MuSpAn: A Toolbox for Multiscale Spatial Analysis

Joshua A. Bull[1*], Joshua W. Moore[1], Eoghan J. Mulholland[2], Simon J. Leedham[2], Helen M. Byrne[1,3]

[1]Wolfson Centre for Mathematical Biology, Mathematical Institute, University of Oxford, Woodstock Road, Oxford, OX2 6GG, Oxfordshire, UK.
[2]Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7DQ, Oxfordshire, UK.
[3]Ludwig Institute for Cancer Research, University of Oxford, Roosevelt Drive, Oxford, OX3 7DQ, Oxfordshire, UK.

*Corresponding author(s). E-mail(s): bull@maths.ox.ac.uk;
Contributing authors: joshua.moore@maths.ox.ac.uk;

Abstract

The generation of spatial data in biology has been transformed by multiplex imaging and spatial-omics technologies, such as single cell spatial transcriptomics. These approaches permit a detailed mapping of cell populations and phenotypes within the tissue context, which reveals that tissues are complex ecosystems that include multiple organisational structures over different length scales. Quantitative methods for maximising the information that can be retrieved from these images have not kept pace with technological advances in platforms, and no standard methodology has emerged for spatial data analysis. Proposed pipelines are often tailored to individual studies, leading to a fragmented landscape of available methods, and no clear guidance about which statistical tools are best suited to a particular research question.

In response to these challenges, we present MuSpAn, a Multiscale Spatial Analysis package designed to provide straightforward access to well-established and cutting-edge mathematical tools for analysing spatial data. MuSpAn provides easy to use, flexible, and interactive access to quantitative methods developed from mathematical fields that include spatial statistics, topological data analysis, network theory, geometry, probability and ecology. Users can construct custom pipelines from tools across these fields to address specific biological problems, or conduct unbiased exploration of data for discovery spatial biology. In summary, MuSpAn is an extensive platform which enables multiscale analysis of spatial data, ranging from the subcellular to the tissue-scale.

Keywords: Spatial analysis, Spatial biology, Data analysis, Applied mathematics, Statistics, Software

1 Introduction

Coordinated interactions between different cell types produce complex architectural tissue structures that may become disrupted in disease. Recent developments in imaging technology allow this complexity to be mapped in increasing detail [1–3]. Patterns in spatial data span biological length scales, from subcellular features (nm) to individual cells (µm) and whole tissue sections (cm) [4–6]. Multiple imaging methodologies capture this multiscale spatial structure. While the methodology underpinning imaging modalities such as those based on fluorescence *in situ* hybridization (FISH) [7, 8], spatial transcriptomics [9, 10], and spatial proteomics [11–16] may differ, each generates detailed maps of a sample containing multiple markers whose complexity presents substantial challenges for analysis. Challenges relating to data storage, visualisation, object segmentation, and cell phenotyping have been well studied, with community approaches often using advanced AI to improve object detection and annotation [17–28]. Once biological entities within an image have been identified, quantitative methods are required to extract biological insight from these tissue structures.

Existing approaches for analysing spatial data are not standardised and are often limited to spatially-averaged metrics, such as cell counts and densities. These methods overlook spatial information, such as cell position and intercellular interactions, which may unlock important biological insight and maximise the potential of using the tissue context for discovery biology. For example, the exclusion of T lymphocytes to the desmoplastic periphery of a tumour is biologically and prognostically distinct from the same number of immune cells infiltrating through a lesion, where they closely interact with the cancer epithelium [29]. A further challenge is that cell interactions can occur across a range of scales - from proximity-based, such as antigen presentation to an immune cell, through to complex organisational structures, such as tertiary lymphoid aggregates and tissue architectural features like intestinal crypts, lung alveoli, and hepatic lobules. Quantitative methods targeted at one length scale may neglect information at other scales, suggesting that multiple complementary tools should be used to converge upon biological insight [30].

Existing software approaches to spatial analysis often focus on a specific application or imaging modality, applied at a specific length scale, and frequently adopt a 'pipeline' approach in which a fixed set of spatial analysis methods are applied sequentially [31–42]. For example, Schürch *et al.* [31] use a sliding window approach to characterise the composition of local cellular neighbourhoods; the Python software SpottedPy leverages Getis-Ord hotspot analysis and distance measurements to describe Visium data [32]; and the R package SPIAT characterises tissues using methods that include the cross-K function, distance measurements, and cluster identification [33]. Each of these examples offers a different approach to spatial data analysis, and is well-suited to investigate different biological hypotheses. Faced with a new biological context, it can be difficult to compare methods and select appropriate analytical tools, as software packages may be specific for individual imaging modalities, implemented in different coding languages, or describe a single analytical approach. Few platforms enable users to interactively select methods, a notable counterexample being SquidPy, a spatial omics framework which combines efficient data structures with a selection of statistical methods focussed primarily on graph permutations [43, 44]. Applying complementary mathematical methods together can improve the robustness of biological feature detection [30], but the diverse implementations of these methods currently limits their simultaneous application to the same dataset.

We therefore introduce a new platform for **Mu**ltiscale **Sp**atial **An**alysis, MuSpAn. This Python toolbox comprises an extensive collection of mathematical methods for characterising spatial relationships across length scales. Users can select the most suitable methods to address their specific biological questions from a library of cutting-edge mathematical tools. These methods can quantify diverse features including cell-cell contact, spatial correlation, tissue geometry, and topology. MuSpAn enables measurements from one spatial scale to inform analyses at another, expanding the range of potential questions that can be addressed (e.g. How does cell shape vary across a tissue? How are subcellular expression levels of a transcript of interest associated with changes in the wider tissue microenvironment?). By providing the infrastructure to relate precise interrogation of datasets with cutting-edge analysis methods, MuSpAn empowers biologists to ask questions at a level commensurate with the complexity of the data.

2 Results

2.1 MuSpAn provides a framework for robust, quantitative analysis of spatial data at multiple scales

MuSpAn is a Python package that focusses on spatial analysis, the final stage of the data analysis pipeline, after cell segmentation and phenotype assignment (Fig. 1A). It takes as input coordinate information (such as cell centres or boundaries) and metadata in the form of generic labels associated with objects (e.g., cell types, marker intensities, morphometrics, and other metadata). Data can be added using spatial coordinates and labels obtained from any generic segmentation and phenotyping pipeline, or imported from commonly used tools including SpatialData [25, 44], QuPath [17] and Xenium explorer (10x). MuSpAn is supported by detailed documentation and tutorials available from https://www.muspan.co.uk. Rigorous unit testing and continuous development ensure the toolbox can integrate the latest mathematical and statistical methods for quantifying spatial relationships as they are developed.

MuSpAn permits flexible, interactive analysis of data at length scales ranging from the subcellular to the tissue scale. For example, spatial transcriptomics platforms such as Xenium (10x) generate spatial information about subcellular transcripts across whole tissue sections (Fig 1B). At the smallest length scales (up to $\approx$ 10µm), spatial analysis may focus on transcript localisation within a single cell. At intermediate length scales ($\approx$ 100µm), attention may focus on patterns of cell-cell contact, where cell boundaries physically touch. At tissue length scales ($\gtrapprox$ 1000µm), the spatial locations of individual transcripts and cell boundaries may be neglected and attention focused on identifying tissue structures (e.g, colorectal crypts, tumour nests, or tissue stroma) and the boundaries that separate them.

Rather than focus on a particular length scale, MuSpAn stores spatially-resolved data as a series of *objects* (points or shapes) which may represent individual transcripts, cell centres, cell boundaries, and user-defined annotations (Fig 1C). Metadata is associated with objects via *labels*. Objects and their associated metadata are stored within a MuSpAn *domain*, a class with which the user interacts via *queries* that allow them to interactively specify subsets of objects or regions based on spatial cues (e.g., 'points inside a hypoxic region', or 'cell boundaries within 50µm of the tumour invasive margin'), labels (e.g., 'macrophages', 'cells with high expression levels of PDL1', or 'cellular neighbourhoods which are enriched for regulatory T cells'), or other object properties. Spatial quantification can be applied to any subset of objects and within any chosen region of

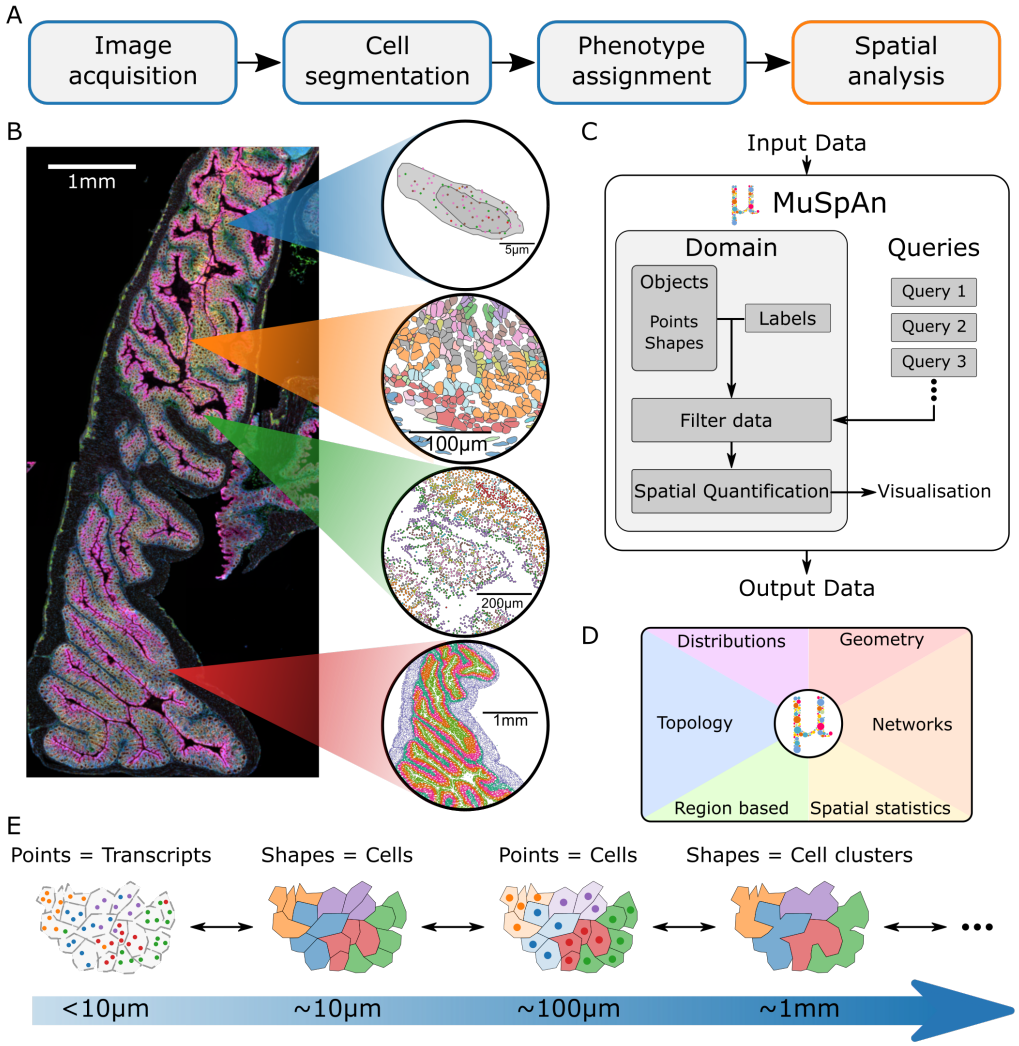

Fig. 1 Overview of MuSpAn.
A: Typical workflow for processing spatial biology imaging data: following cell segmentation and phenotyping, MuSpAn enables extensive spatial analysis of imaging data.
B: A sample image of a healthy mouse colon from the Xenium spatial transcriptomics platform, illustrating spatial structure across multiple length scales. Transcript locations are identified at the subcellular level; at the cellular level, segmentation masks identify physical contact between cell boundaries; collections of cells form higher-order structures such as colorectal crypts; at the whole slide level, different tissue regions can be identified.
C: Schematic showing the internal structure of MuSpAn.
D: Spatial quantification methods within MuSpAn are broadly collected into six modules, based on their mathematical framework.
E: MuSpAn interprets biological objects as points or shapes, and can switch between these interpretations in order to reduce the complexity of data as different length scales of interest are considered. Since each analysis method can be applied to any data of the relevant type (point/shape), regardless of biological interpretation, method selection is agnostic to length scale.

interest, using any method within MuSpAn. Output data are returned for further analysis or data visualisation.

MuSpAn implements a range of mathematical and statistical descriptors of spatial data, broadly divided into six classes (geometry, topology, spatial statistics, networks, region based, and distributions) (Fig. 1D). Each method is agnostic to the spatial scale - they can be applied to objects at different length scales, provided the input data is of the correct type (points or shapes) for the method. Thus, a method for describing spatial correlation between points (for instance, the cross-K function) could be used to describe the spatial correlation between transcripts within a cell, the correlation between cells of different types within a region of interest, or the correlation between locations of cellular neighbourhoods within a whole slide image. This flexibility can be exploited for data exploration, and/or generation of bespoke spatial analysis pipelines to address specific biological hypotheses.

Complex objects can also be converted into simpler ones to facilitate analysis at different length scales (Fig. 1E). Since metadata can be preserved while converting detailed shapes into a centroid, or aggregating multiple points into a single bounding shape, information can be propagated to analyses at different length scales for which less complex representations of the data are more appropriate. For example, quantification of cell morphology may require each cell to be defined by its boundary, while tissue-scale analyses may simply require each cell to be defined by a point at its centroid. For analyses where the location of individual transcripts within a cell is not required, labels containing the transcript composition of the cell may be associated with a single cell boundary shape (obtained by taking the convex hull or an α-shape [45], or by using a segmentation mask).

2.2 Multiscale analysis of spatial transcriptomics data with MuSpAn

We now show how MuSpAn can analyse the same dataset at different length scales. The chosen methods showcase some of the functionality available within MuSpAn, and are not intended as a definitive data analysis pipeline; see Section 2.3 for more details. The methods used in this section are outlined in Table 1, and definitions provided in the supporting references and MuSpAn documentation. We focus on a Xenium sample of healthy mouse colon, and use MuSpAn to describe spatial relationships at length scales ranging from subcellular to whole slide. This dataset provides the spatial locations of segmented cell boundaries and nuclei, with transcript locations of 379 different genes across the tissue. During preprocessing, each cell was assigned one of 19 Celltype IDs, according to its transcripts. See the Methods section for a full description of the dataset.

2.2.1 Spatial analysis of individual cells

We consider a small region of interest (ROI; area $\approx 0.13\text{mm}^2$) from the mouse colon dataset, extracted from a hand-drawn annotation within Xenium Explorer using MuSpAn's Xenium import function (Fig 2A). We focus on ten transcripts: namely, the five strongest indicators of cells with Celltype IDs 1 (Mylk, Myl9, Cnn1, Mgll, Mustn1) and 2 (Ccl9, Sox9, Nupr1, Cldn2, Oit1). At this spatial scale, we view each transcript as a point and identify cell boundaries using the pre-calculated segmentation mask. MuSpAn's `query` functionality enables retrieval of individual cells, their outer boundaries, nuclear boundaries and associated transcripts (Fig. 2B,C). We characterise the cell shape of representative cells of Celltype 1 (Fig. 2B) and 2 (Fig. 2C) using geometric metrics (e.g., perimeter, area, circularity, and principal angle). The distribution of transcripts within each cell is also quantified; we determine the transcript composition within each cell, indicate whether transcripts are cytoplasmic or nuclear, and compute the distribution of distances of cytoplasmic

transcripts from the cell nucleus. As expected, the five transcripts that define Celltype 1 (Mylk, Myl9, Cnn1, Mgll, Mustn1) are elevated in Fig. 2B, while those associated with Celltype 2 (Ccl9, Sox9, Nupr1, Cldn2, Oit1) are elevated in Fig. 2C.

Scalar metrics for each cell can be assigned to the corresponding shape as a new label, making it straightforward to visualise how summary statistics vary across a ROI. For instance, we can create heatmaps in which cell shapes are coloured by each label (Fig. 2D-G). Labels associated with an object may be categorical (Celltype, Fig. 2D) or continuous (perimeter, circularity, principal angle, Fig. 2E-G). The distribution of labels within a ROI can be visualised using histograms and compared using statistical methods (Fig. 2H-K). In this example, while cells of Celltypes 1 and 2 have similar area distributions, Celltype 1 is a stromal cell population of elongated cells (with lower circularity than Celltype 2, Fig. 2J) that have a distinct orientation, indicated by the narrow distribution of principal angles (in contrast to the uniform distribution of principal angles for the epithelial Celltype 2, Fig. 2K). This type of analysis can be useful for quantifying biological properties such as cell polarisation at the interface between a tumour nest and surrounding stroma, or for studying the organisation of fibroblast lamination in tissue fibrosis.

2.2.2 Cell proximity analysis

We now consider a slightly larger ROI annotation (area $\approx 0.38\text{mm}^2$), showing cell boundaries coloured according to the 19 Celltype IDs (Fig. 3A). This representation of the data is suitable for analysing cells in direct contact or close proximity. We consider two cell pairs - Celltypes 1 and 11, and Celltypes 2 and 5. Cells of Celltypes 1 and 11 are found in close proximity; the mean distance from a cell with Celltype 1 to the nearest one with Celltype 11 is 10.9µm (Fig. 3B, C). In contrast, cells of Celltypes 2 and 5 have a mean nearest neighbour distance of 106.0µm (Fig. 3D, E). Using the Average Nearest Neighbour Index [33, 46], these metrics can be converted into z-scores that compare observed values against a statistical null distribution to obtain $z = -13.4$ for Celltypes 1 to 11 (an observed nearest neighbour distance 13.4 standard deviations lower than a null distribution, $p < 10^{-3}$) and $z = 106.4$ for Celltypes 2 to 5 ($p < 10^{-3}$).

We can also use the segmentation mask to identify cell-cell contact. We construct a cell-cell interaction network (Fig. 3F), in which each cell is represented by a node (labelled with the Celltype ID) and nodes are connected if the cell boundaries are within 1µm of one another. Such networks can be analysed using graph theoretical methods, such as the degree distribution of cells of each Celltype (Fig. 3G) and the frequency of contacts between cell types (Fig. 3H, here showing only contacts between Celltypes 1, 2, 5 and 11). This analysis shows that Celltype 1 cells are more isolated than other cell types (mean degree = 2.6), reflecting their localisation in the stroma, at distance from other cells. They most frequently make contact with other Celltype 1 cells, their remaining interactions being primarily with the Celltype 11 cells that line the lamina propria. Similarly, Celltypes 2 and 5 are most likely to be found next to cells of their own type.

A third approach uses region-based metrics to describe cell-cell proximity. Here a lattice of equally-sized 'quadrats' covers the ROI (Fig. 3I), and proximity is determined with respect to a manually specified length scale; here we select a quadrat edge length of 75 µm, ensuring that each quadrat contains sufficient cells to perform a test using the quadrat correlation matrix (QCM, Fig. 3J) [41, 47]. The QCM determines whether each cell-cell combination is more- or less-highly correlated across the distribution of quadrats than would be anticipated under a random reshuffling of labels. A positive standardised effect size (SES) indicates positive correlation between cells (Fig. 3K, Celltypes 1 and 11), while a negative SES indicates exclusion (Fig. 3K, Celltypes 2 and 5).

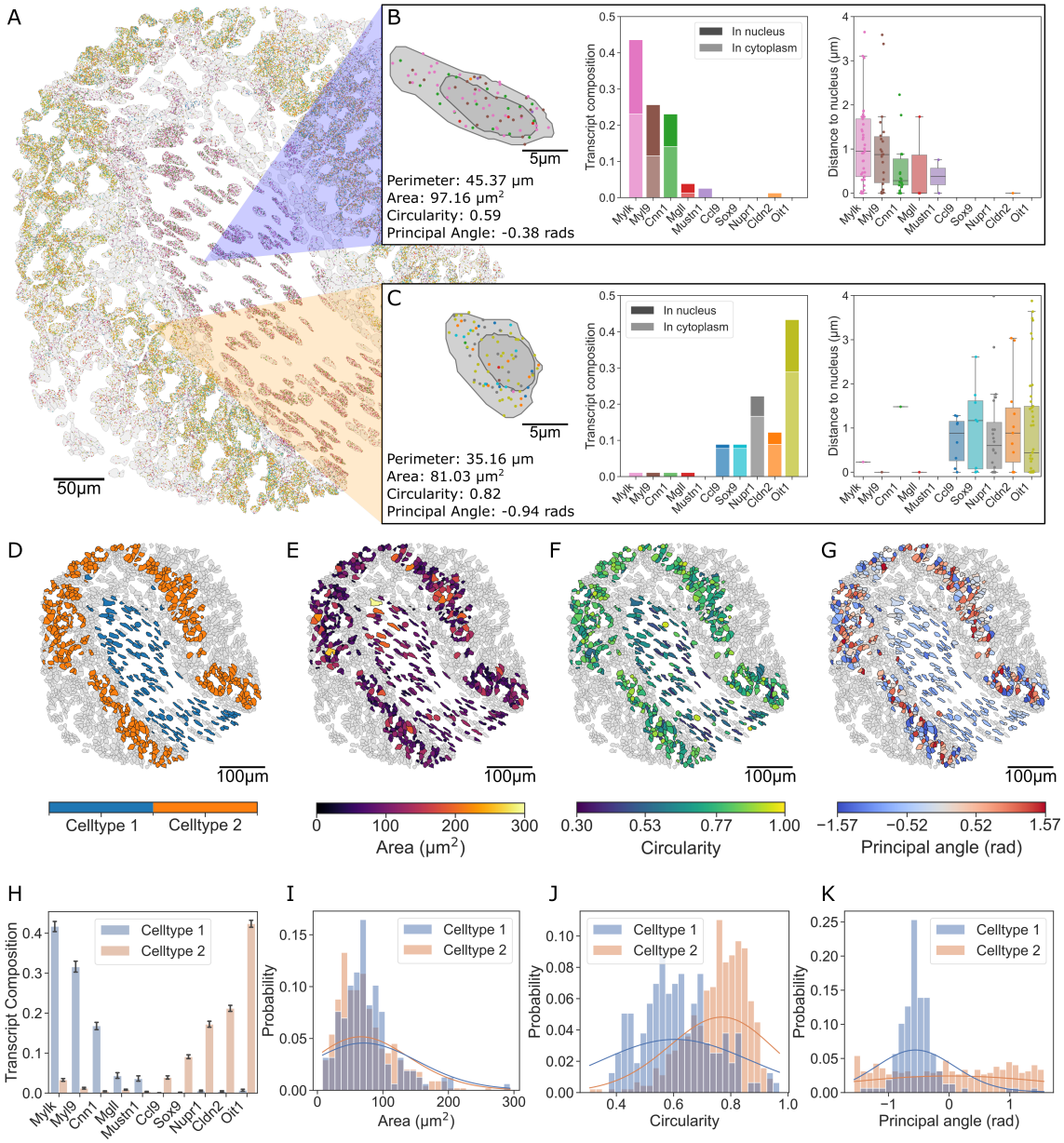

Fig. 2 Quantification at the cell scale.
A: A region of interest (ROI) from the mouse colon image presented in Fig. 1B. Cell boundaries are shown in grey, and coloured points show the top 5 transcripts used to distinguish cells in Celltype 1 (Mylk, Myl9, Cnn1, Mgll, Mustn1) and Celltype 2 (Ccl9, Sox9, Nupr1, Cldn2, Oit1).
B, C: Geometric metrics and transcript summary statistics for representative cells of Celltype 1 (B) and 2 (C) respectively. Shape metrics are calculated from cell boundaries (here perimeter, area, circularity, and principal angle). Summary statistics describe the proportion of transcripts within each cell that are labelled with the 10 markers of interest, their distribution between the nucleus and cytoplasm, and the distance of each cytoplasmic transcript from the edge of the nucleus.
D-G: Morphology metrics, visualised as a heatmap on the cell boundaries for Celltypes 1 and 2. D: Celltype. E: Area. F: Circularity. G: Principal angle.
H-K: Distributions of the same metrics aggregated across the ROI, by Celltype. H: The average proportion of transcripts within each cell reflects those used to define each cell type. I: The distributions of areas are similar for the two cell types ($p = 0.26$, Wilcoxon rank-sum test). J: Cells with Celltype 1 are substantially less circular than those of Celltype 2, despite similar areas ($p < 10^{-3}$, Wilcoxon rank-sum test). K: Celltype 1 cells are more aligned than those of Celltype 2 ($p < 10^{-3}$, Wilcoxon rank-sum test), with a mean principal angle of approximately -0.5 radians, reflecting the orientation of the associated stroma. The distribution of principal angles of Celltype 2 cells is approximately uniform, showing no preferred alignment.

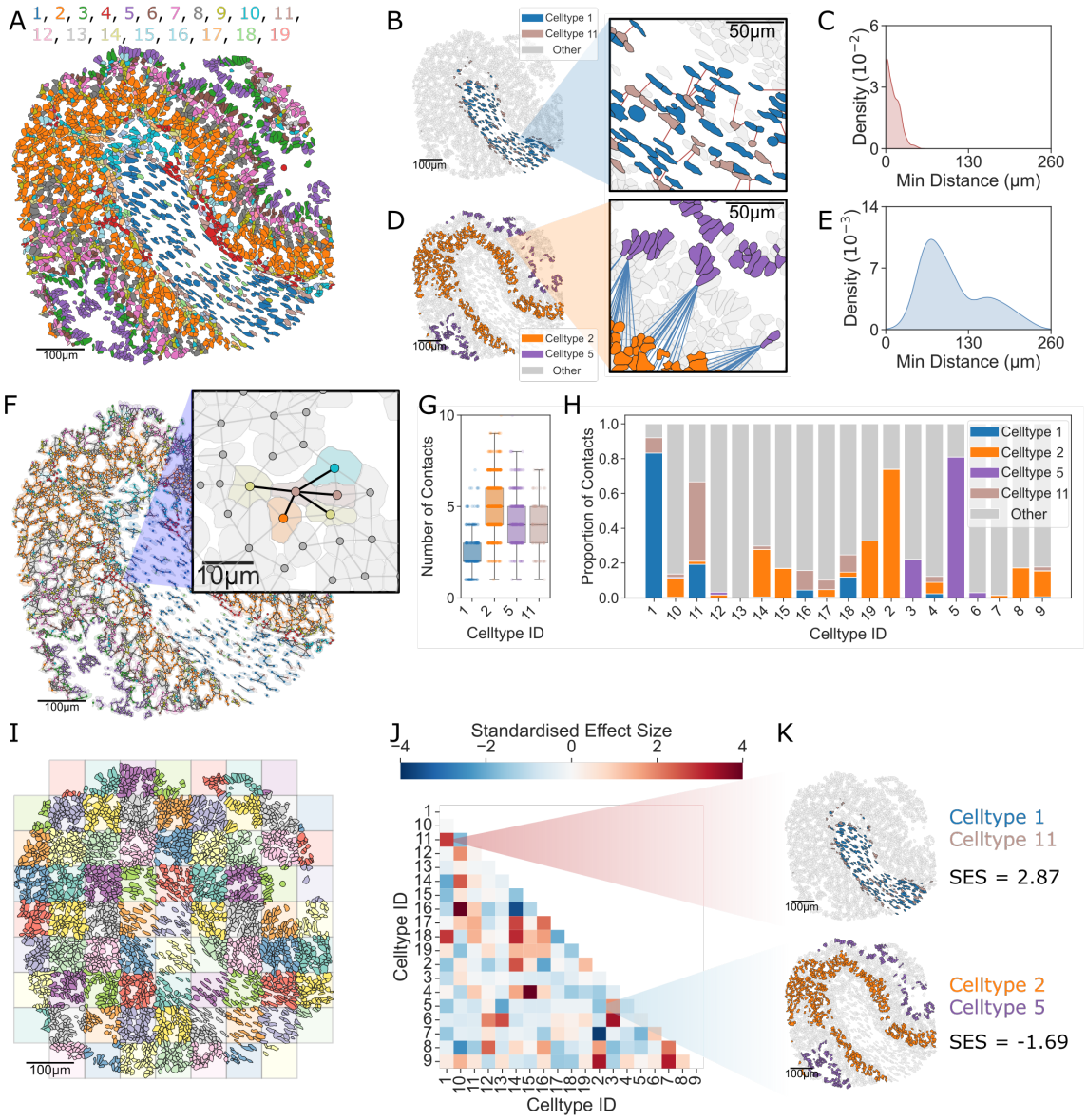

Fig. 3 Quantification of cell-cell proximity.
A: Cells within a ROI, coloured according to their Celltype ID.
B, C: Calculation of nearest neighbour distances from Celltype 1 to Celltype 11 (B) and Celltype 2 to Celltype 5 (C).
D, E: Distribution of nearest neighbour distances for the Celltype pairs considered in B and C. Celltypes 1 and 11 are frequently found in close proximity (mean distance 10.9µm)(D). Celltypes 2 and 5 are rarely found together (mean distance 106.0µm)(E).
F: Proximity network for the ROI in (A). Cell centroids are shown as nodes, and an edge connects two nodes if their cell boundaries are within 1µm.
G: Distribution of the number of neighbours for each node labelled with Celltype 1, 2, 5, and 11 (degree distribution).
H: Average proportion of neighbours of each cell of Celltype X which are Celltype 1, 2, 5, or 11.
I: Division of the ROI into 75µm × 75µm 'quadrats'. Cells are coloured according to the quadrat containing the cell centroid.
J: Quadrat correlation matrix (QCM) quantifying increased/decreased colocalisation of cell pairs (within 75µm length scale), compared to the values expected under uncorrelated cell labelling.
K: Locations of selected cell pairs within the ROI, corresponding to a high SES (Celltypes 1 and 11, SES = 2.87, colocalisation), and a low SES (Celltypes 2 and 5, SES = -1.69, exclusion).

In this section, we have used three different approaches (nearest neighbour, graph based, and QCM) to uncover the same biological relationships between Celltypes 1 and 11 (high cell-cell proximity) and Celltypes 2 and 5 (low cell-cell proximity). Using multiple methods in this way increases the robustness of biological conclusions [30]. Proximity analysis can be important in many biological contexts, such as assessing the effectiveness of immunotherapies at manipulating immune cell populations in cancer treatment.

2.2.3 Analysis of cell-cell relationships across scales

Emergent spatial patterns, such as cell clustering or exclusion, and more complex tissue architecture, may be caused by interactions between cells that are not in physical contact, such as those mediated by production of signalling molecules. Here, we consider a large annotated ROI (Fig. 4A, area $\approx 1.13\text{mm}^2$) and describe each cell using the cell centroid labelled by Celltype ID. We use methods from spatial statistics and topological data analysis (TDA) that are defined using point-like data to analyse cell-cell relationships beyond direct contact.

The cross-pair correlation function (cross-PCF, $g_{A,B}(r)$) is a spatial statistic that determines whether cells of types A and B are positively ($g_{A,B}(r) > 1$) or negatively ($g_{A,B}(r) < 1$) correlated at length scale r when compared with a statistical null distribution. In Fig. 4B, $g_{2,1}(r) < 1$ for $0\mu\text{m} \leq r \leq 80\mu\text{m}$, indicating that cells of Celltypes 1 and 2 are separated by short distances (up to $r \approx 80\mu\text{m}$) less frequently than would be expected if they were randomly distributed. In contrast, $g_{2,2}(r) > 1$ for $0\mu\text{m} \leq r \leq 150\mu\text{m}$, indicating that cells of Celltype 2 are clustered together over length scales $0\mu\text{m} \leq r \leq 150\mu\text{m}$. The cross-PCF can be interpreted quantitatively; for instance, cells of Celltype 2 are approximately 3.8 times more likely to be separated by distances in the range 0-10 µm than would be expected if they were randomly distributed. MuSpAn automatically accounts for edge effects by adjusting contributions to the cross-PCF from the relevant annuli (Fig. 4C,D).

Topological data analysis (TDA) can also identify architectural features across multiple length scales. For example, persistent homology (PH) is a type of TDA which characterises topological features of a set of points. PH considers 'connected components' and 'loops' that are formed and destroyed as the points are gradually connected based on the distances between them (Fig. 4E). Points can be connected using a range of rules; here, we use a method called the Vietoris-Rips (VR) filtration to characterise the topology of the dataset across different 'interaction radii' between the points. For instance, the topological 'loops', or 'voids', in the set of cell centres of Celltype 2 seen at a length scale of 25 µm correspond to individual intestinal crypts (Fig. 4E). PH uses a varying interaction radius to identify the radii at which topological features appear and disappear, with features that persist over a wide range of interaction radii often deemed to be most important. This can be summarised using a 'persistence diagram', which represents each connected component or void as a point depending on the interaction radii at which it is formed ('birth') or disappears ('death') (Fig. 4F). The colorectal crypts in Fig. 4E appear as 'loops' (termed 'H_1 features') that are born at small interaction radii ($r_b \approx 10$) and die at the approximate diameter of a crypt ($r_d \approx 25 - 35\mu\text{m}$). The '$H_0$' features represent distinct 'connected components' within the dataset, and provide a measure of the distribution of average distances between points.

The analyses described in this section can be used in a large number of biological contexts; for example, in quantifying the structural relationship between cells which enclose voids such as intestinal crypts or lung alveoli, or in identifying aggregation or exclusion of immune cells around a tumour nest.

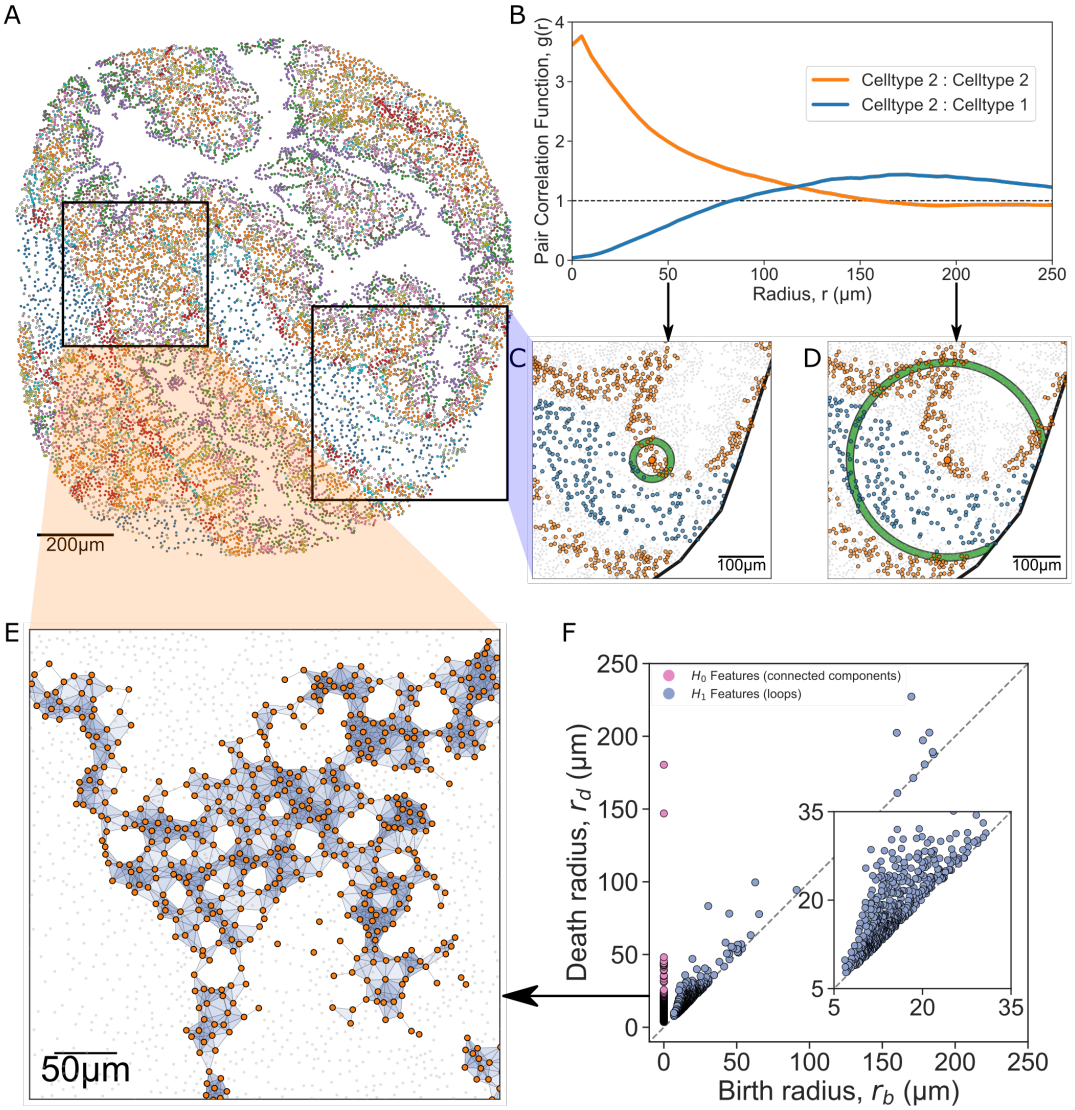

Fig. 4 Quantification of intercellular relationships across length scales.
A: Cell centres within a larger ROI surrounding that in Fig. 3, coloured according to their celltype ID.
B: Cross-PCFs $g_{2,1}(r)$ and $g_{2,2}(r)$. $g_{A,B}(r)$ describes the relative change in pairs of cells separated by distance r with labels Celltype A and Celltype B compared with the number expected under complete spatial randomness. Here, $g_{2,2}(r) > 1$ for $0\mu\text{m} \leq r \leq 150\mu\text{m}$, indicating clusters of Celltype 2 cells with a characteristic width of 150µm. $g_{2,1}(r) < 1$ for $0\mu\text{m} \leq r \leq 80\mu\text{m}$, indicating exclusion between Celltype 1 and 2 cells within approximately 80µm of one another.
C, D: Annuli of radius 50 µm (C) and 200 µm (D) constructed around a representative Celltype 2 cell, used in calculating the cross-PCF. Panel D shows boundary correction within MuSpAn, where contributions of cells whose annuli overlap a boundary are corrected based on the proportion within the analysis boundary.
E: A stage in the construction of the Vietoris-Rips (VR) filtration on cells with Celltype 2 ($r = 25\mu\text{m}$). The VR filtration connects cell centres at distance r or less, and fills any triangles comprising three, pair-wise connected cells. For this value of r, there are many small voids corresponding to the cross-section of colorectal crypts.
F: Representation of the persistent homology (PH) of panel E. across all length scales. 'Birth' and 'Death' refer to the radii at which topological features (here, connected components, H_0; and voids, H_1) appear and disappear from the VR filtration. The cluster of H_1 'voids' born at small radii and with $r_d \approx 25 - 35\mu\text{m}$ (see inset) correspond to the colorectal crypts, and the radius r_d approximates the crypt diameter.

2.2.4 Spatial relationships at tissue scales

MuSpAn can analyse data at length scales much larger than single cells. Here, we show how cell-scale information can be analysed at the tissue-scale using three different approaches.

One approach to categorising large tissue regions is to assign to each cell a 'cellular neighbourhood' [31, 48–51]. We can assign each individual cell in a large tissue region (Fig. 5A) a neighbourhood based on a wide array of methods; in this example, we construct the Delaunay network on the cell centres (with edge distances restricted to a maximum of 30µm) as a proxy for contact. We associate with each cell a vector which records the number of cells of each cell type in its 3-hop neighbourhood (Fig. 5B). Following [31], we use k-means clustering to assign cells with similar cellular neighbourhoods to the same 'microenvironment' (Fig. 5C). Enrichment analysis reveals which cell types define each cluster (Fig. 5D). Mapping these microenvironment IDs (ME IDs) back to the tissue shows large scale spatial structure, which resembles the tissue architecture of the colon (Fig. 5E).

MuSpAn enables us to extend the 'cellular neighbourhood' pipeline by grouping spatially contiguous cells with the same ME ID into distinct shapes (Fig. 5F). These can then be analysed using any shape analysis metric within MuSpAn. Here, we construct a proximity network, representing each shape by a node and connecting nodes whose shapes share a common boundary. We then compute an adjacency significance score for each pair of microenvironments, by comparing the number of MEs that are in contact against a null distribution formed by shuffling the ME IDs. The resulting adjacency network (Fig. 5G) quantifies connectivity between MEs in Fig. 5F, reflecting the ordering of microenvironments from the tissue boundary to its centre (ME2 - ME1 - ME0 - ME4 - ME3).

We can also decompose the tissue into distinct regions, for instance by tiling it with hexagons, and assigning to each tile cell count labels for each Celltype ID (Fig. 5H). We show the normalised cell counts of Celltypes 1, 2 and 11 within 100µm regular hexagonal tiles. Celltypes 1 and 11 are concentrated around the edge of the tissue, and Celltype 2 concentrated centrally. These data can be analysed in numerous ways (e.g., using the QCM; see Fig. 3I-K); here, we identify hotspots of Celltype 2 using the Local Getis-Ord* statistic (Fig. 5I).

An alternative to visualising cell distributions on a lattice is to interpret them as a continuous density. We use kernel density estimation to generate cell density maps (Fig. 5J). These densities can be analysed in a variety of ways. We use the KL-divergence and Wasserstein distance to quantify the similarity of Celltypes 1, 2 and 11 across the sample (Fig. 5K). Low values of these metrics identify Celltypes with similar distributions, and higher values indicate Celltypes with dissimilar distributions.

Exemplar applications of the methods described in this section include defining tissue compartments in multilayered mucosa, or in identifying hotspots of interaction between regulatory and cytotoxic T cells which may contribute to a dampened immune response in many diseases.

2.3 Using MuSpAn for biologically-driven pipeline generation

MuSpAn is designed to facilitate construction of new analysis pipelines that use multiple mathematical methods to test biological hypotheses. While all MuSpAn methods could, in principle, be applied to the same dataset, care is needed to select a subset of methods appropriate for the hypotheses under consideration. Combining such complementary approaches within a pipeline increases confidence in the robustness of any findings [30].

We now illustrate how the convergence of many applicable tools within MuSpAn can be used to analyse three biological scenarios which require different quantitative approaches - quantifying

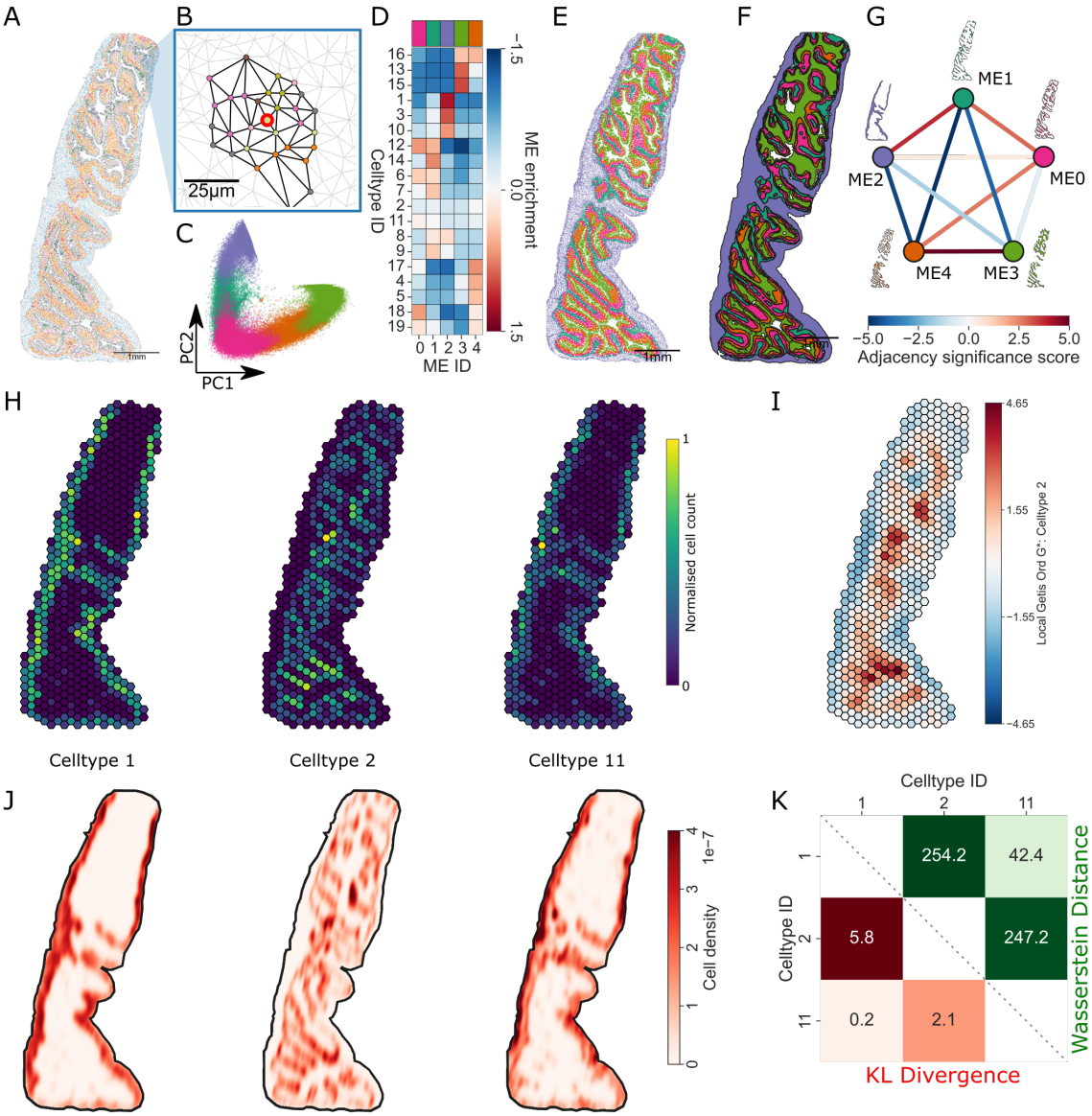

Fig. 5 Quantification of tissue scale features.
A - G: Microenvironment analysis. H - I: Region-based analysis. J - K: Distribution-based analysis.
A: A large section of tissue from the mouse colon dataset, with individual cells coloured according to Celltype ID. B: A 3-hop neighbourhood centred on the cell with the red outline. C: Projection of the Celltype compositions of the 3-hop neighbourhoods onto their first 2 Principal Components, with cells assigned ME IDs via k-means clustering, with $k = 5$. D: Composition of the five MEs, showing which Celltypes are enriched in each ME. E: Individual cell locations within the tissue, coloured by ME ID. F: Shapes formed by grouping spatially contiguous cells with the same ME ID. G: Adjacency analysis quantifies the significance of ME-ME interactions relative to random ME labelling. This graph reflects the ordering of MEs from the centre of the sample to the boundary (ME3 - ME4 - ME0 - ME1 - ME2). H: Region-based distributions of Celltypes 1, 2 and 11 inside a hexagonal lattice with edge length 100µm. Celltypes 1 and 11 are concentrated around the edge of the sample, and Celltype 2 towards the interior. I: Local Getis-Ord* statistic, identifying hotspots of Celltype 2 cells.
J: Kernel density estimates of the Celltypes 1, 2 and 11 distributions. K: The KL-divergence compares the cell densities in panel J (lower triangle, red); it is small for similar distributions (Celltypes 1 and 11) and larger for dissimilar ones (Celltypes 2 and either 1 or 11). The Wasserstein distance also measures the similarity between two distributions (upper triangle, green) and reveals the same trends.

cell-cell contact, characterising spatial cellular neighbourhoods, and describing the shape of tissue architectural features. For each, we select three complementary methods that may inform a robust analysis pipeline, and use the same synthetic dataset to illustrate how they use different aspects of the data. These methods are not intended to be exhaustive, but rather to demonstrate how complementary methods address the same biological question from different perspectives.

2.3.1 Cell-cell contact

We highlight three complementary approaches for characterising cell-cell contact (Fig. 6A). First, adjacency networks provide a natural means to quantify cell-cell contact. They can be generated from the cell segmentation mask (Fig. 3F), or, when cell centres are used, inferred using a Delaunay triangulation. Network edges can be weighted by, for example, cell-cell distance, and cell-cell contact characterised using an adjacency permutation test [52].

A second approach uses the nearest neighbour distribution (Fig. 3B-E). Where network approaches consider all labels simultaneously, nearest neighbour approaches compare pairs of points: here, we show the distance from each green cell to the nearest blue cell (Fig. 6A). The resulting distance distribution can be used to determine whether two cell populations are spatially associated (e.g., using an Average Nearest Neighbour Index [46]).

A Local Indicator of Spatial Association (LISA) can also be used to identify regions where the local correlation between two cell types is higher than expected under spatial randomness [53]. Here, we use the topographical correlation map (TCM) to determine whether green cells are locally positively (red) or negatively (blue) associated with blue cells [54]. MuSpAn incorporates a range of alternative LISAs, including the Local Getis-Ord statistic shown in Fig. 5I.

2.3.2 Neighbourhood characterisation

We highlight three methods that can be used to characterise local neighbourhoods within a ROI (Fig. 6B). Quadrat based approaches tile the ROI into local neighbourhoods, and spatial data is aggregated within each tile. Statistics which can be used to characterise these neighbourhoods include the QCM and the Morisita-Horn index, which measures the similarity of labels between different tiles [55, 56].

Ripley's cross-K function measures the colocalisation of two cell types within a given radius of each other [57]. Extensions that quantify colocalisation of three, or more, cell types within a specified distance of each other have been proposed [54].

Neighbourhood enrichment metrics can be calculated using several approaches (Fig. 5A-E). Here, we use a Voronoi tessellation to tile the domain. Cellular neighbourhoods can then be defined using k-nearest neighbour networks [31].

2.3.3 Tissue architecture

Spatial analysis may be used to identify and characterise architectural structures, such as the circle of blue and orange cells in this example (Fig. 6C). When the edges of a structure can be defined (here by converting cell locations into an α-shape [45]), standard morphological descriptors can be used to quantify its shape (Fig. 2).

When architectural shape boundaries are not specified, cell locations can be used to characterise emergent structures. We construct a Vietoris-Rips complex to show how persistent homology identifies the ring shape in this schematic.

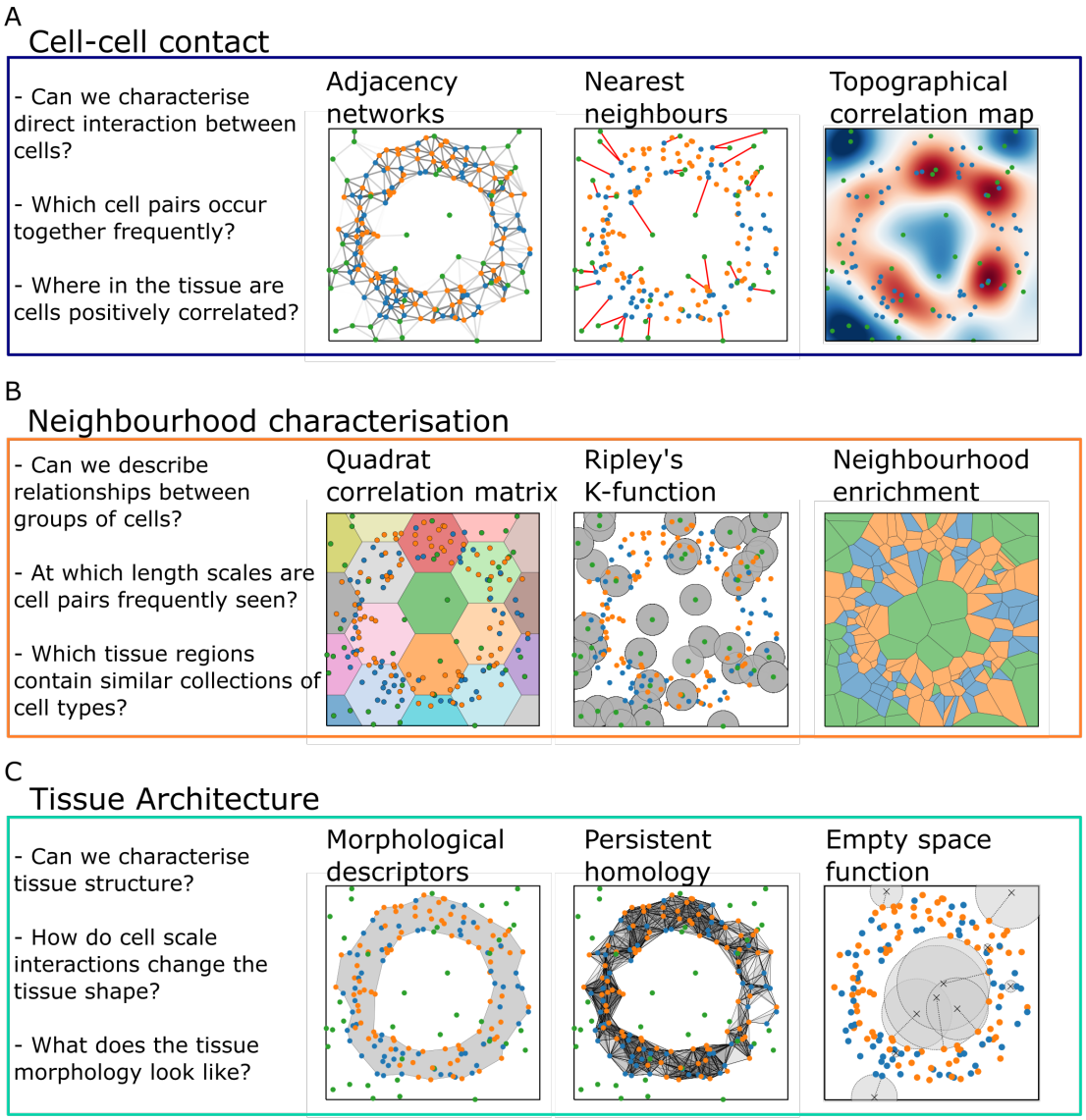

Fig. 6 Biologically-driven pipeline development using MuSpAn
Examples of how complementary mathematical methods can be combined to characterise different biological features. In each example we use the same illustrative dataset to show how the methods use different aspects of the data. This synthetic dataset contains three cell types: blue and orange cells form a circle, and green cells are randomly distributed (see Methods).
A: Cell-cell contact could be described using adjacency networks, nearest neighbour approaches, and topographical correlation maps.
B: Neighbourhood characterisation can be conducted using region-based approaches such as the QCM, spatial statistics like Ripley's K function, and neighbourhood enrichment scores.
C: Tissue architecture can be described using standard morphological descriptors (such as area/perimeter, circularity, convexity), persistent homology, and the empty-space function.

Finally, we use the empty space function to identify the sizes of regions devoid of blue and orange cells within the ROI. This function is similar to the nearest neighbour function (Fig. 3B-E, Fig. 6A), but instead identifies the distance from random locations to the closest cell. The resulting distribution characterises the size of spherical voids within the ROI [58].

3 Discussion

Spatial biology offers the ability to precisely map cell populations, their interactions, and structural arrangements within the complex environments of tissues. By uncovering the patterns of intercellular interaction in different tissues, spatial analysis can become a powerful tool for biological discovery. To facilitate quantitative exploration of such spatial data, we present MuSpAn, a platform for multiscale and multimodal spatial analysis, designed to be compatible with multiplex imaging modalities and spatial-omics technologies. The platform currently implements methods from mathematical fields that include geometry, topology, network theory, spatial statistics, optimal transport and probability. However, it has been engineered to enable easy incorporation of innovations in method development. Using complementary mathematical approaches to address the same biological problems introduces robustness into spatial analysis, and ensures that MuSpAn can be used to reach conclusions about spatial biology which are not sensitive to the limitations of any one specific method [30]. While the examples in this paper focus on single cell spatial transcriptomics technology, MuSpAn can be applied to pre-processed data from multiple modalities, including light microscopy, multiplex IHC, imaging mass cytometry, PhenoCycler, CellDive, and Visium - illustrative analyses can be found in the SI (Appendix C). More generally, data need not originate from cellular biology; MuSpAn can also be applied to spatial data derived from fields such as ecology, astronomy, chemistry and geography, positioning it as an interdisciplinary tool for multiscale spatial analysis.

MuSpAn is supported by extensive online materials that encourage experimentation and exploration. These materials include detailed documentation, installation guides, tutorials, and application case studies (see https://www.muspan.co.uk). As with any analytical toolbox, care is needed to ensure selection of appropriate methods (and parameters) for a particular investigation, and the guides and tutorials help users to make informed decisions. We anticipate that, as demand for spatial quantification increases, standard mathematical methods will emerge as preferred solutions for performing certain analyses. MuSpAn can facilitate this process through the design of pipelines such as those described in Fig. 6.

MuSpAn's infrastructure provides a robust, extensible framework which can be used for low-throughput, detailed data exploration for discovery biology, or for implementing high-throughput analysis pipelines. It is straightforward to use, as users interact only with the MuSpAn domain (and, optionally, queries). Most statistics within MuSpAn can be executed and visualised using a single line of code, reducing the complexity of user scripts, so users can readily access all methods. The `query` infrastructure permits intricate filtering of data for experts, but most users can select common subsets of data using simple syntaxes. These design choices increase software flexibility and encourage experimentation in pipeline development.

MuSpAn's mathematical and statistical methods complement the increasing use of Artificial Intelligence (AI) approaches for data analysis. MuSpAn provides interpretable metrics which can give insight into pre-processed data, while AI approaches commonly make predictions from the image itself. When trained with large datasets, AI approaches can distinguish important and subtle differences between images, but these can be difficult to understand [59]. In contrast, MuSpAn can identify and quantify spatial relationships within a single image, to provide a mechanistic

understanding of spatial patterns. Each of its methods is directly interpretable, linking results to specific aspects of spatial organisation, such as physical contact between cells, cell exclusion from certain tissue regions, and distinctive morphological features. MuSpAn generates output data which can be vectorised and input to machine learning pipelines, an approach that has been used in topological data analysis [60, 61].

There are many open challenges in analysing spatial data and integrating it with multimodal data and wider 'omics' approaches. For example, identifying analytical methods that integrate spatially aggregated data (e.g., scRNAseq) with spatially resolved data remains an open problem in biology and mathematics. MuSpAn is well placed to address these challenges. For example, MuSpAn's metrics could be integrated with patient metadata to understand how spatial environments change as diseases progress. In future work, MuSpAn will be extended to accommodate higher dimensional datasets, e.g., three dimensional spatial data, and longitudinal data which describes dynamic changes in tissue organisation.

In conclusion, standardisation of quantification methods for spatial biology is urgently needed. MuSpAn aims to close the gap between technological advances that have driven the generation of increasingly detailed spatial imaging in biology, and the ability of spatial analysis methods to describe and quantify them. MuSpAn exploits the full range of multiscale and contextual data within biomedical images, and aims to make accessible to biologists a new level of mathematical depth that complements the wider community of tools for exploratory spatial biology [17, 26, 43, 62].

4 Methods

4.1 MuSpAn

4.1.1 Overview

MuSpAn provides a general framework for conducting spatial analyses. In this section, we summarise the structure of MuSpAn, and how this facilitates flexible and targeted spatial analyses which generalise to any imaging modality or data source. We emphasise here that MuSpAn is designed for spatial analysis, and not for preprocessing of raw imaging data. Cell or object identification, region annotation, and labelling should be conducted prior to spatial analysis, using any available pipeline or image analysis software (e.g., Xenium explorer, Qupath, Halo, Deepcell, or similar). MuSpAn can generically import array-like data directly within Python — such as reading CSV files with libraries like Pandas, or using Numpy arrays - and also provides additional 'quality of life' tools to import annotations or labels from popular software like QuPath or Xenium Explorer.

In the following sections, we outline MuSpAn's design, key features, and how it enables highly flexible spatial analysis. While many features are not user-facing, and therefore do not require user understanding, understanding the infrastructure supporting MuSpAn can aid more advanced use.

Overview

All MuSpAn analyses begin with the *domain*, a class which contains data, labels, metadata and more. The domain can be thought of as a container to which data is added, which standardises the internal representation of that data in order for it to be interactively queried or analysed as required by the user. This ensures that analysis methods are not called on incompatible data (e.g., requesting the area of point-like data), that data can easily be filtered as desired for subsequent analysis, and permits straightforward expansion of the spatial analysis toolbox for subsequent software releases.

Spatial data are stored within the domain as *objects*, which represent distinct biological entities as points or shapes - the interpretation of these objects can vary depending on the source of the data. For instance, both transcripts and cell centres are represented as point objects, while cell boundaries, manually specified annotations, and region of interest boundaries are all shape objects. Each object can have metadata associated as *labels*, which can be used to help isolate populations of interest for statistical analysis. By ensuring that all data within a domain is held in a standardised format, the data can be easily interrogated using individual quantitative analysis methods, and the results visualised appropriately within MuSpAn, or returned as raw data for subsequent visualisation or onward use outside of the MuSpAn package.

Domain

A MuSpAn *domain* is a class which standardises the structure of spatial data and metadata, and can be thought of as a container within which objects for analysis are located. Biologically, it represents a single area containing all relevant data for analyses within the same spatial frame of reference, such as a region of interest within an image where all cell (or transcript) locations are represented within the same coordinate system. The domain is enclosed by a continuous closed *boundary*, which defines a simply connected space that fully contains all objects.

When adding new objects to the domain outside of the existing boundary, the boundary of the domain updates by default to be the axis aligned bounding box surrounding all objects; however, the boundary can be easily changed to any non-self intersecting polygon which encloses all objects, such as the convex hull, an alpha shape that surrounds the objects, or an annotation made in external software (for instance, the boundaries of the domains shown in Figs. 2-5 are annotations drawn within Xenium explorer).

The domain also contains helpful metadata, such as colourmaps for visualisation of objects under different criteria (e.g., for colouring objects according to associated labels), or caches to ensure computationally expensive calculations such as distances between objects or network representations of the data can be retained where needed.

Objects

MuSpAn *objects* represent biological entities as either *points* or *shapes*, each assigned a unique Object ID. Points are defined by (x, y) coordinates, while shapes are non-self-intersecting polygons defined by ordered coordinate lists. Shapes can include complex features like internal voids, allowing them to represent diverse regions such as cell boundaries, tumour zones, or areas for exclusion in spatial analyses.

Each object's Object ID links it to metadata, including labels, spatial properties, networks, distances, and other features. Objects can be added or removed from a domain and may store pre-calculated attributes, such as area, perimeter, bounding circles for efficient distance calculations, or relationships to other objects (e.g., parent-child links created through transformations like shape decomposition). The spatial distance between two objects is defined as the minimum Euclidean distance, accommodating both point-like and shape-like objects. For shapes, distances can be calculated using the centroid or via the nearest part of the relevant line segment in the shape boundary.

Objects can belong to one or more collections, which are named subsets of objects (e.g., 'cell boundaries', 'transcripts', or 'damaged tissue'). A collection denotes a subset of objects, and is one way of referring to multiple objects at once. Collections have no restrictions on names or overlapping membership.

Labels

The main mechanism by which MuSpAn associates metadata with an object is through the use of *labels*. Each object may be assigned any number of different labels, which may be categorical (discrete) or numeric (continuous). For example, in Fig. 2D-G, cells are shape objects that have been assigned a categorical label 'Celltype 1' or 'Celltype 2', as well as numerical labels 'Area', 'Principal angle' and 'Circularity'. Labels can be assigned to some subset of objects (for instance, all objects in one collection), either based on metadata obtained from outside of MuSpAn (e.g., labelling cells from multiplex immunohistochemistry according to their mean pixel intensity of a marker of interest) or from within MuSpAn (e.g., labelling cells according to the result of some spatial analysis of interest). In Fig. 2D-G, the celltype label is based on the preprocessing steps conducted in Xenium explorer, while the other labels are derived from geometric summary statistics calculated for each cell within MuSpAn and subsequently added to the shapes. Labels can be used for analysis or visualisation, or as part of the calculation of other spatial metrics.

Moving between spatial scales

As indicated in Fig. 1D, a key way of moving between spatial scales in MuSpAn is by defining new objects based on existing ones. Starting from the lowest spatial scale added to the domain (in our examples, point objects representing transcript locations), we can move up spatial scales by combining transcript locations into a single cell object (using the segmentation mask if available, or by finding, for instance, the convex hull or α-shape of the points). Converting a shape into a single point (e.g., turning a cell boundary into a cell centroid) also moves up spatial scales by 'simplifying' a collection of more complex spatial information into a single object. These processes effectively lower the resolution of the object or objects in question (moving from a set of points to a single shape, or from a detailed shape boundary to a single centroid), which can then be considered in the context of other objects at the same spatial scale without becoming computationally infeasible.

Data can be transferred across spatial scales through the use of labels, which allow a 'coarse grained' object to contain information about the 'fine grained' objects from which is has been derived - for instance, a cell centroid could be labelled according to the circularity or area of the parent cell boundary object. This allows features computed at one scale to be inherited by the new 'child' objects derived from the 'parent'. Parent-child relationships are stored within the domain as new objects are created, which allows easy construction of ancestry trees for multiscale information retrieval.

Queries

MuSpAn provides powerful query tools to filter and analyse objects based on specific properties, enabling both straightforward and advanced data interrogation. While most users benefit from this infrastructure automatically, advanced users can directly create and use queries to extract complex data subsets using the `query` module. A query in MuSpAn specifies criteria for selecting objects. It comprises three components:

1. a property of interest (e.g., the label 'Celltype');
2. a relation (e.g., 'is equal to');
3. a comparison value (e.g., 'T cell').

These components would generate a query that returns any objects with a 'Celltype' label equal to 'T cell'. This query is evaluated at runtime to produce a boolean array (i.e., True for matching

objects, False otherwise). Once created, queries can be reused across various functions, including spatial analysis and visualisation tools.

Users can construct queries explicitly or rely on MuSpAn's intuitive syntax for common cases. For example, passing the tuple ('Celltype', 'T cell') to a function will automatically lead to its interpretation as a query relating to the 'Celltype' label. Advanced queries can combine multiple criteria using query containers, which support logical operations (e.g., AND, OR). These containers allow nesting of conditions, enabling highly detailed filtering. For example, a query container might select all T cells within 20 µm of a tumour boundary, or even more complex combinations of spatial relationships, label values, and even summary statistics.

Queries persist in the Python workspace, allowing consistent use across analysis and visualisation methods. While beginner users can rely on MuSpAn's default behaviour and 'behind the scenes' use of the query infrastructure to conduct basic analyses, leveraging the query system provides a powerful way to explore and analyse data efficiently. Tutorials and examples of advanced queries are provided on the MuSpAn website.

Visualisation

While not primarily a visualisation package, MuSpAn leverages Python's matplotlib package [63] to provide easy access to data visualisation tools through MuSpAn's `visualise` module. In particular, this permits straightforward visualisation of the domain and objects within it, coloured according to labels, collections, or other properties of interest. Domain visualisation accepts query arguments to specify the objects to plot or the metadata used to colour the output, permitting easy publication-ready plotting of complex subsets of objects; almost all the panels of Figs. 2-6 are generated using the visualisation tools available within MuSpAn, and detailed tutorials showing how this is achieved are available on the MuSpAn website.

MuSpAn can also be used to natively plot the outputs of selected spatial analysis methods. Every MuSpAn spatial analysis method returns the output as raw data that can be passed by the user to other analysis or visualisation tools; however, this data can also be passed to dedicated functions within the `visualise` module to straightforwardly generate plots. These functions can also be called automatically using the `visualise_output` keyword argument associated with most spatial analysis methods. While the `visualise` module provides customisation via matplotlib, it is generally recommended that users use raw outputs for creating custom visualisation.

4.1.2 Quantitative analyses

Central to MuSpAn are the quantitative analyses it provides. All of the methods available within MuSpAn v1.0.0 have been previously described in the academic literature, and so are not described in detail here; a complete list of implemented methods, with references to their use in spatial biology where possible, can be found in the supporting information. Where possible, MuSpAn leverages implementations of individual methods from established discipline-specific Python packages [64–66].

The spatial analysis methods provided within MuSpAn are divided between six submodules, named according to the underlying mathematical framework within which the analysis methods are contained. These modules are:

- **Distribution** - containing methods through which spatial data is considered as a continuous distribution, such as a density function.
- **Geometry** - containing methods that describe the morphology or other characteristics of shapes.

- **Networks** - containing methods through which data is represented as spatially embedded networks.
- **Region Based** - containing methods which represent data within spatially-averaged subregions of the region of interest.
- **Spatial Statistics** - containing methods from the field of spatial statistics, which compare observations of point data against statistical null models.
- **Topology** - containing methods from the field of topological data analysis, which describes data in terms of topological invariants.

We reiterate here that each method within these modules is designed to investigate different types of spatial relationships. As such, techniques from any module can be applied at any length scale. To generate pipelines as described in Fig. 6 that are robust to potential error introduced by viewing data through the lens of one particular methodology, we suggest that methods from across multiple submodules should be employed where possible.

Specifying populations

Each of the functions available within MuSpAn for spatial analysis accepts different classes of objects; for example, spatial statistics such as the cross-PCF require two point populations to be defined, while a geometric statistic such as convexity is only defined for shape-like (polygonal) objects. Each statistic allows the population (or populations) to be specified when calculated, for instance by using the MuSpAn `query` module described above. This permits subsets of the data to be isolated and analysed very easily, and for the same population to be passed to multiple functions in series, making pipeline construction simple.

Specifying analysis boundaries

Many statistics are dependent not only on the population of objects they consider, but also on the boundaries of the analysis. As a simple (non spatial) example, the density of a fixed number of cells varies depending on the size of the region in which they are measured. MuSpAn permits analysis boundaries to be specified via queries in the same way as populations are. Users can specify a set of shapes that define the boundaries of a calculation region, or of shapes defining regions that should be excluded from analysis. MuSpAn converts these requests on-the-fly into simply connected spaces for analysis, and where appropriate adjusts statistical analyses to account for boundary corrections. Where no boundary is specified, analyses will use the domain boundary. See Appendix C in the SI for examples of spatial analysis within shape boundaries.

4.2 Datasets

Xenium mouse colon tissue

Spatial analysis was conducted on mouse colon tissue stained with the Xenium Mouse Tissue Atlas Panel (10x Genomics, USA), using dataset 'Fresh Frozen Mouse Colon with Xenium Multimodal Cell Segmentation' sourced from the 10x Genomics website and available under a CC BY license (available from https://www.10xgenomics.com/datasets/fresh-frozen-mouse-colon-with-xenium-multimodal-cell-segmentation-1-standard). Fresh-frozen tissue arrays were prepared from 8-week-old male C57 mice (Charles River Laboratories, USA). Tissue samples were processed according to the Xenium protocols for fresh-frozen tissues, following the Xenium In Situ for Fresh Frozen Tissues - Tissue Preparation Guide (CG000579) and

Fixation & Permeabilisation Guide (CG000581). Staining procedures involved probe hybridisation, washing, ligation, amplification, and cell segmentation staining, carried out in alignment with the Xenium In Situ Gene Expression with Cell Segmentation Staining User Guide (CG000749). Post-instrument processing, including hematoxylin and eosin (H&E) staining, followed the Xenium In Situ Gene Expression - Post-Xenium Analyser H&E Staining Protocol (CG000613). The Xenium Mouse Tissue Atlas Panel, pre-designed by 10x Genomics, includes 379 genes selected to represent major cell types and optimised for spatial gene expression profiling in mouse tissues.

Cell type annotation was conducted by Xenium, assigning 25 unique 'Cluster IDs' to the cells using the 'Graph-based' method (Leiden clustering). Only Cluster IDs 1-19 were considered in our demonstrative analysis due to low observations of cells with Cluster IDs 20-25. We have renamed 'Cluster ID' to 'Celltype ID' throughout this paper to avoid potential confusion with spatial clustering.

Synthetic example dataset

The synthetic dataset used in Fig. 6 is available within MuSpAn as the `'Synthetic-Points-Architecture'` dataset, and can be easily loaded via the command `muspan.datasets.load_example_domain('Synthetic-Points-Architecture')`. The panels shown represent a subset of the full dataset centred around a ring of orange and blue points, with green points randomly placed throughout the domain; for full details on the dataset, see Figure 1 of [30]; code for generating the panels shown is available alongside this paper.

4.3 Spatial analysis methods overview

MuSpAn v1.0.0 includes a diverse collection of 40+ quantitative analysis methods which have been described in the scientific literature. Table 1 summarises the spatial methods used in the exemplar analyses shown in Figures 2–6, and provides references containing detailed methodological descriptions (for additional information see the SI, Appendix B). A complete list of all spatial analysis methods implemented in MuSpAn v1.0.0 is provided in the SI in Table A1. In addition, detailed information on all 150+ MuSpAn functions can be found at https://docs.muspan.co.uk.

Table 1: Overview of spatial analysis methods featured in Figures 2–6. We indicate each method's module within the MuSpAn package and provide a concise description. For full details on the definitions of these methods, please see the references provided. References for standard methods are omitted (denoted by '-').

Figure	**Method** (`MuSpAn module`)	**Summary**	**Ref**
2	Circularity (`geometry`)	A measure of how similar the shape of a polygon is to a circle, with values ≈ 1 indicating shapes that are closer to perfectly round.	-
2	Principal angle (`geometry`)	The angle of the longest line through the center of a polygon that best represents its overall orientation, often used to describe a shape's direction or alignment.	[67]

2	Label composition (`summary_statistics`)	The count of a specific label category, normalised relative to the total number of labels in the region.	-
3	Average nearest neighbour index (ANNI) (`spatial_statistics`)	A statistical measure used to determine whether a spatial distribution of points is clustered, random, or dispersed by comparing observed and expected nearest neighbour distances.	[46]
3	Proximity network (`networks`)	A network that connects nodes based on their spatial distance.	[68]
3	Quadrat correlation matrix (QCM) (`spatial_statistics`)	A measure of correlation between the counts of different cell types within a set of regions.	[30, 42] [41, 47]
4	Cross pair correlation function (cross-PCF) (`spatial_statistics`)	A measure of spatial aggregation between pairs of objects across length scales.	[54, 58] [69, 70]
4	Vietoris-Rips filtration (`topology`)	A method in topological data analysis that constructs a nested sequence of simplicial complexes by connecting points within increasing distances, capturing the shape and features of a dataset across scales.	[64]
5	Delaunay network (`networks`)	A network where nodes in a plane are connected to form triangles such that no point lies inside the circumcircle of any triangle.	[68]
5	k-hop neighbourhood (`networks`)	The set of all nodes that can be reached from a given node within k steps or connections.	-
5	Neighbourhood clustering (`networks`)	Group similar neighbourhoods together based on the enrichment of labels within the neighbourhoods.	[31, 40]
5	Adjacency permutation test (`networks`)	A statistical method that assesses the significance of node label adjacency by randomly permuting node labels and comparing the observed property to the distribution from the permutations.	[42, 52]
5	α-shape (`helpers`)	A method to define the shape of a set of points in space by connecting points that are close to each other, using a parameter α to determine the length scale captured.	[42, 52]
5	Hexgrid (`region_based`)	Generate a mesh of regular hexagons covering the requested space.	-

5	Local Getis-Ord* (`spatial_statistics`)	A statistic that identifies spatial clusters by measuring the degree of local spatial association for a variable, highlighting areas of high or low values relative to their surroundings.	[32, 39] [71]
5	Kernel density estimation (KDE) (`distribution`)	A non-parametric statistical method for estimating the probability density function of a random variable by smoothing data points using a kernel function.	[66]
5	Wasserstein distance (`distribution`)	A metric for comparing probability distributions by measuring the optimal transport cost required to move one distribution to another.	[30, 35]
5	KL-divergence (`distribution`)	A measure of the difference between two probability distributions by quantifying the amount of information lost when approximating one distribution with another.	[72]
6	Topographical correlation map (TCM) (`spatial_statistics`)	A Local Indicator of Spatial Association (LISA) which quantifies the pairwise correlation between two populations of points and is visualised locally as a heatmap.	[41, 54]
6	Ripley's K-function (`spatial_statistics`)	A measure of pairwise spatial correlation between two populations of points, based on comparing the observed numbers of points within a minimum pairwise distance of one another.	[57]
6	Empty space function (`spatial_statistics`)	A function which characterises empty spaces by repeatedly sampling random locations and measuring the distance to the nearest point.	[58, 73]

5 Code Availability

MuSpAn is available free for non-commercial use via the MuSpAn website (https://www.muspan.co.uk/get-the-code) under an Academic Use Licence. Commercial users should contact Oxford University Innovation via their software store (https://process.innovation.ox.ac.uk/software/p/21795/multiscale-spatial-analysis-toolbox/1). Full user guides including installation, documentation, and detailed tutorials are available via https://www.muspan.co.uk.

In addition, scripts which reproduce the Figures in this paper can be found at https://github.com/joshwillmoore1/Supporting_material_muspan_paper. A public Github repository which can be used for issues tracking, feature suggestion, and MuSpAn discussion, is available at https://github.com/joshwillmoore1/MuSpAn-Public.

6 Data Availability

The 'healthy mouse colon' dataset is available from 10x Genomics (https://www.10xgenomics.com/datasets/fresh-frozen-mouse-colon-with-xenium-multimodal-cell-segmentation-1-standard). The synthetic dataset used in Fig. 6 is available within MuSpAn as the 'Synthetic-Points-Architecture' dataset, and can be easily loaded via the command `muspan.datasets.load_example_domain('Synthetic-Points-Architecture')`. All MuSpAn domains used for analysis in this paper can be found at https://github.com/joshwillmoore1/Supporting_material_muspan_paper.

References

[1] Bressan, D., Battistoni, G. & Hannon, G. J. The dawn of spatial omics. *Science* **381**, eabq4964 (2023). URL https://www.science.org/doi/full/10.1126/science.abq4964. Publisher: American Association for the Advancement of Science.

[2] Moffitt, J. R., Lundberg, E. & Heyn, H. The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics* **23**, 741–759 (2022). URL https://www.nature.com/articles/s41576-022-00515-3. Publisher: Nature Publishing Group.

[3] Merritt, C. R. *et al.* Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nature Biotechnology* **38**, 586–599 (2020). URL https://www.nature.com/articles/s41587-020-0472-9. Publisher: Nature Publishing Group.

[4] Jacquemet, G., Carisey, A. F., Hamidi, H., Henriques, R. & Leterrier, C. The cell biologist's guide to super-resolution microscopy. *Journal of Cell Science* **133**, jcs240713 (2020). URL https://doi.org/10.1242/jcs.240713.

[5] Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics* **24**, 494–515 (2023). URL https://www.nature.com/articles/s41576-023-00580-2. Publisher: Nature Publishing Group.

[6] Alon, S. *et al.* Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science* **371**, eaax2656 (2021). URL https://www.science.org/doi/full/10.1126/science.aax2656. Publisher: American Association for the Advancement of Science.

[7] Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641–1659.e26 (2020). URL https://www.cell.com/cell/abstract/S0092-8674(20)30940-5. Publisher: Elsevier.

[8] Takei, Y. *et al.* Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590**, 344–350 (2021). URL https://www.nature.com/articles/s41586-020-03126-2. Publisher: Nature Publishing Group.

[9] Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021). URL https://www.nature.com/articles/s41586-021-03634-9. Publisher: Nature Publishing Group.

[10] Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016). URL https://www.science.org/doi/full/10.1126/science.aaf2403. Publisher: American Association for the Advancement of Science.

[11] Method of the Year 2024: spatial proteomics. *Nature Methods* **21**, 2195–2196 (2024). URL https://www.nature.com/articles/s41592-024-02565-3. Publisher: Nature Publishing Group.

[12] Hickey, J. W. *et al.* Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. *Nature Methods* **19**, 284–295 (2022). URL https://www.nature.com/articles/s41592-021-01316-y. Publisher: Nature Publishing Group.

[13] Gerdes, M. J. *et al.* Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences* **110**, 11982–11987 (2013). URL https://www.pnas.org/doi/abs/10.1073/pnas.1300136110. Publisher: Proceedings of the National Academy of Sciences.

[14] Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968–981.e15 (2018). URL https://www.sciencedirect.com/science/article/pii/S0092867418309048.

[15] Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nature Medicine* **20**, 436–442 (2014). URL https://www.nature.com/articles/nm.3488. Publisher: Nature Publishing Group.

[16] Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods* **11**, 417–422 (2014). URL https://www.nature.com/articles/nmeth.2869. Publisher: Nature Publishing Group.

[17] Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Scientific Reports* **7**, 16878 (2017). URL https://www.nature.com/articles/s41598-017-17204-5. Publisher: Nature Publishing Group.

[18] Valen, D. A. V. *et al.* Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. *PLOS Computational Biology* **12**, e1005177 (2016). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005177. Publisher: Public Library of Science.

[19] Greenwald, N. F. *et al.* Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology* **40**, 555–565 (2022). URL https://www.nature.com/articles/s41587-021-01094-0. Publisher: Nature Publishing Group.

[20] Israel, U. *et al.* A Foundation Model for Cell Segmentation (2024). URL https://www.biorxiv.org/content/10.1101/2023.11.17.567630v3. Pages: 2023.11.17.567630 Section: New Results.

[21] Berg, S. *et al.* ilastik: interactive machine learning for (bio)image analysis. *Nature Methods* **16**, 1226–1232 (2019). URL https://www.nature.com/articles/s41592-019-0582-9. Publisher: Nature Publishing Group.

[22] Schapiro, D. *et al.* histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods* **14**, 873–876 (2017). URL https://www.nature.com/articles/nmeth.4391. Publisher: Nature Publishing Group.

[23] Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022). URL https://www.science.org/doi/10.1126/science.abl5197. Publisher: American Association for the Advancement of Science.

[24] Yuan, Z. *et al.* Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods* **21**, 712–722 (2024). URL https://www.nature.com/articles/s41592-024-02215-8. Publisher: Nature Publishing Group.

[25] Marconato, L. *et al.* SpatialData: an open and universal data framework for spatial omics. *Nature Methods* 1–5 (2024). URL https://www.nature.com/articles/s41592-024-02212-x. Publisher: Nature Publishing Group.

[26] Chiu, C.-L., Clack, N. & the napari community. napari: a Python Multi-Dimensional Image Viewer Platform for the Research Community. *Microscopy and Microanalysis* **28**, 1576–1577 (2022). URL https://doi.org/10.1017/S1431927622006328.

[27] Zhao, T. *et al.* A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods* 1–11 (2024). URL https://www.nature.com/articles/s41592-024-02499-w. Publisher: Nature Publishing Group.

[28] Goldsborough, T. *et al.* InstanSeg: an embedding-based instance segmentation algorithm optimized for accurate, efficient and portable cell segmentation (2024). URL http://arxiv.org/abs/2408.15954. ArXiv:2408.15954.

[29] Bruni, D., Angell, H. K. & Galon, J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nature Reviews Cancer* **20**, 662–680 (2020). URL https://www.nature.com/articles/s41568-020-0285-7. Publisher: Nature Publishing Group.

[30] Bull, J. A. *et al.* Integrating diverse statistical methods to analyse stage-discriminatory cell interactions in colorectal neoplasia (2024). URL https://www.biorxiv.org/content/10.1101/2024.06.02.597010v1. Pages: 2024.06.02.597010 Section: New Results.

[31] Schürch, C. M. *et al.* Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell* **182**, 1341–1359.e19 (2020). URL https://www.cell.com/cell/abstract/S0092-8674(20)30870-9. Publisher: Elsevier.

[32] Withnell, E. & Secrier, M. SpottedPy quantifies relationships between spatial transcriptomic hotspots and uncovers environmental cues of epithelial-mesenchymal plasticity in breast cancer. *Genome Biology* **25**, 289 (2024). URL https://doi.org/10.1186/s13059-024-03428-y.

[33] Feng, Y. *et al.* Spatial analysis with SPIAT and spaSim to characterize and simulate tissue microenvironments. *Nature Communications* **14**, 2697 (2023). URL https://www.nature.com/articles/s41467-023-37822-0. Publisher: Nature Publishing Group.

[34] Peters Couto, B. Z., Robertson, N., Patrick, E. & Ghazanfar, S. MoleculeExperiment enables consistent infrastructure for molecule-resolved spatial omics data in bioconductor. *Bioinformatics* **39**, btad550 (2023). URL https://doi.org/10.1093/bioinformatics/btad550.

[35] Mah, C. K. *et al.* Bento: a toolkit for subcellular analysis of spatial transcriptomics data. *Genome Biology* **25**, 82 (2024). URL https://doi.org/10.1186/s13059-024-03217-7.

[36] Ali, M. *et al.* GraphCompass: Spatial metrics for differential analyses of cell organization across conditions (2024). URL http://biorxiv.org/lookup/doi/10.1101/2024.02.02.578605.

[37] Nirmal, A. J. & Sorger, P. K. SCIMAP: A Python Toolkit for Integrated Spatial Analysis of Multiplexed Imaging Data (2024). URL http://arxiv.org/abs/2405.02076. ArXiv:2405.02076 [q-bio].

[38] Sethi, R. *et al.* ezSingleCell: an integrated one-stop single-cell and spatial omics analysis platform for bench scientists. *Nature Communications* **15**, 5600 (2024). URL https://www.nature.com/articles/s41467-024-48188-2. Publisher: Nature Publishing Group.

[39] Ospina, O. E. *et al.* spatialGE: A user-friendly web application to democratize spatial transcriptomics analysis (2024). URL https://www.biorxiv.org/content/10.1101/2024.06.27.601050v1. Pages: 2024.06.27.601050 Section: New Results.

[40] Stoltzfus, C. R. *et al.* CytoMAP: A Spatial Analysis Toolbox Reveals Features of Myeloid Cell Organization in Lymphoid Tissues. *Cell Reports* **31**, 107523 (2020). URL https://linkinghub.elsevier.com/retrieve/pii/S221112472030423X.

[41] Weeratunga, P. *et al.* Single cell spatial analysis reveals inflammatory foci of immature neutrophil and CD8 T cells in COVID-19 lungs. *Nature Communications* **14**, 7216 (2023). URL https://www.nature.com/articles/s41467-023-42421-0. Number: 1 Publisher: Nature Publishing Group.

[42] Weeratunga, P. *et al.* Temporo-spatial cellular atlas of the regenerating alveolar niche in idiopathic pulmonary fibrosis (2024). URL https://www.medrxiv.org/content/10.1101/2024.04.10.24305440v1. Pages: 2024.04.10.24305440.

[43] Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nature Methods* **19**, 171–178 (2022). URL https://www.nature.com/articles/s41592-021-01358-2. Publisher: Nature Publishing Group.

[44] Virshup, I. *et al.* The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology* **41**, 604–606 (2023). URL https://www.nature.com/articles/s41587-023-01733-8. Publisher: Nature Publishing Group.

[45] Edelsbrunner, H., Kirkpatrick, D. & Seidel, R. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory* **29**, 551–559 (1983). URL https://ieeexplore.ieee.org/document/1056714. Conference Name: IEEE Transactions on Information Theory.

[46] Clark, P. J. & Evans, F. C. Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology* **35**, 445–453 (1954). URL http://doi.wiley.com/10.2307/1931034.

[47] Morueta-Holme, N. *et al.* A network approach for inferring species associations from co-occurrence data. *Ecography* **39**, 1139–1150 (2016). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.01892. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.01892.

[48] Hu, Y. *et al.* Unsupervised and supervised discovery of tissue cellular neighborhoods from cell phenotypes. *Nature Methods* **21**, 267–278 (2024). URL https://www.nature.com/articles/s41592-023-02124-2. Publisher: Nature Publishing Group.

[49] Haviv, D. *et al.* The covariance environment defines cellular niches for spatial inference. *Nature Biotechnology* 1–12 (2024). URL https://www.nature.com/articles/s41587-024-02193-4. Publisher: Nature Publishing Group.

[50] Wu, Z. *et al.* Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering* **6**, 1435–1448 (2022). URL https://www.nature.com/articles/s41551-022-00951-w. Publisher: Nature Publishing Group.

[51] Li, Y., Zhang, J., Gao, X. & Zhang, Q. C. Tissue module discovery in single-cell-resolution spatial transcriptomics data via cell-cell interaction-aware cell embedding. *Cell Systems* **15**, 578–592.e7 (2024). URL https://www.cell.com/cell-systems/abstract/S2405-4712(24)00124-8. Publisher: Elsevier.

[52] Fredrickson, M. M. & Chen, Y. Permutation and randomization tests for network analysis. *Social Networks* **59**, 171–183 (2019). URL https://www.sciencedirect.com/science/article/pii/S0378873318303381.

[53] Anselin, L. Local Indicators of Spatial Association—LISA. *Geographical Analysis* **27**, 93–115 (1995). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1995.tb00338.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.1995.tb00338.x.

[54] Bull, J. A., Mulholland, E. J., Leedham, S. J. & Byrne, H. M. Extended correlation functions for spatial analysis of multiplex imaging data. *Biological Imaging* (2024). URL https://doi.org/10.1017/S2633903X24000011.

[55] Maley, C. C., Koelble, K., Natrajan, R., Aktipis, A. & Yuan, Y. An ecological measure of immune-cancer colocalization as a prognostic factor for breast cancer. *Breast Cancer Research* **17**, 131 (2015). URL https://doi.org/10.1186/s13058-015-0638-4.

[56] Shaban, M. *et al.* A Novel Digital Score for Abundance of Tumour Infiltrating Lymphocytes Predicts Disease Free Survival in Oral Squamous Cell Carcinoma. *Scientific Reports* **9**, 13341 (2019). URL https://www.nature.com/articles/s41598-019-49710-z. Publisher: Nature Publishing Group.

[57] Ripley, B. D. Modelling Spatial Patterns. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 172–192 (1977). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/

j.2517-6161.1977.tb01615.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01615.x.

[58] Bull, J. A. *et al.* Combining multiple spatial statistics enhances the description of immune cell localisation within tumours. *Scientific Reports* **10**, 18624 (2020). URL https://www.nature.com/articles/s41598-020-75180-9. Number: 1 Publisher: Nature Publishing Group.

[59] Sidak, D., Schwarzerová, J., Weckwerth, W. & Waldherr, S. Interpretable machine learning methods for predictions in systems biology from omics data. *Frontiers in Molecular Biosciences* **9** (2022). URL https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2022.926623/full. Publisher: Frontiers.

[60] Vipond, O. *et al.* Multiparameter persistent homology landscapes identify immune cell spatial patterns in tumors. *Proceedings of the National Academy of Sciences* **118**, e2102166118 (2021). URL https://www.pnas.org/doi/abs/10.1073/pnas.2102166118. Publisher: Proceedings of the National Academy of Sciences.

[61] Yoon, I. H. R. *et al.* Deciphering the diversity and sequence of extracellular matrix and cellular spatial patterns in lung adenocarcinoma using topological data analysis (2024). URL https://www.biorxiv.org/content/10.1101/2024.01.05.574362v2. Pages: 2024.01.05.574362 Section: New Results.

[62] Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: Access and store annotated data matrices. *Journal of Open Source Software* **9**, 4371 (2024). URL https://joss.theoj.org/papers/10.21105/joss.04371.

[63] Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95 (2007). URL https://ieeexplore.ieee.org/document/4160265. Conference Name: Computing in Science & Engineering.

[64] Bauer, U. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology* **5**, 391–423 (2021). URL https://doi.org/10.1007/s41468-021-00071-5.

[65] Hagberg, A. A., Schult, D. A. & Swart, P. J. *Exploring Network Structure, Dynamics, and Function using NetworkX*, 11–15 (Pasadena, California, 2008). URL https://doi.curvenote.com/10.25080/TCWV9851.

[66] Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020). URL https://www.nature.com/articles/s41592-019-0686-2. Publisher: Nature Publishing Group.

[67] Kwinta, A. & Bac-Bronowicz, J. Regular polygons in 2D objects shape description. *Geomatics, Landmanagement and Landscape* (2020). URL https://repo.ur.krakow.pl/info/article/UR62ad65e95a3a45efb63ae1d773088563/.

[68] Barthélemy, M. Spatial networks. *Physics Reports* **499**, 1–101 (2011). URL https://www.sciencedirect.com/science/article/pii/S037015731000308X.

[69] Parrillo, C. *et al.* ALOA, a pipeline for preliminary analysis of spatial profiling imaging data. *Computational and Structural Biotechnology Journal* **0** (2024). URL https://www.csbj.org/article/S2001-0370(24)00397-0/fulltext. Publisher: Elsevier.

[70] Baddeley, A., Rubak, E. & Turner, R. *Spatial Point Patterns: Methodology and Applications with R* 1st edition edn (CRC Press, Taylor & Francis Group, Boca Raton, London, New York, 2016). URL https://www.routledge.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/p/book/9781482210200.

[71] Getis, A. & Ord, J. K. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* **24**, 189–206 (1992). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1992.tb00261.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.1992.tb00261.x.

[72] Kullback, S. & Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86 (1951). URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-1/On-Information-and-Sufficiency/10.1214/aoms/1177729694.full. Publisher: Institute of Mathematical Statistics.

[73] van Lieshout, M. N. M. & Baddeley, A. J. A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica* **50**, 344–361 (1996). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1996.tb01501.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9574.1996.tb01501.x.

[74] Sreelekha, M. G., Krishnamurthy, K. & Anjaneyulu, M. V. L. R. Interaction between Road Network Connectivity and Spatial Pattern. *Procedia Technology* **24**, 131–139 (2016). URL https://www.sciencedirect.com/science/article/pii/S2212017316301025.

[75] Saxena, A. & Iyengar, S. Centrality Measures in Complex Networks: A Survey (2020). URL http://arxiv.org/abs/2011.07190. ArXiv:2011.07190.

[76] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008). URL https://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[77] Wills, P. & Meyer, F. G. Metrics for graph comparison: A practitioner's guide. *PLOS ONE* **15**, e0228728 (2020). URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0228728. Publisher: Public Library of Science.

[78] Bull, J. A. & Byrne, H. M. Quantification of spatial and phenotypic heterogeneity in an agent-based model of tumour-macrophage interactions. *PLOS Computational Biology* **19**, e1010994 (2023). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010994. Publisher: Public Library of Science.

[79] Ali, D. *et al.* A Survey of Vectorization Methods in Topological Data Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**, 14069–14080 (2023). URL https://ieeexplore.ieee.org/abstract/document/10235748. Conference Name: IEEE Transactions on

Pattern Analysis and Machine Intelligence.

[80] Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011). URL http://jmlr.org/papers/v12/pedregosa11a.html.

[81] Gatenbee, C. D. *et al.* Immunosuppressive niche engineering at the onset of human colorectal cancer. *Nature Communications* **13**, 1798 (2022). URL https://www.nature.com/articles/s41467-022-29027-8. Number: 1 Publisher: Nature Publishing Group.

[82] Xi, J. & Niles-Weed, J. Distributional Convergence of the Sliced Wasserstein Process. *Advances in Neural Information Processing Systems* **35**, 13961–13973 (2022). URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5a5e9197ea547141b4977a5a198bbaac-Abstract-Conference.html.

[83] Bonneel, N., Rabin, J., Peyré, G. & Pfister, H. Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision* **51**, 22–45 (2015). URL https://doi.org/10.1007/s10851-014-0506-3.

7 Acknowledgements

The authors would like to thank the many researchers (experimentalists, bioinformaticians, mathematicians, and others) involved in testing the MuSpAn package for their valuable feedback in implementing improvements in the usability and performance of the package. Many thanks to Ling-Pei Ho, Praveen Weeratunga, Chris Pugh, and Philip Macklin for collaborations which led to data shown in Section C of the supplementary information. The authors would also like to thank Oxford University Innovation for their help in releasing MuSpAn. JAB and JWM were supported by Cancer Research UK (CRUK) grant number CTRQQR-2021\100002, through the Cancer Research UK Oxford Centre. SJL was supported by CRUK Program Grant (DRCNPG-Jun22\100002) and the MRC Mouse Genetics Network. EJM was supported by the Lee Placito Medical Research Fellowship (University of Oxford).

8 Author Contributions

JAB and JWM were responsible for software development and implementation. JAB, JWM, HMB, and SJL conceptualised the project. All authors wrote and reviewed the manuscript, and approved the final version.

9 Competing Interests

MuSpAn is released under an Academic Use Licence, and is therefore available to licence commercially.

Appendix A Available quantitative tools

Table A1 contains a list of the data analysis methods currently available within each MuSpAn module. For further details about how each method works, we refer the interested reader to the citations in Table A1, or to the documentation available at https://docs.muspan.co.uk/ where additional functions not used here can also be found.

Table A1: A summary of the data analysis methods available within each module of MuSpAn v1.0.0, together with a brief summary of their purpose. For further details about these methods, please see the references provided.

Method name (`MuSpAn Module`)	**Summary**	**References**
Label Count (`Summary Statistics`)	Number of objects with a given label, within a given region .	[30, 35, 69]
Label Density (`Summary Statistics`)	Count, divided by region area.	[35, 41]
Label Entropy (`Summary Statistics`)	Shannon entropy of a specified label.	[33, 56]
Generate Distribution (`Distribution`)	Convert point data into a distribution using general kernel.	-
Kernel Density Estimation (`Distribution`)	Convert point data into a probability density distribution using a Gaussian kernel function.	[66]
Kullback-Leibler Divergence (`Distribution`)	Measures the similarity between two distributions using relative entropy.	[72]
Wasserstein distance (`Distribution`)	Measures the distance between two distributions.	[30, 35]
Summarise Distribution (`Distribution`)	Computes the mean and standard deviational ellipses of spatial point data.	-
Area (`Geometry`)	Area of a shape.	-
Circularity (`Geometry`)	Circularity of a shape.	-
Convexity (`Geometry`)	Convexity of a shape (ratio of the area to the area of its convex hull)	-
Perimeter (`Geometry`)	Perimeter of a shape.	-
Principle Axis (`Geometry`)	Computes the angle and major axis of a polygon relative to the x-axis.	-

Method name (`MuSpAn Module`)	**Summary**	**References**
Adjacency Permutation Test (`Networks`)	Test the correlation between adjacent labels in a network by randomly permuting them (also referred to as 'neighbourhood enrichment').	[42, 52]
Alpha Index (`Networks`)	A measure of connectivity which evaluates the number of cycles in a graph in comparison with the maximum number of cycles. (≈ 0 for a Tree, ≈ 1 for a maximal planar graph).	[68, 74]
Betweenness Centrality (`Networks`)	The sum of the fraction of all shortest paths that pass through each node.	[68, 75]
Closeness Centrality (`Networks`)	The average shortest distance from each node to all other reachable nodes.	[68, 75]
Cluster Neighbourhoods (`Networks`)	Cluster similar neighbourhoods based on the enrichment of labels within the neighbourhoods.	[31, 40]
Community Detection (`Networks`)	Perform community detection based on node connectivity.	[76]
Compactness (`Networks`)	Index in the range $[0, 1]$ that relates the area of a network to the length of its edges.	[68, 75]
Cyclomatic Number (`Networks`)	The number of elementary cycles within a graph.	[68]
Edge Orientation Distribution (`Networks`)	Compute the (symmetric) distribution of edge orientations.	[68]
Gamma Index (`Networks`)	A measure of connectivity that considers the relationship between the number of observed links and the number of possible links.	[68, 74]
Network Generation (`Networks`)	Generate a network based on MuSpAn objects. Available methods include Delaunay, k-Nearest Neighbours, and Proximity networks.	[31, 37, 43, 50, 68]
Katz Centrality (`Networks`)	The influence of a node in a network, determined by considering the total number of walks, weighted by their lengths, that connect the node to all other nodes..	[68]
k-Hop Neighbourhood (`Networks`)	Find the k-hop neighbourhood of a set of source objects	-
k-Nearest Neighbour Label Count (`Networks`)	Count the occurrences of each label within k steps from a set of source objects	-
Laplacian Centrality (`Networks`)	A measure of node connectivity in a graph	-

Method name (`MuSpAn Module`)	**Summary**	**References**
Network Distance (`Networks`)	Measure the similarity of two networks, using a range of available distance metrics.	[77]
Node Degree (`Networks`)	Compute the degree of each node.	-
Generate Quadrats (`Region-based`)	Create a mesh of quadrats covering the requested space.	-
Generate Hexgrid (`Region-based`)	Create a mesh of regular hexagons covering the requested space.	-
Morisita Horn Index (`Region-based`)	A measure of the similarity of observations of a label within a set of regions.	[55, 56]
Quadrat Correlation Matrix (`Region-based`)	Describe correlations between the counts of different cell types within a set of regions.	[30, 41, 42, 47]
Average Nearest Neighbour Index (`Spatial statistics`)	A measure of spatial aggregation based on the average distance between pairs of objects.	[33, 46]
Cross Ripley's K Function (`Spatial statistics`)	A measure of (cumulative) spatial aggregation between pairs of objects across length scales.	[33, 43]
Cross L Function (`Spatial statistics`)	A normalised version of Ripley's K-Function.	[35, 43]
Cross Pair Correlation Function (`Spatial statistics`)	A measure of spatial aggregation between pairs of objects across length scales.	[54, 58, 69, 70]
Getis Ord Statistic (`Spatial statistics`)	A measure of spatial autocorrelation that identifies hot- and cold-spots in the distribution of a continuous label.	[32, 39, 71]
J-Function (`Spatial statistics`)	A test that identifes clustering and dispersion in point patterns. Relates the Spherical Contact Distribution and the Nearest Neighbour Distribution.	[58, 73]
Moran's I (`Spatial statistics`)	Calculate both local and global Moran's I statistic within the domain. The local Moran's I statistic quantifies the spatial autocorrelation for each object in the population. The global Moran's I statistic quantifies the spatial autocorrelation for the entire population.	[33, 39, 43]
Nearest Neighbour Distribution (`Spatial statistics`)	Calculate the distribution of distances between the closest objects in two populations.	[40]
Spherical Contact Distribution (`Spatial statistics`)	Calculate the distance from randomly sampled points to the closest objects in a population. Also referred to as the 'Empty Space Distribution'.	[58]

Method name (`MuSpAn Module`)	**Summary**	**References**
Topographical Correlation Map (`Spatial statistics`)	A spatially-resolved, quantitative summary of colocalisation between two point populations across a region of interest.	[30, 41, 42, 54, 69]
Weighted Pair Correlation Function (`Spatial statistics`)	An extension of the cross PCF which accounts for continuous labels.	[54, 78]
Level Set Filtration (`Topology`)	Calculate the level set filtration of a distribution or image.	[30, 64]
Vietoris-Rips Filtration (`Topology`)	Calculate the Vietoris-Rips filtration of a set of points.	[64]
Vectorise Persistence (`Topology`)	Convert persistence diagram into a vector for subsequent analysis.	[79]
Cluster Data (`Helpers`)	Wrapper for clustering methods implemented in scikit-learn.	[80]
Crop Domain (`Helpers`)	Return a new domain (or domains) based on a set of new boundary shapes (or label values).	-
Combine Shapes (`Shape operations`)	Combine two shapes in the domain using a specified operation.	-
Return Difference of Shapes (`Shape operations`)	Compute the difference between two shapes.	-
Return Intersection of Shapes (`Shape operations`)	Compute the intersection of two shapes.	-
Return Union of Shapes (`Shape operations`)	Compute the union of two shapes.	-

Appendix B Further information on spatial analysis methods

Below, we describe the spatial analysis methods used to generate Figs. 2-5 of the main text. The methods are organised to mirror the subsections of the Results section in the main text.

B.1 Spatial analysis of individual cells

B.1.1 Circularity

Description: Circularity measures how closely a shape resembles a circle.
Definition: Let O be a polygon with area, A, and perimeter, P. Then, the circularity, C, of O is defined as

$$C = \frac{4\pi A}{P^2}. \tag{B1}$$

Application: In Fig 2F, circularity is computed for all 'Cell boundary' objects labelled with Celltype ID 1 or 2.

B.1.2 Principal angle

Description: A 2D shape has two principal axes: the principal major axis is the axis of maximum moment of area (maximising bend resistance), and the principal minor axis is the axis of minimum moment of area (minimum bend resistance) [67].
Definition: The principal angle is the angle between the principal major axis and the line $y = 0$. Let O be a polygon defined by n vertices $\boldsymbol{v}_i = (x_i, y_i) \in \mathbb{R}^2$ $(1 \leq i \leq n)$, ordered anti-clockwise.

The second moments of area, I_x, I_y and I_{xy}, are given by

$$I_x = \frac{1}{12}\sum_{i=1}^{n-1}\left(x_i y_{i+1} - x_{i+1} y_i\right)\left(x_i^2 + x_i x_{x_{i+1}} + x_{i+1}^2\right), \tag{B2}$$

$$I_y = \frac{1}{12}\sum_{i=1}^{n-1}\left(x_i y_{i+1} - x_{i+1} y_i\right)\left(y_i^2 + y_i y_{x_{i+1}} + y_{i+1}^2\right), \tag{B3}$$

$$I_{xy} = \frac{1}{24}\sum_{i=1}^{n-1}\left(x_i y_{i+1} - x_{i+1} y_i\right)\left(x_i y_{i+1} + 2x_i y_i + 2x_{i+1} y_{i+1} + x_{i+1} y_i\right). \tag{B4}$$

They describe how the vertices of a polygon are distributed with respect the x and y-axis. The principal angle, θ_O, of O is defined by

$$\theta_O = \frac{1}{2}\arctan\left(\frac{2I_{xy}}{I_y - I_x}\right). \tag{B5}$$

Application: In Fig 2G, the principal angle is computed for all 'Cell boundary' objects with labels Celltype ID 1 or 2.

B.2 Cell proximity analysis

B.2.1 Nearest neighbour metrics

Description: The Average Nearest Neighbour Distance is the mean minimum distance from an object in one population to an object in another. It is closely related to the Average Nearest Neighbour Index (ANNI), which measures the degree of clustering (or dispersion) of a set of points by comparing the average nearest neighbour distance to the expected nearest neighbour distance in a random distribution of the same density of points, here assumed to be complete spatial randomness (CSR) [46].

Definition: Let Ω be a spatial domain with area α. Let $d_{A,B}$ denote the minimum Euclidean distance between objects A and B, for A, B contained in Ω. Let $\mathcal{O}_A = \{A_1, ..., A_n\}$ and $\mathcal{O}_B = \{B_1, ..., B_m\}$ be two sets with objects A_i $(i = 1, ..., n)$ and B_j $(j = 1, ..., m)$ in Ω. Then the minimum nearest neighbour distance from a specific object A_k to an object in the set $\mathcal{O}_B$ is

$$\mathrm{NN}(A_k; \mathcal{O}_B) = \min_{j=1,...,m} \{d_{A_k,B_j}\}. \tag{B6}$$

The average nearest neighbour distance for objects from the set $\mathcal{O}_A$ to those in $\mathcal{O}_B$ is then

$$\mathrm{ANN}(\mathcal{O}_A, \mathcal{O}_B) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{NN}(A_i; \mathcal{O}_B). \tag{B7}$$

Assuming CSR, the expected average nearest neighbour distance, $\overline{\mathrm{ANN}}(\mathcal{O}_A; \mathcal{O}_B)$, is given by

$$\overline{\mathrm{ANN}}(\mathcal{O}_A; \mathcal{O}_B) = \frac{1}{2}\sqrt{\frac{\alpha}{m}}. \tag{B8}$$

The Average Nearest Neighbour Index, $\mathrm{ANNI}(\mathcal{O}_A; \mathcal{O}_B)$ is expressed as the ratio of the observed to the expected average nearest neighbour distance:

$$\mathrm{ANNI}(\mathcal{O}_A; \mathcal{O}_B) = \frac{\mathrm{ANN}(\mathcal{O}_A; \mathcal{O}_B)}{\overline{\mathrm{ANN}}(\mathcal{O}_A; \mathcal{O}_B)}. \tag{B9}$$

The significance of the difference between the observed nearest neighbour distribution and that expected under CSR is typically quantified using the z-score (standard normal variate), Z, such that

$$Z = \frac{\mathrm{ANN}(\mathcal{O}_A; \mathcal{O}_B) - \overline{\mathrm{ANN}}(\mathcal{O}_A; \mathcal{O}_B)}{\sigma_{\mathcal{O}_A,\mathcal{O}_B}}, \tag{B10}$$

where the variance, $\sigma_{\mathcal{O}_A,\mathcal{O}_B}$, is evaluated as

$$\sigma_{\mathcal{O}_A,\mathcal{O}_B} = \frac{c\sqrt{\alpha}}{M}, \tag{B11}$$

with $c = 0.26136$ [46].

Application: The ANNI was used to investigate the spatial relationships between cells with Cluster ID 1 and 11, and 2 and 5 in Figs 3B-3E.

B.2.2 Quadrat correlation matrix

Description: The quadrat correlation matrix (QCM) describes correlations between counts of two labels of interest (here, Celltype IDs) within a set of 'quadrats', a regular tiling of the domain of interest with non-overlapping areas [30, 41, 47, 81]. For a given pair of labels, the QCM quantifies how different the observed correlation coefficient across the quadrats is from that expected under a random shuffling of the object labels.
Definition: The QCM is calculated by comparing the observed matrix of cell counts within each quadrat, $\mathbf{O}$, against a null distribution of K null matrices, $\mathbf{N^k}$ for $k \in [1, ..., K]$, whose row and column sums are the same as those of $\mathbf{O}$. The standard effect size (SES) for Celltypes i and j is given by

$$SES_{ij} = (P_{O_{ij}} - \mu[P_{N^k_{ij}}])/\sigma[P_{N^k_{ij}}], \tag{B12}$$

where P_A is the partial correlation matrix associated with matrix A, and μ and σ are the mean and standard deviation over all k. The entry of the QCM corresponding to Celltypes i and j is SES_{ij} if the p-value associated with SES_{ij} score is statistically significant after applying the Benjamini-Hochberg correction for multiple comparisons, and 0 otherwise.
Application: In Fig. 3J, the QCM was computed using a square lattice of edge length 75µm and shuffled over 1000 iterations ($K = 1000$). For simplicity, values shown in Fig. 3J are not filtered for statistical significance (so, strictly speaking, this is the SES matrix, rather than the QCM).

B.3 Analysis of cell-cell relationships across scales

B.3.1 Cross pair correlation function

Description: The cross pair correlation function (cross-PCF) is a spatial statistic that relates the observed number of point pairs from populations A and B that are separated by distance r against the expected number under a statistical null distribution.
Definition: Let $\mathbf{x}^A$ and $\mathbf{x}^B$ be the sets of locations of points in populations A and B respectively, and suppose there are N_A and N_B points in each set. Consider the indicator function

$$I_{[p,q)}(r) = \begin{cases} 1 & \text{if } r \in [p, q), \\ 0 & \text{otherwise.} \end{cases} \tag{B13}$$

The cross-PCF, $g_{AB}(r)$, is given by

$$g_{AB}(r) = \frac{1}{N_A} \sum_{a=1}^{N_A} \sum_{b=1}^{N_B} \frac{I_{[0,dr)}(|\mathbf{x}^A_a - \mathbf{x}^B_b| - r)}{\alpha_r(\mathbf{x}_i)} \Big/ \frac{N_B}{\alpha}, \tag{B14}$$

where α is the area of the domain of interest, and $\alpha_r(\mathbf{x})$ is the area of the intersection between the domain and an annulus centred at $\mathbf{x}$ with inner radius r and outer radius dr. This provides a boundary correction term for points close to the domain edge; alternative methods for boundary correction can be found in [70].
Application: In Fig. 4B we use $dr = 15$µm and calculate $g_{AB}(r)$ at discrete values $r_k \in [0, 5, 10, ..., 250]$, excluding any points within 20µm of the region boundary from the calculation.

B.3.2 Vietoris-Rips filtration

Description: A Vietoris–Rips (VR) filtration is a method from persistent homology that can be used to quantify the topological features (e.g. connected components, loops, voids) of point clouds at multiple spatial scales [64].
Definition: Let X denote a finite set of points in a domain, $\Omega \subset \mathbb{R}^2$, and $d_{i,j}$ represent the distance between points x_i and x_j. For a given scale parameter $\epsilon \geq 0$, the VR complex $R_\epsilon(X)$ is a simplicial complex defined as follows. A finite subset $\sigma = \{x_0, x_1, \ldots, x_k\} \subseteq X$ forms a k-simplex in $R_\epsilon(X)$ if and only if the pairwise distances between all points in σ satisfy $d_{i,j} \leq \epsilon$ for all $0 \leq i, j \leq k$. As the scale parameter ϵ increases, the VR complexes are nested, satisfying $R_{\epsilon_1}(X) \subseteq R_{\epsilon_2}(X)$ for any $0 \leq \epsilon_1 \leq \epsilon_2$. This nested sequence forms a filtration of simplicial complexes, denoted by $\{R_\epsilon(X)\}_{\epsilon \geq 0}$.

To quantify these topological features, the homology groups $H_k(R_\epsilon(X))$ of the VR complexes are computed for each dimension k as follows. Each k-simplex in $R_\epsilon(X)$ there exists a boundary operator ∂_k that maps it to a formal sum of its $(k-1)$-dimensional faces. The boundary operators satisfy the composition property $\partial_k \circ \partial_{k+1} = 0$, which ensures that the image of ∂_{k+1} is contained within the kernel of ∂_k. The k-th homology group is then defined as the quotient

$$H_k(R_\epsilon(X)) = \frac{\ker(\partial_k)}{\mathrm{im}(\partial_{k+1})}. \tag{B15}$$

It captures the k-dimensional topological features, such as connected components ($k = 0$), loops ($k = 1$), and higher-dimensional voids. Persistent homology extends this computation across the filtration $\{R_\epsilon(X)\}_{\epsilon \geq 0}$. As ϵ increases, the boundary maps and the resulting homology groups change, leading to the birth and death of topological features.
Application: VR filtrations were used to identify crypt-like structures in Fig. 4F in the Celltype 2 population by restricting our attention 0 and 1-dimensional features ($k \leq 1$) and defining $d_{i,j}$ as the minimum euclidean distance between point objects i and j.

B.4 Spatial relationships at tissue scales

B.4.1 Microenvironment (neighbourhood) clustering

Description: Microenvironment clustering seeks to group objects with similar compositions of labels located within a local neigbbourhood of an object.These methods are inspired by those deployed in [31].
Definition: Let $\mathcal{G}(V, E)$ be a network with nodes $v_i \in V$ and edges $e_{ij} \in E$ encode spatial relationships between nodes. Each node has a categorical label $c_i \in \{C_1, \ldots, C_m\}$ (i.e., Celltype ID). Let the microenvironment, $\mathcal{N}_i$, be the set of nodes about node i defined by a k-hop neighbourhood (see Table 1). For each neighbourhood, $\mathcal{N}_i$, the composition of each label category is computed to generate an observation matrix $\boldsymbol{O} \in \mathbb{N}^{|V| \times m}$ where

$$[\boldsymbol{O}]_{i,s} = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \delta(c_j, C_s) \tag{B16}$$

and

$$\delta(c_i, C_s) = \begin{cases} 1 \text{ if } c_i = C_s, \\ 0 \text{ otherwise.} \end{cases} \tag{B17}$$

Row-wise clustering is performed on $\boldsymbol{O}$ to identify similar microenvironment compositions, which partition the nodes $v_i \in V$ into disjoint sets S_t.

Cluster representatives (i.e., the mean or medoid composition vector of the cluster) are extracted from each set S_t, denoted by M_{i,S_t}.

Define the average composition of label C_i of the sample, M_i, by

$$M_i = \frac{1}{|V|} \sum_{j=1}^{|V|} \delta\left(c_j, C_i\right). \tag{B18}$$

An microenvironment enrichment score, E_{i,S_t}, is then calculated by compare the representatives to the average composition,M_i, define as a log-fold change

$$E_{i,S_t} = \log_2 \left(\frac{M_{i,S_t}}{M_i} \right). \tag{B19}$$

Application: In Fig. 5D, $\mathcal{G}(V, E)$ is a Delaunay network (see Table 1) on point-like objects with edge distance threshold of 30µm. A 3-hop neighbourhood was used to generate microenvironment sets about each node. Clustering was performed using k-means with $k = 5$.

B.4.2 Adjacency permutation test

Description: The adjacency permutation test quantifies the significance of labelled nodes being adjacent in a network. This non-parametric test involves repeatedly shuffling node labels to generate a null distribution of a test statistic under the hypothesis of no spatial association [52].

Definition: Let $\mathcal{G}(V, E)$ be a network with nodes $v_i \in V$ and edges $e_{ij} \in E$ where each node has a categorical label $c \in \{C_1, \ldots, C_m\}$. For every pair of categorical labels C_s and C_t, the number of observed edges, n_{C_s,C_t} , is calculated between nodes with these labels such that

$$n_{C_s,C_t} = \sum_{e_{ij} \in E} \delta\left(c_i, C_s\right) \delta\left(c_j, C_t\right), \tag{B20}$$

where $\delta(c_i, C_s)$ is defined in equation (B17). An observation matrix, $\boldsymbol{N} \in \mathbb{N}_0^{m \times m}$, is constructed using each categorical pair count, $[\boldsymbol{N}]_{s,t} = n_{C_s,C_t}$, for $1 \leq s, t \leq m$.

The node labels are then shuffled and the adjacent pair count procedure is repeated for K iterations, generating a null distribution of count matrices $\tilde{\boldsymbol{N}}_k$ for $1 \leq k \leq K$. The adjacency significance score for each pair of categorical labels, $S_{s,t}$, is calculated as z-score of the observed to the null,

$$S_{s,t} = \frac{[\boldsymbol{N}]_{s,t} - \mu\left([\tilde{\boldsymbol{N}}_k]_{s,t}\right)}{\sigma\left([\tilde{\boldsymbol{N}}_k]_{s,t}\right)} \tag{B21}$$

are the mean and standard deviation over all k.

Application: The Adjacency Permutation Test was used to determine the adjacency significance of the ME regions that are in close proximity in Fig. 5G. After converting points with identical ME labels into shapes (ME regions, Fig. 5F), a proximity network of these shapes was generated where edges between ME regions were drawn if they less than 5µm apart. This proximity network equipped with ME ID labels was then subjected to a Adjacency Permutation Test with 1000 random samples ($K = 1000$).

B.4.3 Local Getis-Ord

Description: The local Getis-Ord statistic G_i^* is measures spatial autocorrelation by relating spatial weights (e.g., distances between objects) with label values [71].
Definition: Let $\boldsymbol{x}_i \in \mathbb{R}^2$ represent the spatial position of the i-th object, $1 \leq i \leq n$, and associated with a continuous label $l_i \in \mathbb{R}$. If $w_{i,j} \in \mathbb{R}$ defines the spatial relationship between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, then the local Getis-Ord statistic, G_i^*, for object i is given by

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij} l_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} l_j}. \tag{B22}$$

G_i^* is the weighted average of label values adjacent to object i. The value of G_i^* relative to the mean identifies whether an object belongs to a *hotspot* (+ve) or *coldspot* (-ve) for that label is typically reported as a z-score, Z,

$$Z = \frac{G_i^* - \mu(G_i^*)}{\sigma(G_i^*)} \tag{B23}$$

where μ and σ are the mean and standard deviation over all i, respectively.
Application: In Fig. 5I, the objects are hexagons, and we assign a weight function for the shapes i and j of the form

$$w_{ij} = \begin{cases} 1 \text{ if } d_{i,j} = 0, \\ 0 \text{ otherwise,} \end{cases} \tag{B24}$$

where $d_{i,j}$ is the minimum distance between objects i and j. The labels l_i are the number of cells of Celltype ID 2 located within hexagon i.

B.4.4 Gaussian kernel density estimation

Description: Kernel density estimation (KDE) is a non-parametric method for estimating a probability density function from observations.
Definition: A Gaussian KDE uses Gaussian kernels centred on each observation, $\boldsymbol{x}_i$, to generate a distribution over a domain $\Omega \subset \mathbb{R}^2$. The resultant probability density function, $P(\boldsymbol{x})$, is determined by

$$P(\mathbf{x}) = \frac{1}{n\sigma\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(\frac{-||\boldsymbol{x} - \boldsymbol{x}_i||^2}{2\sigma^2}\right), \tag{B25}$$

where n is the number of observations and σ is the bandwidth of the kernel. Increasing σ generates smoother representations of the observed point data, $\boldsymbol{x}_i$.
Application: In Fig. 5J, we independently apply Gaussian kernels, centered on points with Celltype IDs 1, 2, and 11, with $\sigma = 0.05$.

B.4.5 Wasserstein distance

Description: The 2-Wasserstein distance (referred to here as the Wasserstein distance), describes the minimum distance required to transform one distribution onto another (see [82] for futher details).
Definition: Consider two discrete probability distributions $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$ and $\boldsymbol{\nu} = (\nu_1, ..., \nu_m)$ with spatial locations $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ within a domain $\Omega \subset \mathbb{R}^2$, respectively.

Then the 2-Wasserstein distance is defined as

$$W_2\left(\boldsymbol{\mu},\boldsymbol{\nu}\right) = \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \sum_{i,j} ||\boldsymbol{x}_i - \boldsymbol{y}_j||_2 \boldsymbol{\pi}_{i,j} \tag{B26}$$

where $||\cdot||_2$ is the l_2-norm and

$$\Pi\left(\boldsymbol{\mu},\boldsymbol{\nu}\right) = \{\boldsymbol{\pi}' \in \mathbb{R}^{n\times m}_{\geq 0} : \boldsymbol{\pi}'\mathbf{1}_m = \boldsymbol{\mu}, \boldsymbol{\pi}'^T\mathbf{1}_n = \boldsymbol{\nu}\} \tag{B27}$$

is the set of all possible transport plans from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$. The mapping $\boldsymbol{\pi}$ provides the optimal transport plan that transforms $\boldsymbol{\mu}$ onto $\boldsymbol{\nu}$ with the minimum work.

Solving for $\boldsymbol{\pi}$ is computationally inefficient for distributions in more than one dimension and therefore we employ the sliced method to determine $\boldsymbol{\pi}$ [83]. We project $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ onto a one-dimensional line drawn through the barycentre of the distributions at an angle $\theta \sim U[0,\pi)$ and then determine the 1-dimensional Wasserstein determined,

$$W_2\left(\mu^\theta,\nu^\theta\right) = \min_{\pi^\theta\in\Pi(\mu^\theta,\nu^\theta)} \sum_{i,j} ||x_i^\theta - y_j^\theta||_2 \pi_{i,j}^\theta. \tag{B28}$$

The Wasserstein distance of the full-dimension dataset is then approximated by the root average of these projected distances such that

$$W_2\left(\boldsymbol{\mu},\boldsymbol{\nu}\right) \approx \sqrt{\frac{1}{t}\sum_{k=1}^{t} W_2\left(\mu^{\theta_k},\nu^{\theta_k}\right)}. \tag{B29}$$

Critically, equation (B29) holds in equality in the limit $t \to \infty$ [82].
Application: In Figure 5k, μ and ν are weighted uniformly such that $\mu_i = \frac{1}{n}$ and $\nu_j = \frac{1}{m}$ for all $1 \leq i \leq n$ and $1 \leq j \leq m$. In addition, $t = 500$ was selected as the number of sliced projections for sufficient coverage of the spatial data.

B.4.6 KL-Divergence

Description: The Kullback-Leibler (KL) divergence is a method for comparing the statistical distance between two probability distributions by measuring the entropy of a distribution relative to the other [72].
Definition: Let P and Q be two probability distributions defined over the sample space $\mathcal{X}$. The KL divergence is defined as

$$\mathrm{D}_{KL}(P||Q) = \sum_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x}) \log\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right). \tag{B30}$$

Thus, $D_{KL}(P||Q)$ measures the expected logarithmic difference between the probabilities P and Q, with the expectation taken under the distribution defined by P.
Application: In Figure 5k, P and Q are the Gaussian KDEs associated with of the point clouds of the corresponding Celltype IDs 1,2 and 11.

Appendix C Application of MuSpAn to other imaging modalities

In this section, we show how MuSpAn can be applied to data from a range of imaging modalities: stacked immunohistochemistry (IHC) slides of head and neck cancer [58], Imaging Mass Cytometry (IMC) images from COVID-19 infected lung tissue [41], and CODEX images of colorectal cancer [31]. In each case, we use methods within MuSpAn to reproduce or augment the existing analyses of these datasets. Code for conducting these analyses is available in the github repository accompanying the supporting information (https://github.com/joshwillmoore1/Supporting_material_muspan_paper).

C.1 Immunohistochemistry - Bull *et al.* (2020)

Fig. C1 uses MuSpAn to analyse a set of stacked immunohistochemistry (IHC) images from a resected human head and neck cancer sample (from [58]). We consider four stains (Fig. C1A), taken from consecutive registered tissue slices: PanCK - a marker for epithelial (tumour) cells, CAIX - a marker for hypoxia, pimonidazole - a marker for severe hypoxia, and CD68 - a marker for macrophages. In [58], the slides were registered to produce a single ROI (Fig. C1B). This example dataset is available within MuSpAn, and can be loaded using `muspan.datasets.load_dataset('Macrophage-Hypoxia-ROI')` (Fig. C1C). Here, the macrophage cell centres are represented as point objects, while the boundaries of the PanCK, pimonidazole (pimo), and CAIX stains are distinct shape objects.

Here, we show how shapes can be used as analysis boundaries using the PanCK shapes. First, we calculate the macrophage-macrophage cross-PCF, which shows the association of macrophages with other macrophages (Fig. C1D). When calculated across the whole ROI (blue dashes, 'ROI'), the cross-PCF indicates extremely short-range exclusion (representing exclusion of neighbouring macrophage cell centres at lengthscales below the diameter of a macrophage), with positive correlation (clustering) at length scales greater than approximately 20 pixels (17.6µm). In particular, macrophages are observed to be around 2.25 times more likely to be found 20 pixels from another macrophage than if they were randomly distributed. We calculate the same cross-PCF within the PanCK shapes (green line, 'tumour') and excluding the PanCK shapes (red line, 'stroma'), with MuSpAn automatically adjusting the calculation of the cross-PCFs based on the relevant regions of interest (Fig. C1D). This analysis shows that macrophages are more strongly clustered within the tumour regions than in the stroma: stromal macrophages are around 1.75 times more likely to be 20µm from another stromal macrophage, while within the tumour this rises to around 2.75 times.

Similarly, we calculate the kernel density estimation (KDE) to estimate macrophage density both across the ROI (Fig. C1E) and within the PanCK shapes (Fig. C1F). Fig. C1E shows that the highest macrophage density is in the upper right of the domain. In contrast, restricting the KDE calculation to the PanCK shapes only shows that the tumour island with the highest macrophage density is in the lower left of the domain (Fig. C1F).

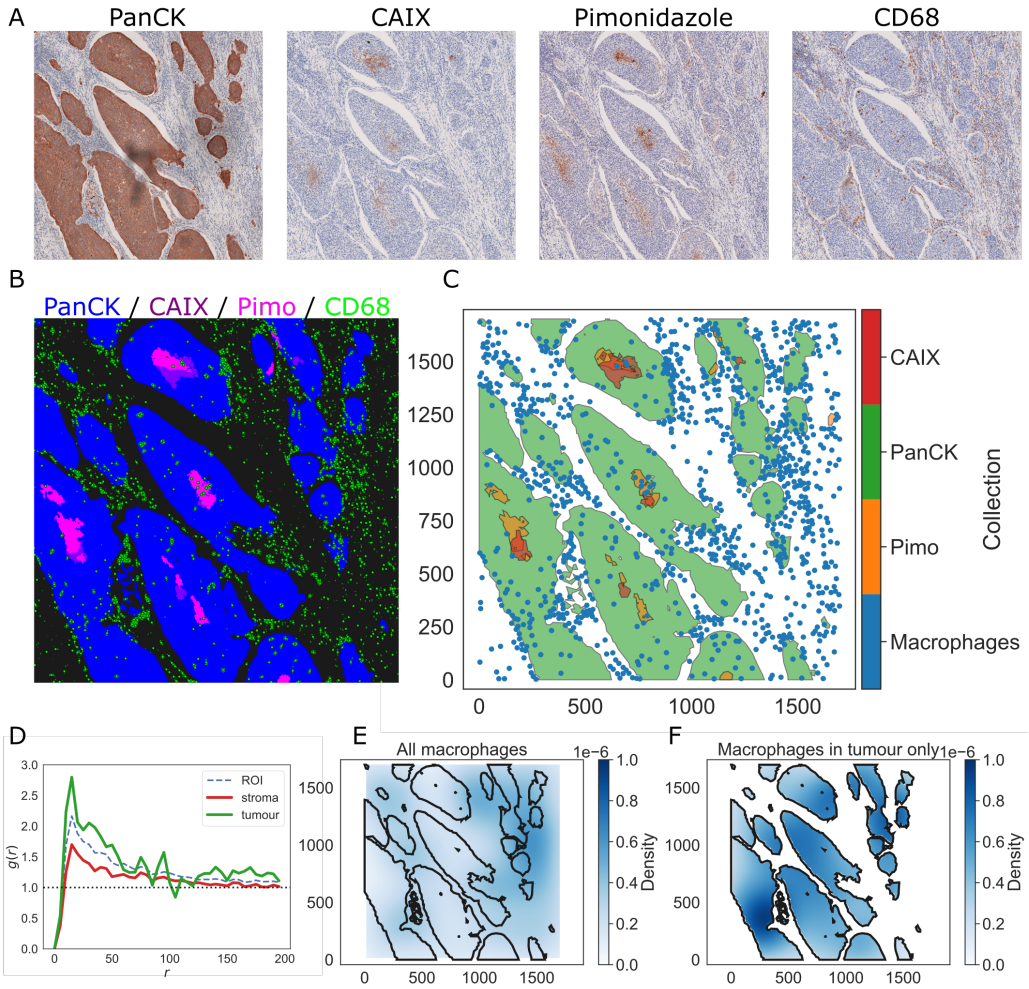

Fig. C1 MuSpAn applied to stacked IHC data.
A: A 1700 × 1700 pixel (1.5mm × 1.5mm) ROI from a human head and neck cancer sample, stained on consecutive slides in the tissue stack to show PanCK (tumour cells), CAIX (hypoxia), pimonidazole (severe hypoxia) and CD68 (macrophages).
B: Composite image showing the alignment of the four markers.
C: Boundaries and points from panel B shown as a MuSpAn domain.
D: Cross-PCFs describing macrophage-macrophage correlation within the domain. The dashed blue line (ROI) shows macrophage-macrophage correlation across the whole ROI. The solid green (tumour) and red (stroma) lines show cross-PCFs for macrophage-macrophage correlation, restricted to inside the PanCK regions (tumour) or outside (stroma).
E: Kernel density estimation (KDE) of macrophage localisation within the ROI. Outlines of PanCK shapes are shown for reference, but not used in the estimation.
F: KDE of macrophage localisation within the PanCK (tumour) regions.

C.2 Imaging Mass Cytometry - Weeratunga *et al.* (2023)

We apply MuSpAn to Imaging Mass Cytomtery (IMC) data from COVID-19 infected human lung tissue [41], considering a single 2mm $\times$ 2mm ROI characterised histologically as 'Organising Pneumonia' (Fig. C2A, B, C). The large number of cell types present makes it challenging to identify spatial relationships by eye; here, we use MuSpAn to reproduce key steps of the SpOOx pipeline [41]. We identify cell pairs of interest by first computing the QCM (Fig. C2D). We then select two cell pairs which display high or low correlation within quadrats of size 100µm: 'Blood vessels'-'UD structural SM' and 'UD-structural SM'-'Myofibroblasts', respectively.

In Fig. C3 and Fig. C4 we use MuSpAn to extend the SpOOx pipeline to characterise cell pairs of interest. In Fig. C3 we consider 'blood vessels' and 'UD structural SM'; in Fig. C4, we consider 'UD structural SM' and 'myofibroblasts'.

We compute nearest neighbour distances between the cell pairs (Fig. C3A/Fig. C4A), and then construct a weighted Delaunay network focussed on these cell pairs (Fig. C3B/Fig. C4B). We calculate the TCM (Fig. C3C/Fig. C4C), and use KDEs to compare the distribution of the cell pairs across the ROI (Fig. C3D, E/Fig. C4D, E). We also calculate two spatial statistics - the cross-K function (Fig. C3F/Fig. C4F) and cross-PCF (Fig. C3G/Fig. C4G).

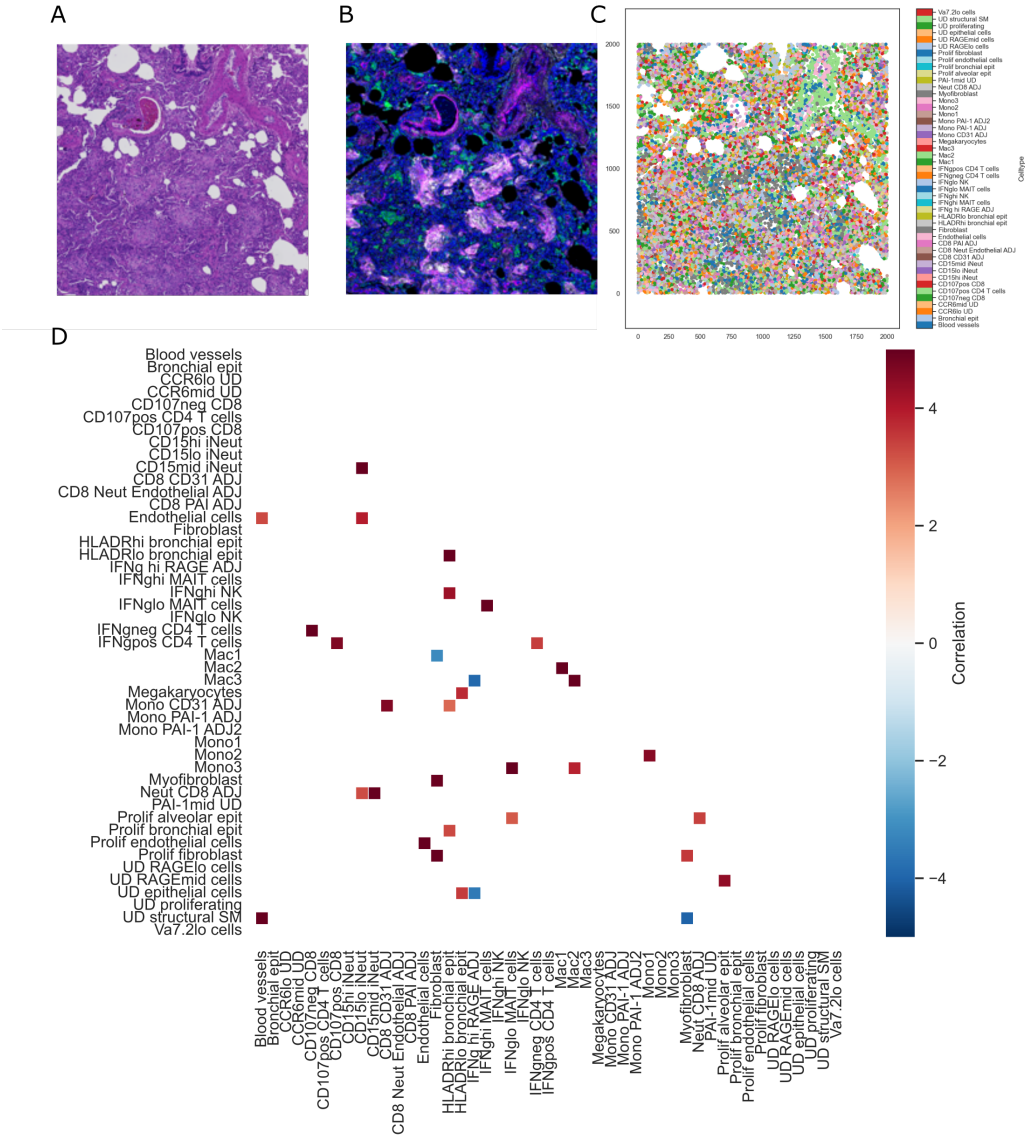

Fig. C2 MuSpAn applied to Imaging Mass Cytometry (IMC) data.
A: H&E staining of a 2mm × 2mm ROI from a human lung resected from a COVID-19 infected patient [41].
B: False colour IMC staining of the same lung ROI.
C: Point cloud of cell centres from the ROI, with annotations from [41].
D: Quadrat correlation matrix (QCM) for the cells in this ROI (square quadrats, edge length 100µm), showing statistically significant values only.

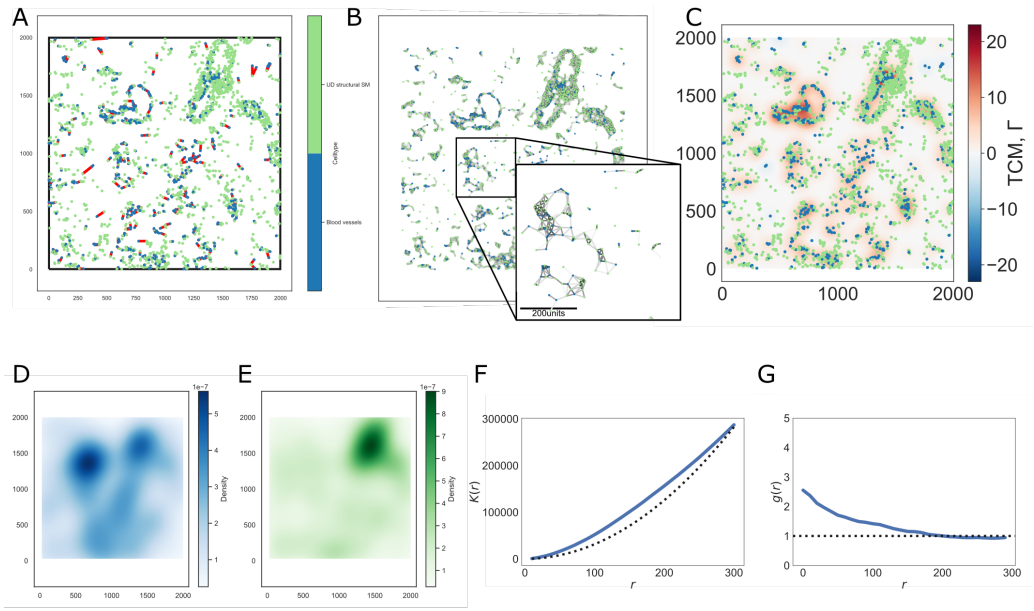

Fig. C3 MuSpAn applied to IMC data - positive spatial correlation.
A: Nearest neighbour distances (red) from blood vessels to the nearest 'UD structural SM' cell, two populations which are highlighted as spatially correlated by the QCM in Fig. C2D.
B: Weighted Delaunay network constructed on the cell centres from the two populations of interest.
C: Topographical Correlation Map (TCM) showing where blood vessels are locally positively or negatively associated with 'UD structural SM' cells.
D: Kernel Density Estimation (KDE) of blood vessels in the ROI.
E: KDE of 'UD structural SM' cells in the ROI.
F: Cross-K function between the two cell populations, indicating increased association at length scales up to 300µm.
G: Cross-PCF between the two cell populations, indicating clusters of the two populations with a characteristic size of approximately 180µm.

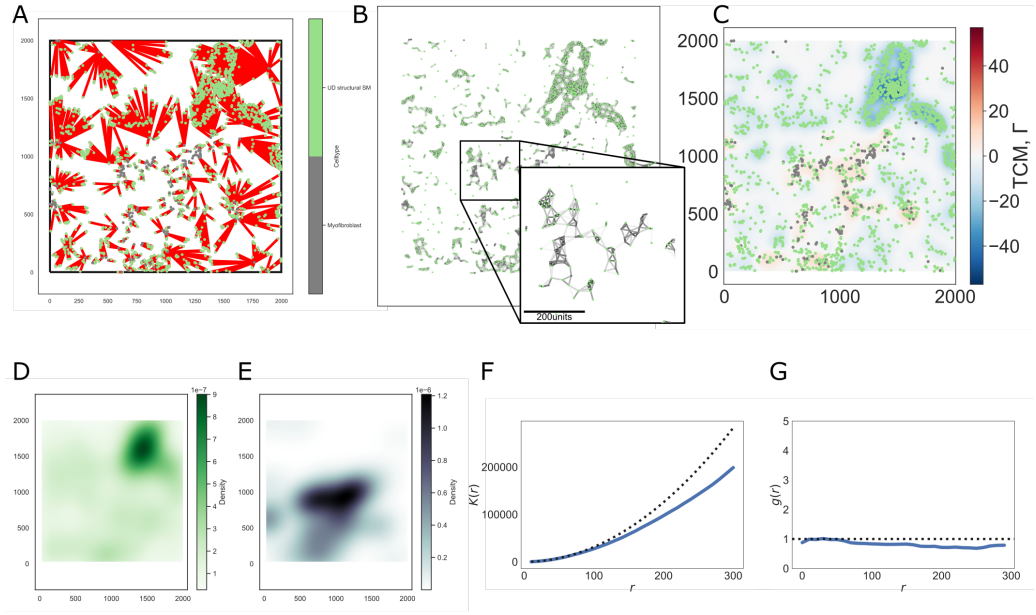

Fig. C4 MuSpAn applied to IMC data - spatial anti-correlation.
A: Nearest neighbour distances (red) from 'UD structural SM' cells to the nearest myofibroblast, two populations which are highlighted as spatially anti-correlated by the QCM in Fig. C2D.
B: Weighted Delaunay network constructed on the cell centres from the two populations of interest.
C: Topographical Correlation Map (TCM) showing where 'UD structural SM' cells are locally positively or negatively associated with myofibroblasts.
D: Kernel Density Estimation (KDE) of 'UD structural SM' cells in the ROI.
E: KDE of myofibroblasts in the ROI.
F: Cross-K function between the two cell populations, indicating decreased association at length scales up to 300µm and beyond.
G: Cross-PCF between the two cell populations, indicating lack of correlation between the populations at short length scales and anti-correlation at length scales beyond 100µm.

C.3 CODEX - Schürch *et al.* (2020)

In this section, we show how MuSpAn can be used to reproduce the analysis of CODEX tissue microarray (TMA) data in [31]. Fig. C5A shows a selection of 68 TMAs ('group 1') from this dataset. We use MuSpAn to conduct a neighbourhood enrichment study based on a k-nearest neighbours network (and assigned 10 k-means clusters using `muspan.networks.cluster_neighbourhoods()` to obtain an enrichment matrix for each cluster, Fig. C6A). The clusters obtained from this analysis are in good qualitative agreement with those obtained by Schürch *et al.* [31]. MuSpAn assigned clusters are shown in Fig. C6B and D, and the corresponding cluster assignments for the same regions from [31] are shown in Fig. C6C and E (note that since neighbourhood cluster assignment is conducted on all samples simultaneously, not every cluster is represented in every TMA).

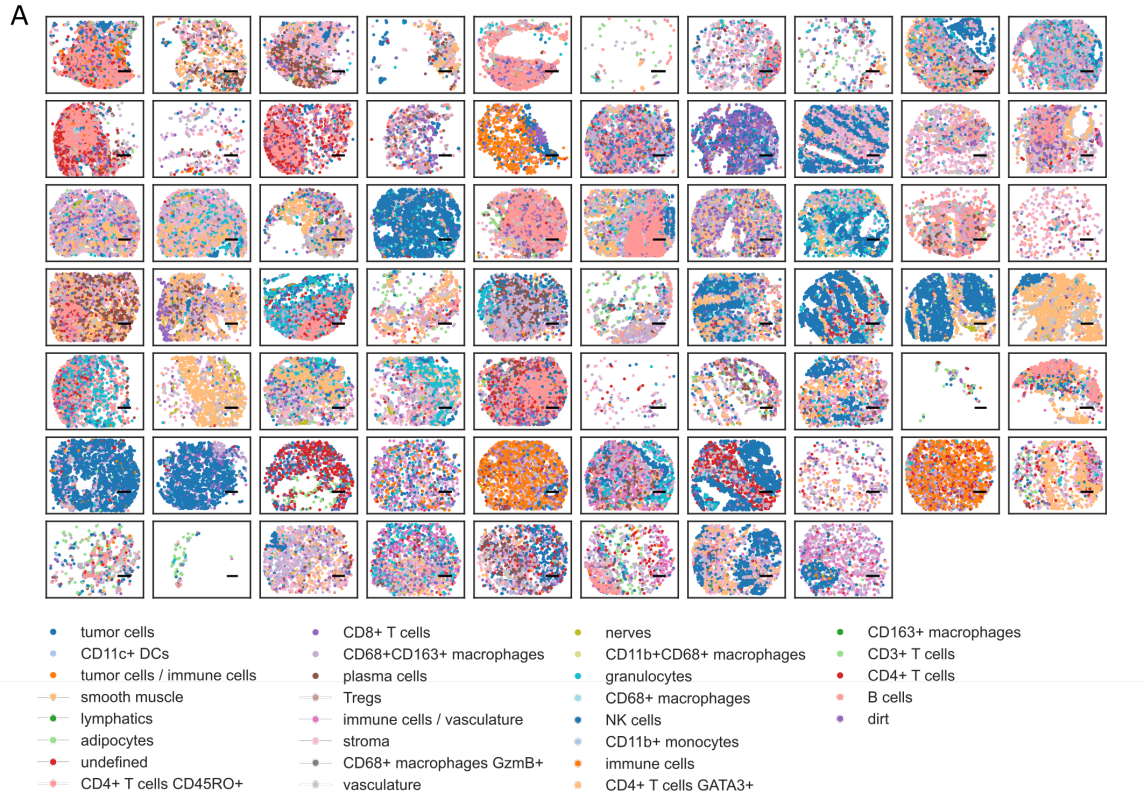

Fig. C5 MuSpAn applied to CODEX data.
A: MuSpAn visualisation of cell centres from the TMAs used in [31].

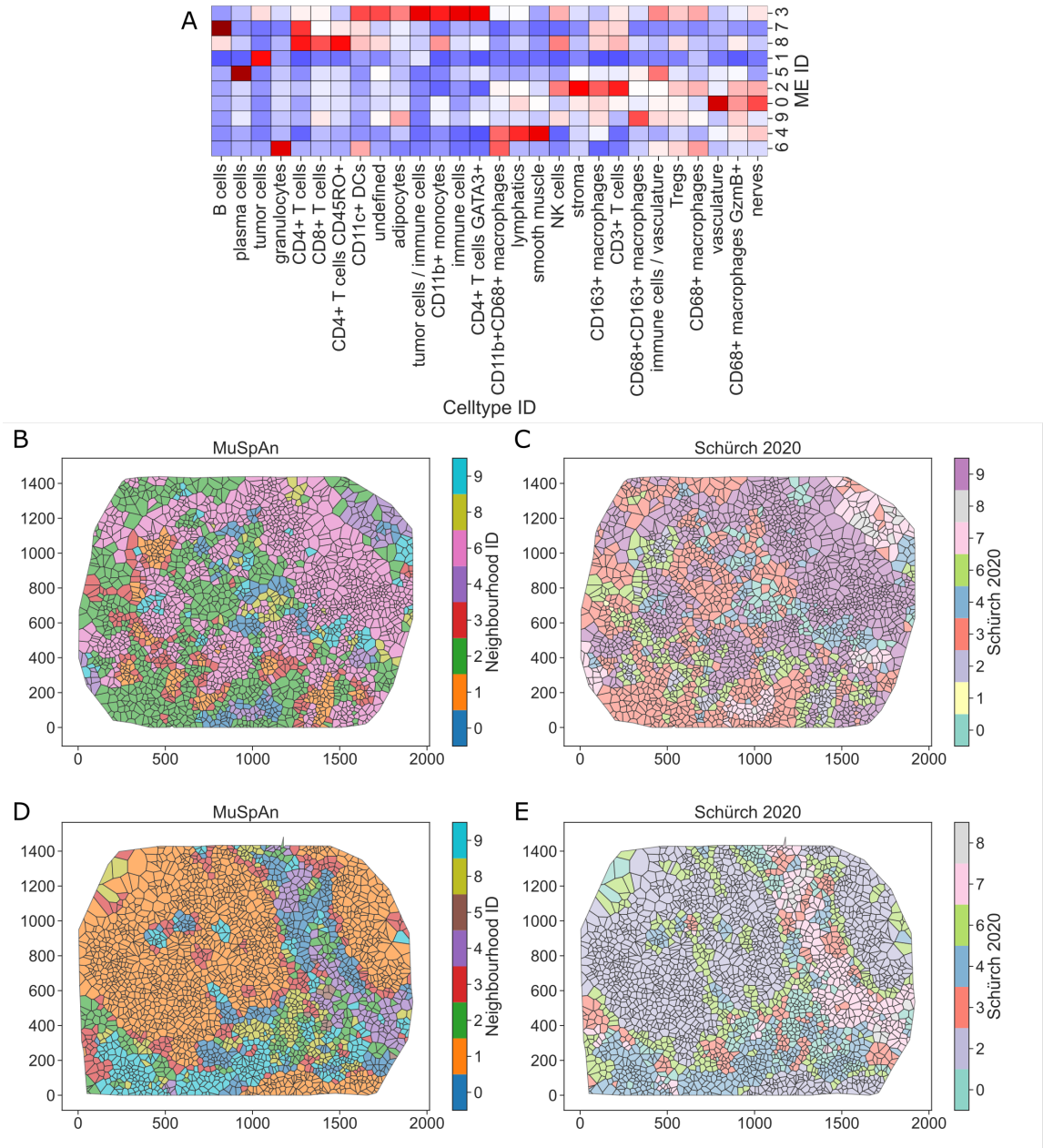

Fig. C6 MuSpAn applied to CODEX data.
A: A microenvironment enrichment matrix showing the celltypes which are enriched in each of the 10 microenvironment clusters (ME ID 0-9) detected in a MuSpAn reproduction of the Schürch *et al.* pipeline.
B-E: Voronoi representations of cell data from two TMAs (B/D and C/E), coloured according to the MuSpAn assigned cellular neighbourhoods (B, D) and to the Schürch *et al.* assigned neighbourhoods (C, E).