Technical University of Munich
TUM School of Life Sciences

Towards Modelling Spatial Components of Cellular Variation

Giovanni Palla

Complete reprint of the dissertation approved by the TUM School of Life Sciences of the Technical University of Munich for the award of the

Doktor der Naturwissenschaften.

Chair: Prof. Dr. Dmitrij Frischmann

Examiners:

1. Prof. Dr. Dr. Fabian Theis
2. Prof. Dr. Oliver Stegle
3. Assoc. Prof. Xiao Wang, Ph.D.

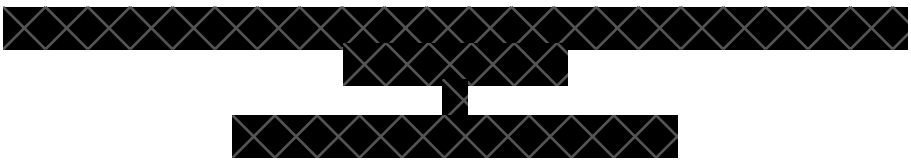
The dissertation was submitted to the Technical University of Munich on 25.03.2024 and accepted by the TUM School of Life Sciences on 06.11.2024

Towards Modelling Spatial Components of Cellular Variation

Computational Methods for Spatial Omics Data

by

Giovanni Palla



Abstract

Single-cell genomics has transformed how we study and understand cellular diversity, enabling an unparalleled resolution that facilitates the creation of maps of cellular variation across tissues and organs. The latest technological advancements have made it possible to examine molecular profiles at the level of individual cells within their native tissue environments, enabling the development of spatially-resolved cellular maps. However, these advancements bring computational challenges that must be addressed to gain biological insights hidden within the data. In this thesis, I employ computational approaches to analyze the spatial organization of cellular niches. By integrating spatial statistics and graph theory, I demonstrate the ability to uncover spatial patterns within tissues, such as the identification of cell neighborhoods that organize in determined tissue niches or the description of how cells are organized spatially. Additionally, I introduce tools designed for the efficient handling and processing of large-scale, multimodal spatial omics datasets. These tools are supported by a robust infrastructure that covers the entire data analysis workflow, from storage to high-performance visualization. Leveraging methods from Optimal Transport, I develop techniques for integrating single-cell and spatial omics data, applying these methods to predict gene and protein levels in spatial omics data of the mouse liver. I also explore how Optimal Transport can align spatial omics datasets and enable the identification of spatially-resolved developmental trajectories, illustrating its utility with adult mouse brain and mouse embryogenesis case studies. These methodologies are distributed in open-source software tools, available for researchers seeking to tackle similar challenges with their data. Overall, the strategies outlined in this thesis advance the methodologies available for spatial omics data analysis, pushing forward the capabilities of machine learning methods and software for tissue biology.

Zusammenfassung

Die Einzelzellgenomik hat die Art und Weise, wie wir zelluläre Vielfalt studieren und verstehen, grundlegend verändert und ermöglicht eine beispiellose Auflösung, die die Erstellung von Karten zellulärer Variationen in Geweben und Organen erleichtert. Die neuesten technologischen Fortschritte haben es möglich gemacht, molekulare Profile auf der Ebene einzelner Zellen in ihrer natürlichen Gewebe Umgebungen zu untersuchen, wodurch die Entwicklung räumlich aufgelöster zellulärer Karten ermöglicht wird. Diese Fortschritte bringen jedoch algorithmische Herausforderungen mit sich, die bewältigt werden müssen, um die in den Daten verborgenen biologischen Einsichten zu gewinnen. In dieser Arbeit setze ich rechnergestützte Ansätze ein, um die räumliche Organisation von zellulären Nischen zu analysieren. Durch die Integration von räumlicher Statistik und Graphentheorie ermögliche ich räumliche Muster in Geweben aufzudecken, wie beispielsweise die Identifizierung von Zellnischen, die sich in bestimmten Gewebenischen organisieren, oder die Beschreibung, wie Zellen räumlich organisiert sind. Darüber hinaus stelle ich Methoden vor, die für die effiziente Handhabung und Verarbeitung von großen, multimodalen räumlichen Omics-Datensätzen konzipiert sind. Diese Methoden werden durch eine robuste Infrastruktur unterstützt, die den gesamten Datenanalyse-Workflow abdeckt, von der Speicherung bis zur Visualisierung. Unter Ausnutzung von Methoden aus dem Transportproblem entwickle ich Algorithmen für die Integration von Einzelzell- und räumlichen Omics-Daten und wende diese Methoden an, um Gen- und Proteinexpression in räumlichen Omics-Daten der Mausleber vorherzusagen. Ich erforsche auch, wie das Transportproblem räumliche Omics-Datensätze ausrichten kann und die Identifizierung von räumlich aufgelösten Entwicklungsverläufen ermöglicht, und illustriere dessen Nutzen mit Fallstudien zum erwachsenen Mausgehirn und zur Mausembryogenese. Diese Methodologien werden in open-source Software verteilt, die für Forscher verfügbar sind, die ähnliche Herausforderungen mit ihren Daten angehen möchten. Insgesamt bringen die in dieser Arbeit umrissenen Strategien die verfügbaren Methoden für die Analyse von räumlichen Omics-Daten voran und treiben die Fähigkeiten von maschinellem Lernen und Software in der Gewebebiologie voran.

Acknowledgments

I want to thank my supervisor, Fabian Theis, for mentorship in all areas of science and career development, for always pushing me to improve and make the most of my research projects. I would also like Fabian for his support and research freedom I was able to enjoy during my PhD. I would also like to thank Prof. Prisca Liberali and Dr. Benjamin Schubert for their academic support as part of my Thesis Advisory Committee. I would also like to thank Prof. Marco Cuturi and Prof. Lorin Crawford for the various collaborations. I was very fortunate to be able to work with many leaders in their field, from whom I learned a lot.

I would like to thank the members of Theislab, for collaborations, chats over coffee, emotional support and for making it a great and stimulating place to do research. The concentration of talent and creativity in the group has always amazed me, and I will cheerish the memories and experiences for a long time.

I would also like to thank the members of the ICB admin: Sabine Kunz, Dr. Anna Sacher, Daniela Herrmann, Karina Kahl, Luka Schwarzlose, and from MUDS: Dr. Julia Schlehe and Mara Kieke. I thank you for the unparalleled support and help to navigate the administrative and bureaucratic difficulties of the Helmholtz Center Munich.

Vorrei infine ringraziare la mia famiglia ed Emma per il loro supporto in questi anni.

Contents

1 Introduction

Spatial Biology is an emerging field that aims at characterizing and modeling molecular determinants of cellular behavior in tissue. It builds on the foundations of cell biology (*Section.* 1.1), and particularly high-throughput approaches such as single-cell RNA sequencing and multiplexed imaging (*Section.* 1.2, *Section.* 1.3). Modern experimental techniques have enabled the investigation of phenomena of tissue biology, such as cell-cell communication and tissue organization, at an unprecedented resolution (*Section.* 1.3). The richness and complexity of the data, paired with the exploding progress of Machine Learning and Artificial Intelligence, have enabled to push the boundaries of the field and address fundamental questions in tissue biology, which represent the focus of this thesis (*Section.* 1.4, *Section.* 1.5). Finally, scaling efforts in both experimental techniques and computational methods will allow building multiscale models of tissue and organs at single cell resolution (*Section.* 1.6).

1.1 Cell Biology

Cells are the foundational units of life. In unicellular organisms, a single cell performs all necessary functions for reacting to environmental shifts, growing, developing, reproducing, and maintaining homeostasis [1, 2]. In contrast, multicellular organisms entail specialized cells that synergistically contribute to forming complex macrostructures. In multicellular organisms, cells collectively constitute tissues and organs, which, taken together, build organisms designed by evolutionary processes to carry out the same functions: react to environmental shifts, grow, develop, reproduce, and maintain homeostasis [1].

Cell biology has been historically pivotal in studying the molecular basis of life. Discoveries in molecular biology considerably advanced our understanding by linking cellular studies to molecular entities like nucleic acids, proteins, and metabolites. Historically, microscopy was the primary tool for analyzing cellular structures and functions [3]. However, the emergence of high-throughput technologies, like sequencing and Mass Spectrometry (MS)-based proteomics and RNA-sequencing, has broadened our perspective on the diversity of cellular phenotypes. Quantifying the abundance of RNA and proteins helps scientists describe cellular phenotypes and derive mechanistic insights into cellular state and function.

In the first decade of the XXI century, the primary approach was bulk RNA sequencing, measuring the RNA abundance in aggregates of cells. However, advancements in single-cell RNA sequencing have extensively augmented our capacity to dissect the phenotypic heterogeneity and composition of tissues and organs [4]. It's crucial to note that while standard microscopy tech-

niques provide single-cell resolution, they usually suffer from throughput limitations that make them unable to sample many observations (many cells, or large tissue sections) or measure a broad spectrum of molecules in terms of diversity and abundance (many genes, with sufficient depth and sensitivity). Conversely, the molecular profiles derived from high-throughput techniques provide a high number of measured cells (or samples) to recover the heterogeneity of cellular states or phenotypes based solely on feature distribution similarity (i.e., which cells express which genes and how abundant). It should be further noted that "bulk" RNA-sequencing, due to measuring aggregated samples, is less prone to noise and recovers a higher number of expressed genes, it is not able to recover important yet rare gene expression signatures particular for low-abundant cell types that entails a small fraction of the bulk composition [5, 6, 7].

1.2 High-Throughput Single-Cell Biology

In the mid-2010s, there was a transformative development in the characterization of cellular phenotypes by introducing droplet-based single-cell RNA sequencing [4, 8]. This technique's primary innovation lies in its capacity to isolate individual cells within droplets, facilitating subsequent library preparation and direct sequencing of the RNA with standard high-throughput sequencing methodologies9, as well as the introduction of Unique Molecular Identifiers (UMI), which consists of random nucleotide sequences that eliminate cDNA amplification noise as well improvement in mRNA capture efficiency [9]. As a result, it has become feasible to profile RNA abundance at a single-cell level and large scale. Rapidly after that, this methodology was adapted to examine various molecular modalities: DNA sequencing for deciphering copy number variations and structural DNA changes [10], Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq) for chromatin accessibility [11], histone modifications [12, 13] as indicators of chromatin states, transcription factor binding occurrences [12], active transcription sites [14], and numerous other post-transcriptional or post-translational events [15]. Moreover, modifications of this technology have permitted the quantification of protein abundance through antibody binding for both surface and intracellular proteins [16, 17], as well as untargeted protein assessments via MS [18]. Nevertheless, all droplet- or plate-based single-cell technologies dissociate the biological samples, removing cells from their native tissue context.

Similarly, new generations of high-throughput, high-content screens, which consist of screening an extensive collection of cell (or organoids) phenotypes (possibly under perturbation) [19, 20, 21, 22] have also been developed. While they are similar in scope: profiling large numbers of cells in various conditions, yet they can recover the original spatial context, they lack the multiplexing ability in terms of the number of molecular entities recovered, as often only a few (between four and a dozen markers) molecular entities can be measured.

Recently, the omics and high-throughput imaging fields have started to converge, borrowing experimental techniques and computational tools from each other and starting to transition to a new view on cellular states that preserve their tissue context: spatial biology [23, 24, 25].

1.3 The Emerging Field of Spatial Biology

Spatial Biology entails profiling single-cell phenotypes at high throughput yet maintaining the contextual information of cellular neighborhood [24, 25, 26, 27, 28]. This approach seeks to evaluate cellular states within their original context by harnessing sophisticated, multi-modal microscopy techniques. This transition can be viewed as reversing to the original tools and investigative methods of Cell Biology: the microscope.

Advancements in high-throughput microfluidics and sophisticated RNA primer designs UMI have predominantly propelled the emergence of single-cell genomics [4]. However, a limited number of advanced research laboratories concentrated on refining imaging techniques that are able to measure single transcripts Fluorescent In Situe Hybridization (FISH) [29, 30] and orthogonal imaging techniques that are able to measure proteins in tissues, such as multiplexed Immuno-Histo-Chemistry (mIHC) [31, 32]. The objective was to parallel the comprehensiveness of profiling cellular states achieved by single-cell genomics while preserving the cellular tissue context. Concurrently, innovative methods have been designed for RNA capture from tissue sections, allowing RNA to be mapped to its spatial coordinates and, hence, its cellular tissue context. Although the foundational papers detailing these techniques coincided with significant milestones in dissociated single-cell genomics, their robust application and scalability necessitated extended development. In the recent two years, commercial solutions have emerged, enabling research teams without expertise in advanced microscopy to utilize these techniques reliably and extensively [33, 34]. The affordability and ease of use of these technologies will revolutionize the study of tissue structure and function in the following years, from development to disease.

Similar to single-cell genomics technologies, spatial omics methodologies can be categorized based on the molecular modality being measured, predominantly RNA or protein, and the experimental technique employed for the measurements [35, 36]. Spatial Transcriptomics (ST) focuses on RNA profiling and can be further categorized into sequencing-based or imaging-based methods [24]. Array-based approaches utilize specialized slides coated with probes that capture the RNA molecular, enabling the subsequent reverse transcription and standard library preparation for commercial sequencing platforms [25]. In contrast, imaging-based methods fundamentally rely on microscopy techniques, pinpointing the location of individual RNA molecules of specific genes within tissue. Here, RNA detection is performed by fluorescence probes, which are either direct evolutions of FISH methodologies or through in situ sequencing, wherein sequential nucleotide additions release detectable fluorophores. For spatial proteomics technologies, the primary measurement tool is antibodies modified with either fluorophores or heavy isotopes. Fluorophore-modified antibodies are analyzed using conventional fluorescence microscopy, a process often termed cyclical mIHC, necessitating iterative cycles of antibody staining and washing [31]. In contrast, heavy isotope-modified antibodies are profiled via MS, where the antibodies are ionized using multiplexed ion beam imaging by time of flight (MIBI-TOF), and the resulting isotopes are subsequently deconvoluted and quantified using standard MS analysis pipelines [23, 37, 38]. It should also be noted that, as for dissociated single-cell technologies, spatial techniques have also been extended to measure additional modalities such as metabolites [39], immune-cell receptor diversity [40], or chromatin organization [41]. *Figure.* 1.1 presents an overview of the trade-off between technologies. Besides the molecular entity being measured, there is an additional

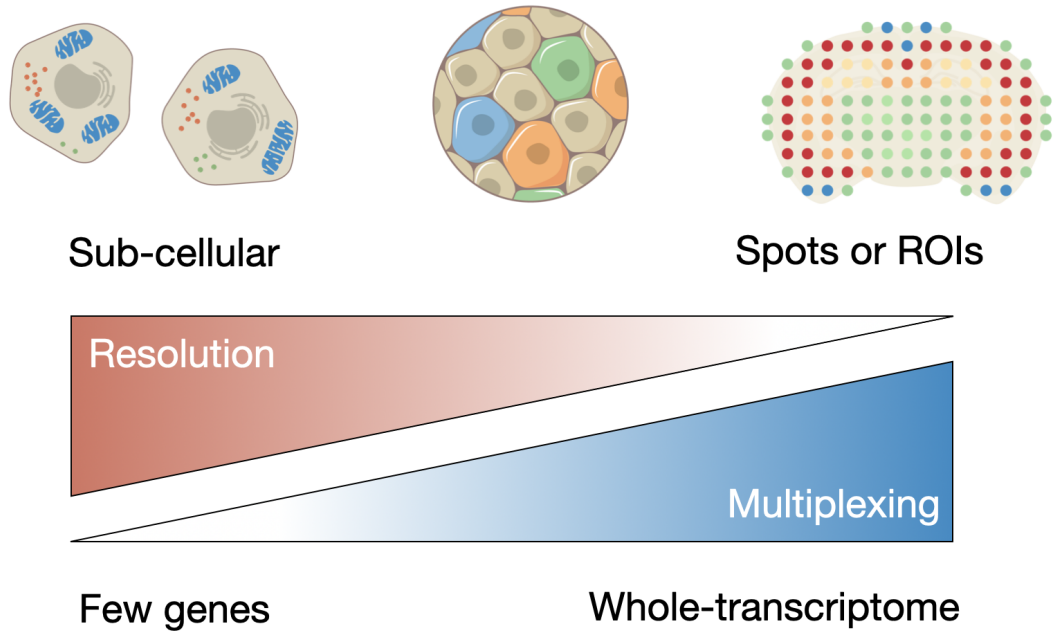


Figure 1.1: **High throughput spatial biology for deep molecular characterization of tissues.** Spatial omics technologies differ vastly in the molecular entities they profile (DNA, mRNA, proteins, metabolites), the resolution (from subcellular to super-cellular regions) and the multiplexing (from few genes to unbiased whole transcriptome). Novel experimental techniques are continuously improving the constraints of this trade off. This figure has been adapted from [42]

trade-off that is important to recognize between these technologies: in fact, there is no experimental technique that excels in high resolution, multiplexing and throughput at the same time, thus posing key limitations in the type of biological phenomena of interest that can be measured.

1.4 Spatial Components of Tissue Biology

Spatial biology methodologies hold significant potential for bridging the comprehensive cellular state characterization enabled by scRNA-seq and the long-standing tradition of multi-scale morphological examination of tissue architecture and functionality [35, 43]. Profiling both morphological and molecular variations at scale facilitates addressing a spectrum of biological problems: from elucidating mechanisms of cell-cell interactions to pinpointing the determining features of cellular microenvironments, from discerning multi-cellular functional assemblies within tissues to uncovering emergent functional properties of tissues, to crafting spatially-resolved atlases of tissues and organs towards building a multi-scale Common Coordinate System (CCS) for entire organisms. The nascent field of tissue biology already showcases numerous illustrative examples where integrating molecular profiling with tissue context has helped unlock complex biological questions. Examples include characterizing the dynamics of the tumor microenvironment [44, 45, 46, 47, 48], mainly how its variability can indicate cancer progression [46, 49], and providing

extensive maps of developmental cellular trajectories across spatial and temporal dimensions [50, 51, 52, 53, 54, 55].

Cell-Cell Communication (CCC) is a key determining factor of multicellular architecture [56, 57]. Coordinating cellular activities in multicellular organisms is essential and hinges on cell-cell interactions. To study the phenomenon of CCC, it is crucial to investigate cellular functions and phenotypes within its original tissue context and disentangle the molecular basis of such communication, from ions to secreted proteins. Some of these molecules, like cell adhesion molecules, support structural interactions, while others, like hormones or neurotransmitters, facilitate CCC [58]. Monitoring messenger molecules and their pathways is crucial for grasping the significance and depth of cellular communication. Mainly because CCC is responsible for triggering downstream signaling cascades that are responsible for rewiring the gene regulatory network of the cell [59]. Traditional methods based on single-cell genomics data offered a generalized overview of the potential communication and are based on the hypothesis that the transcriptome should contain the "footprint" of the CCC, which manifests as gene signatures (both ligand-receptor pairs but also signaling molecules) [60, 61, 62]. Spatial biology data enables novel machine learning algorithms to directly model such interactions [63, 64], thus recovering CCC's spatial and molecular components. CCC determines the emergence of cellular niches, which can be identified as groups of cells with a similar neighborhood composition [63, 65], that is, a similar composition in terms of cell types in its proximities. Tissue niches can be seen as the spatial equivalent of cell types and cell states [66]. Cellular niches respectively give rise to functional tissue units [67, 68]. Such units represent functional tissue elements that carry out specific physiological functions, such as liver lobules, islets in the pancreas, or brain regions. Characterizing such units is paramount to understanding the complexities of tissues and their malfunction in diseases [69]. Functional tissue units are can also be referred to as tissue modules [43], which are recurring cellular groupings in tissues that perform specific tasks and may vary in composition across different locations [70]. These modules, which can vary significantly in size and cell composition, are defined by characteristics unique to tissue areas rather than individual cells. These modules can present challenges for identification. Identifying cellular niches is the first step to characterizing tissue modules, from high-resolution cellular interactions to lower-resolution yet higher spatial context of tissue modules. In fact, as their scale increases, it becomes harder to distinguish them due to overlapping biological processes. Furthermore, the structural organization differs across tissues, making a universal, cross-tissue model challenging to formalize. Metabolic and morphogenetic gradients provide information about developmental and environmental variables [58, 71, 72]. These gradients and biophysical forces in tissues are additional factors contributing to the emergence and shaping of tissue modules.

1.5 Models of Tissue Biology

The single-cell genomics revolution has been jointly shaped by advancements in experimental methodologies and the integration of innovative machine learning methods to model cell-state heterogeneity [66, 73]. This field has witnessed the productive employment of an array of machine learning paradigms, encompassing non-linear dimensionality reduction tools like t-distributed

Stochastic Neighbor Embedding (t-SNE) [74] and Uniform Manifold Approximation and Projection (UMAP) [75], the adaptation of non-linear latent factor models such as Variational Autoencoders (VAE) in biological contexts [76], and the quantitative and mechanistic modelling of cellular differentiation [77, 78]. This latter aspect has catalyzed the proliferation of trajectory-inference algorithms designed to elucidate cellular differentiation processes and lineage hierarchies [79, 80]. Another pivotal challenge in the single cell RNA-sequencing (scRNA-seq) data analysis process has been data integration and batch correction, which concerns the integration of datasets from disparate experimental sources or individuals [73, 81]. The complexities inherent to single-cell genomics data modeling have also inspired the formulation of novel machine-learning methods, motivating innovative algorithms [82] and influential benchmarks for robust predictions [83].

Simultaneously, the single-cell genomics community has also pushed the boundaries of long-standing challenges in statistical modeling for biological data. For instance, ongoing research endeavors are grappling with determining the most appropriate noise model for RNA-seq quantification, subsequently influencing optimal normalization strategies [84]. Differential testing, which pinpoints associations between gene expression and specific covariates (like in a case-control study), has undergone rigorous investigation and has been expanded to various subcategories, including differential abundance [85] of cellular state and donor-centric differential testing utilizing pseudobulking techniques [86].

Spatial biology have expanded the horizon of challenges and opportunities for computational analysts interested in modeling spatial components of cellular variation [43](*Figure.* 1.2). Computational approaches to problems in spatial biology have built on top of the algorithmic and software innovations in single cell genomics [87, 88] and high-throughput imaging [89, 90, 91], and extended them to accommodate for the unique challenges of spatial omics data. For example, the data integration challenge in spatial biology does not only concern the removal of technical variations across samples but also harmonizing different fields of view and tissue cross-sections, highlighting the need to account for variability across biological samples but also across sections of the same biological sample [92, 93]. Furthermore, multimodal alignment or integration is also of interest for spatial omics data analysts and method developers, particularly with the advent of multimodal spatial omics data [94, 95]. Similarly, the integration of single cell data with spatial omics measurements has also been extensively studied, with several methods that aims at either mapping cell types proportions, also called deconvolution methods [96, 97], or in addition also gene expression or multimodal measurements [98, 99], of which one of the methods for such problem is also presented in this thesis [100]. Spatial omics data also requires an additional level of integration, that is the integration of spatial coordinates in consecutive sections or similar sections from different samples. This is very similar to the problem of image or point cloud registration in computer vision and signal processing [101, 102]. Nevertheless, the alignment (or registration) of spatial coordinates between similar (yet deformed) spatial omics measurements has its nuances. Specifically, the signal at each pixel/point is a vector of gene expression variation that the registration algorithms should consider. Various approaches have been suggested [103, 104, 105], of which one is also presented in this thesis [100].

Furthermore, explicitly modeling spatial elements has revived interest in various spatial statistical metrics, such as identifying spatially variable genes [106, 107, 108, 109], demarcation of spatial domains [110, 111], neighborhood enrichment of cellular niches [42], and linear correlations between niche compositions and molecular profiles [63]. Similarly, linear and non-linear factor models have been extended to account for the spatial coordinates of cells, effectively learning latent factors of variations that incorporate the cell's spatial context [112, 113, 114, 115].

Trajectory inference methodologies for spatial omics data have also emerged [100, 116]. Notably, while current spatial profiling techniques provide a snapshot view of biological processes, they provide additional information that influences determinants of cell fate, lineage commitment, and differentiation trajectories [43, 92]. In many physical contexts, due to the restricted mobility of cells within tissues, their spatial adjacency often signifies cellular communication, evident in scenarios such as the relationship between stem cells in an intestinal crypt and the neighboring cells in the villus [72, 117]. Several modeling efforts have also been focused on modeling cellular niches, by either disentangling neighborhood effect on gene expression variability relying with linear models approaches [63] or finding common cell-type neighborhood patterns across the tissue section of the study, thus retrieving potential CCC events and aiding the identification of tissue niches [64, 65, 118], see also *Section.* 1.3 1.4.

The integration and alignment of multiple modalities have evolved beyond solutions initially conceptualized for single-cell genomics, given that spatial omics frequently capture molecular signatures and cellular morphology. The nature of such paired measurements also opens up research avenues on modality alignment and translation [119], providing high-quality data for long-standing research questions such as linking the molecular profile of a cell to its morphology [92]. Exciting advanced in this domain will entail the adoption of current breakthroughs in generative AI and large multimodal models [120, 121, 122]. Finally, the scale and heterogeneity of spatial omics data present formidable obstacles to the data management infrastructures initially designed for single-cell genomics. Conventional solutions tailored for tabular data often need to improve their ability to accommodate the dimensions of tissue images, making developing versatile and scalable systems for efficiently analyzing and visualizing spatial omics data a pressing research need [123, 124, 125]. One solution for such problem is also presented in this thesis [124].

1.6 Multiscale Models of Tissues and Organs

Integrating large and diverse collections of single-cell genomics data has led to the construction of integrated single-cell atlases, which serve as comprehensive and detailed maps of the cell state diversity in specific tissues and organs [126, 127, 128]. As more of such atlases are being built, the next frontier will be the integration of single cell genomics data with spatial omics and imaging data, towards building spatially-resolved integrated atlases and Common Coordinate Framework (CCF) of entire tissues and organs [129].

To achieve this, available single-cell atlases must be integrated with emerging spatial datasets to establish unified Common Coordinate System (CCS) across spatial locations and molecular signatures, subsequently creating comprehensive, spatially-resolved maps of tissues and organs [130,

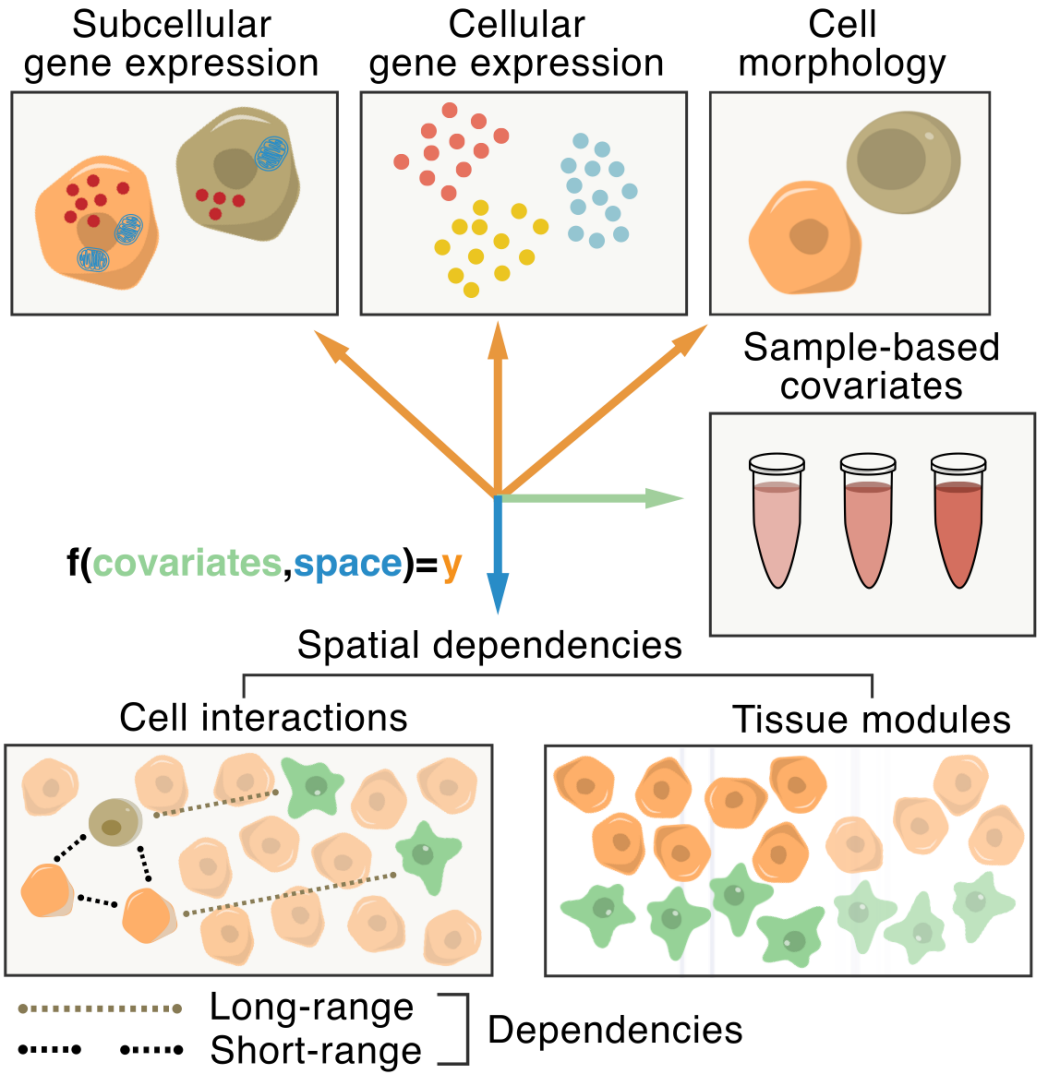


Figure 1.2: **Accounting for spatial dependencies in models of tissue biology**. In the review presented in this thesis [43] we discuss the opportunities and challenges of modeling spatial variation in context with the molecular and sample-level variation. By explicitly accounting for spatial coordinates, proxies, or CCC, we can learn models that disentangle intrinsic and extrinsic factors of cellular variation. We also link modeling concepts in single-cell genomics to the spatial domains and provide an in-depth overview of the current tools and methods proposed for various open problems in spatial biology. This figure has been adapted from [43].

131, 132]. Key components of such problems are significantly related to the computational challenges discussed in *Section.* 1.5, such as the mapping of cell states and genes to spatially-resolved data and the integration of omics modalities with morphology and imaging assays. The Allen Brain Atlas represents a paramount illustration of this integrated framework, encompassing both Mouse and Human Brain [133, 134]. This atlas entails multiscale data across different modalities and biological replicates, which, integrated into a unified perspective, represent a reference representation and offer a spatially and molecularly detailed view of the brain, serving as a foundational resource for brain research. However, there is a crucial conundrum about the sustainability and consistency of this approach, mainly when applied to biological systems where the stereo-typical organization might not be uniformly preserved [68, 135], as opposed to the brain, which retains a precise stereotypical structure across individuals. In fact, the construction of a CCF for other types of biological tissues and organs faces challenges due to the diverse organizational structures across tissue compositions and scales. The tissue's organizational state influences the strategy to create this template [129]. At one end, tissues might have a highly similar structure across samples, as seen in early embryogenesis or in adult brains. Conversely, cells, niches and functional tissue units might be arranged more randomly, a typical example being the tumor microenvironment, which shows the most significant spatial variation across tissue structure. The vast majority of tissues lie somewhere between these extremes. Historically, similar challenges arose in other biological contexts. For instance, in the context of RNA quantification, one method uses a known template to map RNA transcripts when constructing a transcriptome [136]. An alternate method creates a transcriptome without any reference [137, 138]. Furthermore, even within a single organ, the level of structural consistency can vary. For instance, in the gastrointestinal tract, while significant structures like the stomach and intestines are consistent across individuals, finer structures, like those within the small intestine, can vary. This means that a combination of direct measurement as well as data-driven in-silico approaches might be needed even for a single specimen [129].

Multiscale atlases of tissues and organs will be particularly useful not only from a purely descriptive perspective, such as what the stereotypical healthy brain might look like, or what are the stereotypical structures and morphogenic forces shaping a mouse embryo. They will also be useful for directly connecting the spatial organization to the disease state. Excellent examples of this have been the characterization of the TME in breast cancer [46], which has identified specific cellular niches that are associated with disease stage and prognosis. Similar insights have also been gained in lung cancer [139]. Finally, pairing spatial measurements with targeted, in-situ perturbations [140, 141, 142] will enable an even finer-grained characterization of molecular determinants of tissue organization and cell-cell interactions.

1.7 Aims of the Thesis

In this thesis, I have investigated the spatial dependencies of cellular variation and developed novel algorithms that tackle key problems in tissue biology. Specifically, I have worked on the following problems (*Figure.* 1.3):

- **How can spatial patterns within tissues be identified and quantified?** This is a key question that enabled the characterization of tissue organization, useful in many biolog-

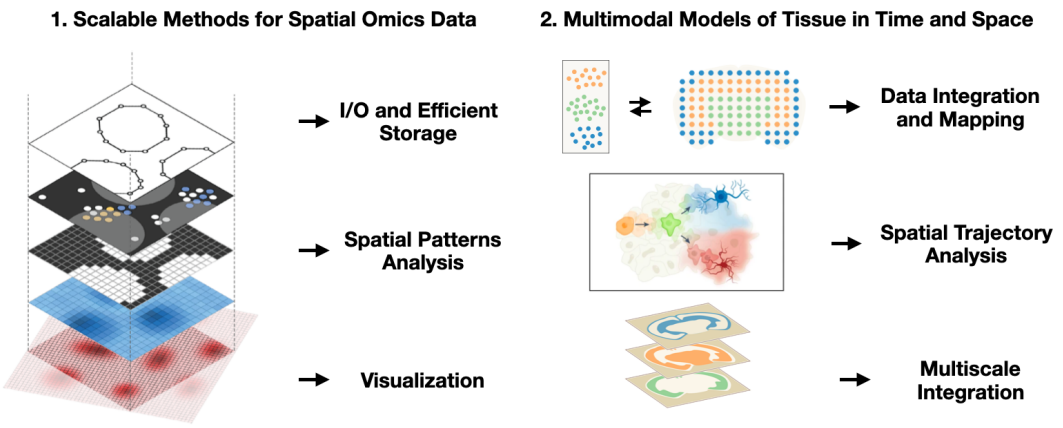


Figure 1.3: **Computational methods for the analysis of spatial omics data.** In this thesis, I developed computational methods for storing, analyzing, and visualizing spatial omics data. I further worked on problems such as data integration between single cell and spatial omics data, developing methods for label transfer, gene expression, and multimodal imputation. I have also proposed a method to learn a spatio-temporal trajectory of developing tissues, which can be used to model tissue dynamics in space and time, and methods for spatial data integration and registration leveraging Optimal Transport. This figure has been adapted from [43, 124]

ical settings such as charting the tumor microenvironment or delineating spatial components of developmental dynamics. In this thesis, I present methods I have developed to address this, by leveraging spatial statistics and graph metrics to discern complex spatial arrangements within tissues. Furthermore, I also present methods that aim to integrate morphological and molecular information, in order to provide a comprehensive view of tissue structures.

- **How can the storage and management of spatial omics data can be optimized?** The diversity in resolution and the sheer scale of spatial omics data pose key challenges in data analysis and visualization of such data. I propose an efficient, scalable, and Findability, Accessibility, Interoperability, and Reusability (FAIR) data structure to address the challenges of handling large spatial omics datasets. Additionally, the creation of efficient visualization tools, both static and interactive, supports the exploration of these datasets, accelerating analysis and insights. Furthermore, I explore how the co-founding and active participation in the open-source software organization scverse underscore the thesis' commitment to collaborative science.

- **How can multimodal and single-cell data be integrated to gain insights into tissue architecture and spatial molecular patterns?** Leveraging the rich abundance of multimodal dissociated data is key to expanding and enhancing the dissection of spatial omics data, mapping annotations, and molecular profiles that have been previously identified to its spatial context. I have developed methods that seek to integrate single-cell genomics with ST data, to enable downstream analysis such as gene imputation, and multimodal imputation of molecular profiles such as Cellular Indexing of Transcriptomes and Epitopes se-

quencing (CITE-seq), and improve cell type annotations. Furthermore, I also investigate how different tissue views can be integrated into spatial coordinates, by solving a registration task that considers both spatial coordinates and gene expression similarity between tissue slices. I have leveraged Optimal Transport (OT) theory and algorithms to develop such methods.

- **How can the temporal dynamics of tissue development be understood through spatial omics?** Time-resolved spatial omics data provide the opportunity to study developmental dynamics across scales, thus improving the characterization of tissue organization in development. In this thesis, I propose learning the spatiotemporal dynamics of tissue development by integrating spatial and gene expression information into novel trajectory inference algorithms. This approach aims to elucidate the developmental pathways of tissues over time. I show how such an approach can be used to investigate the spatiotemporal cellular dynamics in drosophila and mouse embryogenesis. Similar to point 3, the Fused-Gromov-Wasserstein (GW) formulation of OT has been proven useful to address this problem.

This is a cumulative thesis, and I have divided it into the following sections, which correspond to the topics of this thesis and the research papers I have published (either as a journal article, or preprint) on these topics.

In the first part of this thesis, I describe the challenges and opportunities of computational models for spatial omics data. I describe the main biological problems in the field and computational approaches that could be developed to tackle them, as well as review the state-of-the-art in spatial omics data analysis and modeling. Furthermore, I describe the development of squidpy, a software for spatial omics data analysis that I developed, together with colleagues. squidpy represented the first multi-purpose analysis toolkit for spatial omics data in Python, and it is now widely used in the community. I further describe the development of a novel data format and specifications for the storage and analysis of spatial omics data, called SpatialData. I describe my contribution to developing a scalable and cloud-ready imaging format called Open Microscopy Environment (OME)-Zarr. I describe the work I did in open-source software for the biomedicine community and how I have, together with colleagues, founded a new open-source software organization called scverse. I also outline my contribution to a work that summarizes best practices in single-cell and spatial data analysis pipelines.

- **Publication 1 (co-lead author)** ***Spatial components of molecular tissue biology*** This paper reviews the advancements in spatial profiling methods used for examining RNA and protein expressions in cells. We emphasize the importance of framing straightforward biological questions and developing computational tools for practical spatial tissue analysis. We also highlight existing challenges in data interpretation and underscore the necessity for developing analytical tools and models tailored to these experimental techniques.

- **Publication 2 (co-lead author)** ***Squidpy: a scalable framework for spatial omics analysis*** In this paper, we develop Spatial Quantification of Molecular Data in Python (squidpy), a Python framework tailored to handle, analyze, and visualize spatial omics data efficiently.

Our system seamlessly integrates with existing Python libraries of the single-cell analysis ecosystem, ensuring both memory efficiency and computational speed. With thorough documentation and robust testing, we envision squidpy spearheading the evolution of computational methods in spatial omics.

- **Publication 3 (co-lead author)** ***SpatialData: an open and universal data framework for spatial omics*** In this paper, we develop SpatialData, a unified framework to manage large and diverse spatial omics datasets, facilitating spatial annotations and cross-modal analysis. Our system streamlines the creation of common coordinate systems, enabling more robust comparisons across studies and enhancing reproducibility. Future enhancements include extending compatibility to R/Bioconductor, supporting multiscale structures, and cloud-based data access.

- **Additional publication 1 (co-author)** ***OME-Zarr: a cloud-optimized bioimaging file format with international community support*** In this paper, we develop a next-generation file format built on top of OME-Zarr for bioimaging, addressing scalability and heterogeneity challenges in the field. Our approach emphasizes community cooperation to streamline bioimaging data management and analysis by consolidating storage file formats like HDF5, TIFF, and Zarr. With the growth of OME-Zarr tools and resources, we urge users and developers to adopt this format, strengthening the bioimaging ecosystem and promoting the FAIR principles.

- **Additional publication 2 (co-lead author)** ***The scverse project provides a computational ecosystem for single-cell omics data analysis*** In this paper, we introduce "scverse", an open-source project addressing the challenges posed by the rapid growth of single-cell omics tools, aiming to streamline the storage, analysis, and interoperability of single-cell data. We present core data structures, Annotated Data (AnnData) and Multimodal Annotated Data (MuData), designed for unimodal and multimodal data, ensuring interoperability across various programming languages. Additionally, scverse incorporates fundamental toolkits for single-cell genomics tasks, emphasizing the importance of standardized tool development, testing, and community collaboration.

- **Additional publication 3 (co-author)** ***Best practices for single-cell analysis across modalities*** In this paper, we introduce comprehensive best-practice workflows for unimodal and multimodal single-cell analysis based on independent benchmarks, addressing the wide range of tools in the field. We emphasize the importance of large-scale single-cell data sets, highlighting the emerging significance of models considering cellular heterogeneity and predicting a shift from unsupervised analysis towards supervised predictions. Lastly, we discuss the potential of single-cell multi-omics in clinical settings and the need for patient data integration tools. These workflows will be built upon our foundational multimodal single-cell analysis overview.

In the second part of this thesis, I describe how the Optimal Transport (OT) framework can be used to model a variety of computational challenges in spatial omics data analysis. I present moscot, a tool developed together with colleagues that aims at providing an extensive and scal-

able interface to efficient OT algorithms for temporal and spatial omics problems. I also present a preprint on uncertainty quantification for atlas level cell-type transfer, where we study how uncertainty quantification methods can be used to identify novel cell types in query-to-reference integration methods, and how they are better calibrated.

- **Preprint 1 (co-lead author)** ***Mapping cells through time and space with moscot*** We develop multi-omics single-cell optimal transport (moscot) a novel and scalable framework for integrating multimodal single-cell genomics data using optimal transport. We demonstrate its efficacy in reconstructing cellular developmental trajectories, mapping ST, and analyzing the spatiotemporal dynamics of mouse embryogenesis. By incorporating advanced features such as GPU acceleration, linear time and memory complexity, and support for diverse modalities, moscot stands out as a powerful tool to interpret complex biological datasets and provides the community with an efficient Python package.
- **Additional publication 4 (co-author)** ***Unbalanced Low-rank Optimal Transport Solvers*** In this paper, we develop advanced OT methods that merge the benefits of both scalable low-rank solvers and flexible unbalanced variants. We introduce custom algorithms for the linear OT problem and its Fused-Gromov-Wasserstein (FGW) generalization, addressing computational and modeling challenges. Our approach offers improved scalability and flexibility, notably demonstrated in ST matching problems.
- **Preprint 3 (co-lead author)** ***Modeling single-cell dynamics using unbalanced parameterized monge maps*** In this paper, we develop the Unbalanced Parameterised Monge Maps (UPMM) for modeling single-cell dynamics. Our method enhances trajectory inference by accounting for asymmetric shifts in cell state distributions, proving its efficacy on real-world and synthetic data. Additionally, we highlight UPMM's potential for atlas building, multi-modality extensions, and its complementary role with RNA velocity.
- **Preprint 4 (co-lead author)** ***Uncertainty Quantification for Atlas-Level Cell Type Transfer*** This paper studies the importance of uncertainty quantification for cell type classification in single-cell reference atlases. Our findings show that current models lack robustness and calibration, but those that quantify uncertainty better identify previously unseen cell types. We emphasize the need for better models to ensure accurate predictions and inform future research on understanding cellular diversity using scRNA-seq data.

2 Methodology

This section is divided into experimental and computational methods. The experimental techniques describe the biological and molecular principles for the data presented in this thesis. The computational methods section provides an overview of the mathematical and algorithmic background of the computational methods presented in this thesis.

2.1 Experimental Methods

The evolution of single-cell omics has led to the rise of spatially resolved transcriptomics, encompassing techniques and computational strategies to delineate cell locations and their gene expression patterns spatially. The technology has been elected "Method of The Year 2020" by Nature Methods [27]. Here, I provide an overview of these methodologies and their genomic research applications.

2.1.1 Array-based Spatial Transcriptomics

The primary strength of sequencing-based technologies is their non-biased mRNA capture ability. This allows for a comprehensive molecular characterization of tissues, bypassing traditional morphological assessments. A notable example is the Spatial Transcriptomics (ST) technology, which employs spatially barcoded probes to profile tissues. Subsequent developments like Visium expanded ST's capabilities, striving to achieve single-cell resolution. Efforts such as Slide-seq [143], HDST [144], and Stereo-seq [50] have been directed toward enhancing spatial resolution using densely packed bead arrays. Despite challenges like the need for individual barcode decoding, these technologies signal a promising trajectory for spatially resolved transcriptomics. A vital drawback of these technologies is that there is no straightforward way to link the molecular profile to its cell entity. While ST [145] and Visium have capture areas spanning 200 to 50 um, deconvolution methods are usually required to link transcriptional signatures to cellular heterogeneity. In contrast, Slide-seq, HDST, and Stereo-seq provide higher resolution mRNA capture areas. Yet, some level of aggregation across beads is often required at the computational stage for decreasing signal-to-noise and for practical data analysis.

Array-based ST technologies have contributed to elucidating tissue architecture and function in several studies. For instance, studies on the human heart utilized ST to generate developmental transcriptional atlases [53], underlining the importance of spatial context in cellular subclassifications during development phases. Similarly, endeavors in neuroscience have harnessed ST to create molecular 3D atlases of mouse brains [146], highlighting potential applications in the human

cellular atlas initiative146. Spatial contexts deeply influence diseases. Studies using array-based ST have revealed transcriptional changes across disease progression stages [147], illustrating how transcriptomic data can elucidate disease mechanisms. Especially in oncology, the spatial environment plays a pivotal role, with spatially resolved transcriptomics highlighting tumor heterogeneities and revealing potential therapeutic targets. Finally, developmental studies have also benefited from using array-based ST technologies. For example, a recent study reported the first comprehensive spatial characterization of mouse embryogenesis, highlighting the highly structured and spatially organized features of developmental processes in mammal embryos.

The future likely holds the convergence of sequencing technologies, leading to multimodal spatial profiling [94, 95]. With advances in barcoding schemes, a simultaneous assessment of various biomolecules from the same tissue section is already possible, yet not widely commercialized. This spatial multi-omics approach and computational tools will present a comprehensive view of biological mechanisms and interactions within tissue ecosystems.

2.1.2 Imaging-based Spatial Transcriptomics

Advancements in genome-scale imaging have facilitated the spatially resolved single-cell omics analyses within intact cells and tissues. This technological progression permits detailed gene expression profiling at the individual cell level. Such high-resolution techniques elucidate the spatial arrangements of the genome and the transcriptome inside cells.

For Imaging-based ST, two primary methodologies have been delineated: in situ sequencing (ISS) [148] and multiplexed FISH [29, 30]. Each approach possesses unique challenges and common hurdles in attaining genome-scale imaging.

Historically, FISH was employed for imaging genomic loci and RNA transcripts. Using combinatorial color schemes, single-molecule FISH facilitated single-cell gene expression profiling, such as simultaneous imaging of multiple genes. Nevertheless, the finite number of distinct color channels rendered genome-scale imaging challenging. The method of multiplexed error-robust FISH (MERFISH) was conceived to address this [30]. Concurrently, an alternative multiplexed FISH method, seqFISH, was established [29]. This method differentiates RNAs via a sequential color code. Modern seqFISH techniques have incorporated error correction coding schemes and signal amplification strategies, enhancing imaging capabilities. On the other hand, techniques like ISS permit targeted gene analyses, utilizing unique nucleotide sequences as barcodes. Recent innovations have augmented the efficiency and accuracy of in situ sequencing, such as STARmap [148].

Spatially resolved single-cell transcriptomic analyses have facilitated the direct identification and spatial mapping of cell types in intact tissues. Such analyses have unveiled potential cellular interactions and the interplay of different cellular types, with notable examples in the brain. Combining techniques like STARmap and MERFISH with neuronal activity imaging has augmented our understanding of how different cell types react to various stimuli. Furthermore, these analyses have offered valuable insights into conditions like neurodegenerative diseases [147, 149].

2.1.3 Spatial Proteomics

To comprehend tissue functionality and pathology, particularly in the context of cancer diagnosis and treatment, tools have emerged that can retain spatial information while quantifying multiple protein expressions. Although effective to a certain extent, traditional techniques have limitations in in-depth, resolution, or simultaneous protein tracking. In recent years, advancements in Mass Spectrometry (MS) and multiplexed Immuno-Histo-Chemistry (mIHC) have propelled the development of novel techniques for high-throughput measurements of protein abundance in tissue.

There are various techniques in this domain, yet they all fundamentally rely on detecting antibodies bound to proteins via MS or Fluorescent Imaging. Some of the most notable ones are multiplexed ion beam imaging by time of flight (MIBI-TOF) [37], Imaging Mass Spectrometry (IMC) [38], and Cyclic Immunofluorescence (CyCIF) [31]. MIBI-TOF, is based on time-of-flight MS, facilitating subcellular imaging in clinical tissue sections. MIBI-TOF demonstrates improved throughput in imaging formalin-fixed paraffin-embedded (FFPE) human tissue biopsies and with repeat scanning, this technique offers multiplexed three-dimensional imaging. IMC is MS-based imaging that employs defined mass or metal isotope reporters. Like MIBI-TOF, it permits the concurrent detection of up to 40 antigens and nucleic acid sequences in various mediums. It can also integrate multiplex detection with 3D tissue visualization, capturing individual cells and larger entities like blood vessels. CyCIF is a methodology that consists of a fluorophore-bound compound that targets specific proteins, which, when imaged, illuminate their cellular positions. Subsequent fluorophore deactivation and reapplication allow sequential protein mapping. In contrast to the MS base technology, it can more easily reach a higher number of total proteins measured, up to almost one hundred, particularly for variations of mIHC such as 4i [150]. The large body of application of these techniques has been in cancer. Since they can be used to study biopsies and to obtain images that cover a large area of tumors, they provide a new light on the organization of the Tumor Micro-Environment (TME). Preserving the single-cell resolution capabilities was instrumental in characterizing the TME in Breast Cancer [46, 47], and Brain Cancers such as Glioblastoma [151].

2.2 Computational Methods

Spatial omics methodologies provide tools for researchers to probe the spatial organization and function of cells within tissues. In this section, I will introduce key concepts for the methods developed in this thesis. In particular, in *Section.*2.2.1, I introduce critical spatial statistics essential for comprehending cellular and molecular arrangement ts in tissues. In *Section.*2.2.2 I introduce key components of the storage and visualization tools for spatial omics data developed in this thesis. Finally, in *Section.*2.2.3 I introduce notions on Optimal Transport (OT) and describe how they are used in the methods I develop for spatial omics data integration and mapping.

2.2.1 Spatial Statistics

The neighbor enrichment score assesses the association between label pairs in cellular spatial graphs [152]. This association is quantified by counting nodes belonging to each cluster present in the data, labeled as x_{ij}. To determine how this count deviates from a random label configuration while keeping connectivity intact, we shuffle the cluster labels and tally the node numbers, performing this operation over a number of iterations (1000 by default). Using this data, we calculate expected means (μ_{ij}) and standard deviations (σ_{ij}) for each pair, and a Z-score as,

$$Z_{ij} = (x_{ij} - \mu_{ij})/\sigma_{ij} \tag{2.1}$$

This Z-score indicates if there is an over representation or deficiency of cluster pair interactions in the connectivity graph.

Ripley spatial statistics [153] offer a suite of methods to determine if points with specific spatial annotations exhibit random, dispersed, or clustered distributions. Ripley's statistics can be used to describe the spatial patterning of cell clusters in the area of interest.

In squidpy [42], I have implemented 3 Ripley's statistics: F, G and L functions.

Ripley's K function is defined as:

$$K(t) = A \sum_{i=1}^{n} \sum_{j=1}^{n} I(d_{i,j} < t) \tag{2.2}$$

Where $I(d_{i,j} < t)$ is the indicator function, setting the operand to 1 or 0 considering distance $d_{i,j}$ computed at search radius t, whereas A is the average density of point in the area of interest.

The Ripley's L function is a variance-stabilized transformation of the Ripley's K function, defined as:

$$L(t) = \left(\frac{K(t)}{\pi} \right)^{1/2} \tag{2.3}$$

Finally, the Ripley's F and G function are defined as:

$$P(d_{i,j} \leq t) \tag{2.4}$$

Where $d_{i,j}$ indicates the distance from a point to other reference points in space.

Another essential spatial metric in squidpy is the cluster co-occurrence score. It quantifies the co-occurrence of certain cluster groups across spatial domains and is expressed as:

$$\frac{p(\text{exp}|\text{cluster})}{p(\text{exp})} \tag{2.5}$$

where $cluster$ denotes a specific annotation used as a reference for the co-occurrence of all clusters across varying radii in the tissue area. It was inspired by an analysis performed by [154] to investigate tissue organization in the human pancreas with ST.

Finally, spatial autocorrelation statistics are fundamental for analyzing spatial data, assisting in identifying patterns of gene expression as clustered, dispersed, or random. In squidpy, two such statistics are implemented: Moran's I and Geary's C [155]. Moran's I is expressed as:

$$I = \frac{n}{W} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{2.6}$$

whereas Geary's C is defined as:

$$C = \frac{(n-1) \sum_{i,j} w_{i,j} (x_i - x_j)^2}{2W \sum_i (x_i - \bar{x})^2} \tag{2.7}$$

where z_i represents the feature's deviation from the average $(x_i - \overline{X})$, $w_{i,j}$ signifies the spatial weight between observations, n is the total number of spatial units and W is the sum of all $w_{i,j}$. Test statistics and p values are computed from a permutation-based test where edges of the adjacency matrix of the spatial graph are randomly shuffled across iterations.

Though these tools are exploratory in nature, they remain invaluable for generating hypotheses and encapsulating the spatial information within the data.

2.2.2 Storage and Infrastructure for Spatial Omics Data

Analyzing spatial omics data poses unique challenges due to the complex nature and size of the data. This data encompasses tabular data, raster images represented as multidimensional arrays, polygonal data denoting regions of interest, and point clouds, which are large sets of dimensionless points reflecting molecular measurements like detected mRNA transcripts. Furthermore, all these data may or may not correspond to the same spatial coordinate system as cells or tissue regions.

This diversity in data types complicates storage and in-memory representation, particularly when the volume exceeds available memory, necessitating out-of-core or out-of-memory strategies. Here, I detail the creation of a novel data format designed to address spatial omics data storage and manipulation comprehensively. This solution, named SpatialData offers various innovations for the handling of spatial omics data: It abstracts common data types and measurements linked with spatial omics experiments. This abstraction facilitates standardized serialization and efficient, scalable in-memory representation by harnessing cutting-edge methods in Python's scientific computing ecosystem. These data abstractions encompass five elements:

- Raster Images (e.g., microscopy images)
- Raster Labels (e.g., segmentation masks)

- Shapes (polygons symbolizing regions of interest or cells)
- Points (e.g., transcript locations or other tissue annotations)
- Tables (storing molecular profiles, annotations, and related graphs).

For storage, SpatialData employs zarr [156] and parquet [157] formats. In memory, the data is xarray [158, 159] objects for Images and Labels, Dask Dataframes for Points [160], Geopandas Dataframes for Shapes [161], and AnnData for Tables [162].

A standout feature of this approach is its high performance concerning data storage and retrieval, even in cloud environments due to the use of cloud-ready formats (Zarr [156, 163] and Parquet [157]). This is complemented by efficient data manipulation tools, allowing analysts to achieve a multi-scale data view of the data, and to aggregate or extract specific data subsets of interest for their analysis. All these processes are designed for optimal performance, considering both memory utilization and computational efficiency.Finally, a comprehensive suite of visualization tools, both interactive, based on napari [164] and matplotlib [165], which enable users to make publication-ready figures as well as complex visualization, with the addition of interactivity.

The research discussed in this thesis extends beyond the realm of spatial omics. Notably, contributions to the "scverse" platform have broader implications for biomedical data analysis and visualization using Python [166]. Such contributions encompass software developers' typical tasks, including code review and engaging with users on platforms like GitHub. Additionally, I have aided in creating a Project Template for open-source software (OSS) projects that utilize scverse tools. This template offers a pre-configured toolkit for documentation construction, automated testing, and continuous integration. I have also contributed to other projects in scverse such as Scanpy [87] and AnnData [162], as well as many other projects that can be seen on GitHub. It's also important to note that not all data in this thesis are spatial omics data.

2.2.3 Optimal Transport

Optimal Transport (OT) addresses the problem of identifying the best transportation plan P between two sets of points a and b, using a cost matrix C that indicates the distances between the source points a and the target points b. In the discrete scenario, when aiming to match sets $x \in \mathbb{R}^D$ to $y \in \mathbb{R}^D$, the goal is to ascertain a bipartite graph $T : X \mapsto Y$. This corresponds to determining T such that:

$$c(T) := \sum_{x \in X} C(x, T(x)) \tag{2.8}$$

Here, C is known as the cost matrix, representing the cost of transporting a source point x_i to target point y_j. The function $c(T)$ indicates the total cost of applying $T(x)$, and it follows that the optimal transportation plan is unique $P \in \Pi$ for which that cost is minimized [167]. While this formulation has several practical applications, it also poses several limitations, such as the requirement that the source and target points x and y have the same length n, as well as practical implementation challenges with a time complexity of $O(n^3)$ [168]. Modern applications of discrete OT

in Machine Learning have shifted towards formulations such as the Kantorovich relaxation [167, 168]. This approach alleviates the constraint of equating point numbers in both source and target and permits source points to be split and subsequently assigned to target points. Conceptualizing the source and target points as distributions μ and ν linked to their metric spaces X and Y aids in comprehension [168]. Thus, given probability measures μ on X and ν on Y, the OT problem centers on finding a transportation map $T : X \to Y$ minimizing:

$$\inf\{\int_{X\times Y} c(x,y)d\pi(x,y)|\pi \in \Pi(\mu,\nu)\} \tag{2.9}$$

In the discrete case, μ_x the probability mass associated with $x \in X$ and ν_x the probability mass associates with $y \in Y$, the Kantorovich problem is stated as:

$$\min_{\pi\geq 0} \sum_{x\in X}\sum_{y\in Y} c_{xy}\pi_{xy} \tag{2.10}$$

subject to mass conservation constraints $P \in U(x,y)$:

$$\mu = \sum_{y\in Y} \pi_{xy} \quad \nu = \sum_{x\in X} \pi_{xy} \tag{2.11}$$

This also corresponds to minimizing the Wasserstein (W) distance between the two distributions [168].

Recent ML research has seen the Kantorovich OT formulation's significant rise in usage. This is primarily due to the introduction of entropic regularization [169], which estimates the regularized OT map using the Sinkhorn algorithm. However, it is helpful first to make the observation that the OT problem introduced in 2.9 also provides a dual formulation [167, 168], that is the following:

$$\mathcal{L}_c(\alpha,\beta) = \sup\{\int_{\mathcal{X}} \varphi \mathrm{d}\mu + \int_{\mathcal{Y}} \psi \mathrm{d}\nu|(\varphi,\psi) \in \mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y}), \varphi(x)+\psi(y) \leq c(x,y)\} \tag{2.12}$$

Here, φ and ψ are often called "Kantorovich potential". In the discrete case, the Kantorovich problem, similarly to 2.10, is stated as:

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \max_{(\mathbf{f},\mathbf{g})\in\mathbf{R}(\mathbf{C})} \langle \mathbf{f},\mathbf{a}\rangle + \langle \mathbf{g},\mathbf{b}\rangle \tag{2.13}$$

where the set of admissible dual variables is:

$$\mathcal{R}(\mathbf{C}) := \{(\mathbf{f},\mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m | \mathbf{f}\oplus\mathbf{g} \leq \mathbf{C}\} \tag{2.14}$$

The Sinkhorn algorithm leverages the dual formulation of OT 2.13 and introduces the entropy regularization of the coupling matrix P.

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P},\mathbf{C}\rangle - \varepsilon\mathbf{H}(\mathbf{P}) \tag{2.15}$$

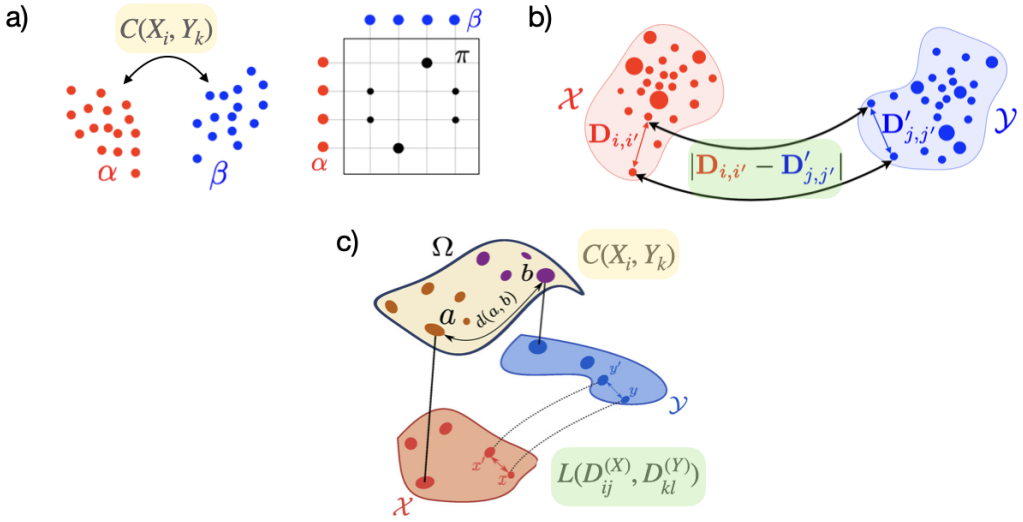


Figure 2.1: **W, GW and FGW formulations of OT.** Schematics of the a) linear (W), b) quadratic (GW) and c) the fused linear and quadratic (FGW) formulation of OT. Figure has been adapted from [168, 170].

If we consider the matrix $\mathbf{K} : X \times Y$ so that $\mathbf{K}_{xy} = e^{(c_{xy})}$, which is also referred to as the Gibbs kernel of the cost matrix $\mathbf{C}_{xy}$ [168], finding the optimal $\mathbf{P}*$ is equivalent to find the two diagonal positive matrices $\mathbf{D_1} = \text{diag}(\mathbf{u})$ and $\mathbf{D_2} = \text{diag}(\mathbf{v})$ with the same size as X and Y so that the marginals contraints a, b (mass conservation) are satisfied: $\mathbf{a} = \text{diag}(\mathbf{u})\mathbf{K}\text{diag}(\mathbf{v})\mathbf{1}_y$ and $\mathbf{b} = \text{diag}(\mathbf{v})\mathbf{K}^T\text{diag}(\mathbf{u})\mathbf{1}_x$.

The Sinkhorn algorithm is the following:

Require: Cost matrix $C \in \mathbb{R}^{n \times m}$
Require: Marginal vectors $a \in \mathbb{R}^n_+$ and $b \in \mathbb{R}^m_+$ with $\sum_i a_i = 1$ and $\sum_j b_j = 1$
Require: Regularization parameter $\epsilon > 0$
1: Initialize $u = \mathbf{1}_n, v = \mathbf{1}_m$
2: Compute the kernel matrix $K = \exp(-C/\epsilon)$
3: **while** not converged **do**
4: $\quad u \leftarrow \frac{a}{Kv}$ ▷ Element-wise division
5: $\quad v \leftarrow \frac{b}{K^T u}$
6: **end while**
7: $P = \text{diag}(u)K\text{diag}(v)$ ▷ Regularized optimal transport plan
8: **return** P

The Sinkhorn algorithm is at the foundation of ML applications of OT, and it essentially represents the foundation also for the the other formulation of OT, Gromov-Wasserstein (GW).

GW OT (*Figure.* 2.1), seeks the optimal match between two distributions that inhabit different metric spaces, making direct cost computation of the optimal transportation plan unfeasible. This method aims at aligning probability measures based on structural similarities within their spaces. The GW distance was originally proposed by [171, 172], and it was expanded by [173], incorporating the calculation of barycenters and a scalable formulation boosted by the entropic regularization.

Given two distance matrices $\mathbf{D}^Y \in \mathbb{R}_+^{M\times M}$ and $\mathbf{D}^X \in \mathbb{R}_+^{N\times N}$, the discrete OT solution is the following:

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{ijkl} L\big(\mathbf{D}^X_{ij}, \mathbf{D}^Y_{kl}\big)\mathbf{P}_{ik}\mathbf{P}_{jl} \tag{2.16}$$

Typical choice of loss $L(a, b)$ is the quadratic loss (or squared euclidean loss) $\frac{1}{2}|a - b|^2$ or the Kullback-Leibler divergence $KL(a|b) = a\log(a/b) - a + b$. This formulation of OT has applications in areas like single-cell and spatial genomics, for tasks such as modality alignment. Here, datasets profiled via diverse technologies (like scATAC-seq and scRNA-seq) need to be matched so that each cell's identity can be identified across modalities.

Finally, in situations where both shared and incomparable spaces exist between the two distributions, the Fused-Gromov-Wasserstein (FGW) (*Figure.* 2.1) is employed.

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \alpha\langle\mathbf{P}, \mathbf{C}\rangle + (1-\alpha)\sum_{ijkl} L\big(\mathbf{D}^X_{ij}, \mathbf{D}^Y_{kl}\big)\mathbf{P}_{ik}\mathbf{P}_{jl} \tag{2.17}$$

where the weight parameter α is introduced to control for the relative contribution of the linear (W) and quadratic (GW) term. In practical terms, the objective function now contains the following cost matrices (with references to the single cell genomics application):

- $C \in \mathbb{R}_+^{N\times M}$: compares reference cells with spatial samples in terms of expression in shared genes.
- $C^X \in \mathbb{R}_+^{N\times N}$: compares reference cells among each other in terms of gene expression.
- $C^Y \in \mathbb{R}_+^{M\times M}$: compared spatial samples among each other in terms of spatial distance.

This is what for example is employed in the method I have implemented of the multi-omics single-cell optimal transport (moscot) mapping problem (*Figure.* 2.2), where in that case the cost $C \in \mathbb{R}_+^{N\times M}$ is the distance between samples in the shared gene space, $C^X \in \mathbb{R}_+^{N\times N}$ is the cost for the unique genes measured in the single cell datasets, and are not present in the spatial dataset, and $C^Y \in \mathbb{R}_+^{M\times M}$ is instead the distances between cells in the spatial datasets based on their spatial coordinates. Analogously, in the moscot alignment and spatiotemporal problems that I have implemented, the cost $C \in \mathbb{R}_+^{N\times M}$ is the cost for the set of genes measured between spatial omics datasets that should be aligned in space, or for which a spatiotemporal mapping is learned for further trajectory analysis, similary to the moscot mapping problem. In contrast, $C^X \in \mathbb{R}_+^{N\times N}$

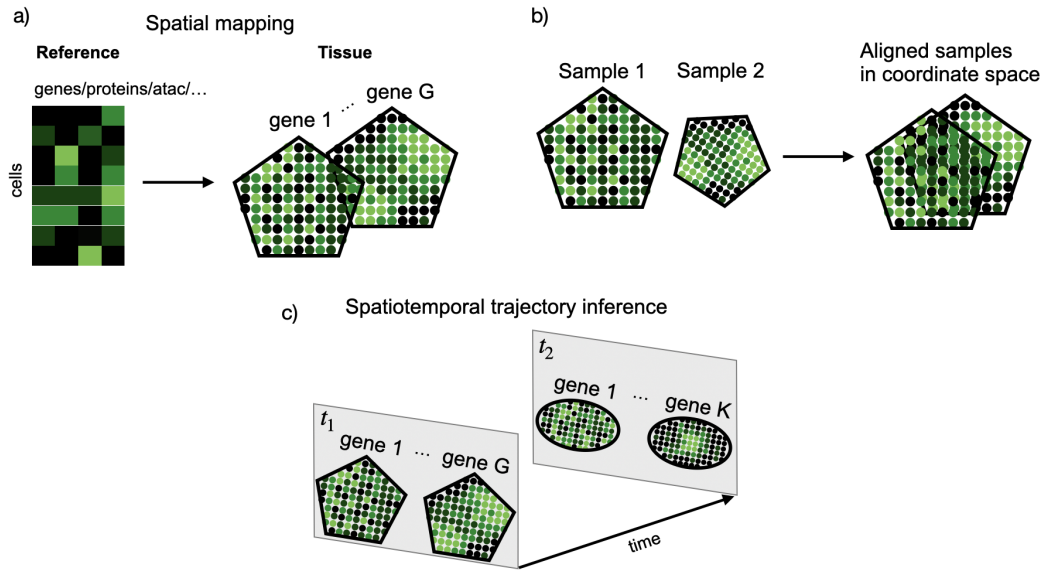


Figure 2.2: **Applications of OT to ST data in moscot** a) Spatial Mapping, where single cell data is mapped to spatial data, b) Alignment of ST slides and c) spatiotemporal trajectory inference.

and $C^Y \in \mathbb{R}_+^{M \times M}$ are the costs for the spatial coordinates, respectively for the source and target dataset of the alignment and spatiotemporal mapping problem.

moscot (*Figure.* 2.2) is also able to leverage more modern formulations such as the low-rank approximations of the W and GW problems introduced in [174, 175, 176, 177].

3 Publication Summaries

This section contains a one-page summary of this thesis's five leading publications and preprints.

3.1 Spatial components of molecular tissue biology

This review paper was published in Nature Biotechnology in March 2022. The full citation is:

Giovanni Palla*, David S Fischer*, Aviv Regev, Fabian J Theis
Spatial components of molecular tissue biology
Nature Biotechnology 2022
https://doi.org/10.1038/s41587-021-01182-1

Contribution

I conceived the project with all my co-authors. I have reviewed all the papers cited in the review and designed the tables and boxes. I have designed the sketch of the figures and helped design them together with the illustrator acknowledged in the review. I have written the manuscript with all my co-authors.

Summary

Our review aimed to elucidate how computational methods can uncover how tissue organization (structure) connects to its function. Single-cell profiling data describe cellular phenotypes through variables like gene expression, protein levels, and chromatin accessibility, which depend on factors like sample type, assay methodology, and sequencing batches. In contrast, spatial cellular models incorporate extra variables, like molecular spatial distribution within cells and cell morphology.

We identified two primary influences on cell states in spatial contexts: direct cell-to-cell communication and tissue context. These dependencies manifest as high or low-frequency patterns across tissues, with high-frequency ones being local, often related to cell-to-cell communication, and low-frequency ones more related to global processes, like morphogenic or metabolic gradients. For instance, in a pathological context like tumors, low-frequency signals may relate to metabolic gradients of oxygen concentration on the tissue. In contrast, high-frequency ones can relate to cytotoxic T-cell activity.

Furthermore, spatial assays probe the effects of short-length scale cell-to-cell interactions, including physical contacts and signaling. Although specific cell-to-cell interactions have been captured using single-cell profiling techniques, many can only be observed and validated via spatial protocols. The spatial neighborhood context offers robust evidence of coexisting molecular and structural phenotypes, paving the way for causal inference frameworks. Such models can decode effects like a T cell's influence on a tumor cell or how a tumor's metabolic environment modifies immune cell protein expression.

On a larger scale, tissue biology is characterized by super-cellular-level spatial variations. We need broader-view spatial technologies and computational methods to decipher tissue's emergent properties and their dysregulation in disease. Tissues consist of recurrent cellular units or "tissue mod-

ules" with specific functions, which can vastly differ in size and function. Determining these modules necessitates both molecular and spatial data. Furthermore, biophysical tissue attributes, influential in systems from development to cancer biology, can be discerned with spatial molecular techniques combined with morphology. By merging these observations with spatial omics, we can potentially identify the molecular causes of mechanobiology.

Finally, we provide an in-depth overview of the current computational challenges and proposed solutions across various problems in handling and modeling spatial omics data.

3.2 Squidpy: a scalable framework for spatial omics analysis

This research paper was published in Nature Methods in February 2022. The full citation is:

Giovanni Palla*, Hannah Spitzer*, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, Fabian J Theis
Squidpy: a scalable framework for spatial omics analysis
Nature Methods 2022
https://doi.org/10.1038/s41592-021-01358-2

Contribution

I conceived the project with Hannah Spitzer and Fabian Theis, I led the software development with Hannah Spitzer, and Michal Klein, with help from all the coauthors. My main focus has been on the graph building and spatial statistics methods, as well as the draft of the napari implementation. I have performed the analysis in the paper with Hannah Spitzer and the rest of the co-authors. The first iteration of the code was developed with the co-authors at a Hackathon in 2020. I wrote the manuscript with Hannah Spitzer and Fabian Theis, and with assistance from all other authors.

Additional supplementary material is available at the publisher's website https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01358-2/MediaObjects/41592_2021_1358_MOESM1_ESM.pdf, All code published in the context of this project can be found on GitHub

- https://github.com/scverse/squidpy
- https://github.com/scverse/squidpy_notebooks
- https://github.com/theislab/squidpy_reproducibility

Summary

Spatial omics technologies enable the study of tissues in situ at both cellular and subcellular levels. They produce vast and varied data, presenting challenges in representation, analysis, and visualization. Most existing frameworks for handling this spatial data focus either on data preprocessing or specific aspects of spatial data analysis. Combining different analysis steps is challenging due to the absence of a consistent data representation and a unified application programming interface.

To address this gap, we introduce Spatial Quantification of Molecular Data in Python (squidpy), a Python-based tool designed to analyze spatially resolved omics data. Squidpy standardizes spatial data into common representations, offering comprehensive analysis and interactive visualization tools. It manages spatial omics data through two primary representations: a neighborhood graph

from spatial coordinates and high-resolution microscopy images of tissues. Both representations are designed to handle two-dimensional and three-dimensional data.

Squidpy also offers tools to decode spatial patterns in tissues, empowering researchers to understand spatial organization and cellular relationships. This capability is highlighted using various datasets and technologies, demonstrating Squidpy's ability to provide interpretable results across diverse experimental methodologies.

Furthermore, Squidpy incorporates functionalities for analyzing and visualizing high-resolution images from spatial omics technologies. These images offer additional information about tissue structure and cellular variations. Squidpy introduces a data object, the ImageContainer, to efficiently manage these images. In addition, Squidpy supports image analysis techniques, enabling the extraction of various features from tissue images.

Squidpy provides a comprehensive, efficient, scalable platform for analyzing spatial molecular data. By interfacing with other popular Python libraries and tools, it offers a robust framework for developing future spatial molecular data analysis methodologies. The extensive documentation and tutorials make it accessible to a broad audience, from novice analysts to seasoned developers. We have anticipated future enhancements, such as further integration with emerging spatial omics tools developed by the community, positioning Squidpy as a pivotal tool to advance spatial omics technologies and bridge the gap between omics and image analysis communities.

Since publication, Squidpy has grown dramatically, reaching approximately 60k downloads on PyPI and more than 330 stars on Github, at the time of writing this document.

3.3 SpatialData: an open and universal data framework for spatial omics

This research paper was published in Nature Methods in March 2024. The full citation is:

Luca Marconato*, Giovanni Palla*, Kevin A Yamauchi*, Isaac Virshup*, Elyas Heidari, Tim Treis, Marcella Toth, Rahul Shrestha, Harald Vöhringer, Wolfgang Huber, Moritz Gerstung, Josh Moore, Fabian J Theis, Oliver Stegle
SpatialData: an open and universal data framework for spatial omics
Nature Methods 2024
https://www.nature.com/articles/s41592-024-02212-x

Contribution

I conceived the project with Luca Marconato, Kevin Yamauchi, and Isaac Virshup, I have led the software development with Luca Marconato, Isaac Virshup and Kevin Yamauchi, as well as help from all the co-authors. Particularly, I have led the development of the storage layout, IO and the design of the data model. I have also contributed to the data loader, as well as the first napari implementation, the IO library and key contributions to the spatialdata plotting library. I wrote the manuscript with Luca Marconato, Kevin Yamuchi, Josh Moore, Oliver Stegle, and Fabian Theis, and with assistance from all other authors.

Additional supplementary material is available at the publisher's website https://www.nature.com/articles/s41592-024-02212-x, All code published in the context of this project can be found on GitHub

- https://github.com/scverse/spatialdata
- https://github.com/scverse/spatialdata-notebooks
- https://github.com/giovp/spatialdata-sandbox

Summary

Our study recognized that a significant challenge lies in integrating data from diverse spatial omics techniques due to the varying data types, file formats, and spatial resolutions. To address this, we developed the SpatialData framework. This tool ensures the FAIR integration of multimodal spatial omics data. SpatialData uses a language-independent storage format that enhances the interoperability of data sources. This framework categorizes spatial datasets into five main elements: Images, Labels, Points, Shapes, and Tables. By storing multiple datasets with a record of their spatial relationships, we can analyze them collectively.

We implemented this framework using a Python library, allowing easy access and manipulation of SpatialData objects. One of the key features introduced in SpatialData is to define Common Coordinate System (CCS) for imaging and spatial coordinates. This enables the alignment of

data representation in space, which was impossible before as samples could only be aligned using a shared index either in the observation or feature space (with AnnData). Once datasets are aligned, they can be efficiently queried and aggregated. Additionally, we incorporated an interactive annotation plugin called napari-spatialdata. This integration allows for practical spatial annotations, facilitating more profound and interactive data exploration. We also include an additional package for static plotting and a package for general reader function for various spatial omics technologies.

To demonstrate the potential of SpatialData, we applied it to a breast cancer study that combined H&E images with assays such as 10X Genomics Visium and Xenium. By employing this framework, we could seamlessly integrate and analyze data from the different modalities, offering insights that wouldn't be achievable by examining datasets individually. For example, we showcase how aggregating cell type proportions across visium spots based on ground-truth cell types obtained from Xenium allows us to compare the obtained cell type proportions with the inferred ones from deconvolution methods. Our results highlighted the concordance and consistency across different datasets and modalities, emphasizing the power of this integrative approach.

In conclusion, SpatialData offers a robust infrastructure for integrating and analyzing spatial omics data. This tool is not only beneficial for complex multi-modal studies but also for single-modality datasets. Our findings pave the way for a more unified and comprehensive exploration of biological systems. We hope that this effort lays the foundation for a new format in the spatial omics domain that could be easily accessed by other programming frameworks, such as R and JavaScript, thus making multi-platform spatial omics data analysis more convenient for the users.

3.4 The scverse project provides a computational ecosystem for single-cell omics data analysis

This correspondence was published in Nature Biotechnology in May 2023. The full citation is:

Isaac Virshup*, Danila Bredikhin*, Lukas Heumos*, Giovanni Palla*, Gregor Sturm*, Adam Gayoso*, Ilia Kats, Mikaela Koutrouli, Bonnie Berger, Dana Pe'er, Aviv Regev, Sarah A Teichmann, Francesca Finotello, F Alexander Wolf, Nir Yosef, Oliver Stegle, Fabian J Theis
The scverse project provides a computational ecosystem for single-cell omics data analysis
Nature Biotechnology 2023
https://doi.org/10.1038/s41587-023-01733-8

Contribution

This correspondence presents the scverse project, an open-source software consortium that provides foundational software for single-cell genomics data analysis in Python. I am a co-founder and co-developer of scverse. I have written the manuscript together with my co-authors.

Summary

Single-cell omics technologies have been pivotal for generating cell atlases across various tissues and species and unraveling the mechanics of development, disease, and homeostasis. However, a surge in computational tools introduced data format inconsistencies, and issues with APIs, and user interfaces. In this correspondence, we present "scverse," an open-source software ecosystem designed for single-cell data storage and analysis. This initiative aims to consolidate the omics analysis community by integrating multiple tools within Python.

Scverse's foundation includes the AnnData and MuData classes, which handle unimodal and multimodal high-dimensional data, and SpatialData for spatial omics data. These structures ensure tight alignment between annotations and learned representations with actual data. Standardizing storage with formats like HDF5 and Zarr assures data interoperability, further augmented by improved metadata management and collaboration with standards like OME-NGFF.

Moreover, scverse offers toolkits for core tasks in single-cell genomics, such as Scanpy, Muon, and Squidpy, which facilitate diverse data analyses. Their joint development ensures consistency, reduced redundancies, and ongoing support.

The community-driven nature of scverse is essential. Multiple tools have been developed based on its foundational packages, forming the scverse ecosystem. Ecosystem packages adhering to best practices are promoted on the scverse website. Our work is hosted on public platforms like GitHub to ensure transparency and community engagement. We promote open discussions via Zulip chat and Discourse forum, fostering collaborations beyond individual research groups.

As we advance into an era of more sophisticated experimental methods and bigger datasets, the need for adaptable and interoperable software becomes paramount. With its core tools and community focus, scverse is primed to champion the forthcoming innovations in single-cell genomics.

3.5 Mapping cells through time and space with moscot

This research paper is a preprint. The full citation is as follows:

Dominik Klein*, Giovanni Palla*, Marius Lange*, Michal Klein*, Zoe Piran*, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Aimee Bastidas-Ponce, Marta Tarquis-Medina, Heiko Lickert, Mostafa Bakhti, Mor Nitzan, Marco Cuturi, Fabian J Theis
Mapping cells through time and space with moscot
Bioarxiv 2023
https://doi.org/10.1101/2023.05.11.540374

Contribution

I conceived the project with Dominik Klein, Marius Lange, Michal Klein, and Zoe Piran, I have led the software development with Dominik Klein and Michal Klein, with help from Marius Lange and Zoe Piran and other co-authors. I performed the analysis of the spatial mapping and spatial alignment problem, as well as some of the analysis of the spatiotemporal problem. I wrote the manuscript with Dominik Klein, Marius Lange, Zoe Piran, and Fabian Theis, and with assistance from all other co-authors.

All code published in the context of this project can be found on GitHub

- https://github.com/theislab/moscot
- https://github.com/theislab/moscot_notebooks
- https://github.com/theislab/moscot-framework_reproducibility
- https://github.com/theislab/moscot_benchmarks

Summary

In prior research, the optimal transport (OT) theory, a mathematical domain for comparing and aligning probability distributions, was utilized for mapping and alignment in cellular biology. While promising, current OT methods in single-cell genomics face three issues: 1) they are primarily for uni-modal data, missing out on multi-modal information; 2) high computational costs, especially with increasing cell numbers; and 3) existing tools have varied backends, complicating their adaptation for novel tasks like spatiotemporal mapping.

Addressing these challenges, we introduce a computational framework, multi-omics single-cell optimal transport (moscot). This tool has three primary features. First, it fully supports multi-modal data, capitalizing on joint cellular representations. Second, its scalability is enhanced, making it suitable for large-scale datasets. Lastly, it unifies prior OT applications in single-cell genomics and introduces a novel spatiotemporal application.

We utilized moscot to address feature mapping and spatial alignment issues often encountered in spatial omics data. Notably, the moscot tool effectively mapped dissociated single-cell profiles to their spatial arrangement, outperforming alternative methods such as Tangram and gimVI in various datasets. We also applied moscot to a Cellular Indexing of Transcriptomes and Epitopes sequencing (CITE-seq) and MERFISH dataset of the mouse liver, where moscot enabled the identification of central veins (CV) and portal veins (PV), enabling the further characterization of liver zonation. Additionally, moscot facilitated the alignment of vast Spatial Transcriptomics (ST) datasets, marking a significant advancement in building a common coordinate framework of biological systems. Our results with the moscot tool demonstrated its robustness and scalability in handling large datasets, providing a holistic view of tissue structures and gene expression within a spatial context.

moscot's capability also enabled us to learn cellular differentiation trajectories integrating spatial coordinates with gene expression data, known as spatiotemporal mapping. This novel approach, which we validated using a mouse embryogenesis atlas, benefits from including spatial information, enhancing mapping precision, enabling high-resolution cell type annotation transfer, and supporting the study of differentiation trajectories.

Finally, the moscot framework helped hypothesize the cell ancestry of delta and epsilon cells during mouse pancreatic development. In conclusion, moscot offers a transformative framework for mapping and aligning single-cell data across various domains and is available as an open-source Python package.

4 Conclusion and Outlook

Here, I discuss the limitations and challenges of the work presented in this thesis in the broader context of spatial omics data analysis and outline an outlook for the next challenges in uncovering spatial components of molecular tissue biology.

4.1 Conclusion

Spatial omics data offers a detailed perspective on tissues and organs' cellular structure. Though the technology is advanced enough for broad application, researchers are just beginning to understand tissue structure and function at such a detailed level.

4.1.1 Challenges in Data Processing

Several challenges persist despite significant advancements in the field. Technically, from a preprocessing and data quality perspective, the robust identification of cellular entities remains unresolved. In the case of image-based transcriptomics, this involves the step of cell segmentation and its associated challenges, such as identifying a cell by its nucleus or membrane and accounting for varied cell shapes across tissues. It is evident that the precise segmentation of an adipocyte and a neuron poses different challenges due to changing cell shapes and cellular density in the tissue region of interest. There have been several recent preprints that look at the general quality of data generated with competing imaging-based spatial transcriptomics platforms [178, 179] and found that a large contributor of the quality of the data is how well the segmentation step is performed. This has also been evaluated in a recent processing workflow benchmark [180]. There's a need to consider whether specialized models for particular biological systems are necessary, or if general-purpose models, similar to the "segment anything" framework [181], are sufficient, or instead methods tailored to the specific biological settings and leveraging additional data modalities are needed [182]. Scalable infrastructure for data storage and handling is also essential, solutions as the one proposed in this thesis will be instrumental to handle the large diversity of spatial omics data (*Section. 3.3 3.2*). Another key challenge in image-based transcriptomics data analysis is normalization. While in single-cell genomics, this problem is still being investigated [84], in spatial transcriptomics, it has recently been considered [183]. A key problem seems to be the fact that due to the limited feature selection, often optimally designed for the tissue of interest [180], standard library size normalization is not optimal, yet utilizing the area of the segmentation mask [184] might also give rise to normalization artifacts. I expect that the computational community interested in data analysis of spatial omics data will spend considerable efforts in delineating the best

normalization methods for image-based spatial transcriptomics, as the technology adoption becomes more widespread. In contrast, array-based technologies like Visium, Stereo-seq, and Slide-seq have made considerable progress in increasing the resolution of the capture areas. However, determining the optimal resolution for aggregating gene expression counts, and subsequently generating the gene expression matrix for analysis, is still debated [185]. These aren't just technical considerations since they influence all subsequent analysis steps, such as identifying the optimal noise model for normalization and differential expression and recovering the cellular heterogeneity of interest, for example, based on the scRNA-seq dataset obtained from the same sample. On this front, there has been a significant amount of work on deconvolution methods [96, 97, 186]. The question is whether such methods will be able to handle such spot resolution, or whether mapping methods [98, 100] will also be competitive, such as the one presented in this thesis [100].

4.1.2 Spatial Domains and Niche Identification

Obtaining detailed molecular profiles across extensive tissue areas allows for defining tissue regions based on molecular profile similarities and spatial layout. This challenge, often termed spatial domain identification, has seen numerous methods proposed (*Section. 3.1 3.2*). However, the benchmarks and evaluations often appear too specific to certain data. They may not be applicable when tissue regions do not maintain a stereotypical organization, such as the one observed in the brain, thus presenting more modular and multi-scale regionalization, like the one observed in the liver, the pancreas, or the TME. While this remains an open issue, identifying spatial domains can yield valuable insights into tissue function. Furthermore, such concepts are tightly related to the "cellular niche" concept, a specific neighborhood composition regarding cell types and/or molecular profiles observed across the tissue. In the literature, the concept of niche is becoming more prevalent [65, 187], and it often refers to a different clustering approach, which essentially labels cells based not only on their molecular profile similarity but also on the similarity of the molecular profile of their neighbors. Therefore, defining the neighbor in spatial coordinates and across coordinates systems is the first crucial step of this analysis, and several tools already rely on algorithms proposed in this thesis [43, 124](*Section. 3.3 3.2*). The problem of niche identification can be related to the cell type definition in single-cell genomics. Similarly to the challenges encountered in single-cell genomics, various methods will likely emerge to identify cellular niches. Yet, it will ultimately require a deep investigation and manual annotation of cellular signatures by the analyst expert in the tissue of interest to settle on the optimal way to define the niche. Nevertheless, it is a compelling problem that will have a tremendous impact on the domain of spatial genomics data analysis.

4.1.3 Spatial Differential Analysis

Understanding tissue organization, especially its alteration in diseases, is crucial. In *Section. 3.3*, we present data analysis techniques that reveal the interdependence of cells in tissues, yet a crucial challenge remains. There's an intrinsic inconsistency in organization of cell types across different sections of the same tissue sample, leading to variability within samples [188]. Additionally, individual differences between samples can further overshadow subtle data insights that characterize

tissue-level cellular variability between conditions. Considering such variability is particularly important Nevertheless, there have been significant findings, for example in the context of cancer research. Specifically, neighborhood enrichment measurement has enhanced patient categorization in breast cancer46 or lung cancer [139], resulting in optimized treatments. Another related challenge is spatial differential expression, which entails assessing spatial variability between groups, such as in a case-control setting. This should not be confused with spatially variable gene detection, such as the one presented in this thesis [43]. Linear modeling approaches that consider both the spatial dependencies as well as the covariates to test represent interesting solutions [189]. Extensions of conventional methods borrowing from generalized linear and additive modeling will likely deliver the most significant insights. Similarly, differential tests for spatial enrichment are also of interest, especially in the context of the identification of disease determinants. While techniques like point pattern analysis or neighborhood enrichment provide valuable information, they need to directly tackle the variance in neighbor distribution spatially. Some approaches that use graph descriptors to perform such analysis have been recently proposed [190].

4.1.4 Spatio-temporal Analysis of Trajectory of Niches

With single-cell genomics, a primary challenge is to infer cell lineage maps, which involves organizing cells from specific moments into continuous transitions that elucidate cell fate decisions. Current spatial profiling techniques also only provide a snapshot of biological processes. However, they offer extra information that affects cell decisions, lineage, and differentiation, which is encoded in spatial coordinates. Particularly, spatial data enables cell-fate-inference methods to interpolate in molecular cell state space and spatial coordinates. From a causal viewpoint, branching events might be clarified by the spatial context. This insight suggests that developmental processes, where the localization of morphogenic factors is tightly regulated, can be modeled more accurately by considering cell-to-cell interactions and other spatial effects. In the work described in *Section. 3.5*, we proposed for the first time a trajectory inference method that builds on the approach of Waddington-OT [191] yet also leverages structural similarity in terms of spatial coordinates [100]. Furthermore, it can be readily interfaced with state-of-the-art algorithms to determine cell fates like CellRank [79]. While this is an one attempt at explicitly modeling extrinsic factors of variations in spatially-resolved single-cell data, it opens various research directions, including lineage information and lifting the RNA-velocity [77, 78] concept to spatial coordinates. RNA velocity can also be directly measured in situ, and these state-space velocities can be linked to biological axes of variation [192]. Ultimately, this discussion underscores the potential to translate, expand, or reframe concepts originally developed for trajectory inference in single-cell genomics by integrating spatial information.

4.1.5 Multimodal Mapping and Integration of Single Cell and Spatial Genomics

This thesis addresses the challenge of integrating dissociated scRNA-seq profiles with their spatial context [100](*Section. 3.5*), though a complete solution remains elusive. While extensive research has been conducted on single-cell data, and some methods have bridged the gap between single-

cell and spatial domains, significant limitations persist. A central issue concerns the feature matching steps. Many methods, including the one discussed in Section 3, match gene names between datasets. However, variations like shifts in cell type or state can result in discrepancies between dissociated and spatial datasets, as it has been clearly demonstrated in several benchmarks and studies [180, 183, 184]. The technique introduced in *Section. 3.5* accounts for dissociation variation, in particular variation in terms of proportions between cell states in single-cell and spatial data, but still requires an initial feature-matching step. A systematic exploration of how the matching performs as features are altered in the shared space is of great interest. Techniques that are able to map gene expression profiles from single-cell data to spatial data are also very useful to impute unobserved single-cell gene expression profiles in space. This limitation, intrinsically due to the complexity of image-based transcriptomics techniques, might be alleviated by novel technologies but hardly completely resolved. That also applies to multimodal information such as proteins and chromatin accessibility, and it is of general interest, specifically in the scope of leveraging large annotated single-cell datasets for mapping them in their spatial context. It's also worth examining how much and where the matching quality improves, especially when extra information on the cellular state is available, like morphology for spatial datasets or chromatin accessibility for single-cell datasets, is incorporated.

4.2 Outlook

In this section, I discuss some of the significant next-generation problems that I believe the field will be approaching.

4.2.1 Multimodal Generative Modelling

Multimodal foundational models, such as Contrastive Language-Image Pretraining (CLIP) [120] and DALL-E [121], have transformed computer vision and natural language understanding into cultural phenomena. They are versatile tools, widely recognized for various applications including text-to-image generation and image descriptions. DALL-E for example is a generative model that takes text as input and generate images as output, where the images correspond to the text description at input [121]. CLIP instead is a model that embeds two modalities jointly, text and images, and learn to pair the two modalities, by training on dataset of figures and captions [120]. The biomedical domain is also beginning to adopt these innovative techniques. For instance, one model was developed to integrate H&E data with text, enhancing question-answering and classification tasks [193, 194]. Another leverages a cross-attention framework like Flamingo [195] to produce text from combined text and image inputs [196]. It's conceptually similar to CLIP yet it consists of cross-attention layers that interleave the modalities, and furthermore, it uses a text decoder and therefore it is trained as a text generative model. Spatial omics data provide a natural pairing between modalities, since because of how the measurement is performed, we have an exact match between the molecular profile and the spatial location of the image that it corresponds to. There are several use cases of generative AI in spatial omics data. A prime example is predicting gene expression from images, which can substantially aid in molecularly characterizing unimodal imaging datasets of pathology slides, especially H&E – a dominant method in biomedical imaging

used extensively in cancer pathology for biomarker identification and diagnosis. Such generative task can be paired with conditional generation [197, 198] to steer the generative model with additional information such as pathologist annotation, additional molecular profiles such as specific markers or probes, or other types of metadata that are useful for the generative process. Furthermore, embedding the different modalities by leveraging the CLIP objective has demonstrated the superior performance of multimodal models over unimodal counterparts regarding image classification accuracy [120]. Another application of multimodal models trained on spatial transcriptomics data will be their performance on image-specific features such as annotation of tissue states. This approach could be impactful for complex tasks like identifying minute, challenging-to-detect pathological tissue states in pathology slides, for which the genomic information instead would provide a stronger signal for the prediction. In summary, foundation models that concurrently learn morphological and gene expression information could be instrumental in various clinical applications, and spatial transcriptomics is poised to become the technology of choice for the data generation efforts required to train these models.

4.2.2 Tissue Functional Genomics

Large-scale perturbations in single-cell genomics have improved the elucidation of gene regulatory mechanisms at a cellular state resolution195. This has offered insights into reversing gene expression profiles of cells, potentially moving from diseased to healthy states, highlighting novel avenues of therapeutic discovery. Moreover, the accessibility of large-scale perturbational studies has surged, particularly in high-throughput imaging. Comprehensive studies like Cell Painting [199] or Optical Screens [141] assess the impact of perturbations on cellular phenotypes morphologically across diverse experimental conditions. Such studies pave the way for phenotypic maps, where nearly all genes have been altered in cells via CRISPR knock-outs and thousands of drugs tested. These efforts aim to chart a phenotypic map of cellular perturbations and might reveal previously unidentified drug mechanisms that produce similar phenotypic outcomes to disrupting specific gene expression signatures. While providing rich information on the cellular phenotypic response under perturbation, they have a crucial limitation that might hamper the discovery potential: the cellular phenotype being perturbed are typically cell lines on a dish, which means a very low heterogeneity in the naive state as well as a lack of meaningful cellular communications and the emergence of cellular niches. Furthermore, while both methods offer robust insights, they face additional unique and orthogonal challenges. Perturbational single-cell genomics excels in characterizing cellular states from a gene regulatory angle. Still, it lacks clarity on whether the perturbation reverted the intended cellular phenotype, often defined by morphological or protein expression changes. Conversely, image-based profiling techniques discern these phenotypical changes but struggle to correlate them with molecular profiles. Spatial genomics bridges this gap, capturing both morphological and molecular data. Furthermore, this method is advantageous as it allows measurements directly in tissues, unveiling extracellular gene functions, especially in tissue contexts. The ability to dissect cellular perturbation in its tissue context is particularly useful because of the possibility to characterize also the cellular neighborhood, and hence precisely disentangling the intrinsic and extrinsic factors of variation on cellular phenotypes. Existing technologies already harness this approach. For example, in cancer research,

CRISPR knock-outs [140], identifiable via fluorescent barcodes, are retrieved from cancer tissues, illuminating the role of specific gene knock-outs in tumor growth, histopathological features, and the TME. Newer techniques enable this approach by measuring transcriptomes instead of protein measurements, by means of highly-multiplexed spatial transcriptomics techniques [142]. From a computational perspective, such measurements will enable the development of predictive models that account for spatial coordinates, neighbor composition as well as molecular profiles of cellular perturbation response [200], possibly paving the way for causal models of cellular phenotypes and niches [201].

4.2.3 Spatial Cell Atlases

Detailed cellular maps have paved the way for reference maps of tissues and organs at single-cell resolution. However, these maps primarily focus on molecular profiles of isolated cells. Integrating these profiles into their natural tissue environments is vital for the next generation of tissue atlases, aiming to create a Common Coordinate Framework (CCF) for organs and organisms [129]. Efforts like the Human Reference Atlas [132] are underway to create a digital representation of the human body, capturing the 3D organization of organs and detailing cellular relationships and unique biomarkers. Arguably, the only existing CCF is the Brain, which in the case of the Human Brain it has been characterized at an anatomical [202] and molecular [128] resolution. This atlas intends to allow in-depth analysis of the human brain, enhancing spatial and semantic studies. Such a reference will improve our grasp of how a healthy body functions and its alterations with age or illness. Yet, constructing atlases for other human tissues and organs presents challenges. Most notably, there's no universally recognized naming system for anatomical structures, cell types, or biomarkers outside the brain [132, 203]. Essential data is scattered across different sources, lacking a consolidated 3D reference. This data fragmentation complicates biomedical advancements due to data management and comparison difficulties. Another pivotal aspect is deciding the granularity of the human CCF. A macro-scale would give a broad view of the major organs, a mesoscale offers insights into specific structures like parts of the lungs or kidneys, and a micro-scale provides a detailed cellular perspective. Most prior atlas projects focused on the mesoscale. The future challenge lies in crafting micro-scale atlases that can seamlessly integrate with broader scales. The depth of scale directly influences data needs and analytical methods, and several computational challenges lie ahead. For example, it will require efficient infrastructure that is able to store the diversity of data but also scalable visualization tools that can visualize the different views and modalities of the data. A multi-level hierarchical structure is an achievable and vital objective for contemporary human atlas research. Computational methods and infrastructure that are able to integrate the multimodal nature of the data will be instrumental in building the next generation of common coordinate frameworks of tissues and organs.

Acronyms

AnnData	Annotated Data
ATAC-seq	Assay for Transposase-Accessible Chromatin with sequencing
CCC	Cell-Cell Communication
CCF	Common Coordinate Framework
CCS	Common Coordinate System
CITE-seq	Cellular Indexing of Transcriptomes and Epitopes sequencing
CLIP	Contrastive Language-Image Pretraining
CyCIF	Cyclic Immunofluorescence
FAIR	Findability, Accessibility, Interoperability, and Reusability
FFPE	formalin-fixed paraffin-embedded
FGW	Fused-Gromov-Wasserstein
FISH	Fluorescent In Situe Hybridization
GW	Gromov-Wasserstein
IMC	Imaging Mass Spectrometry
MIBI-TOF	multiplexed ion beam imaging by time of flight
mIHC	multiplexed Immuno-Histo-Chemistry
moscot	multi-omics single-cell optimal transport
MS	Mass Spectrometry
MuData	Multimodal Annotated Data
OME	Open Microscopy Environment
OT	Optimal Transport
scRNA-seq	single cell RNA-sequencing
squidpy	Spatial Quantification of Molecular Data in Python
ST	Spatial Transcriptomics
t-SNE	t-distributed Stochastic Neighbor Embedding
TME	Tumor Micro-Environment
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique Molecular Identifiers
VAE	Variational Autoencoders
W	Wasserstein

Bibliography

1. B. Alberts, A. D. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell (Sixth Edition) EBook Folder*. en. W. W. Norton, Incorporated, 2020.
2. D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry*. en. W. H. Freeman, 2017.
3. J. W. Lichtman and J.-A. Conchello. "Fluorescence microscopy". en. *Nat. Methods* 2:12, 2005, pp. 910–919.
4. P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. "Quantitative single-cell RNA-seq with unique molecular identifiers". *Nature*, 2014.
5. R. Jiang, T. Sun, D. Song, and J. J. Li. "Statistics or biology: the zero-inflation controversy about scRNA-seq data". en. *Genome Biol.* 23:1, 2022, p. 31.
6. A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg. "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications". en. *Genome Med.* 9:1, 2017, p. 75.
7. M. D. Luecken and F. J. Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". *Mol. Syst. Biol.* 15:6, 2019, e8746.
8. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets". en. *Cell* 161:5, 2015, pp. 1202–1214.
9. S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. "Quantitative single-cell RNA-seq with unique molecular identifiers". en. *Nat. Methods* 11:2, 2014, pp. 163–166.
10. G. D. Evrony, A. G. Hinch, and C. Luo. "Applications of Single-Cell DNA Sequencing". *Annu. Rev. Genomics Hum. Genet.* 22:1, 2021, pp. 171–197.
11. S. Pott and J. D. Lieb. "Single-cell ATAC-seq: strength in numbers". *Genome Biol.* 16:1, 2015, p. 172.
12. M. Bartosovic, M. Kabbe, and G. Castelo-Branco. "Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues". *Nat. Biotechnol.*, 2021.
13. F. J. Rang, K. L. de Luca, S. S. de Vries, C. Valdes-Quezada, et al. "Single-cell profiling of transcriptome and histone modifications with EpiDamID". *Mol. Cell*, 2022.

14. J. Moody, T. Kouno, A. Suzuki, Y. Shibayama, C. Terao, J.-C. Chang, F. López-Redondo, C. W. Yip, J. Severin, H. Suetsugu, Y. Ando, K. Yamamoto, P. Carninci, J. W. Shin, and C.-C. Hon. "Profiling of transcribed cis-regulatory elements in single cells". *bioRxiv*, 2021.

15. M. VanInsberghe, J. van den Berg, A. Andersson-Rolf, H. Clevers, and A. van Oudenaarden. "Single-cell Ribo-seq reveals cell cycle-dependent translational pausing". en. *Nature* 597:7877, 2021, pp. 561–565.

16. C. J. Frangieh, J. C. Melms, P. I. Thakore, K. R. Geiger-Schuller, P. Ho, A. M. Luoma, B. Cleary, L. Jerby-Arnon, S. Malu, M. S. Cuoco, M. Zhao, C. R. Ager, M. Rogava, L. Hovey, A. Rotem, C. Bernatchez, K. W. Wucherpfennig, B. E. Johnson, O. Rozenblatt-Rosen, D. Schadendorf, A. Regev, and B. Izar. "Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion". en. *Nat. Genet.* 53:3, 2021, pp. 332–341.

17. M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. "Simultaneous epitope and transcriptome measurement in single cells". en. *Nat. Methods* 14:9, 2017, pp. 865–868.

18. N. Slavov. "Scaling Up Single-Cell Proteomics". en. *Mol. Cell. Proteomics* 21:1, 2022.

19. N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. "Near-optimal probabilistic RNA-seq quantification". en. *Nat. Biotechnol.* 34:5, 2016, pp. 525–527.

20. L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, A. Asplund, E. Sjöstedt, E. Lundberg, C. A.-K. Szigyarto, M. Skogs, J. O. Takanen, H. Berling, H. Tegel, J. Mulder, P. Nilsson, J. M. Schwenk, C. Lindskog, F. Danielsson, A. Mardinoglu, A. Sivertsson, K. von Feilitzen, M. Forsberg, M. Zwahlen, I. Olsson, S. Navani, M. Huss, J. Nielsen, F. Ponten, and M. Uhlén. "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics". en. *Mol. Cell. Proteomics* 13:2, 2014, pp. 397–406.

21. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén. "Proteomics. Tissue-based map of the human proteome". en. *Science* 347:6220, 2015, p. 1260419.

22. I. Lukonin, D. Serra, L. Challet Meylan, K. Volkmann, J. Baaten, R. Zhao, S. Meeusen, K. Colman, F. Maurer, M. B. Stadler, J. Jenkins, and P. Liberali. "Phenotypic landscape of intestinal organoid regeneration". en. *Nature* 586:7828, 2020, pp. 275–280.

23. E. Lundberg and G. H. H. Borner. "Spatial proteomics: a powerful discovery tool for cell biology". en. *Nat. Rev. Mol. Cell Biol.* 20:5, 2019, pp. 285–302.

24. X. Zhuang. "Spatially resolved single-cell genomics and transcriptomics by imaging". en. *Nat. Methods* 18:1, 2021, pp. 18–22.

25. L. Larsson, J. Frisén, and J. Lundeberg. "Spatially resolved transcriptomics adds a new dimension to genomics". en. *Nat. Methods* 18:1, 2021, pp. 15–18.

26. C. Ortiz, M. Carlén, and K. Meletis. "Spatial Transcriptomics: Molecular Maps of the Mammalian Brain". en. *Annu. Rev. Neurosci.*, 2021.

27. V. Marx. "Method of the Year: spatially resolved transcriptomics". en. *Nat. Methods* 18:1, 2021, pp. 9–14.

28. J. L. Close, B. R. Long, and H. Zeng. "Spatially resolved transcriptomics in neuroscience". en. *Nat. Methods* 18:1, 2021, pp. 23–25.

29. C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, and L. Cai. "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH". en. *Nature* 568:7751, 2019, pp. 235–239.

30. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. "RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells". en. *Science* 348:6233, 2015, aaa6090.

31. J.-R. Lin, M. Fallahi-Sichani, J.-Y. Chen, and P. K. Sorger. "Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging". en. *Curr. Protoc. Chem. Biol.* 8:4, 2016, pp. 251–264.

32. E. K. Neumann, E. S. Rivera, N. H. Patterson, J. L. Allen, M. Brewer, M. deCaestecker, A. B. Fogo, R. M. Caprioli, and J. M. Spraggins. "Highly Multiplexed Immunofluorescence of the Human Kidney using Co-Detection by Indexing (CODEX)". en. 2020.

33. S. He, R. Bhatt, B. Birditt, C. Brown, E. Brown, K. Chantranuvatana, P. Danaher, D. Dunaway, B. Filanoski, R. G. Garrison, G. Geiss, M. T. Gregory, M. L. Hoang, E. E. Killingbeck, T. K. Kim, Y. Kim, M. Korukonda, A. Kutchma, E. Lee, Z. R. Lewis, Y. Liang, J. S. Nelson, G. Ong, E. Perillo, J. Phan, T. Phan-Everson, E. Piazza, T. Rane, Z. Reitz, M. Rhodes, A. Rosenbloom, D. Ross, H. Sato, A. W. Wardhani, C. Williams-Wietzikoski, L. Wu, and J. M. Beechem. "High-Plex Multiomic Analysis in FFPE Tissue at Single-Cellular and Subcellular Resolution by Spatial Molecular Imaging". en. 2021.

34. A. Janesick, R. Shelansky, A. D. Gottscho, F. Wagner, M. Rouault, G. Beliakoff, M. F. de Oliveira, A. Kohlway, J. Abousoud, C. A. Morrison, T. Y. Drennon, S. H. Mohabbat, S. R. Williams, and S. E. B. Taylor. "High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue". en. 2022.

35. L. Moses and L. Pachter. "Museum of spatial transcriptomics". en. *Nat. Methods* 19:5, 2022, pp. 534–546.

36. M. Asp, J. Bergenstråhle, and J. Lundeberg. "Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration". en. *Bioessays*, 2020, e1900221.

37. L. Keren, M. Bosse, S. Thompson, T. Risom, K. Vijayaragavan, E. McCaffrey, D. Marquez, R. Angoshtari, N. F. Greenwald, H. Fienberg, J. Wang, N. Kambham, D. Kirkwood, G. Nolan, T. J. Montine, S. J. Galli, R. West, S. C. Bendall, and M. Angelo. "MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure". en. *Sci Adv* 5:10, 2019, eaax5851.

38. C. Giesen, H. A. O. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schüffler, D. Grolimund, J. M. Buhmann, S. Brandt, Z. Varga, P. J. Wild, D. Günther, and B. Bodenmiller. "Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry". en. *Nat. Methods* 11:4, 2014, pp. 417–422.

39. L. Rappez, M. Stadler, S. Triana, P. Phapale, M. Heikenwalder, and T. Alexandrov. "Spatial single-cell profiling of intracellular metabolomes in situ". en. 2019.

40. J. K. Benotmane, J. Kueckelhaus, P. Will, J. Zhang, V. M. Ravi, K. Joseph, R. Sankowski, J. Beck, C. Lee-Chang, O. Schnell, and D. H. Heiland. "High-sensitive spatially resolved T cell receptor sequencing with SPTCR-seq". en. *Nat. Commun.* 14:1, 2023, pp. 1–17.

41. A. Baysoy, Z. Bai, R. Satija, and R. Fan. "The technological landscape and applications of single-cell multi-omics". en. *Nat. Rev. Mol. Cell Biol.* 24:10, 2023, pp. 695–713.

42. G. Palla, H. Spitzer, M. Klein, D. Fischer, A. C. Schaar, L. B. Kuemmerle, S. Rybakov, I. L. Ibarra, O. Holmberg, I. Virshup, M. Lotfollahi, S. Richter, and F. J. Theis. "Squidpy: a scalable framework for spatial omics analysis". en. *Nat. Methods* 19:2, 2022, pp. 171–178.

43. G. Palla, D. S. Fischer, A. Regev, and F. J. Theis. "Spatial components of molecular tissue biology". en. *Nat. Biotechnol.*, 2022.

44. D. Phillips, M. Matusiak, B. R. Gutierrez, S. S. Bhate, G. L. Barlow, S. Jiang, J. Demeter, K. S. Smythe, R. H. Pierce, S. P. Fling, N. Ramchurren, M. A. Cheever, Y. Goltsev, R. B. West, M. S. Khodadoust, Y. H. Kim, C. M. Schuerch, and G. P. Nolan. "Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma". 2020.

45. F. J. Hartmann, D. Mrdjen, E. McCaffrey, D. R. Glass, N. F. Greenwald, A. Bharadwaj, Z. Khair, S. G. S. Verberk, A. Baranski, R. Baskar, W. Graf, D. Van Valen, J. Van den Bossche, M. Angelo, and S. C. Bendall. "Single-cell metabolic profiling of human cytotoxic T cells". en. *Nat. Biotechnol.*, 2020.

46. H. W. Jackson, J. R. Fischer, V. R. T. Zanotelli, H. R. Ali, R. Mechera, S. D. Soysal, H. Moch, S. Muenst, Z. Varga, W. P. Weber, and B. Bodenmiller. "The single-cell pathology landscape of breast cancer". en. *Nature* 578:7796, 2020, pp. 615–620.

47. L. Keren, M. Bosse, D. Marquez, R. Angoshtari, S. Jain, S. Varma, S.-R. Yang, A. Kurian, D. Van Valen, R. West, S. C. Bendall, and M. Angelo. "A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging". en. *Cell* 174:6, 2018, 1373–1387.e19.

48. D. Cui Zhou, R. G. Jayasinghe, S. Chen, J. M. Herndon, M. D. Iglesia, P. Navale, M. C. Wendl, W. Caravan, K. Sato, E. Storrs, C.-K. Mo, J. Liu, A. N. Southard-Smith, Y. Wu, N. Naser Al Deen, J. M. Baer, R. S. Fulton, M. A. Wyczalkowski, R. Liu, C. C. Fronick, L. A. Fulton, A. Shinkle, L. Thammavong, H. Zhu, H. Sun, L.-B. Wang, Y. Li, C. Zuo, J. F. McMichael, S. R. Davies, E. L. Appelbaum, K. J. Robbins, S. E. Chasnoff, X. Yang, A. N. Reeb, C. Oh, M. Serasanambati, P. Lal, R. Varghese, J. R. Mashl, J. Ponce, N. V. Terekhanova, L. Yao, F. Wang, L. Chen, M. Schnaubelt, R. J.-H. Lu, J. K. Schwarz, S. V. Puram, A. H. Kim, S.-K. Song, K. I. Shoghi, K. S. Lau, T. Ju, K. Chen, D. Chatterjee, W. G. Hawkins, H. Zhang, S. Achilefu, M. G. Chheda, S. T. Oh, W. E. Gillanders, F. Chen, D. G. DeNardo, R. C. Fields, and L. Ding. "Spatially restricted drivers and transitional cell populations cooperate with the microenvironment in untreated and chemo-resistant pancreatic cancer". en. *Nat. Genet.* 54:9, 2022, pp. 1390–1405.

49. A. Irmisch, X. Bonilla, S. Chevrier, K.-V. Lehmann, F. Singer, N. C. Toussaint, C. Esposito, J. Mena, E. S. Milani, R. Casanova, D. J. Stekhoven, R. Wegmann, F. Jacob, B. Sobottka, S. Goetze, J. Kuipers, J. Sarabia Del Castillo, M. Prummer, M. A. Tuncel, U. Menzel, A. Jacobs, S. Engler, S. Sivapatham, A. L. Frei, G. Gut, J. Ficek, N. Miglino, Tumor Profiler Consortium, R. Aebersold, M. Bacac, N. Beerenwinkel, C. Beisel, B. Bodenmiller, R. Dummer, V. Heinzelmann-Schwarz, V. H. Koelzer, M. G. Manz, H. Moch, L. Pelkmans, B. Snijder, A. P. A. Theocharides, M. Tolnay, A. Wicki, B. Wollscheid, G. Rätsch, and M. P. Levesque. "The Tumor Profiler Study: integrated, multi-omic, functional tumor profiling for clinical decision support". en. *Cancer Cell* 39:3, 2021, pp. 288–293.

50. A. Chen, S. Liao, M. Cheng, K. Ma, L. Wu, Y. Lai, X. Qiu, J. Yang, J. Xu, S. Hao, X. Wang, H. Lu, X. Chen, X. Liu, X. Huang, Z. Li, Y. Hong, Y. Jiang, J. Peng, S. Liu, M. Shen, C. Liu, Q. Li, Y. Yuan, X. Wei, H. Zheng, W. Feng, Z. Wang, Y. Liu, Z. Wang, Y. Yang, H. Xiang, L. Han, B. Qin, P. Guo, G. Lai, P. Muñoz-Cánoves, P. H. Maxwell, J. P. Thiery, Q.-F. Wu, F. Zhao, B. Chen, M. Li, X. Dai, S. Wang, H. Kuang, J. Hui, L. Wang, J.-F. Fei, O. Wang, X. Wei, H. Lu, B. Wang, S. Liu, Y. Gu, M. Ni, W. Zhang, F. Mu, Y. Yin, H. Yang, M. Lisby, R. J. Cornall, J. Mulder, M. Uhlén, M. A. Esteban, Y. Li, L. Liu, X. Xu, and J. Wang. "Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays". en. *Cell* 185:10, 2022, 1777–1792.e21.

51. C. Liu, R. Li, Y. Li, X. Lin, K. Zhao, Q. Liu, S. Wang, X. Yang, X. Shi, Y. Ma, C. Pei, H. Wang, W. Bao, J. Hui, T. Yang, Z. Xu, T. Lai, M. A. Berberoglu, S. K. Sahu, M. A. Esteban, K. Ma, G. Fan, Y. Li, S. Liu, A. Chen, X. Xu, Z. Dong, and L. Liu. "Spatiotemporal mapping of gene expression landscapes and developmental trajectories during zebrafish embryogenesis". en. *Dev. Cell* 57:10, 2022, 1284–1298.e5.

52. M. Guilliams, J. Bonnardel, B. Haest, B. Vanderborght, C. Wagner, A. Remmerie, A. Bujko, L. Martens, T. Thoné, R. Browaeys, F. F. De Ponti, B. Vanneste, C. Zwicker, F. R. Svedberg, T. Vanhalewyn, A. Gonçalves, S. Lippens, B. Devriendt, E. Cox, G. Ferrero, V. Wittamer, A. Willaert, S. J. F. Kaptein, J. Neyts, K. Dallmeier, P. Geldhof, S. Casaert, B. Deplancke, P. Ten Dijke, A. Hoorens, A. Vanlander, F. Berrevoet, Y. Van Nieuwenhove, Y. Saeys, W. Saelens, H. Van Vlierberghe, L. Devisscher, and C. L. Scott. "Spatial proteoge-

nomics reveals distinct and evolutionarily conserved hepatic macrophage niches". en. *Cell* 185:2, 2022, 379–396.e38.

53. M. Asp, S. Giacomello, L. Larsson, C. Wu, D. Fürth, X. Qian, E. Wärdell, J. Custodio, J. Reimegård, F. Salmén, C. Österholm, P. L. Ståhl, E. Sundström, E. Åkesson, O. Bergmann, M. Bienko, A. Månsson-Broberg, M. Nilsson, C. Sylvén, and J. Lundeberg. "A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart". en. *Cell* 179:7, 2019, 1647–1660.e19.

54. M. Litviňuková, C. Talavera-López, H. Maatz, D. Reichart, C. L. Worth, E. L. Lindberg, M. Kanda, K. Polanski, M. Heinig, M. Lee, E. R. Nadelmann, K. Roberts, L. Tuck, E. S. Fasouli, D. M. DeLaughter, B. McDonough, H. Wakimoto, J. M. Gorham, S. Samari, K. T. Mahbubani, K. Saeb-Parsy, G. Patone, J. J. Boyle, H. Zhang, H. Zhang, A. Viveiros, G. Y. Oudit, O. A. Bayraktar, J. G. Seidman, C. E. Seidman, M. Noseda, N. Hubner, and S. A. Teichmann. "Cells of the adult human heart". en. *Nature* 588:7838, 2020, pp. 466–472.

55. E. Madissoon, A. J. Oliver, V. Kleshchevnikov, A. Wilbrey-Clark, K. Polanski, N. Richoz, A. Ribeiro Orsi, L. Mamanova, L. Bolt, R. Elmentaite, J. P. Pett, N. Huang, C. Xu, P. He, M. Dabrowska, S. Pritchard, L. Tuck, E. Prigmore, S. Perera, A. Knights, A. Oszlanczi, A. Hunter, S. F. Vieira, M. Patel, R. G. H. Lindeboom, L. S. Campos, K. Matsuo, T. Nakayama, M. Yoshida, K. B. Worlock, M. Z. Nikolić, N. Georgakopoulos, K. T. Mahbubani, K. Saeb-Parsy, O. A. Bayraktar, M. R. Clatworthy, O. Stegle, N. Kumasaka, S. A. Teichmann, and K. B. Meyer. "A spatially resolved atlas of the human lung characterizes a gland-associated immune niche". en. *Nat. Genet.* 55:1, 2023, pp. 66–77.

56. D. Dimitrov, D. Türei, C. Boys, J. Nagai, R. R. Flores, H. Kim, B. Szalai, I. Costa, A. Dugourd, A. Valdeolivas, and J. S. Rodriguez. "Cell-cell communication inference from single-cell RNA-Seq data: A comparison of methods and resources". en. 2021.

57. A. A. Almet, Z. Cang, S. Jin, and Q. Nie. "The landscape of cell-cell communication through single-cell transcriptomics". *Current Opinion in Systems Biology*, 2021.

58. B. G. Bruneau. "Signaling and transcriptional networks in heart development and regeneration". en. *Cold Spring Harb. Perspect. Biol.* 5:3, 2013, a008292.

59. J. Saez-Rodriguez, L. G. Alexopoulos, J. Epperlein, R. Samaga, D. A. Lauffenburger, S. Klamt, and P. K. Sorger. "Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction". *Mol. Syst. Biol.* 5:1, 2009, p. 331.

60. D. Dimitrov, D. Türei, C. Boys, J. S. Nagai, R. O. Ramirez Flores, H. Kim, B. Szalai, I. G. Costa, A. Dugourd, A. Valdeolivas, and J. Saez-Rodriguez. "Comparison of Resources and Methods to infer Cell-Cell Communication from Single-cell RNA Data". en. 2021.

61. M. Efremova, M. Vento-Tormo, S. A. Teichmann, and R. Vento-Tormo. "CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes". en. *Nat. Protoc.* 15:4, 2020, pp. 1484–1506.

62. R. Browaeys, W. Saelens, and Y. Saeys. "NicheNet: modeling intercellular communication by linking ligands to target genes". en. *Nat. Methods* 17:2, 2020, pp. 159–162.

63. D. S. Fischer, A. C. Schaar, and F. J. Theis. "Modeling intercellular communication in tissues using spatial graphs of cells". en. *Nat. Biotechnol.* 41:3, 2022, pp. 332–336.

64. Z. Cang and Q. Nie. "Inferring spatial and signaling relationships between cells from single cell transcriptomic data". en. *Nat. Commun.* 11:1, 2020, p. 2084.

65. D. Haviv, M. Gatie, A.-K. Hadjantonakis, T. Nawy, and D. Pe'er. "The covariance environment defines cellular niches for spatial inference". en. *bioRxiv*, 2023.

66. A. Wagner, A. Regev, and N. Yosef. "Revealing the vectors of cellular identity with single-cell genomics". en. *Nat. Biotechnol.* 34:11, 2016, pp. 1145–1160.

67. A. Wilbrey-Clark, K. Roberts, and S. A. Teichmann. "Cell Atlas technologies and insights into tissue architecture". en. *Biochem. J* 477:8, 2020, pp. 1427–1442.

68. S. Bidanta, K. Börner, B. W. Herr, M. Nagy, K. S. Gustilo, R. Bajema, L. Maier, R. Molontay, and G. Weber. "Functional Tissue Units in the Human Reference Atlas". en. 2023.

69. L. Keren and M. Angelo. "Mapping cell phenotypes in breast cancer". *Nature Cancer* 1:2, 2020, pp. 156–157.

70. R. P. Cunningham and N. Porat-Shliom. "Liver Zonation - Revisiting Old Questions With New Technologies". en. *Front. Physiol.* 12, 2021, p. 732929.

71. K. W. Rogers and A. F. Schier. "Morphogen gradients: from generation to interpretation". en. *Annu. Rev. Cell Dev. Biol.* 27, 2011, pp. 377–407.

72. B. Ladoux and R.-M. Mège. "Mechanobiology of collective cell behaviours". en. *Nat. Rev. Mol. Cell Biol.* 18:12, 2017, pp. 743–757.

73. L. Heumos, A. C. Schaar, C. Lance, A. Litinetskaya, F. Drost, L. Zappia, M. D. Lücken, D. C. Strobl, J. Henao, F. Curion, H. B. Schiller, and F. J. Theis. "Best practices for single-cell analysis across modalities". en. *Nat. Rev. Genet.* 24:8, 2023, pp. 550–572.

74. L. van der Maaten and G. Hinton. "Visualizing Data using t-SNE". *J. Mach. Learn. Res.* 9:86, 2008, pp. 2579–2605.

75. L. McInnes, J. Healy, and J. Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction", 2018. arXiv: 1802.03426 [stat.ML].

76. R. Lopez, A. Gayoso, and N. Yosef. "Enhancing scientific discoveries in molecular biology with deep generative models". en. *Mol. Syst. Biol.* 16:9, 2020, e9198.

77. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko. "RNA velocity of single cells". en. *Nature* 560:7719, 2018, pp. 494–498.

78. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis. "Generalizing RNA velocity to transient cell states through dynamical modeling". en. *Nat. Biotechnol.*, 2020.

79. M. Lange, V. Bergen, M. Klein, M. Setty, B. Reuter, M. Bakhti, H. Lickert, M. Ansari, J. Schniering, H. B. Schiller, D. Pe'er, and F. J. Theis. "CellRank for directed single-cell fate mapping". en. *Nat. Methods* 19:2, 2022, pp. 159–170.

80. M. Setty, V. Kiseliovas, J. Levine, A. Gayoso, L. Mazutis, and D. Pe'er. "Characterization of cell fate probabilities in single-cell data with Palantir". en. *Nat. Biotechnol.* 37:4, 2019, pp. 451–460.

81. R. Argelaguet, A. S. E. Cuomo, O. Stegle, and J. C. Marioni. "Computational principles and challenges in single-cell data integration". en. *Nat. Biotechnol.*, 2021.

82. A. Tong, J. Huang, G. Wolf, D. van Dijk, and S. Krishnaswamy. "TrajectoryNet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics". en. *Proc Mach Learn Res* 119, 2020, pp. 9526–9536.

83. M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and F. J. Theis. "Benchmarking atlas-level data integration in single-cell genomics". en. *Nat. Methods* 19:1, 2022, pp. 41–50.

84. C. Ahlmann-Eltze and W. Huber. "Comparison of transformations for single-cell RNA-seq data". en. *Nat. Methods* 20:5, 2023, pp. 665–672.

85. H. Yi, A. Plotkin, and N. Stanley. "Benchmarking differential abundance methods for finding condition-specific prototypical cells in multi-sample single-cell datasets". en. 2023.

86. J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H. Hutson, R. Hudelle, T. Qaiser, K. J. E. Matson, Q. Barraud, A. J. Levine, G. La Manno, M. A. Skinnider, and G. Courtine. "Confronting false discoveries in single-cell differential expression". en. *Nat. Commun.* 12:1, 2021, pp. 1–15.

87. F. A. Wolf, P. Angerer, and F. J. Theis. "SCANPY: large-scale single-cell gene expression data analysis". en. *Genome Biol.* 19:1, 2018, p. 15.

88. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". en. *Nat. Biotechnol.* 36:5, 2018, pp. 411–420.

89. C. McQuin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, M. Doan, L. Ding, S. M. Rafelski, D. Thirstrup, W. Wiegraebe, S. Singh, T. Becker, J. C. Caicedo, and A. E. Carpenter. "CellProfiler 3.0: Next-generation image processing for biology". en. *PLoS Biol.* 16:7, 2018, e2005970.

90. S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, and A. Kreshuk. "ilastik: interactive machine learning for (bio)image analysis". en. *Nat. Methods* 16:12, 2019, pp. 1226–1232.

91. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona. "Fiji: an open-source platform for biological-image analysis". en. *Nat. Methods* 9:7, 2012, pp. 676–682.

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

93. D. Schapiro, B. Dumitrascu, S. Vickovic, and A. Regev. "In silico tissue generation and power analysis for spatial omics". *Nature*, 2023.

94. Y. Deng, M. Bartosovic, S. Ma, D. Zhang, P. Kukanja, Y. Xiao, G. Su, Y. Liu, X. Qin, G. B. Rosoklija, A. J. Dwork, J. J. Mann, M. L. Xu, S. Halene, J. E. Craft, K. W. Leong, M. Boldrini, G. Castelo-Branco, and R. Fan. "Spatial profiling of chromatin accessibility in mouse and human tissues". en. *Nature* 609:7926, 2022, pp. 375–383.

95. Y. Deng, M. Bartosovic, P. Kukanja, D. Zhang, Y. Liu, G. Su, A. Enninful, Z. Bai, G. Castelo-Branco, and R. Fan. "Spatial-CUT&Tag: Spatially resolved chromatin modification profiling at the cellular level". *Science* 375:6581, 2022, pp. 681–686.

96. A. Andersson, J. Bergenstråhle, M. Asp, L. Bergenstråhle, A. Jurek, J. Fernández Navarro, and J. Lundeberg. "Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography". en. *Commun Biol* 3:1, 2020, p. 565.

97. V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, R. Elmentaite, A. Lomakin, V. Kedlian, A. Gayoso, M. S. Jain, J. S. Park, L. Ramona, E. Tuck, A. Arutyunyan, R. Vento-Tormo, M. Gerstung, L. James, O. Stegle, and O. A. Bayraktar. "Cell2location maps fine-grained cell types in spatial transcriptomics". en. *Nat. Biotechnol.* 40:5, 2022, pp. 661–671.

98. T. Biancalani, G. Scalia, L. Buffoni, R. Avasthi, Z. Lu, A. Sanger, N. Tokcan, C. R. Vanderburg, Å. Segerstolpe, M. Zhang, I. Avraham-Davidi, S. Vickovic, M. Nitzan, S. Ma, A. Subramanian, M. Lipinski, J. Buenrostro, N. B. Brown, D. Fanelli, X. Zhuang, E. Z. Macosko, and A. Regev. "Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram". en. *Nat. Methods* 18:11, 2021, pp. 1352–1362.

99. R. Lopez, A. Nazaret, M. Langevin, J. Samaran, J. Regier, M. I. Jordan, and N. Yosef. "A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements", 2019. arXiv: 1905.02269 [cs.LG].

100. D. Klein, G. Palla, M. Lange, M. Klein, Z. Piran, M. Gander, L. Meng-Papaxanthos, M. Sterr, A. Bastidas-Ponce, M. Tarquis-Medina, H. Lickert, M. Bakhti, M. Nitzan, M. Cuturi, and F. J. Theis. "Mapping cells through time and space with moscot". en. 2023.

101. J. V. Hajnal and D. L. G. Hill. *Medical Image Registration*. en. CRC Press, 2019.

102. X. Huang, G. Mei, J. Zhang, and R. Abbas. "A comprehensive survey on point cloud registration", 2021. arXiv: 2103.02690 [cs.CV].

103. R. Zeira, M. Land, A. Strzalkowski, and B. J. Raphael. "Alignment and integration of spatial transcriptomics data". en. *Nat. Methods* 19:5, 2022, pp. 567–575.

104. X. Liu, R. Zeira, and B. J. Raphael. "PASTE2: Partial Alignment of Multi-slice Spatially Resolved Transcriptomics Data". en. *bioRxiv*, 2023.

105. K. Clifton, M. Anant, O. K. Aimiuwu, J. M. Kebschull, M. I. Miller, D. Tward, and J. Fan. "Alignment of spatial transcriptomics data using diffeomorphic metric mapping". en. *bioRxiv*, 2023.

106. M. Hao, K. Hua, and X. Zhang. "SOMDE: A scalable method for identifying spatially variable genes with self-organizing map". en. *Bioinformatics*, 2021.

107. I. Kats, R. Vento-Tormo, and O. Stegle. "SpatialDE2: Fast and localized variance component analysis of spatial transcriptomics". en. 2021.

108. V. Svensson, S. A. Teichmann, and O. Stegle. "SpatialDE: identification of spatially variable genes". en. *Nat. Methods* 15:5, 2018, pp. 343–346.

109. A. Anderson and J. Lundeberg. "sepal: Identifying Transcript Profiles with Spatial Patterns by Diffusion-based Modeling". en. *Bioinformatics*, 2021.

110. J. Hu, X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara, and M. Li. "SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network". en. *Nat. Methods* 18:11, 2021, pp. 1342–1351.

111. E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. B. Taylor, P. Nghiem, J. H. Bielas, and R. Gottardo. "Spatial transcriptomics at subspot resolution with BayesSpace". en. *Nat. Biotechnol.*, 2021.

112. F. William Townes and B. E. Engelhardt. "Nonnegative spatial factorization", 2021. arXiv: 2110.06122 [stat.ME].

113. B. Velten, J. M. Braunger, D. Arnol, R. Argelaguet, and O. Stegle. "Identifying temporal and spatial patterns of variation from multi-modal data using MEFISTO". en. 2020.

114. F. C. Walter, O. Stegle, and B. Velten. "FISHFactor: A Probabilistic Factor Model for Spatial Transcriptomics Data with Subcellular Resolution". en. 2021.

115. M. Elosua, P. Nieto, E. Mereu, I. Gut, and H. Heyn. "SPOTlight: Seeded NMF regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes". en. 2020.

116. X. Qiu, D. Y. Zhu, J. Yao, Z. Jing, L. Zuo, M. Wang, K. H. (Min, H. Pan, S. Wang, S. Liao, Y. Lai, S. Hao, Y. R. Lu, M. Hill, J. D. Martin-Rufino, C. Weng, A. M. Riera-Escandell, M. Chen, L. Wu, Y. Zhang, X. Wei, M. Li, X. Huang, R. Xiang, Z. Yang, C. Liu, T. Xia, Y. Liang, J. Xu, Q. Hu, Y. Hu, H. Zhu, Y. Li, A. Chen, M. A. Esteban, Y. Gu, D. A. Lauffenburger, X. Xu, L. Liu, J. S. Weissman, S. Liu, and Y. Bai. "Spateo: multidimensional spatiotemporal modeling of single-cell spatial transcriptomics". en. 2022.

117. K. H. Vining and D. J. Mooney. "Mechanical forces direct stem cell behaviour in development and regeneration". en. *Nat. Rev. Mol. Cell Biol.* 18:12, 2017, pp. 728–742.

118. M. Varrone, D. Tavernari, A. Santamaria-Martınez, L. A. Walsh, and G. Ciriello. "CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity". en. *Nat. Genet.*, 2023.

119. L. Bergenstråhle, B. He, J. Bergenstråhle, X. Abalo, R. Mirzazadeh, K. Thrane, A. L. Ji, A. Andersson, L. Larsson, N. Stakenborg, G. Boeckxstaens, P. Khavari, J. Zou, J. Lundeberg, and J. Maaskola. "Super-resolved spatial transcriptomics by deep data fusion". en. *Nat. Biotechnol.*, 2021.

120. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning Transferable Visual Models From Natural Language Supervision", 2021. arXiv: 2103.00020 [cs.CV].

121. A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. "Zero-Shot Text-to-Image Generation". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8821–8831.

122. P. Esser, R. Rombach, and B. Ommer. "Taming Transformers for High-Resolution Image Synthesis", 2020. arXiv: 2012.09841 [cs.CV].

123. L. Moses, K. Jackson, L. Luebbert, and L. Pachter. *Voyager: From geospatial to spatial omics*. 2023.

124. L. Marconato, G. Palla, K. A. Yamauchi, I. Virshup, E. Heidari, T. Treis, M. Toth, R. B. Shrestha, H. Vöhringer, W. Huber, M. Gerstung, J. Moore, F. J. Theis, and O. Stegle. "SpatialData: an open and universal data framework for spatial omics". en. 2023.

125. M. S. Keller, I. Gold, C. McCallum, T. Manz, P. V. Kharchenko, and N. Gehlenborg. *Vitessce: a framework for integrative visualization of multi-modal and spatially-resolved single-cell data*. 2021.

126. L. Sikkema, C. Ramırez-Suástegui, D. C. Strobl, T. E. Gillett, L. Zappia, E. Madissoon, N. S. Markov, L.-E. Zaragosi, Y. Ji, M. Ansari, M.-J. Arguel, L. Apperloo, M. Banchero, C. Bécavin, M. Berg, E. Chichelnitskiy, M.-I. Chung, A. Collin, A. C. A. Gay, J. Gote-Schniering, B. Hooshiar Kashani, K. Inecik, M. Jain, T. S. Kapellos, T. M. Kole, S. Leroy, C. H. Mayr, A. J. Oliver, M. von Papen, L. Peter, C. J. Taylor, T. Walzthoeni, C. Xu, L. T. Bui, C. De Donno, L. Dony, A. Faiz, M. Guo, A. J. Gutierrez, L. Heumos, N. Huang, I. L. Ibarra, N. D. Jackson, P. Kadur Lakshminarasimha Murthy, M. Lotfollahi, T. Tabib, C. Talavera-López, K. J. Travaglini, A. Wilbrey-Clark, K. B. Worlock, M. Yoshida, Lung Biological Network Consortium, M. van den Berge, Y. Bossé, T. J. Desai, O. Eickelberg, N. Kaminski, M. A. Krasnow, R. Lafyatis, M. Z. Nikolic, J. E. Powell, J. Rajagopal, M. Rojas, O. Rozenblatt-Rosen, M. A. Seibold, D. Sheppard, D. P. Shepherd, D. D. Sin, W. Timens, A. M. Tsankov, J. Whitsett, Y. Xu, N. E. Banovich, P. Barbry, T. E. Duong, C. S. Falk, K. B. Meyer, J. A. Kropski, D. Pe'er, H. B. Schiller, P. R. Tata, J. L. Schultze, S. A. Teichmann, A. V. Misharin, M. C. Nawijn, M. D. Luecken, and F. J. Theis. "An integrated cell atlas of the lung in health and disease". en. *Nat. Med.* 29:6, 2023, pp. 1563–1577.

127. G. Eraslan, E. Drokhlyansky, S. Anand, E. Fiskin, A. Subramanian, M. Slyper, J. Wang, N. Van Wittenberghe, J. M. Rouhana, J. Waldman, O. Ashenberg, M. Lek, D. Dionne, T. S. Win, M. S. Cuoco, O. Kuksenko, A. M. Tsankov, P. A. Branton, J. L. Marshall, A. Greka, G. Getz, A. V. Segrè, F. Aguet, O. Rozenblatt-Rosen, K. G. Ardlie, and A. Regev. "Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function". en. *Science* 376:6594, 2022, eabl4290.

128. K. Siletti, R. Hodge, A. Mossi Albiach, K. W. Lee, S.-L. Ding, L. Hu, P. Lönnerberg, T. Bakken, T. Casper, M. Clark, N. Dee, J. Gloe, D. Hirschstein, N. V. Shapovalova, C. D. Keene, J. Nyhus, H. Tung, A. M. Yanny, E. Arenas, E. S. Lein, and S. Linnarsson. "Transcriptomic diversity of cell types across the adult human brain". en. *Science* 382:6667, 2023, eadd7046.

129. J. E. Rood, T. Stuart, S. Ghazanfar, T. Biancalani, E. Fisher, A. Butler, A. Hupalowska, L. Gaffney, W. Mauck, G. Eraslan, J. C. Marioni, A. Regev, and R. Satija. "Toward a Common Coordinate Framework for the Human Body". en. *Cell* 179:7, 2019, pp. 1455–1467.

130. M. Haniffa, D. Taylor, S. Linnarsson, B. J. Aronow, G. D. Bader, R. A. Barker, P. G. Camara, J. G. Camp, A. Chédotal, A. Copp, H. C. Etchevers, P. Giacobini, B. Göttgens, G. Guo, A. Hupalowska, K. R. James, E. Kirby, A. Kriegstein, J. Lundeberg, J. C. Marioni, K. B. Meyer, K. K. Niakan, M. Nilsson, B. Olabi, D. Pe'er, A. Regev, J. Rood, O. Rozenblatt-Rosen, R. Satija, S. A. Teichmann, B. Treutlein, R. Vento-Tormo, S. Webb, and Human Cell Atlas Developmental Biological Network. "A roadmap for the Human Developmental Cell Atlas". en. *Nature* 597:7875, 2021, pp. 196–205.

131. B. W. Herr 2nd, J. Hardi, E. M. Quardokus, A. Bueckle, L. Chen, F. Wang, A. R. Caron, D. Osumi-Sutherland, M. A. Musen, and K. Börner. "Specimen, biological structure, and spatial ontologies in support of a Human Reference Atlas". en. *Sci Data* 10:1, 2023, p. 171.

132. K. Börner, S. A. Teichmann, E. M. Quardokus, J. C. Gee, K. Browne, D. Osumi-Sutherland, B. W. Herr 2nd, A. Bueckle, H. Paul, M. Haniffa, L. Jardine, A. Bernard, S.-L. Ding, J. A. Miller, S. Lin, M. K. Halushka, A. Boppana, T. A. Longacre, J. Hickey, Y. Lin, M. T. Valerius, Y. He, G. Pryhuber, X. Sun, M. Jorgensen, A. J. Radtke, C. Wasserfall, F. Ginty, J. Ho, J. Sunshine, R. T. Beuschel, M. Brusko, S. Lee, R. Malhotra, S. Jain, and G. Weber. "Anatomical structures, cell types and biomarkers of the Human Reference Atlas". en. *Nat. Cell Biol.* 23:11, 2021, pp. 1117–1128.

133. S. W. Oh, J. A. Harris, L. Ng, B. Winslow, N. Cain, S. Mihalas, Q. Wang, C. Lau, L. Kuan, A. M. Henry, M. T. Mortrud, B. Ouellette, T. N. Nguyen, S. A. Sorensen, C. R. Slaughterbeck, W. Wakeman, Y. Li, D. Feng, A. Ho, E. Nicholas, K. E. Hirokawa, P. Bohn, K. M. Joines, H. Peng, M. J. Hawrylycz, J. W. Phillips, J. G. Hohmann, P. Wohnoutka, C. R. Gerfen, C. Koch, A. Bernard, C. Dang, A. R. Jones, and H. Zeng. "A mesoscale connectome of the mouse brain". en. *Nature* 508:7495, 2014, pp. 207–214.

134. Q. Wang, S.-L. Ding, Y. Li, J. Royall, D. Feng, P. Lesnar, N. Graddis, M. Naeemi, B. Facer, A. Ho, T. Dolbeare, B. Blanchard, N. Dee, W. Wakeman, K. E. Hirokawa, A. Szafer, S. M. Sunkin, S. W. Oh, A. Bernard, J. W. Phillips, M. Hawrylycz, C. Koch, H. Zeng, J. A. Harris, and L. Ng. "The Allen Mouse Brain Common Coordinate Framework: A 3D reference atlas". en. *Cell* 181:4, 2020, 936–953.e20.

135. S. Ghose, Y. Ju, E. McDonough, J. Ho, A. Karunamurthy, C. Chadwick, S. Cho, R. Rose, A. Corwin, C. Surrette, J. Martinez, E. Williams, A. Sood, Y. Al-Kofahi, L. D. Falo, K. Börner, and F. Ginty. "3D reconstruction of skin and spatial mapping of immune cell density, vascular distance and effects of sun exposure and aging". en. *Communications Biology* 6:1, 2023, pp. 1–14.

136. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. "Twelve years of SAMtools and BCFtools". en. *Gigascience* 10:2, 2021, giab008.

137. M.-A. Bray, S. Singh, H. Han, C. T. Davis, B. Borgeson, C. Hartland, M. Kost-Alimova, S. M. Gustafsdottir, C. C. Gibson, and A. E. Carpenter. "Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes". en. *Nat. Protoc.* 11:9, 2016, pp. 1757–1774.

138. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. "Salmon provides fast and bias-aware quantification of transcript expression". en. *Nat. Methods* 14:4, 2017, pp. 417–419.

139. M. Sorin, M. Rezanejad, E. Karimi, B. Fiset, L. Desharnais, L. J. M. Perus, S. Milette, M. W. Yu, S. M. Maritan, S. Doré, É. Pichette, W. Enlow, A. Gagné, Y. Wei, M. Orain, V. S. K. Manem, R. Rayes, P. M. Siegel, S. Camilleri-Broët, P. O. Fiset, P. Desmeules, J. D. Spicer, D. F. Quail, P. Joubert, and L. A. Walsh. "Single-cell spatial landscapes of the lung tumour immune microenvironment". en. *Nature* 614:7948, 2023, pp. 548–554.

140. M. Dhainaut, S. A. Rose, G. Akturk, A. Wroblewska, S. R. Nielsen, E. S. Park, M. Buckup, V. Roudko, L. Pia, R. Sweeney, J. Le Berichel, C. M. Wilk, A. Bektesevic, B. H. Lee, N. Bhardwaj, A. H. Rahman, A. Baccarini, S. Gnjatic, D. Pe'er, M. Merad, and B. D. Brown. "Spatial CRISPR genomics identifies regulators of the tumor microenvironment". en. *Cell* 185:7, 2022, 1223–1239.e20.

141. D. Feldman, A. Singh, J. L. Schmid-Burgk, R. J. Carlson, A. Mezger, A. J. Garrity, F. Zhang, and P. C. Blainey. "Optical Pooled Screens in Human Cells". en. *Cell* 179:3, 2019, 787–799.e17.

142. T. Kudo, A. M. Meireles, R. Moncada, Y. Chen, P. Wu, J. Gould, X. Hu, O. Kornfeld, R. Jesudason, C. Foo, B. Höckendorf, H. C. Bravo, J. P. Town, R. Wei, A. Rios, V. Chandrasekar, M. Heinlein, S. Cai, C. S. Lu, C. Celen, N. Kljavin, J. Jiang, J. S. Hleap, N. Kayagaki, F. de Sousa e Melo, L. McGinnis, B. Li, A. Singh, L. Garraway, O. Rozenblatt-Rosen, A. Regev, and E. Lubeck. "Highly multiplexed, image-based pooled screens in primary cells and tissues with PerturbView". en. 2023.

143. R. R. Stickels, E. Murray, P. Kumar, J. Li, J. L. Marshall, D. J. Di Bella, P. Arlotta, E. Z. Macosko, and F. Chen. "Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2". *Nat. Biotechnol.*, 2020.

144. S. Vickovic, G. Eraslan, F. Salmén, J. Klughammer, L. Stenbeck, D. Schapiro, T. Äijö, R. Bonneau, L. Bergenstråhle, J. F. Navarro, J. Gould, G. K. Griffin, Å. Borg, M. Ronaghi, J. Frisén, J. Lundeberg, A. Regev, and P. L. Ståhl. "High-definition spatial transcriptomics for in situ tissue profiling". en. *Nat. Methods* 16:10, 2019, pp. 987–990.

145. P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, Å. Borg, F. Pontén, P. I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, and J. Frisén. "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". en. *Science* 353:6294, 2016, pp. 78–82.

146. J. Langlieb, N. S. Sachdev, K. S. Balderrama, N. M. Nadaf, M. Raj, E. Murray, J. T. Webber, C. Vanderburg, V. Gazestani, D. Tward, C. Mezias, X. Li, D. M. Cable, T. Norton, P. Mitra, F. Chen, and E. Z. Macosko. "The cell type composition of the adult mouse brain revealed by single cell and spatial genomics". en. 2023.

147. W.-T. Chen, A. Lu, K. Craessaerts, B. Pavie, C. Sala Frigerio, N. Corthout, X. Qian, J. Laláková, M. Kühnemund, I. Voytyuk, L. Wolfs, R. Mancuso, E. Salta, S. Balusu, A. Snellinx, S. Munck, A. Jurek, J. Fernandez Navarro, T. C. Saido, I. Huitinga, J. Lundeberg, M. Fiers, and B. De Strooper. "Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease". en. *Cell* 182:4, 2020, 976–991.e19.

148. X. Wang, W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C. Ramakrishnan, J. Liu, G. P. Nolan, F.-A. Bava, and K. Deisseroth. "Three-dimensional intact-tissue sequencing of single-cell transcriptional states". en. *Science* 361:6400, 2018.

149. H. Zeng, J. Huang, H. Zhou, W. J. Meilandt, B. Dejanovic, Y. Zhou, C. J. Bohlen, S.-H. Lee, J. Ren, A. Liu, Z. Tang, H. Sheng, J. Liu, M. Sheng, and X. Wang. "Integrative in situ mapping of single-cell transcriptional states and tissue histopathology in a mouse model of Alzheimer's disease". en. *Nat. Neurosci.* 26:3, 2023, pp. 430–446.

150. G. Gut, M. D. Herrmann, and L. Pelkmans. "Multiplexed protein maps link subcellular organization to cellular states". en. *Science* 361:6401, 2018.

151. J.-R. Lin, B. Izar, S. Wang, C. Yapp, S. Mei, P. M. Shah, S. Santagata, and P. K. Sorger. "Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes". en. *Elife* 7, 2018.

152. D. Schapiro, H. W. Jackson, S. Raghuraman, J. R. Fischer, V. R. T. Zanotelli, D. Schulz, C. Giesen, R. Catena, Z. Varga, and B. Bodenmiller. "histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data". en. *Nat. Methods* 14:9, 2017, pp. 873–876.

153. A. Baddeley, E. Rubak, and R. Turner. *Spatial Point Patterns: Methodology and Applications with R*. en. CRC Press, 2015.

154. L. Tosti, Y. Hang, O. Debnath, S. Tiesmeyer, T. Trefzer, K. Steiger, F. W. Ten, S. Lukassen, S. Ballke, A. A. Kühl, S. Spieckermann, R. Bottino, N. Ishaque, W. Weichert, S. K. Kim, R. Eils, and C. Conrad. "Single-Nucleus and In Situ RNA-Sequencing Reveal Cell Topographies in the Human Pancreas". en. *Gastroenterology* 160:4, 2021, 1330–1344.e11.

155. A. Getis and J. K. Ord. "The analysis of spatial association by use of distance statistics". en. *Geogr. Anal.* 24:3, 2010, pp. 189–206.

156. *Zarr core specification (version 3.0) — Zarr specs documentation*. en. https://zarr-specs.readthedocs.io/en/latest/v3/core/v3.0.html. Accessed: 2023-9-25.

157. Wikipedia contributors. *Apache Parquet*. https://en.wikipedia.org/w/index.php?title=Apache_Parquet&oldid=1173786069. Accessed: NA-NA-NA. 2023.

158. S. Hoyer and J. J. Hamman. "xarray: N-D labeled Arrays and Datasets in Python". en. *J. Open Res. Softw.* 5, 2017.

159. *datatree: WIP implementation of a tree-like hierarchical data structure for xarray*. en.

160. Dask Development Team. *Dask: Library for dynamic task scheduling*. 2016.

161. K. Jordahl, J. Van den Bossche, M. Fleischmann, J. Wasserman, J. McBride, J. Gerard, J. Tratner, M. Perry, A. G. Badaracco, C. Farmer, G. A. Hjelle, A. D. Snow, M. Cochran, S. Gillies, L. Culbertson, M. Bartos, N. Eubank, maxalbert, A. Bilogur, S. Rey, C. Ren, D. Arribas-Bel, L. Wasser, L. J. Wolf, M. Journois, J. Wilson, A. Greenhall, C. Holdgraf, Filipe, and F. Leblanc. *geopandas/geopandas: v0.8.1*. 2020.

162. I. Virshup, S. Rybakov, F. J. Theis, P. Angerer, and F. Alexander Wolf. "anndata: Annotated data". en. 2021.

163. J. Moore, D. Basurto-Lozada, S. Besson, J. Bogovic, J. Bragantini, E. M. Brown, J.-M. Burel, X. Casas Moreno, G. de Medeiros, E. E. Diel, D. Gault, S. S. Ghosh, I. Gold, Y. O. Halchenko, M. Hartley, D. Horsfall, M. S. Keller, M. Kittisopikul, G. Kovacs, A. Küpcü Yoldaş, K. Kyoda, A. le Tournoulx de la Villegeorges, T. Li, P. Liberali, D. Lindner, M. Linkert, J. Lüthi, J. Maitin-Shepard, T. Manz, L. Marconato, M. McCormick, M. Lange, K. Mohamed, W. Moore, N. Norlin, W. Ouyang, B. Özdemir, G. Palla, C. Pape, L. Pelkmans, T. Pietzsch, S. Preibisch, M. Prete, N. Rzepka, S. Samee, N. Schaub, H. Sidky, A. C. Solak, D. R. Stirling, J. Striebel, C. Tischer, D. Toloudis, I. Virshup, P. Walczysko, A. M. Watson, E. Weisbart, F. Wong, K. A. Yamauchi, O. Bayraktar, B. A. Cimini, N. Gehlenborg, M. Haniffa, N. Hotaling, S. Onami, L. A. Royer, S. Saalfeld, O. Stegle, F. J. Theis, and J. R. Swedlow. "OME-Zarr: a cloud-optimized bioimaging file format with international community support". *Histochem. Cell Biol.* 160:3, 2023, pp. 223–251.

164. *Napari — napari*. en. https://napari.org/. Accessed: 2022-7-23.

165. Hunter. "Matplotlib: A 2D Graphics Environment". 9, 2007, pp. 90–95.

166. I. Virshup, D. Bredikhin, L. Heumos, G. Palla, G. Sturm, A. Gayoso, I. Kats, M. Koutrouli, Scverse Community, B. Berger, D. Pe'er, A. Regev, S. A. Teichmann, F. Finotello, F. A. Wolf, N. Yosef, O. Stegle, and F. J. Theis. "The scverse project provides a computational ecosystem for single-cell omics data analysis". en. *Nat. Biotechnol.*, 2023.

167. C. Villani. *Topics in Optimal Transportation*. en. American Mathematical Soc., 2021.

168. G. Peyré and M. Cuturi. "Computational Optimal Transport", 2018. arXiv: 1803.00567 [stat.ML].

169. M. Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances", 2013. arXiv: 1306.0895 [stat.ML].

170. V. Titouan, N. Courty, R. Tavenard, C. Laetitia, and R. Flamary. "Optimal Transport for structured data with application on graphs". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6275–6284. URL: https://proceedings.mlr.press/v97/titouan19a.html.

171. F. Mémoli. "Gromov–Wasserstein Distances and the Metric Approach to Object Matching". *Found. Comut. Math.* 11:4, 2011, pp. 417–487.

172. K.-T. Sturm. *The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces*. 2020. arXiv: 1208.0434 [math.MG].

173. G. Peyré, M. Cuturi, and J. Solomon. *Gromov-Wasserstein averaging of kernel and distance matrices*. http://proceedings.mlr.press/v48/peyre16.pdf. Accessed: 2022-5-29.

174. M. Scetbon, M. Cuturi, and G. Peyré. "Low-Rank Sinkhorn Factorization", 2021. arXiv: 2103.04737 [stat.ML].

175. M. Scetbon, G. Peyré, and M. Cuturi. "Linear-time Gromov Wasserstein distances using low rank couplings and costs". Proceedings of Machine Learning Research, 2021. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, pp. 19347–19365. arXiv: 2106.01128 [cs.LG].

176. M. Scetbon and M. Cuturi. "Low-rank optimal transport: Approximation, statistics and debiasing". *Adv. Neural Inf. Process. Syst.* 35, 2022, pp. 6802–6814.

177. M. Scetbon, M. Klein, G. Palla, and M. Cuturi. "Unbalanced Low-rank Optimal Transport Solvers", 2023. arXiv: 2305.19727 [cs.LG].

178. D. P. Cook, K. B. Jensen, K. Wise, M. J. Roach, F. S. Dezem, N. K. Ryan, M. Zamojski, I. S. Vlachos, S. R. V. Knott, L. M. Butler, J. L. Wrana, N. E. Banovich, J. T. Plummer, and L. G. Martelotto. "A Comparative Analysis of Imaging-Based Spatial Transcriptomics Platforms". en. 2023.

179. H. Wang, R. Huang, J. Nelson, C. Gao, M. Tran, A. Yeaton, K. Felt, K. L. Pfaff, T. Bowman, S. J. Rodig, K. Wei, B. A. Goods, and S. L. Farhi. "Systematic benchmarking of imaging spatial transcriptomics platforms in FFPE tissues". en. *bioRxiv*, 2023.

180. S. M. Salas, P. Czarnewski, L. B. Kuemmerle, S. Helgadottir, C. Matsson-Langseth, S. Tismeyer, C. Avenel, H. Rehman, K. Tiklova, A. Andersson, M. Chatzinikolaou, F. J. Theis, M. D. Luecken, C. Wählby, N. Ishaque, and M. Nilsson. "Optimizing Xenium In Situ data utility by quality assessment and best practice analysis workflows". en. 2023.

181. R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson, Y. Wang, S. Zhao, A. B. Fogo, H. Yang, Y. Tang, and Y. Huo. "Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging", 2023. arXiv: 2304.04155 [eess.IV].

182. V. Petukhov, K. Khodosevich, R. A. Soldatov, and P. V. Kharchenko. "Bayesian segmentation of spatially resolved transcriptomics data". en. 2020.

183. L. Atta, K. Clifton, M. Anant, and J. Fan. "Gene count normalization in single-cell imaging-based spatially resolved transcriptomics". en. *bioRxiv*, 2023.

184. T. Lohoff, S. Ghazanfar, A. Missarova, N. Koulena, N. Pierson, J. A. Griffiths, E. S. Bardot, C.-H. L. Eng, R. C. V. Tyser, R. Argelaguet, C. Guibentif, S. Srinivas, J. Briscoe, B. D. Simons, A.-K. Hadjantonakis, B. Göttgens, W. Reik, J. Nichols, L. Cai, and J. C. Marioni. "Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis". en. *Nat. Biotechnol.*, 2021.

185. C. Gong, S. Li, L. Wang, F. Zhao, S. Fang, D. Yuan, Z. Zhao, Q. He, M. Li, W. Liu, Z. Li, H. Xie, S. Liao, A. Chen, Y. Zhang, Y. Li, and X. Xu. "SAW: An efficient and accurate data analysis workflow for Stereo-seq spatial transcriptomics". en. 2023.

186. R. Lopez, B. Li, H. Keren-Shaul, P. Boyeau, M. Kedmi, D. Pilzer, A. Jelinski, I. Yofe, E. David, A. Wagner, C. Ergen, Y. Addadi, O. Golani, F. Ronchese, M. I. Jordan, I. Amit, and N. Yosef. "DestVI identifies continuums of cell types in spatial transcriptomics data". en. *Nat. Biotechnol.*, 2022.

187. S. Birk, I. Bonafonte-Pardàs, A. M. Feriz, A. Boxall, E. Agirre, F. Memi, A. Maguza, R. Fan, G. Castelo-Branco, O. A. Bayraktar, F. J. Theis, C. Talavera-López, and M. Lotfollahi. "Large-scale characterization of cell niches in spatial atlases using bio-inspired graph learning". en. 2024.

188. E. A. G. Baker, D. Schapiro, B. Dumitrascu, S. Vickovic, and A. Regev. "In silico tissue generation and power analysis for spatial omics". en. *Nat. Methods* 20:3, 2023, pp. 424–431.

189. D. M. Cable, E. Murray, V. Shanmugam, S. Zhang, L. S. Zou, M. Diao, H. Chen, E. Z. Macosko, R. A. Irizarry, and F. Chen. "Cell type-specific inference of differential expression in spatial transcriptomics". en. *Nat. Methods* 19:9, 2022, pp. 1076–1087.

190. M. Ali, M. Kuijs, S. Hediyeh-zadeh, T. Treis, K. Hrovatin, G. Palla, A. C. Schaar, and F. J. Theis. "GraphCompass: Spatial metrics for differential analyses of cell organization across conditions". en. 2024.

191. G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander. "Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming". en. *Cell* 176:4, 2019, 928–943.e22.

192. C. Xia, J. Fan, G. Emanuel, J. Hao, and X. Zhuang. "Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression". en. *Proc. Natl. Acad. Sci. U. S. A.* 116:39, 2019, pp. 19490–19499.

193. M. Y. Lu, B. Chen, A. Zhang, D. F. K. Williamson, R. J. Chen, T. Ding, L. P. Le, Y.-S. Chuang, and F. Mahmood. "Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images", 2023. arXiv: 2306.07831 [cs.CV].

194. S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, M. P. Lungren, T. Naumann, and H. Poon. "Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing", 2023. arXiv: 2303.00915 [cs.CV].

195. J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. "Flamingo: a Visual Language Model for Few-Shot Learning", 2022. arXiv: 2204.14198 [cs.CV].

196. M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar. "Med-Flamingo: a Multimodal Medical Few-shot Learner". In: *Proceedings of the 3rd Machine Learning for Health Symposium*. Ed. by S. Hegselmann, A. Parziale, D. Shanmugam, S. Tang, M. N. Asiedu, S. Chang, T. Hartvigsen, and H. Singh. Vol. 225. Proceedings of Machine Learning Research. PMLR, 2023, pp. 353–367.

197. P. Dhariwal and A. Nichol. "Diffusion Models Beat GANs on Image Synthesis", 2021. arXiv: 2105.05233 [cs.LG].

198. J. Ho and T. Salimans. "Classifier-Free Diffusion Guidance", 2022. arXiv: 2207.12598 [cs.LG].

199. S. N. Chandrasekaran, J. Ackerman, E. Alix, D. Michael Ando, J. Arevalo, M. Bennion, N. Boisseau, A. Borowa, J. D. Boyd, L. Brino, P. J. Byrne, H. Ceulemans, C. Ch'ng, B. A. Cimini, D.-A. Clevert, N. Deflaux, J. G. Doench, T. Dorval, R. Doyonnas, V. Dragone, O. Engkvist, P. W. Faloon, B. Fritchman, F. Fuchs, S. Garg, T. J. Gilbert, D. Glazer, D. Gnutt, A. Goodale, J. Grignard, J. Guenther, Y. Han, Z. Hanifehlou, S. Hariharan, D. Hernandez, S. R. Horman, G. Hormel, M. Huntley, I. Icke, M. Iida, C. B. Jacob, S. Jaensch, J. Khetan, M. Kost-Alimova, T. Krawiec, D. Kuhn, C.-H. Lardeau, A. Lembke, F. Lin, K. D. Little, K. R. Lofstrom, S. Lotfi, D. J. Logan, Y. Luo, F. Madoux, P. A. Marin Zapata, B. A. Marion, G. Martin, N. J. McCarthy, L. Mervin, L. Miller, H. Mohamed, T. Monteverde, E. Mouchet, B. Nicke, A. Ogier, A.-L. Ong, M. Osterland, M. Otrocka, P. J. Peeters, J. Pilling, S. Prechtl, C. Qian, K. Rataj, D. E. Root, S. K. Sakata, S. Scrace, H. Shimizu, D. Simon, P. Sommer, C. Spruiell, I. Sumia, S. E. Swalley, H. Terauchi, A. Thibaudeau, A. Unruh, J. Van de Waeter, M. Van Dyck, C. van Staden, M. Warchoł, E. Weisbart, A. Weiss, N. Wiest-Daessle, G. Williams, S. Yu, B. Zapiec, M. Żyła, S. Singh, and A. E. Carpenter. "JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations". en. 2023.

200. M. Lotfollahi, A. Klimovskaia Susmelj, C. De Donno, L. Hetzel, Y. Ji, I. L. Ibarra, S. R. Srivatsan, M. Naghipourfar, R. M. Daza, B. Martin, J. Shendure, J. L. McFaline-Figueroa, P. Boyeau, F. A. Wolf, N. Yakubova, S. Günnemann, C. Trapnell, D. Lopez-Paz, and F. J. Theis. "Predicting cellular responses to complex perturbations in high-throughput screens". en. *Mol. Syst. Biol.* 19:6, 2023, e11517.

201. A. Tejada-Lapuerta, P. Bertin, S. Bauer, H. Aliee, Y. Bengio, and F. J. Theis. "Causal machine learning for single-cell genomics", 2023. arXiv: 2310.14935 [cs.LG].

202. M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. van de Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, C. Abajian, C. F. Beckmann, A. Bernard, D. Bertagnolli, A. F. Boe, P. M. Cartagena, M. M. Chakravarty, M. Chapin, J. Chong, R. A. Dalley, B. David Daly, C. Dang, S. Datta, N. Dee, T. A. Dolbeare, V. Faber, D. Feng, D. R. Fowler, J. Goldy, B. W. Gregor, Z. Haradon, D. R. Haynor, J. G. Hohmann, S. Horvath, R. E. Howard, A. Jeromin, J. M. Jochim, M. Kinnunen, C. Lau, E. T. Lazarz, C. Lee, T. A. Lemon, L. Li, Y. Li, J. A. Morris, C. C. Overly, P. D. Parker, S. E. Parry, M. Reding, J. J. Royall, J. Schulkin, P. A. Sequeira, C. R. Slaughterbeck, S. C. Smith, A. J. Sodt, S. M. Sunkin, B. E. Swanson, M. P. Vawter, D. Williams, P. Wohnoutka, H. R. Zielke, D. H. Geschwind, P. R. Hof, S. M. Smith, C. Koch, S. G. N. Grant, and A. R.

Jones. "An anatomically comprehensive atlas of the adult human brain transcriptome". en. *Nature* 489:7416, 2012, pp. 391–399.

203. Y. Jain, L. L. Godwin, Y. Ju, N. Sood, E. M. Quardokus, A. Bueckle, T. Longacre, A. Horning, Y. Lin, E. D. Esplin, J. W. Hickey, M. P. Snyder, N. H. Patterson, J. M. Spraggins, and K. Börner. "Segmentation of human functional tissue units in support of a Human Reference Atlas". en. *Commun Biol* 6:1, 2023, p. 717.