Investigating computational strategies for domain identification in spatial transcriptomics

Alice Fabienne Descoeudres

München 2025

Investigating computational strategies for domain identification in spatial transcriptomics

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig–Maximilians–Universität München

vorgelegt von
Alice Fabienne Descoeudres
aus Bern, Schweiz

2025

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Stefan Canzar betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 27. Oktober 2025

.....Alice Descoeudres.....

Dissertation eingereicht am 27. Oktober 2025
Erstgutachter: Prof. Stefan Canzar
Zweitgutachter: Prof. Johanna Klughammer
Mündliche Prüfung am 12. Dezember 2025

Contents

Summary

Within the hierarchy of biological organisation, between cells and organs lies a diverse set of tissues. Soft tissues consist of elaborate cellular arrangements, comprising a variety of cell types, and form the basis of larger-scale biological function. Tissues are further organised in regions with distinct functions, morphology, and molecular composition. These regions can take the form of concentric circular patterns as in kidney glomeruli, the laminar structures of cortical brain layers, or indistinct, complex cancer infiltration. Commonly, these tissue regions are identified through sectioning the tissue and using histological stains to increase the visual contrast of morphologies of interest.

Spatial transcriptomics is a collection of technologies enabling the direct transcriptional profiling of tissue sections. This brings about an opportunity to evaluate spatial dependencies in gene expression and to consider tissue coherence on a molecular level. The number of transcripts quantified in different spatial transcriptomics approaches ranges from a few dozen to the entire transcriptome, exceeding 20'000 genes for human tissues. To handle this wealth of data, computational tools for the identification of regions based on spatial transcriptomics have been developed.

As with all analysis approaches, there is a myriad of possible avenues to arrive at the same goal. Method development for spatial domain identification has rapidly outpaced the ability of both users and tool developers to keep track of options and approaches. In this situation, unbiased, independent, and systematic method comparisons, known as benchmarking studies, are indispensable.

In this thesis, I present the setup of and results from a thorough benchmarking study of methods developed for the identification of domains in spatial transcriptomics data. The benchmark utilises public datasets from diverse technological origins, and additionally entails the creation of a custom approach for generating semi-synthetic benchmarking data. This pipeline is utilised for an extensive and systematic evaluation of the effects of technological and tissue characteristics on method performances.

Specifically, we initially benchmark 26 methods for spatial domain detection on 63 tissue slices, profiled using five different technologies across seven public datasets. First, we identify a simple consensus aggregation of method outputs as a highly stable and competitive alternative to any single method. Additionally, through detailed analyses of method performances, we form hypotheses about dataset characteristics that may affect methods in distinct ways. To enable us to systematically study the effect of these dataset characteristics, we develop an approach combining synthetic tissue locations with transcriptome profiles from a real single-nucleus dataset of a mouse brain. Using this combination, we create over 1000 samples of semi-synthetic spatial transcriptomics data, allowing us to investigate the effects of diverse technology-inherent features. Further, by different expression-level perturbations, we evaluate the effects of transcriptional domain similarity and cellular heterogeneity. Lastly, we consider how the size and shape of tissue domains affect their detection by different methods.

We evaluate method stability using a data reordering approach specifically developed to identify stochastic effects of rerunning methods on the same data. Lastly, the runtime, memory usage, and usability of methods is evaluated. All in all, the work presented in this thesis is a valuable resource for prospective method users and developers interested in the domain-based analysis of spatial transcriptomics data, highlighting where and which methods excel and pointing to potential avenues for improvement.

Chapter 1

Introduction

The research presented in this thesis spans topics from molecular biology, tissue biology, and bioinformatics. The following sections present introductions to the relevant topics from each of these areas, namely spatial transcriptomics, domain identification in tissues, and computational method benchmarking.

1.1 The transcriptome and approaches to its analysis

Since the discovery of the cell as the basic building block of life, scientists have been investigating its biochemical innards. Light microscopy enabled a deep appreciation for the diversity and quantity of subcellular structures – from larger organelles like the nucleus or the Golgi apparatus down to macromolecules of various functions. Arguably the most significant step towards understanding the fundamental mechanisms of life was the postulation of deoxyribonucleic acid (DNA) as the carrier of genetic information. Its molecular structure, the double helix, is ideally suited to the storage and proliferation of information. It is formed by two polynucleotide chains, which are held together by backbones of phosphorylated sugars and joined by hydrogen bridges connecting opposing complementary bases. The sequence of base pairs encodes information using an "alphabet" of four bases (Adenine, Guanine, Cytosine and Thymine). The inherent redundancy of information storage in DNA and evolution-honed precise copying and error-correcting procedures enable the genetic code of each living organism to be stored, copied, queried, and recombined.

However, DNA is not in itself an active agent in cell function, reproduction or communication. Long sequences of base pairs known as genes encode, essentially, building instructions. To create life, information contained in those genes needs to be extracted, interpreted, and converted into proteins, which carry out or catalyse the necessary cellular functions. This flow of information is known as the central dogma[1] of molecular biology: DNA is transcribed into ribonucleic acid (RNA) which is translated into proteins, from where no information is able to flow back into, and change, the DNA (Fig. 1.1). Thus, the division of labour within cells becomes clear: The entirety of the genetic information of any organism is *stored* in the complete set of its genes, known as its genome. Cells are able to survive and *function* thanks to a wide array of proteins carrying out specialised tasks. Between information storage and cellular functioning lies the collection of RNAs produced in each cell, *interpreting* the genome and enabling adaptation and variability.

RNA molecules, like individual DNA strands, consist of chains of nucleotides. The nucleotides are equivalent in form to those of DNA, except for the sugar deoxyribose being replaced by ribose. Additionally, the base Thymine is replaced by Uracil in RNA, keeping the complementarity to Adenine

[1] Certainly a misnomer, as Francis Crick, who coined the phrase, himself also acknowledged [1]. A dogma, as defined by the online Cambridge Dictionary, is "a fixed, especially religious, belief or set of beliefs that people are expected to accept without any doubts" – in order words, fundamentally incompatible with the modern scientific process.

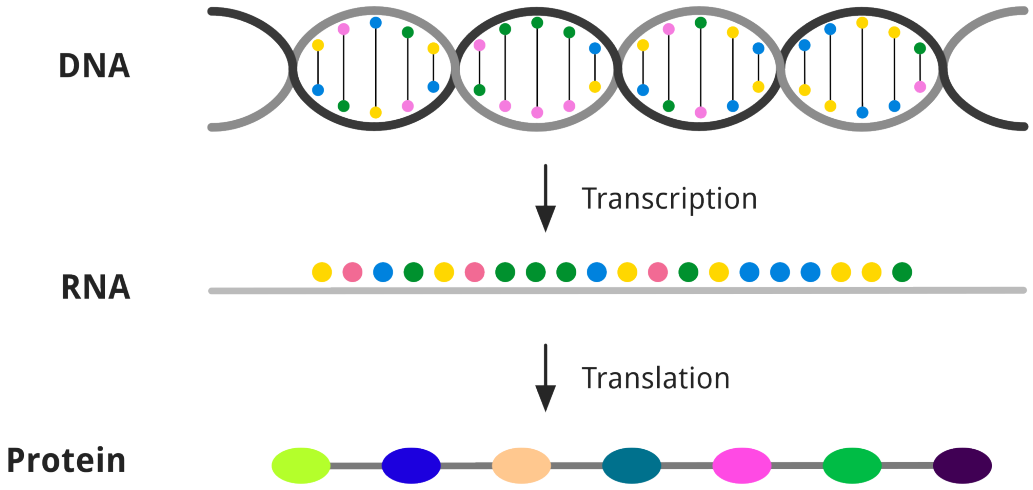

Figure 1.1: **Schematic of the central dogma of molecular biology.** DNA is transcribed into RNA, which is translated into an amino acid chain, and subsequently folded into a protein.

intact. In contrast to DNA, RNA typically occurs in single strands. There are various RNA types and functions – the most abundant being ribosomal (rRNA) or transfer (tRNA) [2]. Both rRNA and tRNA are indispensable parts of the cellular machinery in the process of protein synthesis. However, this thesis focuses exclusively on the analysis of a third type, called messenger RNA (mRNA). If rRNA and tRNA are the construction workers, mRNA is the blueprint: These are the molecules transcribed from coding regions of the genome, in whose sequence proteins are encoded.

Pre-mRNA gets transcribed from a segment of DNA by the enzyme RNA polymerase, as a nucleotide-by-nucleotide copy. The RNA polymerase recognises molecular markers (codons) encoding "transcription start" and "transcription stop", delineating a gene, or a functional fragment of genetic code. The resulting pre-mRNA gets processed further into mature mRNA by the process known as splicing: Parts of the pre-mRNA known as introns are removed[2] and the remaining pieces, called exons, are spliced back together.

Finally, proteins are synthesised from mature mRNA as chains of amino acids, each encoded in a three-nucleotide sequence termed a codon. Once the polypeptide chain is complete, proteins fold into a specific conformation that allows them to carry out their highly specialised functions[3]. The proteome, that is, the set of proteins in a given tissue or cell, is also a highly studied analysis target. A wealth of detection and measurement approaches, frequently based on antibody binding or mass spectrometry, enables detailed explorations of protein abundances and structures, and in some cases allow for sequence-based analyses[4] [6, 7]. However, even discounting challenges relating to abundance and stability, unfolding a protein for sequencing is a challenging task [4, 8]. On the other hand, RNA molecules show a simpler structure[5], and their nucleotide chains are close relatives of the well-studied DNA. The alphabet of amino acids constituting proteins and peptides contains 20 distinct characters, while only four types of nucleotides build up the information stored in RNA. For scientists attempting to decode the molecular phenotype of cells, RNA presents an attractive target. Finally, as RNA molecules can be converted to complementary DNA strands (cDNA) through reverse transcription,

[2]While it is commonly assumed that introns are subsequently degraded back into their constituent nucleotides for reuse, some introns remain stable in cells [3].

[3]Proteins may undergo post-translational modifications such as phosphorylation, potentially changing their function. This further augments the space of possible proteins that can be produced from the fixed genome, already increased by the possibilities of alternative splicing.

[4]Recently, single-molecule protein sequencing approaches have garnered interest, and notably, in 2024 spatial proteomics was pronounced method of the year by Nature Methods [4, 5].

[5]Which is not to say they do not also exhibit secondary and tertiary folding structures [9]. Notably, double-stranded RNA molecules or intrastrand double helices can form.

methods developed for DNA molecules are also applicable to RNA.

The following section gives a brief historical overview of efforts to decode RNA sequences.

1.1.1 A history of RNA sequencing

An inherent property of both RNA and DNA is their tendency to anneal, or hybridise, to complementary nucleotide chains. Using radioactive molecular probes, both Pardue and Gall [10] and John *et al.* [11] were able to localise sequences within *Xenopus* oocytes in 1969.[6] The development of fluorescently tagged probes in the 1980s [13, 14] considerably simplified both synthesis and detection [15]. Thereafter, hybridisation-based techniques were commonly used to map RNA transcripts to previously known regions in the genome [16]. DNA hybridisation microarrays, consisting of thousands of specific sequences attached to a surface, can quantify the occurrence of those sequences in a nucleic acid solution [17]. They were used to map transcripts at a very high genomic resolution, down to several base pairs, and could even detect and quantify differently spliced transcript versions, or isoforms, of mRNA [18, 19]. However, microarray-based methods rely on existing knowledge of the underlying genome. Therefore, they cannot be used to detect novel, previously unknown transcripts. Further, the possibility of cross-hybridisation leading to background noise, as well as signal saturation, combine to make microarrays a suboptimal technique for quantitative transcriptomics [19, 20].

The main early approach to DNA sequencing[7], the chain-termination method, was developed in 1977 by Sanger *et al.* and is now commonly known as Sanger sequencing [23]. Its fundamental principle, shown schematically in Fig. 1.2, is iteratively producing all incremental length sequences of a transcript and fluorescently labelling the terminal nucleotide in each [24]. The resulting fragments are sorted by size through gel electrophoresis and subsequently imaged to reveal the locations of labelled bases. This type of sequencing was used in the Human Genome Project [25, 26].

An enormous jump in sequencing throughput[8] was made possible through the development of flow cell technology, combined with sequencing-by-synthesis [19, 22]. In preparing a flow cell, template adapter oligonucleotides are affixed to a support plate. RNA transcripts to be sequenced are fragmented and reverse transcribed into cDNA strands, to which adapters complementary to those oligonucleotides are ligated. Once attached to the flow cell, this collection of molecules, the "library", is bridge-amplified through a polymerase chain reaction (PCR) procedure. This results in "islands" of clonally amplified cDNA templates. A schematic of these steps is shown in Fig. 1.3. Finally, sequencing occurs by the repeated addition of reversibly fluorescent nucleotides and imaging of each step [20, 24]. This is the most common type of short-read sequencing. Various technologies are commercially available, however, including long-read implementations that skip the fragmentation step to create reads spanning multiple splicing sites [28].

A fundamental limitation of traditional RNA-seq is its bulk nature: The original tissue processed through the RNA-seq pipeline is completely dissociated and processed as one sample. Thereby, on one hand, the diversity inherent in the different cells present in the sample is lost. The development of single-cell RNA-sequencing has revolutionised the study of individual cells and cell types, and enabled the creation of single-cell transcriptomic atlases of various organisms and tissues [29, 30].

On the other hand, in both bulk and single-cell sequencing, the spatial organisation of the sample, or the microenvironment in which single cells reside, are lost in the tissue dissociation step. This information crucially informs various biological processes, however, from cell fate decisions in development

[6] *In situ* hybridisation was concurrently and independently demonstrated by Buongiorno-Nardelli and Amaldi in 1970[12].

[7] Of course, there were earlier approaches. Using two-dimensional paper chromatography to identify short polynucleotides by their migration characteristics, the sequence of the lac operon of *Escherichia coli* was inferred by Gilbert and Maxam in 1973 – and the entire sequence was printed in the paper abstract [21, 22].

[8] In 1991, automated Sanger-based sequence analysis could handle 96 templates in a day [27].

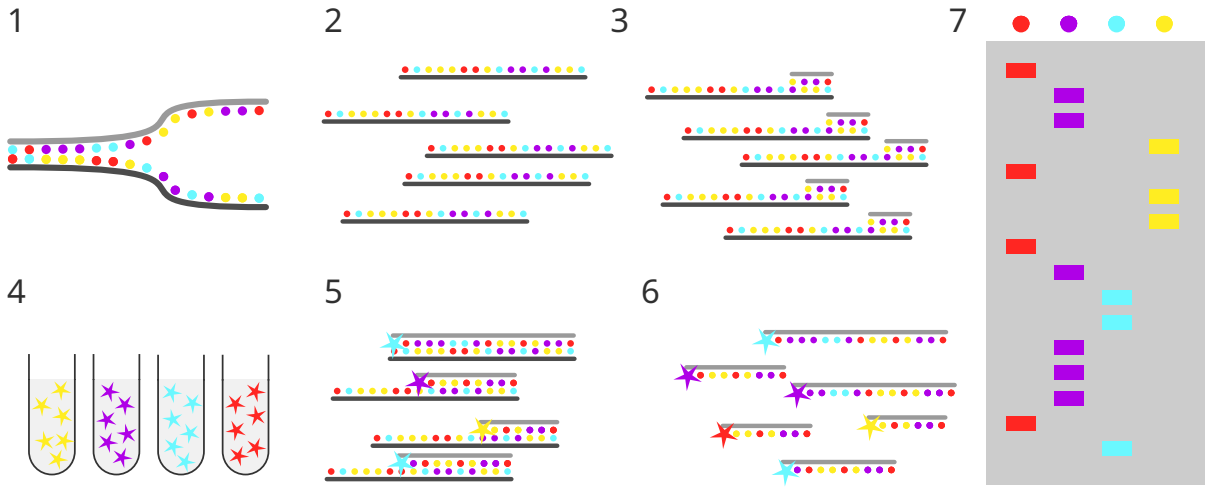

Figure 1.2: **Essential steps of Sanger sequencing.** (1) A fragment of double-stranded DNA is denatured into its constituent strands. (2) One single strand of DNA is amplified into millions of copies, e.g. through Polymerase Chain Reaction (PCR). (3) Short primers, consisting of 20–30 nucleotides, are hybridised to one end of each fragment. (4) The fragments with attached primers are added in equal amounts to four solutions. Each solution contains all four nucleotides, but those of one type are fluorescently labelled to serve as terminal nucleotides. (5) In each solution, a complementary chain to the introduced fragments grows until the random incorporation of a fluorescently labelled terminal nucleotide. (6) The original fragments are denatured from their newly generated complements to obtain a series of single-stranded DNA chains of various lengths. Each strand ends in a fluorescent nucleotide. (7) The DNA chains are separated by length through gel electrophoresis and subsequently imaged to read off the nucleotide sequence.

to disease progression and treatment response, notably in cancer [31–33]. In the following section, I introduce various techniques for the inclusion of spatial information in transcriptional profiling.

1.1.2 Roads to spatiality in transcriptomics

Various technological approaches exist to take into account the spatial distribution of transcripts within tissues. Two different branches of technology development are converging towards high-resolution, unbiased spatial profiling of transcriptional tissue identity. From the high-throughput sequencing side, strategies have been developed to keep information about the tissue context intact and avoid the tissue dissolution necessary for traditional RNA-seq. On the other hand, imaging-based approaches lend themselves naturally to profiling cells in their spatial context. The spatial aspect being thus given, much research has instead gone into the simultaneous detection of multiple molecules, known as multiplexing, and increasing throughput for those technologies. The following paragraphs, and Fig. 1.4, give a brief overview of the different technologies emerging from these broad approaches.

The most straightforward way to spatiality, coming from high-throughput sequencing, is to isolate regions of interest (ROIs), for example through laser capture microdissection (LCM) [35]. Those regions are then dissociated and processed through traditional sequencing. Although a modern extension of the LCM protocol, Geo-seq, is able to reach resolutions up to ten single cells, it is very labour-intensive, limiting throughput [36].

An early alternative approach, the eponymous Spatial Transcriptomics [37], profiles transcripts through an entire tissue slice by unbiased spatial indexing followed by sequencing. The technique, commercialised and further developed by 10x Genomics as Visium, relies on prefabricated glass slides with an array of uniquely barcoded spots. Visium increased the resolution by decreasing the spot size from 100 µm in the original Spatial Transcriptomics to 55 µm, arranged in a hexagonal lattice, and

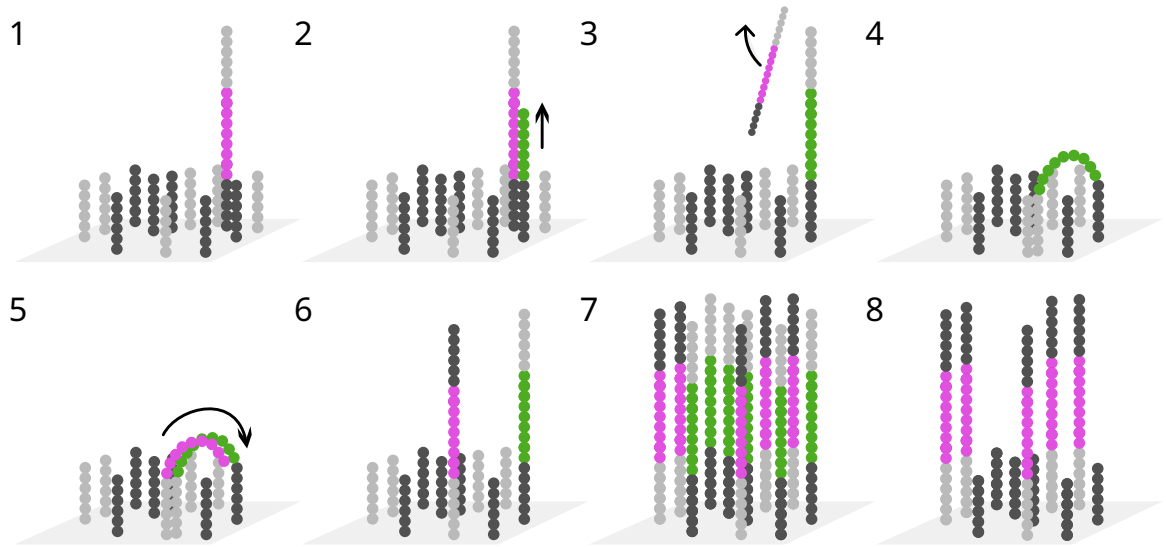

Figure 1.3: **Bridge amplification on a flow cell as used in Illumina sequencing.** (1) The flow cell contains an arrangement of oligonucleotides, to one of which a cDNA fragment hybridises. Previously, adapters were attached to both ends of the fragment (5' and 3'), one of which binds to an oligonucleotide on the surface. (2) Starting from the oligonucleotide, a complementary DNA strand is synthesised along the original fragment. (3) After complete synthesis, the original fragment is denatured and washed away. The flow cell is now prepared for bridge amplification. (4) The second adapter hybridises to another oligonucleotide on the flow cell, bending the DNA fragment. (5) Again, a complementary strand is synthesised, beginning now at the second oligonucleotide sequence. (6) The resulting two strands are separated by denaturation, yet both remain affixed to the flow cell through their adapters. (7) Through repeated bridge amplification, a cluster of short DNA sequences all corresponding to the original fragment is created on the flow cell. (8) Finally, one of the two strand orientations is cleaved from the flow cell and washed away. The resulting arrangement of DNA fragments is then sequenced through repeated imaging after adding fluorescent nucleotides.

the newer Visium HD further reduces the spot size down to 2 µm [38].

In an alternative approach to grid-based barcoding at previously known locations, Slide-seq [39] uses DNA-barcoded 10 µm beads which are dispersed on a glass surface and tightly packed in a monolayer. Bead locations then first need to be identified through *in situ* sequencing. The technique was later extended, with improvements in sensitivity and capture efficiency, into Slide-seqV2 [40]. HDST employs a similar strategy; however, beads are deposited in an ordered well-based array [41]. Finally, Stereo-Seq replaces the beads with DNA nanoballs, created by rolling circle amplification of barcoded primers, and thereby increases the available resolution to 0.2 µm [42].

Imaging-based spatial profiling of the transcriptome has its origins in single molecule fluorescence *in situ* hybridisation (smFISH) [43]. Multiple short fluorescently labelled probes are hybridised to visualise known transcripts in a fixed tissue. This approach is highly sensitive and specific, and can profile a small set of molecules to a spatial resolution defined by the diffraction limit. Cyclic-ouroboros smFISH (osmFISH) is a semi-automated implementation of smFISH capable of handling larger tissue areas [44]. However, as smFISH only visualises one transcript at a time, the number of profiled molecules is still limited by the number of hybridisation rounds.

In sequential FISH (seqFISH), multiple rounds of probe hybridisation, imaging, and stripping allow genes to be identified through sequential colour barcodes [45]. Alternatively, genes can be encoded through binary codes, as in multiplexed error-robust FISH (MERFISH). Separations between barcodes,

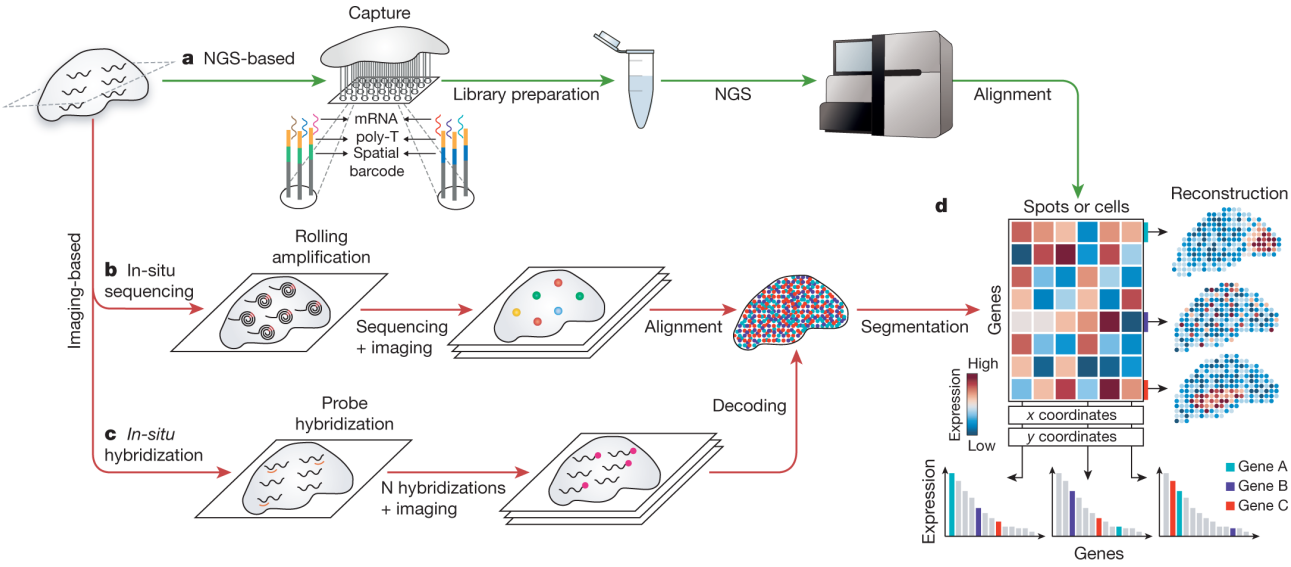

Figure 1.4: **Overview of technological approaches to spatial transcriptomics.** a, high-throughput sequencing-based approaches rely on *in situ* barcoding of transcripts. mRNA molecules are captured within their tissue context and attached to primers corresponding to their location, often within a predefined grid. Subsequently, library preparation and sequencing is performed. b, *in situ* sequencing consists of rolling amplification of transcripts within the tissue context, followed by sequencing and imaging of the amplified DNA balls. c, *in situ* hybridisation-based approaches consist of multiple rounds of fluorescent probe hybridisation and imaging. Optionally, error robustness is improved through the encoding of the sequence in sequential barcodes. As both *in situ* sequencing and *in situ hybridisation* result in molecular resolution, algorithms for cell segmentation are applied to their results. d, From outputs of all approaches to spatial transcriptomics, a count matrix can be defined. This matrix contains the quantified expression of each gene in each profiled spot or cell, with associated spatial coordinates. Figure reprinted with permission from Rao *et al.* [34] © Nature Publishing Group. NGS, next-generation sequencing.

measured in Hamming distance[9], allow for the recognition and correction of sequencing errors [46]. In both seqFISH and MERFISH, only the fluorophores are removed, and probes remain in place between washes, saving time compared to earlier approaches.

Most recently, Xenium is a commercial FISH-based technique based on in situ hybridisation of padlock probes and subsequent rolling circle amplification (RCA) [47]. Padlock probes have the advantage of high specificity, and through RCA, the barcode contained in the probe is highly amplified, increasing the signal-to-noise ratio in subsequent imaging-based readout. Based on the same starting principles of padlock probe hybridisation and RCA, STARmap uses sequencing by ligation (SBL) as a readout technique [48]. Sequencing by ligation, as the name suggests, relies on DNA ligase instead of DNA polymerase for sequencing readout. Fluorescently labelled oligonucleotide probes hybridise to the sequence of interest and are joined by the DNA ligase, resulting in a signal for readout. Technologies such as STARmap are commonly referred to as *in situ* sequencing-based.

For interested readers, a variety of detailed reviews of available technologies for spatial transcriptomics have been published [49–53]. All of the abovementioned technologies have inherent advantages and disadvantages. They are situated within a parameter space opened by spatial resolution, profiled gene panel size, and detection sensitivity. One commonality of all approaches, however, is the wealth of information captured and the trend toward ever higher throughput [54].

[9]The Hamming distance between two barcodes is the number of positions at which the entries differ. It is useful to compare two strings where only substitution errors, not insertions or deletions, are to be expected.

1.1.3 From sequences to data analysis

The advent of large-scale RNA-sequencing-based data generation heralded the need for computational evaluation of the resulting massive datasets, and thus the field of bioinformatics experienced a period of rapid growth[10] [19, 20, 60–62]. The output of a modern sequencing pipeline consists of base calling files. Each sequenced read is stored in the FASTQ format as a sequence of bases, along with a quality score for each base call. Quality scores are calculated based on the probability or odds of a given base having been called correctly, and stored in a single ASCII character. FASTQ files are subjected to quality control concerning quantities like base call quality, GC content, and possible adapter contamination [61]. This ensures that samples with contamination, sequencing errors or PCR artefacts do not affect the final analysis [60]. After trimming and filtering, the thereby processed reads are typically aligned to a known reference genome or transcriptome. If this is not available, or if the aim is to discover novel transcript isoforms, de novo assembly can be undertaken. In this case, reads are first assembled into longer, putative contigs, to which reads are mapped back for quantification. In all cases of alignment, one challenge is the significant fraction of reads that map to multiple locations in the genome, or multiple isoforms in the case of alignment to a transcriptome [60, 62]. Multi-mapping reads pose a difficulty for the quantification of gene, or transcript, expression. For more detailed information about this difficulty and the approaches to overcome it, the interested reader is referred to [62].

Broadly, the annotations of the reference genome or transcriptome are transferred to the aligned reads and used to quantify (count) reads coming from each given gene or transcript isoform. Alternatively, pseudoalignment-based or alignment-free methods forego exact alignment for fast and nevertheless accurate quantification [63]. This quantified expression can be summarised in a count or expression matrix which reports, for each sample, the number of molecules inferred to belong to each gene.

Starting from the expression matrix, various computational analysis tasks can be undertaken. For bulk RNA-seq, the discovery of differently expressed genes (DEGs) between two or more conditions is commonly the next step. As the expression is usually measured, in modern RNA-seq, for many thousands of features, the probability of type I errors (false positives) upon naïve comparison is greatly increased. To avoid this multiple testing problem from distorting significance levels, normalisation and filtering steps need to be carefully applied [62].

As mentioned above, bulk RNA-seq is fundamentally limited to analysis on the tissue level. Since the advent of single-cell RNA-seq (scRNA-seq), assessing gene expression differences and patterns on the cellular level has been possible [64, 65]. A main challenge posed by scRNA-seq is the annotation of individual cells to cell types and states. This can be approached through the evaluation of gene markers. Alternatively, scRNA-seq enables the creation of so-called atlas projects such as the Human Cell Atlas [66], that can be used as reference datasets. Atlases are beginning to be available for a wide range of species [67–69] and tissues [70–72]. Independent researchers can integrate their datasets with these references to annotate their own data using label transfer strategies. Extensive reviews of the challenges and possibilities of scRNA-seq are available [29, 30, 73, 74].

As described in previous sections, an additional limitation of both bulk and single-cell approaches

[10] Bioinformatics as a field had emerged decades earlier, with Ben Hesper and Paulien Hogeweg first coining the term in the beginning of the 1970s to describe "the study of informatic processes in biotic systems" [55]. Even before that, computational approaches to biology had been used in a scientific context [56]. For example, Margaret Dayhoff utilised FORTRAN programs to assist in determining an error-robust amino acid sequence consistent with the known structure of overlapping peptide sequences in 1964, and published the results in a comprehensive and still highly readable article [57]. For the interested reader, I recommend a review of the early days of bioinformatics by Ouzounis and Valencia, published in 2003 [58]. The growth period alongside and after the development of high-throughput technologies is perhaps anecdotally best chronicled by the 2016 article by Jonathan Wren concisely entitled "Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades" [59].

is the nonexistent capture of spatial information. Using spatial transcriptomics approaches, the output of a profiling pipeline is not only an expression matrix containing gene counts for each spot or cell, but also the matrix of associated tissue positions.

This data provides new computational challenges. While researchers employing single-cell resolution spatial transcriptomics approaches (see the previous section for a broad classification) are faced with the problem of cell type assignment detailed above, many technologies aggregate transcriptional profiles over multiple cells. In the Visium technology, with a spot diameter of 50 µm, expression values of typically up to 10 cells can be measured at the same time [75]. Many computational approaches have been developed for the deconvolution of this data and the estimation of cell type composition, often incorporating single-cell reference datasets [76–78].

The immediate next analysis steps branch off into two complementary paths: the delineation of coherent and cohesive transcriptionally defined tissue regions, and the identification of genes whose expression shows a spatial pattern, known as spatially variable genes (SVGs). Approaches for both analyses are manyfold – for SVG detection, methods usually present mathematical models aiming to capture biological signals [79–81]. On the other hand, for spatial region or domain identification, method development has ranged from statistical modelling through clustering-based approaches to the inclusion of sophisticated neural network architectures [82–84]. Further, several approaches have been demonstrated that integrate the two analyses [85, 86]. Of these two fundamental analysis steps, this thesis will focus on approaches for spatial domain identification.

Using spatial transcriptomics approaches, researchers are able to decode spatial dependencies on various length scales, ranging from subcellular transcript distribution up to functional tissue microenvironments [87]. This last type of structure is what the next section will focus on.

1.2 Defining tissue domains

Multicellular organisms are spatially heterogeneous and exhibit some degree of organisational structure – different cell types carry out different tasks. The scale and complexity of cellular organisation range widely, as differentiating tissues develop to fulfil highly specialised functions [88, 89]. Tissues can be told apart by visual or molecular identifiers, as described in the following sections.

1.2.1 Histopathological and molecularly defined regions

Aiming to optimally distinguish different tissues under light microscopic evaluation led researchers to develop advanced histopathological methods, including various staining approaches to increase visual contrast [90]. The most common staining procedure uses Hematoxylin and Eosin (H&E) [90, 91]. Hematoxylin is a cationic basic dye, used as a stain in its oxidised form (haematein) and usually combined with aluminium alum as a mordant [91]. Most prominently, it stains nuclei a blue colour. Complementarily, the anionic acid dye Eosin stains the cell membrane, mitochondria and extracellular matrix pink. Through the combination of the two dyes, fine intra-and intercellular structures can be distinguished in shades of pink and purple (see Fig. 1.5a). On a bigger scale, various tissues within an organ such as the brain appear visually distinct and enable researchers to delineate well-defined regions (see Fig. 1.5b).

In histological approaches, identifying various tissue types is a matter of visual inspection of the morphology after staining. Where they cannot be distinguished visually, tissues can be further characterised by more specialised staining, electron microscopy approaches, or transcriptomic readouts of regions of interest [93, 94]. Spatial transcriptomics, on the other hand, enables the direct assessment of tissue domains in terms of molecular (i.e. transcriptional) identity. Utilised in conjunction with histology, this opens up novel avenues in fields like cancer and neuroscience, for example, the analysis of inter- and intratumour heterogeneity, cortical layers, and tissue development [34, 95]. Beyond animal

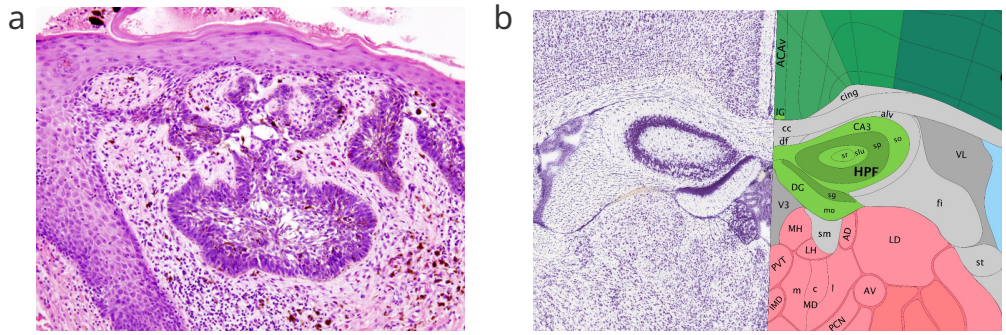

Figure 1.5: **Decoding spatial regions from histological images.** a, H&E stain of basal cell carcinoma of the skin. Cell nuclei are stained in blue-purple, extracellular material is stained in pink. Image from Wikipedia, licenced under CC BY-SA 3.0 (creativecommons.org/licenses/by-sa/3.0/deed.en) [92]. b, Detail from a Nissl stain and the corresponding anatomical annotation of a coronal section of a P56 mouse brain. Allen Mouse Brain Atlas, mouse.brain-map.org [71].

tissue, domain-specific transcriptomic analyses are also applicable to plant systems, though this will not be the focus of this thesis [96].

The evaluation of tissue domains based on molecular tissue identity presents novel challenges. Even for spatial transcriptomics technologies with lower multiplexing capability and thus small panel sizes, such as osmFISH, still many dozens of genes are profiled. For high-throughput sequencing-based approaches, the number of profiled genes can correspond to the size of the entire transcriptome and thus be on the order of 10^4. Individual genes can be easily visualised in space through different colour channels; however, considering more than three genes simultaneously is not possible in this way. Approaches have been developed to visualise the molecular identities of profiled spots or cells using RGB colour channels, through drastically reducing the dimensionality of transcriptomic readouts [97, 98]. These methods, while potentially aiding in the interactive exploration of tissues, may suffer from the general pitfalls of unsupervised dimensional reduction [99, 100]. Therefore, the high dimensionality of the transcriptome necessitates an algorithmic approach to defining spatial domains.

1.2.2 Challenges in defining domains from spatial transcriptomics

To the present day, many approaches and implementations for the identification of spatial domains have been developed. In the effort to categorise methods for spatial transcriptomics analysis in general, multiple databases have been created[11] [54, 101]. Despite the wealth of interest and the rapid tool development, there is no clear consensus on how to transcriptionally define spatial domains (Fig. 1.6a). As reviewed by Walker *et al.*, there are several ways to define what constitutes a domain [102]. Within published approaches, most methods tend to employ one of two definitions: The idea of coherence in gene expression over a spatially contiguous region, or the view of regions with distinct cell type distributions. In both cases, it is crucial to adaptively define thresholds for what is considered incoherent or indistinct.

Further complicating the task of clearly defining spatial domains is the variety of technological approaches to spatial transcriptomics, as outlined in the previous section. Particularly, the disparity in resolution provided by the different approaches makes it challenging to create a technology-spanning definition. In the case of spot-based approaches such as Visium, one spot may capture the transcriptome of multiple cells. This obscures the exact spatial provenance of distinct cells even when

[11] In 2022, Moses and Pachter counted 28 publications relating to spatial regions and in 2023, Chu *et al.* developed the STASH database, showing 65 tools addressing spatial domain identification [54, 101]. In the STASH database, newer methods are not categorised anymore, though the current iteration of the database published by Moses and Pachter counts 148 publications relating to spatial domains ([54], accessed September 2, 2025).

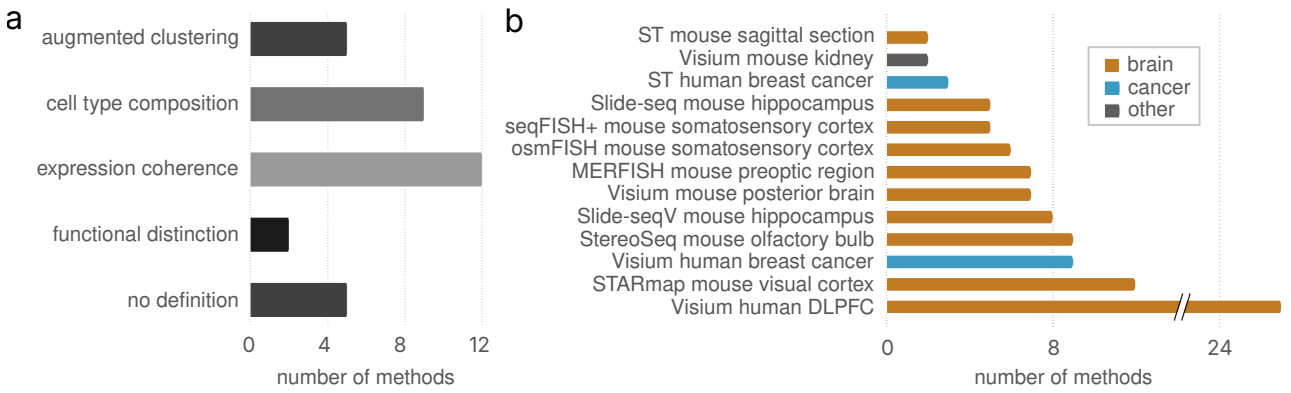

Figure 1.6: **Domain definitions and dataset usage across 34 published methods. a**, Histogram of spatial domain definitions as employed in method publications. Twelve methods emphasise expression coherence, in conjunction with spatial contiguity, while nine methods instead define domains through cell type composition changes. Five method publications view spatial domain identification as spatially augmented clustering, analogous to the clustering of single cells, and two publications focus on the functional distinction between spatial domains. Five publications do not claim a specific definition. Definitions were manually extracted from publications. **b**, Datasets used within method publications for evaluation and benchmarking, sorted by usage frequency. Only datasets that appeared in at least two separate publications are shown. The x axis is broken for ease of visualisation, and bars are coloured according to broad tissue category.

deconvolution of spots into constituent cell types is performed, leading to inexact borders across regions [76, 78, 103]. On the other hand, at single-cell resolution, the challenge becomes exact domain assignment across the diversity of cell types. In heterogeneous tissues such as tumours with a high incidence of infiltrating cells, precise domain delineation may not be possible [104]. Methods have been developed to account for this difficulty by leaving border tissue out of the analysis; however, most methods continue to operate under the assumption of discrete ground truth domain assignments [105].

Finally, considering single-cell resolution technologies, the distinction between cell type clustering and spatial domain identification needs to be an important point of consideration. Cells in the same spatial neighbourhood tend to exhibit common patterns in their gene expression, potentially even across cell types [106–108]. The first method incorporating a spatial proximity constraint into transcriptomics clustering was described in 2014 by Pettit *et al.*, and interestingly, the authors use the method to cluster data into cell types [106]. There are several published methods that can be applied both in a "cell type clustering" and a "domain identification" mode by changing a spatial smoothness parameter [83, 109]. While there has been discussion about exact definitions of "cell types" [110–112], domains should be clearly defined as separate entities.

In the absence of a clear and unified definition of spatial domains, and in the spirit of the purpose of a system being what it does[12], one may turn to the self-evaluation of methods as a way to determine what it is they are looking to find. Particularly, the vast majority of methods benchmark their own performance against that of alternative approaches. A qualitative evaluation is usually carried out, relating the different domain segmentations to biologically meaningful tissue compartments. On the other hand, however, often methods aim to quantify their own superior performance through an accuracy-based comparison to a ground truth.

The ground truth employed in these comparisons is usually defined, at least partially, through the expert annotation of a histological image of the sample in question[13]. However, transcriptional

[12] Stafford Beer coined the phrase "the purpose of a system is what it does" in 2002, in an adress on cybernetics [113].
[13] Although histology images of the same sample are coregistered with molecular profiles in some technologies, notably

markers are often brought in to inform or refine the spot or cell level annotation. The annotation origins of the datasets which are used in this thesis, and in the forthcoming article discussing the benchmarking results, are discussed in Chapter 1.

As the field grew, a core cohort of datasets was established as gold standard resources, commonly used for internal method benchmarking (Fig. 1.6b). These datasets are usually published alongside a ground truth domain annotation, or alternatively, the data is annotated through later, unconnected efforts.

The most prolific dataset by far consists of 12 samples of the human dorsolateral prefrontal cortex (DLPFC) from three different donors, sequenced and published by Maynard *et al.* [114]. The samples are annotated in the original publication into spatial domains corresponding to the anatomical regions defined as white matter and the brain layers L1-L6. The expert annotation is informed by histology images, coregistered automatically with Visium, and layer-specific marker gene expression scores. The continued and widespread use of this dataset for evaluating the accuracy of spatial domain identification indicates the interest in accurately identifying brain layers, corresponding to known anatomical regions, from molecular profiles. Interestingly, of the 13 most frequently used datasets for method evaluation, 10 are datasets of brain tissue (Fig. 1.6b). Of 34 evaluated method publications, 32 evaluate on some type of brain tissue data, while 21 evaluate on at least one cancer dataset.

The disparity between lacking clear definitions and wide purported method applicability is an under-studied area in the field. Particularly, the range of tissue types of interest may necessitate adaptable or varied domain definitions.

1.2.3 Approaches to computational domain identification

The earliest method for identifying spatial domains[14] was published by Zhu *et al.* in 2018 [118]. They describe an approach to identify spatially associated cell subpopulations in seqFISH data using hidden Markov Random fields (HMRFs), a method which would later be integrated in the Giotto framework [87]. Viewed as an application of HMRF, domain identification amounts to modelling an observable, usually a low-dimensional representation of the gene expression, under the assumption of underlying domain labels. A Potts model is employed as a spatial prior in order to impose cluster contiguity. HMRF-based approaches have continued to be developed, with innovations in tunability and applicability to sequencing-based data, multi-sample analysis, and the integration of histological images [82, 119, 120].

Integrating histology into domain recognition based on spatial transcriptomics was pioneered by Pham *et al.* in their method stLearn [121]. In stLearn, morphological information is extracted from an H&E image by an image classification neural network and utilised for spatially-aware normalisation of gene expression values. In subsequent methods, histology images are converted into morphological distances between spots based on RGB values, embedded jointly with gene expression by graph convolutional neural networks, or integrated with gene expression and spatial location through network fusion [122–124]. So far, however, the inclusion of histological information has not been shown to aid in accurate spatial domain identification [117, 125, 126].

Another early approach was demonstrated by Cang *et al.* in SCAN-IT, namely the usage of graph convolutional networks (GCNs) to generate low-dimensional spatially-aware embeddings of gene expression information [127]. In the method SpaGCN, a graph convolutional layer aggregates gene expression, spatial location and histology, where the latter two modalities were previously used to define edge weights in an input graph [122]. With the continued development of the field, graph

Visium, this is not possible in all approaches. In those cases, imaging of adjacent tissue slices may provide close enough analogues to then be able to transfer the annotation to the slice of interest.

[14] Various publications refer to this task as spatial clustering [115–117]. In this thesis, I refer to the more general spatial domain identification, to include clustering-free approaches and emphasise the shared aim of spatial domain contiguity across methods.

attention autoencoders, variational graph autoencoders, along with various data corruption and regularisation strategies, have been applied to spatial domain identification [84, 109, 128–130]. In most of these methods, sophisticated neural networks are applied to the task of generating low-dimensional representations of gene expression, which in some way incorporate information about cellular neighbourhoods. These representations are subsequently subjected to standard clustering algorithms like k-means, model-based clustering as implemented in the R-package mclust, or Leiden clustering, a popular algorithm for single-cell transcriptomics data [29, 131, 132].

A host of other methods also take advantage of well-established clustering algorithms to identify spatial domains. The simplest approach to integrating gene expression and spatial information into a graph for downstream clustering is implemented in TACCO, which simply creates a weighted sum of adjacency graphs for both modalities [76]. Other methods create elaborate expression-aware neighbourhoods or neighbourhood-aware spot networks [83, 133–135]. Yet another approach is taken by SpatialPCA and GraphPCA [136, 137]. As the names suggest, these tools extend principal component analysis (PCA) to be spatially aware by modelling spatial correlation across locations using a kernel matrix or incorporating a spatial constraint in the data reconstruction step. Finally, the extracted spatial PCs are clustered into spatial domains. More recently, spatial dimension reduction has been further extended to multimodal data analysis [138].

The last category of methods approaches domain identification as an image processing problem. Two notable implementations are MULTILAYER and Vesalius [139, 140]. MULTILAYER uses agglomerative clustering of "gexels" to detect contiguous gene expression patterns, which are used to compartmentalise the tissue into domains. Vesalius fully embraces the image processing approach, embedding the transcriptome into an RGB colour space using the nonlinear dimension reduction approach UMAP [98]. Subsequently, spatial domains are identified through iterative smoothing and segmentation.

1.3 Importance of computational methods benchmarking

Benchmarks are commonly used in computer science as a way to quantify and compare the performance of different systems or architectures [141]. In bioinformatics, different algorithms or method implementations are compared in what are commonly known as benchmarking studies [142]. The task of benchmarking, or the comparative evaluation of tools, is often carried out by method developers alongside the publication of new approaches. However, this type of self-evaluation is prone to biases, such as favouring datasets, characteristics or evaluation criteria wherein the authors' own methods excel [143]. Impartial, third-party benchmarking efforts are therefore imperative to the unbiased assessment of methods in a given field and the continued development of relevant and high-performing methods [144].

Many reviews have been devoted to describing best practices and guidelines for benchmarking computational methods. In 2019, Weber *et al.* published a seminal paper on guidelines to computational method benchmarking [145]. More recently in 2023, van Mechelen *et al.* described good benchmarking research practices in an excellent white paper, focusing on the example of clustering methods [146]. Specific to the field of high-throughput measurement in biology commonly termed "omics", Mangul *et al.* authored a comprehensive review in 2019, listing core principles of systematic method evaluations and surveying the state of the art at the time [147]. Brooks *et al.* in 2024 reviewed common oversights and pitfalls in omics benchmarking, and argued for a methodological approach to the reporting of benchmarking pipelines and results [148]. In this section, I will give a brief overview, expanding on the challenges and trade-offs inherent in creating a benchmarking study. Then, I will introduce the context of comparative method evaluation in spatial domains, demonstrating the necessity of a comprehensive and independent analysis.

1.3.1 Best practices and challenges in bioinformatics benchmarking

The core aim of comparative method evaluation is to provide guidelines for prospective users and further method development. For meaningful and informative benchmarking in these contexts, careful consideration must be applied when selecting methods, datasets, and evaluation metrics. Some trade-offs inherent in these selections are listed in Tab. 1.1 and covered in more detail in the following paragraphs. Specifically, I will discuss method selection and hyperparameter setting, data selection, as well as performance metrics and broader evaluation criteria.

Regarding method selection, it is generally considered ideal to include all relevant methods [149]. However, the feasibility of this approach may be constrained by computational power and/or time, as the number of methods exceeds multiple dozen for some tasks[15]. Nevertheless, there have been efforts to benchmark impressive numbers of methods, such as a 2019 publication by Saelens *et al.* that evaluates 45 tools for single-cell trajectory inference [151]. Most benchmarking studies, however, consider significantly fewer individual tools. Two recent meta-evaluations of single-cell benchmarking efforts record a median of 10 methods benchmarked in independent comparisons [152, 153]. On this scale of method evaluation, it is important to choose tools which best represent the state of the art, although this criterion remains ambiguous [154].

Once methods have been selected, the setting of possible hyperparameters plays an important role. A common approach to hyperparameters is leaving them at default values, or setting them to values recommended by the method developers. This reflects common usage and is simple to implement. It may not always be clear how strongly the performance of a tool is affected by a given parameter, and thus, how much effort should be invested into optimising the parameter tuning. This trade-off between ease of optimisation and attainable performance benefit has been investigated in-depth for the case of clustering [155]. However, not all methods provide default or recommended hyperparameter values, and for those that do, not all values generalise to all types of data the method may be applied to. Certain benchmarking studies therefore distinguish between "versions" of tools, implemented using different hyperparameter settings [156]. It is also interesting to separately consider preprocessing and postprocessing steps, which may be common between various tools and may not necessarily be specified in detail [117, 157]. The impact of preprocessing on method performances has been studied for selected analyses, notably in the case of dimensionality reduction [158, 159]. The other end of the spectrum regarding hyperparameter determination, namely, implementing a comprehensive parameter space sweep, is not feasible in most settings. In studies with small numbers of methods or highly standardised hyperparameters, a sweep might be carried out across all methods. Particularly, if a set of parameters is influential to method output and common to all tools, a parameter sweep is indicated [147]. Otherwise, when parameter settings are optimised only selectively or optimised parameters are only relevant in a subset of methods, one runs the risk of unequal treatment of tools. Further, all methods should be provided with the same information about the test data [146].

Once the selection of tools is complete, the benchmark data to be used for their evaluation needs to be selected. Here, the base consideration concerns the types of data commonly analysed in the field, to which the selected tools are likely to be applied. The selected datasets should represent a wide range of applications and conditions [145, 146]. Fundamentally, benchmark datasets can be either real, containing experimentally measured data from a system of interest, or synthetic, created at least partially through computational simulation. Hybrid, semi-synthetic datasets can be generated on the basis of real data, but augmented or transformed in specific ways through simulation [145, 148].

For the evaluation of tool performance on selected datasets, it is indispensable to be able to quantify

[15] In the example of single-cell analysis, the scRNA-tools database at scrna-tools.org currently tracks 1837 tools (> 153 dozen) over all analysis types (status: October 1, 2025) [150]. For clustering alone, 397 tools (> 33 dozen) are recorded. The numbers for spatial transcriptomics tools are lower due to the more recent emergence of the fields, but growing rapidly [54].

performance, in particular, to define what is understood to be a "good" performance outcome [146]. To this end, researchers often employ the comparison to ground truth values for the analysis outcome [145, 147, 148]. In the case of real datasets, the generation of gold-standard information to be used as a ground truth may be included as a part of data acquisition, as a first step in the benchmarking pipeline [148]. Depending on the analysis type, gold-standard evaluation may be commonly published alongside the raw data. This is the case for fundamental processing steps such as annotating cell types in single-cell RNA-seq data. While this ground truth type is usually at least partially attained by manual expert annotation and should have undergone rigorous quality control prior to publication, it is often not feasible to comprehensively check ground truth validity. On the other hand, a ground truth for semi-synthetic or fully synthetic datasets may be generated along with, or underlying, the simulated data [160].

Special care must be taken, notably in the case of synthetic data, to avoid using the same models to generate data that are also used in methods to be evaluated, as this would end up biasing the evaluation towards those methods. A similar effect may also occur when using previously published real datasets, as those may be utilised for evaluation during method development or even included in training datasets for some learning-based methods. In emerging fields, there might be a lack of real data generated with an associated ground truth, leading to overfitting of methods to specific popular benchmarking datasets [161].

Lastly, method performance on benchmarking datasets must be evaluated using an appropriate and comprehensive list of evaluation criteria. Primarily, the performance of a tool is graded by the quality of its outputs, measured as accuracy of classification, correlation or cross-entropy of continuous variables, or root mean square errors, among many further possibilities [145, 147]. A balance must be struck here between including popular, easily interpretable metrics, considering edge cases that are potentially not covered by common metrics, and creating specialised, tailored evaluation criteria for benchmarking. If applicable, it may also be interesting to evaluate tools' error rates in terms of the relative abundance of type I and type II errors (false positives and false negatives). Further, beyond the simple performance evaluation on individual tasks, a comprehensive benchmark should investigate the stability and robustness of tools' performance [146]. Stability analysis can encompass running methods repeatedly on sub-sampled data from the same underlying dataset or, for non-deterministic methods, on the same data for different values of a random seed [145, 146]. On the other hand, method robustness can be tested with respect to data perturbations or hyperparameter settings. Data-level perturbations of interest may include downsampling data, introducing artificial noise or outlier values, or changing parameters in preprocessing steps [146].

These primary quantitative performance evaluation criteria should be complemented by secondary criteria, concerning the quality of the method implementation [145]. Several criteria may be of interest to the end user, such as method runtime, memory usage, scalability, and user-friendliness or usability. Runtime and memory usage should be evaluated and compared using standard computational architectures to enable users to easily compare with their own machines. For scalability analysis, methods should be evaluated on a range of datasets that vary in size but otherwise exhibit shared characteristics. The assessment of usability, encompassing ease of installation, support for different operating systems, and documentation quality, is highly subjective. It can be standardised, to a degree, by using weighted checklists [162].

Finally, it can be useful to summarise metrics into an overall ranking, where a balance has to be struck between weights assigned to the available evaluation criteria. Namely, end method users may not be interested in a high-performing method requiring highly specialised computing architectures or long running times, whereas high speed and computational efficiency might increase the attractiveness of a lower-accuracy method [163].

Concern	Trade-offs
Method selection	Comprehensiveness vs Investment of resources
Parameter tuning	Exhaustiveness vs Investment of resources
Breadth of real data origins	Range of applications vs Necessity of parameter tuning
Ground truth availability	Broad dataset inclusion vs Comprehensive validity check
Synthetic data	Realism vs Tunability and availability
Metric selection	Interpretability vs Specialisation
Metric summarisation	Primary evaluation vs Secondary evaluation

Table 1.1: **Trade-offs inherent in benchmarking study design.** Trade-offs are shown for different areas of concern.

1.3.2 Comparative evaluation of spatial domain identification methods

This thesis deals with the evaluation of methods for spatial domain identification, so in this section, I will review the state of the art of benchmarking in this field. Comparing method performance is only reasonable and possible when multiple tools exist that try to accomplish the task in question. At the advent of any avenue for data analysis, as tools are first being developed, comparative evaluation to previous approaches is carried out within method publications. In the field of spatial domain identification, most methods perform a quantitative comparison to available tools in addition to qualitatively evaluating their own performance, aiming to demonstrate their advantage in select applications. The majority of methods compare their performance to 7 other methods or fewer, with only four methods out of a sample of $n = 33$ benchmarking against 9 or more methods (Fig. 1.7a). Four methods in this informal review do not perform any comparisons to other methods for spatial domain identification [76, 86, 118, 164]. A detailed overview of the surveyed methods is shown in Appendix A (Tab. A.1).

The Adjusted Rand Index (ARI) is chosen for the quantification of clustering accuracy in the vast majority of comparisons (27 out of 33, see Fig. 1.7b). The ARI is introduced in more detail in the following chapter. Briefly, it is a supervised evaluation metric, comparing a putative clustering to a ground truth set of annotations by evaluating pairwise cluster membership. A number of supervised metrics are employed in subsets of method publications, such as the Adjusted and Normalised Mutual Information (AMI and NMI), homogeneity and completeness (HOM and COM), and the Fowlkes-Mallows Index (FMI). All of these metrics will be detailed in Chapter 2. Like the FMI, both the F-score and the Area Under the Curve (AUC) are supervised evaluation metrics based on precision and recall, and are employed in a total of three recent method publications [120, 167, 168]. On the other hand, only a few unsupervised metrics, which evaluate the goodness of a putative clustering without relying on the comparison to a ground truth, are utilised. Among those that are included are the CHAOS and Percentage of Abnormal Spots (PAS) scores (the latter not shown), both adapted from image analysis and used in two publications [136, 169]. Further, the Local Inverse Simpson's Index (LISI) is employed in two method publications to evaluate the mixing of cell types in identified domains [109, 136]. Differentially expressed marker genes and domain-specific SVGs are used in select publications to gauge the quality of domains [120, 122, 127, 130, 170].

One particular well-annotated human brain dataset is employed for method comparison in most cases, as already shown in Fig. 1.6b. This is a dataset of the human dorsolateral prefrontal cortex, sequenced using the Visium technology, introduced in more detail in Sec. 1.2.2. For my purposes here, the conjunction of common dataset and metric usage allows the creation of a directed graph of method comparisons, as shown in Fig. 1.7c. This graph shows methods as nodes, connected by an edge if the methods have been compared in a benchmark published alongside a novel approach. The edges are directed from the reportedly higher- to the reportedly lower-performing method, and weighted by the number of comparisons. As numerical performance values are not published in all cases, in some comparisons, the edge direction had to be inferred from visual estimation.

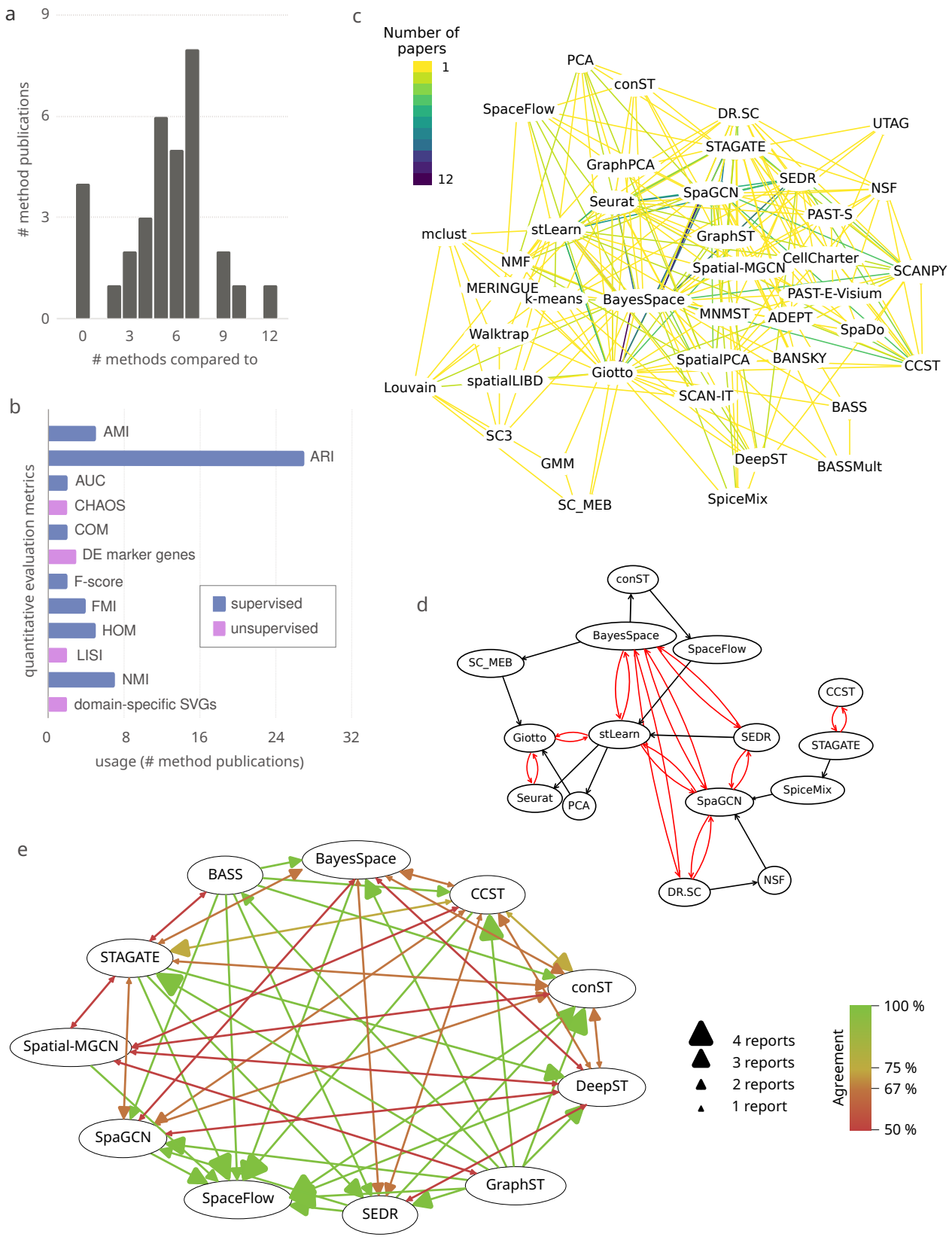
a
method publications
methods compared to
b
quantitative evaluation metrics
AMI
ARI
AUC
CHAOS
COM
DE marker genes
F-score
FMI
HOM
LISI
NMI
domain-specific SVGs
supervised
unsupervised
usage (# method publications)
c
Number of papers
PCA
conST
SpaceFlow
DR.SC
UTAG
STAGATE
GraphPCA
SEDR
NSF
SpaGCN
Seurat
PAST-S
mclust
stLearn
GraphST
NMF
CellCharter
Spatial-MGCN
SCANPY
MERINGUE
k-means
BayesSpace
PAST-E-Visium
MNMST
ADEPT
SpaDo
Walktrap
SpatialPCA
BANSKY
CCST
Louvain
spatialLIBD
Giotto
SCAN-IT
BASS
SC3
GMM
DeepST
BASSMult
SC_MEB
SpiceMix
d
conST
BayesSpace
SpaceFlow
SC_MEB
CCST
Giotto
stLearn
SEDR
STAGATE
Seurat
PCA
SpaGCN
SpiceMix
DR.SC
NSF
e
BASS
BayesSpace
CCST
STAGATE
conST
Spatial-MGCN
DeepST
SpaGCN
GraphST
SpaceFlow
SEDR
4 reports
3 reports
2 reports
1 report
Agreement
100 %
75 %
67 %
50 %

Figure 1.7: **Existing benchmarking efforts for spatial domain identification.** a, Histogram showing the number of comparisons to other methods undertaken in the context of 33 publications containing self-evaluation benchmarking. b, Usage of metrics for quantitative evaluation across the 33 publications. Metrics were only included if they were utilised in at least two separate publications. Bars are coloured according to the evaluation type. c, Method rankings as extracted from within-publication benchmarking efforts. All methods were benchmarked on the human dorsolateral prefrontal cortex dataset published by Maynard *et al.* and performance was evaluated by ARI [114]. Methods are shown as nodes, comparisons as directed edges. Arrows point from better to worse performing methods, and edges are weighted by the number of published comparisons of the node methods. d, Pruned method graph as in c, direct contradictions are marked in red. e, Method comparisons on the same dataset from four published independent benchmarks of spatial domain identification [117, 126, 165, 166]. Methods are only included in the graph if their performance is reported by at least two benchmarks, and edges are only drawn if at least two benchmarks report their relative performance. Relative performance is measured in ARI in all cases. Edges are directed from higher to lower reported performance. In cases with conflicting reporting, the respective edge is bidirectional. The size of the arrow heads corresponds to the number of publications reporting a given ranking, and the colour of the edge details the level of agreement between the reports. Agreement ranges between 100% (all reports agree on the ranking), 75% (one report contradicts three), 67% (one report contradicts two), and 50% (one report contradicts one other).

From Fig. 1.7c, we see that while most method comparisons are only benchmarked once, there is a subset of methods frequently employed in comparisons. These methods tend to be implementations of non-spatial clustering algorithms employed as baselines (such as scanpy and Seurat), or early spatial clustering approaches (like Giotto, stLearn, BayesSpace, SpaGCN, SEDR, CCST, and STAGATE) [84, 87, 109, 119, 121, 122, 170–172]. Between methods with multiple comparisons, contradicting comparison results are possible. That is, edges with weight > 1 can be bidirectional. These are highlighted in a reduced subgraph in Fig. 1.7d. Considering only the comparative method evaluations performed in the context of within-method benchmarking can thus lead to contradictions. In certain cases, these contradictions may be due to conscious or unconscious biased evaluation, as described by Jelizarow *et al.* [154]. For example, some publications evaluate varying hyperparameters for the presented method and not for those to which it is compared [128]. As hyperparameters used for other methods are not reported in many cases, this is difficult to verify. On the other hand, the contradictions in reported performances might also be simply due to method instability or randomness.

In an effort towards independent benchmarking, since the field is maturing, a handful of more independent benchmarking efforts have been published [117, 126, 165, 166]. Notably, some of these publications benchmark spatial domain identification tools previously developed within their working groups [128, 161]. Still, these efforts should avoid any kind of biased evaluation, and within the respective scopes of these comparisons, methods are evaluated more comprehensively. In spite of this, the comparison of method performance by ARI on the aforementioned dataset by Maynard *et al.* is a core part of the evaluation, again allowing a comparison of method comparisons to be drawn (see Fig. 1.7e). Interestingly, even within independent benchmarks, evaluating methods on the same dataset and with the same metrics, contradicting method performances are reported. Between methods compared in two or more independent benchmarking studies, rankings are contradicting in a large proportion of cases, shown in shades of orange and red in Fig. 1.7e.

These contradictions between independent benchmarks should not be due to preferential treatment of select methods. Indeed, all four benchmarks report primarily utilising default parameter settings, or setting hyperparameters according to developers' recommendations. There may, however, be differences in preprocessing steps not covered by method defaults, or ideal values for hyperparameters

identified through parameter space sweeps. Alternatively, inherent randomness on the part of the methods might still play a role.

The present benchmarking effort, as described in this thesis and as will be partially published in the accompanying publication, aims to disentangle possible factors affecting method performance. As method usage with default parameter settings reflects popular usage, the focus is placed less on method parameter space exploration than on the effect of various data characteristics.

1.4 Thesis overview

Spatial transcriptomics has revolutionised the analysis of biological tissues, enabling the direct molecular profiling of cell types, states, and interactions [50]. It has broad implications for medical research, ranging from oncology to nephrology and neuroscience, as well as infectious diseases [173–178]. Further, spatial transcriptomics can aid in fundamental research on development, tissue architecture, and systems biology [173, 179, 180]. A central part of the analysis pipeline for spatial transcriptomics data is the identification of cohesive and characteristic regions within the tissue, commonly termed spatial domains [54, 87, 122, 181, 182]. For this purpose, a wealth of computational approaches has been developed [183]. Efforts have been made to categorise and evaluate groups of tools on public, annotated datasets [117, 126, 165, 166]. However, as described above and in the next chapter, ground truth annotations of real datasets are commonly defined based on manual annotation and thus contain a degree of uncertainty. A comprehensive analysis of method performances, which includes a systematic exploration of different data characteristics, has been missing from the field. The present thesis, alongside the forthcoming corresponding benchmarking publication, attempts to fill this gap.

This thesis presents the cumulative effort of benchmarking computational methods for domain identification in spatial transcriptomics data. The results discussed in the following chapters arise from collaborative work. Chapter 2 introduces the 26 methods that will be investigated, as well as publicly available spatial transcriptomics datasets and evaluation strategies utilised in the benchmarking process. Further, it discusses our benchmarking pipeline and elaborates on results and hypotheses derived from running methods on real data with expert-generated ground truth. Following up on the analysis of method performance on real data, Chapter 3 describes the development of a reliable pipeline for semi-synthetic data generation, commenting on the state of the art in spatial transcriptomics simulation. Additionally, it showcases how method performances are affected by the systematic variation of data characteristics. The effect of technological parameters such as the resolution, the number of profiled genes and the sparsity of the resulting data is investigated. Moreover, the pipeline allows the variation of parameters relating to tissue properties like cell type similarity, molecular heterogeneity, and the size and shape of domains. Additional analyses relating to the method evaluation concerning runtime and memory usage, as well as usability, are presented in Chapter 3. Overall, this thesis presents a significant contribution to the field, functioning as a review of the state of the art and examining in detail various factors affecting the performance of spatial domain identification methods.

Chapter 2

Benchmarking spatial domain identification methods on real datasets

This chapter describes the benchmarking of methods for spatial domain identification using public spatial transcriptomics datasets. I introduce the materials and approaches employed in the benchmarking effort, and show the results of different analysis strategies.

2.1 Prerequisites and implementation

As introduced in Section 1.3, the choice of methods, datasets and metrics is of vital importance for the execution of a well-rounded benchmarking study. The methods selected for comparison should, besides being suited to the task at hand, cover a significant portion of the available approaches, and represent the state of the art. They should be run on data with a well-defined ground truth, chosen from the entire field of possible method applications to ensure a broad method evaluation. Finally, the metrics chosen to represent method performance need to be optimally suited to investigating the specific task under study, interpretable, and relevant to the field. Each of these selections is detailed in a dedicated section. Additionally, the implementation of a reproducible benchmarking pipeline using Snakemake is described.

2.1.1 Method selection

We conducted an informal literature search for spatial domain identification tools, and chose methods for benchmarking based on ad-hoc criteria formulated in Tab. 2.1. Broadly, the criteria cover method relevance to the task and to the field, ease of installation and implementation in the pipeline, and the variety of algorithmic approaches. In the following, for simplicity and brevity, "method" will always refer to a method for spatial domain identification, unless otherwise specified.

As previously introduced in Section 1.2.3, methods range broadly in their approaches. In the present thesis, we will focus on the main categories, which we identify as statistical modelling-based, neural network-based, clustering-based and image processing-based methods. However, methods can be further stratified within and across these categories. In Fig. 2.1a, this subclassification is shown as a graphical introduction to the diversity of approaches included in our study.

The broadest category of methods, which intersects two of the three other main classification groups, is the clustering-based approaches. We characterise methods as clustering-based when the final step of their domain identification strategy implements conventional clustering algorithms such as k-means, mclust, or Leiden [131, 132]. These methods use a variety of strategies to gain a data representation to then cluster into domains. Some use neural network or statistical modelling-based approaches, and are therefore categorised into these respective groups. Methods which are assigned to

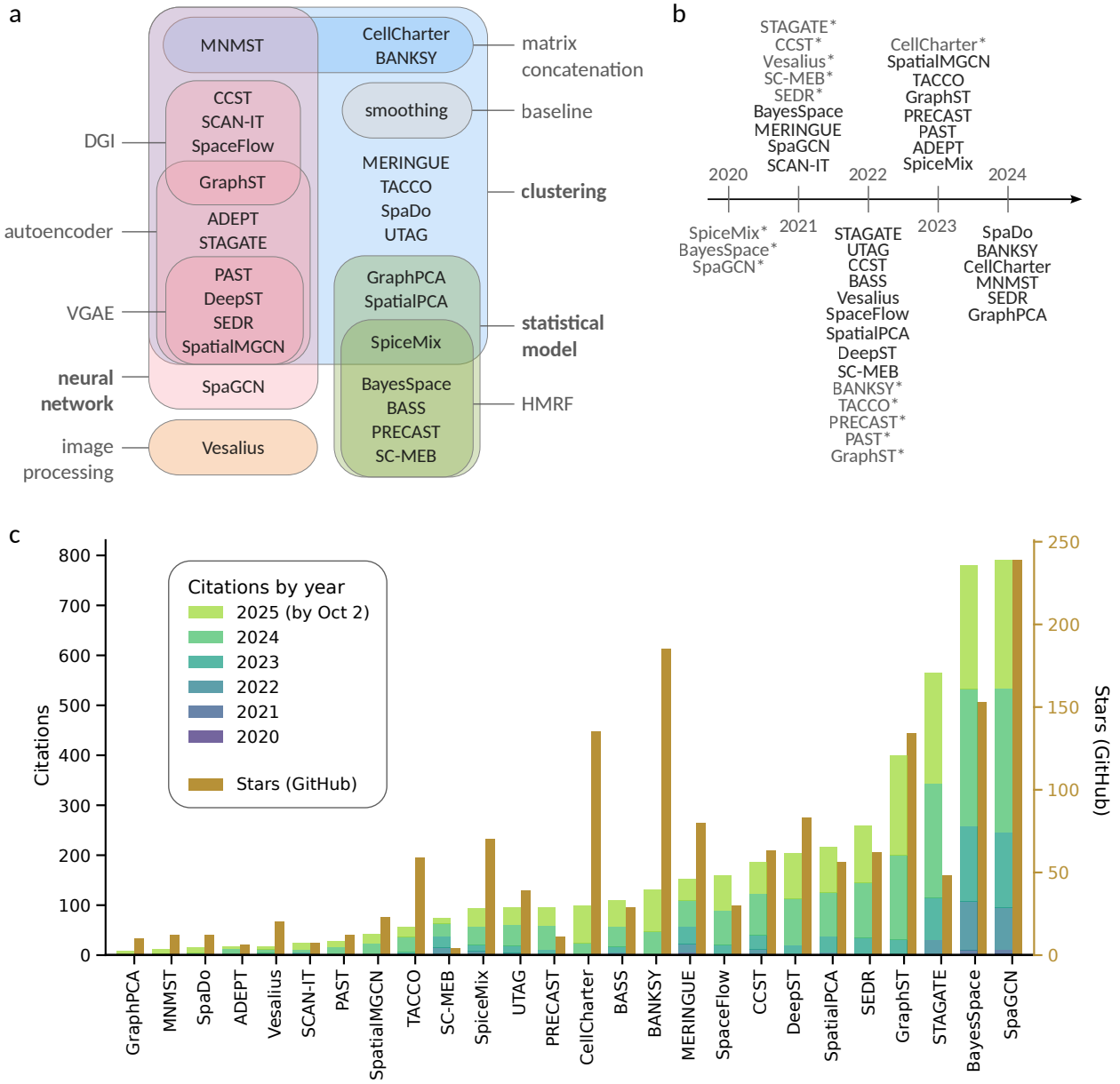

Figure 2.1: **Aspects of methods included in the present benchmarking effort.** Non-spatial baselines are not shown. a, Methods are categorised by their approach broadly into statistical modelling-based, neural network-based, clustering-based and image processing-based. Each of these main categories is assigned a coloured box for visual stratification, and main categories containing more than one method are indicated in bold. Further subclassifications within and across these categories are also indicated by coloured boxes and annotations. b, Coarse timeline of method publication. For methods which were published as preprints before peer review and publication in a journal, the time of preprint publication is indicated in grey and with an asterisk. c, Number of citations (extracted from Google Scholar) and number of Github stars (Status of both metrics: October 2, 2025). Citations are stacked over all years a method has been available, and methods are sorted according to the cumulative value. DGI, Deep Graph Infomax; VGAE, Variational Graph Auto-Encoder; HMRF, Hidden Markov Random Field.

Criterion	Formulation
Relevance	Is the method developed for spatial domain identification, or does it otherwise enable the user to identify spatial domains?
	Is the method referenced and/or benchmarked against in other method development efforts for spatial domain identification?
	Is the method applied for spatial domain identification in published research by biologists or medical researchers?
Usability	Is the method open source, and can it be accessed and installed from a public repository such as GitHub?
	Can the method be run in a script-based, non-interactive manner (i.e., not exclusively through a Graphical User Interface (GUI))
	Is sufficient documentation provided to enable users to install and run the method?
Variety of approaches	Does the method introduce a novelty in its approach to spatial domain identification?
	Does the method claim competitive or superior performance to previous efforts in specified scenarios?
	Does the method belong to an otherwise underrepresented type of approach?

Table 2.1: **Ad-hoc criteria used for method selection.** Criteria cover categories of relevance, usability, and variety of approaches.

the purely clustering-based group include UTAG, TACCO, MERINGUE, and SpaDo. **UTAG** creates a neighbour graph between input coordinates based on a user-defined Euclidean distance threshold [184]. Subsequently, it uses the adjacency matrix of this graph to aggregate features per neighbourhood, either by mean or sum over all neighbours. This augmented feature matrix is the input to the downstream Leiden clustering. **TACCO** calculates two k-nearest neighbour (-NN) graphs of the input spots or cells, based on spatial coordinates and gene expression [76]. It applies Leiden clustering to the weighted sum of the two graphs, where the weight is a user-defined hyperparameter. **MERINGUE** creates a -NN graph based on the gene expression [86]. It calculates a Voronoi tessellation in real space and defines the nearness of two spots or cells via the number of borders crossed when passing from one to the other. The nearness is used as the edge weights to the -NN graph, which is then clustered using the Louvain algorithm. **SpaDo** defines neighbourhoods using a -NN graph for single-cell resolution data and a user-specified radius threshold for grid-based spot resolution data [134]. It represents cell or spot features by the cell type distribution within the local neighbourhood. As a preliminary step, if cell type labels are not available for the input data, it includes automated cell type annotation and, if necessary, spot deconvolution. The distance between spots or cells is calculated using distribution distance measures, namely either the Jensen-Shannon divergence or the Manhattan distance [185]. Domains are identified from the distance matrix by hierarchical clustering.

BANKSY and CellCharter are also clustering-based methods, and both employ the common principle of matrix concatenation. **BANKSY** starts by creating a neighbourhood graph based on real space coordinates, using k-NN, radius-based, or Delaunay triangulation [83]. Edge weights for this graph are generated by default using a $1/r$ kernel, and a neighbourhood representation is calculated as the weighted average of the gene expression. The resulting representation, along with an azimuthal Gabor filter representation, is concatenated with the scaled original gene expression matrix into a neighbourhood-augmented matrix. After dimensionality reduction using Principal Component Analysis (PCA), Leiden clustering is employed to identify spatial domains. **CellCharter** uses a Delaunay triangulation for single-cell resolution data, and a k-NN graph for regular grid-based data, to generate a spatial coordinate network [135]. Features are aggregated within neighbourhoods of different sizes, defined by adjacency steps from the original spot or cell, up to a user-defined number of steps. The aggregation function can be designed for a specific task of interest. The resulting aggregated matri-

ces are concatenated to the original gene expression and clustered using a Gaussian Mixture Model (GMM).

One additional method, MNMST, utilises the matrix concatenation approach. As it incorporates neural networks as a central architecture part, it is assigned to the neural network-based method category. **MNMST** creates a spatial adjacency k-NN graph, and uses pointwise mutual information to define the adjacency [186]. For the gene expression, it utilises the BANKSY approach to augmentation with the local neighbourhood. An adjacency matrix is learned from the gene expression using sparse self-representation learning. MNMST then jointly learns the shared cell features and an affinity graph. The final learned affinity graph is clustered using the Leiden algorithm.

A further group of methods are characterised by their use of the unsupervised graph representation learning method Deep Graph Infomax (DGI) [187]. Based on Graph Convolutional Networks (GCNs), this approach aims to maximise the mutual information between local (patch-level) representations and high-level graph summaries by learning to discriminate between "true" and "corrupted" node relationships. Methods that use DGI include CCST, SCAN-IT, SpaceFlow and GraphST. **CCST** creates a radius-based spatial adjacency graph of spots or cells, and uses a user-defined hyperparameter to set the importance of spatial information for the subsequent embedding [109]. It utilises DGI on the weighted graph to learn an encoder of four graph convolutional layers, creating corrupted samples by permuting edges in the graph. The final embedding obtained by DGI is clustered by k-means after dimensionality reduction by PCA. **SCAN-IT** utilises the alpha complex to define spatial adjacency based on a distance threshold in the Voronoi tesselation [127]. It uses a two-layer GCN as an encoder, trained using DGI using permuted note features as corrupted samples. Finally, a consensus distance matrix from a collection of low-dimensional embeddings is employed to calculate the final representation using metric multidimensional scaling (MDS). This final representation is clustered using k-means or Louvain algorithms. **SpaceFlow** also uses an alpha complex approach to generating a spatial adjacency graph, though a k-NN approach is also implemented [188]. It utilises the DGI framework to train a two-layer GCN using node-permutation for the construction of a corrupted graph. SpaceFlow adds a spatial regularisation term to the loss function to be optimised, such that small distances in embedding space for far-apart spots or cells in real space are penalised. The strength of this spatial regularisation is a hyperparameter. Finally, domains are obtained using Leiden clustering. **GraphST** creates a spatial k-NN graph as an input, and creates a corrupted graph by node feature reshuffling [115]. It utilises a GCN-based encoder-decoder structure, with an objective function comprising the self-reconstruction loss and a DGI-inspired contrastive loss for both the original and corrupted graphs. Finally, the reconstructed gene expression is clustered using mclust [131].

Along with GraphST, most other neural network-based methods employ autoencoder-based frameworks. Autoencoders learn latent factors inherent in the data through self-supervised learning, comparing the final output of the decoder component to the original input. The comparison is usually accomplished by the mean squared error, if not explicitly mentioned otherwise. Two methods, STAGATE and ADEPT, are categorised as purely autoencoder-based, and both of these methods encompass an additional graph attention layer, encoding the relative importance, or similarity, of neighbour features [189]. **STAGATE** constructs a radius-based binary adjacency matrix, with the radius set in grid-based data to detect only nearest neighbours [84]. For grid-based, low-resolution data, STAGATE can optionally prune the graph based on a coarse gene expression pre-clustering by the Louvain algorithm. It then embeds the gene expression matrix, aggregating over the adjacency-defined neighbourhoods using graph attention weights to achieve a spatially-aware embedding. A two-layer network with graph attention is used as the encoder, and the decoder is given by an additional two layers. After minimising the reconstruction loss, the embedding is clustered using mclust when domain numbers are known, or Louvain otherwise [131]. **ADEPT**, after k-NN graph construction, uses a standard graph attention autoencoder to learn a neighbourhood-aware spot or cell embedding [128]. Based on

the resulting embedding, it performs an initial clustering, from which sets of differentially expressed genes (DEGs) are extracted. DEGs are calculated on a one-vs-all basis by a Mann-Whitney U test of expression rankings. For the total list of DEGs, it then creates a full, elementwise nonzero expression matrix by imputation. This matrix is fed to the graph autoencoder, and the final output embedding is clustered to define domains.

Further autoencoder-based methods are PAST, DeepST, SEDR, and SpatialMGCN. These methods specifically implement variational graph autoencoders (VGAEs). This autoencoder type learns stochastic embeddings, modelling the latent space as a probability distribution and thus allowing for inference and data generation based on the embedding. **PAST** uses two parallel modules in the first layer of its encoder architecture, namely a Bayesian neural network (BNN) and an unrestrained fully connected network (FCN) [129]. The output from these two modules is concatenated, and with a k-NN graph of spatial locations, input to two self-attention layers. Finally, two FCNs are used for reparametrisation into a latent Gaussian distribution and create the final embeddings. The decoder consists of a three-layer network, encompassing two self-attention modules and an FCN layer. For large-scale applicability, PAST includes a ripple walk sampling strategy to enable minibatch training. Its objective function finally consists of the reconstruction loss, the Kullback-Leibler (KL) divergence to a standard normal prior, the BNN loss and a metric learning loss. Using the BNN, PAST is able to optionally incorporate reference gene expression data. **DeepST** calculates a modified gene expression matrix, encompassing information from spatial neighbours weighted by their expression correlation and, optionally, morphological similarity [190]. Additionally, it constructs a k-NN graph based on spatial coordinates. Subsequently, DeepST implements a denoising autoencoder as well as a VGAE with reconstruction loss and KL divergence. Domains are identified by Leiden clustering. **SEDR** incorporates a data masking step, whereby the gene expression of randomly sampled spot subsets is masked with learnable vectors [170]. It uses a two-layer encoder to create embeddings from this masked input, and a GCN to embed the spatial information. The two embeddings are concatenated, and a one-layer graph convolutional decoder reconstructs the expression matrix, minimising reconstruction loss. A VGAE learns a graph embedding based on the feature representation from the previous step. The resulting embedding is again concatenated to its input, and an adjacency matrix is reconstructed from this merged representation. The VGAE aims to minimise both the cross-entropy loss for the learned adjacency and the KL divergence for the distribution obtained by reparametrisation from the graph embedding. Latent representations are clustered by default using mclust [131]. **SpatialMGCN** uses a radius-based binary adjacency criterion to create a spatial graph of spots or cells [130]. From the expression values, it generates a k-NN feature graph, measuring gene expression similarity by cosine distance. It then implements a multi-view GCN encoder, consisting of individual convolution of the spatial and feature graphs separately, co-convolution of both by parameter sharing, and finally an attention mechanism applied to the separately generated embeddings. The decoder to reconstruct the expression matrix operates under the assumption of a zero-inflated negative binomial (ZINB) distribution of the gene expression, using the negative log-likelihood of the ZINB distribution as the reconstruction loss. Additionally, SpatialMGCN incorporates a spatial regularisation loss term to minimise the embedding distance between spatial neighbour spots or cells.

The final neural network-based method is SpaGCN, the only non-clustering-based method in this category. **SpaGCN** calculates a complete weighted undirected graph of spots or cells [122]. The edge weights are calculated using a Gaussian kernel from the Euclidean distance in real space, optionally integrating morphological feature information as a third dimension. It uses a single graph convolutional layer to embed the expression matrix, along with the spatial graph structure. Cluster centroids are initialised from this embedding using Louvain, and cluster assignments are refined using an iterative strategy.

All but one of the remaining methods belong to the statistical modelling-based category. SpatialPCA and GraphPCA generate spatially-aware low-dimensional embeddings by statistical means,

analogous to PCA. These embeddings are then fed into a conventional clustering algorithm. **SpatialPCA** follows the probabilistic implementation of PCA in solving a latent factor model [136, 191]. However, instead of assuming independently and identically distributed factors from a standard normal distribution, it uses a Gaussian kernel covariance matrix to model spatial correlation. This aims to encourage similarity in the latent factors of spatial neighbours. Model parameters are inferred by maximum likelihood estimation (MLE), and domains are inferred from the final latent factors through standard clustering. **GraphPCA** calculates a binary spatial adjacency matrix using k-NN [137]. It uses this adjacency to impose a spatial constraint term on the PCA objective function, with a hyperparameter controlling the weight of adjacency. This formulation has a closed-form optimal solution, resulting in low-dimensional representations of the gene expression for each spot or cell. These representations are clustered using k-means.

The majority of modelling-based methods, namely SpiceMix, BayesSpace, BASS, PRECAST, and SC-MEB, employ a hidden Markov random field (HMRF) mechanism. The HMRF represents a latent Markovian variable distribution underlying an observable, and can incorporate prior information. **SpiceMix** combines the HMRF with non-negative matrix factorisation (NMF) [192]. It models the gene expression as a function of underlying metagene mixtures as the latent states. Gene expression is related to the metagenes by an NMF formulation, while the spatial affinity of metagenes is captured in a spatial correlation matrix. The weight of spatial affinities is controlled by a hyperparameter. The parameters of the HMRF model are inferred by MLE, optimised by coordinate ascent. Finally, the inferred metagenes are clustered conventionally into spatial domains. **BayesSpace** is a fully Bayesian method, modelling a low-dimensional representation of the gene expression based on latent cluster affiliations [119]. A Potts model prior encourages similar label assignments to neighbouring spots. Adjacency is defined based on integer coordinates and thus defined only for grid-based data, which BayesSpace was originally designed for. Model parameters and latent variables are inferred using a Markov chain Monte Carlo (MCMC) approach, combining Gibbs sampling for the parameters and a Metropolis-Hastings algorithm to update cluster assignments. **BASS** employs a hierarchical Bayesian framework [82]. It models the relationship between gene expression and underlying spatial domain labels through the intermediate step of cell type labels. Additionally, a Potts model is used to encourage label similarity in neighbouring cells, with neighbourhoods defined through k-NN. A combination of Gibbs sampling and a Metropolis-Hastings algorithm is used for parameter inference. **PRECAST** builds a hierarchical model, connecting gene expression to domain labels through a latent embedding layer [164]. The latent embeddings are modelled using a probabilistic PCA approach, incorporating spatial dependence through a conditional autoregressive approach. Domain labels are then modelled using a GMM, and incorporate a Potts prior within a HMRF formulation. Information on the parameter inference is not available. **SC-MEB** models a dimensionally reduced spot representation as resulting from HMRF latent domain labels, under a spatial smoothness prior [116]. Neighbours are identified using a proximity threshold. The weight assigned to spatial information is adaptively selected via a grid search, while parameters are inferred by an iterative expectation-maximisation (EM) scheme incorporating a pseudo-likelihood maximisation step.

Finally, the last category, which only contains Vesalius, utilises an image processing approach. **Vesalius** starts out by embedding the transcriptome of each spot or cell into an RGB colour space by reducing the dimensionality using PCA, followed by the application of Uniform Manifold Approximation and Projection (UMAP) [98, 140]. Alternatively, it can directly use three selected Principal Components (PCs) for further analysis. From the spot or cell coordinates, it creates a tiling using Voronoi tessellation and converts each tile into a set of pixels via rasterisation. The pixel colours are determined by the colour embedding. Subsequently, image processing is applied using specialised R packages. The pixel array is smoothed by blurring, and colour values are clustered using k-means clustering. These processes may be repeated for different values of hyperparameters and types of blurs. The final colour clusters are then subdivided into spatial domains based on a user-defined

spatial distance threshold.

Fig. 2.1a shows one more method within the clustering-based category, which we call "smoothing". This refers to an optional add-on to non-spatial clustering approaches, implemented as a naïvely spatially aware baseline for the benchmark. It consists of a simple spatial refinement based on local neighbourhood majority voting. Specifically, for all spots or cells in the sample, we define the local spot neighbourhood using k-NN and assign a new spot label as the mode of local neighbourhood labels as assigned by a non-spatial baseline. This results in the removal of visual noise from the domains, as individual spots or cells with locally unique labels get smoothed into the neighbourhood majority. Purely transcriptome-based Leiden clustering, as implemented in the scanpy and Seurat packages, served as a simple baseline for method performance in the entire benchmark [171, 193].

Besides considering the algorithmic variety of method approaches, we are also including methods from a wide temporal range between 2020 and 2024, as indicated by Fig. 2.1b. Methods published after the beginning of 2024 could not be included due to time constraints. As indicated in Fig. 2.1c, we are benchmarking both a number of highly cited methods, as well as more niche approaches. This ensures that our benchmark captures the state of the art, while introducing competitive novel and lesser-known approaches to the broader community. Interestingly, the four most popular methods by number of citations (SpaGCN, BayesSpace, STAGATE, and GraphST) encompass between themselves almost 55% of total citations over all methods. However, recently published methods such as BANKSY and CellCharter show a high proportion of citations in 2025 (status October 2, 2025) compared to their total popularity, possibly indicating an oncoming levelling of the playing field between established methods and novel, optimised approaches. This interpretation is further supported by evaluating the number of "stars" in the methods' respective repositories on GitHub (see Fig. 2.1c). The "stars" metric is utilised here as a proxy for method popularity in the research community, and may reflect method usage in ongoing or future work. It is interesting to consider methods for which the number of stars and the number of citations are discordant. Among others, these may be newly published methods, such as the aforementioned BANKSY and CellCharter, but also GraphPCA, MNMST, or TACCO, which show a disproportionate number of stars compared to total citations. On the other hand, methods like BayesSpace, STAGATE, BASS, PRECAST, or SC-MEB are less often starred than they are cited. This could reflect the GitHub affinity of the respective user bases, as, except for STAGATE, these methods are implemented in the R programming language, rather than in Python.

2.1.2 Dataset selection

For the thorough evaluation of methods, their performance should be assessed in a variety of realistic application scenarios. Particularly, for spatial transcriptomics, we are interested in the method performance across technologies, in order to dissect their broad or differential applicability (see Sec. 1.1.2). Therefore, we aim to include a number of data samples created using different technologies. Many datasets have been made publicly available in the last years, and efforts have been made to categorise and archive those datasets [194–196]. For example, the spatial transcriptomics database (STOmicsDB) lists 361 datasets from 17 species and 128 tissues, while the spatial transcriptomics analysis resource (SOAR) lists 3461 samples from 13 species, 42 tissue types, and 19 technologies (as of September 27, 2025). While many of these datasets are published alongside a cell type level annotation, they by and large do not contain spatial domain annotations. Annotating domains requires expert knowledge and is not yet a standardised processing step in most pipelines. However, an expert annotation of spatial domains, usable as a ground truth, is necessary for accuracy-based method evaluation, as touched upon in Sec. 1.2. This turns out to be the bottleneck for the inclusion of many datasets. We are only able to include datasets annotated with domain labels for each individual spot or cell, so that spot or cell-level method accuracy can be determined.

Upon searching the literature on spatial domain identification, we settle on the inclusion of eight

publicly available datasets for benchmarking as shown in Tab. 2.2. Most of these datasets were published alongside expert domain annotations, while in two cases, domain labels for each spot are sourced from a different publication. We based our dataset selection on the availability of raw count data, binned to single cells in the case of subcellular resolution, and ground truth domain annotation.

The chosen datasets exhibit a wide range of resolutions, from molecular up to spot diameters of 100 µm. Molecular resolution data, binned into single cells for our purposes, is attained by imaging-based approaches like osmFISH, MERFISH and STARmap. On the other hand, sequencing-based techniques profile the gene expression of spots, not necessarily bound by cell boundaries. The spot diameter of the original Spatial Transcriptomics (ST) technique is reduced by half to 55 µm in Visium, and even further by Slide-seq (see Sec. 1.1.2). Besides the resolution, the technique employed for dataset generation also impacts the number of profiled genes. In the imaging-based technologies, the gene panel size ranges from 33 (osmFISH) to over 1000 genes (STARmap). While the sequencing-based technologies Slide-seq, Visium, and ST do not profile a panel of genes but instead unbiasedly sequence the entire transcriptome, the resulting count matrix size varies between 15k and over 30k genes. This is owed to different profiling sensitivity.

Notably, most datasets we were able to include profile tissue sections from the mouse brain. The mouse is a widely employed model organism for the study of mammals, making it unsurprising to see it broadly represented in tissue studies. The brain, on the other hand, is presumably overrepresented in this list of datasets due to its well-studied layer structure, which can be annotated using either accompanying histological images or the expression of well-known marker genes.

2.1.3 Metric selection

After selecting methods and datasets for inclusion in the benchmarking process, we need to decide on criteria for grading method performance. In this section, we will focus on primary evaluation criteria concerning the performance on the spatial domain identification task. Secondary criteria, such as runtime and memory usage, scalability, or method usability, are discussed in Chapter 4 of this thesis.

The performance of computational methods can be evaluated with regard to a diverse set of criteria. These criteria are commonly known as metrics. Contrary to the well-defined mathematical concept of a metric as a distance measure in a metric space, the concept of a metric here refers more broadly to a performance indicator. Generally, metrics can be categorised as "supervised" (that is, method output is compared to a ground truth) or "unsupervised" (considering inherent qualities of the method output). In the context of spatial domain identification, the most common metric types concern the accuracy of label assignments (supervised) and the coherence, or visual smoothness, of those labels (unsupervised).

An overview of supervised and unsupervised metrics considered in this work is shown in Tab. 2.3. For the supervised, accuracy-based evaluation, the Adjusted Rand Index (ARI) is by far the most prevalent metric, as already shown in Fig. 1.7b. Like all supervised metrics, it compares a putative clustering result to a given ground truth to assess the goodness of clustering. Other popular supervised metrics include the Normalised and the Adjusted Mutual Information (NMI and AMI), as well as the Fowlkes-Mallows (FM) index. For basic evaluation, the clustering accuracy (ACC) can be used. On the other hand, there also exist unsupervised metrics, which do not take a ground truth as input and instead compute some inherent criterion for goodness of clustering. Generally, the Silhouette score is a popular evaluation metric in this vein. However, it is not suited to the evaluation of spatial domain assignments in real space due to the approximately uniform distribution of spatial coordinates. It may be utilised for domain evaluation in gene expression space, when gene expression is dimensionally reduced to avoid the curse of dimensionality. For the unsupervised evaluation of domain smoothness, the Percentage of Abnormal Spots (PAS) can be used, adapted from the field of image segmentation.

Dataset name	Technology	Ref. and Download	Origin of annotation	Organism and tissue	# Spl.	# Genes (mean±s.d.)	# Locs (mean±s.d.)
Visium–Maynard	Visium	[114] [a]	expert annotation based on histology, t-SNE and marker gene expression	human, dorsolateral prefrontal cortex	12	33538 ± 0	3944.1 ± 462.0
STARmap–Wang	STARmap	[48] [b]	marker genes and anatomical annotation	mouse, primary visual cortex	1	1020	1207
MERFISH–Zhang	MERFISH	[197] [c] ([198])	marker genes	mouse, primary motor cortex	33	254 ± 0	4890.4 ± 1431.7
MERFISH–Moffitt	MERFISH	[199] [d] ([82])	spatial gene expression patterns and histology from Allen brain atlas [71]	mouse, hypothalamic preoptic region	5	155 ± 0	5663.4 ± 190.4
osmFISH–Codeluppi	osmFISH	[44] [e]	marker genes	mouse, somatosensory cortex	1	33	4839
Visium–Fu	Visium	[f] ([170])	pathological features and cell type annotations	human, breast cancer	1	36601	3798
ST–Stahl	ST	[37] [g]	gene expression	mouse, olfactory bulb	12	15733.3 ± 915.7	259.7 ± 18.8
Slide-seq–Langlieb	Slide-seq	[200] [h]	label transfer and marker genes	mouse, whole brain	10	27719 ± 915.7	3447.0 ± 1008.3

Table 2.2: **Real datasets utilised for benchmarking.** Name, publication of origin and download source of all public datasets included in this study, along with the download source of the respective ground truth annotations, where they differ from the data source. Further information on each dataset includes the organism and tissue of origin, the number of samples it contains, as well as the number of genes and locations (cells or spots) profiled. ST, Spatial Transcriptomics; t-SNE, t-distributed stochastic neighbour embedding [201].
[a] research.libd.org/spatialLIBD [b] zenodo.org/records/10698912 [c] doi.brainimagelibrary.org/doi/10.35077/g.21
[d] datadryad.org/dataset/doi:10.5061/dryad.8t8s248 [e] linnarssonlab.org/osmFISH/availability/
[f] 10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Breast_Cancer_Block_A_Section_1
[g] spatialresearch.org/resources-published-datasets/doi-10-1126science-aaf2403 [h] braincelldata.org

Type	Metric	Ref.	Definition	Characteristics
Supervised	Adjusted Rand Index (ARI)	[202–204]	Compares assignments of pairwise spots in true labelling and putative clustering. Rand index: $$R = \frac{TP + TN}{TP + FP + TN + FN}$$ $$ARI = \frac{R - \langle R \rangle_{\text{chance}}}{R_{\text{maximal}} - \langle R \rangle_{\text{chance}}}$$	• Bounded by 1 (perfect match) and -1 (orthogonal clustering), takes value 0 when cluster assignments are equivalent to random chance • Standard measure, high recognition in many fields (see Fig. 1.7), thus comparable and interpretable • Prefers balanced solutions [205]
Supervised	Normalised Mutual Information (NMI)	[206]	Compares clusterings U and V based on their entropies $H(U)$, $H(V)$ and conditional entropies $H(U\|V)$, $H(V\|U)$. Mutual Information: $$I(U,V) = H(U) - H(U\|V) = H(V) - H(V\|U)$$ $$NMI(U,V) = \frac{I(U,V)}{\sqrt{H(U)H(V)}}$$	• Bounded by 0 (no mutual information) and 1 (perfect match) • Prefers pure clusters [205] • Prefers unbalanced solutions [205]
Supervised	Adjusted Mutual Information (AMI)	[207]	$$AMI = \frac{I(U,V) - \langle I(U,V) \rangle}{\max H(U), H(V) - \langle I(U,V) \rangle}$$	• Bounded by 0 (cluster assignments equivalent to random chance) and 1 (perfect match) • For large number of samples, becomes equivalent to NMI, thus similar characteristics [205]

Table 2.3: Continued on next page.

Type	Metric	Ref.	Definition	Characteristics
Supervised	Fowlkes-Mallows index (FM)	[208]	$FM = \sqrt{\frac{TP}{TP+FP}\frac{TP}{TP+FN}}$	• Bounded by 0 (all samples misclassified) and 1 (perfect match) • Equivalent to the F_1 score up to the choice of mean between precision and recall
Supervised	Accuracy (ACC)		Proportion of correct assignments out of all assignments	• Bounded by 0 and 1 trivially • Needs harmonised label names across clusterings
Unsupervised	Silhouette Score (SI)	[209]	For spot i, define a_i as the mean distance between i and all other spots with the same label and b_i as the mean distance between i and all other spots in the next nearest cluster. Then $SI = \sum_i \frac{b_i - a_i}{\max(a_i, b_i)}$	• Bounded by -1 (incorrect clusters) and 1 (highly dense clustering), takes value 0 when clusters overlap • Generally prefers convex, dense, well-separated clusters
Unsupervised	Percentage of Abnormal Spots (PAS)	[136]	Proportion of spots or cells assigned a different label than over half of their k neighbours	• Bounded by 0 and 1 trivially • Sensitive to setting of hyperparameter k

Table 2.3: **Overview of supervised and unsupervised metrics**. Names, original publications and definitions of metrics which compare to a ground truth (supervised) and which perform stand-alone clustering evaluation (unsupervised). A short list of characteristics is provided. TP, true positive; TN, true negative; FP, false positive; FN, false negative.

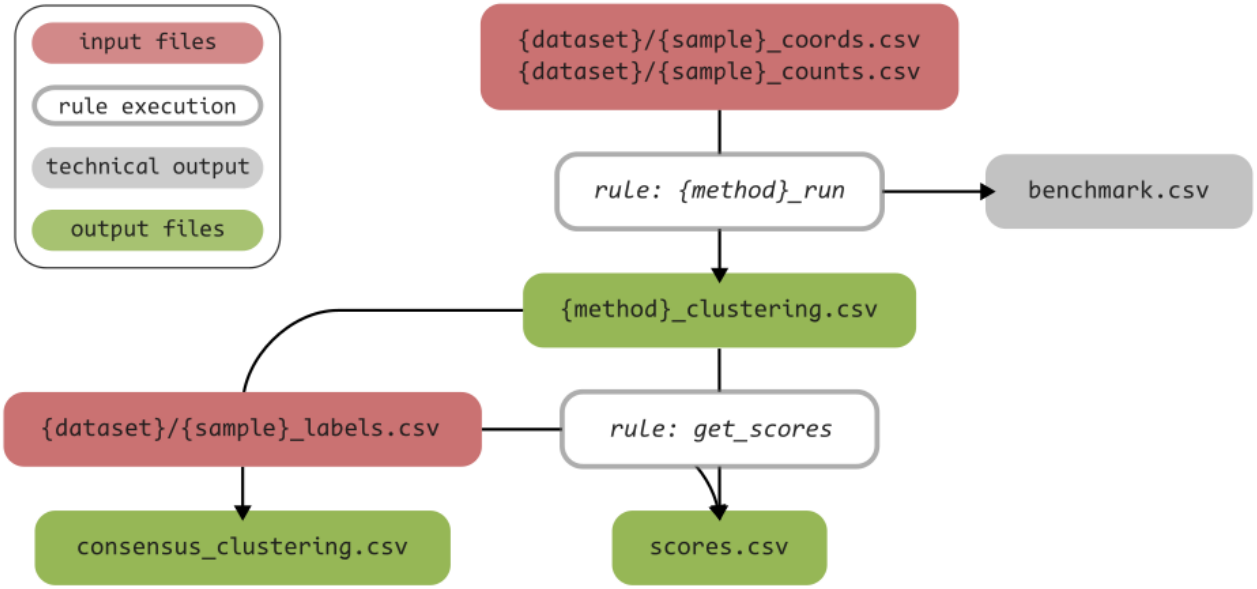

Figure 2.2: **High-level schematic of the benchmarking pipeline.** The benchmarking pipeline shown is implemented using Snakemake [210]. Placeholders for dataset, sample and method names are indicated using curly braces. Input and output files in comma-separated values (CSV) format are shown using solid-colour rectangles, while instances of computing are indicated using a grey border. Red rectangles refer to input files needed by the pipeline, grey rectangles correspond to technical output files provided by Snakemake, and green rectangles refer to individual method outputs and aggregated analysis results. Not shown are the files necessary for the execution of Snakemake rules, such as method scripts and environment files.

2.1.4 Benchmarking pipeline

To create a reproducible and automatised pipeline for benchmarking, we use Snakemake for workflow management [210]. This enables us to define execution rules for each separate method, while alleviating the need to manually run all methods on all included data samples. If the pipeline is restarted, Snakemake automatically recognises which output files already exist, and excludes the corresponding method/data pairs from the rerun. Since every method has specific and unique software environments, we utilise the package manager conda to define method-associated execution environments. Snakemake integrates directly with conda through named YAML files, specifying environment dependencies and packages to be installed.

The final workflow, described in detail in the following paragraphs, consists of Snakemake execution rules, method environment files and scripts, as well as additional rules and scripts to calculate metric scores based on method outputs. A schematic view of the entire pipeline is shown in Fig. 2.2.

For each sample, the pipeline expects a set of files in the comma-separated values (CSV) format. The files necessary for method execution are `coords.csv`, containing a set of (x, y) coordinates for each profiled spot or cell, and `counts.csv`, with the corresponding molecular measurements in the form of a count matrix. Counts and coordinates are matched up by index, so it is paramount to ensure equal ordering of spots or cells between the two files. Further, the `labels.csv` file is required for the calculation of supervised metrics on the corresponding sample. This file should contain the ground truth domain annotations in the form of one column of label assignments for each spot or cell. Again, the ordering must match the other files exactly. If the labels file happens to be missing, unsupervised metrics are still calculated, but supervised metrics measuring method accuracy cannot be evaluated.

For each method, Snakemake requires instructions on the location of both environment files and

the script, and the exact command to execute the method correctly. Environment files are primarily given to conda in the YAML format[1] However, certain packages, notably for the R language, are not available through conda. In these cases, Snakemake accepts a bash script, which it will execute after the installation of all available conda-based packages. Through this script, packages or methods can be installed from the pip package manager, or, using an additional Rscript file called from the bash script, from Bioconductor.

The outputs of the Snakemake-based pipeline are defined within each rule. In our case, they consist primarily of `clustering.csv` files containing label assignments, for each spot or cell. One such file is produced per method and per sample. Additionally, the scores attained by each method on each sample, across the calculated metrics, are summarised in one `scores.csv` file. Finally, each method/sample combination produces a file called `benchmark.csv`, in which Snakemake automatically compiles information about the runtime and memory usage of the corresponding run.

Implementing this comprehensive pipeline provides us with a resource enabling us to combine and cross-evaluate independent methods. Besides benchmarking each method, we therefore investigate two analysis approaches that go beyond individual method output. The results of a chimerisation analysis of neural network-based methods will be published in the article presenting our benchmarking work, and will not be discussed here in more detail. Further, we implement a consensus approach across methods and leverage the ground truth to evaluate spot-wise method agreement. This enables us to distinguish tissue areas that are easily detected as regions by most methods versus other areas that provide a challenge for the state of the art, guiding future method development.

2.2 Evaluation of method accuracy

This section shows the results of evaluating method performance on the previously introduced datasets, focusing on supervised metrics. We first compare the supervised metrics ARI, FM, AMI, NMI, and accuracy (ACC), describing example cases where methods' scores between these metrics do not coincide. Then, we expand on the varying method performance on different datasets, followed by the description of a consensus evaluation approach.

2.2.1 Comparison of supervised metrics

In the evaluation pipeline, we implemented the supervised metrics ARI, FM, AMI, NMI, and ACC. These metrics compare a putative clustering output to the ground truth, using different underlying principles as described in Tab. 2.3. ARI is by far the most prevalent metric in the field (see Fig. 1.7b), and utilising it as the main accuracy metric would therefore enhance the comparability of our results. In order to solidify the validity of this choice, in this section, we compare the behaviour of ARI and the other metrics.

Over all datasets and methods, the supervised metrics FM, AMI, NMI, and ACC show qualitatively the same behaviour as the ARI (Fig. 2.3a). In order to exclude the possibility of dataset-specific or method-specific biases, we evaluate the Pearson correlations per dataset and per method (Fig. 2.3b,c). For all datasets and nearly all methods, the correlations by far exceed 0.8, indicating that it is justified to focus on evaluation by ARI alone. Furthermore, AMI and NMI are correlated with ARI (Pearson > 0.86) across methods.

Both ARI and AMI are adjusted metrics, meaning that they incorporate corrections to the underlying scores. Specifically, they account for chance effects by discounting the expected (non-adjusted) score values for a random clustering. Similarly, NMI is normalised to 0 for no mutual information and

[1] YAML, according to the official website (yaml.org/spec/), stood originally (until Working Draft 10 December 2001) for Yet Another Markup Language, before being changed (from Working Draft 07 April 2002) to YAML Ain't Markup Language. Quite the turnaround.

1 for perfect coincidence. On the other hand, the FM and ACC metrics are not adjusted for chance. This leads them to disagree with ARI in select cases, examples of which are shown in Fig. 2.3d. Disparities between ARI and the adjusted metrics AMI and NMI are rarer and mostly occur in corner cases of small or many domains. The example in Fig. 2.3d demonstrates the advantageous evaluation of pure clusters by the mutual information-based metrics AMI and NMI, as described in Tab. 2.3. In all three examples, ARI is qualitatively better suited to quantifying method performance.

Based on these evaluations, in the following and throughout this thesis, we focus on ARI as the main supervised metric for sample-wide accuracy evaluation.

2.2.2 Accuracy across datasets

For the evaluation of general method accuracy, we utilise the mean ARI score of a method m on a dataset d, which we call $\mathrm{ARI}_{d,m}$. We calculate a method ranking based on aggregating performances across datasets. In order to avoid datasets with highly variable performance between methods from dominating the ranking, scores are standardised per dataset. Standardisation, alternatively known as z-score transformation or simply normalisation, refers to a scaling procedure intended to increase the comparability between groups. Specifically, the original $\mathrm{ARI}_{d,m}$ is scaled to the standardised $\mathrm{ARI}^s_{d,m}$ as

$$\text{standardised ARI} = \mathrm{ARI}^s_{d,m} = \frac{\mathrm{ARI}_{d,m} - \mu_d}{\sigma_d}, \tag{2.1}$$

where μ_d and σ_d are the mean and standard deviation of $\mathrm{ARI}_{d,m}$ on dataset d. The standardised ARIs are subsequently aggregated by mean over all datasets into an overall performance measure for each method. Methods are ranked according to this measure, from best to worst overall performance. This ranking is employed across Figs. 2.4a,b,c.

The standardisation step ensures that datasets on which methods show highly variable perfomance are weighted similarly to less variable datasets. This is necessary primarily to account for method performance on the dataset of the mouse olfactory bulb, sequenced using ST by Stahl *et al.* (see Fig. 2.4a). Compared to the other datasets, methods behave in a highly variable and unique way on ST–Stahl, with the best-performing methods being different to any other dataset. Notably, the non-spatial baseline methods scanpy and Seurat here outperform all but one method. The only spatial method performing better than the baselines, SpiceMix, does so by a marginal amount, as shown in Fig. 2.4b.

In the case of the ST–Stahl dataset, spatial information appears not only not to aid the domain recognition, but might be actively hindering it. Considering the UMAP embedding of transcriptional profiles from one ST–Stahl sample shown in Fig. 2.4d, we see that the ground truth label assignments coincide visually very well with transcriptionally defined clusters. This is confirmed quantitatively by the average Silhouette score of the ST–Stahl ground truth labels exceeding 0.6, in contrast to values of under 0.2 on the other datasets (Fig. 2.4f). On the other hand, in the spatial plot of the same ST–Stahl sample (Fig. 2.4e), the domains are spatially contiguous but very thin, mostly only being one spot wide. Combined, this indicates that the incorporation of spatial information in the non-baseline methods might lead to an overemphasis on contiguous cluster-building in real space, smoothing over relevant transcriptional differences. The influence of spatial smoothing will be considered in more detail in later sections.

In all other datasets, most spatially-informed methods outperform the baselines (see Fig. 2.4b,c). The extent of improvement that is achieved varies substantially across the datasets. In the two Visium datasets, the maximal ARI improvements are 0.1 and 0.2 (for Fu and Maynard, respectively, corresponding to factor improvements of of 1.2 and 1.5). In contrast, for the MERFISH datasets, BASS improved upon the best baseline by 0.5, an increase by a factor of 4.5, in both datasets. Possible reasons for this strong contrast between Visium and MERFISH are discussed in the following sections. SpaDo reached the biggest improvements on the osmFISH and STARmap samples (ARI increases of

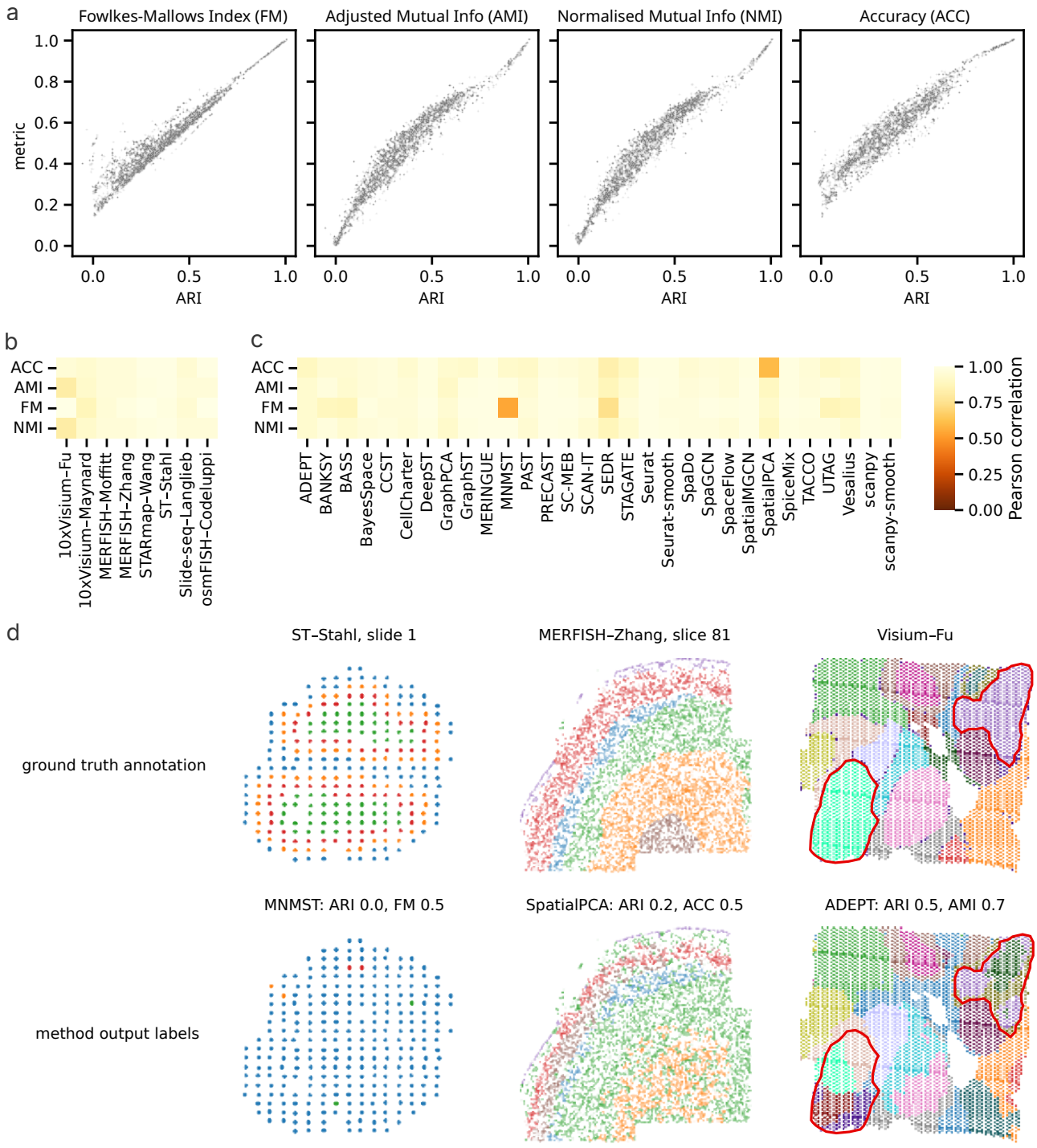

Figure 2.3: **Evaluation of FM, AMI, NMI, and ACC versus ARI.** a, Values across all datasets, samples and methods qualitatively agree with ARI values. b, Stratified by dataset and method, Pearson correlation between the metrics and ARI consistently exceeds 0.8, with few exceptions. c, Example results for methods and datasets where some metrics do not agree with ARI. Top row, ground truth label assignments for three samples from the ST–Stahl, MERFISH–Zhang, and Visium–Fu datasets. Bottom row, example label assignments for those same samples from MNMST, SpatialPCA, and ADEPT, with the corresponding values of ARI and one other metric. The left column illustrates the advantage of ARI over FM in accounting for chance effects in an example with very small domains. In the middle column, the layered structure found by SpatialPCA upon closer inspection does not correspond well to the ground truth. The right column illustrates the tendency of AMI to favour pure clusters. For example, the bottom left (neon turquoise) and the top right (purple) ground truth domains, marked with red borders, are split by ADEPT into multiple constituent clusters.

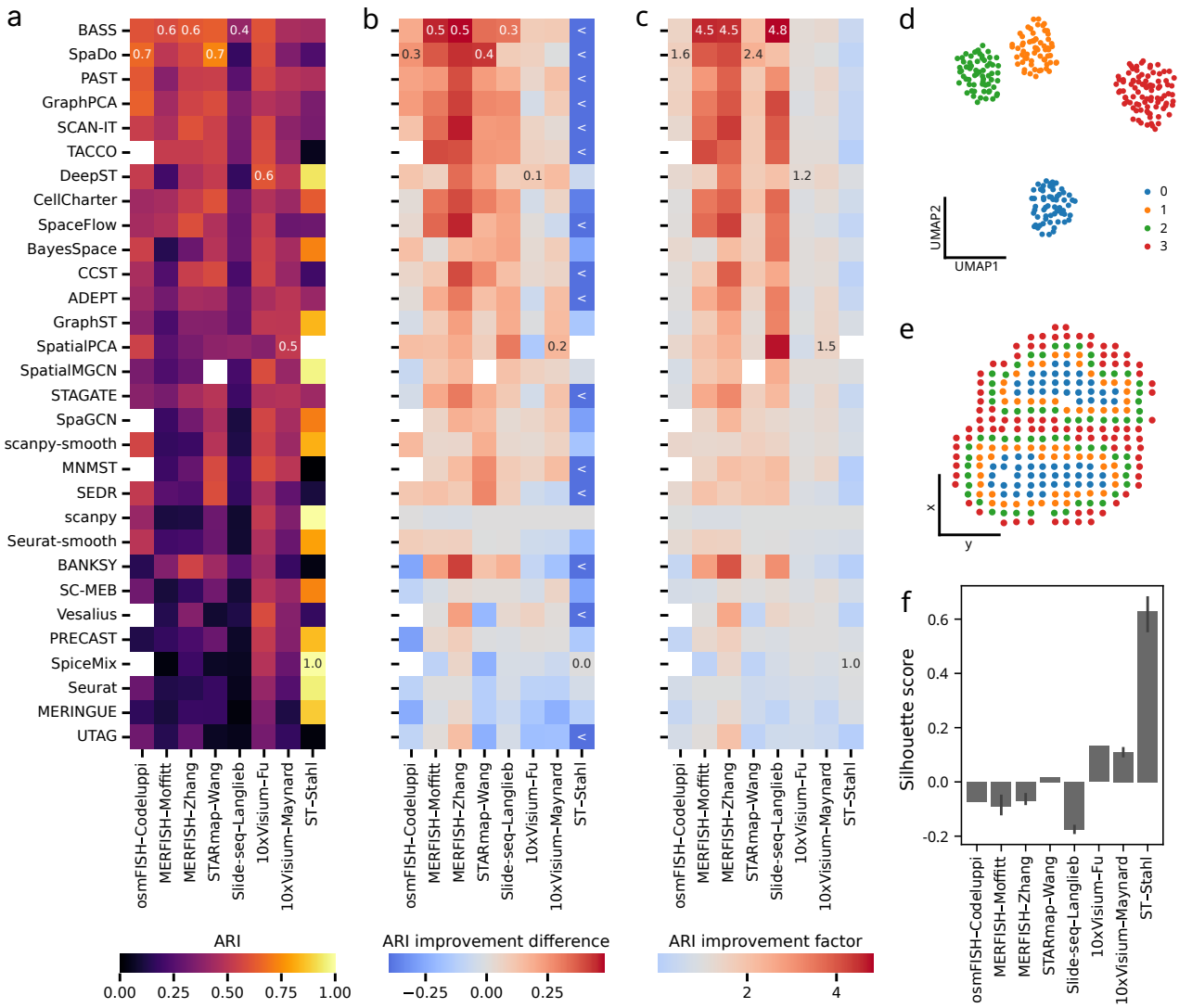

Figure 2.4: **Method accuracy stratified by dataset.** a-c, Accuracy results per method and dataset, aggregated by mean over samples. Methods are sorted by rank according to the mean standardised ARI across datasets, and the best-performing method on each dataset is annotated by value. Datasets are sorted according to their resolution. White squares indicate missing values, i.e. a method failing to give spatial domain assignments. a, Unscaled ARIs. b, Difference in ARI between each method and the best-performing non-spatial baseline (scanpy or Seurat). The colour scale is truncated at -0.4 to emphasise relevant performance differences, and a less-than $<$ indicates that the absolute difference in ARI of a method to the baseline exceeds 0.4. c, Factor of improvement of ARIs of all methods on all datasets over the maximum performance of non-spatial baselines. d, 2D UMAP embedding of spots in ST–Stahl slide 1. e, Spatial 2D-plot of the spots in ST–Stahl slide 1. f, Mean Silhouette scores of ground truth label assignments based on 2D UMAP embeddings. The silhouette score is an unsupervised goodness-of-clustering measure (see Tab. 2.3). The best attainable value is 1; values near 0 indicate overlapping clusters.

0.3 and 0.4, respectively). These improvements are less impressive as factors of the best baseline (1.4 and 2.6, respectively), as baseline performance on these samples is not as low as on MERFISH (see e.g. scanpy in Fig. 2.4a).

Notably, the biggest improvement by factor, 4.8, was reached by BASS on the Slide-seq dataset (see Fig. 2.4c). This is an interesting case, as the final ARI attained by the best method BASS is at 0.4 still quite low, but presents a difference of 0.3 to the best baseline. The Slide-seq technology is known to have a comparatively low capture efficiency[2] [40]. Accordingly, the Slide-seq–Langlieb dataset has a counts sparsity of up to 99%.

As can be expected, the performance of the baseline methods scales with the mean Silhouette score in the transcriptional embedding space (compare Fig. 2.4a, f). Where the Silhouette score is around zero or even negative, as in the case of the high-resolution datasets, non-spatial baselines do not achieve high accuracies. In these cases, the inclusion of spatial coordinates to inform domain identification aids some specialised methods to improve their performance upon the baselines.

2.2.3 Consensus across methods

Combining individual clusterings of the same data into a final consensus, or ensemble, clustering has been shown to improve the robustness and accuracy of the final clustering [126, 206, 211, 212]. We therefore undertake a consensus evaluation across individual methods. To compute the consensus annotation, we take a naïve approach of simply considering the most common label assigned to each spot. Concretely, we compute the mode of labels for each spot or cell in a given sample.

In taking the mode, this approach depends on a coherent labelling of domains across methods. That is, domain A as assigned by method X must correspond, as closely as possible, to domain A as assigned by method Y, for their consensus domain A to remain consistent. However, as the methods do not have any information about the ground truth annotations, they each assign independent domain names with no clear correspondence between them. In our case, as our samples all contain ground truth label assignments, we are able to utilise this ground truth annotation to harmonise method outputs (Fig. 2.5a). This is commonly known within combinatorics as the assignment problem, and can be formulated within graph theory as a maximum weight bipartite matching problem (Fig. 2.5b). Specifically, the domain labels present in the ground truth and in a given clustering output are considered as nodes. Edges between a ground truth label i and a putative label j are weighted according to the number of spots belonging to both in the respective clusterings, creating a cost matrix C with

$$C_{ij} = \sum_{\langle t,p\rangle \in S} \delta_{ti}\delta_{pj}, \tag{2.2}$$

where S is the set of spots or cells in the clustering, and t and p are the true and putative clustering labels of each spot or cell. The problem then is to find a boolean matching matrix X of ground truth labels to putative labels which maximises the total cost. The optimal matching is found by solving

$$\text{argmax}_X \sum_i \sum_j C_{ij}X_{ij}. \tag{2.3}$$

In our implementation, we utilise the linear sum assignment solver by the scipy package [213]. This solver is capable of solving the generalised case in which the number of labels in the ground truth and putative clusterings is different, that is, both C and X are rectangular. While this harmonisation of labels depends on the existence of a ground truth, this could be circumvented by calculating the correspondence of method outputs independently. Also, nota bene, using this simple consensus approach, the resulting domain annotation may not contain all domains present in the ground truth.

[2]While Slide-seqV2 has been shown to compare to Visium in capture efficiency, the original Slide-seq technique captured as little as 10% of transcripts compared to Slide-seqV2 [40].

Name	Optimised for:	Included methods
all	not optimised	all methods
best	overall highest performance	BASS, SpaDo, GraphPCA, SCAN-IT
merfish	MERFISH performance	BANKSY, SpaceFlow, BASS, CCST, SCAN-IT
visium	Visium performance	PAST, GraphST, SpatialPCA, MNMST, GraphPCA

Table 2.4: **Groups of methods investigated using consensus approach.** Besides the unbiased consensus over all method outputs, we evaluate method groups selected for the overall highest performance and technology-specific performances.

We evaluate the consensus across all methods, without filtering for individual performance. Additionally, we select groups of five methods each that excel in the overall ranking, and on MERFISH/Visium data (see Tab. 2.4) and use those to compute the consensus. Astonishingly, as seen in Fig. 2.5, all resulting clusterings outperform all but the top-ranking individual methods on most datasets (Fig. 2.5c). Notably, the selective consensus approaches show better performances than the overall consensus in certain datasets (Fig. 2.5d). Consensus over the method group selected for MERFISH performance outperforms both the consensus over the overall best methods (consensus-best) and the best individual method on the MERFISH–Zhang dataset, while consensus-best narrowly improves upon the others on MERFISH–Moffitt. Interestingly, the consensus evaluation over all individual methods is unique in the magnitude of its improvement on both Visium datasets (mean ARI increase of 0.25 compared to the best non-spatial baseline, and 0.12 compared to the best individual method). Even taking the consensus over methods specifically selected for their high Visium performance (see Tab. 2.4) does not result in a comparable improvement (mean ARI increase of 0.18 over the best non-spatial baseline). On the Visium dataset, consensus-best performs similarly to the consensus over an especially selected method group. Lastly, on ST–Stahl, the aggregation over all methods, uniquely among the consensus approaches, recovers the high performance of the best baseline and best individual method.

Overall, the consensus over all methods matches or outperforms individual methods on all but selected datasets. Except for the MERFISH–Moffitt and STARmap–Wang datasets, this unbiased consensus exhibits highly competitive performance. While the performance on individual datasets can be improved by targeted consensus approaches over method subsets, the unbiased consensus approach is a stable and competitive alternative.

2.3 Visual smoothness effect

From the method performances observed on ST–Stahl, we hypothesised that an exaggerated reliance on visual smoothness and coherence may lead methods to neglect relevant transcriptional differences between clusters. This is contrasted to the performance of our naïvely spatially aware baselines Seurat-smooth and scanpy-smooth, which rank higher than their non-smoothed counterparts in overall performance (by 6 and 3 ranks, respectively, see Fig. 2.4). In this section, therefore, we investigate the interplay between the spatial smoothness of label assignments and method accuracy.

2.3.1 Quantitative evaluation of visual smoothness

For a quantitative evaluation of the visual smoothness of a cluster assignment, we use the Percentage of Abnormal Spots (PAS), as shown in Tab. 2.3. For our purposes, a spot or cell is considered "abnormal" if it is assigned a different label than over half of its spatial neighbours. We assign neighbours based on a 10-nearest-neighbour scheme.

The ground truth PAS of the MERFISH, STARmap and Slide-seq datasets exceeds that of Visium–

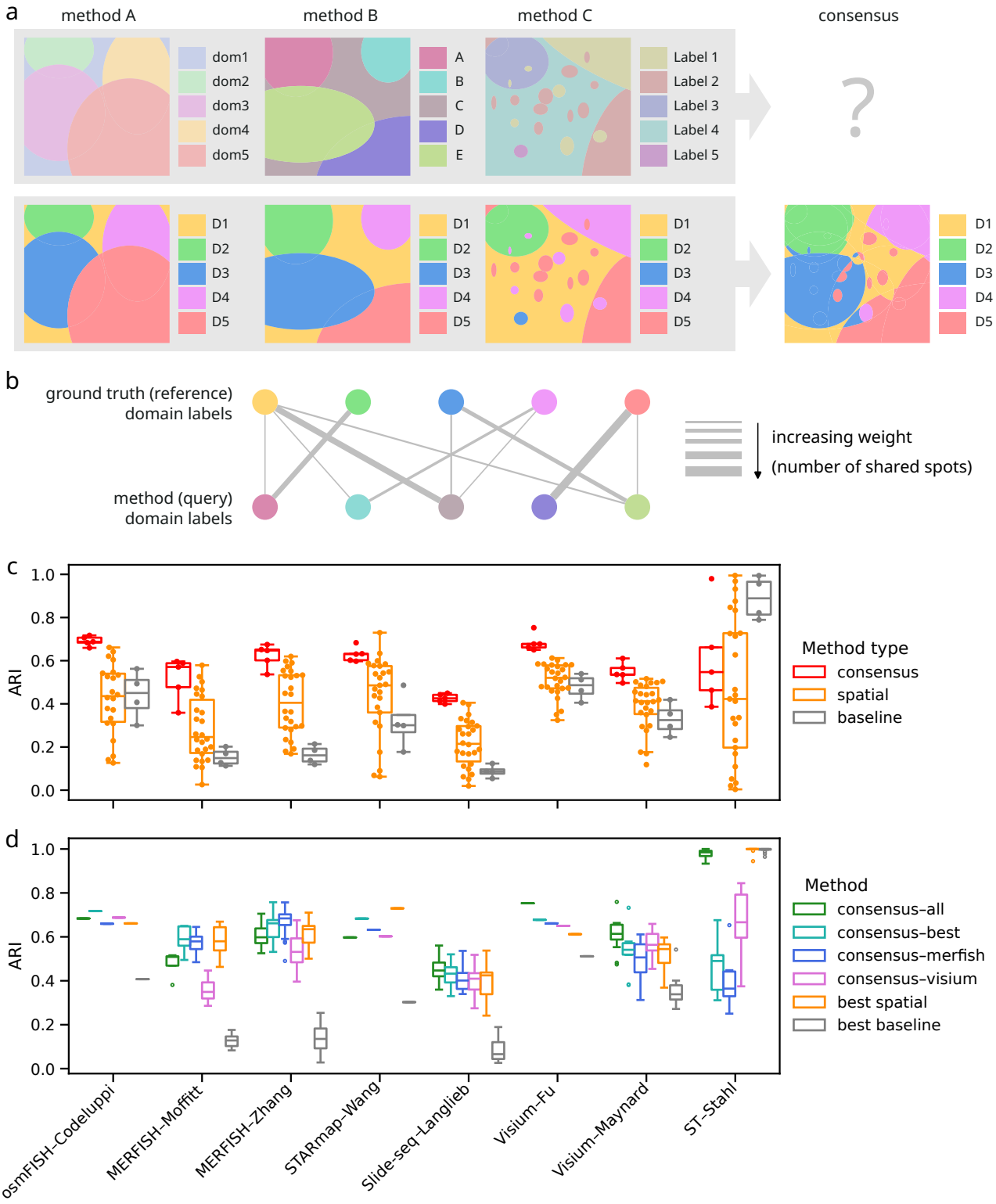

Figure 2.5: **Consensus approach over multiple methods.** a, Domain labels as output from different methods (top) cannot be directly used in our consensus approach; instead, they are harmonised (bottom) to result in a final consensus labelling. b, Harmonisation of method output labels with respect to ground truth domain annotations by maximum weight bipartite matching. c, Consensus approaches outperform individual methods on average, except for baselines on ST–Stahl. Performance is shown per method as the mean over samples per dataset. d, Performance of consensus-all is competitive on all datasets, and improves over other approaches on Visium. Consensus approaches are compared to the best spatial and the best baseline method specific to each dataset, the performance is shown as boxplots over samples in each dataset.

Maynard (see Fig. 2.6c). While Visium–Fu has the highest ground truth PAS of all datasets except ST–Stahl[3], this might also be an effect of Visium–Fu having a comparatively high number of ground truth domains (20, comparing to 4–10 for the other datasets). For evaluating the ground truth, PAS mostly is a measure of border smoothness, as having a higher number of border spots (resulting from having more domains) will lead to an increase in PAS. Notably, Visium–Fu also exhibits some non-contiguous and fragmented domains in the ground truth annotation (see Appendix B), further increasing its PAS. Nevertheless, for all non-ST datasets, the ground truth PAS is under 10%.

2.3.2 Smoothness and accuracy across technologies

Specifically, we focus on the comparison of the datasets MERFISH–Zhang and Visium–Maynard. MERFISH and Visium represent opposite ends of the spectra of both spatial resolution (single-molecule, segmented into single cells, versus spots of 55 µm) and number of profiled genes (gene panel size of 254 versus full-transcriptome). Additionally, both contain over 10 samples, making a statistical analysis more viable, and exhibit comparable laminar tissue structure originating from brain tissue (of mouse or human).

Most methods (23 out of 30) differ significantly in performance between these two datasets (as measured by the Mann-Whitney U test, significance $p < 0.05$, see Fig. 2.6a). The differences seem to primarily be driven by strong intra-dataset performance differences on MERFISH–Zhang. Evaluating the average spatial smoothness per dataset and method, a trend becomes apparent whereby method performance on MERFISH–Zhang appears to correlate negatively with PAS. Concretely, all methods that perform significantly better on MERFISH than on Visium, except the generally less accurate UTAG and MERINGUE, exhibit mean PAS $\leq 13\%$. For methods which perform significantly better on Visium than on MERFISH, on the other hand, PAS values, especially on the MERFISH data, are higher. Interestingly, this set of methods also shows non-negligible PAS on Visium.

Stratifying the datasets by resolution[4] confirms the hypothesis of a strong negative ARI-PAS correlation on all high-resolution samples (median Spearman correlation -0.85, see Fig. 2.6b). For the low-resolution Visium samples, the anticorrelation is much weaker (median Spearman correlation -0.31). Method performance on high-resolution datasets therefore seems to benefit from enforcing high visual smoothness, while this does not aid performance as much in the lower-resolution Visium data. It is not the case that the ground truth PAS on Visium data is generally higher, as seen in the previous subsection.

This leads naturally to the question of whether a simple smoothing step improves method performance to a larger extent on MERFISH data than on Visium. We evaluate this question by considering the difference in performance between the spatially-aware and non-spatially-aware baselines. Indeed, the improvement in ARI attained by a simple smoothing over the non-spatial baseline output is significantly larger on MERFISH–Zhang than on Visium–Maynard (average improvement of 0.071 vs 0.057, see Fig. 2.6d). As expected, the PAS decreases strongly by applying the smoothing. The decrease shows highly significant differences between the two datasets: On MERFISH–Moffitt, where the non-smoothed baselines have a mean PAS of 51%, the smoothing reduces the visual noise dramatically, to a mean PAS of 20% for a total reduction of 31%. On the other hand, on Visium–Maynard, the difference is less pronounced, going from 36% to 12% for a total of 24%.

The average PAS across methods is around 20% for the high-resolution datasets, contrasted with under 10% on Visium–Maynard (see Fig. 2.6c). Interestingly, even methods with very high PAS on MERFISH–Zhang (as shown in Fig. 2.6a) exhibit lower PAS scores on Visium. This is the case for all methods except UTAG. A fundamental technological difference which might account for this disparity

[3]The extraordinarily high PAS values of ST–Stahl are a consequence of its thin domains as shown in Fig. 2.4e.

[4]Visium datasets are categorised as low-resolution, Slide-seq, STARmap and the FISH-based datasets as high-resolution. ST–Stahl is not considered in this evaluation due to its unique behaviour.

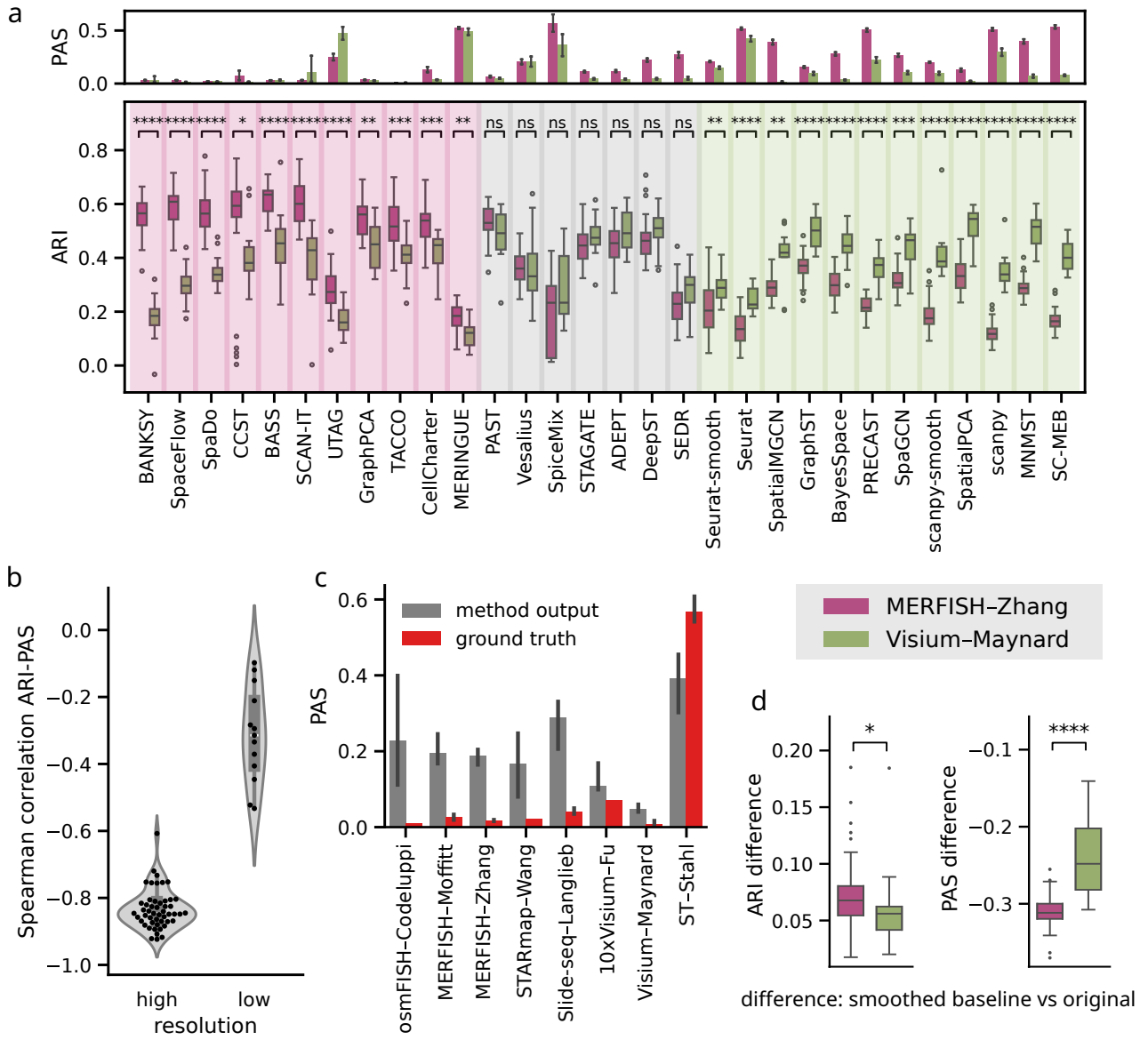

Figure 2.6: **Relationship between accuracy and spatial smoothness.** a, ARI scores of all methods on two datasets generated by MERFISH or Visium technologies. Methods are ordered by the difference in mean ARI between the two technologies. Statistical significance was assessed using two-sided Mann-Whitney U tests. Corresponding PAS values for each method on the same datasets are shown in the top panel. b, Spearman correlation of ARI and PAS over all methods on data stratified broadly by resolution (Visium datasets are designated low resolution, all other technologies high resolution. ST–Stahl is excluded from this analysis). c, Median PAS of method outputs is consistently higher than the median ground truth PAS, except for ST–Stahl. PAS of method outputs is lower for lower-resolution Visium datasets, while no such effect in the ground truth PAS is apparent. d, ARI and PAS of the baselines with and without additional smoothing step, on MERFISH–Zhang and Visium–Maynard.

is the resolution of the respective technologies. We hypothesise that in Visium data, a kind of inherent smoothing is taking place in the gene expression space. Through the measurement of multiple cells in a spot, large transcriptional differences between adjacent spots, which might be reflected in method outputs as different label assignments, will not be registered, thus leading the methods to a lower PAS.

2.4 Domain-specific phenomena

When considering the accuracy attained over an entire tissue slice, method performance can be adequately quantified by ARI or similar metrics. However, ARI does not recognise domain-specific aspects of the results, such as some domains being more easily identified than others. It is particularly interesting for future method development to be able to distinguish domains which are challenging to identify.

To gain a fine-grained view of "hardness of detection", we evaluate the cross-method agreement with the ground truth. That is, for each spot or cell, the number of methods that "agree" with the ground truth annotation is tallied up. The resulting heatmap of the spots or cells in the sample indicates "conflicting" tissue areas, or parts of domains that few methods annotate correctly. This approach explicitly depends on the presence of a ground truth annotation, where the harmonisation of labels is implemented in the same way as for the consensus evaluation.

We apply this method of consensus agreement on the Visium–Maynard dataset of the human dorsolateral prefrontal cortex [114]. Aggregating over the proportion of labels correctly identified per domain and over all samples, we find that the white matter (WM) and L1 domains are consistently recognised throughout (Fig. 2.7a). However, we also find that the L4 domain evades detection. This is confirmed by looking at the example slice 151675 shown in Fig. 2.7b, showing very low method agreement with the ground truth across the entire L4 domain. The observations about WM and L1 are also confirmed visually, with the caveat that at the borders, especially between WM and its neighbouring layer L6, some uncertainty persists.

Applied to MERFISH–Moffitt (see Fig. 2.7b), the domain-level agreement is particularly low for the paraventricular hypothalamic nucleus (PVH) and periventricular hypothalamic nucleus (PV), and to a lesser degree, the medial preoptic area (MPA). This observation is confounded, again, by looking at the example plots (Fig. 2.7b'). Additionally, we can identify hard-to-distinguish border regions around the intersection of the MPA, PV and the otherwise better-distinguished medial preoptic nucleus (MPN).

As a further application, we consider two slices of the MERFISH–Zhang dataset (shown in Fig. 2.7c'). In the overview barplot across all 33 samples of the dataset (Fig. 2.7c), most domains show relatively consistent distinction levels of around 40–70%. However, the white matter layer (WM) exhibits a larger spread than other layers, which is illustrated by the two example slides selected: In slice 131, only 43 spots clustered together along the sample edge are annotated as WM, whereas in slice 180, the WM domain with 734 spots occupies a sizable area in the same region. In slice 180, the large WM domain shows the same pattern as exhibited in the Visium–Maynard dataset and gets recognised easily by the majority of methods. In contrast, for the very small WM domain of slice 131, method agreement with the ground truth is approaching zero. This is an indication that domain size, besides transcriptional distinctness, may have a profound effect on detection.

An interesting phenomenon occurs in sample 151671 of the Visium–Maynard dataset. By evaluation of the method agreement, we are able to identify a highly persistent and coherent subdomain appearing in L3 (see Fig. 2.7d). Due to time constraints, we were not able to evaluate the biological implications of this subdomain thoroughly. However, the UMAP of L3 spots shown in Fig. 2.7d' indicates that there is an expression-level devide between high-agreement and low-agreement spots.

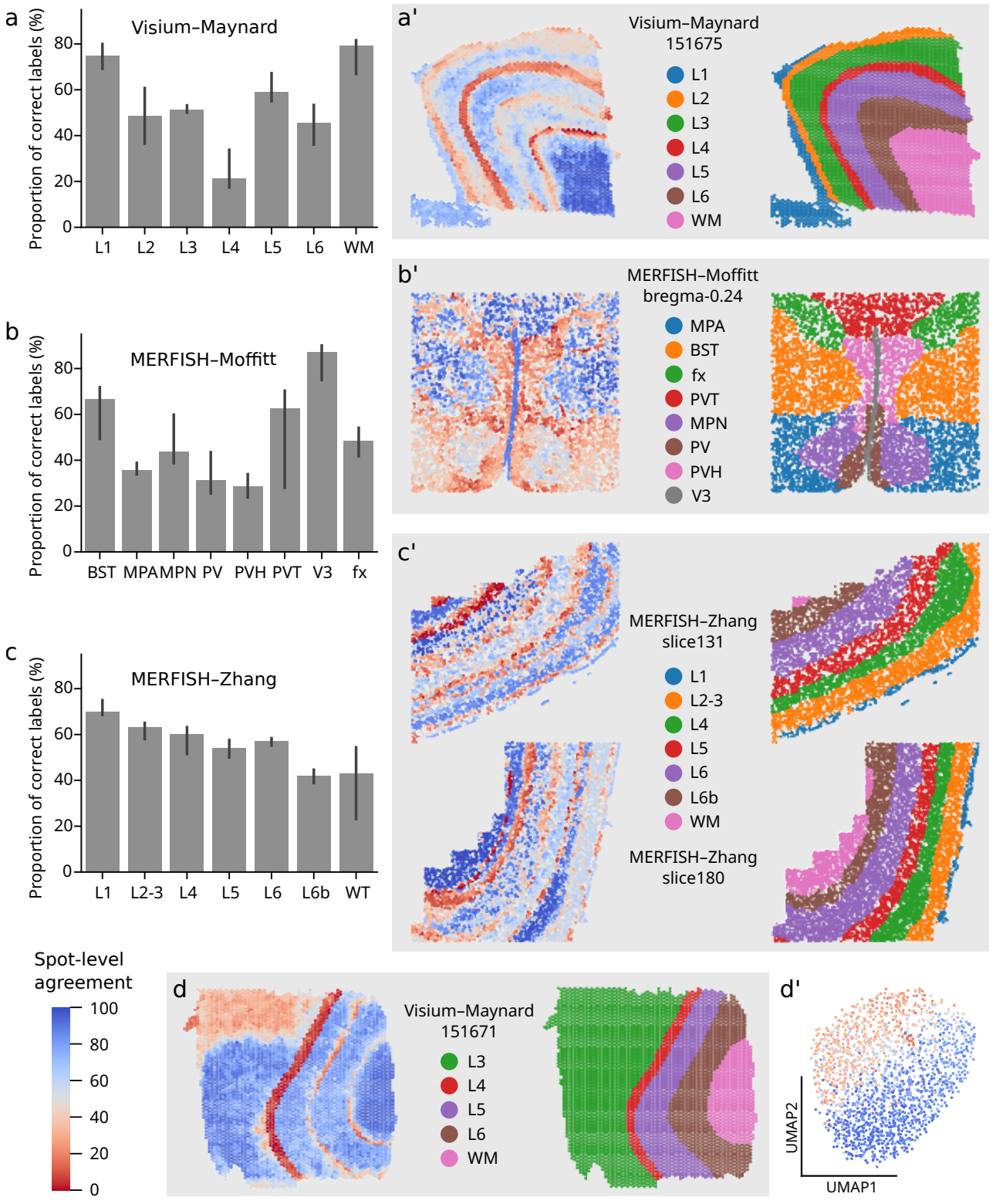

Figure 2.7: **Domain-specific phenomena observed in real datasets.** Selected results are shown for the datasets Visium–Maynard, MERFISH–Moffitt and MERFISH–Zhang. a, b, c, Method agreement with the ground truth across all slices of the three datasets, aggregated per domain by mean. a', b', c', d, Samples from each dataset shown with spot-level method agreement and the corresponding ground truth annotation. d', UMAP representation of spots in the L3 layer of the Visium–Maynard sample 151671, coloured by spot-level agreement.

2.5 Stability with respect to data perturbations

To further evaluate the methods, we test their stability with respect to data perturbations. After establishing a baseline of stochastic effects affecting domain identification, we test the robustness of all methods to the loss of local spatial coherence. We evaluate the stability of methods on the Visium–Maynard and MERFISH–Zhang datasets. These two datasets were chosen as they both contain multiple samples, increasing statistical power, and show a similar laminar domain structure.

2.5.1 Stochastic effects

To establish a baseline measure of within-method variance, methods are run multiple times on the same sample. Since many methods utilise randomness in their implementation, e.g. for initialisation of the clustering or of model parameters, the outputs are often not deterministic [82, 122, 164]. However, it is common practice to fix an internal seed, that is, fix the state underlying random number generation, in order to ensure reproducibility of results. Some methods set the seed internally, so it is not possible to evaluate their stability by simply changing the seed in our implementation [84]. Other methods do not employ randomness in their approach [137].

As a way around this limitation, we devise a strategy for seed-independent evaluation of stochastic method effects. Namely, instead of changing the internal state of method implementations, we change the internal state of the input by reordering the rows of both the count matrix and the tissue locations. We test the effect of this perturbation against changing seeds for four methods containing accessible seeds, and demonstrate that there is no significant change in the distribution of ARI scores (Fig. 2.8a).

For the analysis of pure stochasticity, the inter-sample variability needs to be taken into account. In order to correct for this, ARI scores are scaled linearly per sample such that all sample medians coincide with the dataset median. The results show a wide range of method stability (Fig. 2.8b). CCST stands out for its minimal variability on both datasets, with a mean standard deviation $s.d. < 0.003$. On the other extreme, SpiceMix is particularly unstable, with $s.d. > 0.087$ in the mean across datasets. All other methods are somewhere in between, with notably less variation in standard deviation on the MERFISH dataset. There is no obvious correlation of stability to either high or low method performance, though most overall well-performing methods exhibit consistent standard deviations of around 0.04 across datasets.

2.5.2 Loss of local spatial coherence

Having established a baseline of method stability, we now aim to test the robustness of method performances to perturbation. Specifically, we are interested in the degree to which the loss of local transcriptional coherence between neighbouring spots or cells impairs method performances. We consider local coherence to be given when spatial gene expression patterns are undisturbed.

We again investigate this perturbation on the Visium–Maynard and MERFISH–Zhang datasets. In order to quantify the effect of local coherence loss, we permute count matrix rows within original annotation groups, while keeping physical locations fixed. In this way, each cell gets assigned to new coordinates, but domains are kept intact. The resulting change in the expression patterns of select spatially variable genes is shown in Fig. 2.9a. Briefly, the expression of certain spatially variable genes, previously showing smooth value changes and coherent local maxima, is then spread uniformly over the annotated ground truth domains (shown in the top of Fig. 2.9a'). Interestingly, as also shown in Fig. 2.9a', the local gene expression patterns exhibited by certain genes are reflected in the domains identified by BASS. BASS is selected here as an example because it is the best-performing method over all and performs highly competitively on this dataset. In the domains BASS identifies as L3 and L5, expression patterns from the genes COX6C and SCGB2A2, respectively, seem to be reflected.

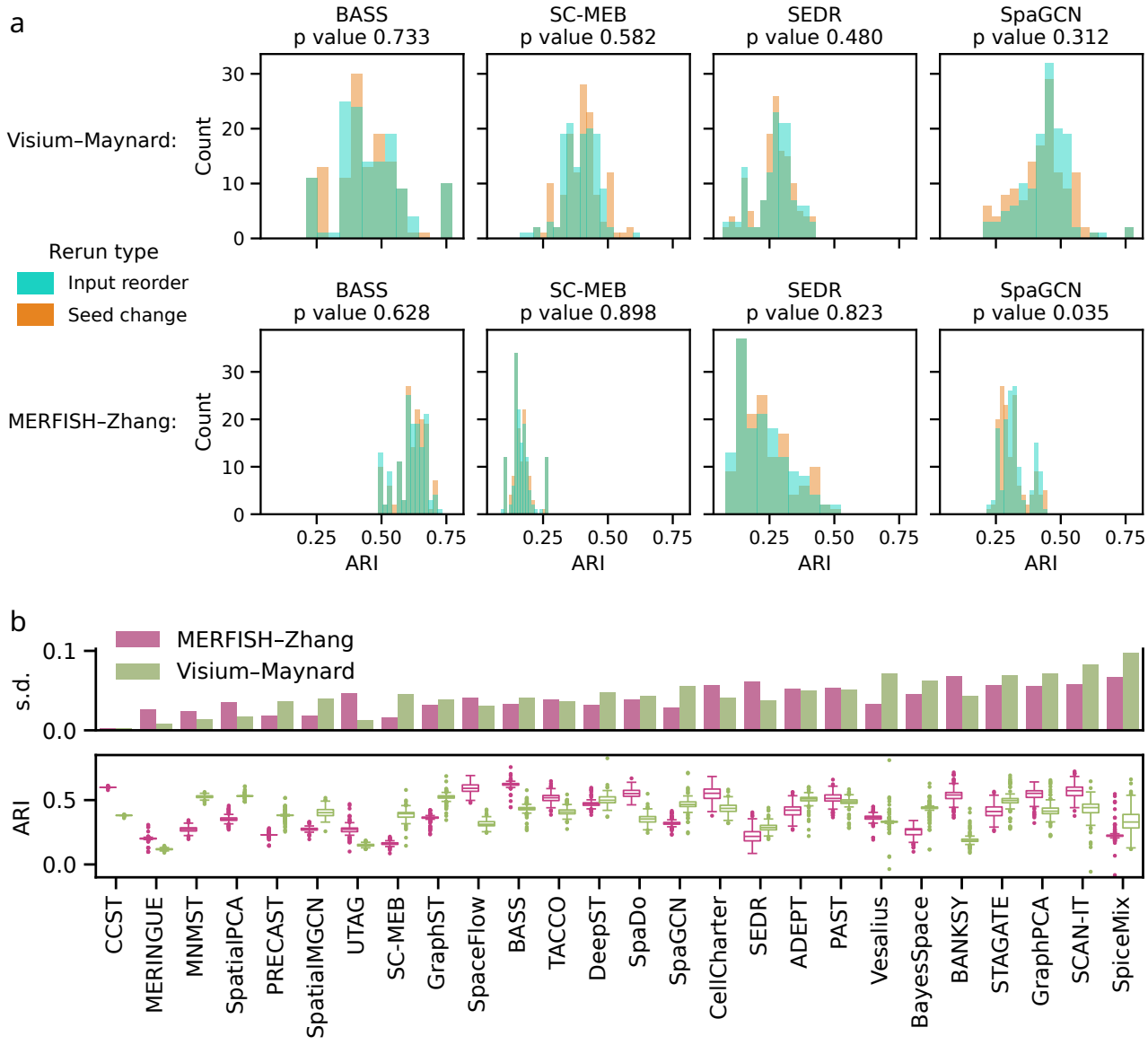

Figure 2.8: **Method stability evaluation on real data.** a, Distribution of results based on seed change versus the proposed method of input reordering. Results are shown for four methods which allow for seed change in their implementation. Twelve input permutations and twelve seed changes are implemented for all samples from Visium–Maynard and twelve samples from MERFISH–Zhang. b, Stochastic variability of domain identification accuracy. ARI scores are reported across 12 random trials of each method on each of the 12 Visium–Maynard samples and 12 MERFISH–Zhang samples. Methods are sorted by their average standard deviation between the two datasets (top bar plot). ARI scores are normalised to correct for inter-sample variability.

Perhaps unsurprisingly, then, the performance of BASS on Visium–Maynard is greatly aided by the removal of these local gene expression patterns (see Fig. 2.9b). While BASS is most strongly affected, increasing in ARI by a median of 0.46, this perturbation has a positive effect on most methods. As expected, the performance of nonspatial baselines is not affected in either dataset, along with methods like MERINGUE, PRECAST, and SC-MEB. Interestingly, the naïvely spatially aware baselines improve more on the Visium–Maynard dataset than on MERFISH–Zhang (by 0.04 ARI). Similarly, BASS, SEDR, MNMST, BayesSpace, and SpatialMGCN exhibit a larger performance improvement on Visium–Maynard. On the other hand, a group of methods comprised of BANKSY, TACCO, UTAG, Vesalius, and GraphPCA shows the opposite behaviour, increasing in performance more strongly on MERFISH–Zhang upon loss of local spatial coherence.

Generally, we observe that the loss of spatial gene expression patterns, which may not necessarily be aligned with annotated spatial domains, affects the majority of methods positively.

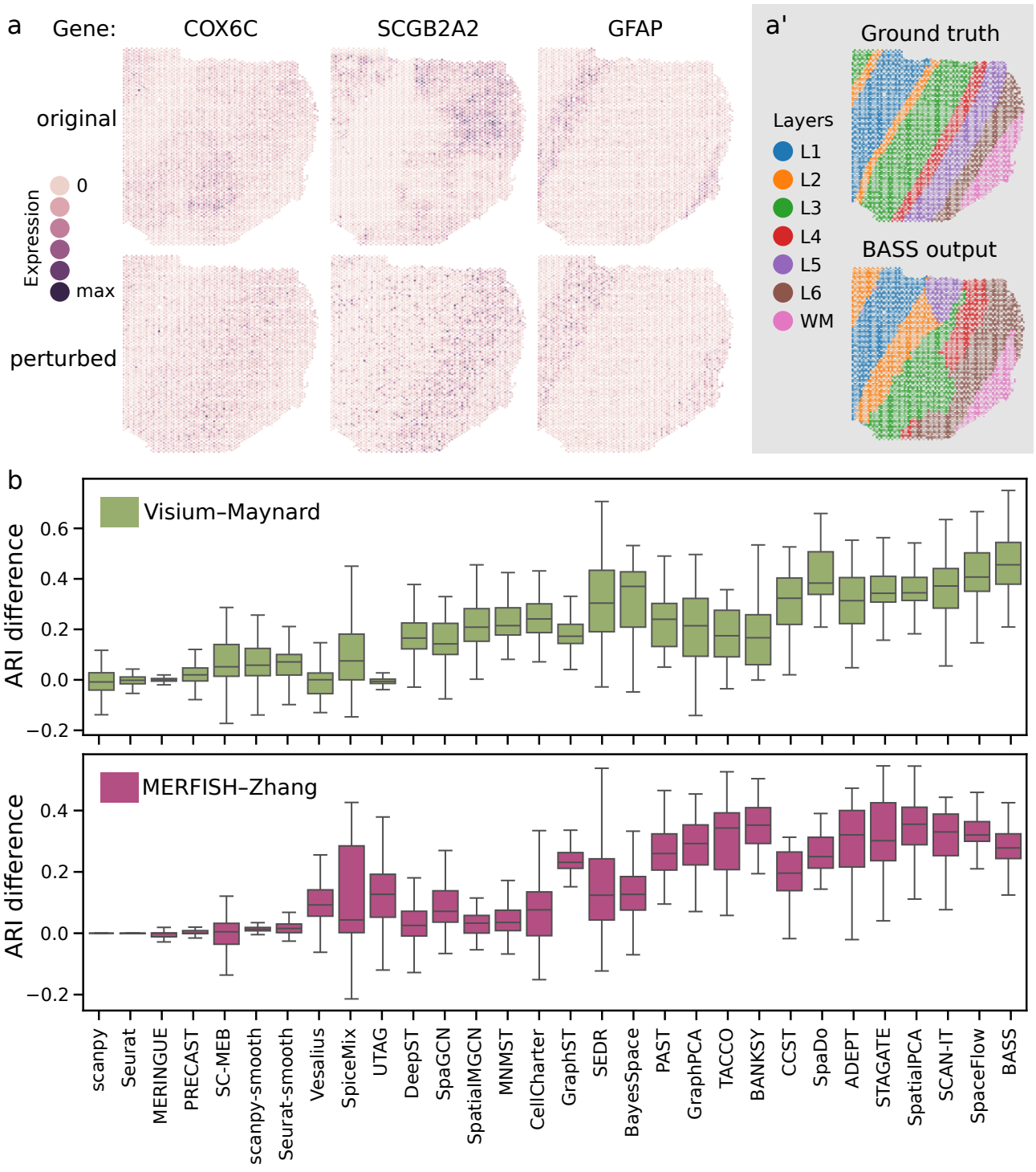

Figure 2.9: **Method robustness to loss of local spatial coherence.** a, Expression of example genes for Visium–Maynard sample 151507, shown in the unperturbed (top) and perturbed (bottom) tissue configurations. a', BASS output on sample 151507 for unperturbed gene expression. Some layers identified by BASS show similarities to gene expression patterns observed in a, namely Layer L5 shows similarity to SCGB2A2 expression and parts of L3 to COX6C. b, Difference in ARI resulting from the perturbation, across 12 samples each of the Visium–Maynard and MERFISH–Zhang dataset. Each sample is perturbed for a total of 12 seeds, the original ARI is subtracted from the perturbed ARI per sample. Methods are sorted according to the mean improvement across datasets.

Chapter 3

Semi-synthetic spatial transcriptomics data for systematic method evaluation

In this chapter, I present the results of running the previously described benchmarking pipeline on semi-synthetic data with tunable characteristics. I introduce several published simulation approaches for spatial transcriptomics data, as well as the pipeline we established for generating semi-synthetic datasets.

3.1 State of the art of spatial transcriptomics simulation

As introduced in Sec. 1.3, synthetic data is an important component of benchmarking efforts. Accordingly, various approaches have been developed aiming to create realistic synthetic spatial transcriptomics data [214, 215]. The resulting synthetic data is often provided alongside ground truth annotations of underlying cell types – it is much rarer to find approaches designed to annotate tissue domains. However, for the benchmarking of domain recognition algorithms, ground truth domain annotations are indispensable, and so methods frequently develop strategies for domain-based spatial transcriptomics simulation.

The following sections give a brief introduction to the approaches used in various method publications for internal benchmarking, as well as the dedicated spatial transcriptomics simulation tool SRTsim [214].

3.1.1 Overview of published simulation approaches with concurrent ground truth domain generation

Many methods for the identification of spatial domains benchmark their performance against other methods based on a simulation approach. The simulation approaches utilised by a subset of methods are briefly summarised in Tab. 3.1. Methods take different approaches to data simulation, some based heavily on real data, while others create fully synthetic datasets (e.g. SpiceMix). Strategies for the definition of spatial domains vary between two main approaches, similarly to the definitions employed in method papers – namely, domains are defined based on their cell type composition or their expression coherence (see Fig. 1.6). The domain definition which is employed in these simulation approaches broadly coincides with the generation of coordinate and count information. Specifically, simulations generating high-resolution datasets define domains based on cell type composition, whereas grid-based, lower-resolution simulated data uses the expression coherence definition. Interestingly, most approaches create layered tissues, based implicitly or explicitly on brain structures. In the list of simulation strategies shown in Tab. 3.1, Vesalius is the only method directly utilising real expression values, measured using Slide-seq, in their synthetic data. All other approaches employ a simulation

step, either using published simulators like scDesign2, scDesign3, or splatter, or an in-house and specially devised method, like one SpiceMix strategy, SC-MEB, and PRECAST [214–217].

The simulations differ widely in their data generation approaches, both in count and coordinate origins as well as in domain definitions. Further, the approaches were devised and implemented to enable the investigation of very different variation scenarios. SC-MEB and PRECAST vary the covariance matrix defining spatial smoothness. Similarly, SpiceMix implements noise types influencing the expression similarity between neighbouring cells, and SpatialPCA varies the neighbour correlation for grid-based data directly using a split-cells approach. GraphPCA investigates a range of technological parameters, ranging from sequencing depth and noise levels to spot and count sparsity. SpatialPCA, Vesalius, and BASS vary the cell type composition heterogeneity of their domains, and BASS additionally changes the proportion and variation magnitude of differentially expressed genes. Interestingly, Vesalius is the only simulation approach which considers different domain layouts.

3.1.2 Simulation with SRTsim

One published simulation method that enables the simultaneous generation of count data, coordinates and the corresponding domain labels is SRTsim [214]. In its *de novo* data generation mode, it takes the desired tissue domain layout as input and allows for the user to input hyperparameter values specifying the distributions from which expression values should be sampled. The approach SRTsim takes to delineating different domains is simple, in that it is based on a user-defined logarithmised fold change of the expression mean of a set of "signal genes". Notably, the definition of domains based on expression mean fold changes is not encountered in other publications to the best of our knowledge.

The SRTsim *de novo* mode is accessible primarily through an R-shiny application[1], providing a GUI. In order to enable serialised, script-based data generation, we adapted the code underlying the application as publicly available on GitHub. We utilised the resulting script to generate data containing a range of different numbers of cells and genes, and varied the signal-to-noise ratio.

Ultimately, we decided not to further pursue the use of this simulation strategy, as for our purposes, the possibility of broader parameter tuning was important. Further, it is not yet possible to evaluate different domain definitions, such as cell type heterogeneity, as an alternative to the inbuilt mean fold change.

As an alternative to the *de novo* approach, SRTsim allows the generation of synthetic spatial transcriptomics data based on reference input data [214]. Using this approach, it is possible to modify certain data parameters, as well as the domain layouts, with respect to the reference. We tested this mode by creating additional samples of the Visium-Maynard dataset and evaluating method performance.

While data from this simulation mode is likely to better reflect real data characteristics in domain composition and gene expression patterns, it also specifically reflects the technology and tissue characteristics of the reference data. This simulation strategy is not suited for comparing method performance across a spectrum of one technological parameter while keeping others at fixed values.

3.2 Construction of the semi-synthetic data generation pipeline

In order to test hypotheses about how various data characteristics affect method performances, we developed a flexible pipeline for generating semi-synthetic spatial transcriptomics data (Fig. 3.1a). Our aim when creating this pipeline was to present a highly tunable approach, able to incorporate and execute perturbations on the level of coordinates, counts and tissue domains. Yet we also wanted to incorporate count data with realistic characteristics, which is hard to achieve in a synthetic approach [218]. One possibility leading to higher realism in the expression data is to simply incorporate counts

[1] It is available at jiaqiangzhu.shinyapps.io/srtsim (accessed on October 12, 2025).

Method	Counts origin	Coords origin	Domain shapes	Domain def.	Variation scenarios
SpiceMix [192]	simulated imaging (metagene-based or scDesign2 from scRNA-seq mouse cortex [216])	simulated (random with minimum distance)	layers	cell type composition	Leakage to neighbouring cells and additive spatially smooth noise
BASS [82]	simulated imaging (using splatter [217])	real (STARmap mouse cortex)	layers	cell type composition	composition heterogeneity, number of genes, proportion differentially expressed and strength of differential expression
Vesalius [140]	real Slide-seq (deconvolved)	simulated random	layers and blobs upon background	cell type composition	shapes and composition heterogeneity
SpatialPCA [136]	simulated sequencing (using splatter [217] from Smartseq2 human DLPFC)	real (random pixel sampling from Visium DLPFC)	layers	cell type composition	composition heterogeneity, non-contiguous domains, added neighbour correlation using split cells
GraphPCA [137]	simulated sequencing (using scDesign3 [215] from Visium sagittal mouse brain)	real (Visium sagittal mouse brain)	complex (sagittal mouse brain)	expression coherence	sequencing depth, noise level, spot sparsity, counts sparsity
SC-MEB [116]	– (generate PCs)	simulated (square lattice) and real (Visium colon)	N/A	expression coherence	covariance matrix
PRECAST [164]	simulated using Potts model, simulated based on PRECAST low-dimensional embeddings extracted from Visium DLPFC, and real from Visium DLPFC (with added pseudocounts)	simulated (square lattice) and real (Visium DLPFC)	simulated (using k-state Potts model) and real (Visium DLPFC)	expression coherence	smoothing parameter and covariance matrix

Table 3.1: **Overview of within-method-publication data simulation approaches.** Characteristics of simulation approaches utilised in some published methods for spatial domain identification. DLPFC, dorsolateral prefrontal cortex

from a publicly available spatial transcriptomics dataset, such as the approach taken by Vesalius (see Tab. 3.1). However, as discussed in the previous section, this data carries with it the exact inherent technological biases we are aiming to study. At the time of the development of our pipeline, to the best of our knowledge, there was no public dataset from a spatial transcriptomics technology with full-transcriptome profiling at true single-cell resolution. Finally, we landed on utilising a well-annotated single-nucleus RNA-seq dataset of the mouse brain to serve as the origin of transcriptional identities for our semi-synthetic data generation [200].

Briefly, we first create archetypal domain shapes and overlay them on an artificial tissue layout of randomly generated cell locations. After transferring the domain identities to cells as label assignments, we proceed to ascribe mixtures of cell types from the mouse brain dataset to define each domain. Counts from those cell types are chosen and assigned to cells within the corresponding domains. Within this pipeline, we are able to vary relevant parameters at all individual steps. Each of these steps is described in more detail in the following sections.

3.2.1 Creating the tissue layout

To generate tissues containing different shapes and arrangements of domains, we utilise the datasets module of the scikit-learn Python package as a basic first step [219]. This module is designed to create locations and cluster assignments for a set of points distributed non-uniformly in a 2D space (shown in Fig. 3.1b). We therefore combine the thus generated points with a new set of cell coordinates, drawn from a 2D uniform distribution (overlaid in Fig. 3.1c). Domain labels are assigned to all cell locations based on label transfer from the nearest point in the sklearn-generated dataset (resulting tissue shown in Fig. 3.1d). Finally, in order to create contiguous domains and remove any outlier labels, a next-neighbour-based smoothing algorithm is applied to the newly created tissue, aligning each spot to the majority-voted neighbourhood label (final synthetic tissue shown in Fig. 3.1e). This strategy has the advantage of being able to create a diverse set of domain shapes and configurations, as shown in Fig. 3.1f, while avoiding the need to define exact domain borders. Additionally, we create a tissue type consisting of parallel stripes to represent a more balanced domain layout, and an archetype that is present in many of the real datasets we included.

3.2.2 Choosing cell types and assigning counts

The single-nucleus mouse brain dataset published by Langlieb *et al.*[2] contains detailed cell type annotations [200]. After subsetting to only cell types containing sufficient numbers of cells, we specifically chose cell types of high pairwise similarity for inclusion in our analysis[3]. Cell type similarities were quantified by the inverse of their distances in a dendrogram published within the scope of the original study [200].

Using this strategy, we settled on a set of five cell types for primary use within the simulation pipeline. Results shown in this thesis, if not otherwise specified, are attained using primarily these cell types to define the domains. After assigning individual cell types or mixtures thereof to define each domain, cells are assigned to each coordinate of the previously generated tissue layout at random.

3.2.3 Implementing variation on different levels

In the course of the steps described in the previous sections, variation can be introduced at different stages. The following paragraphs briefly describe the possible variations.

[2]The dataset is available online at braincelldata.org.

[3]We decided on this approach after preliminary simulations showed excellent domain recognition across all methods applied to data created from transcriptionally dissimilar cell types. We aimed to give the methods more of a challenge.

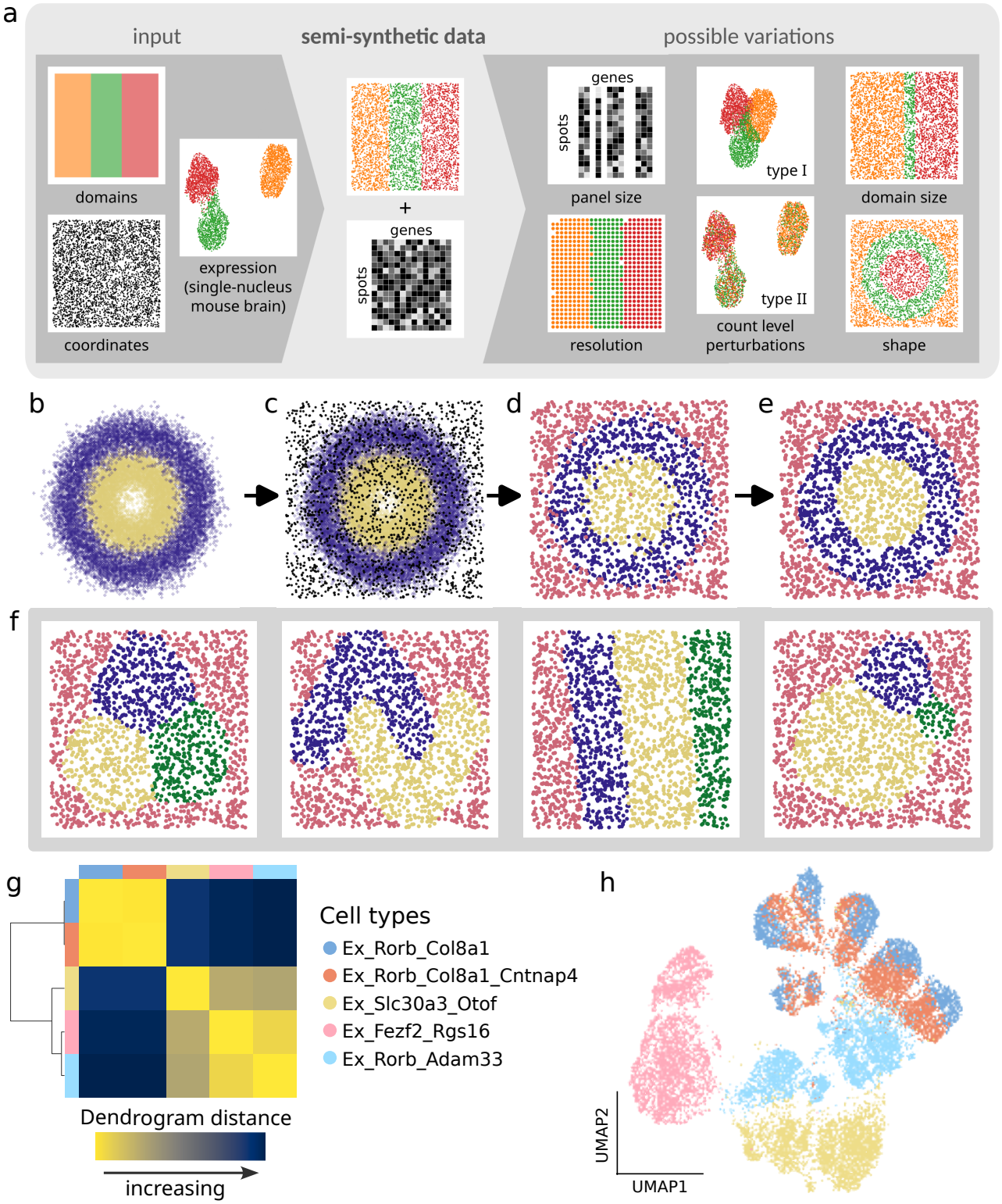

Figure 3.1: **Pipeline developed for semi-synthetic data generation.** a, Overview of the pipeline, along with possible variations it allows us to introduce in the data. b, Point clusters generated using scikit-learn. c, Clusters from scikit-learn, overlaid with randomly generated locations for cells. d, Domains after label transfer from scikit-learn clusters onto cell locations. e, Smoothed domain labels after removal of outlier labels. f, Examples of other shapes the pipeline is able to generate. g, Pairwise distances of the cell types selected by default, shown as a heatmap calculated based on dendrogram data published by Langlieb *et al.* [200]. Cell types are indicated by colours. h, UMAP embedding of cells from all selected cell types

Tissue layout Domain shapes, and the tissue layout in terms of cell density and arrangement, are determined at the start of the pipeline. Besides the default shapes shown in Figs. 3.1e,f, different domain shapes and configurations can be created through scikit-learn or by manual definitions. In addition to the random cell locations simulated by default, grid-based coordinates are also implemented.

Technological parameters: Resolution We are also able to vary the resolution of the resulting tissue after completing the simulation, by overlaying a grid onto the tissue and binning the expression of multiple cells. Specifically, in order to create synthetic data of different resolutions while keeping other parameters fixed, we create and then perturb one basic dataset of single-cell resolution. The original cell locations are chosen randomly in a square tissue of size 100×100 points. Subsequently, for the different resolutions, grids consisting of square tiles are overlaid on this tissue, and cell coordinates are rounded to the nearest tile centroid. From all cells rounded to the same tile, expression counts are aggregated by mean to form the counts of the newly created spot. The range of resolution explored in the synthetic data is large, with tile (spot) side lengths from 0.5 up to 10 points. For reference, at a side length of 5, on average, 9 cells are contained in one spot, comparable to the resolution achieved by Visium. Consequently, a side length of 10 corresponds approximately to the original ST resolution, with 33 cells being aggregated into one spot. This approach creates a more realistic low-resolution dataset than would be possible using an *a priori* grid-based layout, as it mimics the characteristic aggregation of counts across cells of Visium-like technologies.

Technological parameters: Number of genes The gene panel size can also be varied *post facto*, by reducing the number of genes whose expression is included in the final data. Similarly to the resolution strategy, to evaluate the effect of changing gene panel size, we modify a base dataset. Specifically, we decrease the number of included genes to exponentially spaced proportions of the original number (that is, the total gene panel size is reduced to 50%, 20%, 5%, etc., of the original size). From 21899 profiled genes in the original single-nucleus RNA-seq count data, the proportions investigated result in gene numbers close to those of our included real datasets. Once the number of genes to include is calculated, we downsample the count matrix using random sampling. By default, we use a random downsampling strategy to reduce the number of genes, but other strategies are easily implemented.

Technological parameters: Sparsity A different downsampling procedure may be applied to the generated data to increase the count matrix sparsity. We vary the level of sparsity within a range of 0.85 to 0.99. The lower end of this range is set by the sparsity inherent in the single-nucleus RNA-seq count data, while the upper end is comparable to Slide-seq data. Concretely, high dropout levels are simulated through randomly setting counts to zero until the desired sparsity level is reached. Again, various models for high count sparsity could be implemented.

Expression similarity and heterogeneity Finally, we are able to introduce variation on the count level. We investigate two main perturbation types, each applied to the entire tissue and subsequently adapted to apply to all pairs of domains. In the first type of perturbation, starting from domains defined by different cell types, the expression values in the affected domains are modified to increase their similarity. This is achieved in the tissue-wide perturbation by introducing an additional "noise" cell type, to which all domains are made to converge. Concretely, let the original expression vector of spot s in domain d be $\vec{c}_s^d$, defining the expression levels of all genes. Additional gene expression vectors $\vec{n}_s$ for each spot are then generated from the noise cell type n, and gradually replace the original expression by a convex combination as

$$\tilde{\vec{c}}_s^d = (1 - \lambda)\vec{c}_s^d + \lambda\vec{n}_s \tag{3.1}$$

Upon gradual variation of the mixing parameter λ, this approach modifies the count values of the entire tissue, with the expression of all domains increasing in similarity to the "noise" type. In the pairwise domain perturbation, counts are instead created through the convex combination of cell types from both involved domains.

The second perturbation type we investigate approaches the idea of domain similarity through cell type composition. Namely, we introduce cells from either a noise cell type (in whole-tissue perturbation) or from the paired domain (in pairwise domain perturbation), into the tissue in question. In this perturbation, counts are not added to the preexisting cells - instead, the entire expression profile of a certain proportion of cells is replaced. This introduces a degree of expression heterogeneity on the cell type level.

3.3 Investigating technology characteristics

First, we utilise our pipeline to create data with varying technological characteristics, aiming to disentangle their effects on method performances. Specifically, we evaluate the effect of changing resolution, the number of profiled genes, and the count matrix sparsity. In order to avoid any biases arising from specific domain configurations, the following analyses are averaged across the different shapes shown in Fig. 3.1e,f. All experiments are carried out using the same basic assignments of one cell type per domain.

3.3.1 Effect of changing resolution

As described in Sec. 3.2.3, we generate semi-synthetic data samples with a variety of resolutions based on a single-cell resolution base dataset (Fig. 3.2a). The resolution is gradually decreased by increasing the side lengths of the overlaid grid (Fig. 3.2b).

We first examine the correlation of ARI and PAS across all methods. In the real data, we had observed a strong anticorrelation of ARI and PAS in high-resolution data, whereas on lower resolutions, this anticorrelation was weakened (see Fig. 2.6b). This trend is corroborated by the semi-synthetic results, as shown in Fig. 3.2c. Specifically, we find a strong anticorrelation of ARI and PAS (Spearman correlation around -0.5) at small spot side lengths of 0.5 to 3, whereas there is only a weak anticorrelation (Spearman correlation around -0.2) at spot side lengths 4 to 6, which correspond loosely to Visium spot size. Even this level of correlation disappears for even larger spot side lengths.

Concerning the effect of resolution changes on the individual methods, large differences become apparent (Fig. 3.2d). MNMST and BANKSY decline drastically in performance with the aggregation into larger spot sizes. At resolutions only slightly smaller than the Visium equivalent, they end up with domain assignments equivalent to random, indicated by ARIs around 0. MNMST specifically exhibits a sharp performance drop from side lengths 4 to 6, and does not even produce any output for side lengths 9 and 10. The only method showing a similarly abrupt performance drop is CellCharter, with a rapid decrease in ARI from spot side lengths 1 to 2. However, the performance of CellCharter then stabilises at a mid-range ARI value. BANKSY, on the other hand, does not drop in performance from one resolution to the next. Instead, it declines in performance almost monotonically from very small aggregations, and ends up with negligible ARIs from side length 7. Many other methods, from SCAN-IT and SpaDo to GraphST, also show similar gradual performance deterioration from small spot sizes to BANKSY. Unlike BANKSY and MNMST, but like CellCharter, the decrease in ARI of these methods flattens out at about 0.5.

On the other hand, methods like TACCO, PAST, and STAGATE decline only very gradually upon the first aggregation of cells into spots. Only for under Visium-like resolution, that is, side lengths greater than 5, their performance starts to decline more noticeably. Select methods, notably SpiceMix, BASS, and DeepST, uphold a very strong performance with ARIs close to 1 throughout the

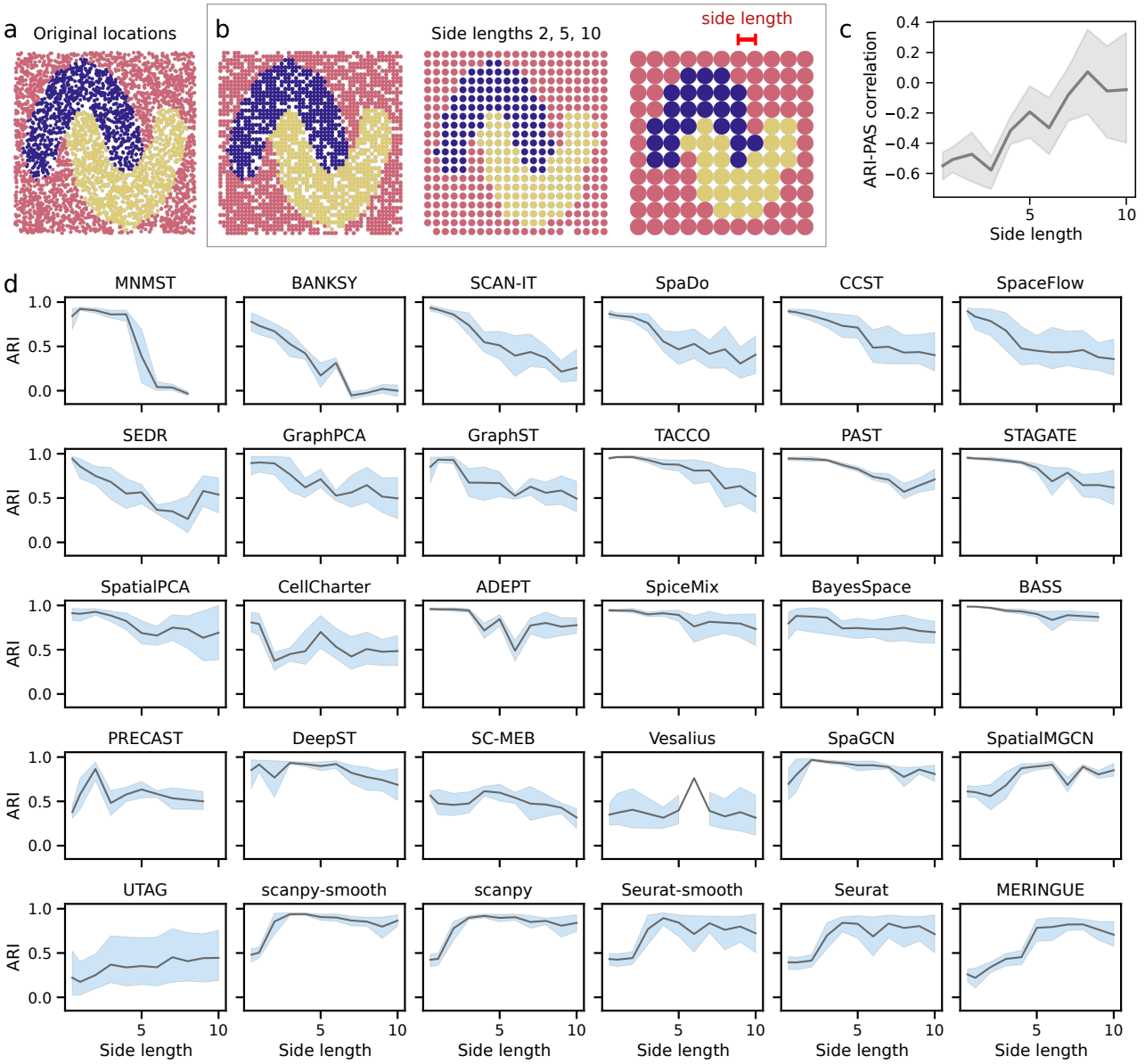

Figure 3.2: **Performance dependence on resolution**. a, The base data, which is subsequently modified to investigate the resolution dependence, is generated with single-cell resolution. One example domain layout is shown. b, Examples of binned data, generated to test resolution dependence, at spot side lengths 2, 5, and 10. c, Spearman correlation of ARI and PAS as a function of the spot side length. The correlation is aggregated over all methods and domain layouts by mean. d, Performance dependence on resolution as parametrised by the spot side length, for all methods. The variance per datapoint is over the different domain layouts. Methods are sorted by the difference in their performance between "big" spots (side length > 6) and "small" spots (side length < 3).

entire resolution range. An interesting behaviour is exhibited by the baseline methods scanpy, Seurat and the respective smoothed implementations, as well as MERINGUE, SpaGCN, and SpatialMGCN. These methods manage to actually improve their performances, in some cases to close to perfect ARIs, upon the first aggregation of cells into spots. Subsequently, the performance stays high as the sample resolution decreases.

The effect of changing resolution may be decomposed into an interplay of two main factors. First, binning multiple cells into one spot leads to diminishing domain sizes, as measured both by the absolute number of spots in each domain and in terms of domain diameters or widths in spot units. For example, consider a domain consisting of 400 spots, arranged in a 20-spot-wide layer, at a given spot side length of a. At a smaller resolution, defined by spot side lengths of $2a$, this domain will contain only about 100 spots, in a layer that is only 10 spots wide. Thus, the aggregation of increasing numbers of cells in each spot causes a decrease in domain sizes. Additionally, this aggregation can itself be viewed as a kind of spatial smoothing operation, converting transcriptional heterogeneity between neighbouring cells to relative spatial homogeneity. Both of these factors will be considered in more detail in later sections.

3.3.2 Effect of changing the number of genes

As a second technological parameter, we use our semi-synthetic data to investigate the effect of changing the number of genes on method performance. Interestingly, several methods do not produce any domain output on samples with small gene numbers (Fig. 3.3). SpatialMGCN, as the most extreme example, only successfully outputs domain labels on samples with over 10'000 genes. SCAN-IT completes runs on samples down to a gene panel size of about 1000 genes, SpatialPCA and SpiceMix down to 200 genes, and some additional methods only fail completely on the smallest gene panel size. These failures do not necessarily correlate with methods failing on real data with small gene numbers, such as the osmFISH and MERFISH datasets.

For the methods which do successfully run over the entire range of gene numbers, the difference between methods does not amount to a total trend reversal, as in the case of the resolution dependence. Rather, while no methods are positively affected by diminishing gene numbers, they differ in the extent of the resulting performance decrease.

UTAG and Vesalius, along with MERINGUE, are only minimally affected at an overall low performance level. Methods like SpatialPCA and GraphPCA also do not exhibit a strong change due to the number of genes, though, as mentioned above, they do not produce any domain labels for the smallest samples. CCST, SpaDo, and BASS, among others, start out with excellent performances of ARIs close to 1 and then decrease monotonically towards smaller panel sizes. These methods end up at mid-range ARIs around 0.5 at the smallest gene numbers. Other methods showing very good performances on the full transcriptome before downsampling, like GraphST, STAGATE, and MNMST, decline gradually in performance to about zero ARI around gene panel sizes of 100. Besides the methods exhibiting a monotonic performance decrease, for some methods, including TACCO, CellCharter, and BANKSY, performance drops sharply at "cutoff" gene panel sizes. While these methods perform highly competitively on samples with large numbers of genes, the domains produced at smaller gene numbers are equivalent to random label assignments. Finally, an interesting behaviour is exhibited by a small group of methods encompassing SpaGCN, SpaceFlow, and notably the spatial baseline scanpy-smooth. These methods do not have their peak performances on samples with the full transcriptome before downsampling, but instead show performance maxima at 500–1000 genes.

The full performance overview of all methods is shown in Fig. 3.3. All in all, while most methods decline in performance on samples containing decreasing numbers of genes, the extent and abruptness of this decline vary widely.

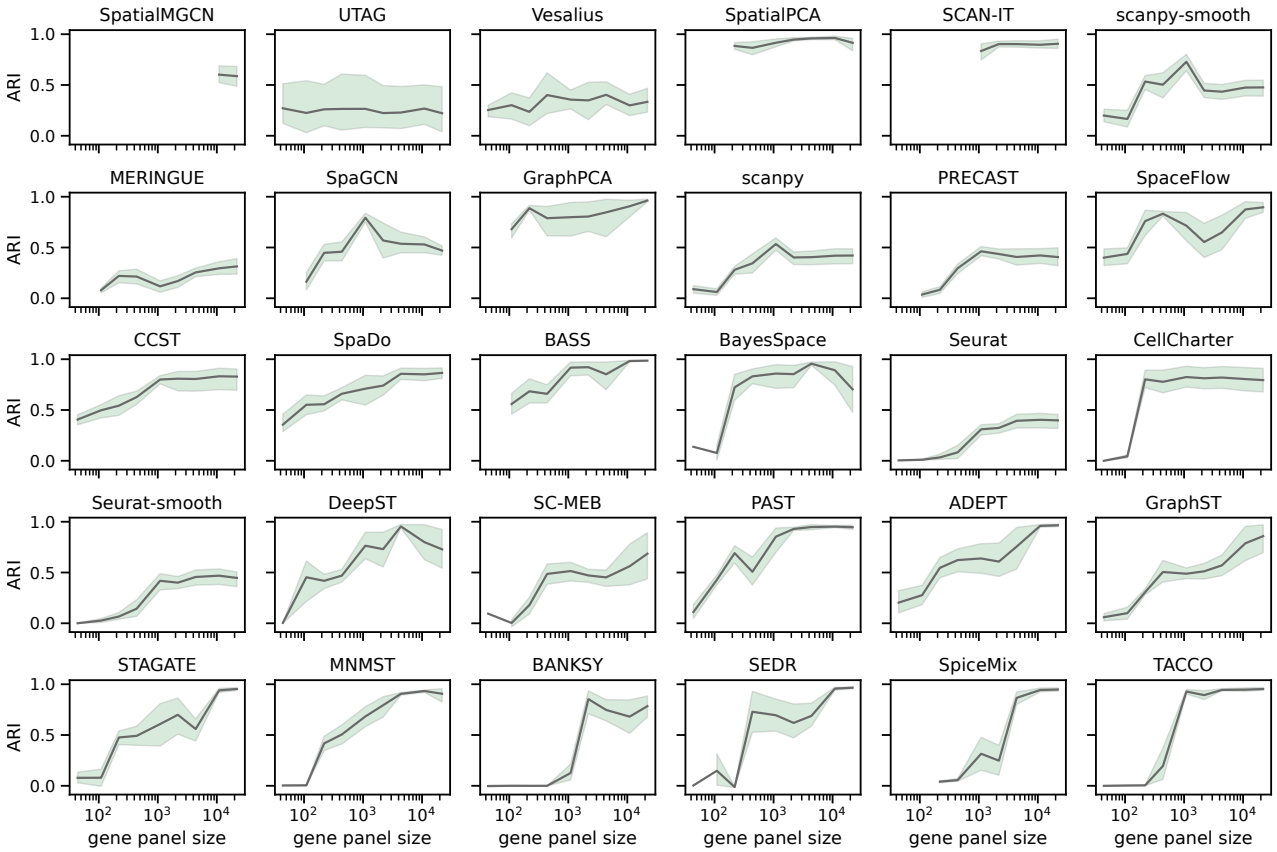

Figure 3.3: **Performance dependence on number of profiled genes.** Method performances in terms of ARI are shown as a function of the number of genes, going from smaller gene panels to full transcriptome profiling. The x axis is on a log scale. Methods are sorted by the difference in their performance between "many" profiled genes (number of genes > 5000) and "few" profiled genes (number of genes < 500). Variance per datapoint is over the shapes.

3.3.3 Effect of changing count matrix sparsity

The third technology characteristic we investigate using our semi-synthetic data is capture efficiency, using the degree of count matrix sparsity as a proxy. The sparsity is varied between 85%, corresponding roughly to the original sparsity of the single-nucleus RNA-seq data, and 99%.

At this highest level of sparsity, only 1% of all counts in the matrix are nonzero. While SpatialPCA and SpiceMix do not produce any output on these samples, a considerable number of methods still reach nonzero ARIs (around 16 of 30, see Fig. 3.4). Notably, GraphPCA, MNMST, and BASS result in ARIs greater than 0.5 at a sparsity of 99%. Besides these methods, SpatialPCA, SpaceFlow and SpaDo are among the least and last affected by the rising proportion of zeros. Notably, MNMST, SpaceFlow, and SpaDo only start to decline in performance at over 95% sparsity. Among the methods that are more strongly affected by sparsity levels are BayesSpace, TACCO, SpiceMix, ADEPT, and PAST. While they perform highly competitively at low sparsity, their ARIs diminish starting at about a sparsity of 90%, and at the highest sparsity levels, these methods end up producing domains equivalent to the results of random label allocation. Similarly, the performance of the scanpy-based baselines, along with SpatialMGCN, SpaGCN, PRECAST, and MERINGUE, decreases to negligible ARIs for high sparsity. Interestingly, Seurat and the corresponding spatial baseline Seurat-smooth are only barely affected by even very high sparsity. Methods like STAGATE and PAST exhibit a plateau in method performance at an ARI around 0.5, which will be discussed in more detail in a later section.

Thus, Fig. 3.4 shows that there are large performance differences between methods concerning

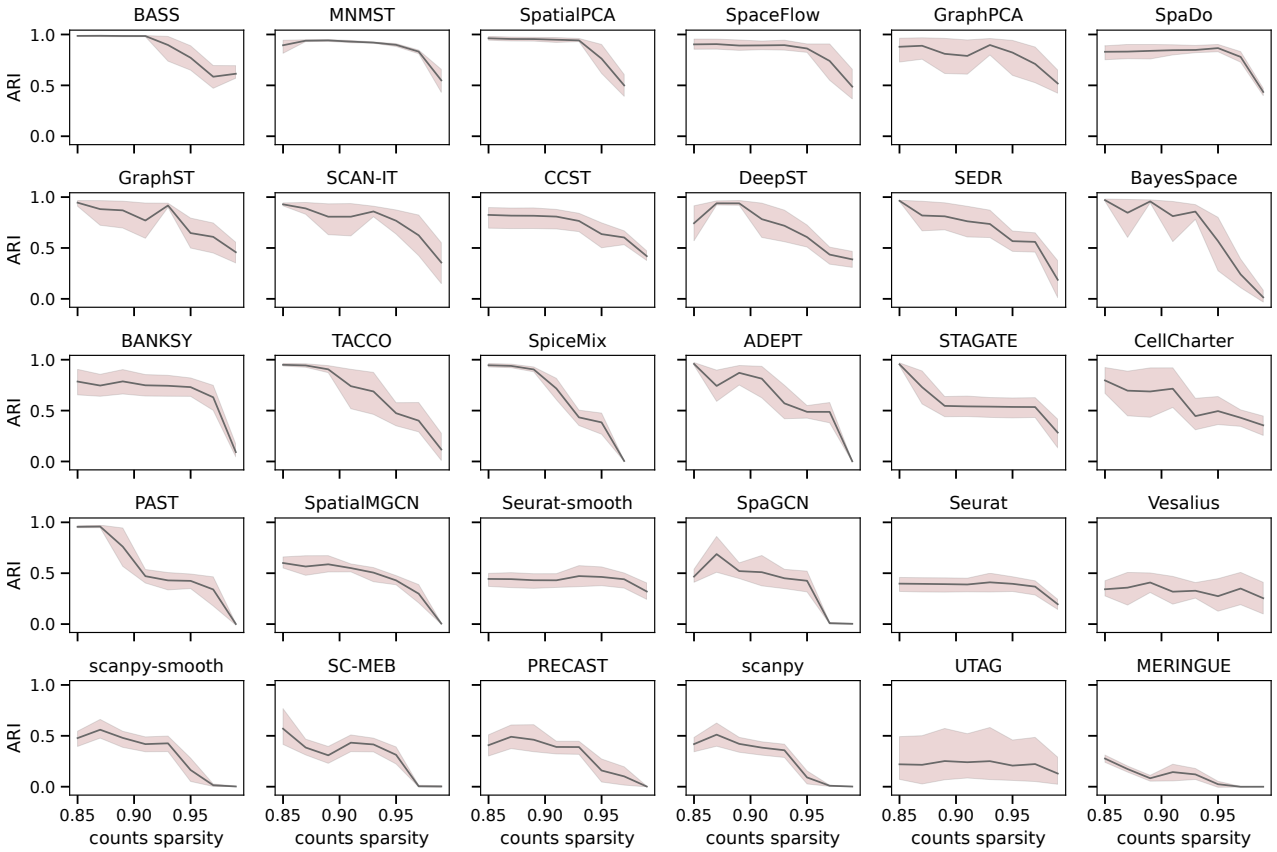

Figure 3.4: **Performance dependence on count matrix sparsity.** Method performances in terms of ARI are shown as a function of the zero percentage of the expression matrix. Methods are sorted according to their mean performance. Variance per datapoint is over the shapes.

the effect of an increasingly sparse expression matrix. Methods like BASS, MNMST, SpaceFlow, and SpaDo stand out for their robustness with respect to high dropout levels.

3.4 Impact of transcriptional similarity and heterogeneity

So far, we have investigated the impact of data characteristics determined by the sequencing or imaging technology. In this section, we analyse instead the influence of two cell type level perturbations (type I and type II) on method performance. As described in Sec. 3.4, in perturbation type I, the domain-defining cell types are gradually modified to increase their similarity. This perturbation can be interpreted as a spectrum of cell types with differing similarities, or as progressive technical contamination through ambient RNA or smearing during an experiment. The other count-level variation we investigate, termed perturbation type II, consists of the addition of cells of a different type throughout the affected domains. This corresponds biologically to the infiltrating behaviour of immune cells, or to cell migratory behaviour during development. Generally, it simulates transcriptional heterogeneity of domains as caused by differences in the cell type composition. A perturbation level of 100% corresponds, in both perturbation types, to zero clustering signal being available to the methods.

The following sections describe the application of these count-level domain similarity modifications on two different levels. We primarily investigate tissue-level perturbation, where all domains are modified to increase overall similarity within the entire tissue. As a further analysis step, we examine the effect of pairwise domain similarity by perturbing two domains at a time, keeping all other domains fixed.

3.4.1 Whole-tissue perturbations

Just as we did for the previous technological parameter investigations, we create a basic dataset with fixed assignments of cell types to domains, containing samples from all shape configurations. Aiming to perturb the entire tissue, we choose an additional "noise" cell type from the group of cell types selected for investigation previously (see Fig. 3.1g,h). Counts originating from cells of this type are added as a uniformly distributed background signal to all cells (type I) or replace a proportion of all cells throughout the tissue (type II). We create samples with "noise" count or cell proportions ranging from 0 to 100%.

First, we investigate the performance behaviour of all methods under increasing proportions of the perturbation type I. Most methods exhibit a gradual decline in ARI scores at increasing type I levels (see Fig. 3.5a). In particular, BASS, ADEPT, and TACCO stand out with negligibly declining ARIs up to perturbation levels of 60%. Other methods decrease in accuracy from the lower proportions, like SpatialPCA, SEDR, and PAST, or from the first added perturbation, like SpaceFlow, STAGATE, and GraphST. As a useful and successful sanity check, methods deteriorate in accuracy down to random label assignments for the highest proportion levels. Still, most methods which start out with ARIs close to 1 at zero perturbation manage to attain ARIs around 0.5 at very high perturbation levels of up to 80%. BASS, ADEPT, and SpatialPCA, the latter of which does not produce any output on 100% perturbed samples, still perform very competitively at 90% type I perturbation. Only select methods beside the spatial and non-spatial baselines, like MNMST, CellCharter, and DeepST, assign labels completely arbitrarily, indicated by ARIs around 0, when there is still some amount of signal to be found within the gene expression. Generally, for the majority of methods, the ARI declines in a concave function of the perturbation proportion.

Additionally, we evaluate the methods using the unsupervised PAS metric. In the real datasets investigated in the previous chapter, as well as in our systematic investigation of the effect of resolution changes using semi-synthetic data, we had found a strong anticorrelation of ARI and PAS on high-resolution data (see Figs. 2.6b and 3.2c). Since we are generating semi-synthetic data with single-cell resolution for this investigation of count-level perturbations, we expect to see this anticorrelated behaviour here. Indeed, increasing PAS levels appear to mirror the ARI decline of many methods, like CellCharter, SpatialMGCN and MERINGUE (see Fig. 3.5a). However, interestingly, a large group of methods, encompassing e.g. SpaDo, CCST, and BANKSY, exhibit PAS values close to zero throughout the range of perturbation levels. In fact, we are able to distinguish three archetypal "modes of method failure" based on the interplay of ARI and PAS, illustrated in Fig. 3.5b. In mode A, exhibited by methods like SpaGCN, SC-MEB, PRECAST, and SpatialMGCN, a performance decrease measured by ARI goes along with an increase in PAS. Methods failing in mode A tend to create visually noisy domains, blurring region boundaries. The second failure archetype, mode B, encompasses methods for which PAS stays low even at high levels of added noise, but which exhibit strictly monotonically decreasing ARI scores. In this failure mode, methods like SpatialPCA, SCAN-IT, SpaDo, and CCST tend to mislabel spatially contiguous groups of spots, leading to a fragmentation of the tissue. Other methods with consistently low PAS, like BASS, STAGATE, and GraphST, exhibit a sharp performance drop followed by a plateau, characterised by ARI scores around 0.5–0.6, from which they again drop off sharply to zero ARI. This failure style indicates a bit flip mode of label misassignment, where an entire domain is mislabelled above a cutoff perturbation level, usually assimilating to a neighbouring domain. Many methods exhibit combinations of these failure modes, with particularly methods like DeepST, SpaceFlow, and MNMST showing combinations of modes A or B with mode C.

Notably, the plateau observed in methods exhibiting mode C failure lies around the maximal ARI values attained by the baselines, among other methods. These less well-performing methods reach a maximal ARI of around 0.5 under zero-perturbation conditions, and in many cases do not decline further in performance until perturbation levels around 50% (see Fig. 3.5a). To elucidate

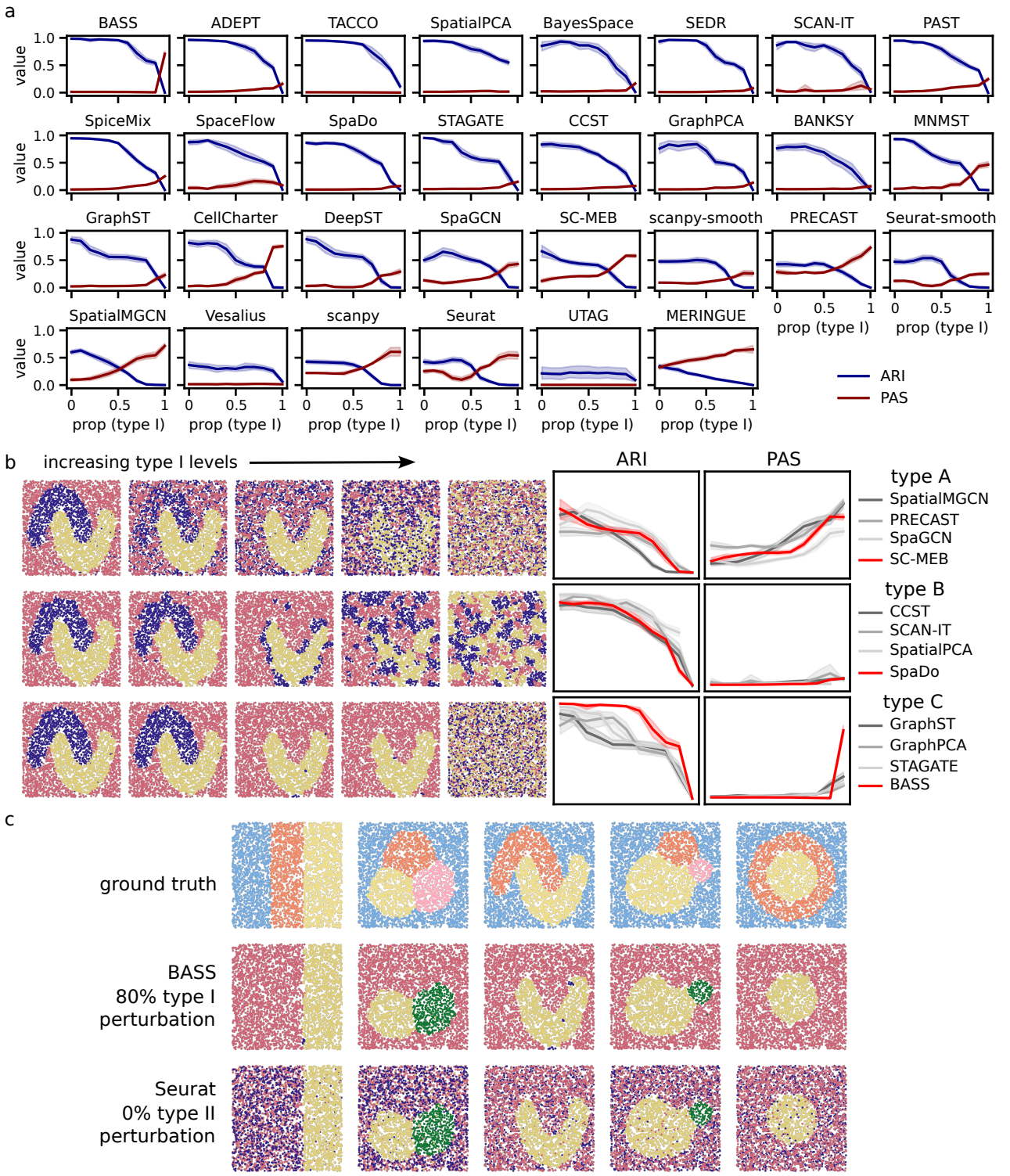

Figure 3.5: **Effect of type I perturbation (increasing transcriptional similarity).** a, ARI and PAS dependence on the level of underlying "ambient RNA" noise (generated by perturbation type I). This perturbation is generated through the expression of each cell representing a convex combination of original counts with those originating from a noise cell type. Methods are sorted by mean performance, and the variance per datapoint is over the shapes and cell assignments. b, Example performances of methods exhibiting three different archetypal failure modes. Method output is shown for the methods highlighted in red, at different type I perturbation levels. Perturbation levels of the shown examples are chosen to represent the progression per method. c, Ground truth label annotations for all shapes, along with example results of BASS at 80% type I perturbation and Seurat at zero perturbation. The light blue and coral domains in the ground truth, defined by cell types Ex_Rorb_Col8a1 and Ex_Rorb_Col8a1_Cntnap4 (see Fig. 3.1g), are indistinguishable in the two example outputs.

this phenomenon, we examine the examples of BASS, as a consistently well-performing method which exhibits a brief plateau in ARI at high perturbation proportions, and the baseline method Seurat (Fig. 3.5c). Considering the example outputs of Seurat on the unperturbed samples and BASS at 80% type I perturbation, it becomes apparent that the same domains evade detection in both methods. Specifically, the methods which are indistinguishable to these methods are defined by the same cell types across all the shapes we investigate (indicated in light blue and orange in the top row of Fig. 3.5c). These are the highly similar cell types Ex_Rorb_Col8a1 and Ex_Rorb_Col8a1_Cntnap4, clustering together both in the dendrogram and the UMAP representation shown in Fig. 3.1g,h.

Next, we evaluate the effect of the type II perturbation on method performances. Interestingly, when applied to tissues with proportions of cellular heterogeneity of just 20%, only a small group of methods are able to hold their performance level (Fig. 3.6a). The majority of methods decline in accuracy immediately upon the addition of any level of type II perturbation. Additionally, we do not observe the tendency toward the "bit flip" failure mode C that we observed on perturbation type I. Rather, more methods exhibit mode A style failures, increasing in PAS gradually as they decrease in ARI. Generally, in the majority of methods, PAS values are significantly higher on samples affected by type II, rather than by type I perturbation (Fig. 3.6b). Specifically, the difference is significant for all methods with non-negligible PAS (thus excluding UTAG, TACCO, BASS, SpaDo, and Vesalius) except SpaceFlow and MERINGUE. As expected, the difference is highly significant for the non-spatial baseline methods, which delineate domains purely based on transcriptional identity. On the other hand, methods like ADEPT, BayesSpace, and PAST also result in significantly higher PAS values on type II- than on type I-perturbed samples. Accordingly, while these methods perform quite competitively under type I perturbation, they are strongly affected by even low levels of type II perturbation (compare Figs. 3.5a and 3.6a). In some extreme cases like GraphST and DeepST, the performance measured by ARI follows a convex function of perturbation proportion.

A few methods stand out for their performance on samples generated with perturbation type II, namely BASS, TACCO, SpaDo, SpaceFlow, SCAN-IT, and SpatialPCA. These methods are able to still attain ARIs around 0.5 for high perturbation levels of 80–90%. Notably, BASS and SpaDo do not exhibit a visible performance decline up to 70–80% of type II perturbation levels. Further, interestingly, SpaceFlow and SCAN-IT outperform most other methods on this perturbation type (as evident by the method sorting in Fig. 3.6a), whereas on type I perturbation, their performance was average within the set of all methods.

Overall, the investigations of tissue-level perturbations of types I and II reveal the different effects of these perturbations. Most methods are immediately affected negatively in their domain identification performance by infiltrating cells (type II perturbation), while transcriptionally highly similar domains can be distinguished. When applied to domains defined by progressively more similar cell types, many methods exhibit a "bit flip" failure style, unable to distinguish pairs of domains after a similarity cutoff. Few methods excel on both perturbations, namely BASS, TACCO, SpatialPCA, and SCAN-IT.

3.4.2 Pairwise domain similarity

Following up on the phenomenon of "bit-flip" failure, and the domains indistinguishable to multiple methods being defined by the same cell types (Fig. 3.5c), we investigate the effect of pairwise domain similarity in more detail. Concretely, as described in Sec. 3.2.3, we gradually increase the pairwise similarity of all domain pairs separately. We consider two types of domain similarity, defined analogously to the previously discussed perturbation types I and II. The type I analogue is defined by what we call "expression mixing", that is, counts in two domains are gradually mixed through convex combination. An alternative formulation of domain similarity, analogous to the type II perturbation, can be defined through "cell shuffling", whereby cells from both affected domains are gradually intermixed, leaving individual cell expression intact.

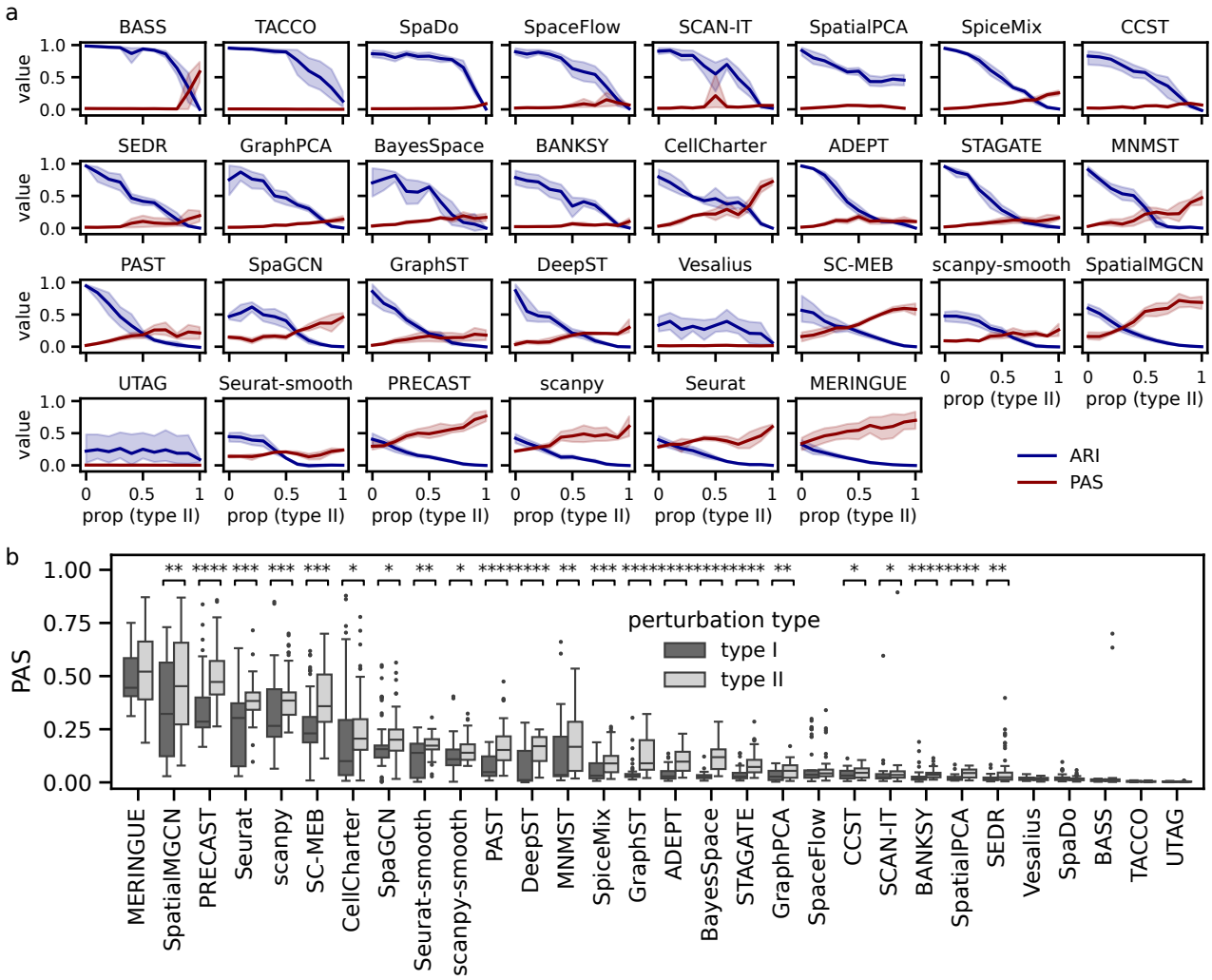

Figure 3.6: **Effect of type II perturbation (increasing cellular heterogeneity).** a, ARI and PAS dependence on the level of "infiltrating cells" noise (generated by perturbation type II). Noise proportion in this case corresponds directly to the proportion of noise cells inserted in the tissue. Methods are sorted by mean performance, and the variance per datapoint is over the shapes and cell assignments. b, PAS values on noise types I and II, aggregated over noise levels, excluding zero noise and 100% noise. Methods are sorted according to the mean PAS across perturbation types. Significance values for PAS values on type II being higher than on type I are calculated using a one-sided Mann-Whitney U test.

Domain-wise evaluation strategy and development of a pairwise confusion metric

As we are now not investigating tissue-wise perturbations anymore, we are also primarily interested in domain-level effects. In order to evaluate method performance on this more granular level, we need to consider a metric that is suited to evaluating the accuracy of individual clusters, rather than an entire clustering result. For this purpose, we utilise the harmonic mean of precision and recall, commonly called the F_1 or simply F score, defined as

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{3.2}$$

We utilise the scikit-learn implementation of the F_1 score, which treats multi-label input as a collection of binary problems [219]. Thus, TP (true positives), FP (false positives) and FN (false negatives) are defined on the level of individual domains, and the metric does not distinguish between erroneous assignments to different labels.

However, besides the evaluation of domain-level accuracy, we are specifically also interested in disambiguating these misassignments, as these will indicate to which domain methods erroneously assign spots, fragments or entire other domains. In other words, we are aiming to evaluate which domains cannot be distinguished (in the following, this is sometimes also referred to as indicating which domains "are confused"). We therefore define a "confusion" metric between domains a and b as follows:

$$\text{confusion} = \frac{r_a^b + r_b^a}{r_a^a + r_b^b}, \quad \text{where } r_i^j = \frac{N_i^j}{\sum_j N_i^j}, \tag{3.3}$$

where N_i^j is the number of cells in ground truth domain i assigned label j. The quantity r_i^j corresponds to the recall when $i = j$. The confusion of a given pair of domains is thus 0 when no cells are mislabelled, and values close to 1 signify "total confusion" or no distinction between domains. In special cases, namely when more spots are mislabelled than labelled correctly, the confusion according to this definition can exceed 1. If there were only two domains involved in the system, this configuration would simply result in the label correspondence to the ground truth flipping to maintain good agreement. This is a consequence of the maximum weight matching algorithm used to calculate label correspondences, as described in Sec. 2.2.3. Confusion values over 1 are thus only possible because our tissue layouts always consist of at least three domains, so there always exists at least one further domain c to which cells can be erraneously assigned. The best overall label correspondence can thus result in the above-described configuration of more mislabelled than correctly labelled spots, namely in the edge case of very small domains. To avoid these outlier confusion values, we cap the confusion metric at a "total confusion" of 1.

Results of pairwise perturbations

Looking first at the cell shuffling perturbation, and stratifying the F_1 results by domain involvement in the perturbation, we can distinguish three method behaviours (shown in Fig. 3.7a,b). Tab. 3.2 details the methods assigned to each behavioural group. Methods in group 1 start out able to distinguish all domains, with only the F_1 scores of perturbed domains subsequently diminishing. The baseline-like group 2 starts out at mean F_1 values around 0.8, confusing the two left-most domains illustrated in Fig. 3.7b. These domains, along with the corresponding ones confused in the other shapes, are defined by the abovementioned highly similar cell types Ex_Rorb_Col8a1 and Ex_Rorb_Col8a1_Cntnap4 (see Fig. 3.1g,h). Even still, similarly to group 1, with increasing proportions of cell shuffling, methods from group II decline in their detection performance of the affected domains, whereas the unperturbed domains continue at similar detection levels.

The last, and most interesting, group of methods is group 3, encompassing SpatialPCA, SEDR, BANKSY, GraphST, and Vesalius. These methods start out at similar average F_1 scores as the

Group 1:	BASS, STAGATE, TACCO, ADEPT, PAST, SpiceMix, MNMST, DeepST, CCST, BayesSpace, SpaDo, SpaceFlow, CellCharter
Group 2:	GraphPCA, SpaGCN, SCAN-IT, SC-MEB, Seurat-smooth, scanpy-smooth, scanpy, Seurat, PRECAST, SpatialMGCN, UTAG, MERINGUE
Group 3:	SpatialPCA, SEDR, BANKSY, GraphST, Vesalius

Table 3.2: **Method groupings based on performance behaviour upon cell shuffling.** The groups exhibit distinct behaviours in F_1 score, as shown in Fig. 3.7a,b.

baseline methods, if slightly lower. However, with increasing levels of perturbation, while the F_1 score of the perturbed domains declines as expected, the detection of domains which are not perturbed increases substantially. Thus, the ability of these methods to recognise domains is affected directly by cell-level heterogeneity, with added heterogeneity in one domain aiding in its distinction. In the example of SEDR shown in Fig. 3.7b, the method is able to distinguish the two middle domains only when 25% of cells are shuffled between the affected domains, at which point it proceeds to confuse domains the second and fourth stripes.

For a different angle of evaluation, we consider instead the confusion metric, stratified by whether the perturbed domains are defined by the *a priori* highly transcriptionally similar cell types (Fig. 3.7c). As indicated above, we see the baselines, along with several other methods like SpaGCN, DeepST, and SpatialPCA, confusing the transcriptionally similar domains independently of any additional perturbation. A newly visible, striking phenomenon is the minimum confusion describing a smooth monotonic function of perturbation, exhibited in methods such as DeepST, STAGATE, SpiceMix, and MERINGUE. This minimum curve coincides broadly with the confusion of not highly similar domains shown by baselines. Additionally, while for the baselines, PRECAST, and SpatialMGCN, among others, the smooth increase in minimum confusion is close to linear, in some cases, such as SpaceFlow, CCST, and GraphPCA, the confusion remains close to 0 until about 25% and subsequently increases steeply. Only BASS, UTAG, TACCO, and Vesalius do not exhibit this smooth minimum behaviour at all. UTAG, TACCO, and Vesalius in fact still display some instances of zero confusion at 50% cell shuffling, with this proportion corresponding to zero signal distinguishing the affected domains. This is thus attributable to a lucky guess and their inherent tendency to smooth domains. BASS, on the other hand, stands out as the only method that does not result in incremental confusion values, indicating its high performance and strict adherence to the bit flip failure mode described in Sec. 3.4.1.

In the case of the expression mixing perturbation, however, many methods show more bit flip-like behaviour (Fig. 3.7d). In fact, the confusion values exhibited across the range of perturbation proportions are nearly exclusively binary for several methods. Besides BASS, this includes, for example, MNMST, STAGATE, and SEDR. However, again, we also see that many methods confuse the *a priori* highly transcriptionally similar domains before the others. The onset of confusion of the other, transcriptionally more distinct domain pairs varies widely between methods. TACCO, BASS, ADEPT, and CCST stand out for zero confusion of those domain pairs before 25% perturbation. In general, more methods are not affected by this perturbation until higher perturbation proportions compared to the cell shuffling.

Generally, this investigation of pairwise domain effects compounds the tissue-wide observations of the previous section. Most methods are more immediately affected by cell shuffling than they are by expression mixing, in which case many methods exhibit a "bit flip" failure style.

3.5 Effect of domain shape and size

The last type of variation we investigate using semi-synthetic data is the size and shape of the individual domains. We evaluate the effect of domain size, first focusing on the thickness of laminar domains,

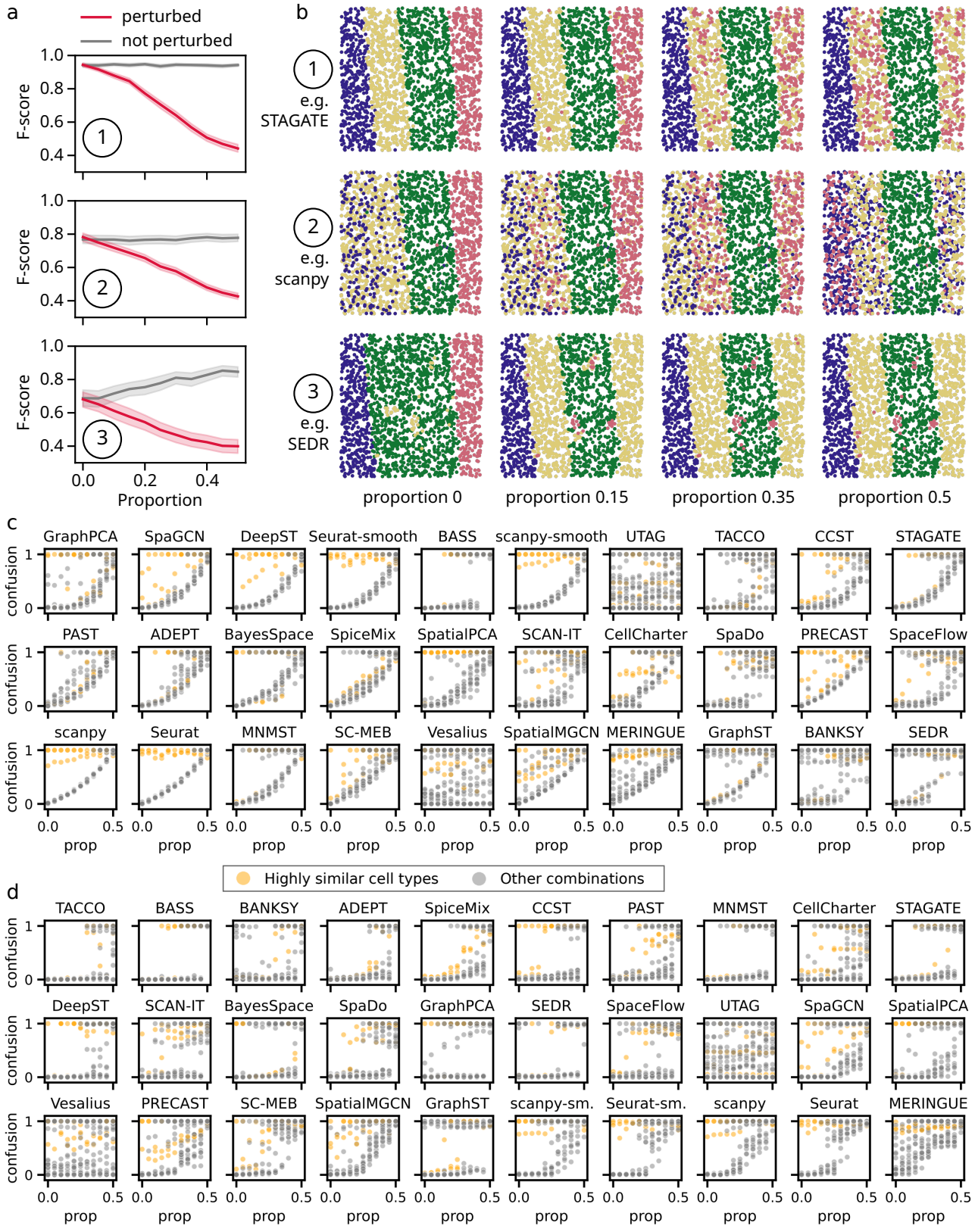

Figure 3.7: **Effect of increasing pairwise domain similarity.** a, Mean F_1 scores of all domains upon cell shuffling, stratified by whether the domain is affected by the perturbation. Methods are aggregated into the groups indicated in Tab. 3.2. b, Examples of method behaviours when shuffling cells between the red and yellow domains as annotated by STAGATE at a perturbation proportion of 0. Method groups are the same as in a. c, d, Confusion upon increasing perturbation levels, stratified by the similarity of the cell types defining perturbed domains. c, Increasing cell shuffling. d, Increasing expression mixing.

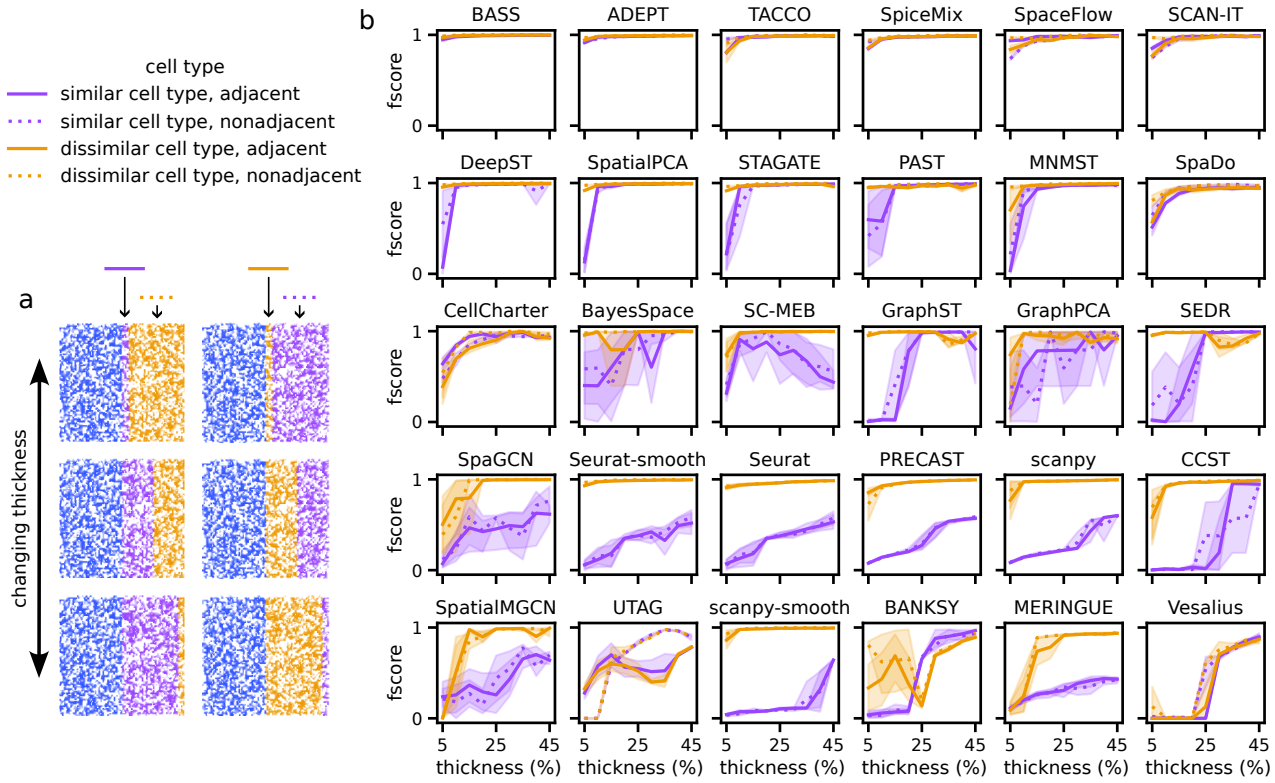

Figure 3.8: **Effect of domain thickness, transcriptional similarity, and adjacency.** a, Example tissue configurations showing the reference domain in blue (left domain), and purple and yellow domains of changing thickness. Configurations are shown for both settings of adjacency to the reference domain; that is, both examples of adjacent highly similar (purple, left column) and dissimilar (yellow, right column) are shown. b, F_1 scores as a function of domain thickness, stratified by transcriptional domain similarity and adjacency to the reference domain. Domain thickness is indicated by a percentage of the total tissue width. Methods are ordered according to mean performance, and the variance per datapoint is over five seeds.

and perform an additional analysis on circular domains of different diameters. Then, we compare method performances on tissues with different, archetypal domain shapes and layouts.

3.5.1 Laminar layer thickness

To investigate the effect of domain size, we create a range of semi-synthetic samples with three laminar layers (Fig. 3.8a). Starting from a thin middle layer, we shift the boundary between the middle and right layers gradually towards the edge of the sample, thereby simultaneously varying the thickness of two domains. Additionally, we perform this experiment for two distinct assignments of cell types to domains, as shown schematically in Fig. 3.8a. In both modes, we assign the cell types Ex_Rorb_Col8a1, Ex_Rorb_Col8a1_Cntnap4, and Ex_Slc30a3_Otof, where the first two cell types are highly similar (Fig. 3.1g,h). Starting from this set of cell types, we fixedly assign cell type Ex_Rorb_Col8a1 to the left, unchanged domain. The domain assignments of the other two cell types are swapped between the middle and right domains, both of which vary in thickness. This allows us to investigate how high transcriptional similarity modulates the effect of domain thickness. On top of that, it also enables the evaluation of adjacency effects between these highly similar domains. In order to increase the statistical power of the analysis, we create five samples for each tissue configuration, varying the random state underlying the assignment of cells to locations within the data generation pipeline.

We evaluate method performance using the F_1 score of the domains with changing thickness.

Additionally, we stratify the results by both transcriptional similarity of the underlying cell type to that of the reference domain and by the adjacency to that same domain. Methods are affected by the domain thickness in very distinct patterns (Fig. 3.8b). BASS, ADEPT, TACCO, SpiceMix, SpaceFlow, and SCAN-IT are barely affected. Some of these methods might show a dip at the smallest domain sizes, but all domains remain detected with F_1 scores above 0.76. For a different set of methods, consisting of DeepST, SpatialPCA, STAGATE, PAST, and MNMST, the dip at the thinnest domain becomes a more pronounced or even total total loss of detection. While these methods are only affected for the thinnest domains, GraphST, SEDR, and CCST decrease abruptly in performance already at larger domain thicknesses. Interestingly, for most of these methods, F_1 scores near zero for the thinnest domain only occur when that domain is highly transcriptionally similar to the reference. Only MNMST and CCST also appear to be affected in their identification of the transcriptionally distinct domains.

SpaDo and CellCharter decrease less abruptly in F_1 score when applied to thinner domains, and the effect is not conditional on the underlying transcriptional similarity levels. Another group of methods exhibit more complex nonlinear dependencies on the domain size, like BayesSpace, SC-MEB, GraphPCA, and BANKSY. The baseline methods and PRECAST show a gradual change in the detection of the transcriptionally similar domains. Similarly, SpaGCN, PRECAST, SpatialMGCN, and MERINGUE gradually decrease for the similar domains, but these methods, in contrast to the previous baseline-like group, are also affected on the dissimilar domains. Finally, UTAG shows a unique behaviour based exclusively on domain adjacency, whereas Vesalius is strongly affected by the domain thickness irrespective of adjacency and similarity, and is only able to distinguish domains from an approximate thickness of one-fourth of the tissue size in our experiment.

3.5.2 Size of circular domains

To corroborate our evaluation of size effects in domain identification performance in laminar tissue configurations, we next investigate size effects in circular domains. To that end, we create two configurations of semi-synthetic tissues containing three circular domains ("blobs"). In configuration I, blob sizes are kept roughly equal, while in the other configuration II, we vary the sizes of the resulting domains by indicating different dispersion parameters for the underlying clusters generated by scikit-learn. The resulting configurations are shown in Fig. 3.9a.

In the previous section, we found that the underlying domain similarity has a strong effect on the detection of small domains. Therefore, we devise a strategy to avoid any bias introduced by the cell type assignments in evaluating the size effect. Specifically, we permute the possible assignments of cell types to domains and evaluate methods on all permutations, finally aggregating the results.

First, we evaluate the methods by means of the F_1 scores attained on each blob across both configurations (Fig. 3.9b). Indeed, configuration II shows a considerable size effect, with larger blob sizes aiding detection across all methods. The biggest circular domain, termed blob A (see Fig. 3.9a), is consistently detected with F_1 scores over 0.8 by all methods except for SpatialMGCN, Vesalius, MERINGUE, and UTAG. The scores decrease for the mid-size blob B, and the smallest domain, blob C, has a median detection of only $F_1 = 0.44$. Only BASS, ADEPT, and TACCO reach F_1 scores of over 0.61 on this smallest domain.

However, by a closer look at the F_1 score results on the equal-sized tissue configuration I, it soon becomes clear that the F-score is a biased evaluation metric for our purposes here (Fig. 3.9b, top). Namely, in the vast majority of methods, the F_1 score of blob B, in configuration I, exceeds the scores of the other domains. This can not be a size-based effect, considering that the three blobs A, B, and C are designed to be of approximately equal size. To elucidate this phenomenon, we perform a thorough investigation of the domain segmentations found by different methods on both tissue configurations. Finally, we found that the tendency towards favourable evaluation of blob B results

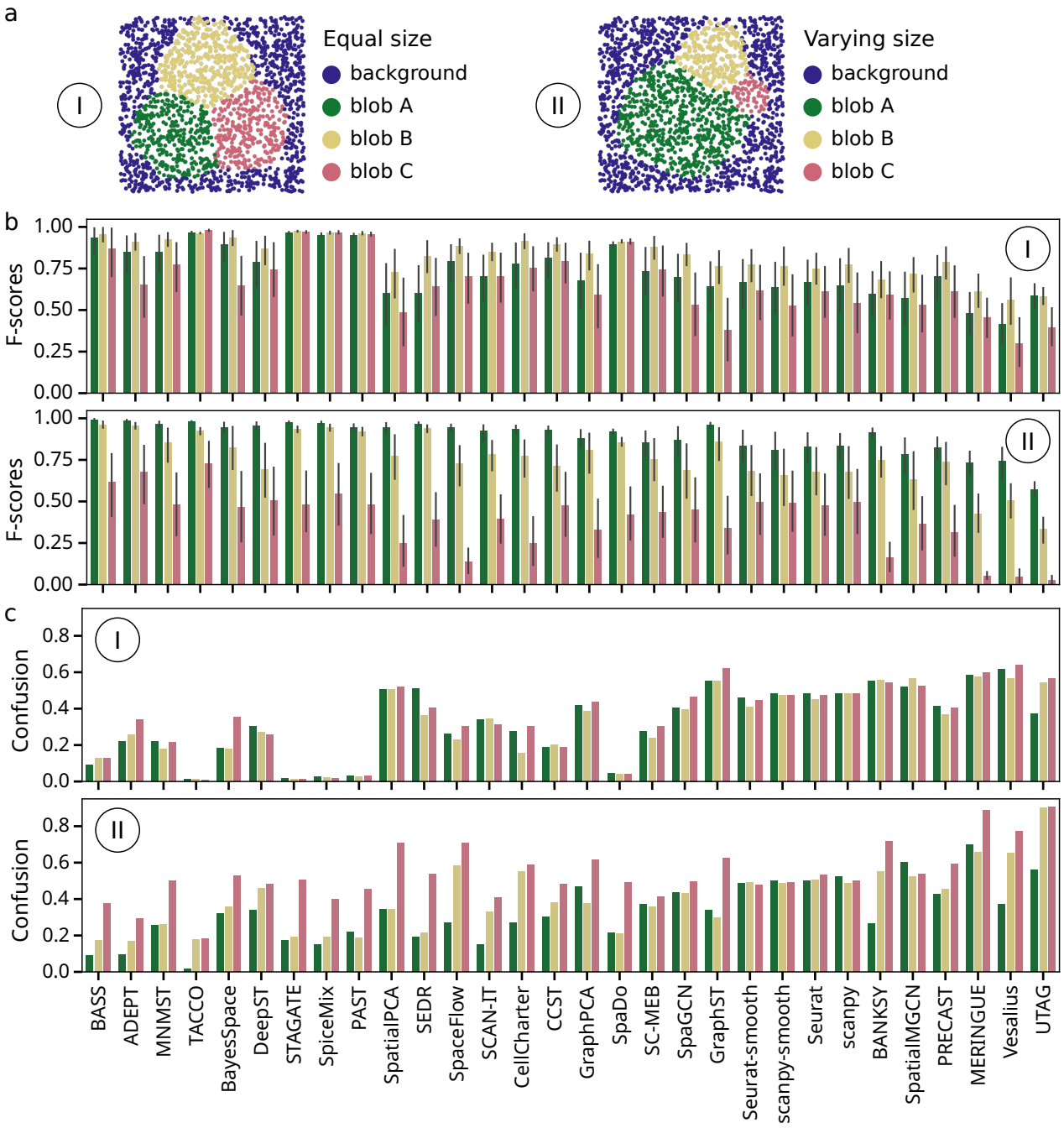

Figure 3.9: **Effect of domain size, evaluated in circular domains.** a, Example samples showing varying tissue configurations consisting of roughly circular domains. Configuration I consists of equal-sized domains, while the domains in configuration II are of varying diameters. b, F_1 scores per method for each circular domain, corresponding to the domains shown in a by colour. Bar plots are shown for configuration I with equal domain sizes (top) and configuration B with varying domain sizes (bottom). Methods are ordered from left to right by descending average ARI, calculated by the mean across both shapes of median performances. The variance per bar plot is over different assignments of cell types to domains. c, Mean domain-wise confusion per method for each circular domain, shown as in b. The domain-wise confusion is given by the maximal pairwise confusion involving the domain in question. Method order is the same as in b.

from an underlying bias in the harmonisation procedure undertaken to ensure label comparability. Concretely, considering integer ground truth labels a and b, each assigned to n_a and n_b spots such that $n_a \approx n_b$ but $n_a > n_b$, putative domain assignments that mix two ground truth domains completely are preferentially assigned label a. In the equal-size domain configuration, we have the case of $n_A = 336$, $n_B = 396$, and $n_C = 315$, such that a method output domain containing all spots from ground truth labels A and B will be assigned label A. This leads to a disproportionate exaggeration of any size effects present.

To circumvent this bias, we focus instead on the evaluation using our confusion metric, as introduced in Sec. 3.4.2. We calculate the pairwise confusion between all domains i, j and define the per-domain confusion as the maximum of its pairwise confusion

$$\text{confusion}_i = \max_{j \neq i} \text{confusion}_{i,j}, \tag{3.4}$$

where confusion_{ij} is calculated between domains i and j. Taking the maximum here is justified because we are not interested in differentiating the number of domains being confused. That is, whether one domain is confused with one or multiple others is irrelevant for our current purpose. Instead, we view the per-domain confusion as a kind of "winner takes it all" metric - for a domain to be fully confused, it is sufficient for it to be indistinguishable from one other domain.

Using this confusion-based approach, we evaluate all method performances and first investigate the baseline methods. We find that the baseline methods now cease to show a size-specific effect in the tissue configuration II (see Fig. 3.9c). This further confirms the validity of our analysis, as we do not expect domain size to impact the behaviour of purely transcriptome-informed clustering methods. As an additional validation in the case of equal-sized blobs generated in configuration I, we evaluate the variation over the confusion of the different blobs. This variation appears randomly distributed between all methods and is thus attributable to chance effects.

Finally, considering tissue configuration II with varying size circular domains, evaluating the impact of domain size through the lens of the per-domain confusion shows a less striking effect than what we found using the F_1 score. However, still, nearly all methods exhibit increasing confusion with decreasing domain sizes, notably for the smallest blob C domain. While many methods don't show a striking difference in confusion between differently sized domains, the only method that does not exhibit the highest confusion for the smallest domain is SpatialMGCN.

Over both the analyses with F_1 score and with our per-domain confusion metric, we thus see a strong effect of the domain size. The most pronounced decreases in detection are seen for "small" domains, with less pronounced effects between mid-size and large domains. We are also able to corroborate some findings from the size evaluation based on layered domains. Specifically, methods showing abrupt performance decline for thin domains in the layered configuration, like SpatialPCA, STAGATE, PAST, and MNMST, show the same confusion levels for the large and mid-size blobs A and B, and only increasingly confuse the smallest blob C.

3.5.3 Domain shape and tissue configuration

Finally, we want to evaluate whether the accurate identification of domains is affected by the domains' shapes and the tissue layout more generally. To this end, we create semi-synthetic samples mimicking four different archetypal of tissue layouts. Specifically, we generate individual circular, layered, and concentric circular domains, and a more complex, interlocking domain configuration, all shown in Fig. 3.10a. As with the previous evaluation of domain size in circular domains, we aim to minimise the bias introduced by cell type assignments do domains. To this end, we again permute over all possible cell type-to-domain assignments and average the results.

Since we are here again interested in whole-tissue performance rather than domain-specific effects, we evaluate method performances using the ARI score. No perturbations are applied to the tissue in

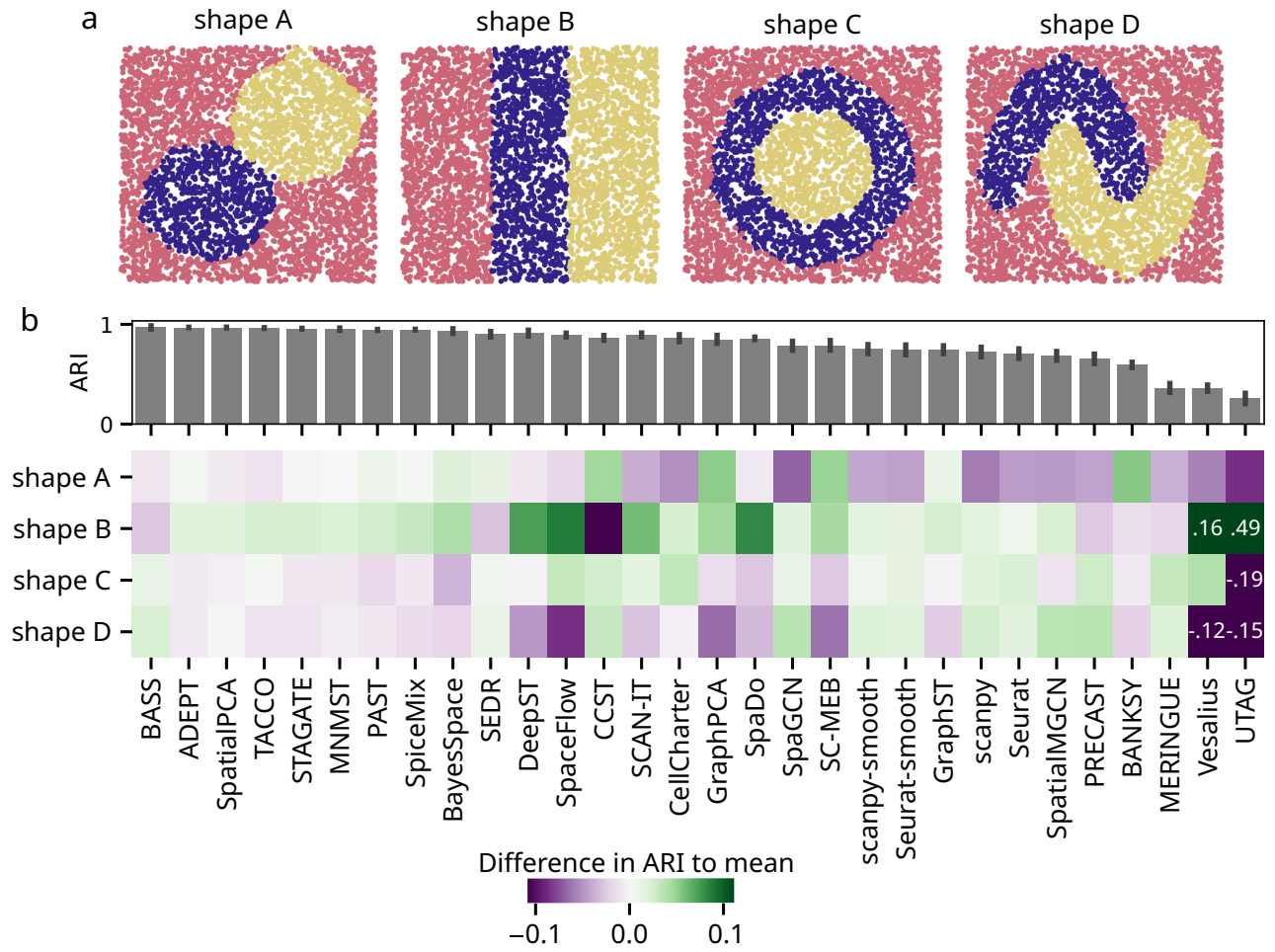

Figure 3.10: **Effect of domain shapes and tissue configurations.** a, Example samples showing the different tissue configurations examined in this experiment. b, Top (bar plot), mean ARI scores per method, averaging over the different tissue layouts (shapes). Bottom (heatmap), difference in ARI to the mean per method. Mean differences are indicated by colour, where positive values (green) indicate better-than-average method performance on a given shape. The colour scale is capped at the range of -0.11 to 0.11 for better visibility; larger differences are indicated within the heatmap by value.

this analysis, leading most methods to perform very well on all samples, as shown by the mean ARI in Fig. 3.10b). Still, we investigate the differences in ARI on each tissue layout to disentangle potential shape effects.

Interestingly, most methods perform the best on the layered configuration. This improvement over the mean, and the general effect of tissue layout on method performances, is modest in most cases, ranging within ARI differences of -0.1 to 0.1. The most strongly affected methods within this range of effects are DeepST, SpaceFlow, SCAN-IT, and SpaDo, which perform best on layered tissues, and CCST, which performs decidedly worse on layered configurations than on all other shapes. In contrast, Vesalius and particularly UTAG show significantly larger effects, markedly favouring the laminar, layered tissue configuration. Specifically, they outperform their average performance by 0.16 and even 0.49, respectively, when applied to layered tissue.

While the Vesalius and UTAG show a strong preference towards finding layered domains, most other methods only show small effect sizes concerning domain shape and tissue layouts.

Chapter 4

Additional results from secondary evaluation criteria

This last chapter presents additional analyses from the benchmarking pipeline introduced thus far, focusing on secondary evaluation criteria. Specifically, I describe how we evaluate runtime, memory usage and usability of the methods, and show results from these investigations.

4.1 Runtime and memory benchmarking

The evaluation of runtime and memory usage is an important part of method benchmarking, complementary to performance reporting. The following sections describe the setup enabling us to perform this analysis, and show runtime and memory usage results across methods and different datasets.

4.1.1 Evaluation setup

Having implemented a comprehensive Snakemake-based method benchmarking workflow directly enables us to measure the runtime and memory usage of the methods on each analysed sample. Specifically, Snakemake rules can take the `benchmark` directive, configuring jobs resulting from these rules to directly output wall clock time and memory usage to a user-specified text file.

Analysing method runtime and memory usage on the real data samples included in the benchmark provides general insights into method behaviour. Additionally, we are interested specifically in the evaluation of method scalability, in terms of these secondary evaluation measures, with respect to the number of cells or spots. To that end, we utilise the SRTsim simulator introduced in Sec. 3.1.2 to generate samples of varying cell numbers [214]. Specifically, we utilise the *de novo* mode of data generation, simulating the expression of 500 "signal" and 500 "noise" genes. The simulator is configured to create random cell locations within a square tissue layout, and we define four rectangular layers as spatial domains. Within that layout, we vary the number of locations in the tissue between 2000 and 100'000. Additionally, to increase statistical power, each sample size is generated for three random seeds. The scalability experiments were evaluated on an AMD Ryzen Threadripper 3990X 64-Core Processor @ 4.3GHz, equipped with a Nvidia GeForce RTX 3090 GPU. For the scalability evaluation, differently to the experiments on real and semi-synthetic data, methods that have the capability of utilising GPUs were configured to do so.

4.1.2 General runtime and memory results

As a first overview, we evaluate the runtime and memory usage of all methods on the real datasets. We find that methods range widely in both quantities, with runtimes between under 20 seconds and

close to 8 hours (Fig. 4.1a), and memory usage between barely over 200 MB and almost 100 GB (Fig. 4.1b).

Generally, it is apparent that both runtime and memory usage are affected across methods by the size of the count matrix. This size is not only determined by the number of cells or spots, but importantly also by the number of profiled genes. Particularly, memory usage, but also runtimes on datasets from the Visium and Slide-seq technologies are significantly increased. Interestingly, Vesalius stands out with both the longest runtime and the highest memory usage of any method on any dataset, with its performance on Visium–Fu. This dataset, in contrast to the other Visium dataset, exhibits a relatively complex domain structure and more ground truth domains than any other dataset we include (see Appendix B).

In terms of the runtime, CCST takes an average of close to 6 hours to run on the high-resolution osmFISH and MERFISH datasets, and MNMST uses a similar amount of time on the Slide-seq data. Considering the memory usage, ADEPT takes about one order of magnitude more than the average method, across all datasets. In terms of fast runtimes and low memory usage, no method clearly stands out. However, a number of methods keep their runtimes under five minutes consistently, and while memory usages are increased across the board for the Visium and Slide-seq datasets, the majority of methods still use under 5 GB of memory.

4.1.3 Scalability

Following up on the analysis using real data, we aim to isolate the effect of an increased number of cells or spots. To this end, we evaluate method performances on simulated samples encompassing a range of sizes. For better distinguishability of individual method trends, we split the methods into four constituent groups based on slope quantiles in both runtime and memory. There is a tradeoff between memory and runtime for some methods, though a large subset of methods shows either steep or shallow increases across both measures.

Among the methods with both good runtime and memory scaling are ADEPT, BANKSY, CellCharter, PAST, STAGATE, SpaceFlow, and TACCO (Fig. 4.2a,a'). These methods range in runtimes from a few seconds to under a minute on the smallest samples (2000 cells), and just over a minute to less than ten minutes on the largest samples (100'000 cells). Memory usage ranges from just over 300 MB to under 4 GB on the smallest, and just under 2 GB to close to 10 GB on the largest samples. The best scaling with respect to runtime is shown by BANKSY, while TACCO stands out for low memory usage. ADEPT is the only method in this group which does not produce output on data with over 10'000 cells.

Methods which show decent scaling with respect to runtime, but increase steeply in memory usage, include BASS, BayesSpace, CCST, PRECAST, SEDR, SpaDo, and SpiceMix (Fig. 4.2b,b'). Among these methods, SpiceMix stands out with high runtimes, starting out at close to 30 minutes for 2000 cells, and steep memory scaling. It does not produce output for samples with over 6000 cells. BayesSpace, CCST, and PRECAST use over 100 GB of memory for the largest samples, while BASS, SpaDo, and SEDR do not produce any output. There are only a few methods showing the opposite behaviour of runtime and memory scaling, namely steep increases in runtime while memory usage remains relatively stable (Fig. 4.2c,c'). Of the three methods showing this behaviour, SpatialMGCN is the only one which results in any output for samples with high numbers of cells, taking over 6 hours to finish running on the largest sample. By contrast, DeepST and SC-MEB show similar increase trends, but cease producing clustering results for samples with over 10'000 and 20'000 cells, respectively.

Lastly, the remaining methods that show favourable scaling neither in runtime nor in memory usage include GraphPCA, GraphST, MERINGUE, MNMST, SCAN-IT, SpaGCN, SpatialPCA, and UTAG (Fig. 4.2d,d'). While all of these methods start out with runtimes of under 4 minutes and memory usages of up to 3 GB on the smallest samples, they increase rapidly in usage of both resources as the

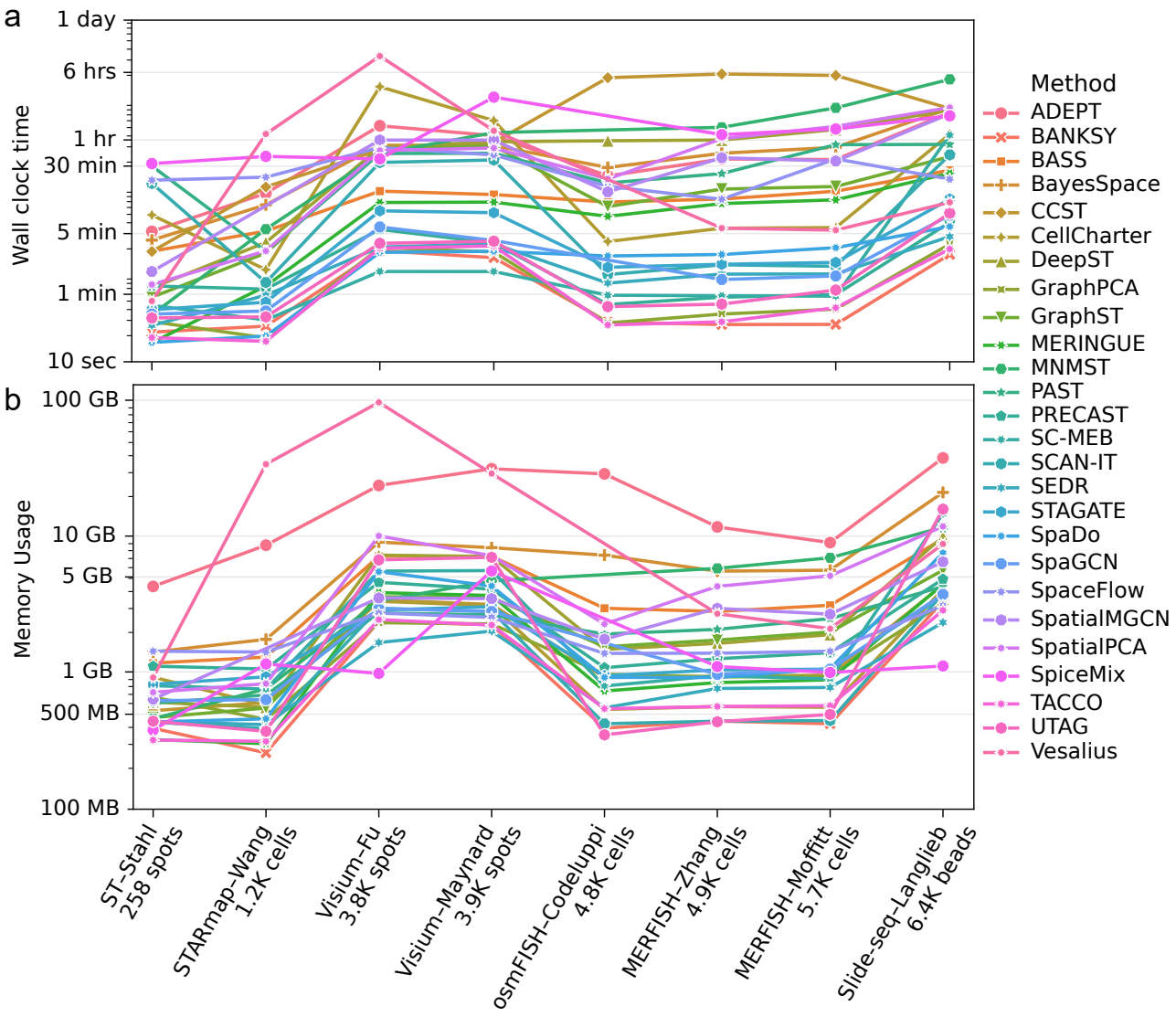

Figure 4.1: **Runtimes and memory usage on real data**. Wall clock runtime (a) and memory usage (b) of all methods on all real samples, aggregated by mean per dataset. Datasets are sorted by the mean number of spots or cells.

number of cells increases. SpatialPCA does not produce any output for samples larger than 6000 cells, MNMST and UTAG cease producing output above 20'000 cells, and GraphST fails to result in an output on the largest sample of 100'000 cells. Except for SpatialPCA, all methods use at least 300 GB of memory on the largest sample for which they result in any output. GraphPCA and MERINGUE stand out with a maximum memory usage of over 200 GB on the largest sample, and MERINGUE also takes the cake in runtime, running for almost two days to generate a result.

Vesalius did not produce any output for the samples generated using SRTsim. We were not able to determine what caused this.

4.2 Usability evaluation

In addition to runtime and memory usage, the usability of methods plays a significant role in their broader adoption by the research community. For completeness of the benchmarking, it is therefore interesting to evaluate methods on usability criteria, encompassing user-friendliness and accessibility. These assessments tend to be subjective, but can be made more objective by the use of predefined checklists. The results shown in this section are preliminary, created through a basic usability checklist.

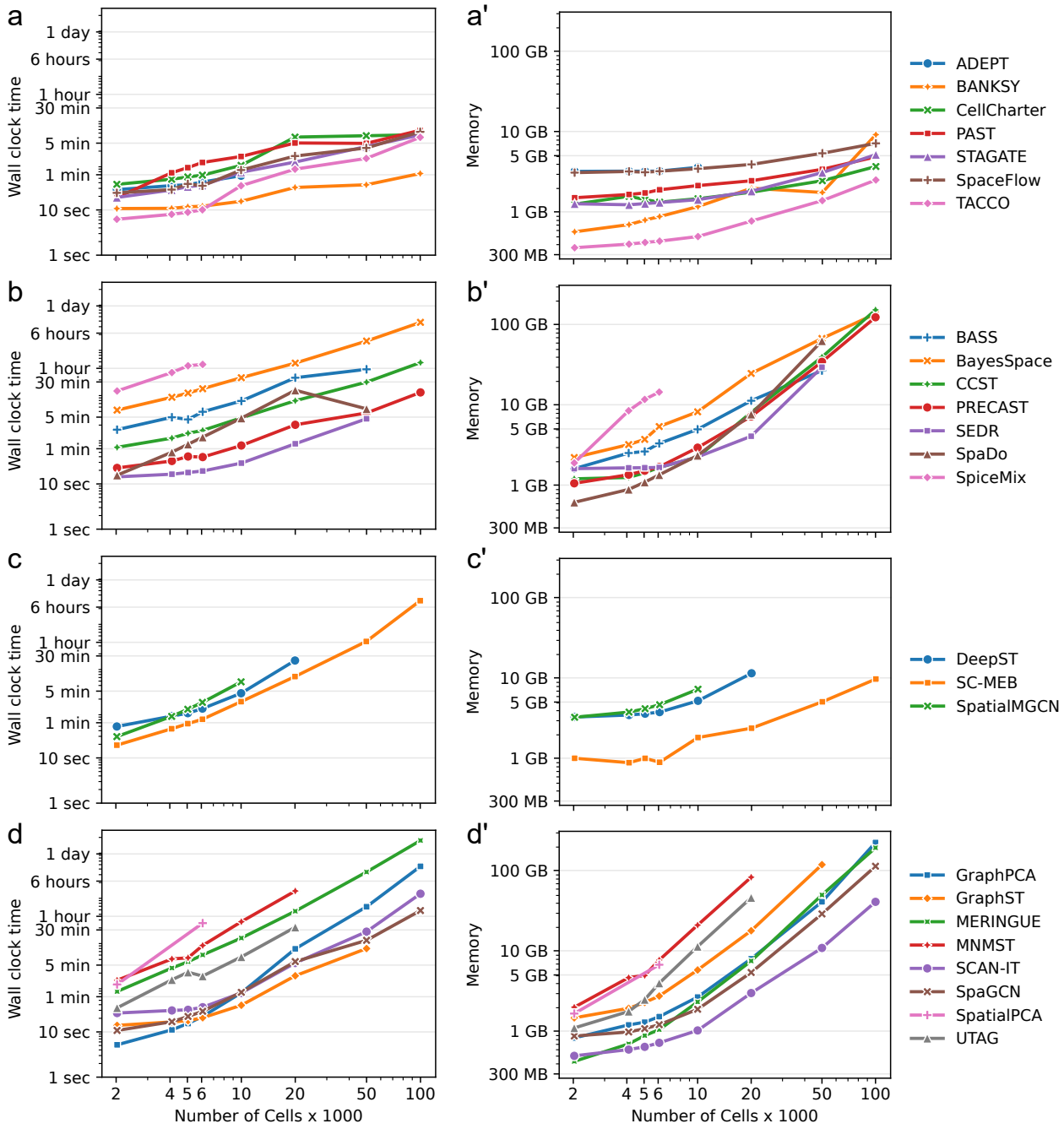

Figure 4.2: **Scalability of methods concerning runtimes and memory usage.** Wall clock time and memory usage by all methods on *in silico* samples with varying numbers of locations. Unprimed subplots (left) show runtime, primed subplots (right) show memory usage. Methods are split into four groups based on scaling trends in each quantity. Both the x and the y axes of all plots are shown on a log scale. a, a', Slow increase in runtime and in memory. b, b', Slow increase in runtime and fast increase in memory. c, c', Fast increase in runtime and slow increase in memory. d, d', Fast increase in runtime and memory. Methods are categorised into slow and fast increases based on percentiles such that the resulting groups contain roughly equal numbers of methods.

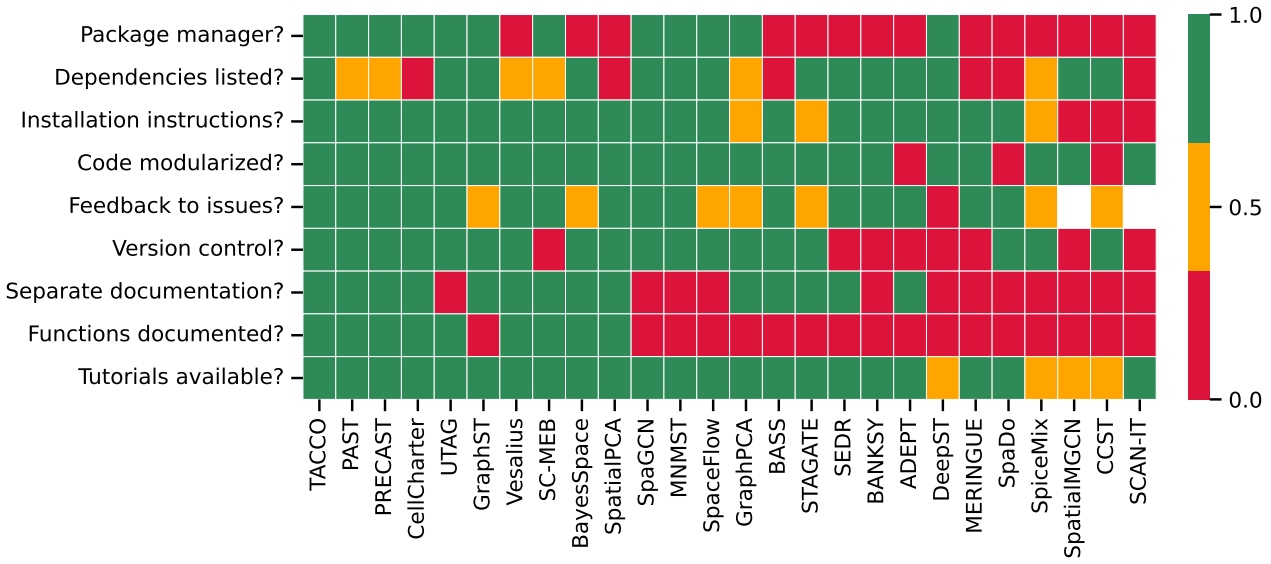

Figure 4.3: **Usability of methods.** Evaluation is carried out using the checklist in Tab. 4.1. Methods are sorted by mean performance over all criteria.

In particular, we took inspiration from a scoring checklist utilised by Duo *et al.* in their benchmarking of single-cell and spatial transcriptomics simulators [162]. We implemented a simplified checklist, adapted to the needs and circumstances of the spatial domain identification methods included in our benchmark. The checklist is shown in Tab. 4.1. We categorise aspects of method usability into Availability, including installation procedures, Maintenance, which incorporates criteria relating to good coding practices and continued support through platforms such as GitHub, and Documentation. Each category of usability is interrogated using three specific questions, the answers to which are mapped to values between 0 and 1, with 1 corresponding to the best and 0 to the worst outcome. An additional value of 0.5 is possible in some questions, indicating partial fulfilment of the criterion. All of the methods included in this evaluation are open-source and freely available through GitHub, so we primarily base our answers to the checklist on the public GitHub pages.

We find that only TACCO completely fulfils all of our usability criteria (Fig. 4.3). While all methods except for CCST, SpatialMGCN, and SCAN-IT provide at least rudimentary installation instructions within the `README.md` files of their GitHub repositories, several methods, including CellCharter, SpatialPCA, BASS, MERINGUE, SpaDo, and SCAN-IT, do not list necessary dependencies in an easily accessible manner. Additionally, only just under half of all methods provide their tool as a package easily installable using conda, pip, or Bioconductor. The remaining methods are primarily installed directly through GitHub, which might pose a challenge to less computationally savvy prospective users.

Concerning maintenance criteria, the majority of methods show good usability. Except for ADEPT, CCST, and SpaDo, all methods publish modularised or otherwise transparently structured code. This is relevant to potential method users for increased understanding of methods' inner workings. Additionally, in some cases, users might need to modify functions locally to work in their own setups[1]. Besides DeepST, all tools for which any GitHub issues had been opened provided full or partial responses to those issues. Lastly, two-thirds of all methods have established version control, encompassing full versioning or publication tags.

On the other hand, many methods do not provide adequate documentation according to our criteria, with 11 out of 26 methods not documented beyond usage tutorials. All methods do provide at least one tutorial, and except for DeepST, SpiceMix, SpatialMGCN, and CCST, even multiple tutorial

[1]For example, the mclust function of SEDR contains a hardcoded setting of the R home directory [170].

Category	Question formulations	Point assignments
Availability	(a) Is the method installable through a package manager/repository (like conda, pypi/pip or Bioconductor)?	yes - 1, no - 0
	(b) Are all dependencies listed in the README file or in a clearly referred to requirements file, including specific required versions?	yes including versions - 1, yes without versions - 0.5, no - 0
	(c) Are installation instructions provided and comprehensive, accessible with basic coding knowledge?	yes accessible - 1, yes not accessible - 0.5, no - 0
Maintenance	(a) Is the code modularised and/or generally structured in a straightforward and transparent manner?	yes - 1, partially - 0.5, no - 0
	(b) Do authors respond to issues raised on GitHub?	yes to all - 1, yes to some - 0.5, no - 0, no issues opened - NA
	(c) Is a type of versioning implemented, either through GitHub or a package manager?	yes - 1, no - 0
Documentation	(a) Is there a dedicated package documentation, e.g. through readthedocs.io?	yes - 1, no - 0
	(b) Are individual functions well-documented, through an API documentation or as code comments?	yes - 1, partially - 0.5, no - 0
	(c) Are tutorials on how to run the method available?	yes multiple - 1, yes one - 0.5, no - 0

Table 4.1: **Criteria for usability evaluation**. Checklist questions relating to the usability of methods, grouped in criteria of availability, maintenance, and documentation.

versions. Additionally, slightly over half of the tools are additionally documented in a dedicated site. However, most methods do not especially excel in terms of function documentation. Only nine methods provide either a dedicated API documentation or thorough code comments to explain function parameters and usage. This indicates that most tool developers rely on the prospective users following tutorials closely to figure out the workings of individual functions. This may, in some instances, significantly complicate the adoption of tools for novel technologies or data types, for which no dedicated tutorial exists.

Overall, while most methods fulfil the majority of our usability criteria, there is considerable room for improvement, especially in the realm of comprehensive documentation.

Chapter 5

Discussion and Conclusions

Within the field of spatial transcriptomics, as in the broader bioinformatics community, new computational analysis approaches evolve alongside technological development. A wealth of different computational approaches to a diverse set of data analysis types has been and continues to be developed. In this context, independent benchmarking studies evaluate existing methods, providing an overview of the current state of the art. They intend to guide the research focus of method developers, as gaps in the literature and application areas become apparent. Further, they serve to educate prospective method users, giving an overview of applicable tools and demonstrating their respective strengths.

This thesis presents the effort of, and results from, a benchmarking evaluation of methods for spatial domain identification. The entire benchmarking pipeline was implemented using Snakemake for workflow management and integration with conda, enabling reproducible and portable analysis. We selected 26 methods for benchmarking and used real, publicly available spatial transcriptomics datasets from a range of technologies for their initial evaluation. After generating a set of hypotheses about the effect of various data characteristics on method performance, we created a pipeline for the tunable generation of semi-synthetic spatial transcriptomics data. This custom pipeline enabled us to vary parameters corresponding to features of spatial transcriptomics technologies, as well as tissue-inherent factors, in turn allowing us to carry out a systematic investigation of how these factors affect method performances. Additionally, we evaluated the stability and robustness of methods to perturbations, and investigated consensus approaches as a competitive and robust alternative to individual methods. Lastly, and importantly for a comprehensive method comparison, we benchmarked the runtime and memory usage of all individual tools with a focus on scalability, and graded the methods on a usability scale.

5.1 Benchmarking setup and pipeline

Methods were selected for evaluation based on informal criteria of relevance, usability, and variety of approaches. We settled on the inclusion of 26 individual methods, which we be broadly categorised into clustering-based, neural network-based, statistical modelling-based, and image processing-based groups. The methods were first published over a number of years, ranging from 2020 to 2024. We only included methods published after June 2024 if they had been previously uploaded to bioRxiv and we had already included the tool based on this preprint version. A number of methods are first made public in the preprint format on platforms like bioRxiv, enabling the community to access tools and resources before the termination of peer review for traditional publication. This creates an opportunity for method developers to make their approaches known to potential users and other interested parties. In certain cases, especially when traditional publication is delayed by various possible factors, preprint publication can lead to methods being widely adopted before their eventual publication. Such is the case for methods like SEDR, which was published by Genome Medicine in 2024 [170]. However,

Dataset	Ref. dataset	First publication of technology, and relation to dataset
ST–Stahl	[37]	[37] (same publication)
Visium–Maynard	[114]	– (commercial technique, first published study)
Visium–Fu	–	– (commercial technique, example data resource)
Slide-seq–Langlieb	[200]	[39] (same research group)
STARmap–Wang	[48]	[48] (same publication)
MERFISH–Moffitt	[199]	[46] (same research group)
MERFISH–Zhang	[197]	[46] (same research group)
osmFISH–Codeluppi	[44]	[44] (same publication)

Table 5.1: **Relationships of datasets and technologies.** The relation of the datasets included in this benchmark to the technologies by which they were generated.

enabled by its having been made accessible as a preprint on bioRxiv in 2021, it is one of the most highly cited methods in the field (ranked 5th out of the 26 methods included in our benchmark, with over 200 citations as of October 2, 2025).

As for the selection of datasets for method evaluation on real data, we were able to include 8 datasets from 6 different technologies. The technologies span a wide range of the available technological parameter space, ranging from low-resolution, full-transcriptome sequencing approaches to high-resolution, targeted smFISH-based techniques. We aimed to include a large number of datasets, but were heavily constrained by the availability of ground truth domain annotations. Of the datasets we were able to include, 6 are of the mouse brain, 1 is of the human brain, and 1 is of a human breast cancer sample. All but two of the datasets were published by the same research group that originally developed the technology utilised in the data acquisition, as detailed in Tab. 5.1. Notably, three of the datasets are published as part of the original technology demonstration. Both datasets that are not directly affiliated with the technology development are generated using Visium, a commercial approach based on Spatial Transcriptomics (ST). One of these datasets, namely the dorsolateral prefrontal cortex dataset published by Maynard *et al.* in 2021, represents the first data published using the (at that point) newly demonstrated technology [114].

It may be interesting to consider the implications of most of our included datasets being generated, if not by, then in direct relation to the original developers of the utilised techniques. This shows, on one hand, that first publications utilising a novel approach may take special care to present data in such a way that it is usable as a resource. On the other hand, it may be an indication that techniques developed by specific research groups may not generalise easily to different circumstances or resource availability. Some techniques may only be applicable in highly specialised research environments, restricting their usability by the interested community.

As for the selection of evaluation metrics, we classify the available quantitative strategies into supervised (utilising the comparison to a ground truth annotation) and unsupervised (based on only the putative clustering). By far the most prevalent metric in the field is the Adjusted Rand Index (ARI), which evaluates the "goodness" of a putative clustering by its correspondence to a "true" data labelling. This supervised approach to spatial domain evaluation carries some issues, notably and most importantly, the necessity of a trustworthy ground truth data annotation. This is problematic in multiple aspects. First, a detailed annotation of the data in question is often performed by experts and thus necessitates a considerable investment of time and resources. Further, the annotation is often based on, or aided by, an accompanying histological image. These images are generated alongside the data acquisition process in technologies like Visium, making them easily accessible for downstream evaluation. However, for other approaches, histological imaging presents an additional step to be completed during data generation, again representing time and resources invested. Ad-

ditionally, histology-based spatial domain annotation may produce a bias in the field through the resulting ground truth – namely, aiming to identify the same structures as visible in histology again, this time through transcriptomics-based strategies. This may obfuscate the identification of purely transcriptionally defined tissue structures, which are not evident by visual examination.

In cases where no histology information is available, domain annotations are frequently inferred by the evaluation of marker genes for known tissue regions. This is relatively straightforward to implement for exceedingly well-studied tissues like the mouse brain, where curated marker gene lists for different structures are available. It poses problems, however, in more complex or understudied tissues. Particularly in samples originating from tumours, there may not be known gene sets available for tissue annotation. An interesting approach is taken by the authors of SEDR in annotating the human breast cancer dataset included in our benchmark [170]. Namely, they base their annotation of spatial domains on the previously annotated cell type labels. Accurate and well-informed cell type labelling is a long-standing focus of the single cell transcriptomics field, and may thus be used as a starting point for the annotation of tissue structures and domains.

This leads to the last, and potentially most relevant, issue concerning the identification of a ground truth for domain identification. As briefly touched upon in the introduction to the present thesis, to the best of my knowledge, *there is currently no consensus in the field about the definition of spatial domains*. As detailed in Fig. 1.6 of the introduction, many tools are published under loose working definitions, citing expression coherence or cell type composition. Other methods simply operate under the mantle of spatially-augmented clustering, avoiding the necessity of defining a specific goal, or give no definition for the structures they aim to identify. Few methods, only 2 out of a sample of 33 interrogated for this thesis, define domains as being functionally distinct from the surrounding tissue. This definition is broad and does not translate directly to a well-formulated aim for method development. Generally, any biologically solid definition for domains identifiable through transcriptomics would have to be "translated" into the language of computing, necessitating an additional level of abstraction. In the scope of this thesis, I am not able to further investigate or attempt to close this gap in the research. However, I am convinced that for purposeful and streamlined method development, and the clear demarcation of the field of applicability of these methods, it is imperative to work toward a well-defined concept of spatial domains.

5.2 Method evaluation on real and semi-synthetic datasets

As a preliminary investigation, we considered the most popularly used supervised metrics for spatial domain evaluation. Comparing the Fowlkes-Mallows index, the Adjusted and Normalised Mutual Information metrics, and the Accuracy to the ARI, we showed that they are largely equivalent in their assessment of clustering correspondence to the ground truth labelling. Using the example of outlier inconsistencies in the domain assessment across metrics, we demonstrated the advantage of the ARI for the purpose of evaluating spatial domain identification performance.

Further, we established a consensus approach, facilitated by the implementation of a host of methods in our comprehensive Snakemake pipeline. We demonstrated that, particularly, the unbiased consensus evaluation over all method outputs represents a stable and competitive alternative to any individual method.

In the following sections, I will discuss insights gleaned from analysing method performances on the real data and connect them to detailed and systematic investigations we performed using our semi-synthetic data generation pipeline.

5.2.1 Technological variation

On the real data, we found that method performances vary widely between different datasets. Particularly the resolution appeared to strongly affect method performances. While the dedicated spatial domain identification methods showed a strong improvement upon the baselines on the single cell-resolved datasets, no strong improvements were attained in the lower-resolution Visium data. Besides the resolution, technologies differ in the number of genes which are profiled, and the sparsity of the resulting count matrix. The number of profiled genes varies widely between our included datasets, from 33 in the case of the osmFISH dataset to full transcriptome profiling for the sequencing-based approaches. Interestingly, in the real data, no direct relationship of the number of genes in a dataset to the method performances was apparent. Particularly, whereas many methods performed worse on the MERFISH datasets (with 200–300 genes in the gene panel) than on the full-transcriptome Visium data, the average performance on the osmFISH dataset was consistently high. To evaluate whether there are more complex relationships at play, we later used our semi-synthetic data to investigate the effect of differing gene numbers. Lastly, the sparsity of the data appeared to play a significant role in method performances. The Slide-seq dataset, with a sparsity of 98%, exhibited very weak method performances across the board.

This strong dependence on the dataset and particularly the technology led us to investigate the effect of technology-level data characteristics systematically. We considered the effect of resolution by binning semi-synthetic single-cell-resolution data into progressively larger "spots", averaging over the individual cells' gene expression levels. We found that of the 9 competitive methods which perform significantly better on MERFISH–Zhang than on Visium–Maynard, 6 were within the top 8 methods exhibiting the strongest declines in performance with decreasing resolution. Conversely, the baseline methods and a number of methods which performed similarly significantly better on Visium–Maynard actually exhibited an increase in performance associated with larger spot sizes, and thus smaller resolutions.

In terms of the effect of changing the number of profiled genes, all methods tended to decline in performance at small numbers. However, they varied in the onset and rapidity of this decline. Methods which performed better on Visium–Maynard than MERFISH–Moffitt in the real data evaluations exhibited the steepest ARI slopes, indicating that they are strongly affected by technological multiplexing capability. Interestingly, BayesSpace, which was originally designed for spot-level Visium or ST data, was only strongly affected by declining gene numbers once the panel size shrank below 100 genes.

Modifying the sparsity of the semi-synthetic data, it is unsurprising that all methods declined in performance when the sparsity neared 100%. Rather, it is interesting to evaluate methods by the onset of the decline, and the performance at extremely low signal availability. Some method performances only started to decline at sparsities of 0.95%, and at 99% sparsity, 7 methods still recovered enough domain-specific signal to reach ARI scores above 0.42. All in all, several methods appeared suited for the analysis of highly sparse data.

5.2.2 Tissue-level perturbation

Besides the direct technological parameters, we also investigated a related phenomenon, which could be contributing to the strong baseline performances on Visium datasets. Essentially, we noted that through the probable aggregation of multiple cells into one spot in low-resolution technologies, a technology-inherent gene expression smoothing operation is performed. The idea is illustrated in Fig. 5.1, considering an example gene which is highly expressed in only one of the semi-synthetic domains. As the resolution of the data is decreased (Fig. 5.1a), and gene expression is aggregated within spots, the region defined by high expression of the example gene becomes more contiguous and visually easier to identify (Fig. 5.1b). This qualitative interpretation is further corroborated by 2D

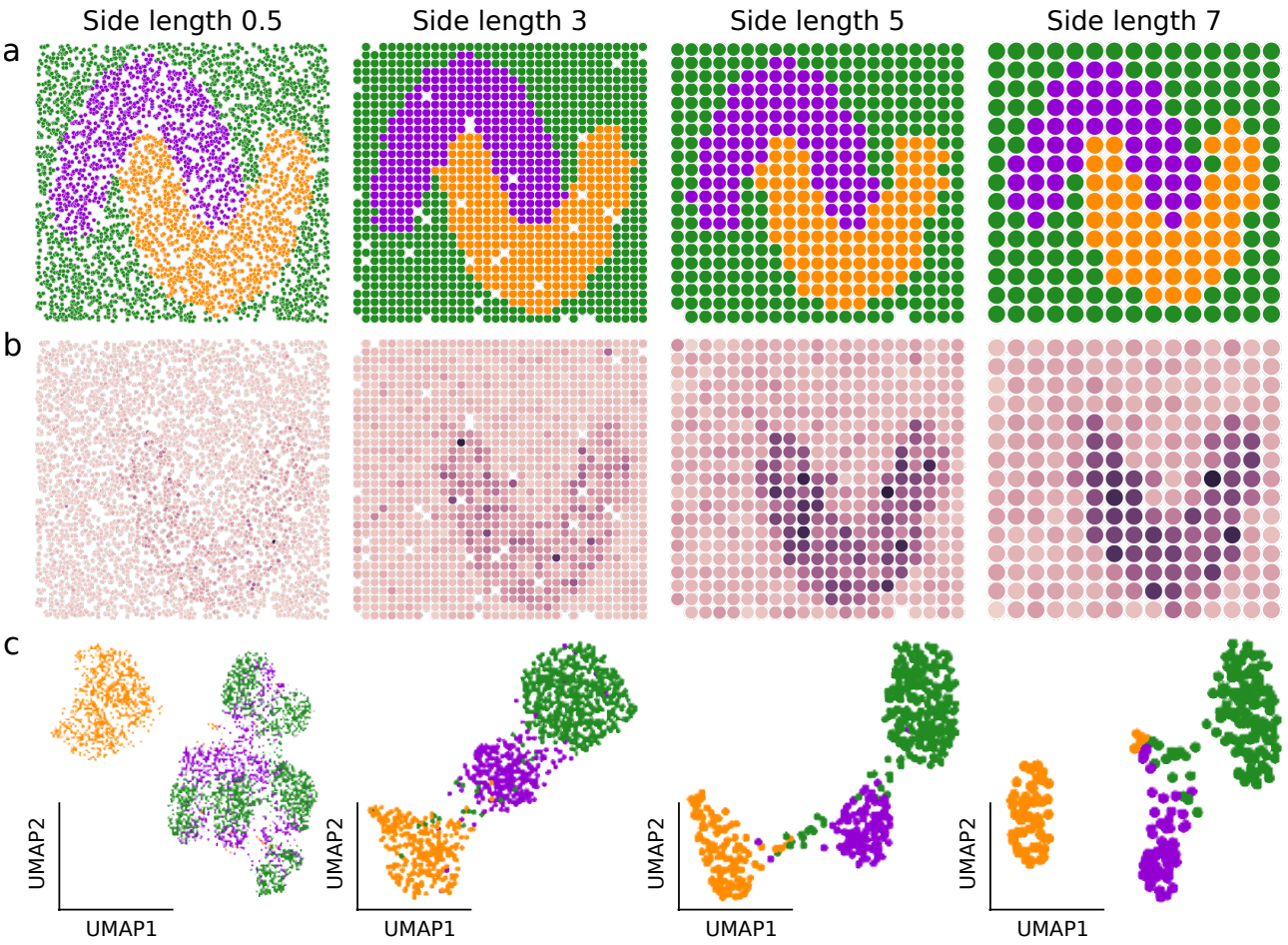

Figure 5.1: **Effect of decreasing resolution on distinguishability of domains.** Synthetic data is shown for four settings of the sample resolution as parametrised by the spot side length, ranging from 0.5 (close to single-cell resolution) to 7 (corresponding to a resolution between the Visium and ST technologies). a, Ground truth domain annotation. b, Expression of example gene Opcml. Expression values are scaled from min to max per resolution; darker colours correspond to higher expression. Within the data generation pipeline, gene expression values for lower resolution (higher spot side length) are created by the mean over cells binned to the corresponding spot. c, UMAP embeddings of the semi-synthetic gene expression at varying settings of the spot side length, showing more visually separated clusters at lower resolutions.

UMAP embeddings of the gene expression space (Fig. 5.1c), which show clearer and more distinct clusters as the resolution is decreased. In the same vein as the clusters being visually more easily distinguished in the UMAP, the baseline expression-based clustering methods would identify these constituent domains more readily.

The hypotheses emerging from these observations are twofold. The first can be formulated as the conjecture that high-performing methods for single-cell resolution data should be able to handle high levels of intercellular transcriptional heterogeneity, which would negatively affect the domain identification performance of baseline-like methods. As for the second hypothesis, we observed that the baseline methods are very adept at distinguishing domains when they correspond to transcriptionally defined clusters. However, excellent spatial methods may be able to distinguish domains with a higher transcriptional similarity to the rest of the tissue, based on subtle, spatially associated gene expression differences.

We investigated both of these hypotheses using our semi-synthetic data generation setup by adding varying levels of different perturbative noise types to the data. Indeed, we found that most of the

spatially-informed methods are able to distinguish domains defined by highly similar cell types, which are indistinguishable to the baseline approaches. We were able to identify three archetypes of method failure on highly perturbed data, distinct in their respective characteristic ARI/PAS curves. Interestingly, we found that many methods decline in performance in a nearly stepwise fashion upon increasing transcriptional similarity, a finding which is slightly obscured by the tissue-wide performance evaluation carried out using the ARI. To elucidate this observation, we additionally performed pairwise domain perturbations, allowing for a more detailed evaluation of method performance as two cell types progressively converge to a common gene expression profile. Here, we found that a large group of methods exhibits a nearly binary confusion performance, with little in-between confusion states.

Investigating the effect of cellular heterogeneity, we encountered a different picture – namely, many methods declining gradually in performance. This, along with the significantly higher PAS values exhibited on this perturbation by the vast majority of methods, indicated that these tools continually found transcriptionally defined clusters, neglecting spatial contiguity as expression heterogeneity increased. A small group of methods stood out for their high performance on heterogeneous domains. Besides the generally high-performing methods BASS, TACCO, and SCAN-IT, this notably included SpaDo and SpaceFlow, which ranked in the top 4 methods considering their robustness to heterogeneity (compared to ranks 10 and 11 in robustness to transcriptional similarity). Similarly, in the pairwise heterogeneity perturbation, they were among the methods maintaining low confusion at the highest perturbation levels. Interestingly, they both performed significantly better on MERFISH–Zhang than on Visium–Maynard in the real data, indicating a possible performance advantage on high-resolution data awarded by robustness to intercellular heterogeneity.

5.2.3 Domain sizes and shapes

Interestingly, while methods designed for spatial domain identification outperformed the spatially unaware baselines on most real data, this was not the case for the ST–Stahl dataset. On ST–Stahl, the non-spatial baseline methods reached performances up to ARI = 1, corresponding to spotwise perfect domain annotation. Investigation of the dataset structure in both real and expression-based UMAP space indicated that the ground truth labels correspond exceptionally closely to well-defined transcriptional clusters. On the other hand, in real space, the annotated domains of this dataset are thin, arranged in concentric, laminar layers that are frequently only one spot wide. These two complementary observations lead us to the hypothesis that some spatially-aware methods may be prone to over-smoothing of transcriptional differences for the sake of contiguous, “blobby” domain annotations. In particular, those methods may exhibit a performance difference on transcriptionally identically defined domains, depending on the size of the domain in question. This hypothesis was also further supported by the observations of method agreement on real data, where we encountered a distinctly lower spot-level agreement in transcriptionally similar domains when they contained smaller numbers of cells or spots.

We investigated the effect of domain size in two different scenarios, consisting of layered structures with shifting widths and circular domains of varying diameter. In both cases, many methods were affected by the size of the domains in their detection. This domain size effect notably played a large role at small sizes.

Evaluating a possible influence of domain shape on method performances, we found that the majority of methods showed a slight bias towards layered structures. Connecting this to the real data, the overall well-performing methods SpatialPCA, SpaceFlow, SCAN-IT, SpaDo, and GraphST all exhibited better performances on the MERFISH-Zhang dataset, which consists of laminar brain layers, than on MERFISH-Moffitt, with a more complex shape.

5.3 Analysis of method stability and secondary evaluation criteria

Besides investigating the effect of technological and tissue-level data characteristics on method performance, we also aimed to complete our comprehensive analysis by secondary evaluation criteria. After a first analysis of method stability and robustness to perturbation, we further benchmarked methods on their runtime, memory usage and scalability. Lastly, method usability was briefly explored.

5.3.1 Stability analysis

The stability of methods was evaluated over multiple independent runs on the same data, and additionally, robustness with respect to the loss of local spatial coherence was investigated. To quantify the stochastic method stability, we developed an approach utilising input reordering to circumvent fixed random states implemented by some methods. We found substantial instability for some methods, as measured by the spread of ARI achieved on reruns of the same, reordered data. One single method, CCST, did not exhibit any variation in its performance.

The investigation of loss of local spatial coherence is motivated by an interest in synthetic data generation. Many attempts to simulate spatial transcriptomics data encompassing a ground truth spatial domain annotation have been published, primarily within the context of method development and for within-method benchmarking purposes. With the exception of certain published simulation software tools (SRTsim, scDesign3), which do not necessarily generate a spatial domain annotation, most published strategies randomly assign cells to spatial locations.

We simulated the effect of this random count allocation in two real datasets, Visium–Maynard and MERFISH–Zhang, by randomly reshuffling the gene expression among spots per domain. Our analysis showed that random count assignment potentially creates a considerable bias in method evaluation based on synthetic data, as methods were strongly and differentially positively affected by the loss of local spatial coherence. Interestingly, except for BASS, all methods which improved more strongly on Visium–Maynard are those which were also shown to perform significantly better on that dataset *a priori* than on MERFISH–Zhang. On the other hand, methods like BANKSY, TACCO, UTAG, and GraphPCA, which showed stronger improvement on MERFISH–Zhang, also generally performed significantly better on that dataset than on Visium–Maynard in the unperturbed states. Thus, the performance improvement attained by methods upon loss of local coherence seems to exhibit a measure of correlation with their baseline method performance.

5.3.2 Runtime, memory usage, and usability investigation

We further evaluated the runtimes and memory usages of all methods, both on the public real datasets and on a dataset of *in silico* samples generated using SRTsim [214]. We found considerable differences between methods in both quantities.

Analysing real data results, a possible bias becomes apparent that is inherent in measuring these secondary evaluation quantities naïvely. Specifically, we are not distinguishing the runtime and memory usage of the method itself from the resources and time needed to simply load the involved matrices into memory, and potentially preprocess the data. We have implemented the same data loading procedures for all R and Python-based methods, respectively, but it is not possible in our setup to avoid a biased evaluation between the two programming languages. Additionally complicating this disambiguation, methods differ in whether they incorporate steps for data preprocessing within their framework, or assume preprocessed data as an input.

Next, we performed a scalability investigation of runtime and memory usage across methods. Overall, the runtimes and memory usages exhibited on the larger synthetic data containing over 10'000 cells were comparable to those on the real datasets. However, the quantities we found on lower cell numbers were lower than those measured on real datasets of similar sizes. This could – beside the

possible effect of the different computing architecture – indicate an effect of the number of genes, or of the lower complexity of the synthetic data.

We found that due to probable memory or runtime constraints on the side of method users, the majority of methods are not suited for the analysis of very large datasets. This is a considerable issue and will likely hamper the adoption of tools in the future, as the trends are towards profiling larger tissue sizes (e.g. StereoSeq) and higher numbers of cells. In the extraordinary accompanying material[1] to their 2022 publication, Moses and Pachter already count a number of published studies profiling over 100'000 cells, profiled mostly using MERFISH and Xenium [54]. The earliest such study listed is from 2018, and in the years since 2023, many more have been published. Spatial domain identification methods which fail to produce a result on sample sizes exceeding even 10'000 cells are unlikely to find wide applicability in this context. Similarly, methods which scale unfavourably in runtime or memory usage may not be feasibly applied to the large datasets which are already being generated.

Lastly, methods were evaluated by usability criteria. This is an indispensable part of thorough method benchmarking, as the adoption of methods by the user group is heavily influenced by ease of use. We utilised a usability checklist inspired by Duo *et al.*, through which we graded methods on criteria of availability, maintenance and documentation. While most tools scored decently on availability and maintenance, a majority of methods did not provide adequate documentation beyond tutorials showcasing specific applications.

5.4 Future directions and outlook

In the research described within this thesis, we have provided a comprehensive overview and benchmarking of the state of the art in spatial domain identification. As is also generally the crux of descriptions of the state of the art, benchmarking projects describing current methods are rapidly out of date [152]. We have attempted to circumvent some of this effect by focusing on not merely describing method performances, but instead aiming to disentangle and thereby explain possible influencing factors. However, since the cutoff date for method inclusion in our benchmark, a multitude of methods have been developed which reportedly outperform existing approaches. In the introduction to the present work, I have outlined reasons to be wary of highly confident performance claims – and thus, further benchmarking efforts are needed to evaluate the existence and extent of performance improvements [143]. However, the creation of evaluation pipelines is time and resource-intensive, and there is an undercurrent of reinventing the wheel with every novel benchmarking effort.

To avoid this recurring trap, one direction to take with this present work is the continued development of an extendable benchmarking framework. In fact, the structure of our pipeline already lends itself to extension, both through the inclusion of new datasets and through novel method implementations. Methods are easily added to the Snakemake workflow through the generation of two files: one describing the prerequisite computing environment in YAML format, and the other containing the script for running the method non-interactively. We already provide a detailed guide to the expected input and output formats, Snakemake rule definition, and the changes in configuration files necessary to fully implement a new method. In this way, we hope to form a resource for method developers to easily compare their new implementations to existing, high-performing approaches. Additionally, this will aid researchers working on generating new datasets through presenting the ability to evaluate a host of pre-implemented methods on their data and find the best fit.

Finding the best fit of methods and parameters for a novel dataset also connects to a different future interest, namely, in creating a coherent and widely applicable definition of spatial domains. Reviewing different avenues for ground truth generation in terms of the underlying, implicit domain

[1] The accompanying material, taking the shape of an evolving online resource in addition to the originally published book, is accessible at pachterlab.github.io/LP_2021.

definitions would form a meaningful starting point for this broad endeavour. It could also be highly informative to evaluate tissue structures characterised by low method agreement with the ground truth annotations, which could help to identify friction points in the current working domain definitions. Generally, there needs to be an increased dialogue between experimental groups that generate data and others focusing more strongly on data analysis. A more direct interchange of ideas between wet-lab biologists and computational method developers would avoid tools being produced for the sake of method development and instead aim collective efforts at solving concrete, existing problems. Ideally, those problems can be formulated as part of one coherent and fixed spatial domain definition. In the more realistic case, namely the vastly different applications and different tissue types leading to disparate formulations, this investigation will still have brought a measure of clarity of purpose to the field.

Aided by well-defined concepts and formulations of spatial domains, the development of well-suited evaluation metrics for the specific task of domain identification could be within reach. Approaches for unsupervised spatial domain evaluation that are specific to this field have not yet been developed. It has been suggested to take inspiration from the geographical sciences in this endeavour, as these already present a tradition of regionalisation methods and their evaluation [220, 221].

Concerning more concrete future paths, the further investigation and development of our consensus approach could be of interest. The unbiased consensus over all methods included in our evaluation is stable and highly competitive, but could potentially be improved by the integration of more sophisticated consensus strategies, such as the Monti consensus clustering algorithm [222]. Detangling the positive performance effects of a consensus approach integrating individually worse-performing methods could yield valuable insights. On a different note, it would be interesting to implement a combination of the consensus approach with our input-reordering stability evaluation strategy. Our investigation highlighted the considerable instability of a large group of overall well-performing methods. Through combining the ability to, through input reordering, repeatedly run a method on the same data, with the subsequent integration of method outputs through taking the consensus, method stability, and potentially also performance, could be enhanced.

Further, leaning on the accumulated knowledge from this comprehensive overview and in-depth evaluation of the field, we could undertake the development of our own stand-alone method for spatial domain identification. This would enable us to directly bring in our expertise to the field, creating an approach to apply insights gleaned from the benchmarking process. One interesting avenue entails the development of a not purely data-driven method, instead integrating in a level of prior knowledge about the dataset in question. This would directly involve expertise in both method development and the question of biological applicability. In the majority of applications, users are experts in the datasets they are analysing, and might already have a measure of knowledge or intuition about the structures they are aiming to identify. One example of such an application could be the molecularly-informed identification of glomeruli in renal tissue [223]. This type of analysis might benefit from the incorporation of priors, for example, to encode informative transcriptional markers or domain shapes of interest. On the other hand, a fully exploratory analysis strategy might be chosen by researchers studying tumour tissue of previously unknown structure. For such an evaluation, it might be useful and informative to focus on a method enabling comprehensive, multi-level analysis through tunable, interpretable hyperparameters. In any future method development approach, it is imperative to especially consider robustness to cellular heterogeneity as a hallmark of good performance on high-resolution spatial transcriptomics data. As technological advances point in the direction of higher resolutions, this tool characteristic is likely to increase in importance.

Concerning the systematic method evaluation undertaken in this work using semi-synthetic data, there are a few avenues for potential further development. Notably, the investigation of the effect of changing the number of profiled genes could be further enhanced in realism. Instead of downsampling to randomly selected gene subsets, we could select for highly variable or spatially variable genes, or

alternatively, utilise an algorithm for marker gene detection. This might be more representative of real data insofar as gene panels in FISH-based, targeted approaches are also selected for biological informativeness. Further, it would be interesting to disentangle the effect of changing resolution on method performance from the effect of domain size, as necessarily the binning of multiple cells into one spot diminishes the number of individual points making up a domain.

There are also numerous avenues for further development considering the generation of semi-synthetic data itself. One starting point could be incorporating a general cell type similarity measure to enable direct quantification of the improvement in distinction attained by utilising spatial information. On the other hand, creating semi-synthetic spatial transcriptomics data that closely mimics different concrete tissue types and structures, but encompasses ground truth domain annotations, could enable both method developers and users to directly tune tools to the characteristics of a tissue of interest. Further, as identified by our investigation of the effect of local spatial coherence loss, the random allocation of cells to spatial locations utilised in many approaches to synthetic data generation places a strong caveat on the interpretability of simulation-based claims concerning method performance. Existing simulators have circumvented this problem by assigning gene expression based on predefined or learned spatial gene expression patterns [214, 215]. However, these published approaches to creating semi-synthetic data often do not lend themselves to the concurrent generation of a spatial domain ground truth. Integrating these different viewpoints, namely also under the consideration of a solid domain definition, could enhance the potential of semi-synthetic data in spatial domain evaluation.

Lastly, we recognise that the usability of methods was not the focus of this thesis nor of the entire benchmarking project. However, considering the central part it plays in the wider and continued use of computational tools, a more in-depth investigation of the state of usability in the field of spatial domain identification would be appropriate. Broad guidelines for increasing usability in bioinformatics software have been developed, for example, by List *et al.* in 2017 [224] and by Mangul *et al.* in 2019 [225]. Adapting these guidelines to the specific challenges in this field would provide a framework for researchers to consider when developing novel tools, in order to maximise the size of their potential user base.

5.5 Conclusions

In this thesis, I have presented a comprehensive view of the context, state of the art and future directions for domain identification in spatial transcriptomics. I report a detailed and systematic benchmarking of published tools, comparing performances on real datasets with ground truth annotations and additionally directly investigating the effect of a host of data characteristics on method performances using semi-synthetic data. Additionally, I have pointed out possible future research avenues within this field, most importantly the necessity of clearly defining the concept of spatial domains, as well as entering a more immediate dialogue with prospective method users, as a prerequisite for goal-oriented and efficient method development.

Independent benchmarkings of computational method performances are invaluable in the current scientific climate, which is marked by overoptimistic self-reporting due to the well-documented phenomenon of publication bias [143, 226–228]. Some guidelines to ameliorate this phenomenon from the perspective of method developers can be found in an excellent 2015 editorial by Boulesteix [229]. However, in a scientific publication context that continues to incentivise unprecedented methodology and reward outstanding reported performances, placing the responsibility for upholding rigorous standards for self-evaluation on authors alone would speak of some naiveté. Independent, post-publication evaluation of tools in benchmarking studies guides future research directions and provides resources for method users overwhelmed by choices. This type of analytical methodological research must find a suitable place alongside the development of novel analyses and approaches.

Appendix A

General overview of tools for spatial domain identification

With the aim of attaining a comprehensive view of the space of spatial domain identification methods, I extracted pertinent information from 33 tool publications. This information is summarised over the following pages in Tab. A.1.

Name	Domain type	Datasets Used	# Co.	Metrics Used
ADEPT [128]	expression coherence	STARmap-Wang-mouse-visual-cortex, Visium-Maynard-human-DLPFC, Visium-demo-human-breast-cancer	5	ARI, qualitative
BANKSY [83]	cell type composition	CosMx-He-human-colon, MERFISH-Moffitt-mouse-hypothalamic-preoptic-region, Merscope-demo-human-colon-tumor, STARmap-Wang-mouse-visual-cortex, Slide-seqV2-Stickels-mouse-hippocampus, Slide-seq-Rodriques-mouse-cerebellum-hippocampus-olfactory-bulb, Visium-Maynard-human-DLPFC	7	ARI, qualitative
BASS [82]	cell type composition	MERFISH-Moffitt-mouse-hypothalamic-preoptic-region, STARmap-Wang-mouse-visual-cortex, Visium-Maynard-human-DLPFC	3	ARI, qualitative
BayeSMART [120]	no definition	STARmap-Wang-mouse-visual-cortex, ST-Andersson-human-breast-cancer, Visium-Maynard-human-DLPFC	9	ARI, AUC, F1score, domain-specific SVGs
BayesSpace [119]	augmented clustering	ST-Thrane-human-melanoma, Visium-Maynard-human-DLPFC, Visium-Zhao-human-breast-cancer, Visium-demo-human-ovarian-cancer	7	ARI, qualitative
CCST [109]	augmented clustering	MERFISH-Xia-human-sarcoma-cell-line, Visium-Maynard-human-DLPFC, Visium-demo-human-breast-cancer, seqFISH+-Eng-mouse-somatosensory-cortex	7	ARI, FMI, LISI, NMI, qualitative
CellCharter [135]	cell type composition	CosMx-He-human-carcinoma, Merscope-demo-human-lung-cancer, Visium-Maynard-human-DLPFC	5	ARI, FMI, qualitative
CytoCommunity [168]	cell type composition	MERFISH-Moffitt-mouse-hypothalamic-preoptic-region	5	AMI, F1score
DeepST [190]	expression coherence	MERFISH-Moffitt-mouse-hypothalamic-preoptic-region, Slide-seqV2-Stickels-mouse-hippocampus, StereoSeq-Chen-mouse-olfactory-bulb, Visium-Maynard-human-DLPFC, Visium-demo-human-breast-cancer, Visium-demo-mouse-posterior-brain	6	ARI, DB, Silhouette, qualitative

Table A.1: Continued on next page.

Name	Domain type	Datasets Used	# Co.	Metrics Used
GraphPCA [137]	no definition	STARmap-Wang-mouse-visual-cortex, Visium-Guilliams-mouse-liver, Visium-Maynard-human-DLPFC	7	AMI, ARI, HOM, qualitative
GraphST [115]	expression coherence	Slide-seqV2-Stickels-mouse-hippocampus, StereoSeq-Chen-mouse-olfactory-bulb, Visium-Maynard-human-DLPFC	7	ARI, qualitative
GRAS4T [230]	expression coherence	MERFISH-Moffitt-mouse-hypothalamic-preoptic-region, STARmap-Wang-mouse-visual-cortex, ST-Andersson-human-breast-cancer, StereoSeq-Chen-mouse-olfactory-bulb, Visium-Maynard-human-DLPFC, Visium-demo-mouse-posterior-brain	5	ARI, NMI, qualitative
HMRF/Giotto [87, 118]	expression coherence	seqFISH-Shah-mouse-visual-cortex	0	qualitative
IRIS [169]	cell type composition	ST-Ji-human-squamous-cell-carcinoma, Slide-seq-Chen-mouse-spermatogenesis, StereoSeq-Chen-mouse-olfactory-bulb, Visium-Maynard-human-DLPFC, Xenium-Janesick-human-breast-cancer	12	ARI, CHAOS, qualitative
MERINGUE [86]	expression coherence	ISH-Fowlkes-drosophila-embryo, MERFISH-Moffitt-mouse-hypothalamic-preoptic-region, ST-Stahl-mouse-coronal-sagittal-posterior-brain, Slide-seq-Rodriques-mouse-cerebellum-hippocampus-olfactory-bulb	0	qualitative
MNMST [186]	expression coherence	STARmap-Wang-mouse-visual-cortex, Slide-seqV2-Stickels-mouse-hippocampus, StereoSeq-Chen-mouse-olfactory-bulb, Visium-Maynard-human-DLPFC, Visium-demo-human-breast-cancer, Visium-demo-mouse-posterior-brain, osmFISH-Codeluppi-mouse-somatosensory-cortex	7	ARI, qualitative
PAST [129]	functional distinction	STARmap-Wang-mouse-visual-cortex, StereoSeq-Chen-mouse-olfactory-bulb, Visium-Maynard-human-DLPFC, Visium-demo-human-breast-cancer, osmFISH-Codeluppi-mouse-somatosensory-cortex	9	AMI, ARI, COM, FMI, HOM, NMI, qualitative

Table A.1: Continued on next page.

Name	Domain type	Datasets Used	# Co.	Metrics Used
PRECAST [164]	augmented clustering	ST-Hildebrandt-mouse-liver, Slide-seqV2-Stickels-mouse-hippocampus, Visium-Liu-human-liver-tumor, Visium-Maynard-human-DLPFC	0	ARI, NMI, qualitative
SC-MEB [116]	augmented clustering	MERFISH-Moffitt-mouse-hypothalamic-preoptic-region, Visium-Maynard-human-DLPFC, Visium-Yang-human-colon-COVID-CRC	4	ARI, qualitative
SCAN-IT [127]	cell type composition	ST-Stahl-mouse-coronal-sagittal-posterior-brain, Slide-seq-Rodriques-mouse-cerebellum-hippocampus-olfactory-bulb, Visium-Maynard-human-DLPFC, osmFISH-Codeluppi-mouse-somatosensory-cortex, seqFISH+-Eng-mouse-somatosensory-cortex, seqFISH-Zhu-mouse-visual-cortex	4	AMI, ARI, DE marker genes, FMI, NMI, comparison to reference tissue, qualitative
SEDR [170]	no definition	StereoSeq-Chen-mouse-olfactory-bulb, Visium-Maynard-human-DLPFC, Visium-demo-human-breast-cancer	10	AMI, ARI, COM, DE marker genes, HOM, V-measure, purity, qualitative
SOTIP [133]	cell type composition	Visium-Maynard-human-DLPFC, osmFISH-Codeluppi-mouse-somatosensory-cortex, seqFISH+-Eng-mouse-somatosensory-cortex	7	ARI, qualitative
SpaceFlow [188]	augmented clustering	ST-Andersson-human-breast-cancer, Visium-Maynard-human-DLPFC	6	ARI, qualitative
SpaDo [134]	cell type composition	STARmap-Wang-mouse-visual-cortex, Visium-Maynard-human-DLPFC, Visium-Meylan-human-renal-cell-cancer, osmFISH-Codeluppi-mouse-somatosensory-cortex, seqFISH+-Eng-mouse-somatosensory-cortex	6	ARI, qualitative
SpaGCN [122]	expression coherence	STARmap-Wang-mouse-visual-cortex, ST-Moncada-human-pancreatic-cancer, Visium-Maynard-human-DLPFC, Visium-demo-mouse-posterior-brain	3	ARI, domain-specific SVGs, qualitative

Table A.1: Continued on next page.

Name	Domain type	Datasets Used	# Co.	Metrics Used
SpatialMGCN [130]	expression coherence	StereoSeq-Chen-mouse-olfactory-bulb, Visium-Maynard-human-DLPFC, Visium-demo-human-breast-cancer	7	ARI, DE marker genes, qualitative
SpatialPCA [136]	cell type composition	Slide-seqV2-Stickels-mouse-hippocampus, Slide-seq-Rodriques-mouse-cerebellum-hippocampus-olfactory-bulb, Visium-Maynard-human-DLPFC, Visium-demo-human-breast-cancer	5	ARI, CHAOS, LISI, PAS, qualitative
SpatialPrompt [167]	no definition	MERFISH-Yao-mouse-brain, STARmap-Wang-mouse-visual-cortex, Slide-seq-Rodriques-mouse-cerebellum-hippocampus-olfactory-bulb, Visium-Maynard-human-DLPFC, Visium-demo-mouse-kidney, Visium-demo-mouse-posterior-brain	5	AUC, NMI, qualitative
SpiceMix [192]	expression coherence	STARmap-Wang-mouse-visual-cortex, Visium-Maynard-human-DLPFC, seqFISH+-Eng-mouse-somatosensory-cortex	2	ARI, qualitative
STAGATE [84]	expression coherence	IMC-Rendeiro-human-lung-infection, Slide-seqV2-Stickels-mouse-hippocampus, StereoSeq-Chen-mouse-olfactory-bulb, Visium-Maynard-human-DLPFC, Visium-demo-mouse-posterior-brain	6	ARI, HOM, NMI, qualitative
TACCO [76]	expression coherence	osmFISH-Codeluppi-mouse-somatosensory-cortex	0	qualitative
UTAG [184]	functional distinction	CyCIF-Rashid-human-lung-cancer, IMC-Damond-human-pancreas, IMC-Lehmann-human-COVID-intestine, IMC-Ohara-human-urothelial-carcinoma, IMC-Rustam-human-lung	4	HOM, RI, qualitative
Vesalius [140]	no definition	Seq-Scope-Cho-mouse-liver-and-colon, Slide-seqV2-Stickels-mouse-hippocampus, Visium-demo-human-breast-cancer, Visium-demo-mouse-kidney, Visium-demo-mouse-posterior-brain, seqFISH-Lohoff-mouse-embryo	6	ARI, VI, qualitative

Table A.1: **Overview of 33 (total) method publications.** Information about domain definitions, datasets used for benchmarking, number of methods compared to, and the metrics used for those comparisons.

Appendix B

Ground truth domain assignments for the included real data samples

In this appendix, all ground truth domain assignments for benchmarked samples are shown. The origin of both the data and the corresponding ground truth labels is detailed in Tab. 2.2.

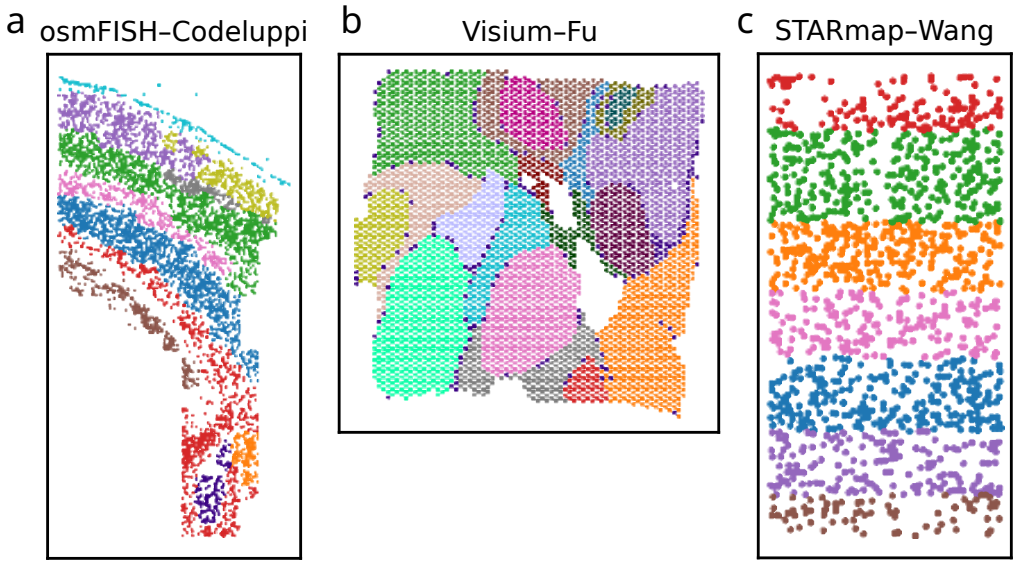

Figure B.1: **Ground truth domain assignments of single-sample datasets.** a, osmFISH–Codeluppi. b, Visium–Fu. c, STARmap–Wang.

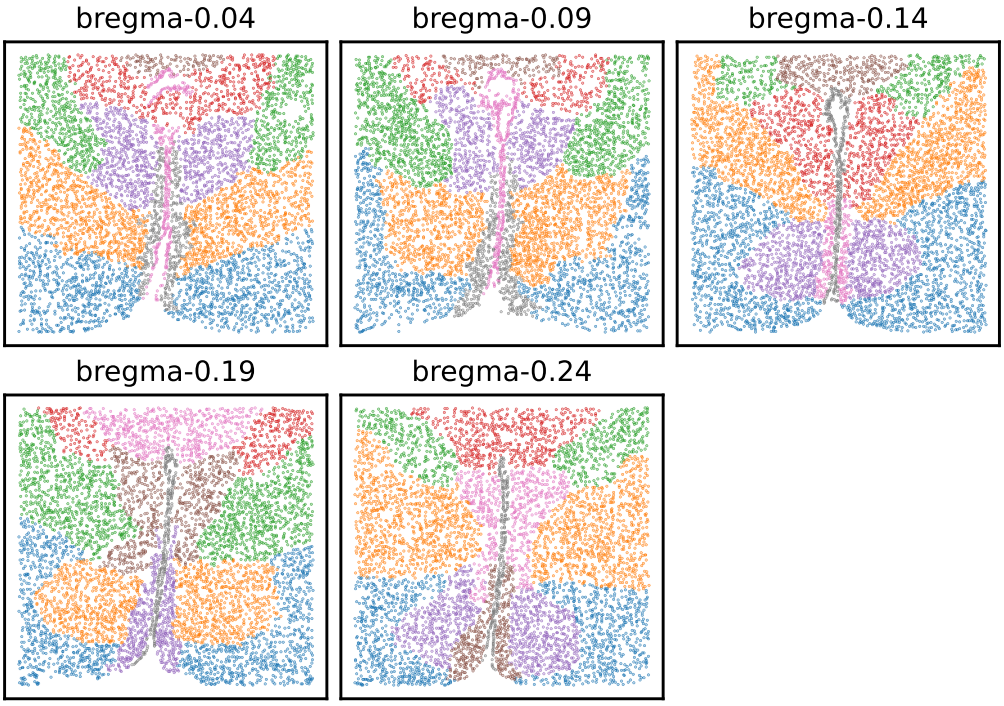

Figure B.2: **Ground truth domain assignments of MERFISH–Moffitt.**

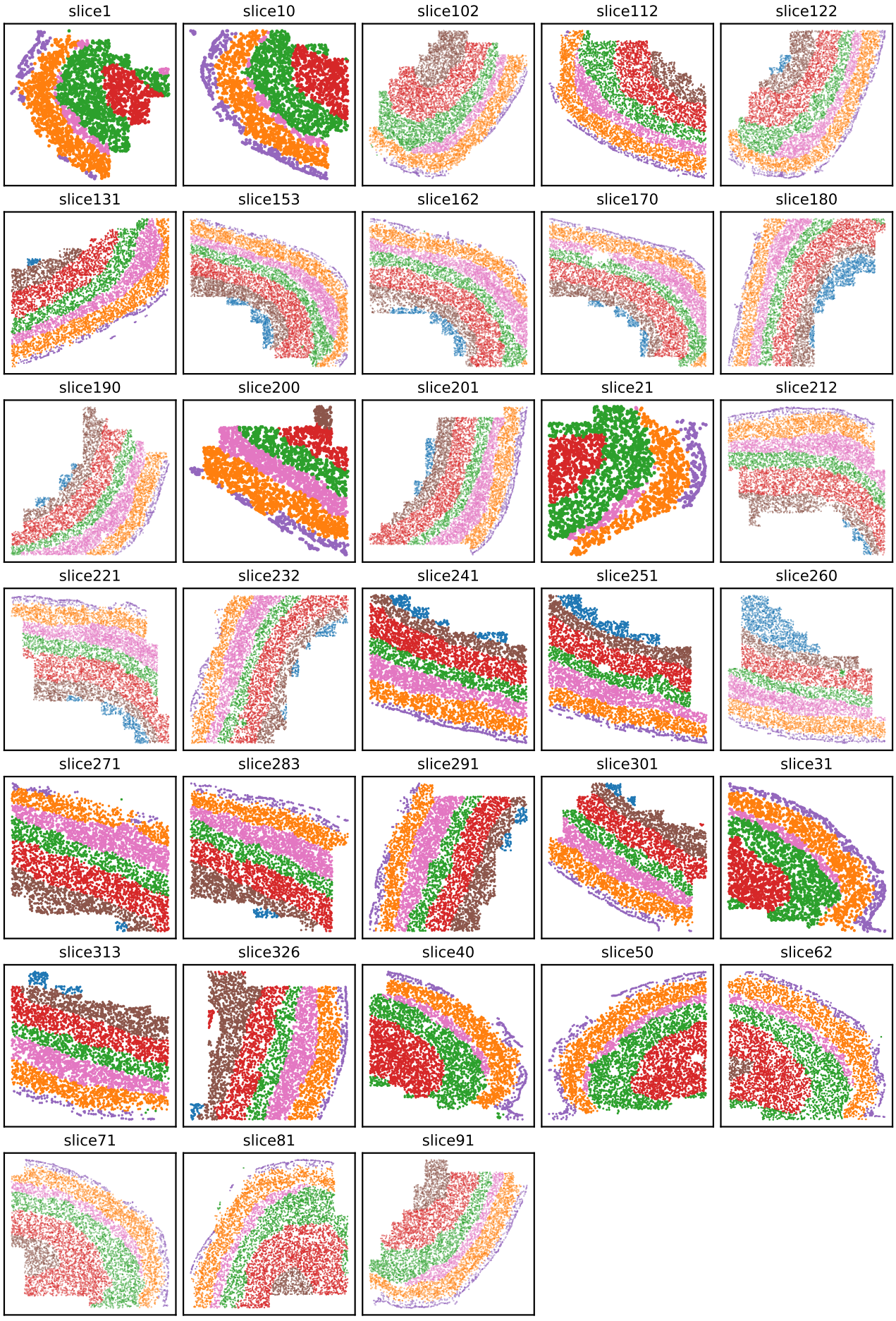

Figure B.3: **Ground truth domain assignments of MERFISH–Zhang.**

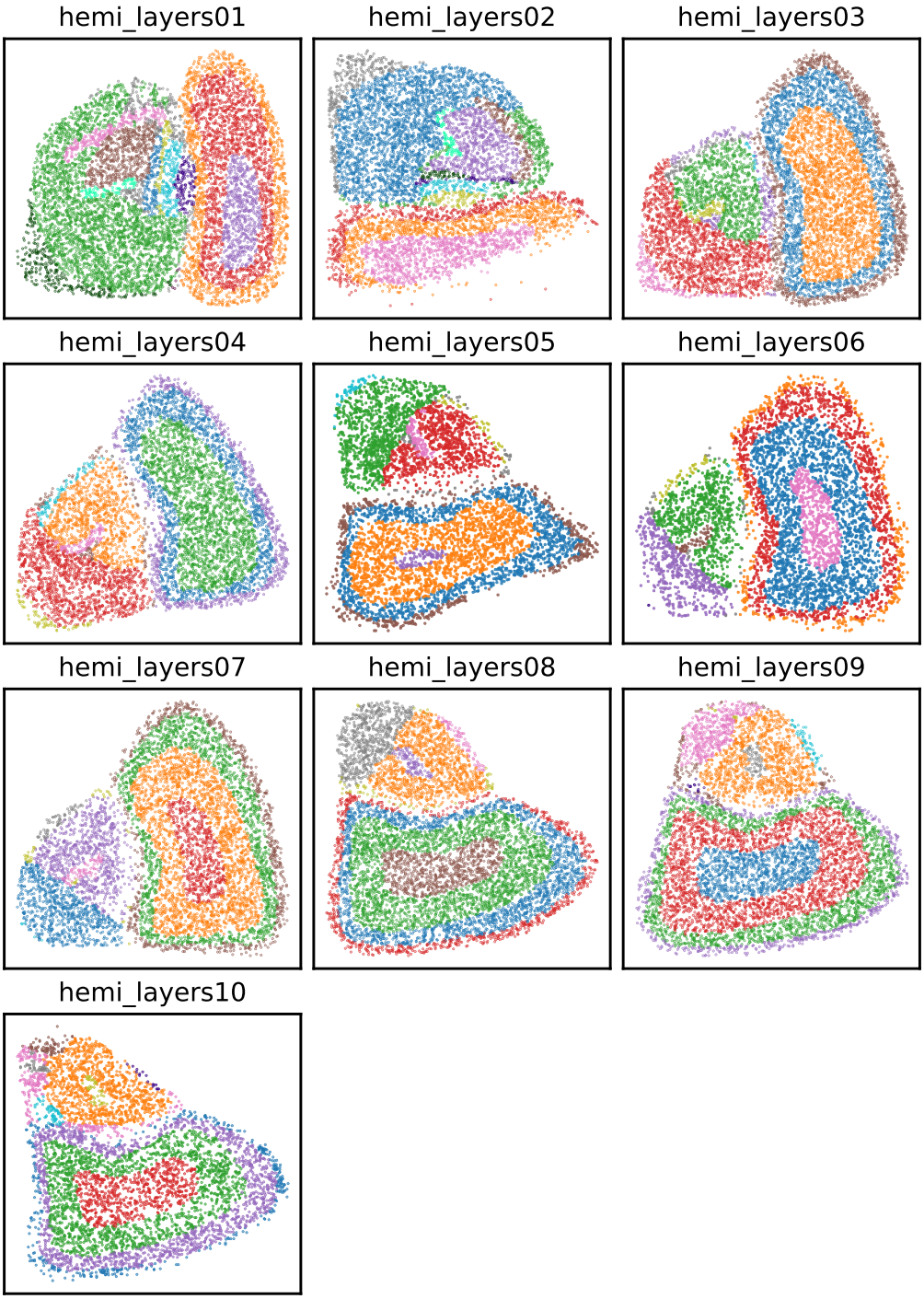

Figure B.4: **Ground truth domain assignments of Slide-seq–Langlieb.**

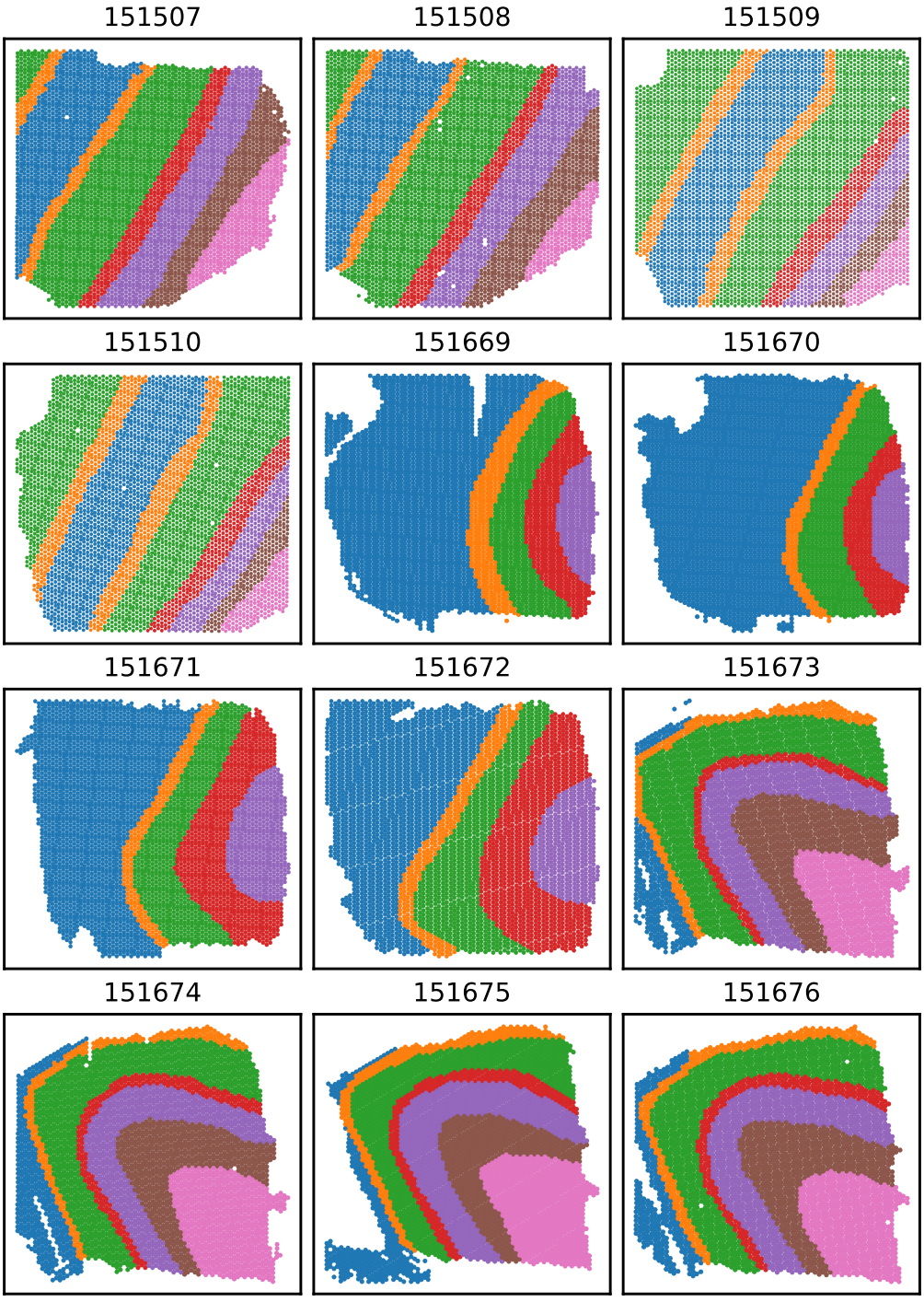

Figure B.5: **Ground truth domain assignments of Visium–Maynard.**

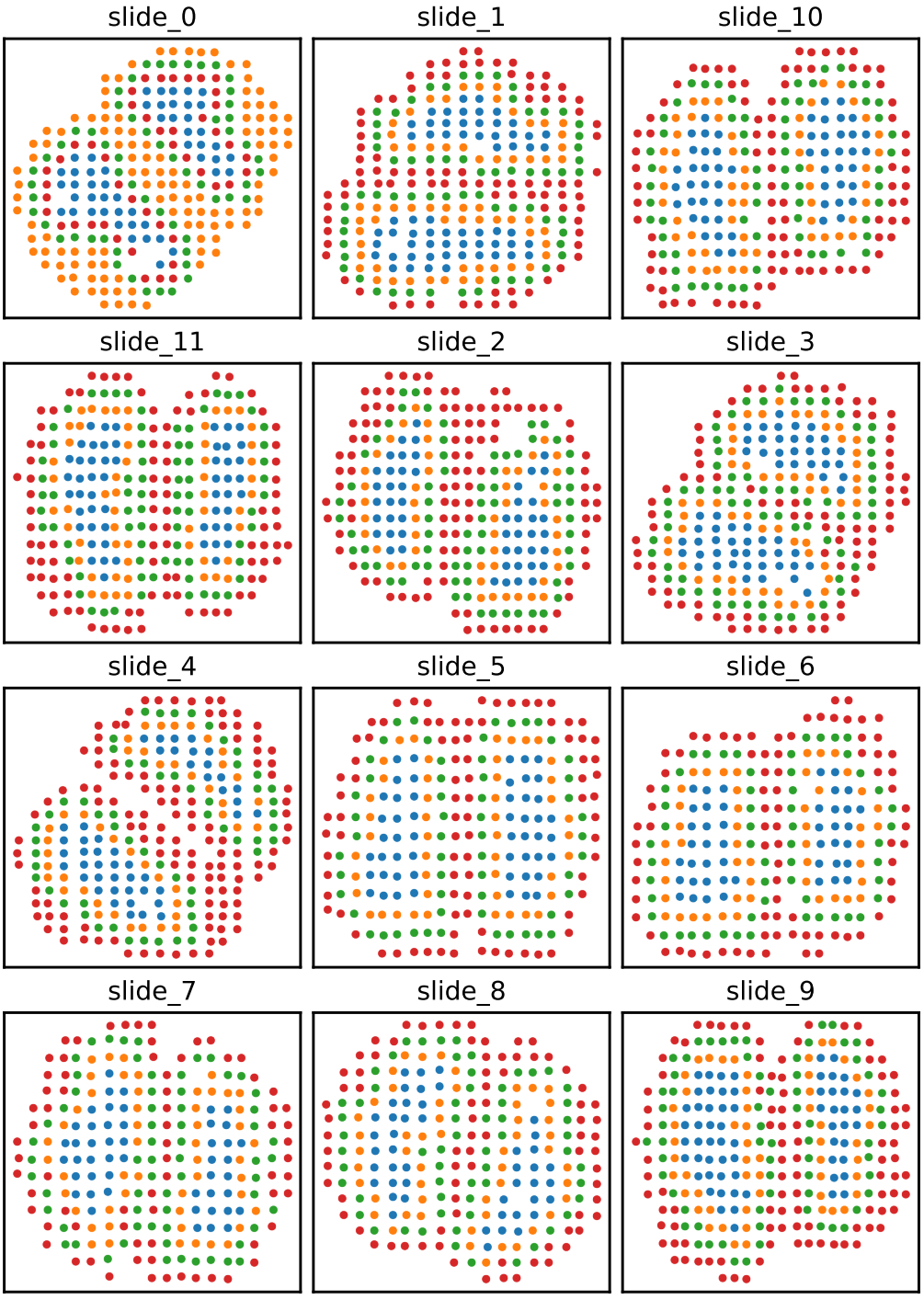

Figure B.6: **Ground truth domain assignments of ST–Stahl.**

Acknowledgements

Thank you to my supervisor, Stefan Canzar, for all the steady support and feedback you offered me over the years. Throughout the two moves, you kept the lab going, and we never even missed more than a couple of weekly meetings!

My collaborators on the benchmarking project have taught me a lot, both scientifically and personally. I specifically want to extend my gratitude to Tomislav Prusina for all of your help. Additionally, I thank Hoan Van Do, Francisca Rojas Ringeling, and Domagoj Matijević for their input and contributions.

To Pablo and Shuang – thank you for your advice, especially early on, and for the long chats at the end of the day. To you, and to all the other current and past members of our group across Germany, Croatia, and the US, I wish nothing but the best.

A particular thank you goes to Johanna Klughammer and Simon Mages, for not only allowing me to socially integrate with your group, but also continuing to support me scientifically and offering valuable insights on my work.

I am incredibly grateful to all the current and past members of the Klughammer and Mages labs for your friendship, reassurance, and for offering your scientific expertise. Especially to Antonia, Mohammad, and Jan, along with the extended, ever-changing lunchtime and kicker group – you truly have been invaluable in helping me to keep going. I don't know what I would have done without you.

Thank you to Augusto for all of your help and encouragement, and for making me want to try to be a better scientist. This thesis would not be what it is without your feedback.

Finally and importantly, I want to thank my family and friends back home for all of your incredible patience and support.

I was supported in my PhD by the Graduate School of Quantitative and Molecular Biosciences Munich (QMB, formerly QBM). This work was supported by the Collaborative Research Center / Transregio (CRC TRR) 338 (LETSIMMUN - Lymphocyte Engineering for Therapeutic Synthetic Immunity), funded by the Deutsche Forschungsgemeinschaft (DFG).

Bibliography

[1] Matthew Cobb. "60 Years Ago, Francis Crick Changed the Logic of Biology". In: *PLOS Biology* 15.9 (Sept. 18, 2017), e2003243. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.2003243.

[2] Alexander F. Palazzo and Eliza S. Lee. "Non-Coding RNA: What Is Functional and What Is Junk?" In: *Frontiers in Genetics* 6 (Jan. 26, 2015). ISSN: 1664-8021. DOI: 10.3389/fgene.2015.00002.

[3] Jay R. Hesselberth. "Lives That Introns Lead after Splicing". In: *WIREs RNA* 4.6 (2013), pp. 677–691. ISSN: 1757-7012. DOI: 10.1002/wrna.1187.

[4] Laura Restrepo-Pérez, Chirlmin Joo, and Cees Dekker. "Paving the Way to Single-Molecule Protein Sequencing". In: *Nature Nanotechnology* 13.9 (Sept. 2018), pp. 786–796. ISSN: 1748-3395. DOI: 10.1038/s41565-018-0236-6.

[5] "Method of the Year 2024: Spatial Proteomics". In: *Nature Methods* 21.12 (Dec. 2024), pp. 2195–2196. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02565-3.

[6] Ruedi Aebersold and Matthias Mann. "Mass-Spectrometric Exploration of Proteome Structure and Function". In: *Nature* 537.7620 (Sept. 2016), pp. 347–355. ISSN: 1476-4687. DOI: 10.1038/nature19949.

[7] Hanno Steen and Matthias Mann. "The Abc's (and Xyz's) of Peptide Sequencing". In: *Nature Reviews Molecular Cell Biology* 5.9 (Sept. 2004), pp. 699–711. ISSN: 1471-0080. DOI: 10.1038/nrm1468.

[8] William C.S. Cho. "Proteomics Technologies and Challenges". In: *Genomics, Proteomics & Bioinformatics* 5.2 (June 1, 2007), pp. 77–85. ISSN: 1672-0229. DOI: 10.1016/S1672-0229(07)60018-7.

[9] Philip C. Bevilacqua, Laura E. Ritchey, Zhao Su, and Sarah M. Assmann. "Genome-Wide Analysis of RNA Secondary Structure". In: *Annual Review of Genetics* 50 (Volume 50, 2016 Nov. 23, 2016), pp. 235–266. ISSN: 0066-4197, 1545-2948. DOI: 10.1146/annurev-genet-120215-035034.

[10] Mary Lou Pardue and Joseph G. Gall. "Molecular Hybridization of Radioactive Dna to the Dna of Cytological Preparations". In: *Proceedings of the National Academy of Sciences* 64.2 (Oct. 1969), pp. 600–604. DOI: 10.1073/pnas.64.2.600.

[11] H. A. John, M. L. Birnstiel, and K. W. Jones. "RNA-DNA Hybrids at the Cytological Level". In: *Nature* 223.5206 (Aug. 1969), pp. 582–587. ISSN: 1476-4687. DOI: 10.1038/223582a0.

[12] M. Buongiorno-Nardelli and F. Amaldi. "Autoradiographic Detection of Molecular Hybrids between rRNA and DNA in Tissue Sections". In: *Nature* 225.5236 (Mar. 1970), pp. 946–948. ISSN: 1476-4687. DOI: 10.1038/225946a0.

[13] J. G. J. Bauman, J. Wiegant, P. Borst, and P. van Duijn. "A New Method for Fluorescence Microscopical Localization of Specific DNA Sequences by in Situ Hybridization of Fluorochrome-Labelled RNA". In: *Experimental Cell Research* 128.2 (Aug. 1, 1980), pp. 485–490. ISSN: 0014-4827. DOI: 10.1016/0014-4827(80)90087-7.

[14] P R Langer-Safer, M Levine, and D C Ward. "Immunological Method for Mapping Genes on Drosophila Polytene Chromosomes." In: *Proceedings of the National Academy of Sciences* 79.14 (July 1982), pp. 4381–4385. DOI: 10.1073/pnas.79.14.4381.

[15] Joseph G. Gall. "The Origin of In Situ Hybridization - a Personal History". In: *Methods (San Diego, Calif.)* 98 (Apr. 1, 2016), pp. 4–9. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2015.11.026. PMID: 26655524.

[16] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray". In: *Science* 270.5235 (Oct. 20, 1995), pp. 467–470. DOI: 10.1126/science.270.5235.467.

[17] Roger Bumgarner. "Overview of DNA Microarrays: Types, Applications, and Their Future". In: *Current Protocols in Molecular Biology* 101.1 (2013), pp. 22.1.1–22.1.11. ISSN: 1934-3647. DOI: 10.1002/0471142727.mb2201s101.

[18] Tyson A. Clark, Charles W. Sugnet, and Manuel Ares. "Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays". In: *Science* 296.5569 (May 3, 2002), pp. 907–910. DOI: 10.1126/science.1069415.

[19] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: A Revolutionary Tool for Transcriptomics". In: *Nature Reviews Genetics* 10.1 (Jan. 2009), pp. 57–63. ISSN: 1471-0064. DOI: 10.1038/nrg2484.

[20] Valerio Costa, Claudia Angelini, Italia De Feis, and Alfredo Ciccodicola. "Uncovering the Complexity of Transcriptomes with RNA-Seq". In: *BioMed Research International* 2010.1 (2010), p. 853916. ISSN: 2314-6141. DOI: 10.1155/2010/853916.

[21] Walter Gilbert and Allan Maxam. "The Nucleotide Sequence of the Lac Operator". In: *Proceedings of the National Academy of Sciences* 70.12 (Dec. 1973), pp. 3581–3584. DOI: 10.1073/pnas.70.12.3581.

[22] Robert A. Holt and Steven J. M. Jones. "The New Paradigm of Flow Cell Sequencing". In: *Genome Research* 18.6 (Jan. 6, 2008), pp. 839–846. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.073262.107. PMID: 18519653.

[23] F Sanger, S Nicklen, and R Coulson. "DNA Sequencing with Chain-Terminating Inhibitors". In: *PNAS* 74.12 (Dec. 1977), pp. 5463–5467.

[24] James M. Heather and Benjamin Chain. "The Sequence of Sequencers: The History of Sequencing DNA". In: *Genomics* 107.1 (Jan. 1, 2016), pp. 1–8. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2015.11.003.

[25] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding

Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, et al. "The Sequence of the Human Genome". In: *Science* 291.5507 (Feb. 16, 2001), pp. 1304–1351. DOI: 10.1126/science.1058040.

[26] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Nae-Chyun Chen, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G. S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Ying Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, et al. "The Complete Sequence of a Human Genome". In: *Science* 376.6588 (Apr. 2022), pp. 44–53. DOI: 10.1126/science.abj6987.

[27] Mark D. Adams, Jenny M. Kelley, Jeannine D. Gocayne, Mark Dubnick, Mihael H. Polymeropoulos, Hong Xiao, Carl R. Merril, Andrew Wu, Bjorn Olde, Ruben F. Moreno, Anthony R. Kerlavage, W. Richard McCombie, and J. Craig Venter. "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project". In: *Science* 252.5013 (June 21, 1991), pp. 1651–1656. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.2047873.

[28] Vivien Marx. "Method of the Year: Long-Read Sequencing". In: *Nature Methods* 20.1 (Jan. 2023), pp. 6–11. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01730-w.

[29] Malte D. Luecken and Fabian J. Theis. "Current Best Practices in Single-Cell RNA-seq Analysis: A Tutorial". In: *Molecular Systems Biology* 15.6 (June 2019), e8746. ISSN: 1744-4292. DOI: 10.15252/msb.20188746.

[30] Geng Chen, Baitang Ning, and Tieliu Shi. "Single-Cell RNA-Seq Technologies and Related Computational Data Analysis". In: *Frontiers in Genetics* 10 (Apr. 5, 2019), p. 317. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00317.

[31] Kyongho Choe, Unil Pak, Yu Pang, Wanjun Hao, and Xiuqin Yang. "Advances and Challenges in Spatial Transcriptomics for Developmental Biology". In: *Biomolecules* 13.1 (Jan. 2023), p. 156. ISSN: 2218-273X. DOI: 10.3390/biom13010156.

[32] Esther Danenberg, Helen Bardwell, Vito R. T. Zanotelli, Elena Provenzano, Suet-Feung Chin, Oscar M. Rueda, Andrew Green, Emad Rakha, Samuel Aparicio, Ian O. Ellis, Bernd Bodenmiller, Carlos Caldas, and H. Raza Ali. "Breast Tumor Microenvironment Structures Are As-

sociated with Genomic Features and Clinical Outcome". In: *Nature Genetics* 54.5 (May 2022), pp. 660–669. ISSN: 1546-1718. DOI: 10.1038/s41588-022-01041-y.

[33] Anna Fomitcheva-Khartchenko, Aditya Kashyap, Tamar Geiger, and Govind V. Kaigala. "Space in Cancer Biology: Its Role and Implications". In: *Trends in Cancer* 8.12 (Dec. 1, 2022), pp. 1019–1032. ISSN: 2405-8033. DOI: 10.1016/j.trecan.2022.07.008. PMID: 35995681.

[34] Anjali Rao, Dalia Barkley, Gustavo S. França, and Itai Yanai. "Exploring Tissue Architecture Using Spatial Transcriptomics". In: *Nature* 596.7871 (7871 Aug. 2021), pp. 211–220. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03634-9.

[35] Michael R. Emmert-Buck, Robert F. Bonner, Paul D. Smith, Rodrigo F. Chuaqui, Zhengping Zhuang, Seth R. Goldstein, Rhonda A. Weiss, and Lance A. Liotta. "Laser Capture Microdissection". In: *Science* 274.5289 (Nov. 8, 1996), pp. 998–1001. DOI: 10.1126/science.274.5289.998.

[36] Jun Chen, Shengbao Suo, Patrick PL Tam, Jing-Dong J. Han, Guangdun Peng, and Naihe Jing. "Spatial Transcriptomic Analysis of Cryosectioned Tissue Samples with Geo-seq". In: *Nature Protocols* 12.3 (Mar. 2017), pp. 566–580. ISSN: 1750-2799. DOI: 10.1038/nprot.2017.003.

[37] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. "Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics". In: *Science* 353.6294 (July 2016), pp. 78–82. DOI: 10.1126/science.aaf2403.

[38] Michelli Faria de Oliveira, Juan Pablo Romero, Meii Chung, Stephen R. Williams, Andrew D. Gottscho, Anushka Gupta, Susan E. Pilipauskas, Seayar Mohabbat, Nandhini Raman, David J. Sukovich, David M. Patterson, and Sarah E. B. Taylor. "High-Definition Spatial Transcriptomic Profiling of Immune Cell Populations in Colorectal Cancer". In: *Nature Genetics* 57.6 (June 2025), pp. 1512–1523. ISSN: 1546-1718. DOI: 10.1038/s41588-025-02193-3.

[39] Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. "Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution". In: *Science* 363.6434 (Mar. 29, 2019), pp. 1463–1467. DOI: 10.1126/science.aaw1219.

[40] Robert R. Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L. Marshall, Daniela J. Di Bella, Paola Arlotta, Evan Z. Macosko, and Fei Chen. "Highly Sensitive Spatial Transcriptomics at Near-Cellular Resolution with Slide-seqV2". In: *Nature Biotechnology* 39.3 (Mar. 2021), pp. 313–319. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0739-1.

[41] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, Gabriel K. Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L. Ståhl. "High-Definition Spatial Transcriptomics for in Situ Tissue Profiling". In: *Nature Methods* 16.10 (Oct. 2019), pp. 987–990. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0548-y.

[42] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, Xin Wang, Huifang Lu, Xi Chen, Xing Liu, Xin Huang, Zhao Li, Yan Hong, Yujia Jiang, Jian Peng, Shuai Liu, Mengzhe Shen, Chuanyu Liu, Quanshui Li, Yue Yuan, Xiaoyu Wei, Huiwen Zheng, Weimin Feng, Zhifeng Wang, Yang Liu, Zhaohui Wang, Yunzhi Yang, Haitao Xiang, Lei Han, Baoming Qin, Pengcheng Guo, Guangyao Lai, Pura Muñoz-Cánoves, Patrick H. Maxwell, Jean Paul Thiery, Qing-Feng Wu, Fuxiang Zhao, Bichao

Chen, Mei Li, Xi Dai, Shuai Wang, Haoyan Kuang, Junhou Hui, Liqun Wang, Ji-Feng Fei, Ou Wang, Xiaofeng Wei, Haorong Lu, Bo Wang, Shiping Liu, Ying Gu, Ming Ni, Wenwei Zhang, Feng Mu, Ye Yin, Huanming Yang, Michael Lisby, Richard J. Cornall, Jan Mulder, Mathias Uhlén, Miguel A. Esteban, Yuxiang Li, Longqi Liu, Xun Xu, and Jian Wang. "Spatiotemporal Transcriptomic Atlas of Mouse Organogenesis Using DNA Nanoball-Patterned Arrays". In: *Cell* 185.10 (May 12, 2022), 1777–1792.e21. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2022.04.003.

[43] Andrea M. Femino, Fredric S. Fay, Kevin Fogarty, and Robert H. Singer. "Visualization of Single RNA Transcripts in Situ". In: *Science* 280.5363 (Apr. 24, 1998), pp. 585–590. DOI: 10.1126/science.280.5363.585.

[44] Simone Codeluppi, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. "Spatial Organization of the Somatosensory Cortex Revealed by osmFISH". In: *Nature Methods* 15.11 (Nov. 2018), pp. 932–935. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0175-z.

[45] Eric Lubeck, Ahmet F. Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. "Single-Cell in Situ RNA Profiling by Sequential Hybridization". In: *Nature Methods* 11.4 (4 Apr. 2014), pp. 360–361. ISSN: 1548-7105. DOI: 10.1038/nmeth.2892.

[46] Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. "Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells". In: *Science* 348.6233 (Apr. 24, 2015), aaa6090. DOI: 10.1126/science.aaa6090.

[47] Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A. Morrison, Michelli F. Oliveira, Jordan T. Sicherman, Andrew Kohlway, Jawad Abousoud, Tingsheng Yu Drennon, Seayar H. Mohabbat, and Sarah E. B. Taylor. "High Resolution Mapping of the Tumor Microenvironment Using Integrated Single-Cell, Spatial and in Situ Analysis". In: *Nature Communications* 14.1 (Dec. 19, 2023), p. 8353. ISSN: 2041-1723. DOI: 10.1038/s41467-023-43458-x.

[48] Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan, Felice-Alessio Bava, and Karl Deisseroth. "Three-Dimensional Intact-Tissue Sequencing of Single-Cell Transcriptional States". In: *Science* 361.6400 (July 27, 2018), eaat5691. DOI: 10.1126/science.aat5691.

[49] Michaela Asp, Joseph Bergenstråhle, and Joakim Lundeberg. "Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration". In: *BioEssays* 42.10 (2020), p. 1900221. ISSN: 1521-1878. DOI: 10.1002/bies.201900221.

[50] Luyi Tian, Fei Chen, and Evan Z. Macosko. "The Expanding Vistas of Spatial Transcriptomics". In: *Nature Biotechnology* (Oct. 3, 2022), pp. 1–10. ISSN: 1546-1696. DOI: 10.1038/s41587-022-01448-2.

[51] Jeffrey R. Moffitt, Emma Lundberg, and Holger Heyn. "The Emerging Landscape of Spatial Profiling Technologies". In: *Nature Reviews Genetics* (July 20, 2022), pp. 1–19. ISSN: 1471-0064. DOI: 10.1038/s41576-022-00515-3.

[52] Lukas Valihrach, Daniel Zucha, Pavel Abaffy, and Mikael Kubista. "A Practical Guide to Spatial Transcriptomics". In: *Molecular Aspects of Medicine* 97 (June 1, 2024), p. 101276. ISSN: 0098-2997. DOI: 10.1016/j.mam.2024.101276.

[53] Yue You, Yuting Fu, Lanxiang Li, Zhongmin Zhang, Shikai Jia, Shihong Lu, Wenle Ren, Yifang Liu, Yang Xu, Xiaojing Liu, Fuqing Jiang, Guangdun Peng, Abhishek Sampath Kumar, Matthew E. Ritchie, Xiaodong Liu, and Luyi Tian. "Systematic Comparison of Sequencing-Based Spatial Transcriptomic Methods". In: *Nature Methods* 21.9 (Sept. 2024), pp. 1743–1754. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02325-3.

[54] Lambda Moses and Lior Pachter. "Museum of Spatial Transcriptomics". In: *Nature Methods* (Mar. 10, 2022), pp. 1–13. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01409-2.

[55] Paulien Hogeweg. "The Roots of Bioinformatics in Theoretical Biology". In: *PLOS Computational Biology* 7.3 (Mar. 31, 2011), e1002021. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002021.

[56] Joel B. Hagen. "The Origins of Bioinformatics". In: *Nature Reviews Genetics* 1.3 (Dec. 2000), pp. 231–236. ISSN: 1471-0064. DOI: 10.1038/35042090.

[57] M. O. Dayhoff. "Computer Aids to Protein Sequence Determination". In: *Journal of Theoretical Biology* 8.1 (Jan. 1, 1965), pp. 97–112. ISSN: 0022-5193. DOI: 10.1016/0022-5193(65)90096-2.

[58] Christos A. Ouzounis and Alfonso Valencia. "Early Bioinformatics: The Birth of a Discipline—a Personal View". In: *Bioinformatics* 19.17 (Nov. 22, 2003), pp. 2176–2190. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg309.

[59] Jonathan D. Wren. "Bioinformatics Programs Are 31-Fold over-Represented among the Highest Impact Scientific Papers of the Past Two Decades". In: *Bioinformatics* 32.17 (Sept. 1, 2016), pp. 2686–2691. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw284.

[60] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. "A Survey of Best Practices for RNA-seq Data Analysis". In: *Genome Biology* 17.1 (Jan. 26, 2016), p. 13. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8.

[61] Aishwarya Gondane and Harri M. Itkonen. "Revealing the History and Mystery of RNA-Seq". In: *Current Issues in Molecular Biology* 45.3 (Feb. 24, 2023), pp. 1860–1874. ISSN: 1467-3037. DOI: 10.3390/cimb45030120. PMID: 36975490.

[62] Koen Van den Berge, Katharina M. Hembach, Charlotte Soneson, Simone Tiberi, Lieven Clement, Michael I. Love, Rob Patro, and Mark D. Robinson. "RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis". In: *Annual Review of Biomedical Data Science* 2.1 (2019), pp. 139–173. DOI: 10.1146/annurev-biodatasci-072018-021255.

[63] Mohammed Alser, Jeremy Rotman, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, Victor Xue, Sergey Knyazev, Benjamin D. Singer, Brunilda Balliu, David Koslicki, Pavel Skums, Alex Zelikovsky, Can Alkan, Onur Mutlu, and Serghei Mangul. "Technology Dictates Algorithms: Recent Developments in Read Alignment". In: *Genome Biology* 22.1 (Aug. 26, 2021), p. 249. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02443-7.

[64] "Method of the Year 2013". In: *Nature Methods* 11.1 (Jan. 2014), pp. 1–1. ISSN: 1548-7105. DOI: 10.1038/nmeth.2801.

[65] Rickard Sandberg. "Entering the Era of Single-Cell Transcriptomics in Biology and Medicine". In: *Nature Methods* 11.1 (Jan. 2014), pp. 22–24. ISSN: 1548-7105. DOI: 10.1038/nmeth.2764.

[66] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. "The Human Cell Atlas". In: *eLife* 6 (Dec. 5, 2017). Ed. by Thomas R Gingeras, e27041. ISSN: 2050-084X. DOI: 10.7554/eLife.27041.

[67] Luke Simpson, Andrew Strange, Doris Klisch, Sophie Kraunsoe, Takuya Azami, Daniel Goszczynski, Triet Le Minh, Benjamin Planells, Nadine Holmes, Fei Sang, Sonal Henson, Matthew Loose, Jennifer Nichols, and Ramiro Alberio. "A Single-Cell Atlas of Pig Gastrulation as a Resource for Comparative Embryology". In: *Nature Communications* 15.1 (June 18, 2024), p. 5210. ISSN: 2041-1723. DOI: 10.1038/s41467-024-49407-6.

[68] Zizhen Yao, Cindy T. J. van Velthoven, Michael Kunst, Meng Zhang, Delissa McMillen, Changkyu Lee, Won Jung, Jeff Goldy, Aliya Abdelhak, Matthew Aitken, Katherine Baker, Pamela Baker, Eliza Barkan, Darren Bertagnolli, Ashwin Bhandiwad, Cameron Bielstein, Prajal Bishwakarma, Jazmin Campos, Daniel Carey, Tamara Casper, Anish Bhaswanth Chakka, Rushil Chakrabarty, Sakshi Chavan, Min Chen, Michael Clark, Jennie Close, Kirsten Crichton, Scott Daniel, Peter DiValentin, Tim Dolbeare, Lauren Ellingwood, Elysha Fiabane, Timothy Fliss, James Gee, James Gerstenberger, Alexandra Glandon, Jessica Gloe, Joshua Gould, James Gray, Nathan Guilford, Junitta Guzman, Daniel Hirschstein, Windy Ho, Marcus Hooper, Mike Huang, Madie Hupp, Kelly Jin, Matthew Kroll, Kanan Lathia, Arielle Leon, Su Li, Brian Long, Zach Madigan, Jessica Malloy, Jocelin Malone, Zoe Maltzer, Naomi Martin, Rachel McCue, Ryan McGinty, Nicholas Mei, Jose Melchor, Emma Meyerdierks, Tyler Mollenkopf, Skyler Moonsman, Thuc Nghi Nguyen, Sven Otto, Trangthanh Pham, Christine Rimorin, Augustin Ruiz, Raymond Sanchez, Lane Sawyer, Nadiya Shapovalova, Noah Shepard, Cliff Slaughterbeck, Josef Sulc, Michael Tieu, Amy Torkelson, Herman Tung, Nasmil Valera Cuevas, Shane Vance, Katherine Wadhwani, Katelyn Ward, Boaz Levi, Colin Farrell, Rob Young, Brian Staats, Ming-Qiang Michael Wang, Carol L. Thompson, Shoaib Mufti, Chelsea M. Pagan, Lauren Kruse, Nick Dee, Susan M. Sunkin, Luke Esposito, Michael J. Hawrylycz, Jack Waters, Lydia Ng, Kimberly Smith, Bosiljka Tasic, et al. "A High-Resolution Transcriptomic and Spatial Atlas of Cell Types in the Whole Mouse Brain". In: *Nature* 624.7991 (Dec. 2023), pp. 317–332. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06812-z.

[69] Jin Li, Jongsu Choi, Xuesen Cheng, Justin Ma, Shahil Pema, Joshua R. Sanes, Graeme Mardon, Benjamin J. Frankfort, Nicholas M. Tran, Yumei Li, and Rui Chen. "Comprehensive Single-Cell Atlas of the Mouse Retina". In: *iScience* 27.6 (June 21, 2024). ISSN: 2589-0042. DOI: 10.1016/j.isci.2024.109916.

[70] Ramón Massoni-Badosa, Paula Soler-Vila, Sergio Aguilar-Fernández, Juan C. Nieto, Marc Elosua-Bayes, Domenica Marchese, Marta Kulis, Amaia Vilas-Zornoza, Marco Matteo Bühler, Sonal Rashmi, Clara Alsinet, Ginevra Caratù, Catia Moutinho, Sara Ruiz, Patricia Lorden, Giulia Lunazzi, Dolors Colomer, Gerard Frigola, Will Blevins, Sara Palomino, David Gomez-Cabrero, Xabier Aguirre, Marc A. Weniger, Federico Marini, Francisco Javier Cervera-Paz, Peter M. Baptista, Isabel Vilaseca, Felipe Prosper, Ralf Küppers, Ivo Glynne Gut, Elias Campo,

José Ignacio Martin-Subero, and Holger Heyn. “An Atlas of Cells in the Human Tonsil”. June 26, 2022. DOI: 10.1101/2022.06.24.497299. Pre-published.

[71] Susan M. Sunkin, Lydia Ng, Chris Lau, Tim Dolbeare, Terri L. Gilbert, Carol L. Thompson, Michael Hawrylycz, and Chinh Dang. “Allen Brain Atlas: An Integrated Spatio-Temporal Portal for Exploring the Central Nervous System”. In: *Nucleic Acids Research* 41.D1 (Jan. 1, 2013), pp. D996–D1008. ISSN: 0305-1048. DOI: 10.1093/nar/gks1042.

[72] Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C. Strobl, Tessa E. Gillett, Luke Zappia, Elo Madissoon, Nikolay S. Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, Marie-Jeanne Arguel, Leonie Apperloo, Martin Banchero, Christophe Bécavin, Marijn Berg, Evgeny Chichelnitskiy, Mei-i Chung, Antoine Collin, Aurore C. A. Gay, Janine Gote-Schniering, Baharak Hooshiar Kashani, Kemal Inecik, Manu Jain, Theodore S. Kapellos, Tessa M. Kole, Sylvie Leroy, Christoph H. Mayr, Amanda J. Oliver, Michael von Papen, Lance Peter, Chase J. Taylor, Thomas Walzthoeni, Chuan Xu, Linh T. Bui, Carlo De Donno, Leander Dony, Alen Faiz, Minzhe Guo, Austin J. Gutierrez, Lukas Heumos, Ni Huang, Ignacio L. Ibarra, Nathan D. Jackson, Preetish Kadur Lakshminarasimha Murthy, Mohammad Lotfollahi, Tracy Tabib, Carlos Talavera-López, Kyle J. Travaglini, Anna Wilbrey-Clark, Kaylee B. Worlock, Masahiro Yoshida, Maarten van den Berge, Yohan Bossé, Tushar J. Desai, Oliver Eickelberg, Naftali Kaminski, Mark A. Krasnow, Robert Lafyatis, Marko Z. Nikolic, Joseph E. Powell, Jayaraj Rajagopal, Mauricio Rojas, Orit Rozenblatt-Rosen, Max A. Seibold, Dean Sheppard, Douglas P. Shepherd, Don D. Sin, Wim Timens, Alexander M. Tsankov, Jeffrey Whitsett, Yan Xu, Nicholas E. Banovich, Pascal Barbry, Thu Elizabeth Duong, Christine S. Falk, Kerstin B. Meyer, Jonathan A. Kropski, Dana Pe’er, Herbert B. Schiller, Purushothama Rao Tata, Joachim L. Schultze, Sara A. Teichmann, Alexander V. Misharin, Martijn C. Nawijn, Malte D. Luecken, and Fabian J. Theis. “An Integrated Cell Atlas of the Lung in Health and Disease”. In: *Nature Medicine* 29.6 (6 June 2023), pp. 1563–1577. ISSN: 1546-170X. DOI: 10.1038/s41591-023-02327-2.

[73] Peter V. Kharchenko. “The Triumphs and Limitations of Computational Methods for scRNA-seq”. In: *Nature Methods* 18.7 (7 July 2021), pp. 723–732. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01171-x.

[74] Changde Cheng, Wenan Chen, Hongjian Jin, and Xiang Chen. “A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell–Cell Communication”. In: *Cells* 12.15 (Jan. 2023), p. 1970. ISSN: 2073-4409. DOI: 10.3390/cells12151970.

[75] *How Many Cells Are Captured in a Single Spot?* 10X Genomics. URL: https://kb.10xgenomics.com/hc/en-us/articles/360035487952-How-many-cells-are-captured-in-a-single-spot (visited on 08/20/2025).

[76] Simon Mages, Noa Moriel, Inbal Avraham-Davidi, Evan Murray, Jan Watter, Fei Chen, Orit Rozenblatt-Rosen, Johanna Klughammer, Aviv Regev, and Mor Nitzan. “TACCO Unifies Annotation Transfer and Decomposition of Cell Identities for Single-Cell and Spatial Omics”. In: *Nature Biotechnology* (Feb. 16, 2023), pp. 1–9. ISSN: 1546-1696. DOI: 10.1038/s41587-023-01657-3.

[77] Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. “Robust Decomposition of Cell Type Mixtures in Spatial Transcriptomics”. In: *Nature Biotechnology* 40.4 (4 Apr. 2022), pp. 517–526. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00830-w.

[78] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. “Cell2location Maps

Fine-Grained Cell Types in Spatial Transcriptomics". In: *Nature Biotechnology* 40.5 (5 May 2022), pp. 661–671. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01139-4.

[79] Jiaqiang Zhu, Shiquan Sun, and Xiang Zhou. "SPARK-X: Non-Parametric Modeling Enables Scalable and Robust Detection of Spatial Expression Patterns for Large Spatial Transcriptomic Studies". In: *Genome Biology* 22.1 (June 21, 2021), p. 184. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02404-0.

[80] Valentine Svensson, Sarah A. Teichmann, and Oliver Stegle. "SpatialDE: Identification of Spatially Variable Genes". In: *Nature Methods* 15.5 (5 May 2018), pp. 343–346. ISSN: 1548-7105. DOI: 10.1038/nmeth.4636.

[81] Guoxin Cai, Yichang Chen, Xun Gu, and Zhan Zhou. "Spanve: An Effective Statistical Method to Detect Spatially Variable Genes in Large-scale Spatial Transcriptomics Data". Feb. 8, 2023. DOI: 10.1101/2023.02.08.527623. Pre-published.

[82] Zheng Li and Xiang Zhou. "BASS: Multi-Scale and Multi-Sample Analysis Enables Accurate Cell Type Clustering and Spatial Domain Detection in Spatial Transcriptomic Studies". In: *Genome Biology* 23.1 (Aug. 4, 2022), p. 168. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02734-7.

[83] Vipul Singhal, Nigel Chou, Joseph Lee, Yifei Yue, Jinyue Liu, Wan Kee Chock, Li Lin, Yun-Ching Chang, Erica Mei Ling Teo, Jonathan Aow, Hwee Kuan Lee, Kok Hao Chen, and Shyam Prabhakar. "BANKSY Unifies Cell Typing and Tissue Domain Segmentation for Scalable Spatial Omics Data Analysis". In: *Nature Genetics* 56.3 (Mar. 2024), pp. 431–441. ISSN: 1546-1718. DOI: 10.1038/s41588-024-01664-3.

[84] Kangning Dong and Shihua Zhang. "Deciphering Spatial Domains from Spatially Resolved Transcriptomics with an Adaptive Graph Attention Auto-Encoder". In: *Nature Communications* 13.1 (1 Apr. 1, 2022), p. 1739. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29439-6.

[85] Jian Hu, Amelia Schroeder, Kyle Coleman, Chixiang Chen, Benjamin J. Auerbach, and Mingyao Li. "Statistical and Machine Learning Methods for Spatially Resolved Transcriptomics with Histology". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 3829–3841. ISSN: 20010370. DOI: 10.1016/j.csbj.2021.06.052.

[86] Brendan F. Miller, Dhananjay Bambah-Mukku, Catherine Dulac, Xiaowei Zhuang, and Jean Fan. "Characterizing Spatial Gene Expression Heterogeneity in Spatially Resolved Single-Cell Transcriptomic Data with Nonuniform Cellular Densities". In: *Genome Research* 31.10 (Jan. 10, 2021), pp. 1843–1855. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.271288.120. PMID: 34035045.

[87] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, Rani E. George, Nico Pierson, Long Cai, and Guo-Cheng Yuan. "Giotto: A Toolbox for Integrative Analysis and Visualization of Spatial Expression Data". In: *Genome Biology* 22.1 (Mar. 8, 2021), p. 78. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02286-2.

[88] Donald E. Ingber. "Mechanical Control of Tissue Growth: Function Follows Form". In: *Proceedings of the National Academy of Sciences* 102.33 (Aug. 16, 2005), pp. 11571–11572. DOI: 10.1073/pnas.0505939102.

[89] Adam J. Engler, Patrick O. Humbert, Bernhard Wehrle-Haller, and Valerie M. Weaver. "Multiscale Modeling of Form and Function". In: *Science* 324.5924 (Apr. 10, 2009), pp. 208–212. DOI: 10.1126/science.1170107.

[90] D. Friday King and Laura A. C. King. "A Brief Historical Note on Staining by Hematoxylin and Eosin". In: *The American Journal of Dermatopathology* 8.2 (Apr. 1986), p. 168. ISSN: 0193-1091.

[91] John K. C. Chan. "The Wonderful Colors of the Hematoxylin–Eosin Stain in Diagnostic Surgical Pathology". In: *International Journal of Surgical Pathology* 22.1 (Feb. 1, 2014), pp. 12–32. ISSN: 1066-8969. DOI: 10.1177/1066896913517939.

[92] Wikipedia contributors. *Histopathology of Basal Cell Carcinoma of the Skin.* Wikipedia. Jan. 9, 2006. URL: https://commons.wikimedia.org/wiki/File:Basal_cell_carcinoma_histopathology_(3).jpg (visited on 10/17/2025).

[93] Murli Krishna. "Role of Special Stains in Diagnostic Liver Pathology". In: *Clinical Liver Disease* 2.S1 (2013), S8–S10. ISSN: 2046-2484. DOI: 10.1002/cld.148.

[94] Rolf Zehbe, Astrid Haibel, Heinrich Riesemeier, Ulrich Gross, C. James Kirkpatrick, Helmut Schubert, and Christoph Brochhausen. "Going beyond Histology. Synchrotron Micro-Computed Tomography as a Methodology for Biological Tissue Characterization: From Tissue Morphology to Individual Cells". In: *Journal of The Royal Society Interface* 7.42 (Mar. 25, 2009), pp. 49–59. DOI: 10.1098/rsif.2008.0539.

[95] Silas Maniatis, Joana Petrescu, and Hemali Phatnani. "Spatially Resolved Transcriptomics and Its Applications in Cancer". In: *Current Opinion in Genetics & Development.* Cancer Genomics 66 (Feb. 1, 2021), pp. 70–77. ISSN: 0959-437X. DOI: 10.1016/j.gde.2020.12.002.

[96] Stefania Giacomello, Fredrik Salmén, Barbara K. Terebieniec, Sanja Vickovic, José Fernandez Navarro, Andrey Alexeyenko, Johan Reimegård, Lauren S. McKee, Chanaka Mannapperuma, Vincent Bulone, Patrik L. Ståhl, Jens F. Sundström, Nathaniel R. Street, and Joakim Lundeberg. "Spatially Resolved Transcriptome Profiling in Model Plant Species". In: *Nature Plants* 3.6 (May 8, 2017), p. 17061. ISSN: 2055-0278. DOI: 10.1038/nplants.2017.61.

[97] Zhiyuan Yuan, Wentao Pan, Xuan Zhao, Fangyuan Zhao, Zhimeng Xu, Xiu Li, Yi Zhao, Michael Q. Zhang, and Jianhua Yao. "SODB Facilitates Comprehensive Exploration of Spatial Omics Data". In: *Nature Methods* (Feb. 16, 2023), pp. 1–13. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01773-7.

[98] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (Sept. 2, 2018), p. 861. ISSN: 2475-9066. DOI: 10.21105/joss.00861.

[99] Vivien Marx. "Seeing Data as T-SNE and UMAP Do". In: *Nature Methods* 21.6 (June 2024), pp. 930–933. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02301-x.

[100] Tara Chari and Lior Pachter. "The Specious Art of Single-Cell Genomics". In: *PLOS Computational Biology* 19.8 (Aug. 17, 2023), e1011288. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011288.

[101] Han Chu, Kun Wang, Hansen Cheng, Wenhao Ma, Liting Dong, Yixiong Gou, Jian Yang, and Haoyang Cai. "Exploring the Landscape of Spatial Transcriptome Analysis: Introducing STASH, a Database of Spatial Transcriptome Tools". Apr. 21, 2023. DOI: 10.1101/2023.04.20.537419. Pre-published.

[102] Benjamin L. Walker, Zixuan Cang, Honglei Ren, Eric Bourgain-Chang, and Qing Nie. "Deciphering Tissue Structure and Function Using Spatial Transcriptomics". In: *Communications Biology* 5.1 (Mar. 10, 2022), p. 220. ISSN: 2399-3642. DOI: 10.1038/s42003-022-03175-5.

[103] Roopali Singh, Xi He, Adam Keebum Park, Ross Cameron Hardison, Xiang Zhu, and Qunhua Li. "Retrofit: Reference-Free Deconvolution of Cell-Type Mixtures in Spatial Transcriptomics". June 9, 2023. DOI: 10.1101/2023.06.07.544126. Pre-published.

[104] Kazumasa Kanemaru, James Cranley, Daniele Muraro, Antonio M. A. Miranda, Siew Yen Ho, Anna Wilbrey-Clark, Jan Patrick Pett, Krzysztof Polanski, Laura Richardson, Monika Litvinukova, Natsuhiko Kumasaka, Yue Qin, Zuzanna Jablonska, Claudia I. Semprich, Lukas Mach, Monika Dabrowska, Nathan Richoz, Liam Bolt, Lira Mamanova, Rakeshlal Kapuge, Sam N. Barnett, Shani Perera, Carlos Talavera-López, Ilaria Mulas, Krishnaa T. Mahbubani, Liz Tuck, Lu Wang, Margaret M. Huang, Martin Prete, Sophie Pritchard, John Dark, Kourosh Saeb-Parsy, Minal Patel, Menna R. Clatworthy, Norbert Hübner, Rasheda A. Chowdhury, Michela Noseda, and Sarah A. Teichmann. "Spatially Resolved Multiomics of Human Cardiac Niches". In: *Nature* 619.7971 (July 2023), pp. 801–810. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06311-1.

[105] Linhua Wang, Mirjana Maletic-Savatic, and Zhandong Liu. "Region-Specific Denoising Identifies Spatial Co-Expression Patterns and Intra-Tissue Heterogeneity in Spatially Resolved Transcriptomics Data". In: *Nature Communications* 13.1 (1 Nov. 14, 2022), p. 6912. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34567-0.

[106] Jean-Baptiste Pettit, Raju Tomer, Kaia Achim, Sylvia Richardson, Lamiae Azizi, and John Marioni. "Identifying Cell Types from Spatially Referenced Single-Cell Expression Datasets". In: *PLOS Computational Biology* 10.9 (Sept. 25, 2014), e1003824. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003824.

[107] Mor Nitzan, Nikos Karaiskos, Nir Friedman, and Nikolaus Rajewsky. "Gene Expression Cartography". In: *Nature* 576.7785 (7785 Dec. 2019), pp. 132–137. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1773-3.

[108] Dylan Kotliar, Adrian Veres, M Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A Melton, and Pardis C Sabeti. "Identifying Gene Expression Programs of Cell-Type Identity and Cellular Activity with Single-Cell RNA-Seq". In: *eLife* 8 (July 8, 2019). Ed. by Alfonso Valencia, Naama Barkai, Elisabetta Mereu, and Berthold Göttgens, e43803. ISSN: 2050-084X. DOI: 10.7554/eLife.43803.

[109] Jiachen Li, Siheng Chen, Xiaoyong Pan, Ye Yuan, and Hong-Bin Shen. "Cell Clustering for Spatial Transcriptomics Data with Graph Neural Networks". In: *Nature Computational Science* 2.6 (June 2022), pp. 399–408. ISSN: 2662-8457. DOI: 10.1038/s43588-022-00266-5.

[110] "What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?" In: *Cell Systems* 4.3 (Mar. 22, 2017), pp. 255–259. ISSN: 2405-4712, 2405-4720. DOI: 10.1016/j.cels.2017.03.006. PMID: 28334573.

[111] Jonas Simon Fleck, J. Gray Camp, and Barbara Treutlein. "What Is a Cell Type?" In: *Science* 381.6659 (Aug. 18, 2023), pp. 733–734. DOI: 10.1126/science.adf6162.

[112] Jeff J. Doyle. "Cell Types as Species: Exploring a Metaphor". In: *Frontiers in Plant Science* 13 (Aug. 22, 2022). ISSN: 1664-462X. DOI: 10.3389/fpls.2022.868565.

[113] Stafford Beer. "What Is Cybernetics?" In: *Kybernetes* 31.2 (Mar. 1, 2002), pp. 209–219. ISSN: 0368-492X. DOI: 10.1108/03684920210417283.

[114] Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, Matthew N. Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E. Kleinman, Thomas M. Hyde, Nikhil Rao, Stephanie C. Hicks, Keri Martinowich, and Andrew E. Jaffe. "Transcriptome-Scale Spatial Gene Expression in the Human Dorsolateral Prefrontal Cortex". In: *Nature Neuroscience* 24.3 (3 Mar. 2021), pp. 425–436. ISSN: 1546-1726. DOI: 10.1038/s41593-020-00787-0.

[115] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, Li Zeng, Huazhu Fu, Min Wu, Lina Hsiu Kim Lim, Longqi Liu, and Jinmiao Chen. "Spatially Informed Clustering, Integration, and Deconvolution of Spatial Transcriptomics with GraphST". In: *Nature Communications* 14.1 (1 Mar. 1, 2023), p. 1155. ISSN: 2041-1723. DOI: 10.1038/s41467-023-36796-3.

[116] Yi Yang, Xingjie Shi, Wei Liu, Qiuzhong Zhou, Mai Chan Lau, Jeffrey Chun Tatt Lim, Lei Sun, Cedric Chuan Young Ng, Joe Yeong, and Jin Liu. "SC-MEB: Spatial Clustering with Hidden Markov Random Field Using Empirical Bayes". In: *Briefings in Bioinformatics* 23.1 (Jan. 17, 2022), bbab466. ISSN: 1477-4054. DOI: 10.1093/bib/bbab466. PMID: 34849574.

[117] Teng Liu, Zhao-Yu Fang, Zongbo Zhang, Yongxiang Yu, Min Li, and Ming-Zhu Yin. "A Comprehensive Overview of Graph Neural Network-Based Approaches to Clustering for Spatial Transcriptomics". In: *Computational and Structural Biotechnology Journal* 23 (Dec. 1, 2024), pp. 106–128. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2023.11.055. PMID: 38089467.

[118] Qian Zhu, Sheel Shah, Ruben Dries, Long Cai, and Guo-Cheng Yuan. "Identification of Spatially Associated Subpopulations by Combining scRNAseq and Sequential Fluorescence in Situ Hybridization Data". In: *Nature Biotechnology* 36.12 (12 Dec. 2018), pp. 1183–1190. ISSN: 1546-1696. DOI: 10.1038/nbt.4260.

[119] Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam, Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, Jason H. Bielas, and Raphael Gottardo. "Spatial Transcriptomics at Subspot Resolution with BayesSpace". In: *Nature Biotechnology* 39.11 (11 Nov. 2021), pp. 1375–1384. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00935-2.

[120] Yanghong Guo, Bencong Zhu, Chen Tang, Ruichen Rong, Ying Ma, Guanghua Xiao, Lin Xu, and Qiwei Li. "BayeSMART: Bayesian Clustering of Multi-Sample Spatially Resolved Transcriptomics Data". In: *Briefings in Bioinformatics* 25.6 (Nov. 1, 2024), bbae524. ISSN: 1477-4054. DOI: 10.1093/bib/bbae524.

[121] Duy Pham, Xiao Tan, Brad Balderson, Jun Xu, Laura F. Grice, Sohye Yoon, Emily F. Willis, Minh Tran, Pui Yeng Lam, Arti Raghubar, Priyakshi Kalita-de Croft, Sunil Lakhani, Jana Vukovic, Marc J. Ruitenberg, and Quan H. Nguyen. "Robust Mapping of Spatiotemporal Trajectories and Cell–Cell Interactions in Healthy and Diseased Tissues". In: *Nature Communications* 14.1 (Nov. 25, 2023), p. 7739. ISSN: 2041-1723. DOI: 10.1038/s41467-023-43120-6.

[122] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. "SpaGCN: Integrating Gene Expression, Spatial Location and Histology to Identify Spatial Domains and Spatially Variable Genes by Graph Convolutional Network". In: *Nature Methods* 18.11 (11 Nov. 2021), pp. 1342–1351. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01255-8.

[123] Daoliang Zhang, Na Yu, Zhiyuan Yuan, Wenrui Li, Xue Sun, Qi Zou, Xiangyu Li, Zhiping Liu, Wei Zhang, and Rui Gao. "stMMR: Accurate and Robust Spatial Domain Identification from Spatially Resolved Transcriptomics with Multimodal Feature Representation". In: *GigaScience* 13 (Jan. 1, 2024), giae089. ISSN: 2047-217X. DOI: 10.1093/gigascience/giae089.

[124] Yiran Shan, Qian Zhang, Wenbo Guo, Yanhong Wu, Yuxin Miao, Hongyi Xin, Qiuyu Lian, and Jin Gu. "TIST: Transcriptome and Histopathological Image Integrative Analysis for Spatial Transcriptomics". In: *Genomics, Proteomics & Bioinformatics* (Dec. 19, 2022). ISSN: 1672-0229. DOI: 10.1016/j.gpb.2022.11.012.

[125] Zhiyuan Yuan, Fangyuan Zhao, Senlin Lin, Yu Zhao, Jianhua Yao, Yan Cui, Xiao-Yong Zhang, and Yi Zhao. "Benchmarking Spatial Clustering Methods with Spatially Resolved Transcriptomics Data". In: *Nature Methods* (Mar. 15, 2024), pp. 1–11. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02215-8.

[126] Liping Kang, Qinglong Zhang, Fan Qian, Junyao Liang, and Xiaohui Wu. "Benchmarking Computational Methods for Detecting Spatial Domains and Domain-Specific Spatially Variable Genes from Spatial Transcriptomics Data". In: *Nucleic Acids Research* 53.7 (Apr. 16, 2025), gkaf303. ISSN: 0305-1048. DOI: 10.1093/nar/gkaf303. PMID: 40240000.

[127] Zixuan Cang, Xinyi Ning, and Jing Zhang. "SCAN-IT: Domain Segmentation of Spatial Transcriptomics Images by Graph Neural Network". In: BMVC 2021. Nov. 2021.

[128] Yunfei Hu, Yuying Zhao, Curtis T. Schunk, Yingxiang Ma, Tyler Derr, and Xin Maizie Zhou. "ADEPT: Autoencoder with Differentially Expressed Genes and Imputation for Robust Spatial Transcriptomics Clustering". In: *iScience* 26.6 (June 16, 2023), p. 106792. ISSN: 2589-0042. DOI: 10.1016/j.isci.2023.106792.

[129] Zhen Li, Xiaoyang Chen, Xuegong Zhang, Rui Jiang, and Shengquan Chen. "Latent Feature Extraction with a Prior-Based Self-Attention Framework for Spatial Transcriptomics". In: *Genome Research* 33.10 (Jan. 10, 2023), pp. 1757–1773. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.277891.123. PMID: 37903634.

[130] Bo Wang, Jiawei Luo, Ying Liu, Wanwan Shi, Zehao Xiong, Cong Shen, and Yahui Long. "Spatial-MGCN: A Novel Multi-View Graph Convolutional Network for Identifying Spatial Domains with Attention Mechanism". In: *Briefings in Bioinformatics* 24.5 (Sept. 1, 2023), bbad262. ISSN: 1477-4054. DOI: 10.1093/bib/bbad262.

[131] Luca Scrucca, Chris Fraley, T. Brendan Murphy, and Adrian E. Raftery. *Model-Based Clustering, Classification, and Density Estimation Using Mclust in R*. New York: Chapman and Hall/CRC, Apr. 20, 2023. 268 pp. ISBN: 978-1-003-27796-5. DOI: 10.1201/9781003277965.

[132] V. A. Traag, L. Waltman, and N. J. van Eck. "From Louvain to Leiden: Guaranteeing Well-Connected Communities". In: *Scientific Reports* 9.1 (Mar. 26, 2019), p. 5233. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41695-z.

[133] Zhiyuan Yuan, Yisi Li, Minglei Shi, Fan Yang, Juntao Gao, Jianhua Yao, and Michael Q. Zhang. "SOTIP Is a Versatile Method for Microenvironment Modeling with Spatial Omics Data". In: *Nature Communications* 13.1 (1 Nov. 28, 2022), p. 7330. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34867-5.

[134] Bin Duan, Shaoqi Chen, Xiaojie Cheng, and Qi Liu. "Multi-Slice Spatial Transcriptome Domain Analysis with SpaDo". In: *Genome Biology* 25.1 (1 Mar. 19, 2024), pp. 1–23. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03213-x.

[135] Marco Varrone, Daniele Tavernari, Albert Santamaria-Martínez, Logan A. Walsh, and Giovanni Ciriello. "CellCharter Reveals Spatial Cell Niches Associated with Tissue Remodeling and Cell Plasticity". In: *Nature Genetics* 56.1 (1 Jan. 2024), pp. 74–84. ISSN: 1546-1718. DOI: 10.1038/s41588-023-01588-4.

[136] Lulu Shang and Xiang Zhou. "Spatially Aware Dimension Reduction for Spatial Transcriptomics". In: *Nature Communications* 13.1 (1 Nov. 23, 2022), p. 7203. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34879-1.

[137] Jiyuan Yang, Lu Wang, Lin Liu, and Xiaoqi Zheng. "GraphPCA: A Fast and Interpretable Dimension Reduction Algorithm for Spatial Transcriptomics Data". In: *Genome Biology* 25.1 (Nov. 7, 2024), p. 287. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03429-x.

[138] Mo Chen, Ruihua Cheng, Jianuo He, Jun Chen, and Jie Zhang. "SMOPCA: Spatially Aware Dimension Reduction Integrating Multi-Omics Improves the Efficiency of Spatial Domain Detection". In: *Genome Biology* 26.1 (May 21, 2025), p. 135. ISSN: 1474-760X. DOI: 10.1186/s13059-025-03576-9.

[139] Julien Moehlin, Bastien Mollet, Bruno Maria Colombo, and Marco Antonio Mendoza-Parra. "Inferring Biologically Relevant Molecular Tissue Substructures by Agglomerative Clustering of Digitized Spatial Transcriptomes with MULTILAYER". In: *Cell Systems* 12.7 (July 2021), 694–705.e3. ISSN: 24054712. DOI: 10.1016/j.cels.2021.04.008.

[140] Patrick C N Martin, Hyobin Kim, Cecilia Lövkvist, Byung-Woo Hong, and Kyoung Jae Won. "Vesalius: High-Resolution in Silico Anatomization of Spatial Transcriptomic Data Using Image Analysis". In: *Molecular Systems Biology* 18.9 (Sept. 2022), e11080. ISSN: 1744-4292. DOI: 10.15252/msb.202211080.

[141] R.P. Weicker. "An Overview of Common Benchmarks". In: *Computer* 23.12 (Dec. 1990), pp. 65–75. ISSN: 1558-0814. DOI: 10.1109/2.62094.

[142] Mohamed Radhouene Aniba, Olivier Poch, and Julie D. Thompson. "Issues in Bioinformatics Benchmarking: The Case Study of Multiple Sequence Alignment". In: *Nucleic Acids Research* 38.21 (Nov. 1, 2010), pp. 7353–7363. ISSN: 0305-1048. DOI: 10.1093/nar/gkq625.

[143] Raquel Norel, John Jeremy Rice, and Gustavo Stolovitzky. "The Self-assessment Trap: Can We All Be Better than Average?" In: *Molecular Systems Biology* 7.1 (Jan. 2011), p. 537. ISSN: 1744-4292. DOI: 10.1038/msb.2011.70.

[144] Anne-Laure Boulesteix, Sabine Lauer, and Manuel J. A. Eugster. "A Plea for Neutral Comparison Studies in Computational Sciences". In: *PLOS ONE* 8.4 (Apr. 24, 2013), e61562. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0061562.

[145] Lukas M. Weber, Wouter Saelens, Robrecht Cannoodt, Charlotte Soneson, Alexander Hapfelmeier, Paul P. Gardner, Anne-Laure Boulesteix, Yvan Saeys, and Mark D. Robinson. "Essential Guidelines for Computational Method Benchmarking". In: *Genome Biology* 20.1 (June 20, 2019), p. 125. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1738-8.

[146] Iven Van Mechelen, Anne-Laure Boulesteix, Rainer Dangl, Nema Dean, Christian Hennig, Friedrich Leisch, Douglas Steinley, and Matthijs J. Warrens. "A White Paper on Good Research Practices in Benchmarking: The Case of Cluster Analysis". In: *WIREs Data Mining and Knowledge Discovery* 13.6 (2023), e1511. ISSN: 1942-4795. DOI: 10.1002/widm.1511.

[147] Serghei Mangul, Lana S. Martin, Brian L. Hill, Angela Ka-Mei Lam, Margaret G. Distler, Alex Zelikovsky, Eleazar Eskin, and Jonathan Flint. "Systematic Benchmarking of Omics Computational Tools". In: *Nature Communications* 10.1 (Dec. 2019), p. 1393. ISSN: 2041-1723. DOI: 10.1038/s41467-019-09406-4.

[148] Thomas G. Brooks, Nicholas F. Lahens, Antonijo Mrčela, and Gregory R. Grant. "Challenges and Best Practices in Omics Benchmarking". In: *Nature Reviews Genetics* (Jan. 12, 2024), pp. 1–14. ISSN: 1471-0064. DOI: 10.1038/s41576-023-00679-6.

[149] Bjoern Peters, Steven E. Brenner, Edwin Wang, Donna Slonim, and Maricel G. Kann. "Putting Benchmarks in Their Rightful Place: The Heart of Computational Biology". In: *PLOS Computational Biology* 14.11 (Nov. 8, 2018), e1006494. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006494.

[150] Luke Zappia, Belinda Phipson, and Alicia Oshlack. "Exploring the Single-Cell RNA-seq Analysis Landscape with the scRNA-tools Database". In: *PLOS Computational Biology* 14.6 (June 25, 2018), e1006245. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006245.

[151] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. "A Comparison of Single-Cell Trajectory Inference Methods". In: *Nature Biotechnology* 37.5 (5 May 2019), pp. 547–554. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0071-9.

[152] Anthony Sonrel, Almut Luetge, Charlotte Soneson, Izaskun Mallona, Pierre-Luc Germain, Sergey Knyazev, Jeroen Gilis, Reto Gerber, Ruth Seurinck, Dominique Paul, Emanuel Sonder, Helena L. Crowell, Imran Fanaswala, Ahmad Al-Ajami, Elyas Heidari, Stephan Schmeing, Stefan Milosavljevic, Yvan Saeys, Serghei Mangul, and Mark D. Robinson. "Meta-Analysis of (Single-Cell Method) Benchmarks Reveals the Need for Extensibility and Interoperability". In: *Genome Biology* 24.1 (May 17, 2023), p. 119. ISSN: 1474-760X. DOI: 10.1186/s13059-023-02962-5.

[153] Yue Cao, Lijia Yu, Marni Torkel, Sanghyun Kim, Yingxin Lin, Pengyi Yang, Terence P. Speed, Shila Ghazanfar, and Jean Yee Hwa Yang. "The Current Landscape and Emerging Challenges of Benchmarking Single-Cell Methods". Jan. 31, 2025. DOI: 10.1101/2023.12.19.572303. Pre-published.

[154] Monika Jelizarow, Vincent Guillemot, Arthur Tenenhaus, Korbinian Strimmer, and Anne-Laure Boulesteix. "Over-Optimism in Bioinformatics: An Illustration". In: *Bioinformatics* 26.16 (Aug. 15, 2010), pp. 1990–1998. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq323.

[155] Siddhartha Mishra, Nicholas Monath, Michael Boratko, Ariel Kobren, and Andrew McCallum. "An Evaluative Measure of Clustering Methods Incorporating Hyperparameter Sensitivity". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.7 (June 28, 2022), pp. 7788–7796. ISSN: 2374-3468. DOI: 10.1609/aaai.v36i7.20747.

[156] Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. "Benchmarking Atlas-Level Data Integration in Single-Cell Genomics". In: *Nature Methods* 19.1 (1 Jan. 2022), pp. 41–50. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01336-8.

[157] Yue You, Luyi Tian, Shian Su, Xueyi Dong, Jafar S. Jabbari, Peter F. Hickey, and Matthew E. Ritchie. "Benchmarking UMI-based Single-Cell RNA-seq Preprocessing Workflows". In: *Genome Biology* 22.1 (Dec. 14, 2021), p. 339. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02552-3.

[158] Shiquan Sun, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. "Accuracy, Robustness and Scalability of Dimensionality Reduction Methods for Single-Cell RNA-seq Analysis". In: *Genome Biology* 20.1 (Dec. 10, 2019), p. 269. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1898-6.

[159] Felix Raimundo, Celine Vallot, and Jean-Philippe Vert. "Tuning Parameters of Dimensionality Reduction Methods for Single-Cell RNA-seq Analysis". In: *Genome Biology* 21.1 (Aug. 24, 2020), p. 212. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02128-7.

[160] Tim P. Morris, Ian R. White, and Michael J. Crowther. "Using Simulation Studies to Evaluate Statistical Methods". In: *Statistics in Medicine* 38.11 (2019), pp. 2074–2102. ISSN: 1097-0258. DOI: 10.1002/sim.8086.

[161] Yunfei Hu, Manfei Xie, Yikang Li, Mingxing Rao, Wenjun Shen, Can Luo, Haoran Qin, Jihoon Baek, and Xin Maizie Zhou. "Benchmarking Clustering, Alignment, and Integration Methods for Spatial Transcriptomics". In: *Genome Biology* 25.1 (Aug. 9, 2024), p. 212. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03361-0.

[162] Hongrui Duo, Yinghong Li, Yang Lan, Jingxin Tao, Qingxia Yang, Yingxue Xiao, Jing Sun, Lei Li, Xiner Nie, Xiaoxi Zhang, Guizhao Liang, Mingwei Liu, Youjin Hao, and Bo Li. "Systematic Evaluation with Practical Guidelines for Single-Cell and Spatially Resolved Transcriptomics Data Simulation under Multiple Scenarios". In: *Genome Biology* 25.1 (June 3, 2024), p. 145. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03290-y.

[163] Paul P. Gardner, James M. Paterson, Stephanie McGimpsey, Fatemeh Ashari-Ghomi, Sinan U. Umu, Aleksandra Pawlik, Alex Gavryushkin, and Michael A. Black. "Sustained Software Development, Not Number of Citations or Journal Choice, Is Indicative of Accurate Bioinformatic Software". In: *Genome Biology* 23.1 (Feb. 16, 2022), p. 56. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02625-x.

[164] Wei Liu, Xu Liao, Ziye Luo, Yi Yang, Mai Chan Lau, Yuling Jiao, Xingjie Shi, Weiwei Zhai, Hongkai Ji, Joe Yeong, and Jin Liu. "Probabilistic Embedding, Clustering, and Alignment for Integrating Spatial Transcriptomics Data with PRECAST". In: *Nature Communications* 14.1 (1 Jan. 18, 2023), p. 296. ISSN: 2041-1723. DOI: 10.1038/s41467-023-35947-w.

[165] Congcong Hu, Nana Wei, Jiyuan Yang, Hua-Jun Wu, and Xiaoqi Zheng. "STCC: Consensus Clustering Enhances Spatial Domain Detection for Spatial Transcriptomics Data". Feb. 28, 2024. DOI: 10.1101/2024.02.25.581996. Pre-published.

[166] Zhiyuan Yuan. "MENDER: Fast and Scalable Tissue Structure Identification in Spatial Omics Data". In: *Nature Communications* 15.1 (Jan. 5, 2024), p. 207. ISSN: 2041-1723. DOI: 10.1038/s41467-023-44367-9.

[167] Asish Kumar Swain, Vrushali Pandit, Jyoti Sharma, and Pankaj Yadav. "SpatialPrompt: Spatially Aware Scalable and Accurate Tool for Spot Deconvolution and Domain Identification in Spatial Transcriptomics". In: *Communications Biology* 7.1 (May 25, 2024), p. 639. ISSN: 2399-3642. DOI: 10.1038/s42003-024-06349-5.

[168] Yuxuan Hu, Jiazhen Rong, Yafei Xu, Runzhi Xie, Jacqueline Peng, Lin Gao, and Kai Tan. "Unsupervised and Supervised Discovery of Tissue Cellular Neighborhoods from Cell Phenotypes". In: *Nature Methods* (Jan. 8, 2024), pp. 1–12. ISSN: 1548-7105. DOI: 10.1038/s41592-023-02124-2.

[169] Ying Ma and Xiang Zhou. "Accurate and Efficient Integrative Reference-Informed Spatial Domain Detection for Spatial Transcriptomics". In: *Nature Methods* 21.7 (July 2024), pp. 1231–1244. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02284-9.

[170] Hang Xu, Huazhu Fu, Yahui Long, Kok Siong Ang, Raman Sethi, Kelvin Chong, Mengwei Li, Rom Uddamvathanak, Hong Kai Lee, Jingjing Ling, Ao Chen, Ling Shao, Longqi Liu, and Jinmiao Chen. "Unsupervised Spatially Embedded Deep Representation of Spatial Transcriptomics". In: *Genome Medicine* 16.1 (Jan. 12, 2024), p. 12. ISSN: 1756-994X. DOI: 10.1186/s13073-024-01283-x.

[171] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis". In: *Genome Biology* 19.1 (Feb. 6, 2018), p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0.

[172] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. "Spatial Reconstruction of Single-Cell Gene Expression Data". In: *Nature Biotechnology* 33.5 (5 May 2015), pp. 495–502. ISSN: 1546-1696. DOI: 10.1038/nbt.3192.

[173] Jun Du, Yu-Chen Yang, Zhi-Jie An, Ming-Hui Zhang, Xue-Hang Fu, Zou-Fang Huang, Ye Yuan, and Jian Hou. "Advances in Spatial Transcriptomics and Related Data Analysis Strategies". In: *Journal of Translational Medicine* 21.1 (May 18, 2023), p. 330. ISSN: 1479-5876. DOI: 10.1186/s12967-023-04150-2.

[174] Yang Jin, Yuanli Zuo, Gang Li, Wenrong Liu, Yitong Pan, Ting Fan, Xin Fu, Xiaojun Yao, and Yong Peng. "Advances in Spatial Transcriptomics and Its Applications in Cancer Research". In: *Molecular Cancer* 23.1 (June 20, 2024), p. 129. ISSN: 1476-4598. DOI: 10.1186/s12943-024-02040-9.

[175] Lijia Yu, Yue Cao, Jean Y. H. Yang, and Pengyi Yang. "Benchmarking Clustering Algorithms on Estimating the Number of Cell Types from Single-Cell RNA-sequencing Data". In: *Genome Biology* 23.1 (Feb. 8, 2022), p. 49. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02622-0.

[176] Sanjay Jain and Michael T. Eadon. "Spatial Transcriptomics in Health and Disease". In: *Nature Reviews Nephrology* 20.10 (Oct. 2024), pp. 659–671. ISSN: 1759-507X. DOI: 10.1038/s41581-024-00841-1.

[177] Ed Lein, Lars E. Borm, and Sten Linnarsson. "The Promise of Spatial Transcriptomics for Neuroscience in the Era of Molecular Cell Typing". In: *Science* 358.6359 (Oct. 6, 2017), pp. 64–69. DOI: 10.1126/science.aan6827.

[178] Le Zhang, Zhenqi Xiong, and Ming Xiao. "A Review of the Application of Spatial Transcriptomics in Neuroscience". In: *Interdisciplinary Sciences: Computational Life Sciences* 16.2 (June 1, 2024), pp. 243–260. ISSN: 1867-1462. DOI: 10.1007/s12539-024-00603-4.

[179] Ran Zhou, Gaoxia Yang, Yan Zhang, and Yuan Wang. "Spatial Transcriptomics in Development and Disease". In: *Molecular Biomedicine* 4.1 (Oct. 9, 2023), p. 32. ISSN: 2662-8651. DOI: 10.1186/s43556-023-00144-0.

[180] Andreas E Moor and Shalev Itzkovitz. "Spatial Transcriptomics: Paving the Way for Tissue-Level Systems Biology". In: *Current Opinion in Biotechnology*. Systems Biology • Nanobiotechnology 46 (Aug. 1, 2017), pp. 126–133. ISSN: 0958-1669. DOI: 10.1016/j.copbio.2017.02.004.

[181] Zexian Zeng, Yawei Li, Yiming Li, and Yuan Luo. "Statistical and Machine Learning Methods for Spatially Resolved Transcriptomics Data Analysis". In: *Genome Biology* 23.1 (Dec. 2022), p. 83. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02653-7.

[182] Boxiang Liu, Yanjun Li, and Liang Zhang. "Analysis and Visualization of Spatial Transcriptomic Data". In: *Frontiers in Genetics* 12 (2022). ISSN: 1664-8021.

[183] Jessica Gillespie, Maciej Pietrzak, Min-Ae Song, and Dongjun Chung. "A Meta-Review of Spatial Transcriptomics Analysis Software". In: *Cells* 14.14 (July 10, 2025), p. 1060. ISSN: 2073-4409. DOI: 10.3390/cells14141060. PMID: 40710313.

[184] Junbum Kim, Samir Rustam, Juan Miguel Mosquera, Scott H. Randell, Renat Shaykhiev, André F. Rendeiro, and Olivier Elemento. "Unsupervised Discovery of Tissue Architecture in Multiplexed Imaging". In: *Nature Methods* 19.12 (Dec. 2022), pp. 1653–1661. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01657-2.

[185] Jianhua Lin. "Divergence Measures Based on the Shannon Entropy". In: *IEEE Transactions on Information Theory* 37.1 (1991), pp. 145–151. ISSN: 0018-9448. DOI: 10.1109/18.61115.

[186] Yu Wang, Zaiyi Liu, and Xiaoke Ma. "MNMST: Topology of Cell Networks Leverages Identification of Spatial Domains from Spatial Transcriptomics Data". In: *Genome Biology* 25.1 (May 23, 2024), p. 133. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03272-0.

[187] Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. "Deep Graph Infomax". Dec. 21, 2018. DOI: 10.48550/arXiv.1809.10341. arXiv: 1809.10341 [cs, math, stat]. Pre-published.

[188] Honglei Ren, Benjamin L. Walker, Zixuan Cang, and Qing Nie. "Identifying Multicellular Spatiotemporal Organization of Cells with SpaceFlow". In: *Nature Communications* 13.1 (1 July 14, 2022), p. 4076. ISSN: 2041-1723. DOI: 10.1038/s41467-022-31739-w.

[189] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. "Graph Attention Networks". Feb. 4, 2018. DOI: 10.48550/arXiv.1710.10903. arXiv: 1710.10903 [stat]. Pre-published.

[190] Chang Xu, Xiyun Jin, Songren Wei, Pingping Wang, Meng Luo, Zhaochun Xu, Wenyi Yang, Yideng Cai, Lixing Xiao, Xiaoyu Lin, Hongxin Liu, Rui Cheng, Fenglan Pang, Rui Chen, Xi Su, Ying Hu, Guohua Wang, and Qinghua Jiang. "DeepST: Identifying Spatial Domains in Spatial Transcriptomics by Deep Learning". In: *Nucleic Acids Research* 50.22 (Dec. 9, 2022), e131. ISSN: 0305-1048. DOI: 10.1093/nar/gkac901.

[191] Michael E. Tipping and Christopher M. Bishop. "Probabilistic Principal Component Analysis". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3 (Sept. 1, 1999), pp. 611–622. ISSN: 1369-7412. DOI: 10.1111/1467-9868.00196.

[192] Benjamin Chidester, Tianming Zhou, Shahul Alam, and Jian Ma. "SpiceMix Enables Integrative Single-Cell Spatial Modeling of Cell Identity". In: *Nature Genetics* 55.1 (1 Jan. 2023), pp. 78–88. ISSN: 1546-1718. DOI: 10.1038/s41588-022-01256-z.

[193] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. "Integrated Analysis of Multimodal Single-Cell Data". In: *Cell* 184.13 (June 24, 2021), 3573–3587.e29. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2021.04.048. PMID: 34062119.

[194] Zhicheng Xu, Weiwen Wang, Tao Yang, Ling Li, Xizheng Ma, Jing Chen, Jieyu Wang, Yan Huang, Joshua Gould, Huifang Lu, Wensi Du, Sunil Kumar Sahu, Fan Yang, Zhiyong Li, Qingjiang Hu, Cong Hua, Shoujie Hu, Yiqun Liu, Jia Cai, Lijin You, Yong Zhang, YuXiang Li, Wenjun Zeng, Ao Chen, Bo Wang, Longqi Liu, Fengzhen Chen, Kailong Ma, Xun Xu, and Xiaofeng Wei. "STOmicsDB: A Comprehensive Database for Spatial Transcriptomics Data Sharing, Analysis and Visualization". In: *Nucleic Acids Research* 52.D1 (Jan. 5, 2024), pp. D1053–D1061. ISSN: 0305-1048. DOI: 10.1093/nar/gkad933.

[195] Zhen Fan, Runsheng Chen, and Xiaowei Chen. "SpatialDB: A Database for Spatially Resolved Transcriptomes". In: *Nucleic Acids Research* 48.D1 (Jan. 8, 2020), pp. D233–D237. ISSN: 0305-1048. DOI: 10.1093/nar/gkz934.

[196] Yiming Li, Saya Dennis, Meghan R. Hutch, Yanyi Ding, Yadi Zhou, Yawei Li, Maalavika Pillai, Sanaz Ghotbaldini, Mario Alberto Garcia, Mia S. Broad, Chengsheng Mao, Feixiong Cheng, Zexian Zeng, and Yuan Luo. "SOAR Elucidates Disease Mechanisms and Empowers Drug Discovery through Spatial Transcriptomics". Dec. 15, 2023. DOI: 10.1101/2022.04.17.488596. Pre-published.

[197] Meng Zhang, Stephen W. Eichhorn, Brian Zingg, Zizhen Yao, Kaelan Cotter, Hongkui Zeng, Hongwei Dong, and Xiaowei Zhuang. "Spatially Resolved Cell Atlas of the Mouse Primary Motor Cortex by MERFISH". In: *Nature* 598.7879 (Oct. 2021), pp. 137–143. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03705-x.

[198] Hao Xu, Shuyan Wang, Minghao Fang, Songwen Luo, Chunpeng Chen, Siyuan Wan, Rirui Wang, Meifang Tang, Tian Xue, Bin Li, Jun Lin, and Kun Qu. "SPACEL: Deep Learning-Based Characterization of Spatial Transcriptome Architectures". In: *Nature Communications* 14.1 (Nov. 22, 2023), p. 7603. ISSN: 2041-1723. DOI: 10.1038/s41467-023-43220-3.

[199] Jeffrey R. Moffitt, Dhananjay Bambah-Mukku, Stephen W. Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D. Perez, Nimrod D. Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, and Xiaowei Zhuang. "Molecular, Spatial, and Functional Single-Cell Profiling of the Hypothalamic Preoptic Region". In: *Science* 362.6416 (Nov. 16, 2018), eaau5324. DOI: 10.1126/science.aau5324.

[200] Jonah Langlieb, Nina S. Sachdev, Karol S. Balderrama, Naeem M. Nadaf, Mukund Raj, Evan Murray, James T. Webber, Charles Vanderburg, Vahid Gazestani, Daniel Tward, Chris Mezias, Xu Li, Katelyn Flowers, Dylan M. Cable, Tabitha Norton, Partha Mitra, Fei Chen, and Evan Z. Macosko. "The Molecular Cytoarchitecture of the Adult Mouse Brain". In: *Nature* 624.7991 (Dec. 2023), pp. 333–342. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06818-7.

[201] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data Using T-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. ISSN: 1533-7928.

[202] William M. Rand. "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66.336 (1971), pp. 846–850. ISSN: 0162-1459. DOI: 10.2307/2284239. JSTOR: 2284239.

[203] Leslie C. Morey and Alan Agresti. "The Measurement of Classification Agreement: An Adjustment to the Rand Statistic for Chance Agreement". In: *Educational and Psychological Measurement* 44.1 (Mar. 1, 1984), pp. 33–37. ISSN: 0013-1644. DOI: 10.1177/0013164484441003.

[204] Lawrence Hubert and Phipps Arabie. "Comparing Partitions". In: *Journal of Classification* 2.1 (Dec. 1, 1985), pp. 193–218. ISSN: 1432-1343. DOI: 10.1007/BF01908075.

[205] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. "Adjusting for Chance Clustering Comparison Measures". In: *Journal of Machine Learning Research* 17.134 (2016), pp. 1–32.

[206] Alexander Strehl and Joydeep Ghosh. "Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions". In: *Journal of Machine Learning Research* 3 (Dec 2002), pp. 583–617. ISSN: ISSN 1533-7928.

[207] Nguyen Xuan Vinh, Julien Epps, and James Bailey. "Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?" In: (2009).

[208] E. B. Fowlkes and C. L. Mallows. "A Method for Comparing Two Hierarchical Clusterings". In: *Journal of the American Statistical Association* 78.383 (Sept. 1, 1983), pp. 553–569. ISSN: 0162-1459. DOI: 10.1080/01621459.1983.10478008.

[209] Peter J. Rousseeuw. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1, 1987), pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.

[210] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. "Sustainable Data Analysis with Snakemake". In: 10:33 (Apr. 19, 2021). DOI: 10.12688/f1000research.29032.2.

[211] Sandro Vega-Pons and José Ruiz-Shulcloper. "A Survey of Clustering Ensemble Algorithms". In: *International Journal of Pattern Recognition and Artificial Intelligence* 25.03 (May 2011), pp. 337–372. ISSN: 0218-0014. DOI: 10.1142/S0218001411008683.

[212] Vladimir Yu Kiselev, Kristina Kirschner, Michael T. Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N. Natarajan, Wolf Reik, Mauricio Barahona, Anthony R. Green, and Martin Hemberg. "SC3: Consensus Clustering of Single-Cell RNA-seq Data". In: *Nature Methods* 14.5 (May 2017), pp. 483–486. ISSN: 1548-7105. DOI: 10.1038/nmeth.4236.

[213] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[214] Bokai Zhu, Shuxiao Chen, Yunhao Bai, Han Chen, Guanrui Liao, Nilanjan Mukherjee, Gustavo Vazquez, David R. McIlwain, Alexandar Tzankov, Ivan T. Lee, Matthias S. Matter, Yury Goltsev, Zongming Ma, Garry P. Nolan, and Sizun Jiang. "Robust Single-Cell Matching and Multimodal Analysis Using Shared and Distinct Features". In: *Nature Methods* (Jan. 9, 2023), pp. 1–12. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01709-7.

[215] Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun, and Jingyi Jessica Li. "scDesign3 Generates Realistic in Silico Data for Multimodal Single-Cell and Spatial Omics". In: *Nature Biotechnology* (May 11, 2023), pp. 1–6. ISSN: 1546-1696. DOI: 10.1038/s41587-023-01772-1.

[216] Tianyi Sun, Dongyuan Song, Wei Vivian Li, and Jingyi Jessica Li. "scDesign2: A Transparent Simulator That Generates High-Fidelity Single-Cell Gene Expression Count Data with Gene Correlations Captured". In: *Genome Biology* 22.1 (May 25, 2021), p. 163. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02367-2.

[217] Luke Zappia, Belinda Phipson, and Alicia Oshlack. "Splatter: Simulation of Single-Cell RNA Sequencing Data". In: *Genome Biology* 18.1 (Sept. 12, 2017), p. 174. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1305-0.

[218] Helena L. Crowell, Sarah X. Morillo Leonardo, Charlotte Soneson, and Mark D. Robinson. "The Shaky Foundations of Simulating Single-Cell RNA Sequencing Data". In: *Genome Biology* 24.1 (Mar. 29, 2023), p. 62. ISSN: 1474-760X. DOI: 10.1186/s13059-023-02904-1.

[219] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928.

[220] Orhun Aydin, Mark. V. Janikas, Renato Martins Assunção, and Ting-Hwan Lee. "A Quantitative Comparison of Regionalization Methods". In: *International Journal of Geographical Information Science* 35.11 (Nov. 2, 2021), pp. 2287–2315. ISSN: 1365-8816. DOI: 10.1080/13658816.2021.1905819.

[221] Lambda Moses, Pétur Helgi Einarsson, Kayla Jackson, Laura Luebbert, A. Sina Booeshaghi, Sindri Antonsson, Nicolas Bray, Páll Melsted, and Lior Pachter. "Voyager: Exploratory Single-Cell Genomics Data Analysis with Geospatial Statistics". Aug. 20, 2023. DOI: 10.1101/2023.07.20.549945. Pre-published.

[222] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data". In: *Machine Learning* 52.1 (July 1, 2003), pp. 91–118. ISSN: 1573-0565. DOI: 10.1023/A:1023949509487.

[223] Minoru Takemoto, Liqun He, Jenny Norlin, Jaakko Patrakka, Zhijie Xiao, Tatiana Petrova, Cecilia Bondjers, Julia Asp, Elisabet Wallgard, Ying Sun, Tore Samuelsson, Petter Mostad, Samuel Lundin, Naoyuki Miura, Yoshikazu Sado, Kari Alitalo, Susan E Quaggin, Karl Tryggvason, and Christer Betsholtz. "Large-scale Identification of Genes Implicated in Kidney Glomerulus Development and Function". In: *The EMBO Journal* 25.5 (Mar. 8, 2006), pp. 1160–1174. ISSN: 0261-4189. DOI: 10.1038/sj.emboj.7601014.

[224] Markus List, Peter Ebert, and Felipe Albrecht. "Ten Simple Rules for Developing Usable Software in Computational Biology". In: *PLOS Computational Biology* 13.1 (Jan. 5, 2017), e1005265. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005265.

[225] Serghei Mangul, Lana S. Martin, Eleazar Eskin, and Ran Blekhman. "Improving the Usability and Archival Stability of Bioinformatics Software". In: *Genome Biology* 20.1 (Feb. 27, 2019), p. 47. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1649-8.

[226] Anne-Laure Boulesteix, Veronika Stierle, and Alexander Hapfelmeier. "Publication Bias in Methodological Computational Research". In: *Cancer Informatics* 14s5 (Jan. 1, 2015), CIN.S30747. ISSN: 1176-9351. DOI: 10.4137/CIN.S30747.

[227] Fujian Song, Lee Hooper, and Yoon K Loke. "Publication Bias: What Is It? How Do We Measure It? How Do We Avoid It?" In: *Open Access Journal of Clinical Trials* 5 (July 4, 2013), pp. 71–81. ISSN: null. DOI: 10.2147/OAJCT.S34419.

[228] Arielle Marks-Anglin and Yong Chen. "A Historical Review of Publication Bias". In: *Research Synthesis Methods* 11.6 (2020), pp. 725–742. ISSN: 1759-2887. DOI: 10.1002/jrsm.1452.

[229] Anne-Laure Boulesteix, Robert Hable, Sabine Lauer, and Manuel J. A. Eugster. "A Statistical Framework for Hypothesis Testing in Real Data Comparison Studies". In: *The American Statistician* 69.3 (July 3, 2015), pp. 201–212. ISSN: 0003-1305. DOI: 10.1080/00031305.2015.1005128.

[230] Yang Gui, Chao Li, and Yan Xu. "Spatial Domains Identification in Spatial Transcriptomics Using Modality-Aware and Subspace-Enhanced Graph Contrastive Learning". In: *Computational and Structural Biotechnology Journal* 23 (Oct. 22, 2024), pp. 3703–3713. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2024.10.029. PMID: 39507820.