce discrete topics, and neither Spatial LDA nor UTAG can generate median microenvironments like our method, we omit the selection of biomarkers based on $\mathcal{X}^2$ enrichment tests, and the computation of similarity scores using the median microenvironment. Instead, we use all produced clusters and compute the average similarity score against all microenvironments within a cluster. For Spatial LDA, we fit a model on the training tissues and produce topic mixtures. We infer testing tissue topic mixtures with the same model without retraining. For each patient, the average topic mixture is computed across all of its microenvironments to form a single feature vector. For UTAG, we generate a spatially aggregated feature matrix on the training tissues and cluster. We then compute the average cosine similarity between spatially aggregated features within a training cluster to spatially aggregated features on the testing tissues. The average cosine similarity between all the microenvironments in a testing patient and all the microenvironments within a training cluster is used as a single entry, and the average similarities across all clusters forms a feature vector. For our method, we apply the same procedure as UTAG, but replace cosine similarity with the similarity scores as defined previously in Equation 3.14.

We train a Logistic Regression Classifier on the training features, and evaluate on the testing features. To estimate the variance of each method’s performance, we construct leave-one-out training and testing sets. More precisely, the first training is constructed using microenvironments extracted from patients $\{1, 2, \dots, 8\}$, the second from patients $\{1, 2, \dots, 7, 9\}$, and so on to generate 9 unique training sets. After fitting a classifier to a given training set, we produce $n = 10$ prediction scores. We first calculate the AUROC using true and predicted labels $\{1, 2, \dots, 9\}$, then with $\{1, 2, \dots, 8, 10\}$ and so on to produce 10 unique testing scores. Combined, this allows us to estimate the average and variance of the AUROC over $n = 90$ scores. The leave-one-out procedure on the training set allows us to learn biomarkers without the

influence of the removed patient, while the leave-one-out procedure on the testing set allows us to evaluate test performance without the influence of the removed patient. Together, these procedures limit the influence of outliers, and allow us to construct a more accurate evaluation of performance while estimating its variability.

Using this procedure, we computed the AUROC scores using each method and present the results in Figure 3.4 (c). We observe that our method significantly outperforms the tissue composition baseline and Spatial LDA. While our performance does not surpass UTAG to the same extent, we confirm its advantage using a two-sided Mann-Whitney U-Test with a Bonferroni multiple comparison p -value correction. As expected, the median accuracy of the Composition baseline is better than that of a random classifier (AUROC = 0.5), since tumor cell prevalence and immune density are likely correlated with patient outcome. However, we can clearly observe that microenvironment characterization provides additional insight beyond cell count summaries. For our model, combining cell type and spatial organization to characterize disease-specific microenvironments leads to almost 20% AUROC improvement in prediction performance.

Beyond prediction accuracy, we can also examine the microenvironment biomarkers to determine what characteristics are indicative of patient prognosis. While this is possible using the pairwise χ^2 tests introduced earlier, we can use the weights in our Logistic Regression model to confirm our insights. A Logistic Regression Classifier produces predictions with the following model,

$$p(y = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p))}$$

Here $\beta_1, \cdots \beta_p$ are the weights that correspond to features $x_1, \cdots x_p$, while β_0 is the intercept. Using the law of total probability and exponentiation properties, we can

rewrite this as an *odds ratio*,

$$\frac{p(y = 1)}{1 - p(y = 1)} = \frac{p(y = 1)}{p(y = 0)} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p)$$

In this formulation, we can see that increasing x_i increases the relative likelihood of predicting $y = 1$ by a factor of $\exp(\beta_i)$. In our model, $y = 1$ indicates a non-NED or negative patient outcome, and x_i represents the average similarity of a tissue to microenvironment biomarker $\mathcal{G}_i$. Thus, β_i indicates the relative impact of biomarker $\mathcal{G}_i$ towards a negative patient outcome.

We perform cell enrichment on every microenvironment cluster using the method outlined in Equation 3.13, and sort the microenvironments in increasingly positive patient prediction as determined by the Logistic Regression Classifier weights. In Figure 3.4 (d), we observe a range of microenvironment enrichment profiles and a range of weights. Most weights are near 0, and indicate that these microenvironments are not indicative of NED or non-NED outcomes; they can be viewed as neutral. However, some microenvironments are associated with positive and negative weights of large magnitudes, indicating that they are extremely potent indicators of non-NED or NED patient outcomes respectively. We observe that the weights for negative patient outcomes are associated with microenvironments that have severe immune depletion and tumor enrichment. For example, the microenvironment cluster most strongly associated with a negative outcome (indicated by the red triangle marker in Figure 3.4 (d)) has an approximately three-fold log enrichment for Tumor cells, and five-fold log depletion for Lymph Vessels, Naive Immune cells, and B-cells. In contrast, the microenvironment cluster most strongly associated with a positive outcome (indicated by the blue pentagon marker in Figure 3.4 (d)) has five-fold log depletion of Tumor Cells, and three-fold enrichment of Macrophages, Granulocytes, CD4 T cells, and Antigen Presenting Cells (APCs). These findings are consistent with

clinical knowledge of HNSCC and immune response to tumor progression [2, 10]. In addition, these two clusters in particular are exactly the microenvironments shown in Figure 3.4 (b) previously, that were found using pairwise χ^2 tests. Together, these results offer compelling evidence that the biomarkers identified through our method are useful in discerning biologically meaningful functions with clinical relevance.

While our results are highly promising, we caution that their significance is constrained by the limited HNSCC patient cohort size ($n = 19$). In addition, CODEX data was only collected from tumor microarray cores sourced from the larger tumor tissues removed during patient surgery [30]. If possible, our methods should be further validated on a larger HNSCC patient cohort, and using multiplex data measured from larger portion of the tumor tissue to better evaluate the prognostic signal generated by cell microenvironment analysis. Furthermore, careful attention should be dedicated to addressing potentially confounding factors, such as Human Papillomavirus (HPV) infection [1], and tobacco or alcohol use [43]. These factors should be used to stratify training and testing cohorts, to quantify the information gained by including cell microenvironment analysis in addition to broader risk factors. An interesting complementary experiment would be to repeat the preceding analysis and predict risk factor status (HPV/alcohol/tobacco) instead of outcome, in order to determine if certain risk factors induce changes in or are associated with certain cell microenvironments.

3.5 Conclusion and Discussion

In this chapter, we introduced a computational framework for characterizing cell microenvironments from multiplex images. Our contributions are threefold. First, we have introduced a representation of microenvironments that captures cell-cell interactions using weighted spatial graphs. Secondly, we described a statistical metric for

comparing microenvironments that accounts for cell type information as well as spatial organization with Graph Optimal Transport. Thirdly, we demonstrated how these comparisons can be used to characterize microenvironments through clustering, dimensionality reduction, cell enrichment, spatial domain mapping, median microenvironment identification, and condition specificity. Together, these contributions define a generalized, flexible framework for identifying and describing microenvironments.

We evaluated our method on real and synthetic datasets, and benchmarked its performance against two state-of-the-art algorithms, UTAG and Spatial LDA, as well as informative baselines. Using synthetic data, we demonstrated that our method could characterize microenvironments across both spatial organization and cell type composition. On a dataset of annotated lung tissues, we identified unique microenvironments that mapped to distinct spatial domains in the lung tissue architecture. On HNSCC tissues, we identified disease relevant microenvironments that were highly predictive of patient outcome. On each dataset, we outperformed both the relevant baselines as well as UTAG and Spatial LDA. In the next chapter, we apply microenvironment characterization in conjunction with the spatial integration methods introduced earlier on a novel dataset of mouse tissues.

Chapter 4

Case Study: Microenvironment Remodeling of Colorectal Cancer through TGF- β Inhibition

4.1 Introduction

In this chapter, we present a case study that showcases the practical application of our spatial integration and microenvironment characterization methods on a novel dataset of mouse colorectal cancer tissues profiled with multiplex IHC. Colorectal cancer (CRC) is a leading cause of cancer-related deaths; in the UK, it is the fourth highest in incidence, but the second in mortality [98]. Despite its prevalence, around 75 - 95% of CRC cases occur in people with little to no genetic risk [25]. Instead, studies have shown that microenvironment remodeling is linked to CRC progression and prognosis [48], and in recent years, advancements in multiplex imaging have allowed researchers to profile and identify cell neighborhoods and spatial domains closely associated with distinct patient outcomes [96]. Thus, understanding the interplay between cell microenvironments, spatial organization, and tissue function is

an increasingly important and clinically relevant task for improving CRC treatment.

In this case study, we focus on the role of Transforming Growth Factor β (TGF- β). In healthy cells, the TGF- β signalling pathway plays a critical role in regulating tissue development, proliferation, differentiation, and apoptosis, and is a known tumor suppressor [31]. However, in CRC, the role of TGF- β is more complex and context dependent. Early CRC progression is marked by dysregulated TGF- β signalling, which causes the proliferation rate of epithelial cells to increase. In late CRC stages, TGF- β acts as a tumor promoter and induces metastasis, which is known as the TGF- β paradox [106]. In this stage, TGF- β is known to promote the activation of cancer associated fibroblasts (CAFs) and suppress immune function [47, 92]. The relationship between TGF- β and CAFs is complex and bidirectional; CAFs are both activated by and secrete TGF- β [104]. Thus, TGF- β plays an important yet complex role in shaping CRC tumor microenvironments.

In this case study, we investigate the role of TGF- β in late stage CRC by analyzing a novel dataset of $n = 10$ mouse CRC tissues profiled with multiplex IHC. Half of the samples were treated with a TGF- β inhibitor, and half left untreated. Using our quantized-POT spatial integration algorithm, we aligned adjacent WSI tissue slices stained with optimized marker panels for immune and CAF populations, allowing for a joint analysis of 16 different cell types across massive tissues each comprising of hundreds of thousands of cells. Finally, using our graph-OT microenvironment characterization methods, we identified distinct cell microenvironments with unique cell enrichment compositions and spatial domains that were correlated to TGF- β inhibition.

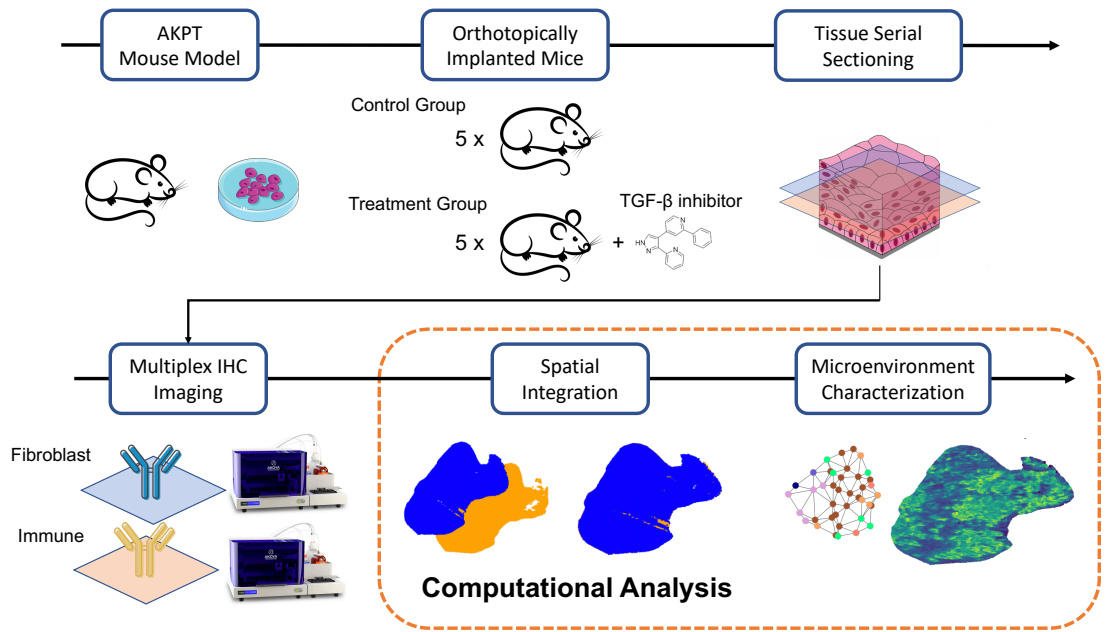


Figure 4.1: **Experimental Outline.** Tumor organoids from an AKPT mouse model of CRC were orthotopically implanted into $n = 10$ mice, half of which were treated with a TGF- β inhibitor. Adjacent tissue slices were resected and stained with IHC markers for fibroblast and immune cell populations, and subsequently imaged and preprocessed (cell segmentation and type assignment). Following data acquisition, we applied our algorithms for spatial integration and microenvironment characterization.

4.2 Methods

Tumor organoids sourced from a mouse with *Apc*, *Kras*, *Trp53*, and *TGF- β* (AKPT) alterations were used as a model for CRC. These alterations model the classic route to late stage CRC, which is marked by the consecutive dysregulation of tumor suppressor genes (*Apc*, *Trp53*, *TGF- β*) and activation of oncogenes (*Kras*) [115, 48]. The organoids were orthotopically implanted into $n = 10$ mice. $n = 5$ mice were orally dosed with *ALK5* inhibitors, and the others were left untreated. *ALK5* (Activin Receptor-Like Kinase 5) is a receptor responsible for transmitting TGF- β signals into the cell, and by blocking *ALK5*, TGF- β signalling is inhibited [109]. The mice were then sacrificed, and two adjacent tissue slices were resected from each mouse. A total of 12 markers were applied across two immunohistochemistry panels. One tissue slice was stained with an immunohistochemistry panel for profiling CAF populations, including markers such as PDGFRA, CD74, pSMAD3, SMA, and pSTAT3. The second slice was stained with an immunohistochemistry panel for profiling immune and tumor populations, and included markers such as CD4, CD8, FoxP3, and E-cadherin. Both tissues were then treated with DAPI (nuclear marker), and imaged using the Vectra Polaris system (Akoya Biosciences). Cell locations and types were extracted from each multichannel WSI by biologists using HALO (Indica Labs) software (Figure 4.1, Appendix A.3).

4.3 Results

4.3.1 Spatial Integration of Immune and Fibroblast Panels

As previously mentioned, TGF- β signalling has complex interactions with tumor cells, CAFs, and immune cells. Thus, CRC microenvironments contain diverse cell populations [96]. While it is possible to profile tissue sections with high-plex technologies

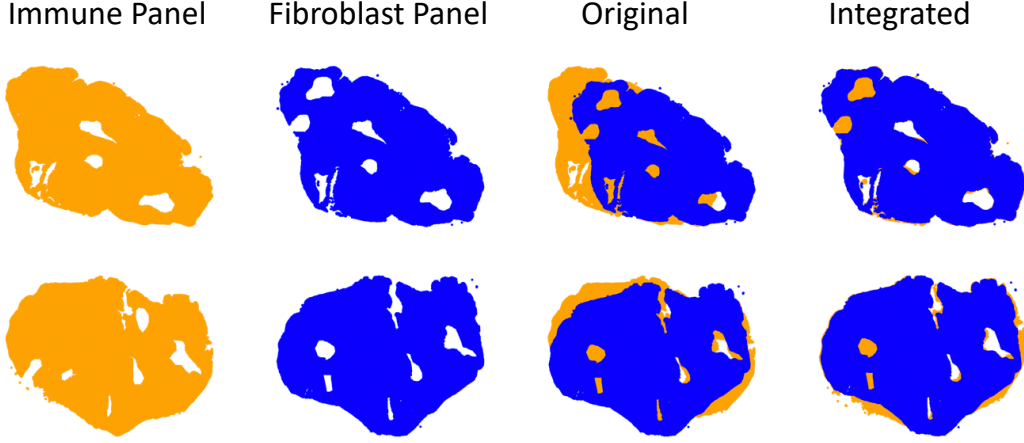


Figure 4.2: **Spatial Integration of Immune and Fibroblast Panels.** Tissues from two mice are shown, with tissues stained with immune markers in blue, and fibroblast markers shown in orange. Their positions and orientations pre and post integration are shown. Note that tears and holes are present that are not shared across slices.

such as CODEX or IMC, these technologies are only able to measure protein expression over very small areas, typically only containing a few thousand cells [102]. In this case study, we are interested in characterizing the CRC microenvironments over diverse cell types, and also over global tissue architectures. As such, multiplex IHC imaging was applied to two adjacent tissue slices, allowing for the identification of immune, tumor, and CAF populations over massive WSIs. In order to perform a joint analysis over these diverse cell types and spatial regions, we turn to spatial integration.

The mean and median tissue sizes were 192,589.25 and 162,306.5 cells respectively, with the largest slice containing 500,345 cells and the smallest 35,815. The size of these tissues renders full POT optimization computationally difficult, and we elected to employ our quantized optimization method. Instead of fixing the number of iterations, we performed quantized optimization until the transformation parameters converged e.g. $\|\Theta^{(k+1)} - \Theta^{(k)}\|_2 \leq 10^{-4}$. On average, this procedure took approxi-

mately 1000 – 3000 iterations, or equivalently, 5 – 10 minutes per alignment.

While the true alignment positions are unknown, based on our past experiments (see Figure 2.2), we anticipate: (i) that the POT loss is highly correlated (Spearman $\rho \geq 0.9$) with the mean squared error from the optimal alignment, and (ii) the quantized gradient, though a stochastic approximation of the true gradient, minimizes this loss in the limit as the number of iterations increases (Figure 2.4). Since our gradient descent optimization algorithm is guaranteed to reduce the POT loss, a convergence in transformation parameters is a necessary (but not sufficient) condition for successful alignment. A notable exception to this case is if the POT loss converges to a local rather than global minimum. To prevent this, we rerun each alignment 5 times with varying initialization parameters, and observe the final results for consistency. For all tissues, our method produces reasonable alignments, even when multiple holes and tears are present (see Figure 4.2).

4.3.2 Graph Optimal Transport Identifies Highly Stable Microenvironment Clusters

After spatial integration, each tissue contains approximately 100,000 cells assigned to 16 unique cell types. To identify and characterize cell microenvironments, we convert each tissue into a spatial interaction network using the procedure outlined in Section 3.3.1. From each tissue, we randomly sample 1000 cell microenvironments defined as three-hop local subgraphs, with a median size of 40 cells per microenvironment. This tissue-stratified sampling process allows us to control for the contribution of each tissue in the microenvironment characterization process. Since the largest tissue (post-integration) contains more than 800,000 cells, and is more than ten times larger than the smallest, which contains fewer than 70,000 cells, this process prevents more populated tissues from dominating our findings. From this collection of 10,000 unique microenvironments, we compute a graph optimal transport distance matrix using the

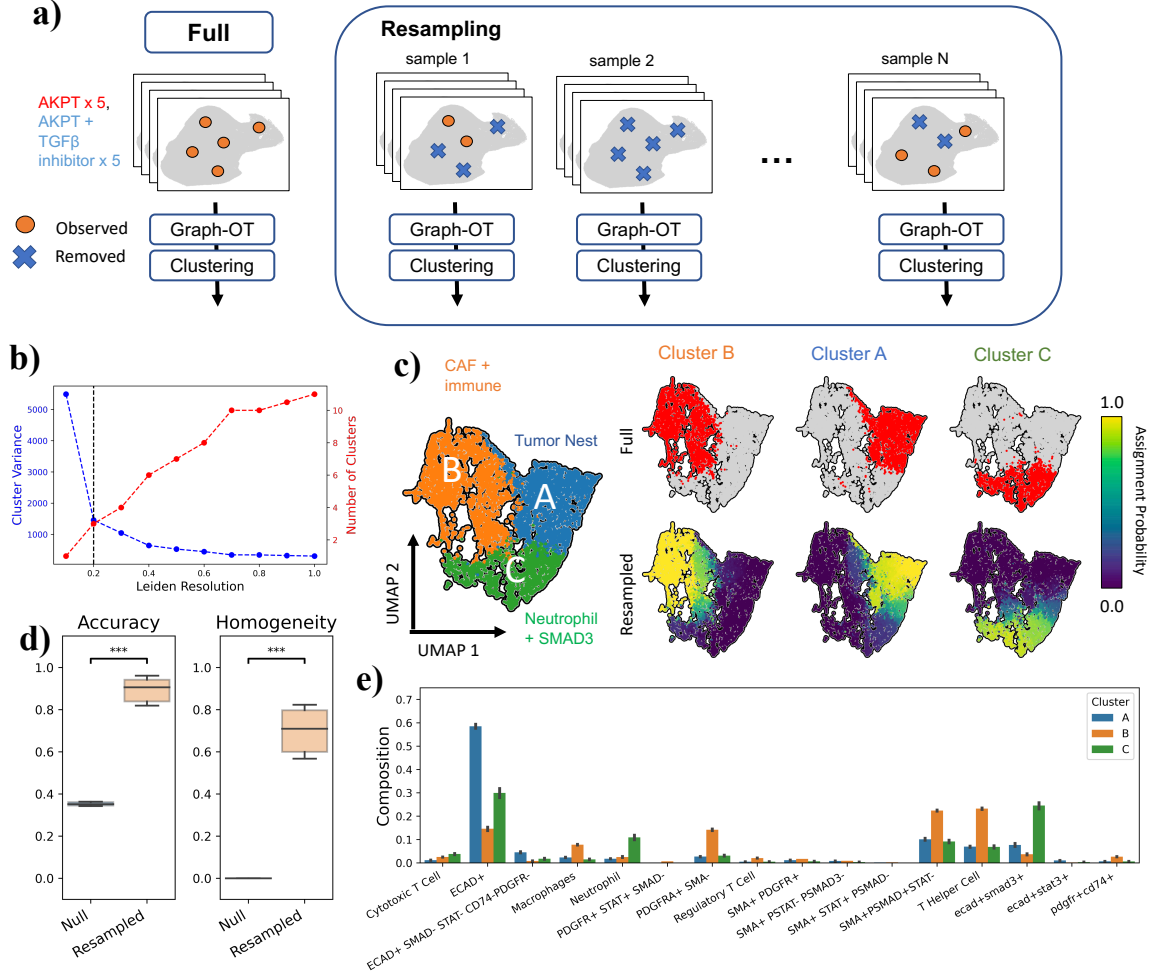


Figure 4.3: Cluster Stability Analysis. (a) By randomly removing microenvironments across and within tissues, we can estimate cluster stability by reclustering on the perturbed datasets. (b) Using an intra-cluster variance curve, we can identify the optimal clustering resolution by selecting the "knee point". The graph shown is for cluster selection on the full dataset. (c) On the left, UMAP visualization of microenvironment clusters. The bottom panel shows the probability of cluster assignment across the resampled datasets, while the top panel shows the cluster assignment on the full dataset. Most microenvironments are almost always uniquely assigned to a single cluster. (d) Cluster agreement between the full and resampled datasets, compared with randomized cluster assignment. Significance bars were calculated using a two-sided Mann-Whitney U Test. (e) Average microenvironment cluster composition across all resampled datasets, with error bars indicating a 95% confidence interval. Microenvironment clusters have small variance and show distinct patterns of cell composition.

procedure defined in Equation 3.8. We then cluster the results using the Leiden Method.

To determine the optimal number of clusters, and to evaluate the robustness of our resulting clusters, we employ the clustering stability strategy introduced by [60], outlined in Figure 4.3 (a). From our original $N = 10^4$ observations, we randomly removed 10%, or 1,000 samples, and recluster on the remaining observations, and repeat this process 100 times. This “resampling” process extends the leave-one-out evaluation method introduced in Section 3.4.3 to estimate prediction variance in HNSCC patient outcomes. Unlike leave-one-out evaluation, which removes a single patient’s microenvironments, randomly removing observations occurs both across and within tissues. Since our tissues are sourced from genetically identical AKPT tumor organoids, inter-sample (i.e. mouse) heterogeneity is less severe. However, since each WSI is an order of magnitude larger than previously analyzed tissue patches, intra-sample, or spatial heterogeneity plays a much greater role. By removing 1000 observations, or the equivalent of a stratified sample from a single mouse, across and within tissues, we can estimate inter as well as intra-sample variance.

We first use the resampled observations to estimate the optimal Leiden resolution parameter for clustering. For each of the 100 resampling iterations, we cluster the corresponding Graph-OT distance matrix using the Leiden Method with a resolution parameter varying from 0.1 to 1.0 with a step size of 0.1. This analyzes microenvironment clusters at varying granularity, from a minimum of 1 cluster to a maximum of 11. At each resolution, we compute the average intra-cluster variance which is defined as the within cluster sum of squares [55], and apply the “knee point method” to select the optimal clustering resolution. The “knee point method” is a popular heuristic that selects the optimal clustering resolution as the “point of diminishing returns”, i.e. the point at which adding a new cluster only marginally reduces the intra-cluster variance [105, 32]. To calculate the “knee-point” in the resolution vs

intra-cluster variance curve, we select the resolution that has the greatest absolute second derivative. In Figure 4.3 (b), the curve and corresponding knee point are shown for the clustering of the original, “full” dataset of $N = 10^4$ observations. The optimal clustering resolution parameter is 0.2 and yields 3 clusters.

To estimate the robustness of this approach, we compare the agreement between the clusters obtained on the “full” dataset and the clusters obtained on each of the 100 resampled datasets. Since each resampled dataset contains 1000 fewer microenvironments than the full datasets, we compare clustering agreement only on microenvironments shared between the full and resampled dataset, and use the homogeneity and accuracy scores introduced in Section 3.4.2. As before, computing an accuracy score requires cluster assignment, as a fixed number of categories or labels must be shared between the two datasets. To compute assigned clusters, we generate a correspondence matrix between the clusters from the full dataset and the clusters from the resampled dataset, and assign each resampled cluster to its most likely cluster in the full dataset. From the assigned clusters, we compute an accuracy score, with 0.0 and 1.0 indicating perfect disagreement and agreement respectively. The homogeneity score does not require cluster assignment, and measures how *uniquely* one set of clusters maps to the other. A homogeneity score of 1 indicates that each “resampled” cluster maps to exactly one “full” cluster or vice versa, while a score of 0 means that clusters map to each other randomly. Our results are summarized in Figure 4.3 (d), and show a high degree of agreement between the resampled and full dataset clusters. We compare this agreement to a null model of randomized clustering, which is generated by randomly assigning microenvironments to clusters based on the prevalence of each cluster. Using a two-sided Mann-Whitney U Test, we show that the microenvironment clusters identified on the resampled and full datasets are not random, and instead quantify distinct groups of microenvironments.

We can also visually observe microenvironment clusters and cluster uncertainty

using dimensionality reduction. As described in section 3.3.3, the Graph-OT distance matrix between microenvironments can be visualized using dimensionality reduction, which attempts to position points in a two or three-dimensional space such that their pairwise Euclidean distances respect the relationships encoded in the original distance matrix. In Figure 4.3 (c), we show the UMAP visualization of the full clustering in the far left panel, with the three clusters labeled as A,B,C, and annotated by their respective microenvironment compositions. Each point in the UMAP visualization represents a microenvironment. To visualize cluster uncertainty, for each microenvironment, we can compute the probability that it is placed in cluster A,B or C over all the resampled datasets in which it was sampled. Since resampled clusters may contain more or fewer than three clusters, these probabilities are calculated post cluster assignment. This process yields the heatmaps shown in Figure 4.3 (c). The top panel illustrates the full cluster positions on the UMAP plot, while the lower panel illustrates the aggregated probabilities over the resampled datasets. While the microenvironments on the "border" between clusters have slightly lower probabilities, we can clearly see that most microenvironments are almost always uniquely assigned to a single cluster.

Lastly, we estimate the variance in microenvironment clusters by analyzing the average cell composition of each microenvironment cluster. For each iteration, we perform cluster assignment to map the resampled clusters into the three full clusters (A,B,C), and compute the average cell composition within the microenvironment using Equation 3.11. Our results are shown in Figure 4.3 (e), where the error bars indicate a 95% confidence interval. From this analysis, we observe that microenvironment variability across resampling is extremely low, and that each cluster exhibits a distinct pattern of cell type composition. Together, this provides compelling evidence that the clusters identified with our method are clearly defined and well-separated, revealing distinct cell microenvironments.

4.3.3 TGF- β Remodels Tumor Microenvironments in AKPT CRC tissues

From the stability analysis, we have confirmed that the clusters identified represent distinct cell microenvironments. In this section, we analyze the cell types and spatial organization of each cluster. From Figure 4.3 (e), we observe that each microenvironment cluster is characterized by a unique cell type composition. While some cell types are very obviously localized in one cluster versus another, discerning trends in cell types with smaller populations can be challenging. The prominence of certain cell types overshadows the behavior of less abundant ones, making it difficult to comprehend their characteristics. For example, it can easily be observed that cancerous epithelial cells (ECAD+) and immune T helper cells comprise 60% and 10% of microenvironments in cluster A, but the contributions of less prominent CAF cell types such as PDGFRA+ STAT+ SMAD- cells are extremely difficult to discern.

To address this challenge, we elected to focus on cell *enrichment* rather than absolute proportions, using the enrichment scores introduced in Section 3.3.3 and described by Equation 3.13. Briefly, an enrichment score measures whether a cell type is “over” or “under” populated in a particular microenvironment cluster relative to uniform distribution over all clusters. An enrichment score of 0 implies that the proportion of a particular cell type within that cluster is exactly equivalent to the proportion obtained if that cell type was evenly distributed across all clusters. In contrast, large positive and negative enrichment scores imply that clusters are significantly concentrated or depleted for certain cell types. This score allows researchers to identify cell-cell interactions by observing which cell types are jointly or oppositely enriched, and is unbiased to the overall population of a particular cell type. In Figure 4.4 (c), we compute the enrichment profiles for each of the three clusters, and observe significantly divergent profiles. Cluster A is heavily enriched with tumor cell types (e.g. ECAD+, ECAD+ STAT3+) and can be best interpreted as a tumor nest. Clus-

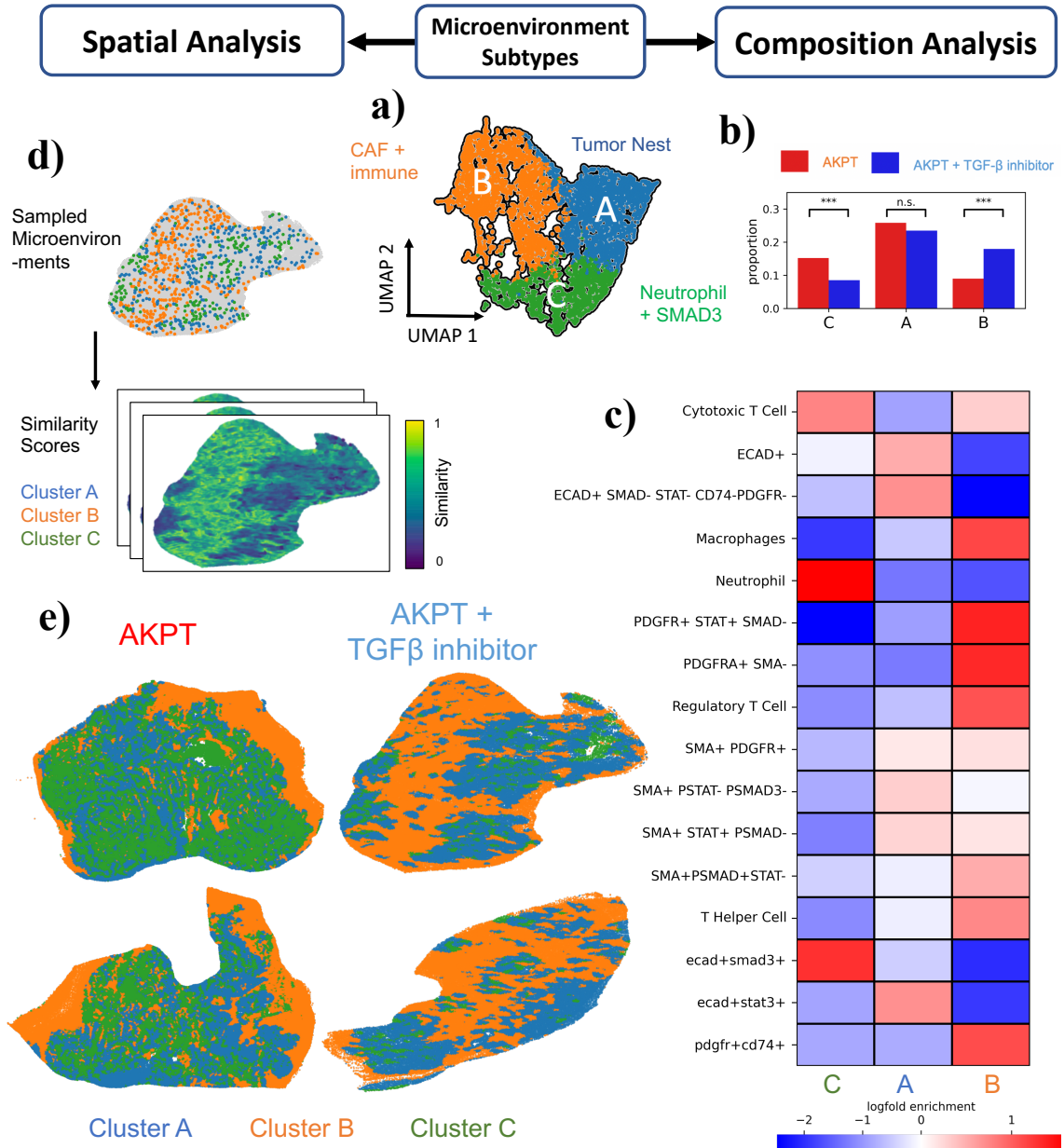


Figure 4.4: TGF- β Remodels the Tumor Microenvironment Landscape. (a) UMAP visualization of microenvironment clustering. (b) Cluster condition specificity, significance bars are calculated using a Bonferroni corrected pairwise χ^2 test. While cluster A is not significantly enriched for either condition, B is significantly enriched in untreated AKPT tissues, and C is significantly enriched in TGF- β inhibited AKPT tissues. (c) Microenvironment cell enrichment profiles. We observe that each cluster has a distinct cell enrichment profile showing unique cell-cell interactions. Using spatial similarity scoring (d), we can construct discrete spatial domains (e) and observe the change in microenvironment locations as a function of TGF- β inhibition.

ter B is heavily enriched with different CAF cell types, as well as immune cells such as Macrophages and Regulatory T-cells. Meanwhile, Cluster C is heavily enriched with Neutrophils, Cytotoxic T-cells, and ECAD+ SMAD3+ cells, while relatively depleted for all other CAFs and immune cells.

Beyond cell type characterization, we can also map microenvironment clusters back to their locations on their original tissue to identify spatial domains. Since the clustered microenvironments represent a stratified sample across each tissue, not every microenvironment in a tissue will be assigned to a cluster. To identify spatial domains across an entire tissue, we turn to the similarity scores introduced in Section 3.4.3 and defined in Equation 3.14. Briefly, a similarity score between two microenvironments is the inverse of their Graph Optimal Transport distance, with 1.0 indicating a distance of 0 or perfect similarity, and 0.0 indicating infinite distance or perfect dissimilarity. By computing the similarity between all possible microenvironments in a tissue and a microenvironment assigned to a cluster, we can determine the spatial prevalence of that cluster across the tissue. Instead of selecting a random microenvironment from each cluster, we select the median or “average” microenvironment as introduced in Section 3.3.3 and defined in Equation 3.9. To obtain a more robust estimate, we isolate a collection of 10 median subgraphs by randomly removing 10% of the microenvironments in each cluster and determining the median on the remaining 90%, and repeating 10 times. This allows us to compute an average spatial score across the 10 median subgraphs per cluster, for all clusters, and the resulting spatial similarity maps are demonstrated in Figure 4.4 (d).

After computing similarity scores, each microenvironment in every tissue can be described by a vector $s \in \mathbb{R}^3$, where each entry describes its similarity to clusters A, B, and C respectively. While it is possible to describe the spatial domains as a “mixture” of clusters, for ease of analysis, we assign each microenvironment to its most similar cluster, i.e. $\arg \max_i s_i$. From this assignment, we can observe the

spatial domains of each respective cluster across AKPT tissues with and without TGF- β inhibition. The results are shown in Figure 4.4 (e), and reveal distinct spatial architectures across conditions. In AKPT tissues, we observe that the CAF-immune microenvironments (cluster C), are pushed towards the tissue edge and away from the tumor nest (cluster A) and SMAD3-Neutrophils (cluster B). In AKPT tissues treated with TGF- β inhibition, we observe that the CAF-immune microenvironments (cluster C) infiltrate the tissue centers. In both treated and untreated tissues, we observe that microenvironments in A and C tend to mix and intermingle while remaining segregated from cluster B.

After observing the cell enrichment profiles and spatial domains of each cluster, we then determine whether the clusters are more highly enriched in treated or untreated tissues. In Figure 4.4 (b), we employ pairwise χ^2 tests as introduced in Section 3.3.3 to identify condition specific microenvironment clusters. To correct for multiple hypothesis testing, we employ a Bonferroni p -value correction. After observing the cell enrichment profiles and spatial domains of each cluster, we then determine whether the clusters are more highly enriched in treated or untreated tissues. We observe that cluster A (tumor nest, Ecad+, Ecad+ Stat3+ enriched) is equally prevalent within control AKPT tissues and those treated with TGF- β inhibition. In contrast, cluster B (CAF-immune enriched) is significantly enriched in TGF- β inhibited AKPT tissues, while cluster C (Neutrophil, Ecad+ Smad3+ enriched) is significantly enriched in untreated AKPT tissues. Since Smad3 is a downstream protein in the TGF- β signalling pathway, the increased prevalence of the Smad3-enriched microenvironment (cluster C) in non-inhibited TGF- β tissues is a biologically plausible and expected result, providing mechanistic validation for our spatial analysis. Together, these results show that TGF- β drives the formation of unique tumor microenvironments with distinct cell enrichment profiles and spatial domains, and that this effect can be countered with TGF- β inhibition. While Smad3 is a protein known to directly interact with

TGF- β , the mechanistic effect of TGF- β signalling on other cell populations is not well understood, and the strong colocalization of Neutrophils with Smad3 active cells in TGF- β active tissues among other associations uncovered through spatial analysis is highly promising and warrant further investigation.

4.4 Discussion and Conclusion

In this chapter, we have provided a case study for the application of our methods for spatial integration and microenvironment characterization. On a novel dataset of murine AKPT CRC tissues, we integrated tissue slices stained with IHC panels optimized for fibroblast and immune populations. Using Quantized POT Alignment, we enabled the analysis of 16 unique cell types across large tissues with hundreds of thousands of cells. From this data, we applied our microenvironment characterization methods to characterize the spatial landscape of AKPT tissues with and without TGF- β inhibition. Using Graph Optimal Transport and cluster stability metrics, we identified three highly stable microenvironment clusters that corresponded to unique cell enrichment profiles and spatial domains. Furthermore, we showed that two of these clusters were condition specific, with one highly enriched in TGF- β inhibited AKPT tissues, and the other in untreated AKPT tissues. Together, this computational analysis provides evidence that TGF- β remodels the tumor microenvironment landscape in late stage CRC. The implications of this case study are highly intriguing and warrant further investigation. A natural extension to our work would be to characterize microenvironments at greater clustering resolutions, to identify more specific microenvironments. For example, while cluster B is currently highly enriched with multiple CAF cell types and immune cells, it would be interesting to see if there are additional CAF sub-clusters with distinct spatial or compositional profiles. This extension could be readily implemented, since our framework allows for re-clustering

at varying resolutions.

In future experiments, a repeated spatial analysis without consideration of Smad3 staining would allow us to better isolate the interactions between other cell states and TGF- β signalling, as well as provide a held-out validation signal. More precisely, it would be interesting to merge the Ecad+ and Ecad+ Smad3+ cell states and repeat our analysis. However, rather than performing χ^2 tests to identify significant changes in microenvironment cluster prevalence between treated (TGF- β inhibited) and control (AKPT) tissues, we would instead “score” microenvironments clusters by the average level of Smad3 activity in their respective microenvironments. These “scores” would allow us to *rank* individual microenvironments by their level of TGF- β activity using Smad3 as a proxy, and provide a more granular analysis rather than relying on tissue-level cohort definitions.

Chapter 5

Conclusion and Outlook

In this thesis, we set out to explore the complex spatial interactions between cells that drive tissue function and disease. To do so, we developed novel computational methods for analyzing highly multiplexed spatial proteomics data. Multiplex imaging technologies are at the forefront of modern spatial biology, and this thesis illustrates the biological insights that can be unlocked by combining multiplex technology with computational tools. In particular, this thesis presented two core contributions: the *enhancement* of multiplexed spatial information using spatial integration, and the *analysis* of this information with microenvironment characterization. While we introduce unique algorithms for these two tasks, both methods rely on and leverage principles from Optimal Transport (OT). OT is a branch of statistics that compares distributions by identifying the cost of the most efficient way of transporting one distribution into another. This approach is highly flexible yet rigorous, and offers a powerful framework for analyzing multiplexed spatial data without the need for supervised machine learning algorithms or custom engineered spatial statistics.

5.1 Review of Contributions

In Chapter 2, we introduced a method for integrating spatial information by framing this task as a point set registration problem and designing an alignment algorithm based on partial optimal transport (POT). Our method allows researchers to profile massive tissues over diverse cell types by performing multiplex imaging on adjacent tissue slices stained for different cell types, and analyzing the integrated spatial profiles. On simulated data, we showed that our method was resilient to the spatial distortions and artefacts induced by sample collection and tissue variation commonly encountered in multiplex datasets, and outperformed Iterative Closest Point (ICP) and Coherent Point Drift (CPD), two state-of-the-art algorithms for point set registration. Furthermore, we introduced a statistical estimator for POT, and showed that it allowed our method to efficiently scale to tissues with hundreds of thousands of cells; achieving similar performance as the original formulation but at a fraction of the cost.

In Chapter 3, we introduced a method for characterizing cell microenvironments by framing them as labeled subgraphs, and comparing them with graph optimal transport (graph-OT). Our methods allow researchers to uncover distinct spatial domains in complex tissue architectures, discern unique patterns of cell enrichment and spatial organization, and identify disease and condition relevant microenvironments. We evaluated our method on synthetic data and two real-world datasets: pathologist annotated lung tissues profiled by multiplex IMC, and CODEX profiled Head Neck Squamous Cell Carcinoma (HNSCC) tumor tissues from patients with distinct survival outcomes. On synthetic data, we showed that our method was capable of capturing both spatial organization and cell type composition. On the lung IMC dataset, graph-OT accurately identified spatial domains confirmed by pathologists, and in some tissues, provided increased granularity and accuracy. On CODEX HNSCC tissues, graph-OT identified microenvironments that are highly predictive of pa-

tient outcome. On both real-world datasets, graph-OT identified microenvironments directly linked to known cell type interactions and functional tissues. We compared our method against two state-of-the-art competitive algorithms, UTAG and Spatial LDA, and exceeded both as well as baseline methods on all performance benchmarks.

Finally, in Chapter 4 we presented a case study of the practical application of both our methods, and investigate the role of TGF- β in late stage CRC by analyzing a novel dataset of $n = 10$ mouse CRC tissues profiled with multiplex IHC, half of which were treated with a TGF- β inhibitor, and half left untreated. Using quantized POT, we aligned adjacent tissue slices stained with optimized marker panels for immune and CAF populations, enabling a joint analysis of 16 different cell types across massive tissues each comprising of hundreds of thousands of cells. From this data, we applied graph-OT and identified three highly stable microenvironment clusters that corresponded to unique cell enrichment profiles and spatial domains. Furthermore, we showed that two of these clusters were condition specific, with one highly enriched in TGF- β inhibited AKPT tissues, and the other in untreated AKPT tissues. Together, our computational analysis provides evidence that TGF- β remodels the tumor microenvironment landscape in late stage CRC.

5.2 Limitations

While we have shown that our methods are highly promising, we also provide a brief summary of their limitations. For spatial integration, a key constraint is our choice of transformation class. In this work, we have presented results restricted to *rigid* transformations. This choice was motivated by our application: the enhancement of multiplex information by enabling the profiling diverse cell types over large tissues. In this task, *adjacent* tissues slices stained with multiplex markers for different cell types are sectioned from the *same* biological sample. While spatial artefacts (rips, holes,

tears etc.) may be induced by the sample collection process, each slice will contain approximately homogeneous anatomical characteristics, preventing the need for more plastic deformations. This anatomical homogeneity will not be preserved for tissues sourced from *different* biological samples, which is required for example in cross-individual spatial analysis. As such, the methodology for rigid spatial integration introduced in this thesis is limited to aligning slices sectioned from the same tissue.

For our microenvironment characterization methods, we caution that our treatment of multiplex information is *statistical* and not *mechanistic*. This choice of paradigm was driven by the limitations of multiplex technology. As the sample collection process is destructive, multiplex images can only be obtained at a single time point, which prevents the observation of continuous processes such as cell motility or molecule transport. While it may be possible to infer cell-cell signalling from protein markers, the diversity of possible protein panels, combined with the limited number of protein markers per image weakens our ability to comprehensively evaluate receptor-ligand interactions. Subsequently, our methods do not consider any biological domain knowledge beyond the cell types and spatial coordinates included in the multiplex images. In particular, our analysis of microenvironment cell-cell interactions is derived from spatial proximity and does not consider mechanistic communication.

5.3 Future Work

There are multiple avenues for extending our methods in new directions, or applying them to new domains. For spatial integration, one of the most promising domains is the comparative analysis of spatial organization across individuals. As mentioned previously, our analysis focused on spatial integration between adjacent slices, and employed rigid transformations to preserve anatomical homogeneity. However, when tissue slices come from different individuals, whether patients, diseases, or even time

points, the constraints imposed by rigid registration are no longer suitable since the organs may differ in size and shape. In this setting, spatial integration is still possible, as different patients will have corresponding spatial regions (e.g. brain regions from one mouse to another), but elastic deformations are needed to generate alignments that enable comparative analysis at matched spatial locations. This task has been encountered at the image level in the registration of MRI and PET images [54], but with the advent of ever higher-resolution multiplex technologies [35, 20], will become increasingly important at the cell scale. A natural extension to our work would be to apply our method for spatial integration with more plastic transformations, such as diffeomorphic functions. This extension does not require any alterations to our optimization procedure; we would still compute a Partial Optimal Transport Loss using a Quantized Estimator, and minimize it using Gradient Descent. Indeed, the only modification would be the transformation function; while a rigid transformation relies on three parameters (angle, vertical and horizontal shifts), an elastic deformation would entail additional parameters. This generalized spatial integration procedure would allow us to perform cross-sample spatial alignment and interrogate spatial variation as a function of disease.

For microenvironment characterization, a further application of our methods beyond multiplex imaging is spatial transcriptomics. In recent years, technologies such as SeqFISH [35] and MERFISH [20] have enabled whole transcriptomic sequencing at single cell spatial resolution. As such, they provide a comprehensive profile of cell state and behavior across tens of thousands of molecules, orders of magnitude larger than what can be measured using multiplex imaging. Our methods can be directly applied to single-cell, spatial transcriptomics data without modifications. In this setting, we would still construct spatial tissue interaction graphs and compare microenvironments using graph optimal transport, and the only modification would be an increase in the dimensionality of node features (from 10^1 proteins to 10^3 genes).

However, a more powerful extension would be to incorporate mechanistic information derived from gene expression information. Computational tools such as CellChat [49] and CellPhoneDB [33] are capable of inferring cell communication from single cell, non-spatial transcriptomic data using receptor-ligand databases. We anticipate that this extension would not require much alterations to our comparisons or analysis between microenvironment; we would still compute Graph Optimal Transport distances, cluster distance matrices to find niches, and map them tissue locations to find spatial domains etc. Instead, this extension would largely be focused on refining our tissue interaction networks. Instead of defining edges between nodes through spatial proximity, we would also need to factor in receptor-ligand cell communication. This extended procedure would allow us to characterize microenvironments using both statistical and mechanistic information, and provide a more complete and powerful analysis.

Appendix A

Additional Methods

A.1 Software Implementations

The Optimal Transport metrics used in this thesis were computed using the "POT: Python Optimal Transport" Python Library [38]. In particular, Partial Optimal Transport was computed using the `ot.partial.partial_wasserstein2` function and Graph Optimal Transport was computed using `ot.gromov.fused_gromov_wasserstein2`. Leiden Clustering was applied using the `leidenalg` python package [110]. To benchmark our methods for microenvironment characterization against other methods, we ran Spatial Latent Dirichlet Allocation (Spatial LDA) [21], and Unsupervised Tissue Architecture Discovery with Graphs (UTAG) [53]. UTAG was applied from <https://github.com/ElementoLab/utag>, while Spatial LDA was computed using the `tl.spatial_lda` function from "Scimap: Single-Cell Image Analysis Package" [78], available at <https://github.com/labsyspharm/scimap>. Accuracy and homogeneity scores were computed using `metrics.accuracy_score` and `metrics.homogeneity_score` from `sklearn` [82]. Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) visualizations were implemented using the `umap-learn` python package [69]

A.2 Data Availability

The published datasets used in this thesis are publicly available at the following repositories from their original publications:

- healthy lung IMC: <https://doi.org/10.5281/zenodo.6376766>
- CODEX HNSCC: <https://doi.org/10.5281/zenodo.7796393>

The mouse CRC dataset is unpublished, and available upon reasonable request.

A.3 Mouse Multiplexing and Image Preprocessing

Akoya Biosciences OPAL Protocol (Marlborough, MA) was employed for multiplex immunofluorescence staining on FFPE tissue sections of 4- μ m thickness. The staining was performed on the Leica BOND RXm auto-stainer (Leica Microsystems, Germany). A total of six consecutive staining cycles were conducted using primary antibody-Opal fluorophore pairs.

Fibroblast Panel: (1) PDGFRA (1:500, ab134123; Abcam)–Opal 480; (2) CD74 (1:3000, ab289885; Abcam)–Opal 570; (3) pSMAD3 (1:100, ab52903; Abcam)–Opal 520; (4) SMA (1:1000, ab5694; Abcam)–Opal 620; (5) pSTAT3 (1:400, 9145S; Cell Signaling)–Opal 690; and (6) E-cadherin (1:500, 3195; Cell Signaling)–Opal 780.

Immune panel: (1) Ly6G (1:300, 551459; BD Pharmingen)–Opal 540; (2) CD4 (1:500, ab183685; Abcam)–Opal 520; (3) CD8 (1:800, 98941; Cell Signaling)–Opal 570; (4) CD68 (1:1200, ab125212; Abcam)–Opal 620; (5) FoxP3 (1:400, 126553; Cell Signaling)–Opal 650; and (6) E-cadherin (1:500, 3195; Cell Signaling)–Opal 690.

The tissue sections were incubated with primary antibody for an hour, and the BOND Polymer Refine Detection System (DS9800, Leica Biosystems, Buffalo Grove, IL) used to detect the antibodies. Epitope Retrieval Solution 1 or 2 was applied to retrieve the antigen for 20 min at 100°C, in accordance with the standard Leica

protocol, thereafter each primary antibody was applied. The tissue sections were subsequently treated with spectral DAPI (FP1490, Akoya Biosciences) for 10 minutes and mounted with VECTASHIELD Vibrance Antifade Mounting Medium (H-1700-10; Vector Laboratories) slides. The Vectra Polaris (Akoya Biosciences) was used to obtain whole-slide scans and multispectral images. Batch analysis of the MSIs from each case was performed using inForm 2.4.8 software, and the resultant batch-analysed MSIs were combined in HALO (Indica Labs) to create a spectrally unmixed reconstructed whole-tissue image. Cell segmentation and phenotypic density analysis was conducted thereafter across the tissue using HALO.

Appendix B

Additional Figures

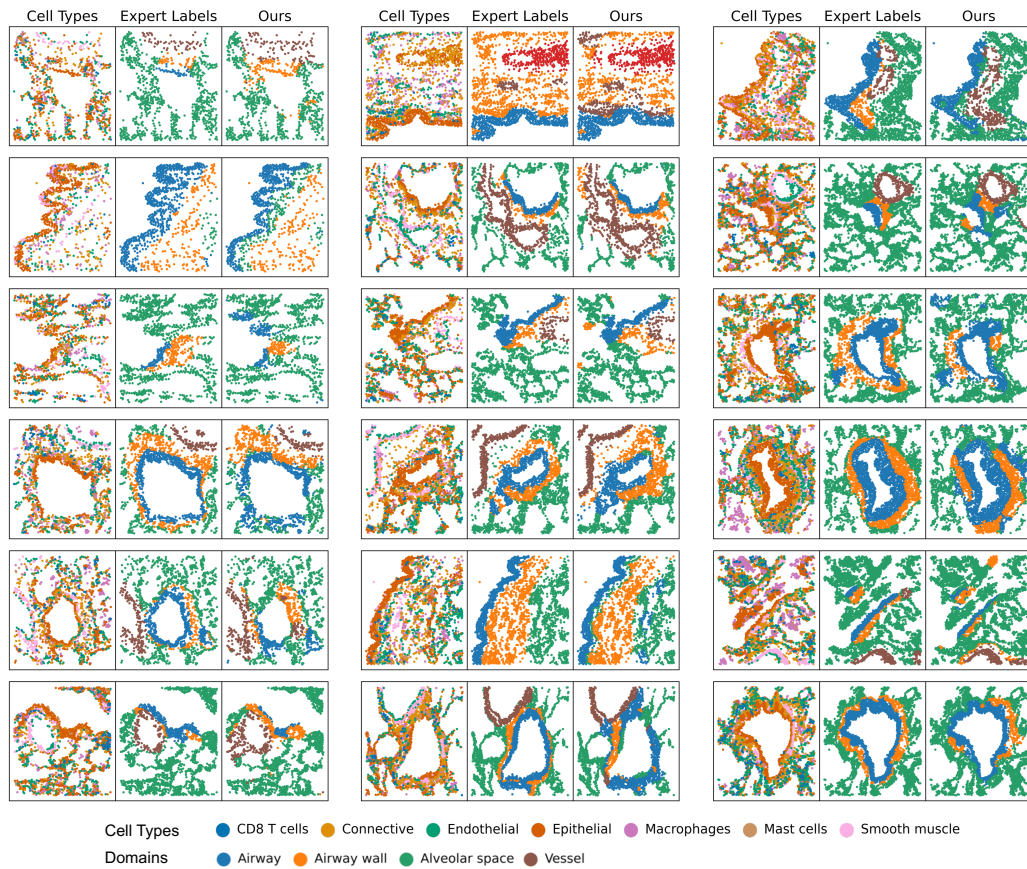


Figure B.1: **Graph-OT Spatial Domain Mapping of Healthy Lung Tissues** For each of the three panels, the first column indicates the cell type, the second pathologist annotations, and the third spatial domains identified using Graph Optimal Transport.

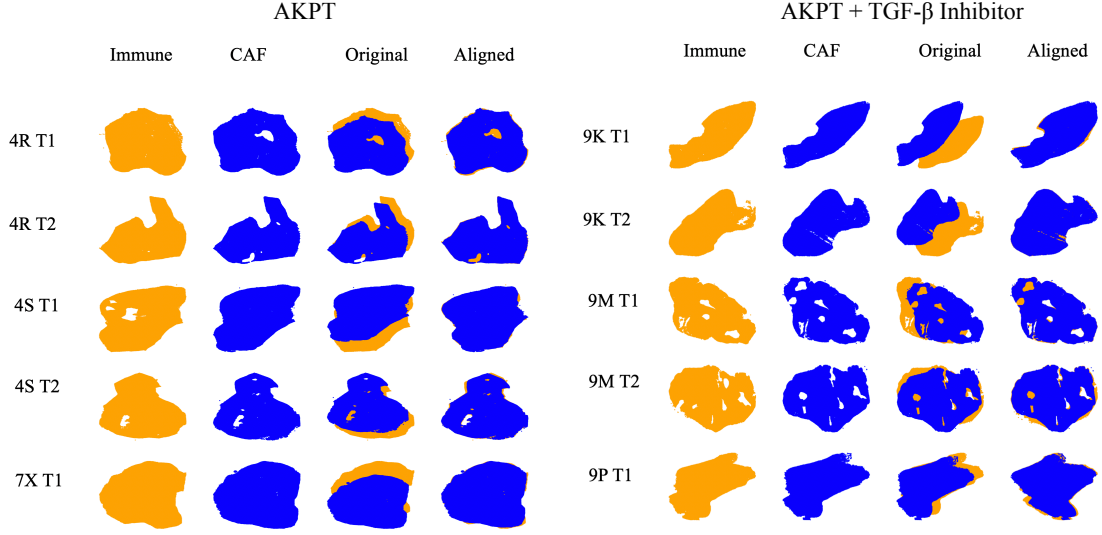


Figure B.2: **Spatial Integration of Mouse Fibroblast and Immune Panels** Adjacent tissues from all $n = 10$ mice are shown, with tissues stained with immune markers in blue, and fibroblast markers shown in orange. Row labels indicate sample IDs. Positions and orientations pre and post integration are shown.

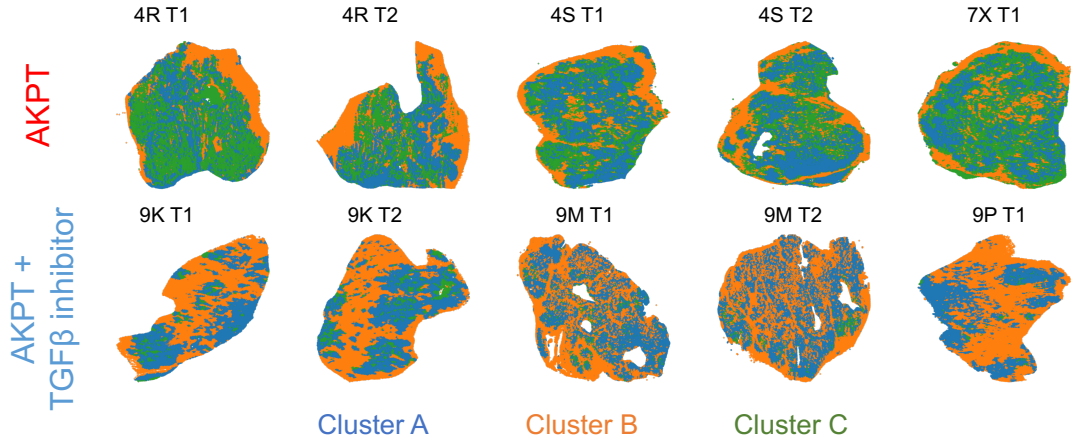


Figure B.3: **Spatial Domain Mapping of Mouse CRC Tissues** In each tissue, all cell microenvironments were assigned to their most similar microenvironment cluster using the procedure outlined in Section 4.3.3. Detailed cluster descriptions, including cell enrichment and condition specificity, are provided in Figure 4.4.

Bibliography

- [1] David J Adelstein, John A Ridge, Maura L Gillison, Anil K Chaturvedi, Gypsyamber D’Souza, Patti E Gravitt, William Westra, Amanda Psyrri, W Martin Kast, Laura A Koutsky, et al. Head and neck squamous cell cancer and the human papillomavirus: summary of a national cancer institute state of the science meeting, november 9–10, 2008, washington, dc. *Head & Neck: Journal for the Sciences and Specialties of the Head and Neck*, 31(11):1393–1422, 2009.
- [2] Athanassios Argiris, Michalis V Karamouzis, David Raben, and Robert L Ferris. Head and neck cancer. *The Lancet*, 371(9625):1695–1709, 2008.
- [3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- [4] Adrian Baddeley, Imre Bárány, and Rolf Schneider. Spatial point processes and their applications. *Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004*, pages 1–75, 2007.
- [5] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

- [6] Mariem Ben-Said. Spatial point-pattern analysis as a powerful tool in identifying pattern-process relationships in plant ecology: an updated review. *Ecological Processes*, 10(1):1–23, 2021.
- [7] 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.
- [8] Gaspard Beugnot, Aude Genevay, Kristjan Greenewald, and Justin Solomon. Improving approximate optimal transport distances using quantization. In *Uncertainty in artificial intelligence*, pages 290–300. PMLR, 2021.
- [9] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.
- [10] Mikhail Binnewies, Edward W Roberts, Kelly Kersten, Vincent Chan, Douglas F Fearon, Miriam Merad, Lisa M Coussens, Dmitry I Gabrilovich, Suzanne Ostrand-Rosenberg, Catherine C Hedrick, et al. Understanding the tumor immune microenvironment (time) for effective therapy. *Nature medicine*, 24(5):541–550, 2018.
- [11] Sarah Black, Darci Phillips, John W Hickey, Julia Kennedy-Darling, Vishal G Venkatarahaman, Nikolay Samusik, Yury Goltsev, Christian M Schürch, and Garry P Nolan. Codex multiplexed tissue imaging with dna-conjugated antibodies. *Nature protocols*, 16(8):3802–3835, 2021.
- [12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [13] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

- [14] Joshua A Bull and Helen M Byrne. Quantification of spatial and phenotypic heterogeneity in an agent-based model of tumour-macrophage interactions. *PLOS Computational Biology*, 19(3):e1010994, 2023.
- [15] Joshua A Bull, Philip S Macklin, Tom Quaiser, Franziska Braun, Sarah L Waters, Chris W Pugh, and Helen M Byrne. Combining multiple spatial statistics enhances the description of immune cell localisation within tumours. *Scientific reports*, 10(1):18624, 2020.
- [16] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE transactions on knowledge and data engineering*, 30(9):1616–1637, 2018.
- [17] Yongqiang Cai, Qianxiao Li, and Zuowei Shen. A quantitative analysis of the effect of batch normalization on gradient descent. In *International Conference on Machine Learning*, pages 882–890. PMLR, 2019.
- [18] Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.
- [19] Fenxiao Chen, Yun-Cheng Wang, Bin Wang, and C-C Jay Kuo. Graph representation learning: a survey. *APSIPA Transactions on Signal and Information Processing*, 9:e15, 2020.
- [20] 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.
- [21] Zhenghao Chen, Ilya Soifer, Hugo Hilton, Leeat Keren, and Vladimir Jojic. Modeling multiplexed images with spatial-lda reveals novel tissue microenvironments. *Journal of Computational Biology*, 27(8):1204–1218, 2020.

- [22] Cassandra Chow. An implementation of the iterative closest point algorithm using opencv. <https://github.com/casychow/Iterative-Closest-Point.git>, 2022.
- [23] Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261, 2004.
- [24] Albert H Coons, Hugh J Creech, and R Norman Jones. Immunological properties of an antibody containing a fluorescent group. *Proceedings of the society for experimental biology and medicine*, 47(2):200–202, 1941.
- [25] David Cunningham, Wendy Atkin, Heinz-Josef Lenz, Henry T Lynch, Bruce Minsky, Bernard Nordlinger, and Naureen Starling. Colorectal cancer. *The Lancet*, 375(9719):1030–1047, March 2010.
- [26] Haskell B Curry. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, 2(3):258–261, 1944.
- [27] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
- [28] Esther Danenberg, Helen Bardwell, Vito RT Zanutelli, Elena Provenzano, Suet-Feung Chin, Oscar M Rueda, Andrew Green, Emad Rakha, Samuel Aparicio, Ian O Ellis, et al. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature genetics*, 54(5):660–669, 2022.
- [29] George B Dantzig. Application of the simplex method to a transportation problem. *Activity analysis and production and allocation*, 1951.
- [30] Monica T Dayao, Alexandro Trevino, Honesty Kim, Matthew Ruffalo, H Blaize D’Angio, Ryan Preska, Umamaheswar Duvvuri, Aaron T Mayer, and Ziv Bar-

- Joseph. Deriving spatial features from in situ proteomics imaging to enhance cancer survival analysis. *Bioinformatics*, 39(Supplement_1):i140–i148, 2023.
- [31] Rik Derynck, Rosemary J Akhurst, and Allan Balmain. Tgf- β signaling in tumor suppression and cancer progression. *Nature genetics*, 29(2):117–129, 2001.
- [32] Anthony WF Edwards and Luigi Luca Cavalli-Sforza. A method for cluster analysis. *Biometrics*, pages 362–375, 1965.
- [33] Mirjana Efremova, Miquel Vento-Tormo, Sarah A Teichmann, and Roser Vento-Tormo. Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols*, 15(4):1484–1506, 2020.
- [34] David W Eggert, Adele Lorusso, and Robert B Fisher. Estimating 3-d rigid body transformations: a comparison of four major algorithms. *Machine vision and applications*, 9(5-6):272–290, 1997.
- [35] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulana, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239, 2019.
- [36] Danielle J Fassler, Shahira Abousamra, Rajarsi Gupta, Chao Chen, Maozheng Zhao, David Paredes, Syeda Areeha Batool, Beatrice S Knudsen, Luisa Escobar-Hoyos, Kenneth R Shroyer, et al. Deep learning-based image analysis methods for brightfield-acquired multiplex immunohistochemistry images. *Diagnostic pathology*, 15(1):1–11, 2020.
- [37] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

- [38] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boissunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.
- [39] Giorgio Gaglia, Megan L Burger, Cecily C Ritch, Danae Rammos, Yang Dai, Grace E Crossland, Sara Z Tavana, Simon Warchol, Alex M Jaeger, Santiago Naranjo, et al. Lymphocyte networks are dynamic cellular communities in the immunoregulatory landscape of lung adenocarcinoma. *Cancer Cell*, 41(5):871–886, 2023.
- [40] Ralucca Gera, Lázaro Alonso, Brian Crawford, Jeffrey House, JA Mendez-Bermudez, Thomas Knuth, and Ryan Miller. Identifying network structure similarity using spectral graph theory. *Applied network science*, 3:1–15, 2018.
- [41] Noah F Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology*, 40(4):555–565, 2022.
- [42] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.
- [43] Mia Hashibe, Paul Brennan, Shu-chun Chuang, Stefania Boccia, Xavier Castellsague, Chu Chen, Maria Paula Curado, Luigino Dal Maso, Alexander W Daudt, Eleonora Fabianova, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the international head and neck

- cancer epidemiology consortium. *Cancer Epidemiology Biomarkers & Prevention*, 18(2):541–550, 2009.
- [44] John W Hickey, Elizabeth K Neumann, Andrea J Radtke, Jeannie M Camarillo, Rebecca T Beuschel, Alexandre Albanese, Elizabeth McDonough, Julia Hatler, Anne E Wiblin, Jeremy Fisher, et al. Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. *Nature methods*, 19(3):284–295, 2022.
- [45] Henry S Horn. Measurement of “overlap” in comparative ecological studies. *The American Naturalist*, 100(914):419–424, 1966.
- [46] Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover’s distance. *Advances in neural information processing systems*, 29, 2016.
- [47] Yoshiro Itatani, Kenji Kawada, and Yoshiharu Sakai. Transforming growth factor- β signaling pathway in colorectal cancer and its tumor microenvironment. *International journal of molecular sciences*, 20(23):5822, 2019.
- [48] Rene Jackstadt, Sander R van Hooff, Joshua D Leach, Xabier Cortes-Lavaud, Jeroen O Lohuis, Rachel A Ridgway, Valerie M Wouters, Jatin Roper, Timothy J Kendall, Campbell S Roxburgh, et al. Epithelial notch signaling rewires the tumor microenvironment of colorectal cancer to drive poor-prognosis subtypes and metastasis. *Cancer cell*, 36(3):319–336, 2019.
- [49] Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. Inference and analysis of cell-cell communication using cellchat. *Nature communications*, 12(1):1088, 2021.

- [50] Minoru Kanehisa. The kegg database. In *‘In silico’ simulation of biological processes: Novartis Foundation Symposium 247*, volume 247, pages 91–103. Wiley Online Library, 2002.
- [51] Leonid V Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422, 1960.
- [52] Ryan T Kelly. Single-cell proteomics: progress and prospects. *Molecular & Cellular Proteomics*, 19(11):1739–1748, 2020.
- [53] Junbum Kim, Samir Rustam, Juan Miguel Mosquera, Scott H Randell, Renat Shaykhiev, André F Rendeiro, and Olivier Elemento. Unsupervised discovery of tissue architecture in multiplexed imaging. *Nature Methods*, 19(12):1653–1661, 2022.
- [54] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1):196–205, 2009.
- [55] Wojtek J Krzanowski and YT Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34, 1988.
- [56] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- [57] Robert R Langle and Isaiah J Fidler. Tumor cell-organ microenvironment interactions in the pathogenesis of cancer metastasis. *Endocrine reviews*, 28(3):297–321, 2007.
- [58] Felix Järemo Lawin, Martin Danelljan, Fahad Shahbaz Khan, Per-Erik Forssén, and Michael Felsberg. Density adaptive point set registration. In *Proceedings*

- of the *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3829–3837, 2018.
- [59] Der-Tsai Lee and Bruce J Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980.
- [60] Erel Levine and Eytan Domany. Resampling method for unsupervised estimation of cluster validity. *Neural computation*, 13(11):2573–2593, 2001.
- [61] Jia-Ren Lin, Mohammad Fallahi-Sichani, Jia-Yun Chen, and Peter K Sorger. Cyclic immunofluorescence (cycif), a highly multiplexed method for single-cell imaging. *Current protocols in chemical biology*, 8(4):251–264, 2016.
- [62] Candace C Liu, Erin F McCaffrey, Noah F Greenwald, Erin Soon, Tyler Risom, Kausalia Vijayaragavan, John-Paul Oliveria, Dunja Mrdjen, Marc Bosse, Dmitry Tebaykin, et al. Multiplexed ion beam imaging: insights into pathobiology. *Annual Review of Pathology: Mechanisms of Disease*, 17:403–423, 2022.
- [63] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.
- [64] Irvin J Lustig, Roy E Marsten, and David F Shanno. Interior point methods for linear programming: Computational state of the art. *ORSA Journal on Computing*, 6(1):1–14, 1994.
- [65] Derek Magee, Yi Song, Stephen Gilbert, Nicholas Roberts, Nagitha Wijayathunga, Ruth Wilcox, Andrew Bulpitt, and Darren Treanor. Histopathology in 3d: From three-dimensional reconstruction to multi-stain and multi-modal analysis. *Journal of pathology informatics*, 6(1):6, 2015.

- [66] Natalia Garcia Martin, Stefano Malacrino, Marta Wojciechowska, Leticia Campo, Helen Jones, David C Wedge, Chris Holmes, Korsuk Sirinukunwatana, Heba Sailem, Clare Verrill, et al. A graph based neural network approach to immune profiling of multiplexed tissue samples. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3063–3067. IEEE, 2022.
- [67] Matthew McCormick, Xiaoxiao Liu, Julien Jomier, Charles Marion, and Luis Ibanez. Itk: enabling reproducible research and open science. *Frontiers in neuroinformatics*, 8:13, 2014.
- [68] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [69] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.
- [70] Al Mead. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(1):27–39, 1992.
- [71] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.
- [72] Jeffrey R Moffitt, Emma Lundberg, and Holger Heyn. The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics*, 23(12):741–759, 2022.
- [73] Jesper Møller and Rasmus P Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.

- [74] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.
- [75] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.
- [76] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010.
- [77] Masoud Najafi, Nasser Hashemi Goradel, Bagher Farhood, Eniseh Salehi, Somaye Solhjoo, Heidar Toolee, Ebrahim Kharazinejad, and Keywan Mortezaee. Tumor microenvironment: Interactions and therapy. *Journal of cellular physiology*, 234(5):5700–5721, 2019.
- [78] Ajit Johnson Nirmal, Yu-An Chen, and Artem Sokolov. labsyspharm/scimap: Release-0.19.0, April 2022.
- [79] James B Orlin. Max flows in $\mathcal{O}(nm)$ time, or better. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 765–774, 2013.
- [80] Edwin Roger Parra. Methods to determine and analyze the cellular spatial distribution extracted from multiplex immunofluorescence data to understand the tumor microenvironment. *Frontiers in Molecular Biosciences*, 8:668340, 2021.
- [81] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

- [82] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [83] 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.
- [84] 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.
- [85] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.
- [86] André F Rendeiro, Hiranmayi Ravichandran, Yaron Bram, Vasuretha Chandar, Junbum Kim, Cem Meydan, Jiwoon Park, Jonathan Foox, Tyler Hether, Sarah Warren, et al. The spatial landscape of lung pathology during covid-19 progression. *Nature*, 593(7860):564–569, 2021.
- [87] Brian D Ripley. The second-order analysis of stationary point processes. *Journal of applied probability*, 13(2):255–266, 1976.
- [88] Brian D Ripley. *Spatial statistics*. John Wiley & Sons, 2005.
- [89] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

- [90] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.
- [91] Samir Rustam, Yang Hu, Seyed Babak Mahjour, Andre F Rendeiro, Hiranmayi Ravichandran, Andreacarola Urso, Frank D’Ovidio, Fernando J Martinez, Nasser K Altorki, Bradley Richmond, et al. A unique cellular organization of human distal airways and its disarray in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, 207(9):1171–1182, 2023.
- [92] Erik Sahai, Igor Astsaturov, Edna Cukierman, David G DeNardo, Mikala Egeblad, Ronald M Evans, Douglas Fearon, Florian R Greten, Sunil R Hingorani, Tony Hunter, et al. A framework for advancing our understanding of cancer-associated fibroblasts. *Nature Reviews Cancer*, 20(3):174–186, 2020.
- [93] Denis Schapiro, Hartland W Jackson, Swetha Raghuraman, Jana R Fischer, Vito RT Zanutelli, Daniel Schulz, Charlotte Giesen, Raúl Catena, Zsuzsanna Varga, and Bernd Bodenmiller. histocat: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods*, 14(9):873–876, 2017.
- [94] Denis Schapiro, Artem Sokolov, Clarence Yapp, Yu-An Chen, Jeremy L Muhlich, Joshua Hess, Allison L Creason, Ajit J Nirmal, Gregory J Baker, Maulik K Nariya, et al. Mcmicro: a scalable, modular image-processing pipeline for multiplexed tissue imaging. *Nature methods*, 19(3):311–315, 2022.
- [95] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

- [96] Christian M Schürch, Salil S Bhate, Graham L Barlow, Darci J Phillips, Luca Noti, Inti Zlobec, Pauline Chu, Sarah Black, Janos Demeter, David R McIlwain, et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359, 2020.
- [97] A Shapiro and Y Wardi. Convergence analysis of gradient descent stochastic algorithms. *Journal of optimization theory and applications*, 91:439–454, 1996.
- [98] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, Ahmedin Jemal, et al. Cancer statistics, 2021. *Ca Cancer J Clin*, 71(1):7–33, 2021.
- [99] Faranak Sobhani, Azam Hamidinekoo, Allison H Hall, Lorraine King, Jeffrey R Marks, Carlo Maley, Hugo M Horlings, E Shelley Hwang, and Yinyin Yuan. Automated dcis identification from multiplex immunohistochemistry using generative adversarial networks. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.
- [100] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.
- [101] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.
- [102] Wei Chang Colin Tan, Sanjna Nilesch Nerurkar, Hai Yun Cai, Harry Ho Man Ng, Duoduo Wu, Yu Ting Felicia Wee, Jeffrey Chun Tatt Lim, Joe Yeong, and Tony Kiat Hon Lim. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications*, 40(4):135–153, 2020.

- [103] Mattia Tantardini, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi. Comparing methods for comparing networks. *Scientific reports*, 9(1):17557, 2019.
- [104] Daniele VF Tauriello, Sergio Palomo-Ponce, Diana Stork, Antonio Berenguer-Llargo, Jordi Badia-Ramentol, Mar Iglesias, Marta Sevillano, Sales Ibiza, Adrià Cañellas, Xavier Hernando-Momblona, et al. Tgf β drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature*, 554(7693):538–543, 2018.
- [105] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.
- [106] Maozhen Tian and William P Schiemann. The tgf- β paradox in human cancer: an update. 2009.
- [107] M Titford. The long history of hematoxylin. *Biotechnic & histochemistry*, 80(2):73–78, 2005.
- [108] 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.
- [109] Masayoshi Tojo, Yoshio Hamashima, Aki Hanyu, Tetsuya Kajimoto, Masao Saitoh, Kohei Miyazono, Manabu Node, and Takeshi Imamura. The alk-5 inhibitor a-83-01 inhibits smad signaling and epithelial-to-mesenchymal transition by transforming growth factor- β . *Cancer science*, 96(11):791–800, 2005.
- [110] Vincent Traag, Fabio Zanini, Ryan Gibson, Oren Ben-Kiki, and Dafne van Kuppevelt. vtraag/leidenalg 0.8.2, September 2020.

- [111] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.
- [112] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alexander Bronstein, and Emmanuel Müller. Netlsd: hearing the shape of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2347–2356, 2018.
- [113] Milad R Vahid, Erin L Brown, Chloé B Steen, Wubing Zhang, Hyun Soo Jeon, Minji Kang, Andrew J Gentles, and Aaron M Newman. High-resolution alignment of single-cell and spatial transcriptomes with cytospace. *Nature Biotechnology*, pages 1–6, 2023.
- [114] MNM Van Lieshout and AJ1422574 Baddeley. A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50(3):344–361, 1996.
- [115] Nuria Vaquero-Siguero, Nikolai Schleussner, Julia Volk, Manuel Mastel, Jasmin Meier, and Rene Jackstadt. Modeling colorectal cancer progression reveals niche-dependent clonal selection. *Cancers*, 14(17):4260, 2022.
- [116] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.
- [117] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. 2019.
- [118] Ewald R Weibel and Domingo M Gomez. Architecture of the human lung: use of quantitative methods establishes fundamental relations between size and number of lung structures. *Science*, 137(3530):577–585, 1962.

- [119] TL Whiteside. The tumor microenvironment and its role in promoting tumor growth. *Oncogene*, 27(45):5904–5912, 2008.
- [120] Zhenqin Wu, Alexandro E Trevino, Eric Wu, Kyle Swanson, Honesty J Kim, H Blaize D’Angio, Ryan Preska, Gregory W Charville, Piero D Dalerba, Ann Marie Egloff, et al. Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering*, 6(12):1435–1448, 2022.
- [121] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.
- [122] Yaochen Xie, Zhao Xu, Jingtun Zhang, Zhengyang Wang, and Shuiwang Ji. Self-supervised learning of graph neural networks: A unified review. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2412–2429, 2022.
- [123] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.
- [124] Ron Zeira, Max Land, Alexander Strzalkowski, and Benjamin J Raphael. Alignment and integration of spatial transcriptomics data. *Nature Methods*, 19(5):567–575, 2022.
- [125] Weiruo Zhang, Irene Li, Nathan E Reticker-Flynn, Zinaida Good, Serena Chang, Nikolay Samusik, Saumyaa Saumyaa, Yuanyuan Li, Xin Zhou, Rachel Liang, et al. Identification of cell types in multiplexed in situ images by combining protein expression and spatial information using celesta. *Nature Methods*, 19(6):759–769, 2022.

- [126] Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision*, 13(2):119–152, 1994.
- [127] Ping Zhu and Richard C Wilson. A study of graph spectra for comparing graphs. In *BMVC*, 2005.