COMPLEX SPATIAL STRUCTURES IN BIOMEDICAL IMAGES: INVESTIGATING THE INTERACTION OF CELLS WITH THEIR MICROENVIRONMENT

by

SEBASTIAN GEORGE BELTON GILBERT

A thesis submitted to
The University of Birmingham
for the degree of
PHD WITH INTEGRATED STUDY IN APPLIED MATHEMATICS
(TOPOLOGICAL DESIGN)

School of Mathematics
College of Engineering and Physical Sciences
The University of Birmingham
December 2023

UNIVERSITY OF BIRMINGHAM

University of Birmingham Research Archive
e-theses repository



This unpublished thesis/dissertation is under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:



Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

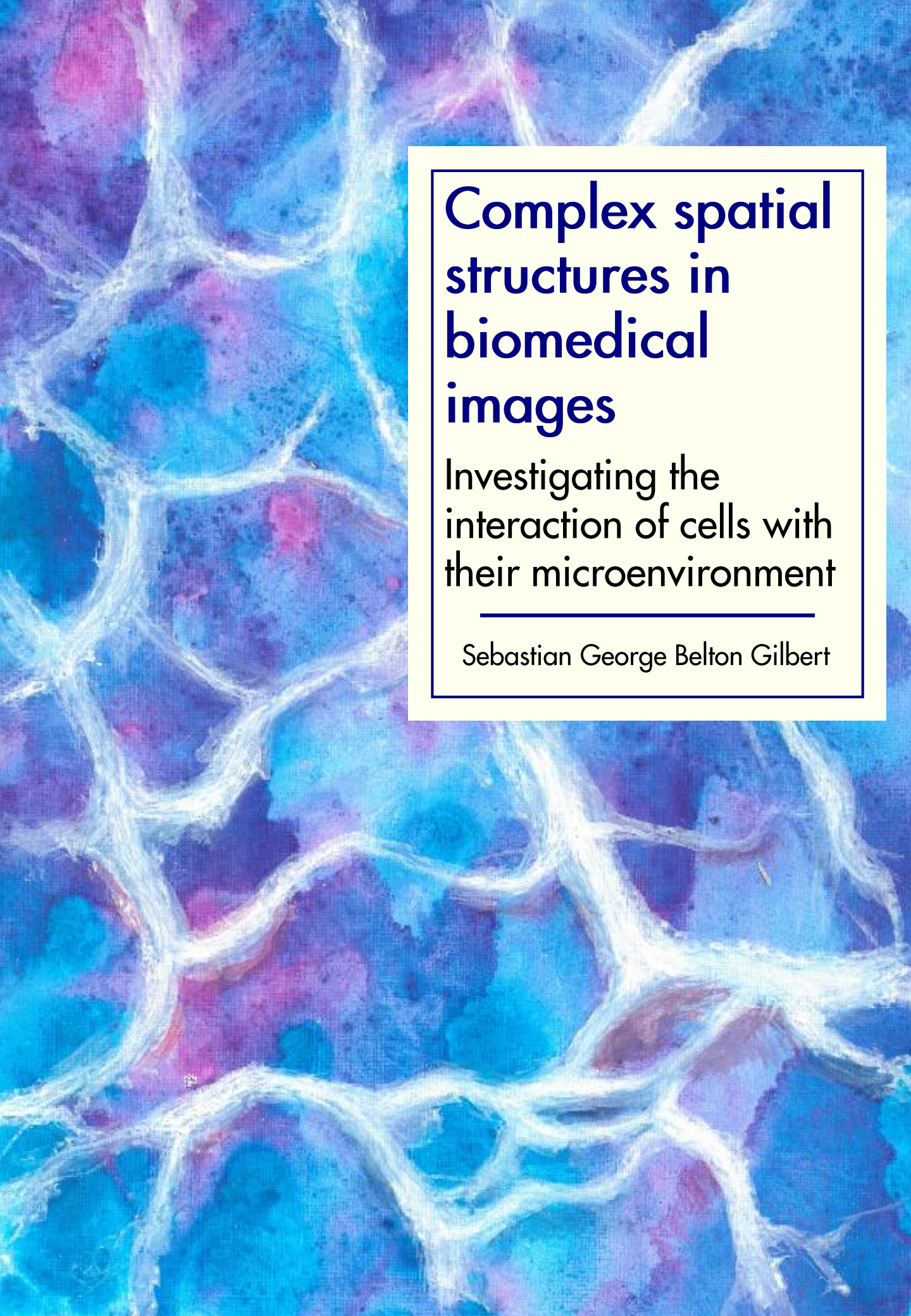
No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Unless otherwise stated, any material in this thesis/dissertation that is cited to a third-party source is not included in the terms of this licence. Please refer to the original source(s) for licencing conditions of any quotes, images or other material cited to a third party.



Complex spatial structures in biomedical images

Investigating the interaction of cells with their microenvironment

Sebastian George Belton Gilbert

From the artist…

After having the thesis explained to me by Seb, I based my design on the amazing colours and images seen through a microscope of cell interaction. For the design, I wanted a fusion of colour for the background with a white vein-type shadow going over the surface. This was based on the image seen through the microscope. For the background, I ran coloured paints over the canvas letting them mix as they travelled over the surface. When almost dry I used a dropper to run white paint lines over the mixed colours and highlighted them when dry.
My thanks to Seb for asking me to produce the design for this concept and I would like to wish him and his colleagues my sincere gratitude for the outstanding work they do.

Peter Howe

Abstract

Imaging modalities can resolve biological systems with a precision that can spatially locate individual cells of multiple cell phenotypes. Identifying meaningful patterns in heterogeneous, multi-agent and high-dimensional images grows in importance as high-throughput versions of this technology develop. Mathematical constructs offer simplifications which are then able to interpret a myriad of complex spatial structures, including point patterns, spatial complex networks and homological features. These constructs provide quantification which enables predictive power for patient disease progression and to infer optimal biological function.

Before the characterisation of images into mathematical representations, ascertaining a reliable signal amidst noise is a foundational step in describing the features of images accurately. Recent developments in machine learning have provided a powerful set of techniques for the automated analysis of complex structures in images. Leveraging these, we integrate the use of advanced image processing techniques, combining tools to construct effective processing workflows. For instance, in the identification of cell boundaries and their respective phenotypes from 2-dimensional multiplex immunofluorescence images or blood vessel structures from 3-dimensional confocal fluorescence microscopy.

Spatial analyses then encapsulate pairwise interactions, multi-phenotypic neighbourhoods and structured cellular neighbourhoods, a novel concept introduced in this research. In this thesis, we apply our techniques to 2-dimensional and 3-dimensional imaging modalities investigating features in three biological systems: Colorectal carcinoma, Sjögren's syndrome and self-organised microvascular models.

In colorectal carcinoma, we identify several immune cell phenotypes of interest for patient prognosis, such as M2 macrophages with cancerous epithelial cells. In Sjögrens syndrome we automatically identify tertiary lymphoid structures, otherwise known as ectopic lymphoid organs, a prognostic predictor of patient progression in this disease as well as in many cancers. Finally, the characterisation of self-organised microvascular models as complex spatial

networks improved the classification of samples according to the medium environment and related subsequent morphology to analytical simulations of oxygen and nutrient flow.

I dedicate this to my parents, brother and partner, for your persistent love and support.

ACKNOWLEDGEMENTS

Thank you foremost to the long list of people who have supported me through my PhD. This work would not have been possible otherwise.

First and foremost, I would like to express my sincere gratitude to my supervisor Fabian Spill, for his guidance, enthusiasm and support over the last few years. From initial explorations of projects to the years spent extracting data, coding, debugging, facing what felt like every institutional hurdle and to the thesis slowly taking its final form.

I would like to thank my second supervisor Iain Styles for his navigation of the interdisciplinary landscape, guidance on topological data analysis, and for his feedback on this thesis.

My industrial supervisor Arthur Lewis allowed me to view my research from a completely different perspective. Thank you for the many helpful discussions, interesting research visits, and the opportunity to experience pharmaceutical research from the inside.

In all parts of this thesis, I rely on experimental data from individuals who have seen the potential of my work. For Chapter 2, I am thankful to Harris Vince for identifying the potential in me and my work and to Christopher Bagnall who then had to deal with the subsequent barrage of biological questions about the data set. For Chapter 3, I am grateful to Elena Pontarini, Michele Bombardieri and Claudio Mauro for their enthusiastic depictions of tertiary lymphoid structures and help in providing me with the materials I needed. For Chapter 4, I would like to thank Sarah Shelton, Cynthia Hajal and Roger Kamm for access to the data set and guidance on pattern formation in blood vessels.

I am grateful for the co-funding from AstraZeneca after the success of my industrial placement. This allowed me to take subsequent trips to the offices at Granta Parks and Academy House, Cambridge. I was fortunate to meet the most welcoming people and would like to

especially thank Azam Hamidinekoo, Yeman Brhane Hagos, Irina Pshenichnaya, Maria Stuart, Daniel Sutton, and Lee Brown for keeping me entertained during my visits. Also, to Arthur Kennedy-Cochran-Patrick, Euan Smithers, Harriet Burdett, and Matthew Hopwood for keeping me entertained on my visits after-hours.

Thank you to all those who offered their feedback on iterations of this thesis. To my father Dave Gilbert, who spent evenings (over Zoom or in the pub) translating technical jargon into understandable science. To my mother Jenny Gilbert, for her excellent command of the English language which keeps me in check. To Sarah Friend, for her keen grammatical eye and for curbing my enthusiasm for commas. To Devi-Fetri Ferdania and Kylie Savoye who provided feedback in the hectic first few weeks of their PhDs. To Robert Maher, whose jovial support and mathematical criticisms I always look forward to.

I would like to thank Alexander Krolak, Barn Hawthorn, Emily Nevinson, Jay Morris and Ornestė Buitkutė as members of my cohort, and fellow guinea pigs, of the EPSRC Centre for Doctoral Training in Topological Design for their camaraderie in our joint endeavour. I am also thankful to Hannah Roberts, Rachel Payne, Qun Wang, Swapna Khandavalli, Ilija Rašović, Mark Dennis and the rest of the topological design family who have provided an unrivalled opportunity to pave a unique PhD experience. This would not have been possible if not for the funding from the UK Engineering and Physical Sciences Research Council (grant EP/S02297X/1).

I would like to extend my gratitude to the BEAR Services team at the University of Birmingham for providing the hardware, software, training and help to use the high-performance computer BlueBEAR. Specifically, I would like to thank Aslam Ghumra for supporting the endeavour to help me and other postgraduate researchers become better coders.

I would like to thank dungeon dwellers past and present in the School of Mathematics. You have ensured that the last few years have been enjoyable ones and there are too many of you to list.

The final word goes to the two people I love the most, my brother Alex Gilbert and partner Bethany Friend. Without you both vying to distract me from the rabbit hole of mathematics,

I would not have remained the well-rounded individual I am today.

CONTENTS

1 Introduction **1**

1.1 Literature Review of Cellular Spatial Organisation in 2-Dimensions 3

1.2 Literature Review of Blood Vessel Networks in 3-Dimensions 10

1.3 Research Questions 14

1.3.1 Identifying Biologically Significant Cellular Spatial Organisation from 2-Dimensional Immunofluorescence Imaging 16

1.3.2 Delineating Live Microvascular Models and their Structure from 3-Dimensional Confocal Imaging 18

1.4 Thesis Contributions 19

2 Exploring Multiple Phenotypic Intercellular Spatial Proximity in 2-Dimensional Multiplex Immunofluorescence Imaging **22**

2.1 Motivation 22

2.2 Biological Context 24

2.2.1 Immunology 24

2.2.2 The HALO1079 data set 26

2.3 Clinical Measures 28

2.3.1 Grading 28

2.3.2 TNM Classification of Malignant Tumours 29

2.3.3 Microsatellite Instability/Stability Status 31

2.3.4 Immunoscore 32

2.3.5 Outlook 33

2.4 Processing the Images . . . 34
2.4.1 Acquisition . . . 34
2.4.2 Data structure . . . 35
2.5 Areas, Counts and Densities . . . 37
2.5.1 Areas . . . 37
2.5.2 Counts . . . 41
2.5.3 Densities . . . 42
2.6 Compartment Interaction . . . 43
2.7 Spatial Correlation within Multitype Point Patterns . . . 47
2.7.1 Pair Correlation Function . . . 48
2.7.2 Comparing Many Permutations / Combinations of Phenotypes using Spatial Correlation Functions . . . 55
2.8 Constructing Graphs from Point Coordinates . . . 57
2.8.1 Modelling Juxtacrine Signalling . . . 58
2.8.2 Modelling Paracrine Signalling . . . 59
2.8.3 k-Nearest Neighbour Graph and Adaptations . . . 61
2.8.4 Delaunay Triangulation and Adaptations . . . 62
2.9 Cellular Neighbourhoods . . . 63
2.9.1 Infiltration Score . . . 64
2.9.2 Coordinated Cellular Neighbourhoods . . . 65
2.9.3 Neighbourhood Analysis Score . . . 67
2.10 Implementation . . . 69
2.11 Results . . . 70
2.11.1 Clinical Measures . . . 71
2.11.2 Areas, Counts and Densities . . . 74
2.11.3 Heterogeneity Scores . . . 78
2.11.4 Compartment Interaction . . . 80
2.11.5 Spatial Correlation within Multitype Point Patterns . . . 83

2.11.6 Cellular Neighbourhoods . . . 95
2.12 Summary and Outlook . . . 98
2.12.1 Summary . . . 98
2.12.2 Future work . . . 100
2.12.3 Conclusion . . . 102

3 Identifying Higher-Order Structures in 2-Dimensional Targeted Immunofluorescence Imaging 103

3.1 Motivation . . . 103
3.2 Biological Context . . . 105
3.2.1 Ectopic / Tertiary Lymphoid Structures / Organs . . . 105
3.2.2 The TRACTISS data set . . . 107
3.3 Processing the Images . . . 108
3.3.1 TRACTISS data acquisition . . . 108
3.3.2 Cell Border Identification . . . 108
3.3.3 Cell Centroid Approximation . . . 110
3.3.4 Cell Phenotype Identification . . . 110
3.4 Structured Cellular Neighbourhoods . . . 112
3.4.1 Identifying High-Density Aggregations . . . 113
3.4.2 Identifying the Boundary of Complex Geometries . . . 116
3.4.3 Density Gradients for Multitype Point Patterns . . . 123
3.4.4 Persistent Homology for Loop Identification . . . 129
3.5 Implementation . . . 133
3.6 Results . . . 134
3.6.1 Artificial Data Set . . . 135
3.6.2 TRACTISS Data Set . . . 142
3.7 Summary and Outlook . . . 147
3.7.1 Summary . . . 147
3.7.2 Future Work . . . 147

3.7.3 Conclusion 151

4 Quantifying 3-Dimensional Self-Organised Microvascular Models as Spatial Graphs for Transport Simulations 152

4.1 Motivation 152

4.2 Biological Context 154

4.2.1 The Blood Vessel Microenvironment 154

4.2.2 The Self-Organised Microvascular Model Data Set 155

4.3 Processing the Images 155

4.3.1 Workflow 1: Segmentation via Global Thresholding 157

4.3.2 Workflow 2: Segmentation via Ilastik 158

4.3.3 Validation of Image Processing 160

4.3.4 Further transformations 160

4.4 Complex Spatial Network Analysis 163

4.4.1 Geometry Metrics 163

4.4.2 Network Metrics 167

4.4.3 Dimensional Reduction Techniques 169

4.5 Analytical Solution of Flow in a Network 171

4.5.1 Hagen–Poiseuille flow from the Navier–Stokes equations 171

4.5.2 Deriving Hagen–Poiseuille flow for circular and elliptical cross-sections . 172

4.5.3 Solving Hagen–Poiseuille flow in a network 175

4.6 Results 177

4.6.1 Analytical Flow Simulations 178

4.6.2 Individual Metric Analysis 183

4.6.3 Dimensional Reduction Analysis 191

4.7 Outlook 195

4.7.1 Assumptions in Geometry Measurements 195

4.7.2 Extracting Geometry Measurements to Estimate Vessel Conductance, c_{jj} 196

4.7.3 Application of Topological Data Analysis 196

4.7.4 Conclusion . . . 197

5 Discussion and Conclusion **198**

5.1 Discussion . . . 198

5.2 Conclusion . . . 202

A Useful Mathematical Definitions **203**

A.1 Complex Spatial Networks . . . 203

A.1.1 Microscopic Network Features . . . 205

A.1.2 Mesoscopic Network Features . . . 206

A.1.3 Macroscopic Network Features . . . 208

A.2 Topological Data Analysis . . . 210

A.2.1 Motivation . . . 212

A.2.2 Complexes . . . 212

A.2.3 Building Simplicial Complexes from Point Clouds . . . 214

A.2.4 The Nerve Theorem . . . 215

A.2.5 Homology . . . 218

A.2.6 Persistent Homology . . . 220

A.2.7 Visualising Persistence . . . 221

A.3 Heterogeneity Scores . . . 225

A.3.1 Richness . . . 225

A.3.2 Shannon Index . . . 225

A.3.3 Gini-Simpson Index . . . 226

A.3.4 Hill numbers and the Diversity Profile . . . 226

A.3.5 Renyi entropy . . . 227

A.3.6 Rao's quadratic entropy . . . 228

A.4 Alternative Spatial Correlation Metrics for Multitype Point Patterns . . . 228

A.4.1 Nearest Neighbour Distribution with the G-function . . . 228

A.4.2 J-function and F, the Empty Space Distribution . . . 235

A.4.3 Ripley's K-function 240
A.4.4 Besag's L-function 245
A.5 Algorithms for Identifying High-Density Aggregations 245

B Additional Information and Results **259**

B.1 Exploring Multiple Phenotypic Intercellular Interactions in 2D Multiplex Immunofluorescence Imaging 259
B.2 Identifying Higher-Order Structures in 2D Targeted Immunofluorescence Imaging 261
B.2.1 Clustering Evaluation Scores on Benchmark Data Set 261
B.2.2 Parameters used in the Application of the Structured Cellular Neighbourhood Workflow 264
B.2.3 Boundary Evaluations of the Individual Structured Cellular Neighbourhoods identified in the example Artificial Data Set 264
B.3 Quantifying 3-Dimensional Self-Organised Microvascular Models as Spatial Graphs for Transport Simulations 265
B.3.1 Image Processing Validation 265
B.3.2 Individual Measurements 269
B.3.3 Multi-Dimensional Reduction analysis 271

List of References **277**

LIST OF FIGURES

1.1 An outline of the work completed in each chapter. In general, each chapter considers a respective data set which is then processed into its constituent elements before applying current and novel metrics. (a) The HALO1079 data set. (b) The constituent cell coordinates and respective phenotypes. (c) The graph construction, via Delaunay Triangulation, of the cell coordinates. (d) A figure from the novel Extent of Compartment Interaction analysis. (e) A figure from the Neighbourhood Analysis Score analysis. (f) The TRACTISS data set. (g) An example of the Loop Identification step in the Structured Cellular Neighbourhood workflow. (h) A figure of the output from the Structured Cellular Neighbourhood workflow. (i) The Self-Organised Microvascular Model data set. (j) A figure of the constituent nodes (branching points) and edges (vessels) in the 3D graph overlaid on a 2D slice of the segmented image. (k) A figure of the novel Analytical Solution of Flow in a Network analysis. (l) A figure of the Dimensional Reduction Analysis output beased on the geometric and topological features of the samples. 4

1.2 From Grigoryan et al. [1]. Reprinted with permission from AAAS. Original caption: Entangled vascular networks. (A to D) Adaptations of mathematical space-filling curves to entangled vessel topologies within hydrogels (20 wt % PEGDA, 6 kDa): (A) axial vessel and helix, (B) interpenetrating Hilbert curves, (C) bicontinuous cubic lattice, and (D) torus and (3,10) torus knot (scale bars, 3 mm). (E) Tessellation of the axial vessel and its encompassing helix along a serpentine pathway. The photograph is a top-down view of a fabricated hydrogel with oxygen and RBC delivery to respective vessels. During perfusion, RBCs change color from dark red (at the RBC inlet) to bright red (at the RBC outlet) (scale bar, 3 mm). Boxed regions are magnified in (F) (scale bar, 1 mm). (G) Perfused RBCs were collected at the outlet and quantified for SO2 and PO2. Oxygen flow increased SO2 and PO2 of perfused RBCs compared with deoxygenated RBCs perfused at the inlet (dashed line) and a nitrogen flow negative control. 15

2.1 Processed data of the HALO1079_18 sample. (a) Validated mIF image (b) The point cloud for the CT and IM. 35

2.2 The annotation boundaries of a sample (HALO1079_18) for the central tumour (orange) and the invasive margin (green) with which the areas intersect each other (yellow). 38

2.3 Comparison of identified boundaries (yellow line) for sample HALO1079_18 on the whole slide, and for a ROI (green rectangle). Whole slide images are presented using the (a) convex hull, (b) concave hull (ratio = 0.01), and (c) alpha shape ($\alpha = 75\mu$m). A ROI is also presented for the same boundary identified by the (d) convex hull, (e) concave hull, and (f) alpha shape. 40

2.4 Counts of CD68+ cells of a distance from the IM boundary. 45

2.5 Areas of binned distance, shown for the HALO1079_18 sample, with the maximum distance of each binned distance shown as the colour of the area within (-200, 400). Beyond this range, the CT and IM are coloured orange and green, respectively. The black solid line represents the identified CT-IM boundary interface for which distances are measured. 46

2.6 The spatial point patterns used to illustrate the interpretation of the following spatial correlation metrics. (a) The spatial point pattern if the points are uniformly random in the domain, following a spatial Poisson point process with similar cell type densities to our observed point patterns. The subsequent plots are when Cell Type 1 has 75% of the points distributed according to a Gaussian distribution with mean at $(0, 0)$ and standard deviation $\sigma = 50$. The remaining 25% of points are uniformly randomly distributed. Then, Cell Type 2 is distributed such that (b) it has the same underlying distribution as Cell Type 1, (c) it has a placement probability proportional to the density of Cell Type 1, or (d) it has a placement probability inversely proportional to the density of Cell Type 1. All spatial point patterns consider the points as centroids of hard-shelled cells where the radius is $r = 10$. 49

2.7 The PCF plot for the (a) independent, (b) proportional, and (c) inverse spatial point pattern. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level. 51

2.8 The spatial point patterns used to illustrate the use of spatial correlation metrics to interpret spatial heterogeneity. Cell Type 1 has 75% of the points distributed according to a Gaussian distribution with centres on a hexagonal lattice (marked here with a black cross) and standard deviation $\sigma = 10$. The remaining 25% of points are uniformly randomly distributed. Then, Cell Type 2 is distributed such that it has a placement probability proportional to the density of Cell Type 1. All spatial point patterns consider the points as centroids of hard-shelled cells where the radius is $r = 5$. 53

2.9 Results of spatial correlation metrics to interpret spatial heterogeneity. The plots for the multitype (a) - (c) Ripley's K-function, (d) - (f) Besag's L-function, and (g) - (i) PCF pairwise comparisons. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level. 54

2.10 ROI to exemplify graph constructions. (a) The point cloud of HALO1079_18 with an overlay (green) of the ROI. (b) Focused view on the ROI. 58

2.11 ROI from sample HALO1079_18 for which juxtacrine signalling has been modelled as interactions occurring when $r < 10\mu$m. 60

2.12 ROI from sample HALO1079_18 for which paracrine signalling has been modelled as (a) interactions occurring for $r < 35\mu$m, and (b) the diffusion coefficient defined by Eq. 2.6 is attributed to the weight of edges, depicted as the proportional relationship to the transparency of an edge. For (b) a threshold on edge weight has been applied for eligibility, meaning only the interactions with a significantly large enough diffusion coefficient remain. 61

2.13 ROI from sample HALO1079_18 for which the kNN graph is constructed, with (a) k$= 5$,and (b) k$= 10$. 62

2.14 ROI from sample HALO1079_18 for which the Delaunay Triangulation is constructed, with (a) no maximum edge length, and (b) maximum edge length $\max(l) = 35\mu$m. 63

2.15 An exemplary set of cells a related graph construction. We consider a single neighbourhood centred on a single cell (orange star) and its neighbours (solid black lines) as well as the global average for all other interactions (dashed black lines). (a) A permutation of cell phenotypes where immune cells (purple speckled frilled cells) are excluded from most epithelial cell neighbourhoods (red homogeneous smooth cells). (b) A permutation of cell phenotypes where immune cells (purple speckled frilled) are mixed with epithelial cell neighbourhoods (red homogeneous smooth cells). 68

2.16 Grading of $n = 90$ patients, comparing the Responders to Non-Responders. The χ^2 test was performed at the 95% confidence level. 72

2.17 TNM Classification of the $n = 90$ patients, comparing the percentage of Responders and percentage of Non-Responders in each stage of TNM Classification. 72

2.18 The Microsatellite Status for the $n = 90$ patients, comparing Responders and Non-Responders. 73

2.19 The Immunoscore-like percentile for the $n_{(\text{CT, IM})} = 43$ patients, based on only the CD8 cells, comparing Responders and Non-Responders. 74

2.20 Area measurements of the (a) convex, (b) classified area of the central tumour by HALO, (c) concave (ratio = 0.01). The measurements of (a) and (b) represent the upper and lower measurements of the area, respectively, while the concave hull measurement represents a more accurate estimate. 75

2.21 Counts of the cell phenotypes for $n_{(\text{CT, IM})} = 43$ according to the different decompositions shown in Table 2.4. (a) The series of figures that represent the counts of all the cell marker combinations that are possible in the HALO1079 data set. (b) The series of figures that represent the counts of all the cell types. (c) The total number of all cells, regardless of cell marker (including cells for which only the nuclei stain DAPI is identified). (d) The number of cells that are either HLA-DR+ or HLA-DR-. (e) The number of cells that present cell markers categorised as immune or epithelial phenotypes. Through this analysis, we were able to identify poor quality cell marker identification for sample HALO1079_025 and HALO1079_063, with their corresponding IM annotations HALO1079_026 and HALO1079_64, which have since been excluded. 76

2.22 Densities of the cell phenotypes, for $n_{(\text{CT, IM})} = 43$ patients, according to the different decompositions shown in Table 2.4. The counts were normalised by the area from the concave hull. (a) The series of figures that represent the densities of all the cell marker combinations that are possible in the HALO1079 data set. (b) The series of figures represent the densities of all the cell types. (c) The density of all cells, regardless of cell marker (including cells for which only the nuclei stain DAPI is identified). (d) The density of cells that are either HLA-DR+ or HLA-DR-. (e) The density of cells that present cell markers is categorised as immune or epithelial phenotypes. 77

2.23 Heterogeneity scores for the $n_{(\text{CT, IM})} = 43$ patients, comparing Responders and Non-Responders. The figures represent the values calculated for the samples by the (a) Richness, (b) Shannon Index, (c) Simpson Index, (d) Renyi Entropy, (e) Rao Entropy. 79

2.24 Heterogeneity scores for the $n_{(\text{CT, IM})} = 43$ patients, comparing Responders and Non-Responders. We calculate the Hill numbers for the set of exponents $\alpha \in \{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$ 80

2.25 Mean ECI curves of epithelial cells for $n_{(\text{CT, IM})} = 43$ patients. 81

2.26 Mean ECI curves of T cells for $n_{(\text{CT, IM})} = 43$ patients. 82
2.27 Mean ECI curves of macrophages for $n_{(\text{CT, IM})} = 43$ patients. 84
2.28 Summary of PCF values between patient responses for pairs of cell markers. Some cell markers have been excluded from PCF analysis due to their low frequencies across most samples. The weighted area difference for pairs of cell markers is symmetric along the diagonal due to the symmetric property of the PCF. Weighted area difference above zero indicates that the value of the PCF is more different from $g(r) \equiv 1$ in Responders than Non-Responders. Weighted area difference below zero indicates values of the PCF is more different from $g(r) \equiv 1$ in Non-Responders than Responders. 86
2.29 PCF plots for cell marker pairs that produced the highest weighted area differences, i.e. PCF values with larger differences to the spatial Poisson point process in Responders than Non-Responders for $r \in [0, 100]$. Here we present the pairwise comparisons for (a) CD8+ cells to CD68+ CD163+ cells (weighted area difference $= 52.3$), (b) CD8+ cells to CD68+ cells (weighted area difference $= 38.4$), (c) CD68+ HLA-DR+ cells to CD163+ cells (weighted area difference $= 32.7$), and (d) CD8+ cells to CD163+ cells (weighted area difference $= 28.2$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group. 87

2.30 PCF plots for cell marker pairs that produced the lowest weighted area differences, i.e. PCF values with larger differences to the spatial Poisson point process in Non-Responders than Responders for $r \in [0, 100]$. Here we present the pairwise comparisons for (a) CD163+ cells to CD68+ CD163+ cells (weighted area difference $= -224.5$), (b) CD163+ cells to CD163 CD163+ HLA-DR+ cells (weighted area difference $= -144.2$, (c) CD8+ cells to CD8+ HLA-DR+ (weighted area difference $= -75.0$, and (d) CD8+ cells to CD68+ CD163+ HLA-DR+ cells (weighted area difference $= -56.3$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group. 88

2.31 PCF plots for epithelial markers (i.e. CK+ and CK+ HLA-DR+) when comparing between patient responses. Here we present the pairwise comparisons for (a) CK+ cells to CK+ cells (weighted area difference $= -2.69$), (b) CK+ cells to CK+ HLA-DR+ cells (weighted area difference $= 11.5$), and (c) CK+ HLA-DR+ cells to CK+ HLA-DR+ cells (weighted area difference $= -15.5$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group. . . 91

2.32 PCF plots for cell marker pairs that produced the largest absolute weighted area differences with CK+ phenotypes when comparing between patient responses. Here we present the pairwise comparison for CK+ cells to (a) CD163+ HLA-DR+ cells (weighted area difference $= 15.5$), (b) CD163+ cells (weighted area difference $= 13.0$), (c) CD8+ cells (weighted area difference $= 12.4$), and (d) CD68+ HLA-DR+ cells (weighted area difference $= 10.2$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group. 92

2.33 PCF plots for within cell phenotype comparison to confirm the characteristic distribution of CK+ cells to immune cells of interest from Figure 2.31. Here we present the pairwise comparison within (a) CD163 HLA-DR+ cells (weighted area difference $= 23.2$), (b) CD163+ cells (weighted area difference $= 11.1$), and (c) CD68+ HLA-DR+ cells (weighted area difference $= 18.3$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group. 93

2.34 PCF plots for cell marker pairs that produced the largest absolute weighted area differences with CK+ HLA-DR+ phenotypes when comparing between patient responses. Here we present the pairwise comparisons for CK+ HLA-DR+ cells to (a) CD68+ CD163+ HLA-DR+ cells (weighted area difference $= -25.2$, and (b) CD8+ cells (weighted area difference $= -14.4$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group. 94

2.35 Neighbourhood Analysis Scores of macrophage neighbours to CK+ cells comparing patient response for the $n_{(\text{CT, IM})} = 43$ patients. 97

2.36 Neighbourhood Analysis Scores for between sub-populations of T cells comparing patient response for the $n_{(\text{CT, IM})} = 43$ patients. 98

2.37 Neighbourhood Analysis Scores of the remaining statistically significant differences, at the 95% confidence level, between patient responses for the $n_{(\text{CT, IM})} = 43$ patients. 99

3.1 An example TLS from the TRACTISS data set. (a) A focused view on a TLS from Figure 3.2. (b) - (d) The individual markers used for the identification of TLS in the TRACTISS data set. The markers are (b) DAPI, (c) FITC, and (d) TRITC. They identify the nucleus of cells positive for CD20 and CD3, which indicate B cells and T cells, respectively. 106

3.2 Whole slide image of a patient's salivary gland sample for the inspection of Sjögren's syndrome. 107

3.3 The application of cell border detection algorithms to the TLS in Figure 3.1. (a) The algorithms were applied to the ROI. (b) - (d) presents the outline (red) of the nucleus used as the centroid of the cells. We applied the (b) Watershed, (c) StarDist [2], and (d) Cellpose [3] algorithms. 111
3.4 Ground truth for our benchmark dataset. 118
3.5 Evaluations of the convex hull boundaries. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively. 119
3.6 Evaluations of the concave hull boundaries. The respective ratio parameters found to produce the highest F_1 score were (a) $\varphi = 0.21$, (b) $\varphi = 0.30$, and (c) $\varphi = 0.17$. The evaluation scores were computed for ratio, $\varphi \in \{n\Delta x \quad | \quad n \in 1, 2, ..., 100\}$ and $\Delta x = 0.01$, with holes allowed in each case. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively. 121
3.7 Evaluations of the alpha hull boundaries. The respective alpha parameters found to produce the highest F_1 score were (a) $\alpha = 0.904$, (b) $\alpha = 0.511$, and (c) $\alpha = 0.753$. The evaluation scores were computed for alpha, $\alpha \in \{$all edge lengths in the Delaunay triangulation of the point coordinates$\}$. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively. . . 122
3.8 Evaluations of the alpha shape boundaries. The respective alpha parameters were (a) $\alpha = 0.733$, (b) $\alpha = 0.418$, and (c) $\alpha = 0.522$. The evaluation scores were computed for alpha, $\alpha \in \{$all edge lengths in the Delaunay triangulation of the point coordinates$\}$. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively. 122
3.9 Results of the density gradient on the artificial data set, presented in Chapter 2, Section 2.7. The plots show the density gradient results for each cell type for the high-density aggregations identified and their respective boundary (black line) in the (a) independent, (b) proportional, and (c) inverse point patterns. Only points within 500μm of the aggregation boundary are shown. 127

3.10 Loop identification for the proportional point pattern and the intermediary computation steps. (a) PH is computed on the area surrounding the aggregation, excluding the points within the aggregation, and identifies all the features of the 0-th dimensional homology group (orange circles) and 1-st dimensional homology group (blue triangles). (b) Excluding the points within the border of the high-density aggregation (yellow circles/triangles), we construct all loops of the Cell Type 2 point cloud which represent the midpoint of vertices that construct a cycle. (c) A threshold is applied for persistence, $\tau = 20$ (black dashed line) and a threshold is applied for the minimum vertex-vertex distance (gold dash-dot line) between points which construct the cycle, obtaining an area (yellow region) in which possible loops are considered, and (d) is the identified loop (blue line) for the set of applied thresholds, which has been confirmed to encompass the coordinates of the original high-density aggregation (yellow circles/triangles). 132

3.11 An example of the artificial data set. (a) Presents the cell types present. (b) Presents the ground truth for the cores of each TLS. 137

3.12 The results of the structured cellular neighbourhood workflow on a sample from the artificial data set. The core of the structured cellular neighbourhood is highlighted in orange with an index overlaid. The threshold for the inside and outside density gradients are presented, respectively. 138

3.13 Individual structured cellular neighbourhoods identified by our workflow. The high-density aggregation boundary is outlined. Only points within 100μm of the aggregation boundary are shown. 'Index' refers to the Index given to the structured cellular neighbourhood in Figure 3.12. 'Cluster param.' refers to the set of cluster parameters used to identify the specific aggregation. Generally, higher values of 'Cluster param.' refer to larger values of the threshold followed by the radius considered for nearby points in the ToMATo clustering algorithm[4]. The 'Aggregation' refers to which aggregation identified for the given 'Cluster param.' is presented. Finally, the observed values for the density gradient for inside, i.e. for Cell Type 1, and outside, i.e. for Cell Type 2, are shown. 139

3.14 Boundary evaluations for individual identified structured cellular neighbourhoods. The F_1, precision and recall scores are presented in Appendix B (Table B.7). Index refers to the aggregation identified by the structured cellular neighbourhood workflow. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively 141

3.15 The results of the structured cellular neighbourhood workflow on a sample from the TRACTISS data set. Here we show the N00176_001B_2 sample. (a) The point cloud, and (b) the identified structured cellular neighbourhoods. (c) mIF image of the validation markers CD21 and CD138. 143

3.16 Individual structured cellular neighbourhoods identified by our workflow. The high-density aggregation boundary is outlined. Only points within 100μm of the aggregation boundary are shown. 'Index' refers to the Index given to the structured cellular neighbourhood in Figure 3.15b. 'Cluster param.' refers to the set of cluster parameters used to identify the specific aggregation. Generally, higher values of 'Cluster param.' refer to larger values of the threshold followed by the radius considered for nearby points in the ToMATo clustering algorithm[4]. The 'Aggregation' refers to which aggregation identified for the given 'Cluster param.' is presented. Finally, the observed values for the density gradient for inside, i.e. for CD20, and outside, i.e. for CD3, are shown. 145

3.17 Individual TLSs identified in sample N00176_001B_2 of the TRACTISS data set with cell markers CD21 (red) and CD138 (yellow). (a) - (c) Focused view on each TLS. 146

4.1 An illustration of the experimental set-up and data set acquired for the following section. (a) A diagram of the microfluidic set-up with an endothelial cell and pericyte co-culture. (b) Image slice captured by brightfield microscopy of the microfluidic setup from the endothelial cell and pericyte co-culture sample. (c) Image slice captured by confocal microscopy via fluorescence of the endothelium, using a lectin live staining of the luminal wall. (d) Image slice captured by confocal microscopy while the fluorescent dye is perfused through the blood vessels network. Images (b) and (d) are captured at the same time point, while (c) is captured at an earlier time point. Scale bars indicate 200 microns 156

4.2 The results of pre-processing on a ROI. (a) The original ROI with indexed markers indicating voids outside the blood microvasculature. (b) The original ROI with indicated line segment. (c) The post-processed ROI with indicated line segment. The yellow line segment measures 150 microns across. 159

4.3 The relative pixel intensity values from the yellow line segment in Figure 4.2. Showing the original ROI and the same ROI after pre-processing. The grey horizontal line shows an example threshold where pixel intensities are separated above and below for inside and outside, respectively, after applying the Kuwahara filter but not for the line segment in the original ROI. 159

4.4 A comparison of unprocessed and processed images via Workflow 1. (a) and (c) are unprocessed 2D slices taken from the 3D volumetric image stack. (b) and (d) are the processed final binary image used in subsequent processing and analysis. White (i.e. pixel value equal to 1) is the blood vessel volume and black (i.e. pixel value equal to 0) for the surroundings. (a) - (b) EC+PC-1. (c) - (d) EC+PC-2. Scale bars indicate 200 microns. 161

4.5 An example of the further transformations performed on one of the data sets, namely the Skeletonisation and annotations according to the Graph representation. (a) A slice of the original image after processing, as specified in Section 4.3. (b) The 3D skeletonisation of the whole blood vessel network with annotations according to the graph representation. Here, nodes with only one edge (coloured black) are coloured cyan while nodes with more than one edge connected are coloured yellow with respective edges coloured red. (c) An overlay of the 3D skeletonisation on a slice of the original image. 164

4.6 A volume rendering of a processed image (i.e. binary) where the cross-section appears elliptical rather than circular. The whitest area on the left is the cross-section and the rest of the vessel curves behind and to the right. 166

4.4 The results of the analytical flow results on the skeleton of the network. The volumetric flow was calculated for each edge according to the analytical derivation. Each edge was then normalised according to the largest volumetric flow rate and coloured according to the `jet` colour scheme, where the warmest colours (e.g. red) being the largest flows and the coolest (e.g. blue) being the lowest with black as near zero flow rates. Nodes are coloured according to whether they are an inlet (red), outlet (blue) or neither (magenta). (a) and (b) are for the microassay cultures with only endothelial cells; (c) and (d) are for the microassay cultures with endothelial and astrocytes; (e) and (f) are for the microassay cultures with endothelial and pericytes; (g) and (h) are the microassay cultures with all three cell types (i.e. endothelial, astrocyte and pericyte). 181

4.5 Relative volumetric flow of individual vessels defined by the skeletonisation. We have the distributions for (a) Endothelial cells only, (b) Endothelial and Astrocyte cells, (c) Endothelial and Pericyte cells, and (d) Endothelial, Astrocyte and Pericyte cells. To obtain the results as a relative measurement within each sample, all vessel flows have been divided by the maximum flow calculated for a given vessel. During this computation the conductance of every edge was set equal to 1. 182

4.6 Distributions of vessel radii estimates, including the (a) Minimum distance from the skeleton voxel to the nearest background voxel in 3D, (b) Local Thickness at the skeleton voxels, and (c) Minimum distance from the skeleton voxels to the nearest background pixel in the 2D X-Y Cross Section. 186

4.7 Distributions of vessel length for individual microassay samples. 187

4.8 Results from PCA multi-dimensional reduction where (a) excludes the network-based metrics and (b) includes the network-based metrics. 193

4.9 Results from UMAP multi-dimensional reduction where (a) excludes the network-based metrics and (b) includes the network-based metrics. 195

A.1 From left to right: a vertex, an edge, a triangle, and a tetrahedron; which represent the 0-, 1-, 2- and 3-dimensional simplex, respectively. 213

A.2 The (a) Čech complex, $\mathcal{C}_\epsilon$, and (b) Vieterois-Rips complex, $\mathcal{R}_\epsilon$, for the same point cloud data set for the same ϵ. Note the difference in dimension for the two simplicial complexes formed under the two different definitions. For $\mathcal{C}_\epsilon$, we have four edges and two triangles but for $\mathcal{R}_\epsilon$ we have one edge, one triangle and a tetrahedron. 216

A.3 The different ways the results of PH can be visualised and fed into further analysis of the topological features. (a) - (d) Depiction of the Čech complex is visualised for significant values of $\epsilon = \{1, 5, 7, 9.5\}$, respectively. Inspired by the example from [5]. (e) A barcode for the point cloud data set is shown in (a) - (d). (f) A persistence diagram for the point cloud data set is shown in (a) - (d), with a threshold at $t = 3$ (black dashed line). In (e) and (f) we show features of the 0-th dimensional homology group (orange bars / orange circles) and features of the 1-st dimensional homology group (blue bars / blue triangles). 223

A.4 Landscape and Silhouette of the 1-st dimensional homology group features of the example. 224

A.5 G-function plots for (a) - (d) independent, (e) - (h) proportional, and (i) - (l) inverse spatial point patterns. Since pairwise comparisons are asymmetric, we have the G-function plot for each point pattern for Cell Type 1 to Cell Type 1, Cell Type 2 to Cell Type 2, Cell Type 1 to Cell Type 2, and Cell Type 2 to Cell Type 1. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level. 231

A.6 J-function plots for (a) - (d) independent, (e) - (h) proportional, and (i) - (l) inverse spatial point patterns. Since pairwise comparisons are asymmetric, we have the J-function plot for each point pattern for Cell Type 1 to Cell Type 1, Cell Type 2 to Cell Type 2, Cell Type 1 to Cell Type 2, and Cell Type 2 to Cell Type 1. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level. 237

A.7 Ripley's K-function plots for (a) - (c) independent, (d) - (f) proportional, and (g) - (i) inverse spatial point patterns. Since pairwise comparisons are symmetric, we have the K-function plot for each point pattern for Cell Type 1 to Cell Type 1, Cell Type 2 to Cell Type 2, and Cell Type 1 to Cell Type 2. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level. 242

A.8 Besag's L-function plots for (a) - (c) independent, (d) - (f) proportional, and (g) - (i) inverse spatial point patterns. Since pairwise comparisons are symmetric, for each point pattern we have the L-function plot for Cell Type 1 to Cell Type 1, Cell Type 2 to Cell Type 2, and Cell Type 1 to Cell Type 2. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level. . . . 246

A.9 The heuristics used to determine k for the k-Means clustering algorithm where (a) presents the Total Within-Cluster Sum of Squares value, (b) presents the Gap Statistic, (c) presents the Silhouette Score for $k \in \{1, ..., 10\}$. (d) Ground truth clustering of the 'aggregation' point pattern [6] with the results of the k-Means algorithm for (e) $k = 4$, and (f) $k = 6$, determined by the Gap Statistic and Silhouette Score, respectively. 249

A.10 Computation steps in the DBSCAN clustering algorithm on the 'aggregation' point pattern [6]. (a) Detection density estimates are calculated by counting the number of other points within `Eps` = 1.00. (b) Classification of core, reachable and noise points for `MinPts` = 5. 254

A.11 Results of the DBSCAN algorithm on an artificial point pattern. (a) Ground truth clustering of the 'aggregation' point pattern [6]. (b) Clusters identified by the DBSCAN clustering algorithm, for Eps = 1.070047 and MinPts = 5, (ARI = 0.959). 254

A.12 Computation steps in the ToMATo clustering algorithm on the 'aggregation' point pattern [6]. (a) Detection density estimates are calculated by counting the number of other points within $r = 1.097725$. (b) Peaks are identified via the graph-based hill-climbing step. (c) Persistence diagram, where we indicate the number of clusters with the same birth-death values. 258

A.13 Results of the ToMATo algorithm on an artificial point pattern. (a) Ground truth clustering of the 'aggregation' point pattern [6]. (b) Clusters identified by the ToMATo clustering algorithm, for r = 1.097725 and threshold = 3.183673, (ARI = 0.987). 258

B.1 Results from the ECI analysis. (a) CD163+. (b) CD163+ HLA-DR+. (c) DC-LAMP+ HLA-DR+ (d) DC-LAMP+ XCR1+ HLA-DR+. 261

B.2 Results from the ECI analysis. (a) XCR1+. (b) XCR1+ HLA-DR+. (c) XCR1+ CD163+ HLA-DR+. 262

B.3 Results from the image processing. (a) - (b) EC-1. (c) - (d) EC-2. Scale bars indicate 200 microns. 266
B.4 Results from the image processing. (a) - (b) EC+AC-1. (c) - (d) EC+AC-2. Scale bars indicate 200 microns. 267
B.5 Results from the image processing. (a) - (b) TC-1. (c) - (d) TC-2. Scale bars indicate 200 microns. 268
B.6 Distributions of tortuosity for individual samples, for three different measures of tortuosity, (a) Arc-Chord ratio / Distance Measure, (b) Number of Inflections plus one, and (c) ICM. 271
B.7 Distributions of network metrics for individual samples, including the (a) Degree, (b) Clustering coefficient, (c) Normalised Betweenness, and (d) Flow coefficient. 272

LIST OF TABLES

2.1 Table of markers used in the mIF imaging of the HALO1079 data set and the cell it is associated to in this study, with a description of the role of the cell in the TME. 27

2.2 Reprinted with permission from American Cancer Society's website [7]. The system described is the most recent American Joint Committee on Cancer system effective January 2018 [8]. It uses the pathologic stage (also called the surgical stage) which is determined by examining tissue removed during an operation. This is also known as surgical staging. This is distinct from clinical staging, which takes into account the results of a physical exam, biopsies, and imaging tests, before surgery. The following additional categories are not listed in the table above: TX, main tumour cannot be assessed due to lack of information; T0, no evidence of a primary tumour; NX, Regional lymph nodes cannot be assessed due to lack of information. 30

2.3 The range of values for each clinical measure and their likely prognosis. Acronyms: NA, Not Applicable. 33

2.4 Table of all possible cell marker combinations in the HALO1079 data set, and their decomposition into cell type, Immune/Epithelial phenotype and HLA-DR status. 42

2.5 Table of the Neighbourhood Analysis Scores for Immune cell neighbours of Epithelial cells for both permutations (*Excluded* and *Mixed*) of the exemplary set of cells with a graph construction. For each permutation of cell phenotype, the classic, HistoCAT and proportional averages of the Neighbourhood Analysis Score were calculated. To aid interpretation, the score is presented for a single neighbourhood centred on the indicated epithelial cell (*Local*) and then the average of the whole network (*Global*). 68

3.1 Number of data sets which either ToMATo,DBSCAN, or both produced the highest ARI score. ARI for individual data sets are presented in Table B.2 . . . 115

3.2 The evaluation score for each algorithm and each underlying distribution to identify given the parameter which produced the highest F_1 score. 123

4.1 Table of measurements for the volume and surface area for the largest functional microvasculature, assumed to be the largest connected endothelial positive object, in comparison to the rest of the endothelial negative volume, also known as background. 185

4.2 Table of summary statistics for the measurement of vessel lengths, in voxels. . 187

4.3 Table of summary statistics for the measurements of tortuosity. 190

B.1 Table of phenotype similarity used to calculate Rao's entropy for the HALO1079 data set. 260

B.2 Comparison of ARI for the ToMATo and DBSCAN clustering algorithms. . . . 263

B.3 Table of parameters specific to the identification of high-density aggregations. 264

B.4 Table of parameters specific to the identification of boundaries of high-density aggregations. 264

B.5 Table of parameters specific for the quantification of gradient densities from the individual boundaries of high-density aggregations. 265

B.6 Table of parameters specific to identify a loop around high-density aggregations to validate a surrounding cell phenotype. 265

B.7 The values of the boundary evaluations of the individual structured cellular neighbourhoods identified on the example artificial data set. TP, FP, and FN stand for True Positive, False Positive and False Negative, respectively. 269
B.8 Table of measurements for the radii estimations of the microassay samples. . . 270
B.9 Table of measurements for the topology of the microassay samples. 270
B.10 Table of the standard deviations, proportion of variance and cumulative proportion of the Principal Components in the PCA without network metrics. Standard deviation is the square roots of the eigenvalues of the covariance/correlation matrix. 271
B.11 Table of the 1-st Principal Component loadings for the PCA without network metrics. 273
B.12 Table of the 2-nd Principal Component loadings for the PCA without network metrics. 274
B.13 Table of the standard deviations, proportion of variance and cumulative proportion of the Principal Components in the PCA with network metrics. Standard deviation is the square roots of the eigenvalues of the covariance/correlation matrix. 274
B.14 Table of the 1-st Principal Component loadings for the PCA with network metrics. 275
B.15 Table of the 2-nd Principal Component loadings for the PCA with network metrics. 276

LIST OF ALGORITHMS

1 k-Means Algorithm 248
2 DBSCAN algorithm 252
3 regionQuery Algorithm (DBSCAN) 253
4 expandCluster Algorithm (DBSCAN) 253
5 ToMATo Algorithm 257

LIST OF ABBREVIATIONS

CRC	Colorectal Carcinoma	3
mIF	Multiplex Immunofluorescence	3
TME	Tumour Microenvironment	5
kNN	k-Nearest Neighbour	5
2D	2-Dimensional	3
TIL	Tumour-Infiltrating Lymphocytes	6
TLS	Tertiary/Ectopic Lymphoid Structures/Organs	6
ICM	Inflection Count Measure	11
3D	3-Dimensional	12
SOAM	Sum of Angles Metric	12
PD-1	Programmed Cell Death Protein 1	23
PD-L1	Programmed Death-Ligand 1	23
T cells	Thymus cells	23
MHC	Major Histocompatibility Complex	25
B cells	Bursa cells	25
CD8	Cluster of Differentiation 8	26
CD68	Cluster of Differentiation 68	26
CD163	Cluster of Differentiation 163	26

DC-LAMP Dendritic Cell - Lysosomal Associated Membrane Glycoprotein 26

XCR1 Chemokine Receptor 1 26

HLA-DR Human Leukocyte Antigens - DR Isotype 26

CK Cytokeratin 26

ECI Extent of Compartment Interaction 28

TNM Tumour-Node-Metastasis 29

MSS Microsatellite Stability 31

MSI Microsatellite Instability 31

DNA Deoxyribonucleic Acid 31

MMR Mismatch Repair 31

CD3 Cluster of Differentiation 3 32

CT Central Tumour 32

IM Invasive Margin 32

CSR Complete Spatial Randomness 47

PCF Pair Correlation Function 48

ROI Region of Interest 58

χ^2 **test** Chi-Squared test 71

PH Persistent Homology 104

CD20 Cluster of Differentiation 20 107

UMAP Uniform Manifold Approximation and Projection 112

t-SNE t-Distributed Stochastic Neighbor Embedding 112

DBSCAN Density Based Spatial Clustering And Noise 113

ToMATo Topological Mode Analysis Tool 113

ARI Adjusted Rand Index 114

TP True Positive 117

TN True Negative 117

FP False Positive 117

FN False Negative 117

NA Not Applicable 128

CD21 Cluster of Differentiation 21 142

CD138 Cluster of Differentiation 138 142

VEGF Vascular Endothelial Growth Factor A 154

BBB Blood-Brain Barrier 154

PCA Principal Component Analysis 170

TDA Topological Data Analysis 210

CHAPTER 1

INTRODUCTION

We live in a world where we are constantly influenced and affecting the environment we live in. The impressions made upon us and on those around us can be deconstructed into a system where individual *elements* can form *connections* to others in their surrounding *environment*, with each connection determined by a set of criteria. This framework is useful in characterising the most important properties of a large number of interacting elements, such as the biological cells inside our bodies. Here, we have a microscopic *cell* as an element, and connections are replaced with *signalling pathways* which cells use to interact with each other within a *microenvironment*. In mathematics, we refer to these analogous constructs as *complex spatial networks*. A complex spatial network reduces a system into universal constituent components which mathematical techniques can be applied to summarise higher-order organisation or identify unexpected localised dynamics. This information can then be used to inform mathematical models beyond this thesis, such as agent-based models [9], to assign parameters for individual agents based on the statistics obtained from spatial complex networks.

There are many attributes to a complex spatial network which differentiate it from simpler networks. Because of this, the network requires different perspectives to summarise the most influential or significant properties. One perspective is through the lens of topology. Topology is a branch of mathematics. Simply put, topology describes the properties of objects which are preserved under continuous deformations. The properties that are *invariant*, those that do not change under continuous deformations, are referred to as topological properties. For example,

a torus (the common example being a doughnut) has one hole. Without stitching the hole up or tearing a new one, we can twist, crumple, stretch and bend the shape but the hole would still exist, so we say it still has the same topology as a torus. This is then distinguished from a sphere, which has no hole and, thus, has a different topology. The topology of a complex spatial network can then be compared to other networks with a similar topology, irrespective of geometric deformation or scale. This is especially helpful when the topology describes features which we can relate to the underlying criteria that dictate connections. It is with this lens, that topology can describe patterns of connections in a perceptive way.

The scale we might consider the principal elements in our system might differ depending on the observations we expect to examine. For instance, it could be relevant to understand the use of urban green spaces by observing the movement of pedestrians through those spaces [10]. Similarly, it could be useful to investigate the movement of individuals to track the spread of an epidemic in a community [11]. On the other hand, when considering an epidemic spread, surveilling every individual quickly leads to resource limitations and ethical concerns. As a result, studies might characterise the principal elements as regions or census output areas instead [12]. In our studies, we will be considering principal elements magnitudes smaller for our research: biological cells. Even at this scale, the metrics used in other contexts can equip an investigator with perceptive interpretations of complex spatial networks.

Complex spatial networks and topological data analysis have provided an invaluable viewpoint to the study of *systems biology*, the analysis and modelling of complex biological systems. These fields have encapsulated dynamics in neural networks of the brain [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27], structures in bone architectures [28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39] and organisation in the lymphatic system [40, 41, 42, 43, 44]. The theoretical foundations of complex spatial networks have found a natural application for the inquiry of numerous, complex and multi-scaled management of many specialised biological cells.

The research conducted in the following work specialises in the application of two areas of systems biology. The first is the study of *immunology*. Immunology is the study of immunity.

Immunity refers to all the mechanisms used by the body to protect itself from infection, disease and cancer [45]. We analyse two data sets which aim to interpret our immune response: one studying the spatial relationship of many immune cells with cancerous tumours of Colorectal Carcinoma (CRC) patients, the other, characterising structures of immune response in salivary tissue for a disease called Sjögren's syndrome. The second area of study is the blood vascular system. Specifically, the data set acquired investigates the growth of new blood vessels, i.e. vasculogenesis, by observing microvascular vessels grown *in-vitro* under different environmental conditions.

Considering the natural division in applications, the remaining chapters of this thesis will detail the work carried out on each data set, as demonstrated in Figure 1.1. A theoretical foundation for the topics of complex spatial networks and topological data analysis are presented in Appendix A (Sections A.1 and Section A.2, respectively) which the chapters reference. The remainder of Chapter 1 will discuss the literature relating to the analysis covered in the remaining chapters. These will give context to the metrics and measures used to analyse images in immunology and blood vessel research, respectively. Then, the specific research questions of this work will be detailed, referring to the limitations of previous studies and finally the extensions provided by the contributions of this work.

1.1 Literature Review of Cellular Spatial Organisation in 2-Dimensions

Multiplex Immunofluorescence (mIF) imaging allows the simultaneous detection of multiple biological markers on a single tissue section. Image processing then allows for the detection of the cell borders. Once identified, their colocalisation with markers determines their cell phenotype. Our two data sets utilise this technology to interpret cellular communication on 2-Dimensional (2D) tissue sections. New protocols endeavour to increase the number of antigens that can be identified [46, 47] making it possible to monitor evermore complex cellular phenotypes. However, with this added complexity comes difficulties in targeted yet transferable

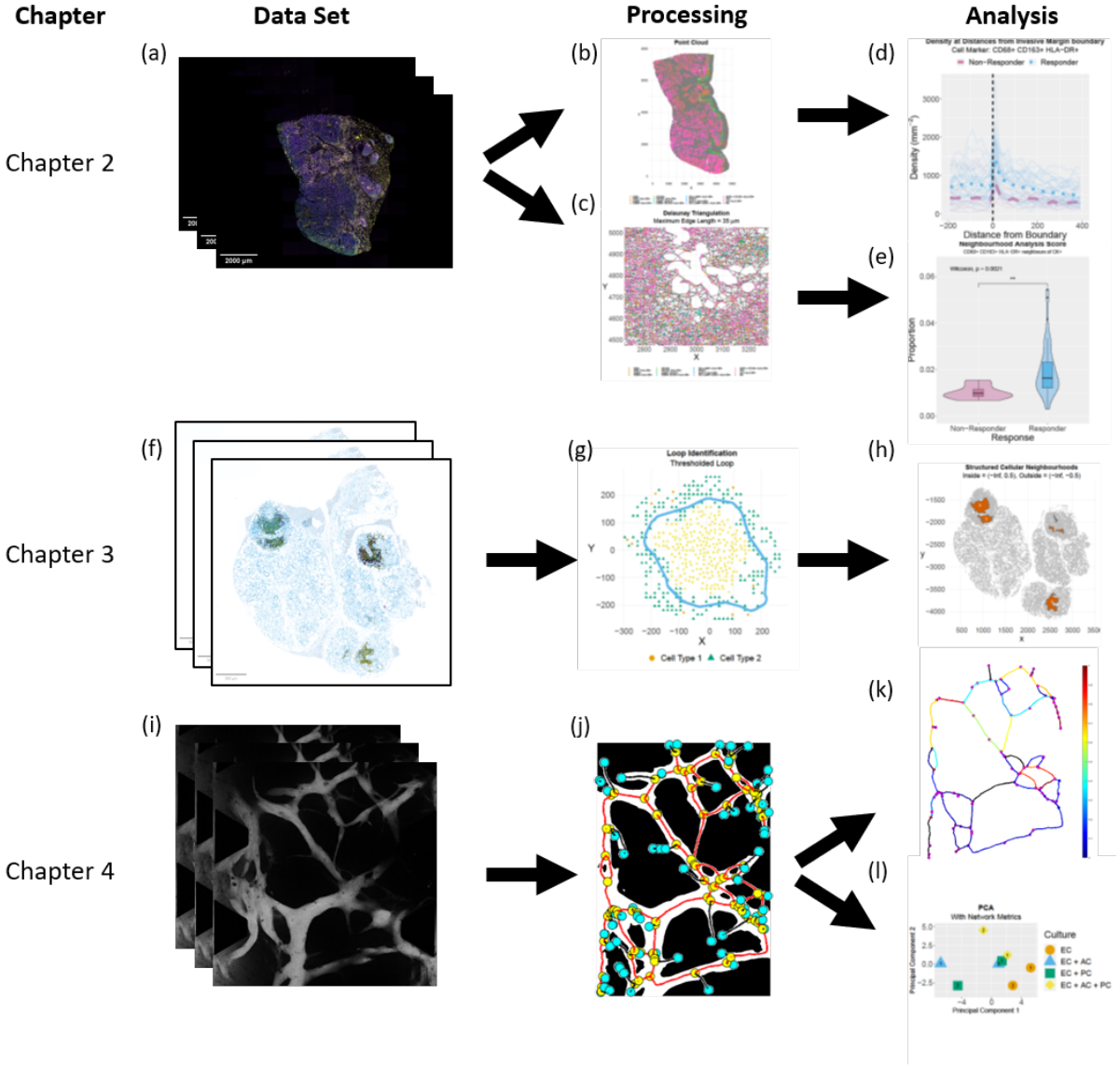


Figure 1.1: An outline of the work completed in each chapter. In general, each chapter considers a respective data set which is then processed into its constituent elements before applying current and novel metrics. (a) The HALO1079 data set. (b) The constituent cell coordinates and respective phenotypes. (c) The graph construction, via Delaunay Triangulation, of the cell coordinates. (d) A figure from the novel Extent of Compartment Interaction analysis. (e) A figure from the Neighbourhood Analysis Score analysis. (f) The TRACTISS data set. (g) An example of the Loop Identification step in the Structured Cellular Neighbourhood workflow. (h) A figure of the output from the Structured Cellular Neighbourhood workflow. (i) The Self-Organised Microvascular Model data set. (j) A figure of the constituent nodes (branching points) and edges (vessels) in the 3D graph overlaid on a 2D slice of the segmented image. (k) A figure of the novel Analytical Solution of Flow in a Network analysis. (l) A figure of the Dimensional Reduction Analysis output beased on the geometric and topological features of the samples.

analyses to characterise the abundant possible interactions [48]. Thus, analytical techniques need to be found to understand the different levels of cellular cooperation across multiple cell types [49, 50, 51, 52].

Machine learning methods have been considered in the characterisation of the Tumour Microenvironment (TME). Algorithms such as random forest [53, 54, 55, 56, 57], gradient boosting [58], *k-Means Clustering*[1] [59], k-Nearest Neighbour (kNN) [60], decision trees [61] and support vector machines [54, 61] have been applied to characterise cellular organisation. Barriers still exist, such as insufficient or limited quality data sets; lack of context provided to algorithms for resolving relevant hypotheses; and ethical concerns associated with applications of machine learning [52].

Many studies in exploratory data analysis rely on spatial correlation metrics, measuring the spatial distribution of one cell type to another cell type. We explore methodologies used in previous studies to classify cancer samples to understand the capabilities and limitations of these techniques.

At its simplest, measuring the distance of each cell type to every other cell type will produce a series of distributions for the spatial localisation of multiple cell types. Bortolomeazzi et al. [62] then compare to known cancer subtypes with this distribution via a two-sided Wilcoxon test. However, as we discuss in Section 2.7, Subsection 2.7.2, this can quickly become cumbersome when comparing multiple phenotypes.

A popular metric for point coordinates is *Ripley's K-function* [63, 64]. Briefly, this is a measurement of the density of another cell type around any other cell type for each increment in distance radially outwards. This is then compared to the theoretical values of a spatial Poisson point process. We explain this metric in more detail in Chapter 2, Subsection 2.7. To address the problem with multiple phenotypes previously noted we also provide a solution to summarising multiple pairwise comparisons which identifies the most distinct cellular distributions.

In the previous examples, the Euclidean metric space is utilised in the methodology to

[1] See Chapter 2, Subsection 2.7 for definition.

describe a physical spatial correlation of different cell types. Lin et al. [65] instead consider using a different metric space. Their metric is based upon the proliferating cell nuclear antigen after discovering its effectiveness in whole-tumour characterisation via kNN-classification. The gradients then inform the distances between cells of a similar stage in the proliferation cycle, thus, the surrounding cells of a particular niche could be measured.

Some methods rely on graph constructions, whereby each node is a cell coordinate and follows a rule or set of rules to be considered connected, via an edge, to another cell in the domain. A common rule is to state that if the distance to another cell is within a pre-specified distance of the cell, then it is considered interacting. For example, Failmezger et al. [66], consider tumour cells to be interacting if the two cells are within 35 microns of one another. This threshold was determined by calculating the distance of cancer cells to their closest stromal neighbour in regions of the tumour–stroma interface. Then, the mean value of this distribution was assigned as the interaction threshold distance. Otherwise, a range of distances can be arbitrarily set to recognise the differences in interaction strength, such as in Küçükköse et al. [67], where three distinct distance-based bins were defined: "intra-epithelial" ($r \leq 0\,\mu$m), "proximal" ($1\,\mu\text{m} \leq r \leq 25\,\mu$m), and "distant" ($r > 25\,\mu$m). To illustrate the arbitrary definition of relevant cell-cell interaction distances, Feichtenbeiner et al. [68] choose a different set of relevant cell-to-cell distances. They chose a range of 20 to 50 μm for within immune cell phenotypes versus their Tumour-Infiltrating Lymphocytes (TIL) counterparts and between immune phenotype distances were found to be most significant at 30μm to 100 μm. Regardless, these methodologies simplify the actions that cells go through when communicating with one another by categorising all cells within a range of distances having the same influence, i.e. not over multiple scales or dependent on cell phenotype.

Whether measured using the implicit Euclidean space, overlaying a new metric space from additional information, constructing a graph representation of the data or even a combination of a few of these approaches [69], many statistical measures can be implemented and compared in reviews on the subject [51, 70, 71, 72].

Higher-order structures, such as Tertiary/Ectopic Lymphoid Structures/Organs (TLS), are

important features to characterise, as discussed in Chapter 3. Therefore, identifying neighbourhoods of interacting cells can identify the intriguing hierarchical ordering of cell types to perform complex mechanisms that go beyond any one cell. Barmpoutis et al. [73] do identify TLSs explicitly, although in H & E images. In addition, this requires an annotated data set which has not been made available in this work.

One implicit approach complements the format many images are acquired by considering an image as a series of *tiles*, non-overlapping tessellating subdomains of the original image. This approach was used by Javed et al. [74] who then labelled each tile with a phenotype determined by k-means clustering. The tiles with similar compositions are then compared using a chi-squared distance to discern similar phenotypes. Essentially, the authors construct an n-dimensional feature vector for each cluster, where n is the number of cell types measured which describes the proportion of each cell type. Then, a distance matrix is constructed using the chi-squared distance:

$$d(i,j) = \exp\left(-\frac{1}{\sigma}\sum_{k=1}^{m}\frac{(\mathbf{h}_{i,k}-\mathbf{h}_{j,k})^2}{\mathbf{h}_{i,k}+\mathbf{h}_{j,k}}\right) \tag{1.1}$$

where $\mathbf{h}_i \in \mathbb{R}^m$ and $\mathbf{h}_j \in \mathbb{R}^m$ are two feature vectors, and σ is a weight decay control parameter. After all pairwise aggregations are completed, the result is converted to a geodesic distance matrix whereby all pairs' shortest path is calculated between every feature vector. This can provide more meaning in case the data is on a non-linear manifold. Assuming the network represented by the distance matrix is fully connected, the distance between two geodesic vectors $\mathbf{g}_i$ and $\mathbf{g}_j$ is defined as,

$$D(i,j) = \sqrt{(\mathbf{g}_i-\mathbf{g}_j)^T\mathbf{W}(\mathbf{g}_i-\mathbf{g}_j)} \tag{1.2}$$

where $\mathbf{W} \in \mathbb{R}^{n\times n}$ is a diagonal matrix containing weights for each dimension of the geodesic distance vector. These weights are adjusted such that the local and global structure of the

network become equally important [75],

$$\mathbf{W}_{p,q} = \begin{cases} \frac{1}{2} & , \quad \text{if} \quad p = q = \{i, j\}, \\ \frac{1}{n-2} & , \quad \text{if} \quad p = q \neq \{i, j\}, \\ 0 & , \quad \text{if} \quad p \neq q. \end{cases} \tag{1.3}$$

The result is a tile-level graph with the nodes representing a tile and the edge length the similarity between aggregations. This is followed by grouping tiles using the k-Means algorithm into specific phenotypes after a density gradient is applied and the nodes are allowed to drift towards maximum density regions of tiles that they are considered most similar to. However, this approach relies on the scale of neighbourhoods to be associated with the size of the tile.

Alternatively, Schurch et al. [46] consider a "cellular neighbourhood" to contain 10 cells, where there is one central cell and their 9 nearest neighbours. Next, a pre-determined number of cellular neighbourhood phenotypes are assigned as input to the subsequent k-Means clustering which matches similar cellular neighbourhoods. A further implementation of this methodology by Phillips et al. [76] was able to identify key spatial correlations which culminated in the creation of a new metric called *SpatialScore*. SpatialScore proved to have predictive power in identifying responsive and non-responsive patients to the pembrolizumab immunotherapy treatment.

A final method of interest for monitoring neighbourhoods is presented by Chen et al. [77]. They proposed a hierarchical cell phenotyping and graph modelling algorithm to profile the cellular architecture. A graph of the cells is first constructed via a Ranking Operation-based clustering algorithm. Then, *Supercell phenotyping*, separating cells into communities, is calculated via the spectral clustering algorithm. Finally, the Delaunay triangulation of the image identifies neighbouring Supercells that are considered to interact and the number of similar or different interacting Supercells is then measured.

Topological tools have also been successfully applied to investigate the TME. The first discussed here, applied in prostate cancer, uses persistent homology to characterise tumours

and compare them to the established Gleason grading scale [78]. In essence, the method relies on the natural homological structures that occur in the prostate. The Gleason grading uses the glands and their regular and circular organisation as a prognostic variable for prostatic adenocarcinoma, where a more irregular and incomplete circularity indicates a higher grade of tumour and thus a negative response. Therefore, using persistent homology naturally identifies these homological structures with more persistent 1st Betti number features indicating a lower Gleason grade.

A similar example in three-dimensional space is presented in breast cancer, where a topologically-derived biomarker called "TopoTxR" is demonstrated [79]. To capture the third dimension the cancer was captured by magnetic resonance imaging. When mIF can image in this new dimension this shows that the topological perspective can provide a useful insight to characterise the TME.

Persistent homology has been the main flavour of topological data analysis referenced thus far, however, an extension of the algorithm was proved necessary by Vipond et al. [80] to account for the stochasticity of real biological data. Their research investigated the infiltration of macrophages into tumour spheroids. They introduced the use of multi-parameter persistent homology. Specifically, instead of relying on one radial parameter a second parameter, called "codensity", filters macrophages based on the neighbouring density of cells. This meant that macrophages in denser regions of the point cloud were included earlier so that less densely arranged outlier macrophages were less likely to contribute as noise.

By understanding the methods being used by the aforementioned studies, we can understand the benefits and limitations of the metrics for our application. In the relevant chapters, we evaluate several techniques on artificial data before considering the most relevant to the analysis of our images. Also, we consider extensions and additional methods to provide new perspectives into the analysis of mIF images.

1.2 Literature Review of Blood Vessel Networks in 3-Dimensions

The objective of the following review on vascular networks is to identify accurate, comparable and intuitive metrics that measure the geometry and/or topology of a network. Of particular interest were the measurements recorded by those studies to characterise a system's morphology. This was part of a larger literature review across many disciplines which looked at 3D data sets of spatial networks, systems where proximity and connectivity of components are equally important. These included data sets such as neuronal networks, bone structures, geological structures and manufactured materials. Understanding the structure of these networks informs our study for the effective transport of oxygen and other nutrients by the microvasculature, whether *in silico* [81] or *in vitro* [82].

A standard calculation for most studies is to measure the vessel diameter [83, 84, 85, 39, 86, 87]. This is often not a calculation specific to analysing the network but a way to identify the hierarchical structure of a microvasculature, where diameters usually over 10 μm are considered arteries or veins and less than 10 μm as capillaries.

Next, reporting the distribution of vessel lengths has given richer information about a system of blood vessels. This information has been used to infer vessel maturation and to provide insight into the hierarchical organisation of a network [39, 88, 86, 87].

Then, a popular measurement amongst studies, and one that plays directly into the 3D properties of the data, is the percentage of the observed volume that is filled by blood vessels or a similar measure quantifying the density of the vessel structures [83, 84, 85, 39, 89, 88, 86, 87, 90]. The interpretation of this metric follows a Goldilocks principle, the 'normal' result being that between certain margins rather than at either extreme. Specifically, if the density is too low then a network is considered to not have fully matured and not carry out its key duty of oxygen and nutrient transport to the surrounding medium. Likewise, if most of the space is filled with the network then the network is indicated as overcompensating to some stressful stimuli.

An alternative approach to measuring the direct volume of a microvasculature to quantify

the density is to dilate the network *in silico* in an attempt to measure the spatial distribution. This is the approach first formulated by Righi et al. [91]. They computed the number of cycles needed for black voxels to undergo to fill at least 95% of the volume according to a rhombicuboctahedral expansion scheme. The greatest initial volume observed among all of the volumes to be analysed was used as the normalisation value. The number of initial dilation cycles needed to reach the normalisation value was then subtracted from the total number of cycles needed. This then produces the metric referred to as the normalised volumetric Halo Index for 95% (nHv 95%). Vascular amounts and dispersion data from each point of accumulated vascular arrangements were then plotted with the percentage of occupied volume versus the number of expansion cycles, where a lower number of cycles (i.e. a lower nHv 95% value) indicated a better spatial distribution of the vascular tree. The result is a series of values which interpret the amount and dispersion of large vessels, total vessels and microvascular density. Also, a manual threshold on 'vessel calibre' determined different phenotypes of microvasculatures in normal and diseased states across multiple data sets [90, 92, 93].

In an attempt to calculate the eccentricity of normal microvasculatures we see the calculation of tortuosity and vessel alignment as a favoured metric [83, 84, 94, 85, 89, 86, 95]. Tortuosity is the property of a vessel delineating from the straight line between the start and the end point of that vessel. There are multiple ways to calculate the tortuosity, however, all studies considered reporting at least one metric of tortuosity as the arc-chord ratio, also known as the Distance Metric, $\tau = C/L$, where C is the centreline length and L is the Euclidean distance between endpoints. Therefore, a measurement of 1 is a straight line and a limit to infinity is estimated as a circle. In some studies, the interpretation becomes one related to the growth response of a microvasculature to different diseased states, where higher values are considered more 'eccentric' which relate to some sensing mechanism in the vessel being disrupted [86]. Sun et al. [85] is an example of a study which quantified multiple tortuosity measurements. This included the Inflection Count Measure (ICM), which could help distinguish between a "S"-shaped curve and a "C"-shaped curve by multiplying the Distance Metric

by the number of inflexions. They also considered the Sum of Angles Metric (SOAM) which is defined as the sum of the curvatures of all voxels along the skeleton of the blood vessel and normalised by the actual path length of the vessel. Bullit et al. [96] compared these three measures of tortuosity (Distance Metric, ICM, and SOAM) and validated their use to intracerebral blood vessels in 3-Dimensional (3D) images. ICM appeared most effective in recognising the first 2 of 3 types of abnormal tortuosity (type I: sinuous curves in long, normally straight vessels and type II: tightly packed clusters with erratic directional changes). In contrast, SOAM seemed to be the most effective in recognising the third type of abnormality, characterised by high-frequency, low-amplitude coils. The vessel alignment measures presented here consider a reference vessel that is straight between endpoints and not how a vessel might orientate according to a gradient, like oxygen concentration, between system inputs and outputs. An example of this is to consider overall vessel alignment with a global average, as performed by Freiman et al. [94]. Here, they related vessel maturity to the variation in average alignment. The approach used the Hough transform and binned points into intervals of orientation based on this output. Despite being practical for summarising vessel alignment without the necessity of defining a vessel centreline, the orientation of vessels cannot then be related to individual vessels in the spatial domain.

Previously, we discussed how tortuosity is applied to record the deviation from an expected lattice configuration that would otherwise fill the volume optimally. A step towards this has been to use specific measurements quantifying the fractal dimension and lacunarity of a network [83, 94, 97]. Fractal dimension refers to the measure of fractal patterns by quantifying their complexity as a ratio of the change in detail to the change in scale [98]. This contrasts to lacunarity which aims to summarise the extent to which a pattern fills the space [99]. In addition, lacunarity can be used to quantify other aspects of patterns such as rotational invariance and more generally, heterogeneity.

So far the metrics mentioned have concentrated on the geometrical features of a microvasculature. These measurements have compared the properties of a graph according to some length scale or spatial distribution. The following illustrates the quantification of properties

that are invariant to length scale and thus topological. The majority of these depend on reducing the blood vessel network to their centreline (i.e. the one-pixel-wide line going through the centre of the vessel). From here, the centreline is transformed into a graph representation where edges are the vessels and where a vessel splits into two or more vessels as nodes, or referred to as *branching points*. Then, statistics from complex networks [100] can be employed to describe some of the emergent properties in the connectivity of the network.

The most trivial metric is counting the number of edges and branching points [90, 89, 97]. Beyond that, the distribution and mean degree of nodes can also be calculated [88, 86]. The degree of a node then refers to the number of vessels that converge to a shared intersection and so the distribution can describe the number of interconnections within a network. These metrics on their own summarise some properties of the size and complexity of a network.

To work out whether the network is efficient in oxygen and nutrient transport, some metrics computing the distance between pairs of nodes across the whole network are used. Despite being a large area of interest in neuroscience (see Nelson et al. [101]) only Kaufhold et al. [95] perform such a calculation in blood vessels with the Dijkstra's shortest path between pairwise nodes. On the other hand, a related measurement is computed in Blinder et al. [87] with the use of network resistance. This combines the diameter and length of each edge in the shortest path between nodes by calculating the Hagen-Poiseuille law for the granular nature of blood in each edge. The network resistance across pairs of vertices was then computed. They found that the resistance asymptotically approached as a function of the Euclidean separation distance which implied that the microvasculature formed a highly interconnected irregular lattice in all directions.

The identification of community structures within larger connected components of the network was identified via the Louvain number [89, 87]. Specifically, Blinder et al. [87] wanted to identify the number of community structures formed in the microvasculature of each cortical column in the brain. In addition, they plot the size of these communities to determine if any individual community is more essential as well as how interconnected they were to monitor their confinement to cortical columns.

As an alternative to measuring emergent properties of the graph representation, Grigoryan et al. [1] construct artery and vein networks that possess ideal topological structures. The objective of their study was to identify effective space-filling topologies for oxygenation. They include an axial vessel with helix orientation, interpenetrating Hilbert curves, bicontinuous cubic lattice, and torus with (3,10) torus knot (see Figure 1.2).

This literature review has provided insight into the methodologies previously carried out by studies into the analysis of blood vessel networks. We see interest in understanding a combination of geometrical properties of blood vessels with their topological structure. From the studies that we have analysed, we have not seen how studies have combined both geometric and topological information into a single analysis. The literature has shown some promising measures to use for specific properties of microvasculature networks but this complex structure requires a more holistic description of its mechanisms to characterise it fully. Since all studies are interested in the organisation of these structures concerning oxygen and nutrient transport, we look for a way to model this mechanism directly. Then, we suspect that a combination of geometry and network metrics can identify the most crucial mechanisms which change depending on differences in culture microenvironment. We discuss our data set in the relevant chapter and choose the most promising metrics from this literature review and complex spatial networks to distinguish properties crucial to oxygen and nutrient transport.

1.3 Research Questions

Due to the ubiquity of complex real-world systems and the flexibility of mathematical descriptions to model them, this thesis focuses on the feature extraction of complex spatial structures. Before we can focus on feature extraction, the images need to be refined to constituent mathematical elements informed by the biological context. This thesis provides insight into descriptive statistics for the biological interpretation of micro-, meso- and macroscopic features suitable for performing further analyses on disease progression or optimal vascular construction. The following sections introduce the major research questions that this thesis

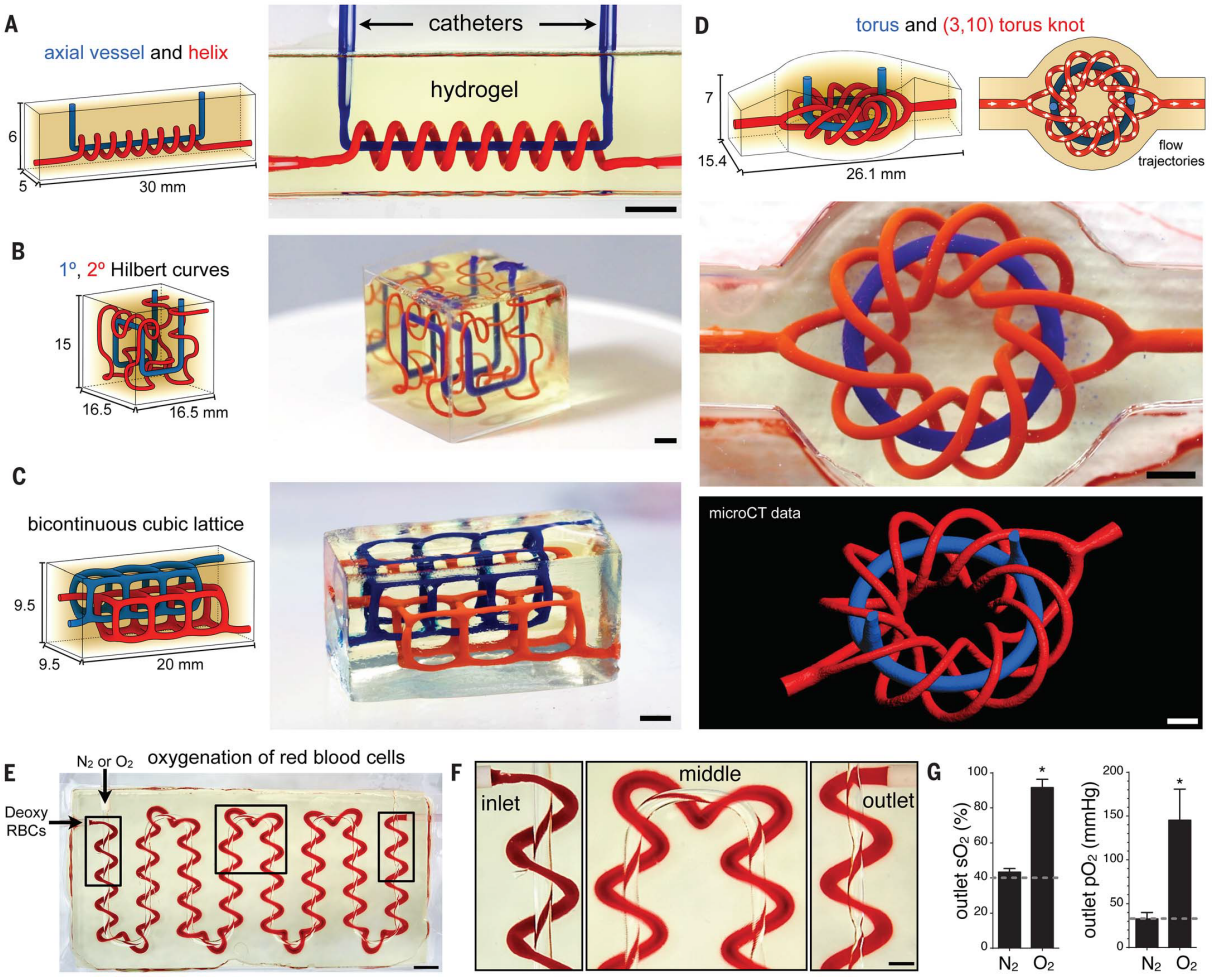


Figure 1.2: From Grigoryan et al. [1]. Reprinted with permission from AAAS. Original caption: Entangled vascular networks. (A to D) Adaptations of mathematical space-filling curves to entangled vessel topologies within hydrogels (20 wt % PEGDA, 6 kDa): (A) axial vessel and helix, (B) interpenetrating Hilbert curves, (C) bicontinuous cubic lattice, and (D) torus and (3,10) torus knot (scale bars, 3 mm). (E) Tessellation of the axial vessel and its encompassing helix along a serpentine pathway. The photograph is a top-down view of a fabricated hydrogel with oxygen and RBC delivery to respective vessels. During perfusion, RBCs change color from dark red (at the RBC inlet) to bright red (at the RBC outlet) (scale bar, 3 mm). Boxed regions are magnified in (F) (scale bar, 1 mm). (G) Perfused RBCs were collected at the outlet and quantified for SO2 and PO2. Oxygen flow increased SO2 and PO2 of perfused RBCs compared with deoxygenated RBCs perfused at the inlet (dashed line) and a nitrogen flow negative control.

has investigated for the respective biological applications. The resulting statistics characterise specific features relating to the principal biological hypotheses that motivate each data set. For each project, the overarching theme has been to infer the interaction of the cells, whether in the formation of blood vessel structures or the many specialised agents in the immune system, through spatial proximity and how that affects change in the microenvironment.

1.3.1 Identifying Biologically Significant Cellular Spatial Organisation from 2-Dimensional Immunofluorescence Imaging

Two of the data sets relate to the study of Immunology. Again, immunity describes many mechanisms by which the body can protect itself from unwanted substances. There is a complex system of many cell types which are involved in our immune response. To begin to understand this complex interplay, data sets have been designed to infer specific cellular interactions through spatial proximity within the context of two diseases.

The observation technique used in both of these data sets uses mIF technology. In each data set specific markers have been identified in surgical biopsies of human tissues. These are imaged as 2D spaces. From these images, we can discern the coordinates of cells and the cell type for a pre-specified panel of markers. It is with this information which we expect to interpret cellular interaction in the tissue sample through the spatial organisation of the coordinate and their cell type. It is by observing the spatial correlation in multiple patients, recording immune organisation, and then comparing to a patient's outcome that we can gain insight into functional immune responses. The results of our work will then aid in identifying targets for future medicine that can bolster impotent immunological mechanisms. Based on this, the first research question of this thesis is:

- How does the spatial cellular organisation of immune cells predict patient outcome for CRC and Sjörgens disease?

Chapter 2 proposes multiple methods to be appropriate in answering this question. The methods can be characterised into broader perspectives and are separated into *Areas, Counts*

and Densities (Section 2.5), *Compartment Interaction* (Section 2.6), *Spatial Correlation within Multitype Point Patterns* (Section 2.7) and *Cellular Neighbourhoods* (Section 2.9), the application of which is then presented in Section 2.11.

Chapter 2 concentrates on the application of these methods to one of our data sets, which we refer to as the HALO1079 data set. In this data set, we have many markers in the panel which may or may not have some significance to disease progression, in this case CRC. This differs from our TRACTISS data set which has only a couple of markers which have been explicitly chosen to identify specific immunological structures. Therefore, the HALO1079 data set is suited to exploratory investigation and characterisation of cellular spatial correlation between multiple markers.

The duration of the TRACTISS data set is differentiated by its pre-specified goal to identify immunological structures, called TLS, and its investigation into Sjörgen's disease. TLS structures are classified by both a specific set of cellular types and their subsequent spatial organisation. The identification of these structures indicates positive disease progression outcomes in Sjörgen's disease. In addition, TLS are observed as indicating a positive disease progression outcome in breast carcinoma, gastric cancer, non-small-cell lung cancer, melanoma, oral squamous cell carcinoma, pancreatic cancer, renal cell cancer, and colorectal cancer [102]. As such, we focus on these structures as a known indicator of positive disease progression. Once the identification algorithm is proven to work on the TRACTISS data set, we look towards the application of the algorithm to our HALO1079 data set for the identification of similar immuno-aggregations. Therefore, the next research question of this thesis is as follows:

- Can we identify higher-order structures of cellular organisation that predict disease progression outcomes?

After reiterating the motivation for the chapter in Section 3.1, we begin Chapter 3 with a summary of the specific biology identified in the images in Section 3.2, including the criteria necessary for an immuno-aggregation to be considered a TLS. In Section 3.3, technical details are then provided for the conversion of the images into data architectures which are commonly used as an input in these analyses. Then, we identify in Section 3.4 the most accurate

techniques considered for the identification of structured cellular neighbourhoods of immuno-aggregations that correspond with the properties of TLS. The application of the final workflow is then validated on an artificial data set and the TRACTISS data set in Section 3.6.1 and Section 3.6.2, respectively.

The aforementioned research questions motivate the initial chapters of this thesis. The immediate opportunities and future implications of this work are discussed at the end of the chapter as well as further discussion in Chapter 5.

1.3.2 Delineating Live Microvascular Models and their Structure from 3-Dimensional Confocal Imaging

Blood is essential for the distribution of necessary nutrients and oxygen for mammals and other vertebrates. Blood is transported around the body via a system of hierarchical vessels which permeate vital substances to every living cell. These are called blood vessels.

There are five types of blood vessels. First, we have the arteries, which carry the blood away from the heart. Then, we have the arterioles, which transport blood to the limbs of the body. From these, the vein and venule vessels carry blood from the arteries and arterioles, respectively. The smallest scale in which blood is transported is through the capillaries, where the exchange of water and chemicals between the blood and the tissues occurs.

The focus of our data set concentrates on the capillaries. Capillary comes from the Latin word capillaries, meaning "of or resembling hair", with use in English beginning in the mid-17th century [103]. A human hair diameter can range from 17μm to 181μm. However, the diameter of a capillary ranges a fraction of that of a human hair, ranging from 5μm to 10μm. It is at this scale that imaging acquisition is pushed to the extreme when particularly capturing living vessels without exerting toxic effects on them. This is before considering the technicalities of imaging samples in 3D.

Tissues and organs are interlaced with capillaries forming the capillary bed. During embryological development, new capillaries are formed by *vasculogenesis*, the process of blood vessel formation occurring by *de novo* production of endothelial cells and their formation into

vascular tubes. In contrast, the term *angiogenesis* denotes the formation of new capillaries from pre-existing blood vessels [104].

In our scenario, we have acquired live microvascular models which have been grown *in vitro*. Each model has been set within a different culture of cells associated with vasculogenesis. It is this distinction in microenvironment that we will determine from the resulting structures and their effect on oxygen and nutrient transport. It is with this in mind that we propose the following research question:

- Which geometric and/or topological metrics characterise the transport of oxygen and nutrients in networks undergoing vasculogenesis in different microenvironments?

Chapter 4 concentrates on resolving this question. First, the processes by which we have converted the images into data which can be used as input for both geometrical and topological analysis are described in Section 4.2.2. Then, we consider several measures describing the geometry (Section 4.4.1) and topology (Section 4.4.2) which provide an intuitive statistic relating to blood vessel architectures. Following this, we apply multi-dimensional reduction techniques which identify the most important measures and summarise a combination of the measures in Section 4.4.3. Next, we derive an analytical model for fluid flow in a network in Section 4.5. Finally, in Section 4.7, we compare the geometrical and topological measures to analytical flow simulations to relate measures to properties of nutrient and oxygen transport.

The immediate opportunities and future implications of this work are discussed at the end of the chapter as well as further discussion in Chapter 5.

1.4 Thesis Contributions

In this thesis, we will present several methods to first process different modalities of imaging data into data structures where we can perform geometric and topological analyses. Automatic detection of cells and features in large 2D images, and even more so in 3D images, is a challenging task requiring careful choice and combinations of techniques. After presenting our

developed pipeline for extraction, we then present metrics and novel methods to identify and compare features of significant biological relevance. This thesis is structured as follows:

- **Chapter 2: Exploring Multiple Phenotypic Intercellular Spatial Proximity in 2D Multiplex Immunofluorescence Imaging.**
- **Chapter 3: Identifying Higher-Order Structures in 2D Targeted Immunofluorescence Imaging.**
- **Chapter 4: Quantifying 3D Self-Organised Microvascular Models as Spatial Graphs for Transport Simulations.**
- **Chapter 5: Discussion and Conclusion.**

In summary, the key contributions of this thesis are the following,

- We integrate the use of advanced image processing techniques, combining tools to construct effective workflows for the identification of cell boundaries and cell phenotypes from 2D multiplex immunofluorescence images.
- We identify the extent of spatial statistics applicable in multitype point patterns
- We apply pairwise comparisons and nearest neighbour spatial correlations to identify cell phenotypes, specifically M2 macrophages in CRC, that indicate predictive prognostic power in patient response.
- We propose a workflow, with stages rooted in topological data analysis, for the automatic detection of higher-order cellular aggregations, which we refer to as structured cellular neighbourhoods, which remain agnostic to cell types yet accurately identify the germinal centres of TLSs.
- We utilise advanced image processing techniques to construct effective workflows for the recovery of live blood vessel structures from 3D images.

- We improve the classification of self-organised microvascular models grown from different environment mediums by quantifying properties associated with complex spatial networks.
- We design an analytical simulation of oxygen and nutrient transport in identified blood vessels.
- Relate the simulation of oxygen and nutrient transport to the properties of the blood vessels as complex spatial networks.

The main purpose of the work proposed in this thesis is to provide frameworks that can better interpret intricate images, yet are suitable for high-throughput biological experiments. Our methods extract useful and intuitive features from complex network models – measurements that will allow for the automatic generation of predictions or further study by domain experts. In the final chapter, we discuss our observations from this thesis in the wider context of biomedical image analysis and identify future opportunities.

CHAPTER 2

EXPLORING MULTIPLE PHENOTYPIC INTERCELLULAR SPATIAL PROXIMITY IN 2-DIMENSIONAL MULTIPLEX IMMUNOFLUORESCENCE IMAGING

2.1 Motivation

Throughout life, an individual will develop cancer that normal immunological mechanisms will eradicate without notice or consequence [45]. However, it is when 'normal' immunological responses fail that we observe the harmful results of cancer. Moreover, a cancer can elicit pro-tumour immunological responses. The adaptation of our immune system to interact with tumours can have a significant impact on the development, progression, and treatment response of an individual.

A tumour resides within an altered system of cells, tissues, and organisms. Most notably, this includes blood vessels, immune cells, fibroblasts, signalling molecules, and the extracellular matrix. We call the tumour and its immediate vicinity the Tumour Microenvironment (TME). Tumours can influence the microenvironment by releasing extracellular signals, promoting tumour angiogenesis (i.e. the growth of blood vessels from existing ones) and inducing immune tolerance, while the microenvironment can affect the growth and evolution of cancerous cells.

The TME has been a rich site for the identification of therapeutic immuno-oncology targets. Immune checkpoint inhibitors are emerging as a leading example of this [105] which have

most successfully targeted Programmed Cell Death Protein 1 (PD-1) and Programmed Death-Ligand 1 (PD-L1) inhibitors leading to a litany of approved treatments including Nivolumab, Pembrolizumab, Atezolizumab, Avelumab, Durvalumab, Cemiplimab, Dostarlimab, and Retifanlimab [106, 45, 107, 108, 109, 110]. The success of these treatments derives from blocking the cancerous cells from presenting PD-L1 to evade detection from PD-1 receptors on immune cells. If not blocked, their interaction with Thymus cells (T cells) causes dysfunction, exhaustion, neutralisation and the release of anti-inflammatory proteins [105]. Understanding the TME is, therefore, a critical area of research to map and understand the complex spatial interplay between different immune, stromal, and tumour cell populations and help us identify new immuno-oncology therapeutic targets outside of immune checkpoint inhibitors.

Many other immune cells and their subsequent effect on the tumour microenvironment remain unknown. Current treatments rely on targeting a known presentation protein on cancerous cells. To predict the effectiveness of immuno-therapies, the studies previously referenced often identify immune activation in T cells as a necessity to elicit a positive response to treatment. However, what about the plethora of other immune cells and their contribution to disease progression? The following aims to better understand this aspect. As such, the work is at the forefront of our biological understanding, thus, prior knowledge about the individual mechanisms could be altered or radically different when considered in tandem with many other interacting cells. Because of this, the following work relies on the principle of exploratory data analysis. Exploratory data analysis encourages the exploration of the data and the continual formulation of hypotheses that could lead to new data collection and experiments. The goal: infer the complex system of interactions through spatial proximity within the TME and provide insight on how this impacts a patient's prognosis in Colorectal Carcinoma (CRC).

2.2 Biological Context

2.2.1 Immunology

The following is a summary of the dominant agents involved in the immune system. An explanation of their role is provided here to give biological context to the results presented in the chapter. This information is based on work by Coico et al. [45] and resources from the Biology LibreTexts [111].

Leukocytes, also known as white blood cells, are an umbrella of cells whose primary function is to identify and eliminate pathogens and are important mediators in the activation of the adaptive immune system. Under this term, there exist five distinct families of cells: phagocytes, mast cells, eosinophils, basophils, and natural killer cells. All of them derive from multipotent cells in the bone marrow known as hematopoietic stem cells. For our study, we further describe the role phagocytes play in the innate immune system.

Phagocytes that are pervasive in the body are neutrophils and macrophages, which are constantly pursuing invading pathogens. Neutrophils largely travel within the blood and are the most abundant type of phagocyte. During the initial phase of inflammation, the neutrophils emigrate out of the blood vessels and towards a site of inflammation, engulfing detected pathogens. However, due to some pathogens being indigestible, they might not be able to resolve infections without the assistance of other types of immune cells. Unlike neutrophils which engulf pathogens, macrophages release a wide array of chemicals to eradicate pathogens. As well as travelling via blood vessels, they also reside within the tissue. Their location in the tissue accounts for their role of searching for worn-out cells and other debris. Additionally, they act as antigen-presenting cells to activate the adaptive immune system. Macrophages are often further divided into two phenotypic types, broadly defined as pro-tumour ('M1') and anti-tumour ('M2') macrophages, first proposed by Mills et al. [112]. This phenomenon has come under particular scrutiny for the prognosis of patient survival in some cancers [113, 114, 115, 116]. For example, in non-small cell lung cancer, high infiltration rates of M1 macrophages into tumour islets, but not tumour stroma, have been associated with increased

patient survival [117].

A less common type of phagocyte includes dendritic cells. These are key cells in activating the adaptive immune system by presenting antigen signals to T cells. These cells can be situated in tissues that are in contact with external environments, mainly in the skin, nose, lungs, stomach, and intestines.

Lymphocytes are another category of white blood cells from the vertebrate immune system. The principal types of lymphocytes consist of T cells and Bursa cells (B cells). These two types of cells are the major cellular components of the adaptive immune response.

Once T cells have processed an antigen, presented in combination with Major Histocompatibility Complex (MHC) molecules which identifies cells to be from the individual, a "non-self" target can be recognised. There are three important sub-types of T cells. The first, referred to as killer T cells, kill cells that are infected with viruses or cells which are otherwise damaged or dysfunctional. The second, called helper T cells, regulates both innate and adaptive immune responses the body makes to inflammation. This means that helper T cells cannot directly kill infected cells, clear pathogens or have any cytotoxic activity. Finally, γ T cells recognise intact antigens not bound to MHC receptors.

B cells become activated when its receptor recognises an antigen and binds to it. In most cases, however, the B cells activation is dependent on a second factor mentioned above — stimulation by activated helper T cells. Once helper T cells have been activated by antigens, they become capable of activating B cells that have already encountered the same antigen. The helper T cells also secrete cytokines, which can interact with the B cells and provide additional stimulation. Antigens that induce a response in this manner, the typical method for activation in B cells, are called T-dependent antigens. Each lineage of B cells expresses a different antibody, so B cells and their antigen receptors represent the complete set of antibodies that the body can manufacture.

2.2.2 The HALO1079 data set

We explored the TME in a cohort of primary CRC surgical resection samples. The data set included images from mIF digital pathology technology that images up to 7 different immunofluorescent markers in the whole slide tissue samples. The markers include Cytokeratin (CK), Cluster of Differentiation 8 (CD8), Cluster of Differentiation 68 (CD68), Cluster of Differentiation 163 (CD163), Dendritic Cell - Lysosomal Associated Membrane Glycoprotein (DC-LAMP), Chemokine Receptor 1 (XCR1), and Human Leukocyte Antigens - DR Isotype (HLA-DR) . This panel was designed to image the interaction of myeloid cell populations within CRC and other important adaptive immune responses. Table 2.1 presents which markers refer to which of the previously summarised immunological agents and a short description of their role.

From the whole slide images, collaborators implemented an image analysis processing workflow to segment cells and generate cell centroid location and corresponding marker stain using the HALO®platform. Because of this, it is possible to analyse the complex spatial characteristics of these cells within the tissues.

In our analysis, we first consider measures already in clinical practice. We consider their development and discuss the limitations of these approaches. Later, we report on the application of these clinical measures to show the predictive power of these measures in classifying pathological complete responses. We emphasise the limitations of broad classifications from clinical practice to both predict and interpret the specific interactions which determine a positive prognosis. To address this, we introduce the processing necessary to transform the samples into the appropriate input for our new analysis. In the first instance, we choose metrics recovered from the literature discussed in Chapter 1. Then, subsequent exploratory data analysis measures are categorised according to the perspective they provide on the interaction between cells. For example, some techniques rely solely on the spatial coordinates of different cell phenotypes. That is, they do not require the notion of a neighbourhood and instead use spatial statistics to quantify correlation compared to a standard pattern. This compares to techniques which require a transformation from coordinates to a spatial network, utilising the notion of

Marker	Associated Cell Phenotype	Cell Phenotype role
CK	Epithelial tissue	It is a thin, continuous, protective layer of compactly packed cells with little extracellular matrix. Epithelial tissues line the outer surfaces of organs and blood vessels throughout the body, as well as the inner surfaces of cavities in many internal organs.
CD8	Cytotoxic T cells	White blood cell of the adaptive immune response, that kills cancer cells, cells that are infected by intracellular pathogens, or cells that are damaged in other ways.
CD68	Macrophages	White blood cell of the innate immune response that can engulf pathogens (such as cancer cells) as well as recruit other immune cells, release pro- and anti-inflammatory chemokines and remodel the extracellular matrix.
CD163	Scavenger receptors	Used as a biomarker in different diseases such as inflammation and as a valuable diagnostic parameter for monitoring macrophage activation.
CD68 CD163	M2-type Macrophages	A phenotype of macrophage typically related to protumour mechanisms, e.g. anti-inflammatory mechanisms, metastasis aiding chemo-attractants.
DC-LAMP	Mature dendritic cells	Act as antigen-presenting cells: they activate helper T cells and killer T cells as well as B cells by presenting them with antigens derived from the pathogen, alongside non-antigen specific costimulatory signals. Dendritic cells can also induce tolerance (unresponsiveness) to T cells.
XCR1	Cross-presenting dendritic cells	As above. Specifically, the dendritic cells that are readily able to present antigens to adaptive immune cells.
HLA-DR	Immune recognition and antigen presentation	The primary function of HLA-DR is to present peptide antigens, potentially foreign in origin, to the immune system that eventually lead to the production of antibodies against the same peptide antigen. Antigen-presenting cells (macrophages, B cells and dendritic cells) are the cells in which DR are typically found. Increased abundance of DR 'antigen' on the cell surface is often in response to stimulation, and, therefore, DR is also a marker for immune stimulation.

Table 2.1: Table of markers used in the mIF imaging of the HALO1079 data set and the cell it is associated to in this study, with a description of the role of the cell in the TME.

neighbourhoods. In each case, we discuss the theoretical perspective each measure provides. Each is then validated on artificial data and a decision is made on the metrics most useful, and subsequently applied, to the HALO1079 data set. We also consider extensions to the existing techniques, specifically highlighting the most important cell-cell interactions amongst multiple cell phenotype comparisons. In addition to established metrics, we introduce a novel method which exploits information unique to the data set provided. Referred to as the Extent of Compartment Interaction (ECI), the technique expands on the pathology annotations of an invasive margin and central tumour in our images. Therefore, we can interpret the 'invasive' nature of a cell phenotype into a tumour across this interface. As a consequence, the project aims to infer patient response of the TME in our CRC cohort and provide insight into the interaction of different cells.

2.3 Clinical Measures

The following describes the calculation of measures currently widespread in clinical practice, and that accompany the HALO1079 data set.

2.3.1 Grading

The grade of a cancer tells you how much the cancer cells look like normal cells. This is often discerned by a pathologist observing a whole slide image under hematoxylin and eosin staining, also known as H&E staining. The grade gives an idea of how a cancer might behave and what treatment is then administered.

The grades of bowel cancer cells are from 1 to 4. Grades 1 and 2 are considered "low" grades where the cells look to be under slow growth, inferred from the well-differentiated and typical characterisation of cells by a pathologist. On the other hand, Grades 3 and 4 are considered "high" grades with cells undergoing rapid growth, inferred from the cells' abnormal morphology and poor differentiation.

- Grade 1: well differentiated, the cells look most like normal cells.
- Grade 2: moderately differentiated – the cells look a bit like normal cells.
- Grade 3: poorly differentiated, the cells look very abnormal.
- Grade 4: undifferentiated, the cells look completely different from normal cells.

Grading is a World Health Organisation classification that has been developed over time. For further discussion on the measure's validity, the reader is encouraged to consult Barresi et al. [118]. In summary, since the adaptations by Hamilton et al. [119] into the two tiers of low and high grade, there has been improved reproducibility in assessment [120]. However, subjectivity in specific assessment areas and locations introduces variability. In addition, Grading has been shown to have improved prognostic power when completed in tandem with other measurable tests, such as Microsatellite Status[1].

2.3.2 TNM Classification of Malignant Tumours

Most cancers are initially characterised by an established system of staging, known as the Tumour-Node-Metastasis (TNM) Classification of Malignant Tumours, with nations having their medical boards reviewing standards with time. The foundation of the stages is based on three pivotal features of the cancer environment: tumour (T), lymph nodes (N), and metastasis (M).

The "tumour" relates to the size and spread of the tumorous region at the primary organ. The "lymph nodes" refers to the presence of nearby lymph nodes. Finally, the "metastasis" denotes whether the tumour has been identified in other organs of the body and thus spread beyond the primary site. For each item, a level is assigned. Then, the combination of assigned T, N, and M reflects a stage whereby lower stages (i.e. I, II) relate to the early growth of cancer, the middle stage (i.e. III) relates to maturing cancer growth and the highest stage (i.e. IV) the full progression of cancer. Table 2.2 details the current guidelines for CRC by the American Joint Committee on Cancer.

[1]See Section 2.3, Subsection 2.3.3

AJCC Stage	Stage Grouping	Stage description
0	Tis N0 M0	The cancer is in its earliest stage. This stage is also known as carcinoma in situ or intramucosal carcinoma (Tis). It has not grown beyond the inner layer (mucosa) of the colon or rectum.
I	T1 or T2 N0 M0	The cancer has grown through the muscularis mucosa into the submucosa (T1), and it may also have grown into the muscularis propria (T2). It has not spread to nearby lymph nodes (N0) or to distant sites (M0).
IIA	T3 N0 M0	The cancer has grown into the outermost layers of the colon or rectum but has not gone through them (T3). It has not reached nearby organs. It has not spread to nearby lymph nodes (N0) or to distant sites (M0).
IIB	T4a N0 M0	The cancer has grown through the wall of the colon or rectum but has not grown into other nearby tissues or organs (T4a). It has not yet spread to nearby lymph nodes (N0) or to distant sites (M0).
IIC	T4b N0 M0	The cancer has grown through the wall of the colon or rectum and is attached to or has grown into other nearby tissues or organs (T4b). It has not yet spread to nearby lymph nodes (N0) or to distant sites (M0).
IIIA	T1 or T2 N1/N1c M0	The cancer has grown through the mucosa into the submucosa (T1), and it may also have grown into the muscularis propria (T2). It has spread to 1 to 3 nearby lymph nodes (N1) or into areas of fat near the lymph nodes but not the nodes themselves (N1c). It has not spread to distant sites (M0).
	OR	
	T1 or T2 N2a M0	The cancer has grown through the mucosa into the submucosa (T1). It has spread to 4 to 6 nearby lymph nodes (N2a). It has not spread to distant sites (M0).
IIIB	T3 or T4a N1/N1c M0	The cancer has grown into the outermost layers of the colon or rectum (T3) or through the visceral peritoneum (T4a) but has not reached nearby organs. It has spread to 1 to 3 nearby lymph nodes (N1a or N1b) or into areas of fat near the lymph nodes but not the nodes themselves (N1c). It has not spread to distant sites (M0).
	OR	
	T2 or T3 N2a M0	The cancer has grown into the muscularis propria (T2) or into the outermost layers of the colon or rectum (T3). It has spread to 4 to 6 nearby lymph nodes (N2a). It has not spread to distant sites (M0).
	OR	
	T1 or T2 N2b M0	The cancer has grown through the mucosa into the submucosa (T1), and it might also have grown into the muscularis propria (T2). It has spread to 7 or more nearby lymph nodes (N2b). It has not spread to distant sites (M0).
IIIC	T4a N2a M0	The cancer has grown through the wall of the colon or rectum (including the visceral peritoneum) but has not reached nearby organs (T4a). It has spread to 4 to 6 nearby lymph nodes (N2a). It has not spread to distant sites (M0).
	OR	
	T3 or T4a N2b M0	The cancer has grown into the outermost layers of the colon or rectum (T3) or through the visceral peritoneum (T4a) but has not reached nearby organs. It has spread to 7 or more nearby lymph nodes (N2b). It has not spread to distant sites (M0).
	OR	
	T4b N1 or N2 M0	The cancer has grown through the wall of the colon or rectum and is attached to or has grown into other nearby tissues or organs (T4b). It has spread to at least one nearby lymph node or into areas of fat near the lymph nodes (N1 or N2). It has not spread to distant sites (M0).
IVA	Any T Any M1a	The cancer may or may not have grown through the wall of the colon or rectum (Any T). It might or might not have spread to nearby lymph nodes. (Any N). It has spread to 1 distant organ (such as the liver or lung) or distant set of lymph nodes, but not to distant parts of the peritoneum (the lining of the abdominal cavity) (M1a).
IVB	Any T Any N M1b	The cancer might or might not have grown through the wall of the colon or rectum (Any T). It might or might not have spread to nearby lymph nodes (Any N). It has spread to more than 1 distant organ (such as the liver or lung) or distant set of lymph nodes, but not to distant parts of the peritoneum (the lining of the abdominal cavity) (M1b).
IVC	Any T Any N M1c	The cancer might or might not have grown through the wall of the colon or rectum (Any T). It might or might not have spread to nearby lymph nodes (Any N). It has spread to distant parts of the peritoneum (the lining of the abdominal cavity), and may or may not have spread to distant organs or lymph nodes (M1c).

Table 2.2: Reprinted with permission from American Cancer Society's website [7]. The system described is the most recent American Joint Committee on Cancer system effective January 2018 [8]. It uses the pathologic stage (also called the surgical stage) which is determined by examining tissue removed during an operation. This is also known as surgical staging. This is distinct from clinical staging, which takes into account the results of a physical exam, biopsies, and imaging tests, before surgery. The following additional categories are not listed in the table above: TX, main tumour cannot be assessed due to lack of information; T0, no evidence of a primary tumour; NX, Regional lymph nodes cannot be assessed due to lack of information.

TNM classification has been useful for decades as a systematic characterisation of an individual's cancer. Based on this, each stage usually receives the same treatment. However, there still exists heterogeneity in such broad categories making it insufficient to differentiate the survival discrepancies of TNM stage I/II CRC patients. For instance, the prognosis of CRC patients at stage II was worse than that at stage III [121]. Therefore, investigating whether the immune system is performing in a suppressive nature or instead complements tumour progression, could help interpret the prognosis of a patient's cancer more accurately and in turn personalise their treatment. In addition, many of the guidelines are based on subjective judgements of biological structures now abnormal from tumour growth. Instead, we introduce metrics which aim to associate fully measurable and objective quantities that can be universally applied to CRC, as well as other cancers.

2.3.3 Microsatellite Instability/Stability Status

Microsatellite Instability (MSI) is the condition of genetic hypermutability, and as such, the predisposition to mutate that results from impaired Deoxyribonucleic Acid (DNA) Mismatch Repair (MMR). The presence of MSI represents phenotypic evidence that the MMR is not functioning normally [122]. For a given CRC, the sample can be allocated into one of three categories: tumours that are classed to have a dMMR–MSI-H signature, where H represents those with a much higher mutation burden, tumours that are classed as pMMR–MSI-L, where L represents those with a much lower mutation burden, or, are considered to have Microsatellite Stability (MSS). Most notably for our investigation, dMMR-MSI-H, (i.e. where there is a high mutation burden) generally refers to tumours with a high infiltration of immune cells, namely CD8+ TILs, T helper 1, CD4+ TILs and macrophages [123].

Defective DNA MMR can be detected either by the lack of immunohistochemical staining of the MMR proteins or by Polymerase chain reaction-identified alterations in the lengths of microsatellites between a patient's tumour and a sample of normal tissue or blood [124]. The ability to measure MSI using deep learning directly from whole slide images has also been attempted [125].

There is evidence that high-grade CRC have poor prognosis only if they are MSS [126], while they behave as low-grade tumours if they show high MSI [119]. Hence, a grading system combining histological differentiation with MSI status was proposed by Rosty et al. [126]. This combined grading system was shown to predict a patient's survival with higher accuracy compared to the current WHO grading.

Considering the established relationship between immune response and microsatellite status, the metric will be a relevant measurement to compare during this investigation.

2.3.4 Immunoscore

The most ratified predictor of CRC-specific prognosis is called Immunoscore®[127]. This metric takes into account the host immune response against cancer by measuring the *in situ* immune cell infiltrate in tumours based on associations identified with clinical response [128]. Tumours are measured for the density of Cluster of Differentiation 3 (CD3) T cells and CD8+ T cells in both the Central Tumour (CT) and Invasive Margin (IM). Then, each density result is compared to a global data set where it is converted to a percentile for each density measure, to rank it to the density found in the the cohort of patients. The overall Immunoscore is then a rank from 0 to 4, whereby the previous densities are averaged and the average percentile can be categorised to the respective Immunoscore: $\mathrm{I0} = (0\% - 10\%], \mathrm{I1} = (> 10\% - 25\%], \mathrm{I2} = (25\% - 70\%], \mathrm{I3} = (70\% - 95\%], \mathrm{I4} = (95\% - 100\%]$. A higher Immunoscore relates to a better clinical response. Often in studies validating the accuracy of Immunoscore, Intermediate and High Immunoscores, i.e. I2, I3, I4, are grouped into a single High Immunoscore characterisation. The global data set to which percentile comparisons are made is propriety so accurate thresholds cannot be presented here. However, approximate percentiles inferred from publications imply a density of cells which follow a logarithmic scale going from 0.1mm^{-2} at 0%, 1,000mm^{-2} at 50%, and 10,000mm^{-2} at 100%.

Extensions of Immunoscore have been suggested to account for the lack of spatial resolution. *Immunomap* [129], is one such example, which presents a technique to characterise the spatial distribution of the specific immune cells in the aforementioned regions. Specifically,

Clinical Measure	Grading	TNM Classification	Microsatellite Status	Immunoscore percentile
Value Range	1, 2, 3, 4	I, IIA, IIB, IIC, IIIA, IIIB, IIIC, IVA, IVB, IVC	MSS, MSI	I0, I1, I2, I3, I4
Likely Positive Prognosis	1	I	MSI	I3, I4 (High Immunoscore)
Intermediate Prognosis	2, 3	IIA, IIB, IIC, IIIA, IIIB, IIIC	NA	I2 (Intermediate Immunoscore)
Likely Negative Prognosis	4	IVA, IVB, IVC	MSS	I0, I1 (Low Immunoscore)

Table 2.3: The range of values for each clinical measure and their likely prognosis. Acronyms: NA, Not Applicable.

whether the cells are evenly distributed or from highly localised aggregations.

In contrast to Immunoscore and Immunomap, the analysis performed in this investigation is agnostic to specific cell types. In addition, this does not provide insight into the spatial organisation of higher-order structures despite research suggesting that these features, such as TLS, are key to understanding the suppressive effects of the immune system [130, 102, 131, 132]. Therefore, the following analysis plans to go beyond this technique and gain insight into the complex spatial network that is the TME of CRC.

2.3.5 Outlook

The values which each of the clinical scores can take, including Immunoscore, are presented in Table 2.3.

The clinical measures presented here aim to categorise the likelihood of a positive or negative prognosis, depending on a perception of the most important biological mechanisms. For example, cell differentiation based on its morphology as interpreted by a pathologist determines the grade of a cancer. This relies on the assumption that rapid cell growth presents observable abnormal morphology. However, this subjective criteria is prone to inter-operator variability. Instead, Microsatellite Status presents a quantitative metric for prognosis. Yet, the technology suggests mutation burden and does not indicate either cell phenotypes that are influential or infer interaction between cell phenotypes. Without this spatial information mechanisms which can affect localised features of a tumour are enough to influence a positive prognosis are concealed as MSS[1]. TNM Classification does provide some spatial insight on the

[1]A hypothesis based upon the results presented in Section 2.11, Subsection 2.11.1.

spatial relationship of tumours but at a tissue level, such as spatial relationships to lymph nodes or structures of the colo-rectum, unable to relate prognosis to specific cell-cell interactions. Therefore, it can be understood why Immunoscore has become a popular tool, quantifying two specific cell phenotypes which have provided a more accurate indication of prognosis. Nevertheless, whether these two cell phenotypes characterise the whole of immuno-tumour response or are confounding factors from a complex interplay of other immune phenotypes remains a question. That is why we envisage the following analysis to not only indicate prognosis but unveil specific spatial distributions indicative of individual cell interactions that might better characterise immunological response in CRC.

2.4 Processing the Images

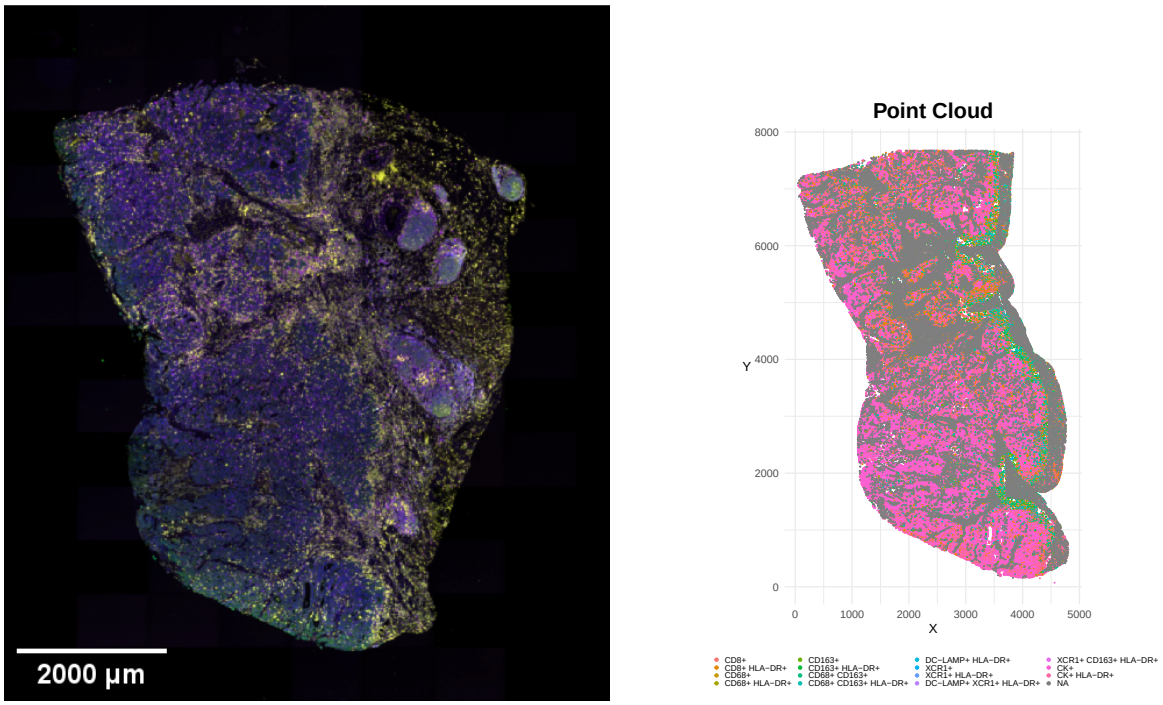
To carry out our analysis, we describe the origin of the data and how surgical biopsies have been prepared for multi-agent, multi-region, and spatially-resolved analysis.

2.4.1 Acquisition

The samples on which the HALO1079 data set is based are propriety data of AstraZeneca®. Based on a cohort ($n = 96$ patients) of primary tumour surgical resection CRC samples initially analysed by Indivumed®. The data is supplemented by information about a patient's surgery date, relapse-free time, patient response, patient survival, medical prescriptions, chemo-therapies, hormone therapies, targeted therapies, previous cancers if metastatic and to which locations if so. The relevant supplementary data for comparison between responding and non-responding patients are Grading, TNM Classification, Multisatellite Status and an Immunoscore-like metric.

The data set was then re-imaged by collaborators at AstraZeneca using mIF imaging with the Vectra Polaris system [133]. This is referred to as the HALO1079 data set.

Sample preparation, background correction and photo-bleaching were quality assured by an expert microscopist using PhenoChart Whole Slide Viewer [134], resulting in an image for



(a) (b)

Figure 2.1: Processed data of the HALO1079_18 sample. (a) Validated mIF image (b) The point cloud for the CT and IM.

each patient sample similar to Figure 2.1a. Then, cell centroid locations and respective marker association were then performed using HALO [135] with an in-house workflow, as visualised in Figure 2.1b.

2.4.2 Data structure

The format in which we receive the information of each sample includes a file with information regarding the region referred to as the central tumour, and often with a supplementary file of the region referred to as the invasive margin. Each file contains a table with information pertaining to the cells identified in these regions. The available categories of information are as follows:

- The image file path with which the processing was performed upon,
- The pathological region, i.e. central tumour or invasive margin, with which the processing was performed upon,

- Object ID of a cell,
- The bottom-left and upper-right coordinate that defines the bounding box of the identified cell,
- A series of binary values referring to a cell's marker identification,
- A series of marker values both within the identified cell's nucleus and cytoplasm,
- Individual cell geometry metrics such as cell area, cytoplasm area, nucleus area, nucleus perimeter, and nucleus roundness,
- The results of the cell classifier model locating cells as either in the tumour or stroma, and
- The parameters used in the cell classification model.

In our analysis, we reduce this information to a reduced form which is commonly used in spatial biomedical imaging research [136, 137]. Specifically, where we obtain the coordinates of the cell and the respective marker associated with it. Also, we import the pathological region which cells are situated within. Therefore, cells can be allocated to reside in the central tumour or invasive margin. Cell geometry was not considered in this analysis, however, it is another variable which can be imported for future analyses. We refer to the final structure as the *column data* of the sample. It is with this information that we have carried out our analysis to characterise the TME.

The primary objective of this analysis is to characterise the spatial patterning of multiple cell types. As a consequence, one would expect that differences in spatial patterning could reflect the differences in patient response. Patient response is different to patient survival since the supplementary data accompanying this data set does not follow patients for a long enough time. We are instead, interested in the prognosis of cancer and the identification of cellular distributions in TME which indicate a positive treatment response that is observable in subsequent clinical visits. The supplementary data categorises patients into one of the four

categories for response: complete response, partial response, progressive disease and unknown. We simplify this categorisation into two groups, those that experience a pathological complete response, and as such we call *Responders*, and those that do not, hereby referred to as *Non-Responders*. This binary status for patients is based on the last known pathological assessment of a patient. It is with this categorisation that we expect to compare subsequent metrics to classify prognosis.

2.5 Areas, Counts and Densities

A preliminary analysis of any study analysing whole slide images of cancer surgical resections is to begin with the number of cells identified with a specific marker. This is then normalised by the area to obtain the density of each marker within each sample. Despite appearing as a trivial measure, there are some consequential decisions to be made. The results of this will be used as inputs to subsequent techniques, such as computing significant local densities between cell phenotypes which contrast to global averages.

2.5.1 Areas

The primary decision is how to calculate the area of a sample. In our scenario, we do have access to the area of samples provided by the HALO software platform, for which the processing of the data set was performed (see Section 2.4, Subsection 2.4.1). However, it is noted that each sample has been split into the central tumour and the invasive margin[1]. It is in this format that we receive the information about the area. Therefore, to calculate the entirety of the area of the entire sample which we hope to utilise, we must combine both regions and calculate the whole area. Unfortunately, this is not as simple as adding the area of each as the annotations for an invasive margin can intersect with the area considered the central tumour, as demonstrated by Figure 2.2.

[1]See Section 2.4, Subsection 2.4.2

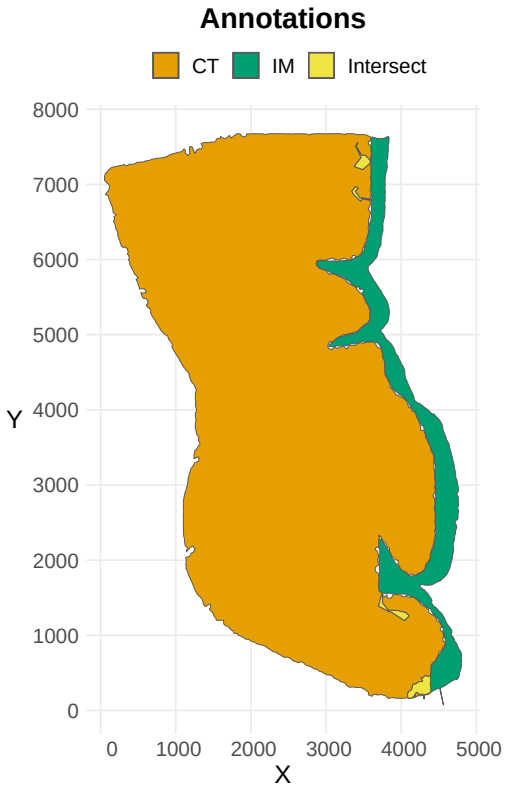


Figure 2.2: The annotation boundaries of a sample (HALO1079_18) for the central tumour (orange) and the invasive margin (green) with which the areas intersect each other (yellow).

To reconcile this, we must add the column data for both annotations together, remove duplicate cells and recalculate the area for both regions based on the point coordinates of the cells. First, this requires identifying the boundary of the sample. This is a well-recognised problem in spatial analyses and we discuss the particulars in Chapter 3 [1]. Here, a brief introduction will be given for the three algorithms since we apply all the methods to our samples to obtain bounds on the true area.

The first algorithm considered is the *convex hull*. One can consider the release of a rubber band around all the points; the perimeter of the final rubber band's position is defined as the convex hull for that set of points. This is the most straightforward computation of a polygon's boundary from point coordinate data since no additional parameters are needed. However, the underlying assumption that the polygon is convex can overestimate the true area by generalising concave boundaries as straight lines. An example of this can be seen in Figure 2.3.

Instead of assuming a convex polygon, a concave polygon allows for exterior angles at vertices of the polygon to be less than 180° or, equivalently, the interior angles of any vertex being greater than 180°. In most cases, this can allow for boundaries to have a more accurate approximation of the true boundary of a series of point coordinates. In other words, the concave hull obtains a boundary equivalent to "shrink-wrapping" a set of points. So where boundaries curve inwards and back out again, unlike the convex hull which draws a line between the outermost start and endpoints, the concave boundary would draw the boundary between points that form the curve of this boundary.

The second algorithm we consider is therefore named after this principle. It is called the *concave hull* algorithm. In reality, there are many ways to reach a concave hull of a set of point coordinates, such as our third algorithm. However, the following algorithm, as defined by the GEOS software package [138], is widely utilised as **the** concave hull algorithm. To summarise, the algorithm computes the Delaunay triangulation[2] of the set of points. Then, an input parameter, referred to as the *ratio*, is used to remove all edges with a length larger than

[1] See Section 3.4, Subsection 3.4.2
[2] See Section 2.8, Subsection 2.8.4

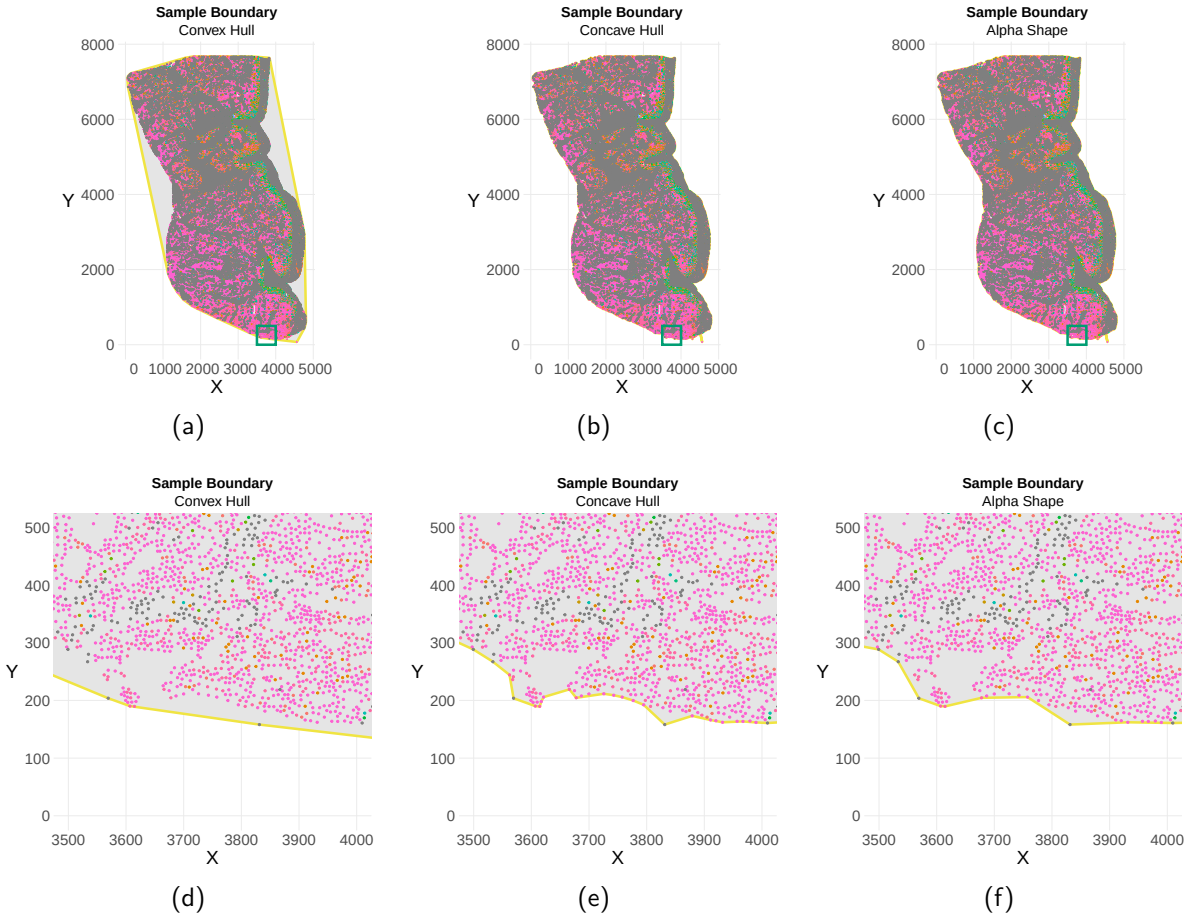


Figure 2.3: Comparison of identified boundaries (yellow line) for sample HALO1079_18 on the whole slide, and for a ROI (green rectangle). Whole slide images are presented using the (a) convex hull, (b) concave hull (ratio $= 0.01$), and (c) alpha shape ($\alpha = 75\mu$m). A ROI is also presented for the same boundary identified by the (d) convex hull, (e) concave hull, and (f) alpha shape.

the proportion divided by the shortest edge in the Delaunay triangulation. Given a reasonable ratio, this can give a much better approximation of a boundary to a set of points. However, the limitation is the necessity of such a parameter.

The third algorithm considered here is called the *alpha hull* algorithm. The principles of the algorithm have a similar basis to the concave hull algorithm described before. For example, in the calculation of the alpha hull, the Delaunay triangulation is computed as before. Thereafter, the algorithm finds the mathematical definition of a set of point coordinates called the alpha hull, as set out by Edelsbrunner et al. [139]. The intuition of the alpha hull is as follows: consider a series of circles (when in 2D) with radius, α, encircling the set of points, and, if there are holes, then around the boundary of such holes. The result is a series of circles where each adjacent circle is fixed between two points from the original point set and contains no other point from the set. The alpha hull boundary is then the smaller perimeter between the pairs of fixed points. An alternative to the perimeter is instead the chord between the two fixed points, this is referred to as the *alpha shape*.

We measure the accuracy of each of the aforementioned methods in more detail in Chapter 3. Here, we discuss the properties of each algorithm which benefit different point sets where we might have some prior information on their construction. Also, we identify extreme scenarios in which the algorithms reach their limits. In our analysis, we apply all the methods to a sample and thus obtain bounds on the true area.

2.5.2 Counts

Previously we have described the panel of markers used in the HALO1079 data set, as presented in Table 2.1. The markers are not distinct in association, therefore some cells can have multiple markers. Therefore, we end up with all the possible combinations of markers for any one cell in the HALO1079 data set in Table 2.4.

Depending on the hypothesis in question, analyses can draw comparisons between different classes of cell phenotype. *Heterogeneity Scores*, consider the relative abundance of different cell phenotypes in a sample. For completeness, we present the results of these metrics when

Cell Marker	Cell Type	Immune/ Epithelial	HLA-DR (+/-)
CK+	CK	Epithelial	-
CK+ HLA-DR+	CK	Epithelial	+
CD8+	CD8	Immune	-
CD8+ HLA-DR+	CD8	Immune	+
CD68+	CD68	Immune	-
CD163+	CD163	Immune	-
CD163+ HLA-DR+	CD163	Immune	+
CD68+ CD163+	CD68 CD163	Immune	-
CD68+ CD163+ HLA-DR+	CD68 CD163	Immune	+
DC-LAMP+ HLA-DR+	DC-LAMP	Immune	+
DC-LAMP+ XCR1+ HLA-DR+	DC-LAMP XCR1	Immune	+
XCR1+	XCR1	Immune	-
XCR1+ HLA-DR+	XCR1	Immune	+
XCR1+ CD163+ HLA-DR+	XCR1	Immune	+

Table 2.4: Table of all possible cell marker combinations in the HALO1079 data set, and their decomposition into cell type, Immune/Epithelial phenotype and HLA-DR status.

applied to the HALO1079 data set but leave the theoretical derivations of the individual metrics to Appendix A (Section A.3). However, counting does not take into account the size of the biological resection being observed. Also, since cell-cell interactions are more likely when cells are in closer contact and therefore have a more significant effect and so consider normalising according to the area.

2.5.3 Densities

Considering the decisions made for both the areas and the counts, the density is usually considered the simple division of the count by the respective area and is applied here. One could imagine a distinction between the global average and the variation in density across the sample at a more local level. For example, an attempt to divide the computed area from before into polygons, similar to the Morista-Horn Index [59, 140], and the variation in density can be compared between samples. For this analysis, we present only the global density in the

pursuit of more mathematically rigorous descriptions of spatial organisation for multiple cell markers.

2.6 Compartment Interaction

The CT and IM are widely known to be important regions in CRC [128, 141]. It is for this reason that the HALO1079 data set is structured in such a way that divides the identified cells into the respective regions. Immunoscore[1] has already been discussed to utilise the density of specific cell markers in these regions.

The following analyses look to combine information from both and combine them into one representative curve for each of the cell types monitored in the HALO1079 data set. The representative curve is referred to here as the Extent of Compartment Interaction (ECI) curve and aims to describe the density of cell markers at distances from the CT-IM boundary interface. The hypothesis is that the invasive margin harbours the majority of the immune response to the inflammation caused by the tumour. Therefore, understanding the extent to which the immune cells have managed to penetrate a tumour could be useful in distinguishing patients who are expected to have a positive prognosis compared to those who will not.

The curve is the ECI function at a range of distances that represents the density value at binned distances from the CT-IM boundary interface. Mathematically this is,

$$\mathrm{ECI}(r, \mathrm{d}r) = \frac{1}{A_{[r-\mathrm{d}r,\, r+\mathrm{d}r]}} \sum_{i=1}^{N} I(r - \mathrm{d}r < \min\{d(b, x_i)\} < r + \mathrm{d}r) \tag{2.1}$$

where r is the distance from the CT-IM boundary interface, $\mathrm{d}r$ is a small step in r, $A_{[r-\mathrm{d}r,\, r+\mathrm{d}r]}$ represents the area for the subset of the domain which points are counted, N is the total number of points in the set X, and $I(r - \mathrm{d}r < \min\{d(b_j, x_i)\} < r + \mathrm{d}r)$ is the indicator function which equals 1 if the point x_i is within the binned distance r from the closest point on the CT-IM boundary interface, $b \in B$, which we calculate the area for.

Computing the boundary of a set of points and the subsequent area have been discussed

[1] See Section 2.3, Subsection 2.3.4

earlier in the chapter[1]. In addition, examples and detailed analysis of the different methods are investigated in Chapter 3[2]. Therefore, we continue to build upon these techniques in this section.

Once the boundary of the IM and CT is obtained, we need to then identify the distances of every cell from the CT-IM boundary interface. To distinguish the cells that are within the IM and the CT, we consider cells within the IM to have a negative value and the remaining cells in the CT to measure a positive value of distance. One could simply plot this result and suggest that this represents the extent to which immune cells from the IM have invaded the CT, such as in Figure 2.4. However, to factor the variation in geometries for binned areas from the CT-IM boundary interface we divide by the area to obtain a density value which can then be compared between samples. Because an estimate for the area is necessary, we require a method to iteratively dilate the CT-IM boundary interface and obtain the area between binned distances. This is referred to as *offsetting* or *buffering*.

Offsetting, also known as buffering, is the method by which a polygon is dilated by a fixed distance from the original boundary [142, 143, 144]. This is necessary as we intend to obtain areas whose difference represents the area for which we are counting the set of binned distances. Offsetting allows for this by encompassing a new area and can subsequently be taken away from the area of the previous dilation distance, and so acquiring the area over that binned distance. For regular 2D polygons, such as a circle or a square, obtaining the area of the dilated version is trivial since well-defined formulas exist for such objects when the difference in dilation is known. In our scenario, this is not the case as can be observed by the IM for one of the samples in Figure 2.2. Hence, a method to obtain the new boundary at a fixed distance from the last for any complex geometry is required. It is also observed that the CT is not covered by all of the new offset areas of the IM. For example, in Figure 2.2 there exists no CT region to the right of the IM. That is why we measure the distance from the CT-IM boundary interface instead of the whole IM boundary. As a result, an example of the binned areas for which we measure density from the CT-IM boundary interface can be seen in

[1] See Section 2.5, Subsection 2.5.1
[2] See Section 3.4, Subsection 3.4.2

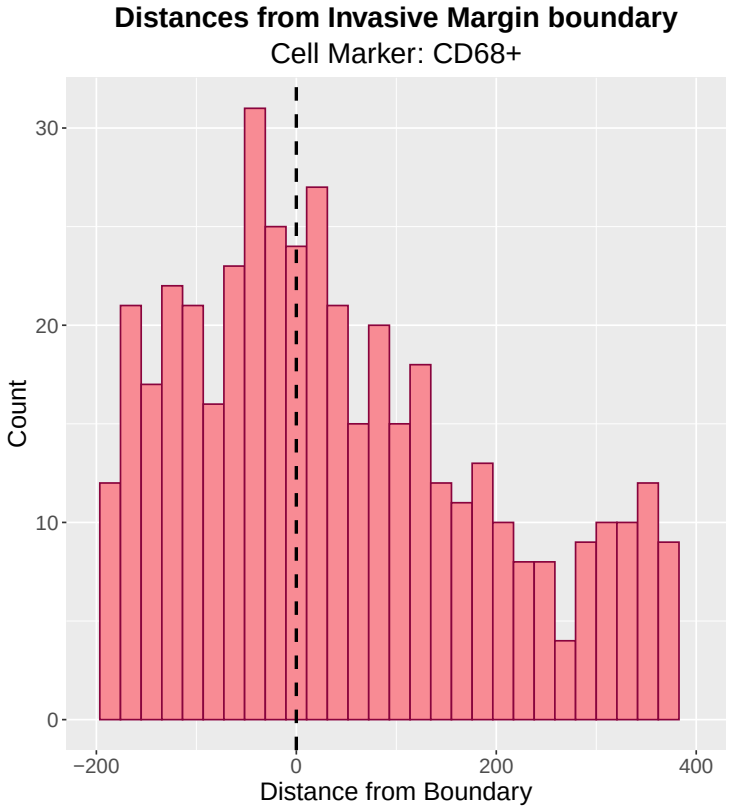


Figure 2.4: Counts of CD68+ cells of a distance from the IM boundary.

Figure 2.5. We only consider the area which is the intersect of the new offset IM boundary and the CT region. Finally, the count for each binned distance is calculated and so the density from the invasive margin is interpolated. The result is the ECI curve. This technique is a novel contribution to the analysis of immune cells. The method quantifies the spatial localisation of cells specific to a biologically relevant region of many cancers. Furthermore, it provides quantitative information to researchers which has previously not been addressed. The ECI curve allows an interpretation of the dynamic interaction of cells with the central tumour by observing a single timeframe and extrapolating density gradients as a moving wave of invasion. The application to the HALO1079 will provide useful results on the most active cell phenotypes interacting across this interface.

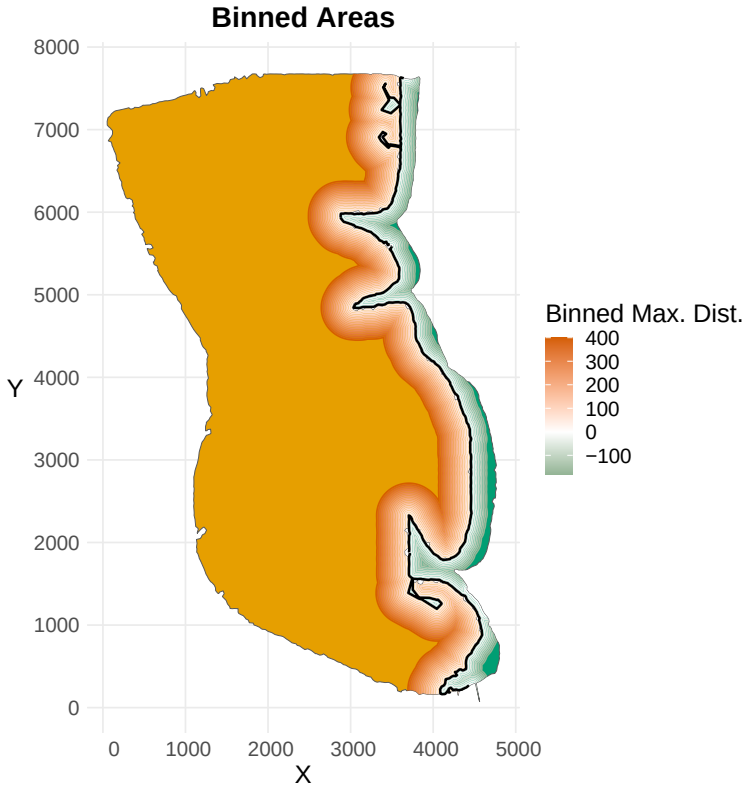


Figure 2.5: Areas of binned distance, shown for the HALO1079_18 sample, with the maximum distance of each binned distance shown as the colour of the area within (-200, 400). Beyond this range, the CT and IM are coloured orange and green, respectively. The black solid line represents the identified CT-IM boundary interface for which distances are measured.

2.7 Spatial Correlation within Multitype Point Patterns

One objective is to understand which of the cells are more or less likely to be in spatial proximity with one another, whether that is within the same phenotype or with other phenotypes. Assuming spatial proximity can infer interaction, an aggregation of interacting cells can initiate a cascade of signalling pathways over a larger length scale. As a consequence, the following methods consider the spatial organisation within cell phenotypes or between pairs of phenotypes over a range of length scales. The set of coordinates and their accompanying attribute, sometimes called marks, is an example of a *multitype point pattern*.

Correlation (whether spatial or not) is a statistical measure of similarity between events. If observations fail to reject the Null Hypothesis, then we assume to have a Null Distribution. In spatial statistics, the Null Distribution often used is the *homogeneous spatial Poisson point process* for the same average densities of the different cell types within the observed point pattern. The name derives from the fact that if a collection of random points is distributed uniformly randomly in some domain, then the number of points in a region of finite size is a random variable with a Poisson distribution. The term homogeneous refers to the distribution of the density for the given domain. The constant, known as the rate or intensity, is the average density of the points in the Poisson point process in a finite subset of the domain. The resulting point process is called a *homogeneous* or *stationary* Poisson point process. In the second case, the point process is called an *inhomogeneous* or *non-homogeneous* Poisson point process, and the average density of points relates to an underlying function on the domain. The Poisson point process has the property that each point is stochastically independent of all other points in the point pattern, which is why it is sometimes referred to as Complete Spatial Randomness (CSR). An example of a homogeneous Poisson point process is shown in Figure 2.6a.

We have constructed three unique artificial spatial point patterns to present the information that can be inferred from different spatial correlation metrics. These are displayed alongside an example of a spatial Poisson point process in Figure 2.6. These three spatial point patterns have been chosen as distinctive features which we see in our mIF images. These three cases are

of particular interest as they concern instances where points are either at high density within the same neighbourhood (see Figure 2.6b) or the localisation of one cell type depends on another cell type (see Figure 2.6c and 2.6c). Specifically, we have chosen Gaussian distributions as examples of higher-density clusters and introduced some random points to initially show the sensitivity of the following metrics.

A comparison of common spatial correlation metrics applied to the point patterns presented in Figure 2.6 are discussed in Appendix A (Section A.4). In the following, we discuss the application of the chosen spatial correlation metric. There is also a series of case studies by Stoyan and Stoyan [145] that illustrate the interpretation of spatial correlation metrics for many other types of spatial point patterns.

2.7.1 Pair Correlation Function

The spatial correlation metric focused upon in our study measures the second-order product density of a point pattern, called the Pair Correlation Function (PCF). The PCF of a stationary point process is expressed as,

$$g(r) = \frac{\varrho'(r)}{2\pi r} \tag{2.2}$$

where $\varrho'(r)$ is the derivative of $\varrho(r)$, the second-order product density. For example, for Ripley's K-function we would have $\varrho(r) = K(r)$. The value of the statistic, $g(r)$, is then compared to the expected value if the two cell types were Poisson randomly distributed. If true, then the distribution would match that calculated via a Poisson distribution, e.g. $K(r) = \pi r^2$, and the pair correlation function would be identically equal to 1 for all r. Otherwise $g(r) < 1$ indicates a negative spatial correlation at distance r, i.e. spatial dispersion, between the two cell types; whereas $g(r) > 1$ indicates a positive spatial correlation at distance r, i.e. spatial clustering.

The PCF has a value at each distance r and as such can be visualised via plots such as those in Figure 2.7. Similarly to the previous spatial correlation metrics, we have the significance interval at the 5% significance level obtained in the same way as the aforementioned Nearest

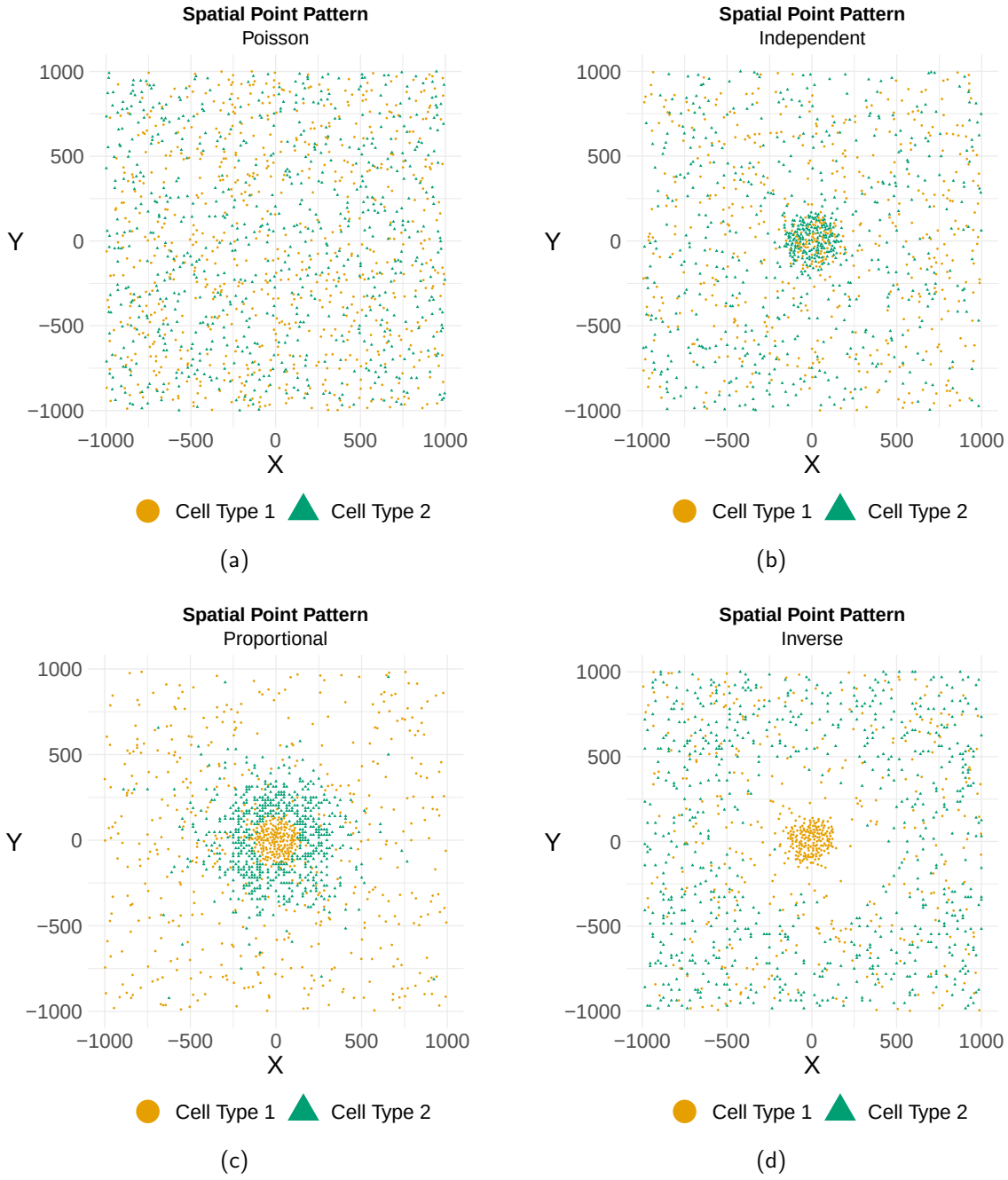


Figure 2.6: The spatial point patterns used to illustrate the interpretation of the following spatial correlation metrics. (a) The spatial point pattern if the points are uniformly random in the domain, following a spatial Poisson point process with similar cell type densities to our observed point patterns. The subsequent plots are when Cell Type 1 has 75% of the points distributed according to a Gaussian distribution with mean at $(0, 0)$ and standard deviation $\sigma = 50$. The remaining 25% of points are uniformly randomly distributed. Then, Cell Type 2 is distributed such that (b) it has the same underlying distribution as Cell Type 1, (c) it has a placement probability proportional to the density of Cell Type 1, or (d) it has a placement probability inversely proportional to the density of Cell Type 1. All spatial point patterns consider the points as centroids of hard-shelled cells where the radius is $r = 10$.

Neighbour Distribution[1]. Note that the definition of L_{ij} is symmetric in i and j, as it is based on the K-function. Therefore, we only need to consider all combinations of the different cell types, resulting in 3 L-function plots for each artificial point pattern.

Across all artificial point patterns, the PCF values are near zero for $r \in [0, 10]$ due to the hard-shelled property imposed on points. It is from herein that we can characterise the different properties of our artificial point patterns.

For the first artificial point pattern we can see in the observed value of $g(r)$ that the within-type comparison and the between-type comparison are considered significant for small values for approximately $r \leq 250$. In addition, the value is inversely proportional to the value of r being measured, so the value falls from a peak to reach a value that fails to reject the Null Hypothesis at the 5% significance level. This reflects not only the diameter of the high-density aggregation but also the type of spatial correlation being experienced. This characteristic is important since we will see in the analysis of the proportional point pattern how we distinguish a different type of spatial correlation. It should be noted that the value that the Null Hypothesis is not rejected in each comparison, although related to the diameter of the high-density aggregation, the value cannot be used as the equivalent of the diameter.

We now look to analyse the observations for $g(r)$ of the proportional point pattern. Similar to before, the within Cell Type 1 high-density aggregation is presented as the characteristic curve discussed in the independent point pattern scenario, with the same approximate values of r considered significant at the 5% significance level. After, we consider the within Cell Type 2 distribution. Here, we initially can observe values of the PCF similar to a high-density aggregation. However, there are two differences in the properties of the PCF. First, the values of $g(r)$ stay higher than the upper bound of the significance interval for larger values of r. Second, the gradient is not consistently decreasing, as shown by a slight flattening from $r \approx 250$ onwards. These properties represent the high-density annulus that Cell Type 2 forms around the high-density Gaussian distribution formed by Cell Type 1. That is, within the radial bandwidth of the annulus, the values of $g(r)$ reflect the values of a high-density aggregation.

[1]See Section 2.7, Subsection A.4.1

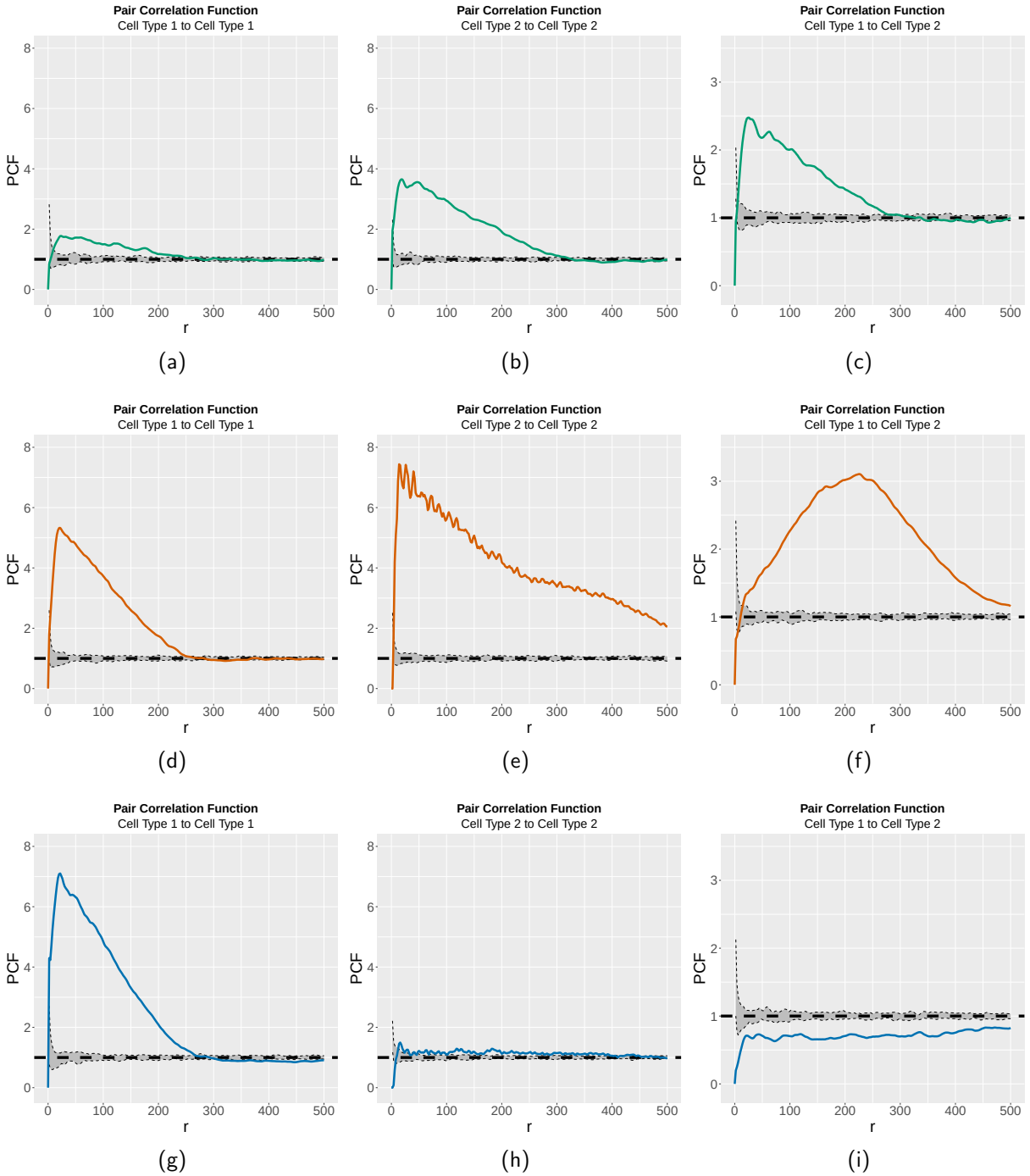


Figure 2.7: The PCF plot for the (a) independent, (b) proportional, and (c) inverse spatial point pattern. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level.

It is then beyond this length scale that Cell Type 2 is only present in a portion of the ever-expanding radius forming the annulus. Therefore, the annulus shape is being measured in the PCF. For the between-type comparison of cell types, we observe a unique distribution which matches intuitively with the construction of the point pattern. That is, from any typical Cell Type 1 to Cell Type 2, at small r values the effect of self-clustering limits the value of the PCF this then rises to a peak in PCF value representative of the radius which most points in the high-density aggregation of Cell Type 1 are within range of the high-density annulus formed by Cell Type 2 around it, and vice versa. It is after this value of r that the PCF decreases again to compensate for a few other points added to the expanding radius, reducing the overall density between Cell Types. Not only do we have now a way to interpret the diameter of the high-density aggregation and annulus, but the relationship between the two and some confirmation of the spatial relationship between the structures.

For the last artificial point pattern, we are expecting to see the same PCF for Cell Type 1 but differences in the remaining two distributions that separate it from previous point patterns. Similar to before, the within Cell Type 1 high-density aggregation is presented as the characteristic curve discussed in the independent point pattern scenario, with the same approximate values of r considered significant at the 5% significance level. It is for the within Cell Type 2 distribution that we begin to identify some specific information that other spatial correlation metrics have not yet revealed. That is the independent distribution within Cell Type 2, with a small bias at some values of r for positive correlation as a result of the point type exclusion from the high-density area of Cell Type 1. Lastly, the between-type comparison reflects the spatial inhibition between point types. The observed PCF rejects the Null Hypothesis for all values of r at the 5% significance level.

In conclusion, the interpretation provided by the PCF goes beyond that of the Nearest Neighbour Distribution and its derivations and the K-function from which it is derived [1]. Therefore, in our analysis of the HALO1079 data set, we analyse the images for these features with the knowledge provided by analysing this series of artificial point patterns.

[1] See Appendix A, Section A.4.

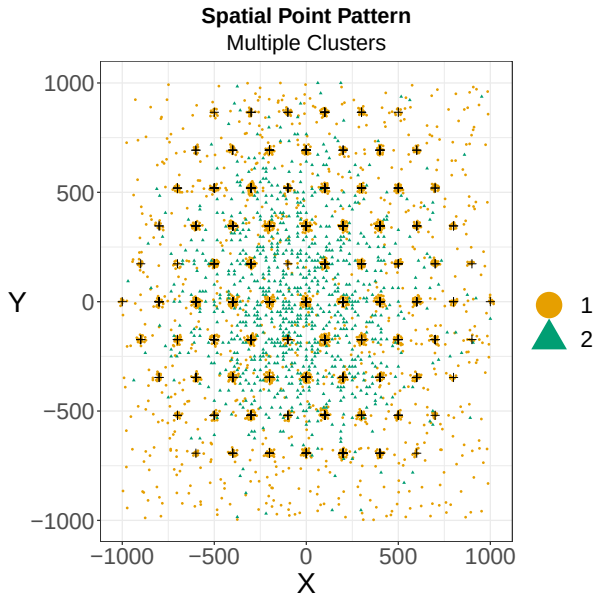


Figure 2.8: The spatial point patterns used to illustrate the use of spatial correlation metrics to interpret spatial heterogeneity. Cell Type 1 has 75% of the points distributed according to a Gaussian distribution with centres on a hexagonal lattice (marked here with a black cross) and standard deviation $\sigma = 10$. The remaining 25% of points are uniformly randomly distributed. Then, Cell Type 2 is distributed such that it has a placement probability proportional to the density of Cell Type 1. All spatial point patterns consider the points as centroids of hard-shelled cells where the radius is $r = 5$.

A comment should be made on the use of the PCF to determine spatial heterogeneity. Often Ripley's K-function and variations thereof are used to infer spatial heterogeneity [146, 147, 148]. To illustrate this, we give an example in the form of the spatial point pattern presented in Figure 2.8. The results for Ripley's K-function, Besag's L-function and PCF are presented in Figure 2.9.

The results of the spatial correlation metrics provide various degrees of insight into the construction of the point pattern, again with the PCF providing the largest signal for spatial heterogeneity. That is, all spatial correlation metrics identify the density heterogeneity with spatial regularity for the within Cell Type 1 measurements. Furthermore, the metrics can reflect the distances which separate the centres of the Gaussian distributions. To do this, we note that the increase in gradient for Ripley's K-function and Besag's L-function begin at $r \approx 200$ and that this is reflected in the PCF as the local peaks in $g(r)$. However, Ripley's K-function and Besag's L-function do not highlight the signal for Cell Type 1 to Cell Type

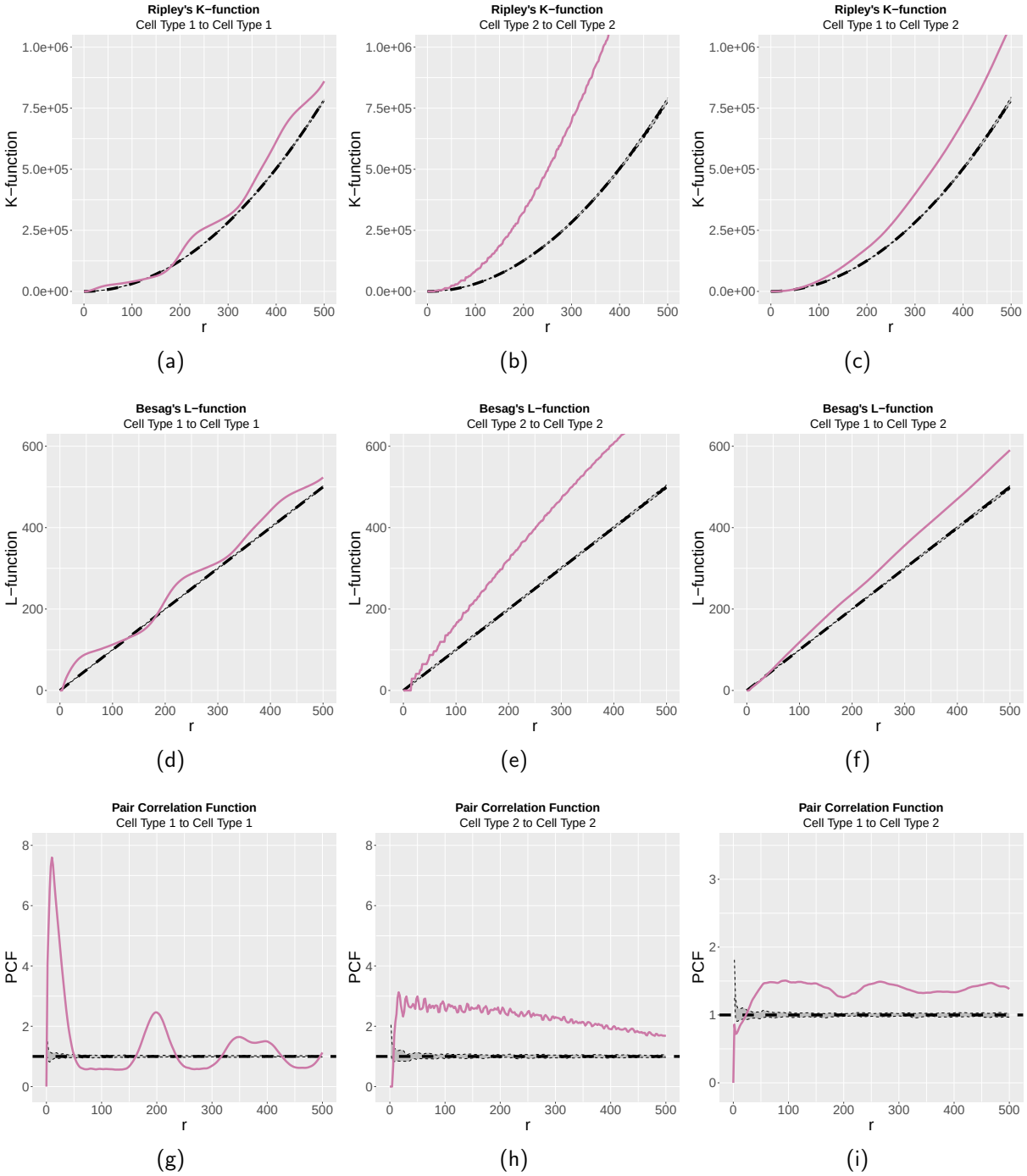


Figure 2.9: Results of spatial correlation metrics to interpret spatial heterogeneity. The plots for the multitype (a) - (c) Ripley's K-function, (d) - (f) Besag's L-function, and (g) - (i) PCF pairwise comparisons. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level.

2 spatial correlation. It is only when inspecting Figure 2.9i that we see a signal for spatial correlation that if overlaid with the within Cell Type 1 observation is the inverse relationship. Simply put, Cell Type 2 is spatially organised close to Cell Type 1 but between the spaces in which the high-density Cell Type 1 aggregations exist.

2.7.2 Comparing Many Permutations / Combinations of Phenotypes using Spatial Correlation Functions

In practice, we have n cell types which implies the number of pairwise comparisons to analyse are either the total number of combinations, such as the case for Ripley's K-function, Besag's L-function and the PCF, or the total number of permutations, as is the case for the Nearest Neighbour G-function or J-function. When comparisons are symmetrical, we have $n!/r!(n-r)!$ combinations of pairwise analyses, where r is the number of point types to be compared. As a result, when comparing the $n = 15$ unique cell phenotypes in the HALO1079 data set, and $r = 2$, the total number of symmetric comparisons produces 105 figures. Otherwise, the total number of asymmetrical permutations is calculated as $n!/(n-r)!$, which comes to 210 figures. Therefore, methods to identify the most important cell-cell correlations become pertinent to save time-consuming manual inspection to verify distributions individually.

To account for this problem, we overlay the average of each patient response and summarise the difference between the two groups. Then, we summarise the difference between the two patient response groups to identify the most significant pairwise comparisons of cell phenotypes. For example, whether there is a positive or negative correlation between the phenotypes compared to the expected spatial Poisson point process; how significant the behaviour is; and, at which length-scale of r is this relationship prominent.

Summarising the G-function has already been attempted by considering the area under the curve for individual pairwise comparisons [149] or by comparing the absolute difference in value between the observed value of the G-function and the equivalent spatial Poisson point process. However, the area under a PCF function cannot as easily be applied. This is due to the scale that the PCF function is measured. Unlike the other metrics where the area under

the curve is directly proportional to some probability or the average number of cells within that radius, the PCF is the derivative of the coefficient to the global density for the radius around the average point. Therefore, it indicates the multiples for which the observed density is changing in gradient. Specifically, the PCF gives scales relative to the multiple of cells counted at a given r compared to that of a theoretical spatial Poisson point process. As a consequence of being a multiplicative coefficient, calculating the area under the curve to the line $g(r) = 0$ with a value of $g(r) < 1$ results in a small area. The equivalent coefficient for spatial clustering, the inverse of the spatial inhibition, would result in a value much greater than 1 and similarly in area. Therefore, it becomes necessary to consider both values with equal weighting to characterise the equivalent contribution of spatial inhibition with spatial clustering.

We suggest calculating the area under the curve whereby values of $g(r) < 1$ contribute to a weighted area under the curve metric as the inverse value. This then gives an equivalent weighting of spatial inhibition to spatial clustering. This is because the metric now considers how many multiples the value is less than the spatial Poisson point process in a similar way that values of $g(r) > 1$ represent multiples of density more. To account for singularities, we re-assign values of $g(r) \to 0$ to values which are equal to the maximum of the function before taking the inverse. It is then, that we can take the difference between the two patient response groups with equal weighting for significant dispersal and/or clustering compared to CSR. It is with the aid of this metric that the most interesting pairwise combinations of cell phenotypes can be identified before the original PCF distributions can be analysed for the features which contribute to the weighted area difference.

To re-iterate the weighted area for a given PCF, $g(r)$, is as follows,

$$wA\,(g) = \begin{cases} \int_{r\in R} g(r)\,\mathrm{d}r & \text{if } g(r) \geq 1, \\ \int_{r\in R} \frac{1}{g(r)}\,\mathrm{d}r & \text{if } g(r) < 1,\ g(r) \not\to 0 \\ \int_{r\in R} \max\limits_{r\in R} g(r)\,\mathrm{d}r & \text{if } g(r) \to 0. \end{cases} \tag{2.3}$$

where R is the range of r being considered.

This means the weighted area difference is equivalent to,

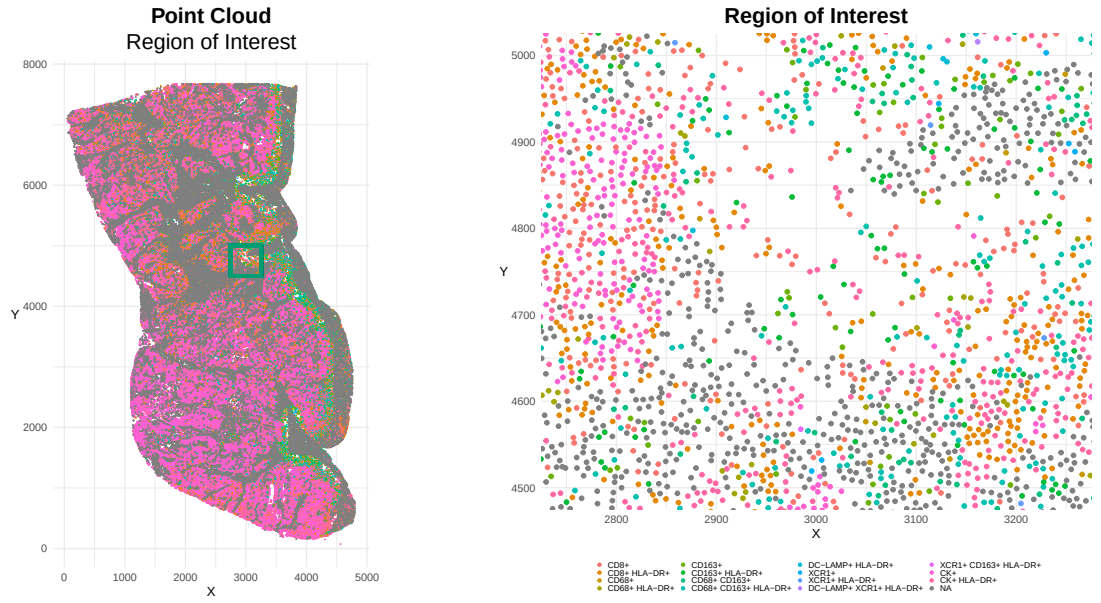
$$wA\left(g_1\right) - wA\left(g_2\right) \tag{2.4}$$

where g_1, g_2 are the PCF of the respective patient response groups. Larger magnitudes of weighted area difference highlight pairs of cell phenotypes with more compelling differences between spatial distributions requiring further investigation. The sign of the weighted area difference also reflects which of the two groups has a larger difference from CSR, with positive values reflecting a larger difference from CSR in $wA(g_1)$ than $wA(g_2)$ and a negative values reflecting the opposite.

The ability to compare multiple pairs of spatial distribution, with the most significant highlighted, allows the quantification of their influence on prognosis when relationships would have otherwise been hard to evaluate.

2.8 Constructing Graphs from Point Coordinates

So far in this chapter, we have a set of point coordinates and their associated cell markers. Instead, one might consider only the inferred interactions between cells, with no reference to their embedding in space. Once constructed, analysis based on this graph, where nodes are individual cells and the edges between them the rule on proximity to consider two cells interacting, can then be computed. However, the results are influenced by the different methodologies of graph construction as well as limited by their computational expense. Cells influence and interact via one of four signalling mediums which are, from the shortest length-scale to the largest: autocrine, juxtacrine, paracrine and endocrine signalling. In our scenario, autocrine is not considered relevant since this relates to intracellular signalling. In addition, although one could argue it could be considered relevant, endocrine signalling interacts with cells at an organ-wide level rather than at an intercellular-level scale. Therefore, the subsequent techniques for constructing interaction graphs consider modelling the two main types of intercellular



(a) (b)

Figure 2.10: ROI to exemplify graph constructions. (a) The point cloud of HALO1079_18 with an overlay (green) of the ROI. (b) Focused view on the ROI.

signalling relevant in mIF images: juxtacrine and paracrine signalling.

Juxtacrine signalling is a type of cell–cell or cell-to-extracellular matrix signalling in multicellular organisms that require close contact. In paracrine signalling, a cell produces a signal to induce changes in nearby cells, altering the behaviour of those cells. Signalling molecules known as paracrine factors diffuse over a relatively short distance (local action), as opposed to cell signalling by endocrine factors, hormones which travel considerably longer distances via the circulatory system. We discuss how juxtacrine and paracrine signalling have been modelled, as well as other intuitive graph constructions. To visualise the results, we focus on the Region of Interest (ROI) in the HALO1079_18 sample shown in Figure 2.10.

2.8.1 Modelling Juxtacrine Signalling

Juxtacrine signalling refers to the direct contact between cells via membrane receptors, junctions or extracellular matrix glycoproteins. Ideally, the interaction between cells would be calculated directly from the processed mIF images. This would enable the identification of "touching" cells by first, offsetting the cell borders by a micron, or a few to allow for sen-

sitivity to precision, and then considering all overlapping cell borders as direct contacts. An example of this approach in breast cancer is presented in [150]. However, since the input data for our analysis is only the point coordinates, we must infer juxtacrine signalling via other means. The first is to refer to metrics from spatial correlation[1]. When computing the PCF[2], we note that the radii of the points can be inferred due to a characteristic absence of cells for the lowest values of r observed. We refer to this as the hard-shelled property of the cells. Since we carry out PCF analysis on the HALO1079 data set, we can infer this property from our computations of PCF to infer a reasonable pairwise distance between cell centroids for adjacent cells. Referring to the results of this analysis[3], it is noted that the PCF curves present such a characteristic hard-shelled cell property for $r \leq 8$, meaning $r \approx 10$ could provide a reasonable estimate for adjacent cells based on their inter-centroid distances. Therefore, an alternative method for computing juxtacrine cells is available in processed mIF images based solely on the point coordinates. Figure 2.11 represents the graph construction for modelling juxtacrine signalling in our reference ROI. However, it is considered more desirable to obtain exact cell geometries as cell phenotypes can vary in size within cell phenotype, let alone across cell phenotypes.

2.8.2 Modelling Paracrine Signalling

When considering the TME, the space is assumed to consist of cells secreting molecules into the extracellular environment. These molecules then diffuse from the source cells and signal to surrounding cells, with their activation or suppression acting in proportion to the concentration of molecule present. Since this type of signalling happens beyond cell-cell contact but depletes with increasing radius, a simplification is to consider an area of influence in which the secretion of molecules is effective [66]. However, when considering interaction as the diffusion of molecules a weight can be attributed from one cell to every other cell following a mathematical interpretation of diffusion. One such interpretation is to assume Fickian diffusion

[1] See Section 2.7
[2] See Section 2.7, Subsection 2.7.1
[3] See Section 2.11, Subsection 2.11.5

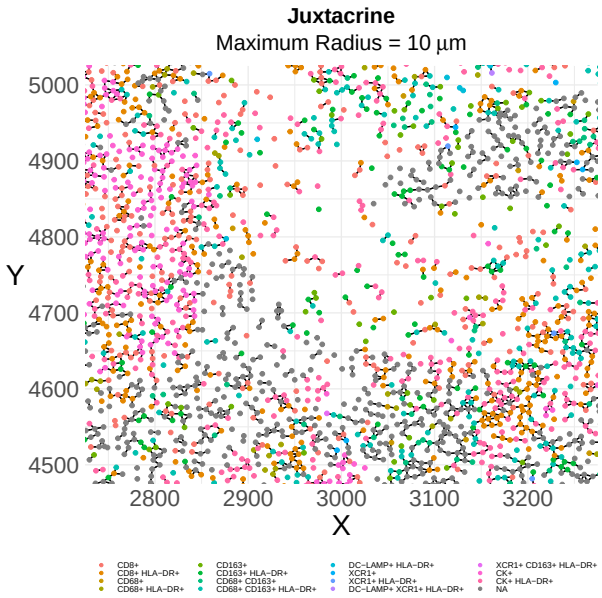


Figure 2.11: ROI from sample HALO1079_18 for which juxtacrine signalling has been modelled as interactions occurring when $r < 10\mu$m.

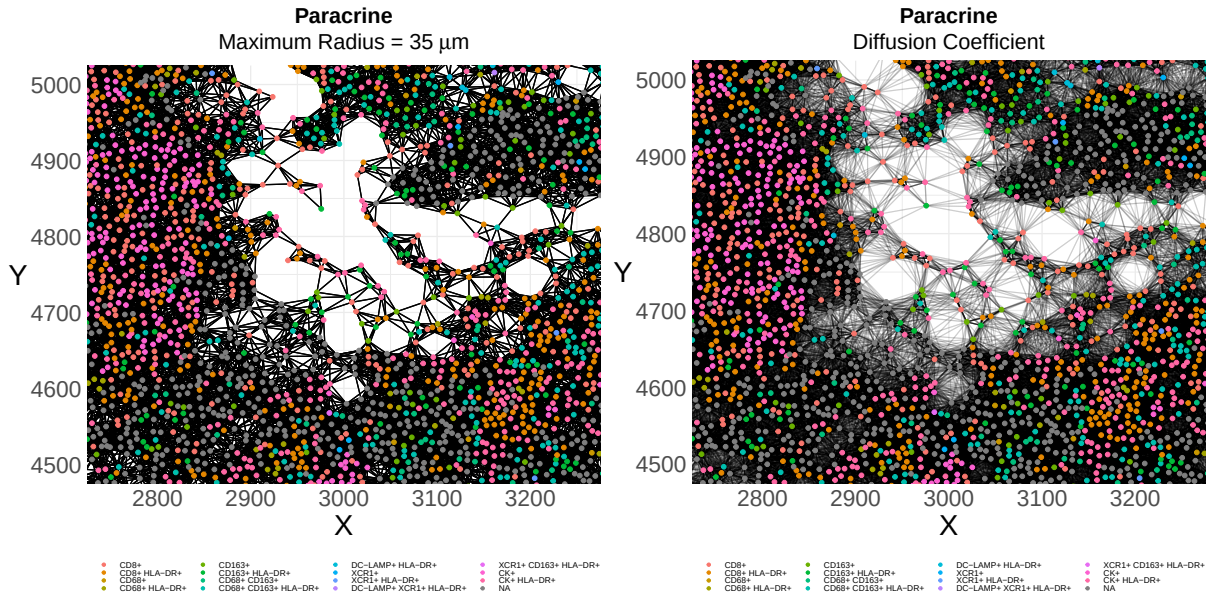
[151] and then measure interaction through Fick's Second Law of Diffusion,

$$c(x,t) = \frac{1}{\sqrt{4\pi D}} \exp\left(\frac{-x^2}{4Dt}\right) \tag{2.5}$$

where $c(x,t)$ is the concentration at x for time t, and D is the diffusion coefficient.

Since we are considering a fixed time point, we can consider t to be fixed and only consider variations in x. Also, we simplify this calculation by considering the diffusion coefficient to be equivalent for molecules secreted by all cell phenotypes. The latter assumption could be adapted to assign diffusion coefficients for specific cell phenotypes relating to the most influential molecules that cell phenotype is known to release. One approach, utilised by Francis and Palsson [152], would be to assign biologically relevant parameters to this derivation of diffusion and to estimate the range for which chemokines are considered effective. However, we assume no such *a priori* knowledge. Considering these simplifications, the weight between any two cells can be attributed based on the distance between them with the following transformation:

$$w_{i,j} = \frac{1}{a} \exp^{\frac{-x^2}{a}} \tag{2.6}$$



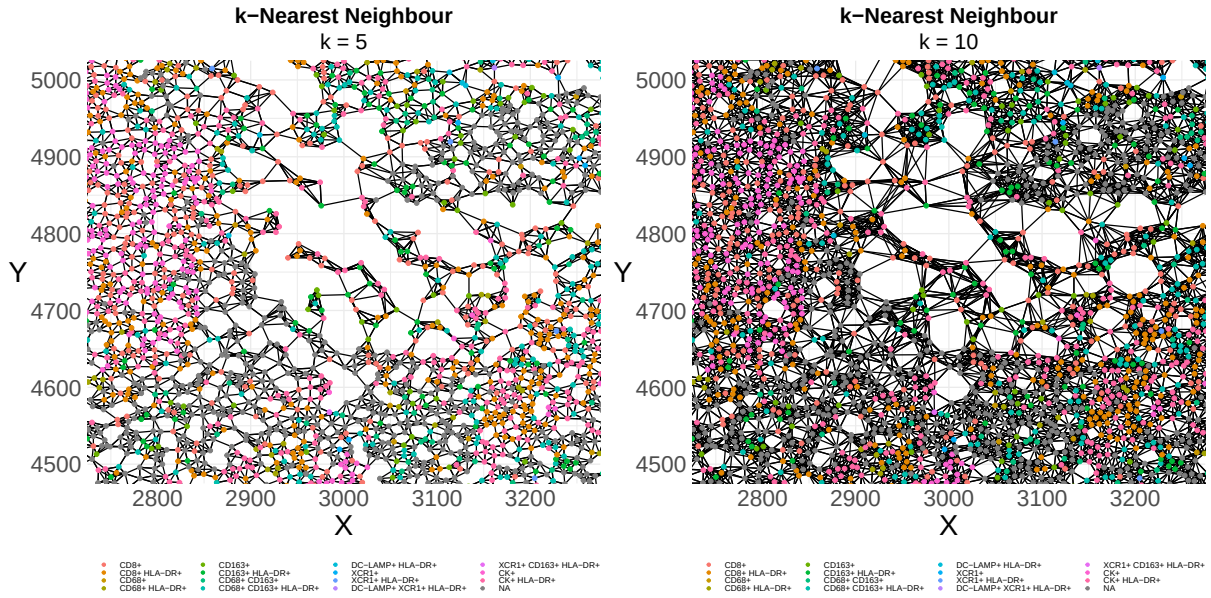
(a) (b)

Figure 2.12: ROI from sample HALO1079_18 for which paracrine signalling has been modelled as (a) interactions occurring for $r < 35\mu$m, and (b) the diffusion coefficient defined by Eq. 2.6 is attributed to the weight of edges, depicted as the proportional relationship to the transparency of an edge. For (b) a threshold on edge weight has been applied for eligibility, meaning only the interactions with a significantly large enough diffusion coefficient remain.

where $w_{i,j}$ is the weight for the edge between the i-th cell and j-th cell, and a is a coefficient estimating the length-scale of diffusion. As a result, all cells are now considered to be interacting with one another but with diminishing influence depending on their intercellular distance. The results of such a graph construction are visualised in Figure 2.12b. Albeit the most biologically relevant graph construction, the computation of pairwise distances and subsequent edge weight attributes becomes computationally expensive. As a consequence, other computationally-optimised graph constructions are adopted. We discuss two such approaches, namely the kNN graph and Delaunay triangulation.

2.8.3 k-Nearest Neighbour Graph and Adaptations

The kNN graph has been a popular construction of graph from point coordinates in biomedical image analysis [46, 76, 153]. The construction considers the distance from every cell to every other cell and then assigns the closest k neighbours as those interacting. However, many



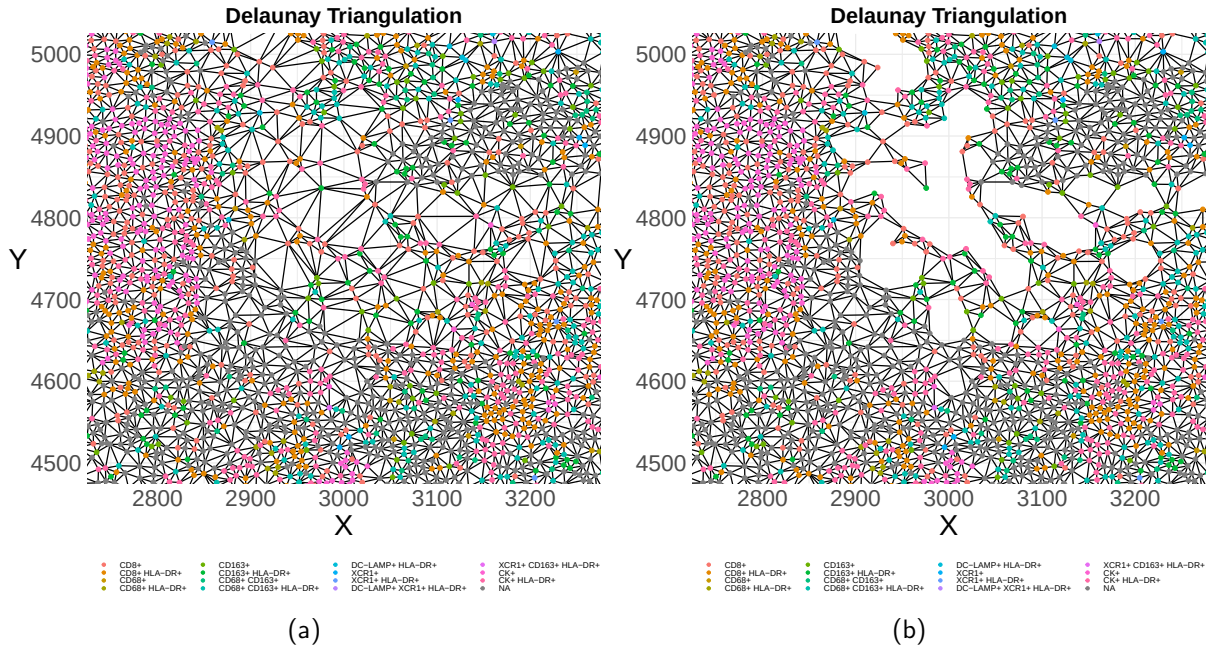
(a) (b)

Figure 2.13: ROI from sample HALO1079_18 for which the kNN graph is constructed, with (a) k$= 5$,and (b) k$= 10$.

algorithms are more efficient at calculating the k nearest neighbours by not having to compute every pairwise distance. From a biological perspective, this representation is the most arbitrary, since it is not restricted by spatial distances or local cell densities. Subsequently, edge lengths can have a less meaningful biological relevance. As the value of k increases, the number of these edge lengths increases too. Accordingly, an optimal k is often sought. Alternatively, an adaptation to the final graph construction can be applied where all edge lengths above a threshold distance are pruned. Finally, we note that the interaction is a directed relationship. For example, for any two cells i, j, the fact that cell i is a neighbour of cell j does not necessarily imply that cell j is a neighbour of cell i.

2.8.4 Delaunay Triangulation and Adaptations

A Delaunay triangulation for a set of point coordinates, also known as a Delone triangulation, is a triangulation such that no point is inside the circumcircle of any triangle in the Delaunay triangulation. Delaunay triangulations maximise the minimum of all the angles of the triangles in the triangulation and so are more likely to avoid sliver triangles. Representing the interaction



(a) (b)

Figure 2.14: ROI from sample HALO1079_18 for which the Delaunay Triangulation is constructed, with (a) no maximum edge length, and (b) maximum edge length $\max(l) = 35\mu$m.

between cells as a Delaunay triangulation results in the graph construction presented in Figure 2.14a. To provide a more realistic biological interpretation, a parameter can be introduced to reflect the maximum area of influence, as discussed for paracrine signalling. Therefore, arbitrarily long edges which represent distances too large for cells to influence over are removed. This has proven useful in applications to CRC data sets [74, 154]. We present this graph construction in Figure 2.14b. This combines features of paracrine modelling, i.e. the maximum area of influence a cell might have, whilst interpreting interactions as a propagating influence through a cell's neighbours, resulting in an accumulative effect of neighbourhoods. Therefore, we consider this construction in our calculations for graph-based metrics.

2.9 Cellular Neighbourhoods

Niches, Cliques and Neighbourhoods are all terms used to refer to the organisation of cells into aggregations. They can often refer to a specific scale of interest. For example, in graph theory

3- or 4-cliques[1] in a graph are often investigated since there are not too many permutations to identify computationally. Then, the observed presence can be contrasted with the probability of occurrence in a randomly generated graph. Therefore, it is often assumed that a clique is the smallest unit of a neighbourhood, as cells in a clique are often in immediate proximity to each other for all of them to be each other's 1-st neighbour[2]. Niches can have specific definitions in graph theory, such as niche graphs where a niche graph, $G = (V, E)$, is a graph such that two vertices x and y are adjacent in G if and only if they have a common vertex in D, a corresponding acyclic digraph $D = (V, A)$ [155]. However, when the term niche is used in the context of biology, the definition reflects aggregations of cells which interact to perform an organ-specific specialised role [156, 157]. These aggregations tend to refer to aggregations with 8 to 50 cells. The broadest term in both classification and scale, is a neighbourhood which encompasses all of the above and larger structures. In general, a definition for a cellular neighbourhood can be considered as a collection of cells which have some input, which then function collectively to produce a specific output. The following measures in this section are based upon this concept of neighbourhood, where interacting cells, i.e. neighbours, are determined by the edge drawn between two cells depending on the graph construction[3].

2.9.1 Infiltration Score

A term often referred to in computational pathology is whether a tumour is considered a *hot* tumour or a *cold* tumour [110]. These terms refer to the invasive nature of immune cells in the CT. A categorisation proposed by Schapiro et al. [69] considers three states of immune cell distribution in the CT. They consider the tumour architectures to either be:

- **Cold:** a small number of immune cells are present in the tumour,
- **Mixed:** a large number of immune cells dispersed evenly across the whole tumour, or

[1] See Appendix A, Def. A.1.12 for definition.
[2] See Appendix A, Def. A.1.11 for definition.
[3] See Section 2.8

- **Compartmentalised:** tumour and immune cells self-cluster in separate, non-intermixed compartments.

To quantify the extent to which these distributions are observed, an infiltration score is suggested, which calculates the ratio of the number of immune-tumour interactions to the number of immune-immune interactions. Generalised extensions of this score have been implemented, one being the ATHENA library [158]. In the ATHENA package, the *Infiltration Score* is defined as

$$\frac{\text{Number of neighbours between Cell Phenotype } i \text{ to Cell Phenotype } j}{\text{Number of neighbours between Cell Phenotype } i \text{ to Cell Phenotype } i} \tag{2.7}$$

where i and j refer to the pair of arbitrary cell phenotypes considered. A neighbour is determined by the edge drawn between two cells depending on the selection of the graph construction[1]. So in the context of hot or cold tumours, and application to the HALO1079 data set, this would consider the Immune-Epithelial phenotypes. However, as demonstrated by the generality of Eq. 2.7 this can compare any pairwise comparison of cell phenotypes and so we consider the calculation of infiltration score through the computation of the Neighbourhood Analysis Score presented in Subsection 2.9.3

2.9.2 Coordinated Cellular Neighbourhoods

When first visualising mIF images, the complexity and quantity of cells which are characterised can be overwhelming. Viewing each cell irrespective of phenotype and then attempting to quantify differences is a task that requires years of training as a pathologist.

To simplify this task, Schürch et al. [46] consider grouping neighbouring cells and then computing phenotypic neighbourhoods based on the composition of the tumour. Their approach allocated every cell to its 10 nearest spatial neighbours, which was referred to as its *window*. The cell phenotype composition was then determined for each window. Then, coordinated cellular neighbourhoods were then identified for each window by converting each window

[1] See Section 2.8, Subsection 2.8.4

into a vector of the cell phenotype frequencies, and the windows were subsequently clustered using the k-Means algorithm, with $k = 10$. To validate the identification of the coordinated cellular neighbourhoods, the algorithm was applied separately to opposite ends of sample prognostic classification in a CRC cohort. The identified coordinated cellular neighbourhoods were reported to be comparable across patient groups, meaning the common coordinated cellular neighbourhoods were therefore not an artefact of the data-merging process. Therefore, they conclude that, although visually distinct as a result of the presence of TLS, shared underlying architectural components could be defined as universal coordinated cellular neighbourhoods.

The approach was later reproduced by Philips et al. [76], for which they identified a specific composition of coordinated cellular neighbourhoods to be apparent in patients who responded to the immunotherapy pembrolizumab than those who did not. This observation led to the creation of *SpatialScore*, which was defined as the average of the ratio between the distance from CD4+ T cells to the nearest Tumour cell and the distance from CD4+ T cells to the nearest Treg cell. Mathematically, we can express this as

$$\text{SpatialScore} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{T cell } i \text{ effector activity}}{\text{T cell } i \text{ suppressive activity}} \tag{2.8}$$

where,

$$\text{T cell } i \text{ effector activity} = \min_{\forall x_{\text{Tumour}} \in X_{\text{Tumour}}} d(x_i, x_{\text{Tumour}}) \tag{2.9}$$

$$\text{T cell } i \text{ suppressive activity} = \min_{\forall x_{\text{Treg}} \in X_{\text{Treg}}} d(x_i, x_{\text{Treg}}) \tag{2.10}$$

with $d(x_i, y)$ as the distance between two points, with $x_i \in X_i \subset X$ representing point of the set of all CD4+ T cell points, with cardinality n, $x_{\text{Tumour}} \in X_{\text{Tumour}} \subset X$ and $x_{\text{Treg}} \in X_{\text{Treg}} \subset X$.

2.9.3 Neighbourhood Analysis Score

Neighbourhood analysis [69], also referred to as spatial enrichment analysis [110], enables the measurement of cell phenotype interaction by comparing the expected and then observed composition of neighbours for a given cell phenotype. A neighbour is determined by the edge drawn between two cells depending on the selection of the graph construction[1]. *Neighbourhood Analysis Score* is such a metric where the expected number or proportion of a neighbouring cell phenotype is then compared to the random permutation of cell phenotypes. The Null Distribution of neighbourhood composition is then used to compute the significance of the observed neighbourhood composition for a given cell phenotype. A choice in averaging neighbourhood compositions is necessary since previous studies have found use cases for each of their nuances [69, 110, 150]. Three versions of averages can be computed, with similar interpretations. These are: *classic*, *HistoCAT* and *proportional*. They are defined as,

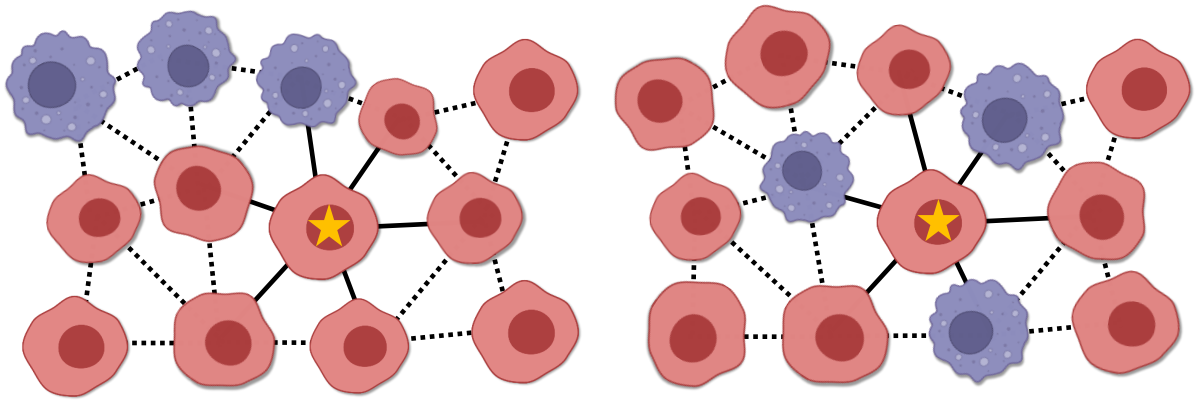
- **Classic:** For the classic averaging of the neighbourhood analysis score, the average number of the neighbouring cell type is computed.
- **HistoCAT:** For the HistoCAT averaging the neighbourhood analysis score, the average number of the neighbouring cell type is computed only for neighbourhoods where at least one cell of the neighbouring type considered is in the neighbourhood.
- **Proportional:** The proportional flavour computes the average proportion of the neighbouring cell type in the neighbourhood, i.e., divides the number of neighbours of the neighbouring cell type by all cells considered in the neighbourhood.

In contrast to the classic or HistoCAT averages, the proportional average is not influenced by the number of cells in a neighbourhood and the score is bounded in the range of [0,1].

We present in Figure 2.15 an exemplary set of cells with a related graph construction. The results of each of the averages computed for such a neighbourhood are displayed in Table 2.5.

Understanding the derivation of the above metrics, we can observe that the Neighbourhood Analysis Score encompasses results reported on by both the Infiltration Score and quantifying

[1] See Section 2.8



(a) (b)

Figure 2.15: An exemplary set of cells a related graph construction. We consider a single neighbourhood centred on a single cell (orange star) and its neighbours (solid black lines) as well as the global average for all other interactions (dashed black lines). (a) A permutation of cell phenotypes where immune cells (purple speckled frilled cells) are excluded from most epithelial cell neighbourhoods (red homogeneous smooth cells). (b) A permutation of cell phenotypes where immune cells (purple speckled frilled) are mixed with epithelial cell neighbourhoods (red homogeneous smooth cells).

	Local		Global	
	Excluded	Mixed	Excluded	Mixed
Classic	1	3	1.400	1.300
HistoCAT	1	3	1.500	1.444
Proportional	0.167	0.500	0.190	0.372

Table 2.5: Table of the Neighbourhood Analysis Scores for Immune cell neighbours of Epithelial cells for both permutations (*Excluded* and *Mixed*) of the exemplary set of cells with a graph construction. For each permutation of cell phenotype, the classic, HistoCAT and proportional averages of the Neighbourhood Analysis Score were calculated. To aid interpretation, the score is presented for a single neighbourhood centred on the indicated epithelial cell (*Local*) and then the average of the whole network (*Global*).

Coordinated Cellular Neighbourhoods. The Infiltration Score can be computed by simply applying the Neighbourhood Analysis Score but for the class of discrete phenotypes that identify cells as tumour or immune. The Neighbourhood Analysis Score reports a similar value as the Coordinated Cellular Neighbourhood but instead of a discrete neighbourhood, the value is attributed to each cell. This means the arbitrary allocation of points into discrete neighbourhoods is not necessary, omitting the choice of central cells which define them as well as the number of nearest neighbours. Instead, we have a value for every cell and the neighbourhood it might influence with no additional confounding factors for neighbourhood delineation. Then, the phenotype proportions of neighbouring cells would be input to high-dimensional clustering algorithms, such as the algorithm introduced by Javed et al. [74], to identify the discrete categories of phenotype interaction proportions. As a consequence, the discrete categories could then be overlaid back onto the images as their colour. With this, and our translation of the Infiltration Score, we see that the computation of the Neighbourhood Analysis Score can be used to evaluate the same insight for similar-scale neighbourhoods. In our analysis, we report on the proportional Neighbourhood Analysis Score since we are interested in the prominence of phenotypes in a neighbourhood which are different between response groups. The alternative perceptions are then available as mentioned previously from the single computation, providing a more robust result in comparison to Coordinated Cellular Neighbourhoods.

2.10 Implementation

All scripts to produce the following results have been written in the coding language R [159]. The scripts have been compiled into a library which we call *TSUNAMI*. The acronym TSUNAMI stands for "Topological and Spatial Utensils for Neighbourhood Analysis in Multiplex Images".

The core functionality of the library is a series of algorithms that compute the metrics discussed in this chapter. The library has explicit dependencies on the sf [160, 161], igraph [162], polyclip [163], RSMLM [164] and ggplot2 [165] libraries.

The computations described were performed using the University of Birmingham's BlueBEAR High-Performance Computer service, which provides a High-Performance Computing service to the University's research community.

2.11 Results

We present the results that apply our exploratory analysis of the HALO1079 data set. The analysis is compared to the binary response of each patient[1], i.e. Responder or Non-Responder.

The data set is formed of $n = 115$ whole slide images. However, as discussed earlier in the chapter[2], $n = 90$ patients have information about their response, where 72 patients are considered to be Responders and the remaining 18 are Non-Responders. However, only 45 have a corresponding annotation for the IM. Therefore, some results are restricted to this subset of the data set instead. Of the 45, it becomes clear when verifying the counts of associated markers, that two samples (namely, HALO1079_025 and HALO1079_063) have poor-quality cell marker identifications too. In many of the analyses, we are then left with a remaining $n_{(\mathsf{CT},\,\mathsf{IM})} = 43$ samples, where 36 are considered Responders and 7 are considered Non-Responders.

Our approach is to apply a series of metrics in an attempt to discern differences between the Responders and Non-Responders. The metrics will be presented in the order they are presented earlier in the Chapter. To summarise, we first consider clinical scores that are currently used to discern patient prognosis. Then, we outline the make-up of the cohort being analysed by presenting the areas, counts and densities. Next, we consider heterogeneity scores which only consider the proportions of cells in each sample, although localised versions of these metrics are available. From then on, we consider metrics that take into account the spatial organisation of the cells, beginning with our ECI curve comparison and spatial correlation. Later, we introduce our graph constructs to consider the concept that cells interact more prominently with their neighbours. Finally, we perform our cellular neighbourhood analysis whereby the perspective

[1] See Section 2.4, Subsection 2.4.1
[2] See Section 2.4, Subsection 2.4.2

of complex spatial networks comes to fruition through the Neighbourhood Analysis Score.

2.11.1 Clinical Measures

There are three clinical scores which we have discussed earlier in the chapter[1]. For all but the Immunoscore-like percentile, this was presented as the metadata alongside the HALO1079 data set. As such, the clinical scores are calculated for all $n = 96$ patients, except for the Immunoscore-like percentile which is calculated for the $n_{(\text{CT, IM})} = 43$ subset.

We present the results of the Grading[2] associated with each patient in Figure 2.16. We can observe that the association of Grading between Responders and Non-Responders is distributed such that a higher percentage of individuals who responded are identified to have a "low" Grade 2 score compared to that considered Grade 3. The distribution of Non-Responders appears more evenly split between Grades 2 and Grades 3. In addition, a small number of patients are in both the lowest grade (G1) and the highest grade (G4), with Responders having a similar percentage in the lowest grade but none in the highest. We compare the result to the Null Hypothesis that Grades present equal percentages of Responders as Non-Responders. We performed the Chi-Squared test (χ^2 test) to determine significance. The results of this test show that we cannot reject the Null Hypothesis at the 5% significance level.

The results of the TNM Classification[3] shares similarities with the Grading results, however, it is important to explicitly state that they do not represent the same classification system. First, we can observe in Figure 2.17, is that the two groups can be broadly distinguished by stages I and II being primarily comprised of Responders and then Non-Responders being classified as either III or IV. This is reinforced by the majority of Responders having been classified as stage IIA. On the contrary, the Non-Responders have their highest percentage of patients classified as stage IIIB, closely followed by stage IVA. The χ^2 test cannot be performed as the contingency table invalidates the assumption on minimal expected observations necessary for such a Hypothesis test.

[1] See Section 2.3
[2] See Section 2.3, Subsection 2.3.1
[3] See Section 2.3, Subsection 2.2

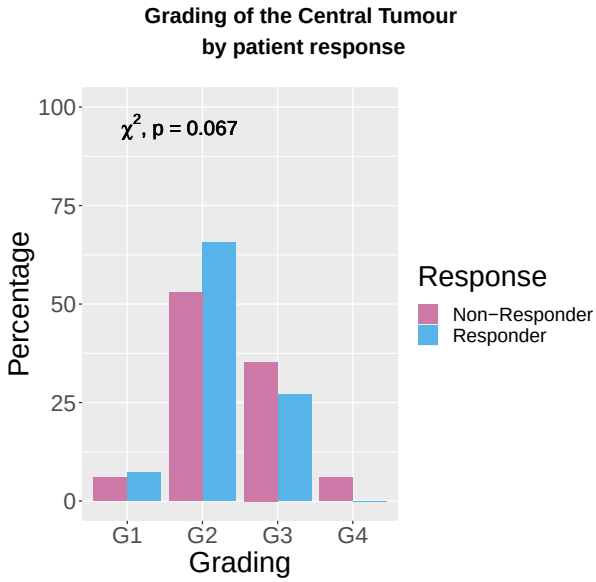


Figure 2.16: Grading of $n = 90$ patients, comparing the Responders to Non-Responders. The χ^2 test was performed at the 95% confidence level.

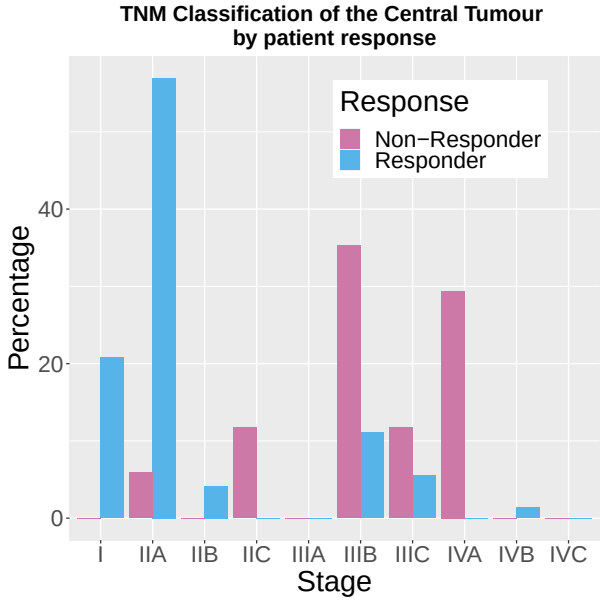


Figure 2.17: TNM Classification of the $n = 90$ patients, comparing the percentage of Responders and percentage of Non-Responders in each stage of TNM Classification.

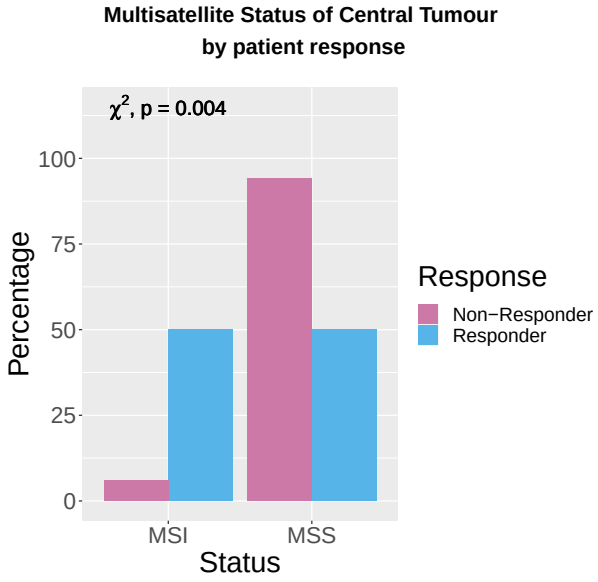


Figure 2.18: The Microsatellite Status for the $n = 90$ patients, comparing Responders and Non-Responders.

Of all the Clinical Scores performed, the Microsatellite Status is the only one to have rejected the Null Hypothesis at the 95% confidence level. As shown in Figure 2.18, almost all of the Non-Responders are considered to be classed as MSS. As such, those patients classified as MSI are almost entirely Responders. In comparison, those patients that are classified to have MSS tumours, are more than likely to be Non-Responders but cannot entirely be distinguished from the Responder patients since half of them reside in both categories. This result supports the predictive power of this genetic test of the tumour. On the other hand, as discussed in the relevant section[1], does not reveal further insight into which cell phenotypes influence this result.

Finally, we consider an Immunoscore-like percentile of the two patient groups. The results presented in Figure 2.19 show the average of the CD8 percentiles for both the CT and IM. A higher percentile represents a higher density of CD8 cells, compared to a centralised database [141], in one or both of the regions. As a consequence, we can observe that the Responder patient group has on average a higher Immunoscore percentile than that of Non-Responders. However, the significance of the Wilcoxon test shows that the Null Hypothesis, where the

[1]See Section 2.3, Subsection 2.3.3

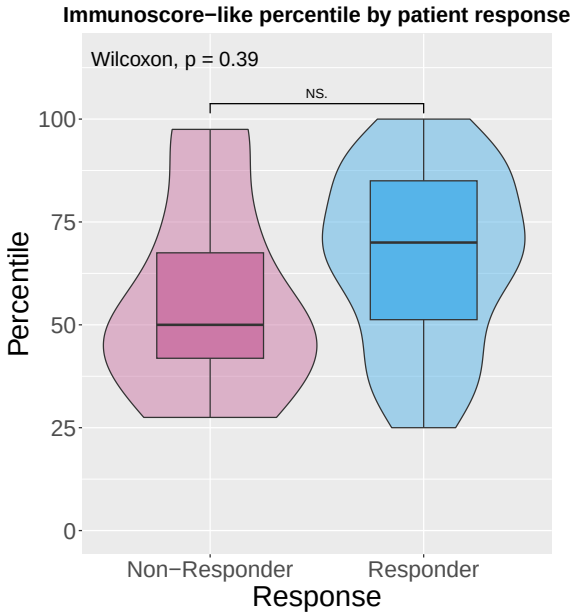


Figure 2.19: The Immunoscore-like percentile for the $n_{(\text{CT, IM})} = 43$ patients, based on only the CD8 cells, comparing Responders and Non-Responders.

distribution of both the Responders and Non-Responders is considered the same, cannot be rejected at the 5% significance level.

2.11.2 Areas, Counts and Densities

The following measurements are based on the $n_{(\text{CT, IM})} = 43$ patients that contain both information on the CT as well as the IM.

We first consider the area calculations of samples. The convex, concave, alpha shape and alpha hull are computed for the combined central tumour and invasive margin. We also compare these to the central tumour areas computed by the HALO software for the respective samples. As a consequence, we obtain an upper bound for the area from the convex hull, a lower bound from the area calculated for only the central tumour and then values understood to be closer to the true area are obtained from the concave hull calculations. The results of all the calculations can be seen in Figure 2.20. The concave hull of the samples was computed for ratio $= 0.01$. In summary, all area calculations appear to be consistent with one another and provide both upper and lower bounds on the true area and confirm our more accurate methods to be suitable for further analyses.

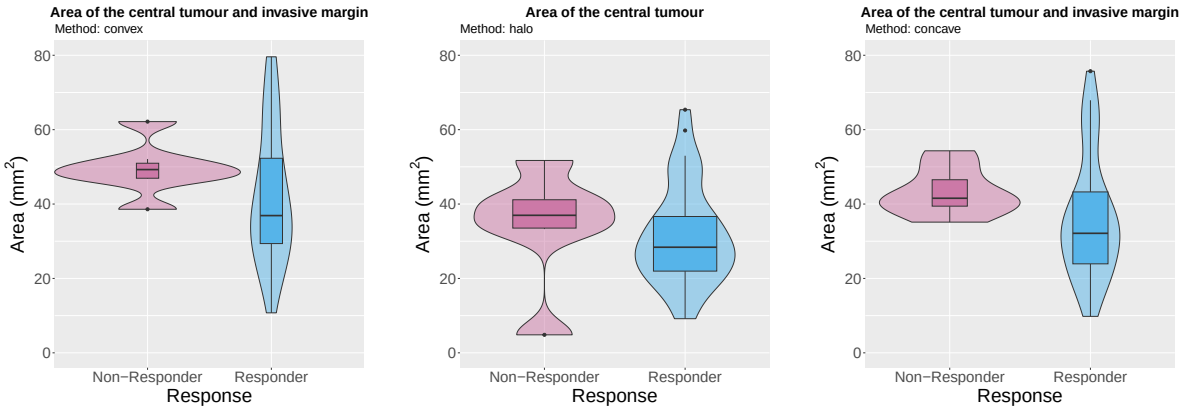


Figure 2.20: Area measurements of the (a) convex, (b) classified area of the central tumour by HALO, (c) concave (ratio = 0.01). The measurements of (a) and (b) represent the upper and lower measurements of the area, respectively, while the concave hull measurement represents a more accurate estimate.

Figure 2.21 displays the counts of the $n_{(\text{CT, IM})} = 43$ cell phenotypes according to the different decompositions shown in Table 2.4. Considering that samples vary by approximately 8 times in area at most, in contrast, the counts vary across multiple magnitudes. Not only does this variation happen across cell phenotypes but also between samples for the same cell phenotype. However, it was from checking the proportions of cell types that two samples were distinguished for having unexpected counts. Through this analysis, we were able to identify poor quality cell marker identification for samples R1079_025 and R1079_063 which have since been excluded, with their corresponding IM.

We can see in Figure 2.22 the densities of the cell phenotypes according to the different decompositions shown in Table 2.4. The counts have been normalised by the area computed by the concave hull algorithm[1]. Despite the distributions between the patient response groups looking different, the difference in medians and interquartile ranges remain similar. The only density values that appear significantly different between the two patient groups are that of the cell type CD68+ CD163+. When looking at a finer scale the difference appears to be largest for CD68+ CD163+ that is also HLA-DR+ positive. This is to some extent observed in the density values for all HLA-DR+ cells, but less clear.

[1]See Section 2.5, Subsection 2.5.1

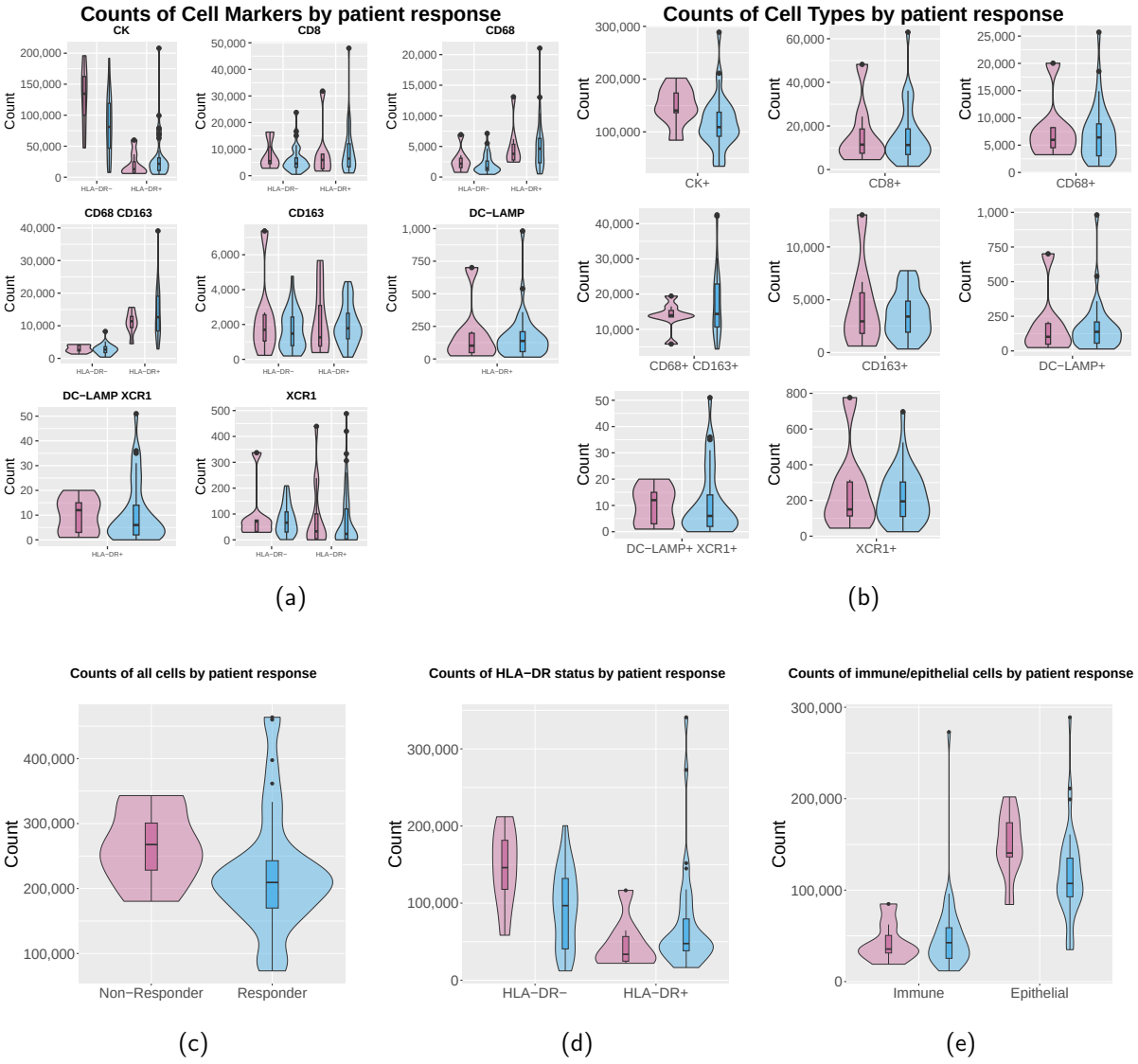


Figure 2.21: Counts of the cell phenotypes for $n_{(\text{CT, IM})} = 43$ according to the different decompositions shown in Table 2.4. (a) The series of figures that represent the counts of all the cell marker combinations that are possible in the HALO1079 data set. (b) The series of figures that represent the counts of all the cell types. (c) The total number of all cells, regardless of cell marker (including cells for which only the nuclei stain DAPI is identified). (d) The number of cells that are either HLA-DR+ or HLA-DR-. (e) The number of cells that present cell markers categorised as immune or epithelial phenotypes. Through this analysis, we were able to identify poor quality cell marker identification for sample HALO1079_025 and HALO1079_063, with their corresponding IM annotations HALO1079_026 and HALO1079_64, which have since been excluded.

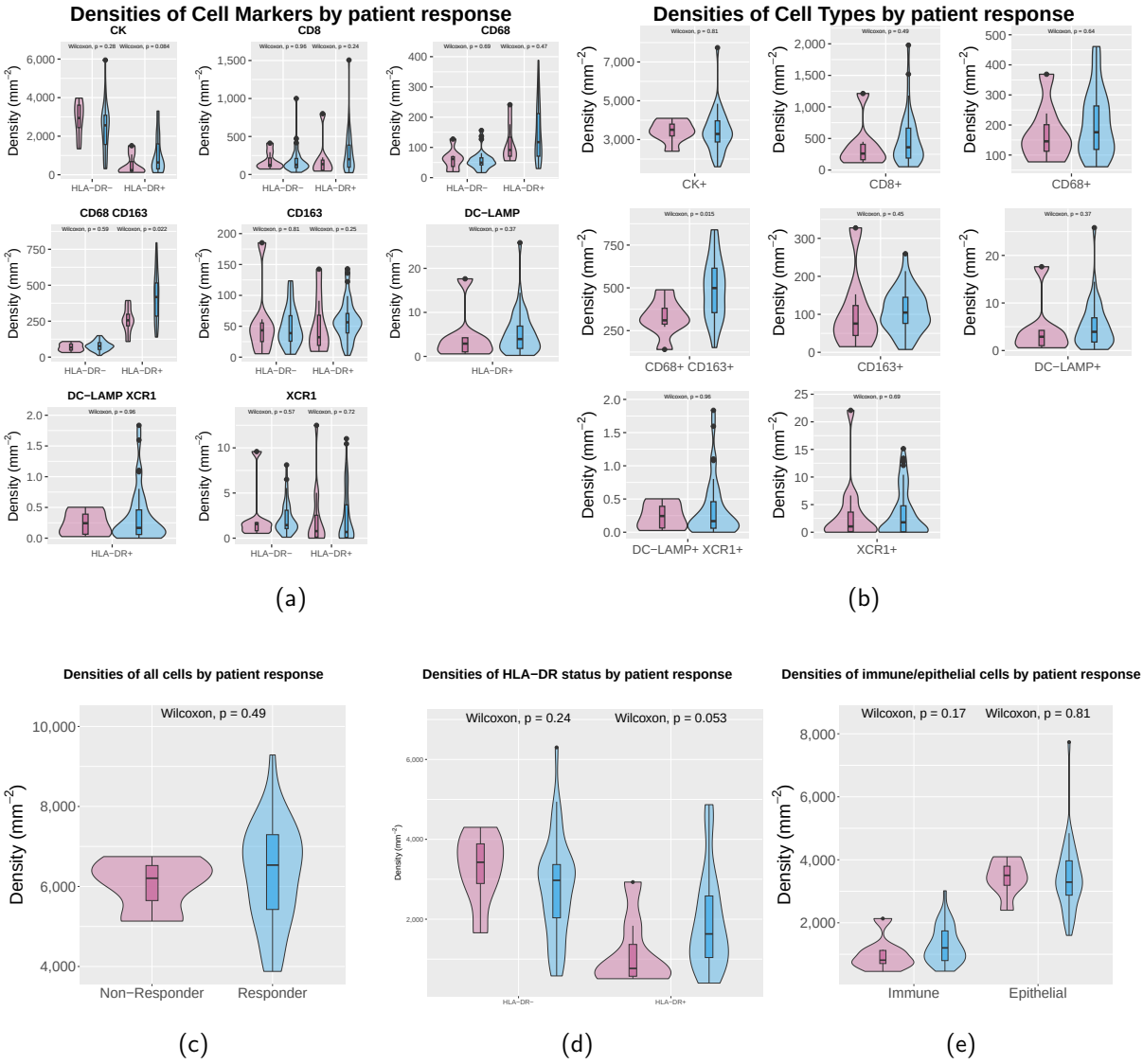


Figure 2.22: Densities of the cell phenotypes, for $n_{(\text{CT, IM})} = 43$ patients, according to the different decompositions shown in Table 2.4. The counts were normalised by the area from the concave hull. (a) The series of figures that represent the densities of all the cell marker combinations that are possible in the HALO1079 data set. (b) The series of figures represent the densities of all the cell types. (c) The density of all cells, regardless of cell marker (including cells for which only the nuclei stain DAPI is identified). (d) The density of cells that are either HLA-DR+ or HLA-DR-. (e) The density of cells that present cell markers is categorised as immune or epithelial phenotypes.

2.11.3 Heterogeneity Scores

We calculated the heterogeneity scores for all defined scores described in Appendix A (Section A.3). For some of the measures, an exponent is taken as an additional parameter. For our computations, we use the exponent for the Hill numbers as the set $\alpha \in \{n/2 \,|\, n \in \mathbb{N}, n \leq 10\}$, the same as that chosen in the established technique [166], and the Renyi exponent $\alpha \to \infty$. In addition, Rao's quadratic entropy requires a phenotype similarity matrix. Using our prior knowledge of the cell marker types, we consider a hierarchy of phenotypes whereby cells of the same marker type but different HLA-DR status are the most similar. Then, phenotypes that are categorised as immune to be considered the next most similar, and similarly for epithelial cell phenotypes. As a result of this logic, we end up with the phenotype similarity matrix presented in Appendix B (Table B.1). The inverse of the phenotype similarity matrix is then the distance metric as required by the Rao quadratic entropy[1].

Figure 2.23 presents the results of the individual heterogeneity scores used in our analysis of the HALO1079 data set. In each implementation of the heterogeneity scores, we have a Null Hypothesis that the distributions between Responders and Non-Responders are the same. We perform the Wilcoxon signed-rank test between the populations at the 5% significance level. The test in each implementation of the heterogeneity score could not reject the Null Hypothesis. When observing the distributions, this might be surprising since several of the metrics do appear to have a predominant number of samples skewed in opposite directions, e.g. Shannon and Simpson Indexes. However, samples in both groups take values in the same range as one another, making a distinction between the two difficult to separate. By implementing these indexes we show that such non-spatial metrics based on proportions of cell phenotypes have not been able to reconcile the differences between patient response. Before we conclude this, we consider a combination of heterogeneity scores to see if weightings to particular abundances might reveal differences in patient groups instead.

The Diversity Profile is a way to represent the spectrum of heterogeneity, by considering a range of exponents. Smaller exponents prioritise the differences in low-proportion phenotypes

[1] See Appendix A, Subsection A.3.6

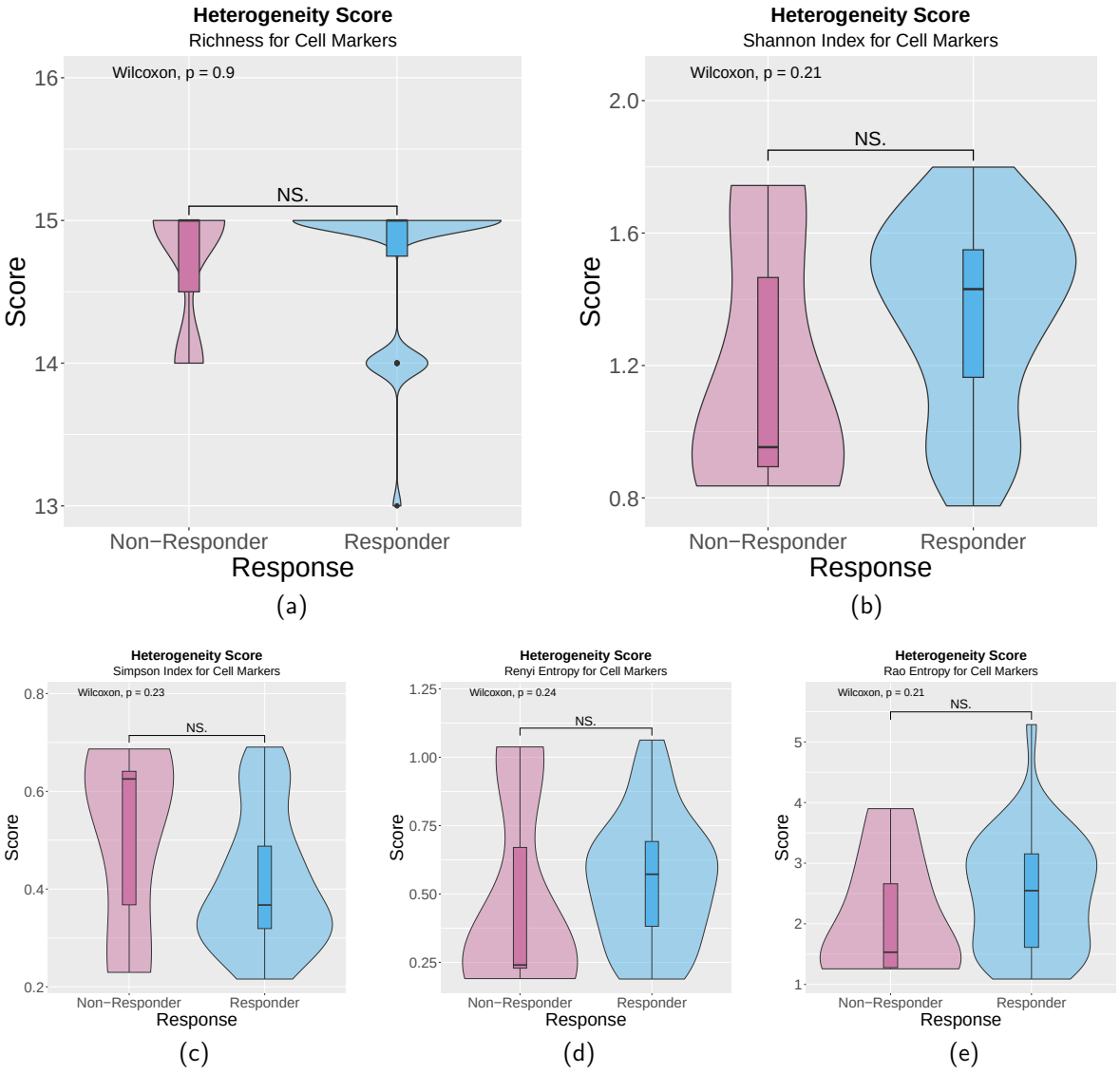


Figure 2.23: Heterogeneity scores for the $n_{(\text{CT, IM})} = 43$ patients, comparing Responders and Non-Responders. The figures represent the values calculated for the samples by the (a) Richness, (b) Shannon Index, (c) Simpson Index, (d) Renyi Entropy, (e) Rao Entropy.

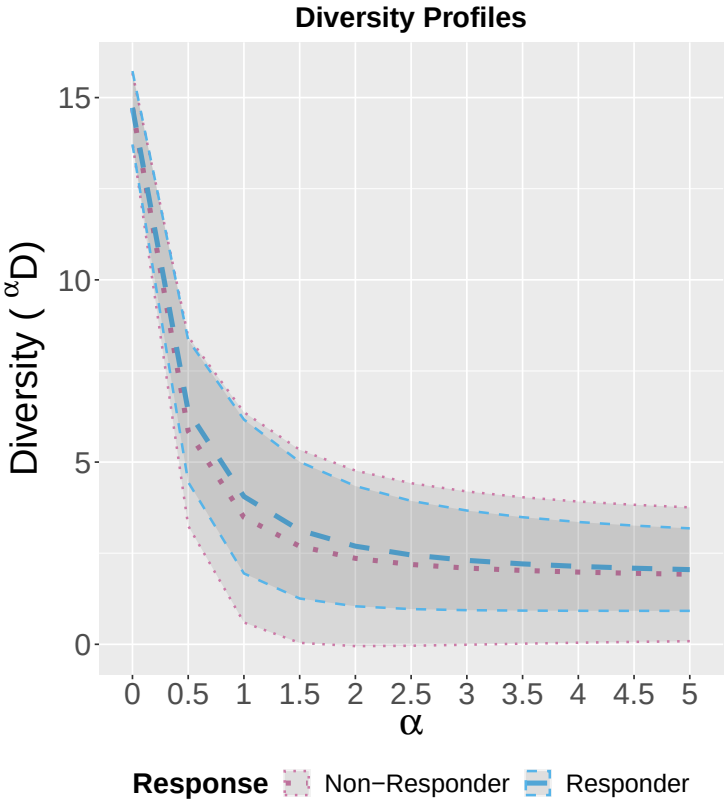


Figure 2.24: Heterogeneity scores for the $n_{(\text{CT, IM})} = 43$ patients, comparing Responders and Non-Responders. We calculate the Hill numbers for the set of exponents $\alpha \in \{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$

and larger exponents weight the metric to the heterogeneity within larger-proportion phenotypes. Figure 2.24 presents such a plot for $n_{(\text{CT, IM})} = 43$ patients. Here, we calculate the Hill number for each sample for the exponents $\alpha \in \{n/2 \,|\, n \in \mathbb{N}, n \leq 10\}$. We have presented all of the individual heterogeneity scores before overlaying the average profile for each patient group, with 95% confidence intervals. Similar to the individual scores, there is no cross-over in diversity for different weightings of heterogeneity which distinguishes between the two patient groups.

2.11.4 Compartment Interaction

The ECI curves describe the density of cell phenotypes that have been able to penetrate the CT, assuming that invasion occurs from the CT-IM boundary interface. We discuss the results for epithelial, T cells, and macrophage phenotypes. The results of the remaining phenotypes

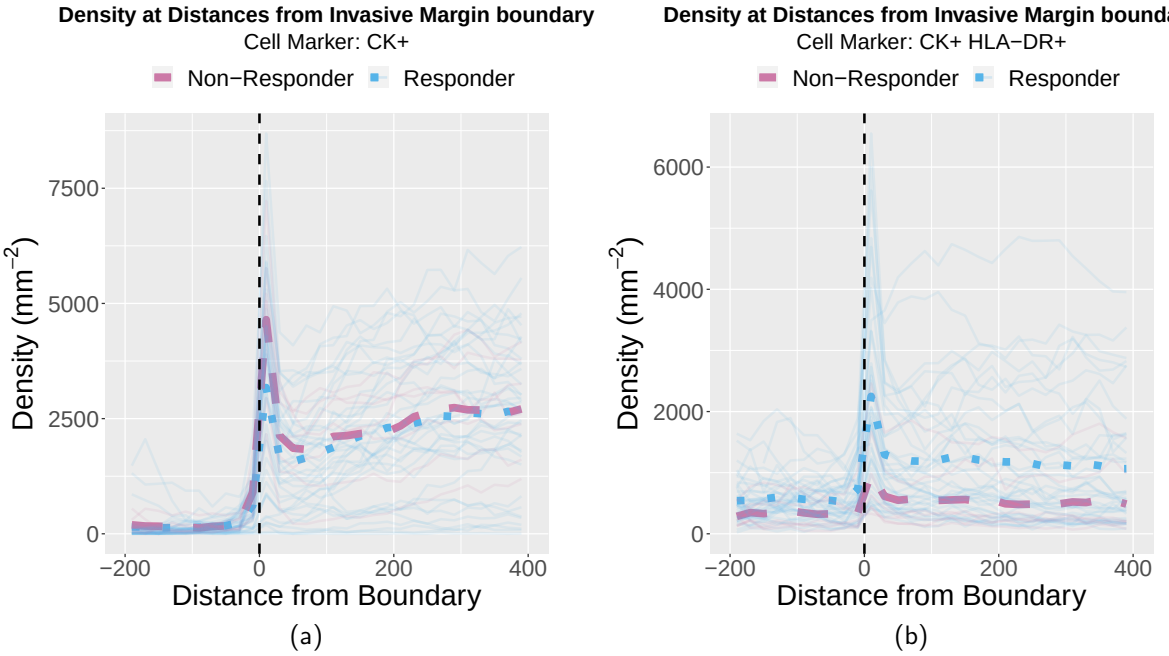


Figure 2.25: Mean ECI curves of epithelial cells for $n_{(\text{CT, IM})} = 43$ patients.

are presented in Appendix B (Figure B.1 and Figure B.2). These results have been selected since they present interesting differences between the patient response groups.

Figure 2.25 show the ECI curves for the epithelial phenotypes, CK+ and CK+ HLA-DR+, respectively. For CK+, there is little to no difference in either the density distribution of CK+ cells in the IM or in the CT. The only notable property for the ECI curve is that the density increases in gradient for distances from the boundary greater than 200. This might refer to the increasing density further into the core of the CT where the proliferation of the tumour was initiated. On the other hand, we see that there is a consistently higher density of CK+ HLA-DR+ cells for all measured distances from the CT-IM boundary interface in Responders. This is contrary to earlier results (see Figure 2.22) presenting no significant difference in observed density for CK+ HLA-DR+ between Responders and Non-Responders.

We now investigate the differences in distribution in T cells relative to the CT-IM boundary interface. Similar to the epithelial case, we can begin to see in Figure 2.26 some difference between the distribution of T cells between Responders and Non-Responders, specific to cell presenting the HLA-DR+ marker. It is in Figure 2.26b that we can see some difference in the mean density between patient response groups. The difference is more pronounced in the

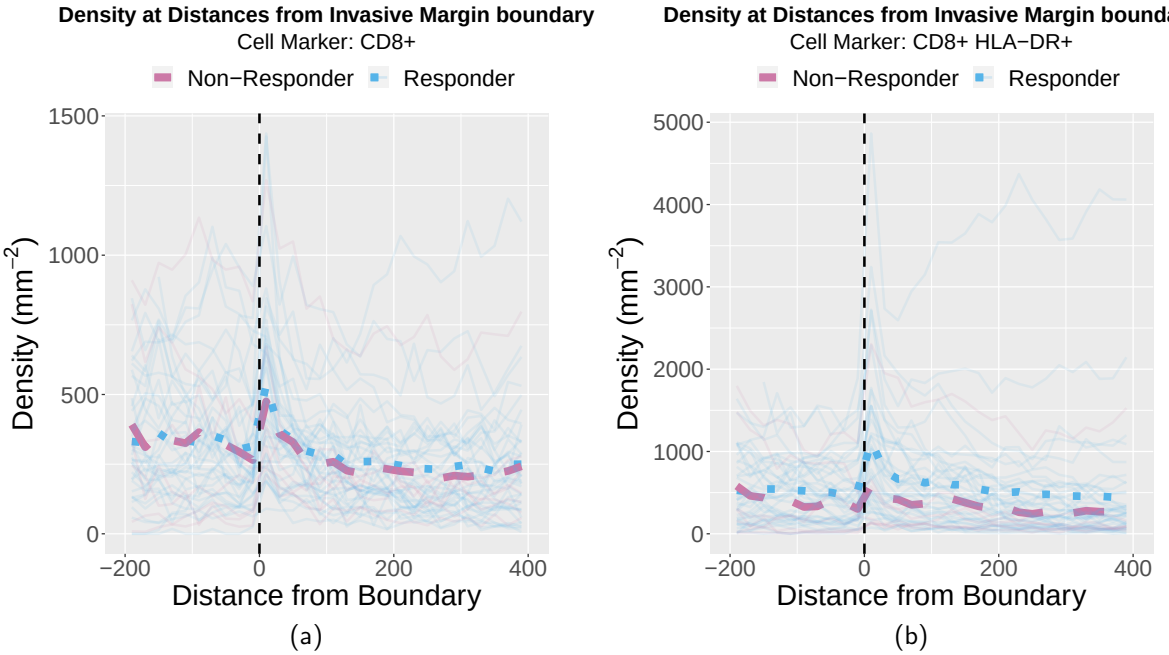


Figure 2.26: Mean ECI curves of T cells for $n_{(\text{CT, IM})} = 43$ patients.

CT than in the IM. However, looking at the individual curves of the sample, we can see that the mean might be biased by a single result in the Responder group with outstanding density values across the observed range. As a consequence, the interpretation of these results is inconclusive.

Finally, we discuss the interpretation of the ECI for macrophages, as presented in Figure 2.27. First, we consider the ECI curve for the CD68+ CD163+ cell phenotype, interpreted as M2 macrophages. It can be noted that there exist higher densities of M2 macrophages within the IM for Responders than exist within the IM for Non-Responders. The difference observed does not appear to persist into the CT, or at least for larger distances from the CT-IM boundary interface. Instead, the density of M2 macrophages decreases steadily in both patient response groups from an initial peak value at the CT-IM boundary interface. Adversely, when we consider the density distribution for CD68+ CD163+ HLA-DR+ (see Figure 2.27b) we instead detect a larger density of this sub-population of M2 macrophages in Responders than Non-Responders across all distances from the CT-IM boundary interface. This in itself is not surprising considering earlier results for the density of this cell phenotype showed significant differences between the groups. Nonetheless, the notable feature of the Responder ECI curve

is the change in gradient from the CT-IM boundary interface towards the centre of the CT. We can observe that there is an initial drop in the density value of Responders but this flattens out soon after, presenting a more persistent density of M2 macrophages with the HLA-DR+ marker. This might suggest a more persistent penetration of this specific cell phenotype than observed in other cell phenotypes. This difference is highlighted by the unremarkable similarities that can be observed for both macrophages not of M2 status when presenting HLA-DR+ or not (see Figure 2.27c and Figure 2.27d). Furthermore, when comparing similar drops in density gradients emanating from the CT-IM boundary interface into the CT, the value appears to recover to a baseline value of density within 100 microns, except for CK+ which steadily increases for larger distances. This suggests that the difference observed for M2 macrophages which present HLA-DR+ markers could be more mobile or at the very least penetrating the CT for which it is having some effect on patient outcome.

2.11.5 Spatial Correlation within Multitype Point Patterns

Since mIF images allow us to investigate tumours at the cellular resolution, we can summarise the most significant intercellular spatial correlations. The extrapolation of such distributions provides insight into the extent to which cell phenotypes are spatially correlated at different length scales. These distributions can then be analysed to compare the intercellular spatial correlation between Responders and Non-Responders.

In our analysis, we only consider the PCF as the metric for spatial correlation. This decision stems from the interpretive power of the PCF over other spatial correlation metrics for features we expect to identify in our samples[1]. Also, we do not use the PCF for quantifying spatial heterogeneity since the length-scale of such features is likely to exceed the maximum range of r that we can consider for accurate values of PCF given the area of the sample [167]. As such, we only consider $r \in [0, 100]$ for the following PCFs so that distributions between all samples remain comparable.

For most cell markers, Figure 2.28 summarises the comparison of many intercellular spatial

[1]See Section 2.7

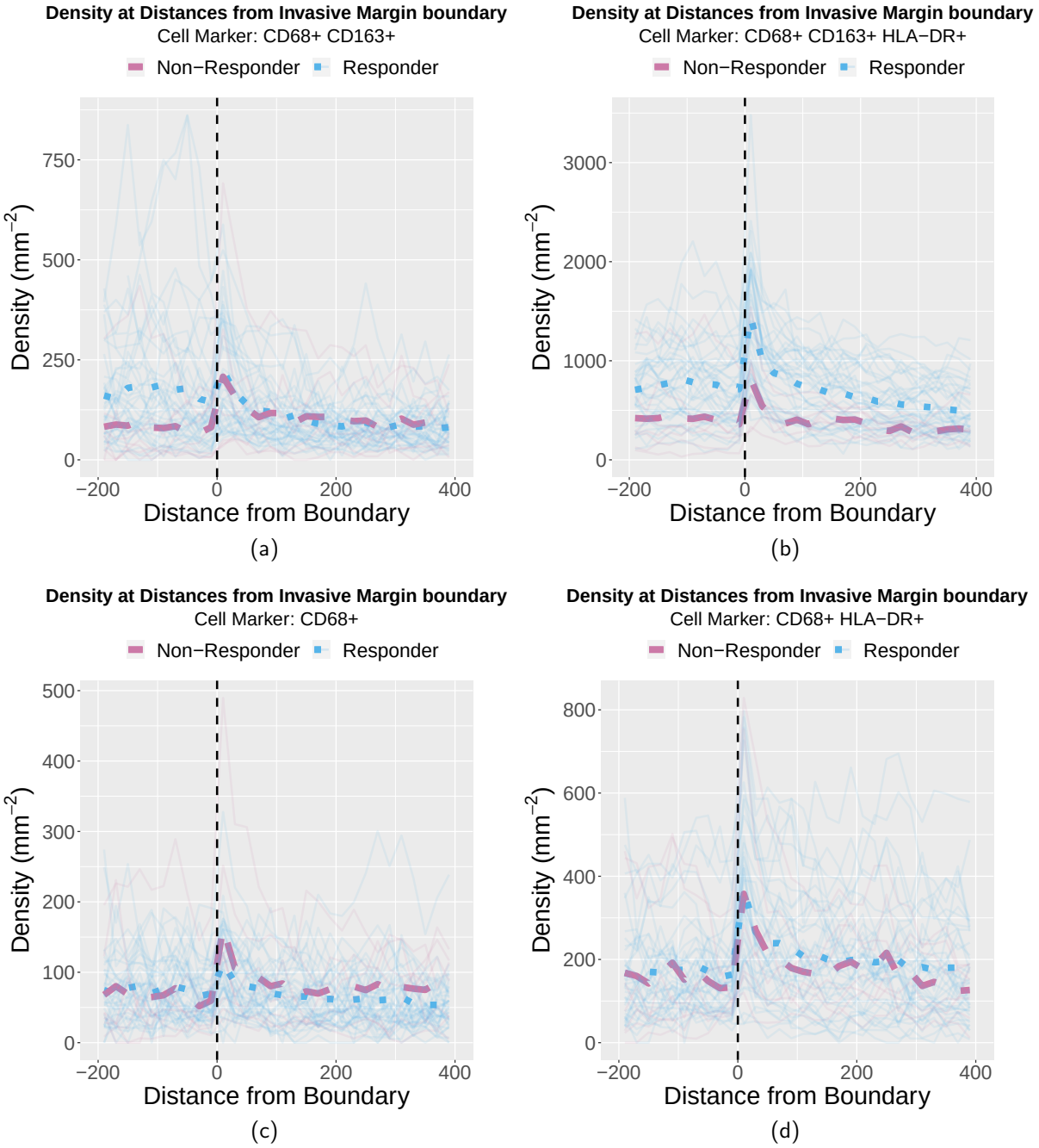


Figure 2.27: Mean ECI curves of macrophages for $n_{(\text{CT, IM})} = 43$ patients.

distribution and their difference between patient responses[1]. The cell markers not computed for PCF comparisons have been excluded since they have frequencies considered too small for accurate statistical comparisons [167, 145] (see Figure 2.21). In our analysis, this constitutes minimum frequencies of below 250 cells of a given phenotype within each samples.

For each pair of cell phenotypes, the weighted area of the PCF curve[2] and subsequent weighted area difference[3] have been calculated from Responders to Non-Responders. Therefore, high values of weighted area difference indicate that the value of the PCF is most different from the spatial Poisson point process value $g(r) = 1$ in Responders than it is in Non-Responders, for the range of $r \in [0,\ 100]$. The lowest values of weighted area difference indicate Non-Responders are more different than Responders from the spatial Poisson point process, for the range of $r \in [0, 100]$. This summary statistic highlights the distributions of the PCF which there is a larger difference in value between patient responses. As such, we will focus on PCF curves that produce the highest (Figure 2.29) and lowest (Figure 2.30) weighted area differences. Also, the largest weighted area differences for CK+ cells (Figure 2.32), and CK+ HLA-DR+ cells (Figure 2.34) will highlight the distributions with the largest differences in immune cell spatial correlation with the tumour.

The PCF with the highest weighted area difference and the lowest weighted area difference are shown in Figure 2.29 and Figure 2.30, respectively. Before examining each individually, we can compare the difference in scale between the highest values of weighted area difference and then those with the lowest weighted area difference. The highest values of weighted area difference are smaller in magnitude than those with the lowest. This can be seen as a reflection on the range of values of the PCF, with the lowest weighted area differences expressing a larger difference in PCF for Non-Responders to the expected spatial Poisson point process than Responders. This might then suggest that the most profound differences between the patient responses are better characterised by the distribution of cell phenotypes with larger effect sizes in Non-Responders than is observed in Responders.

[1] See Section 2.7, Subsection 2.7.2
[2] See Eqn. 2.3
[3] See Eqn. 2.4

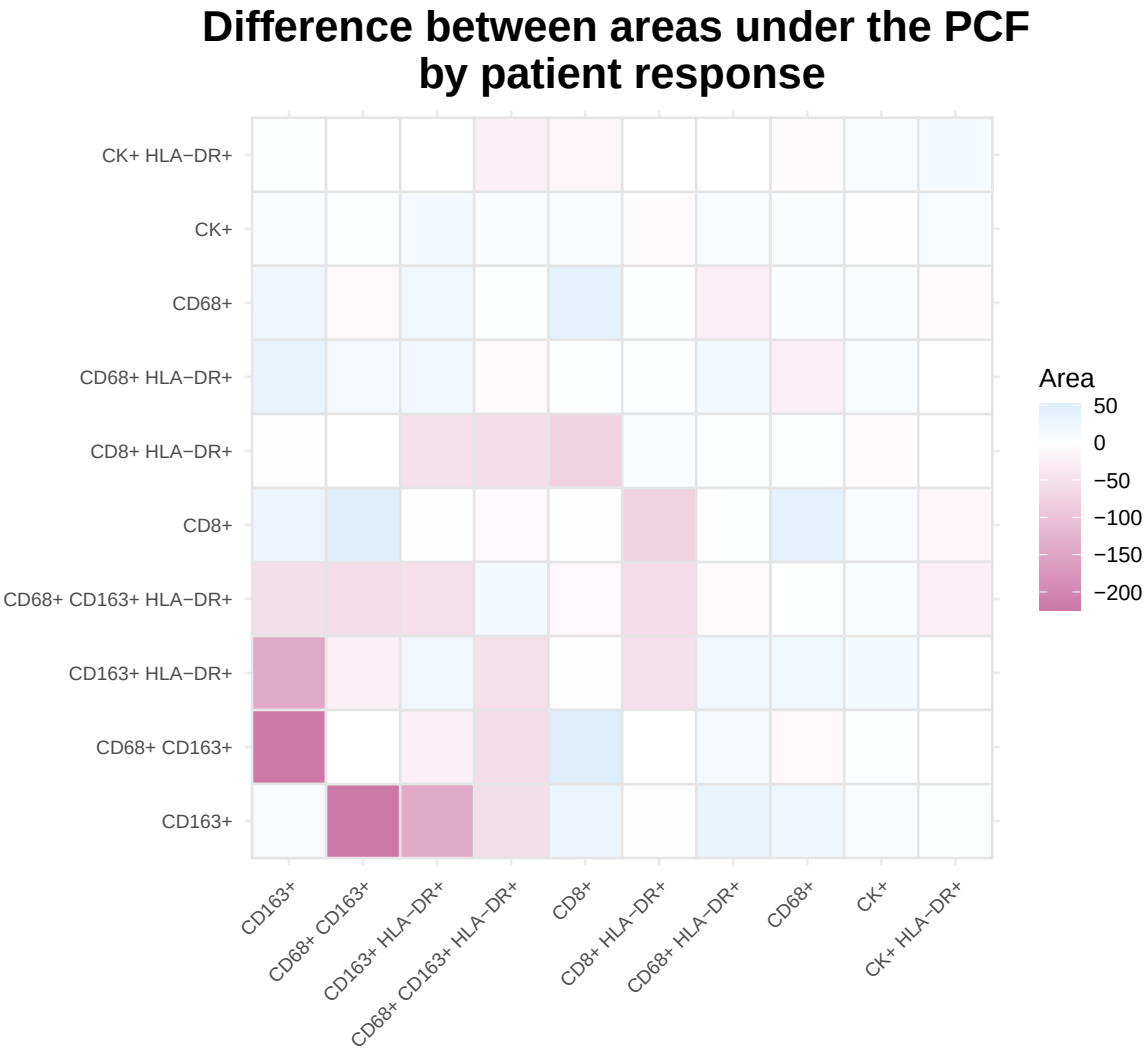


Figure 2.28: Summary of PCF values between patient responses for pairs of cell markers. Some cell markers have been excluded from PCF analysis due to their low frequencies across most samples. The weighted area difference for pairs of cell markers is symmetric along the diagonal due to the symmetric property of the PCF. Weighted area difference above zero indicates that the value of the PCF is more different from $g(r) \equiv 1$ in Responders than Non-Responders. Weighted area difference below zero indicates values of the PCF is more different from $g(r) \equiv 1$ in Non-Responders than Responders.

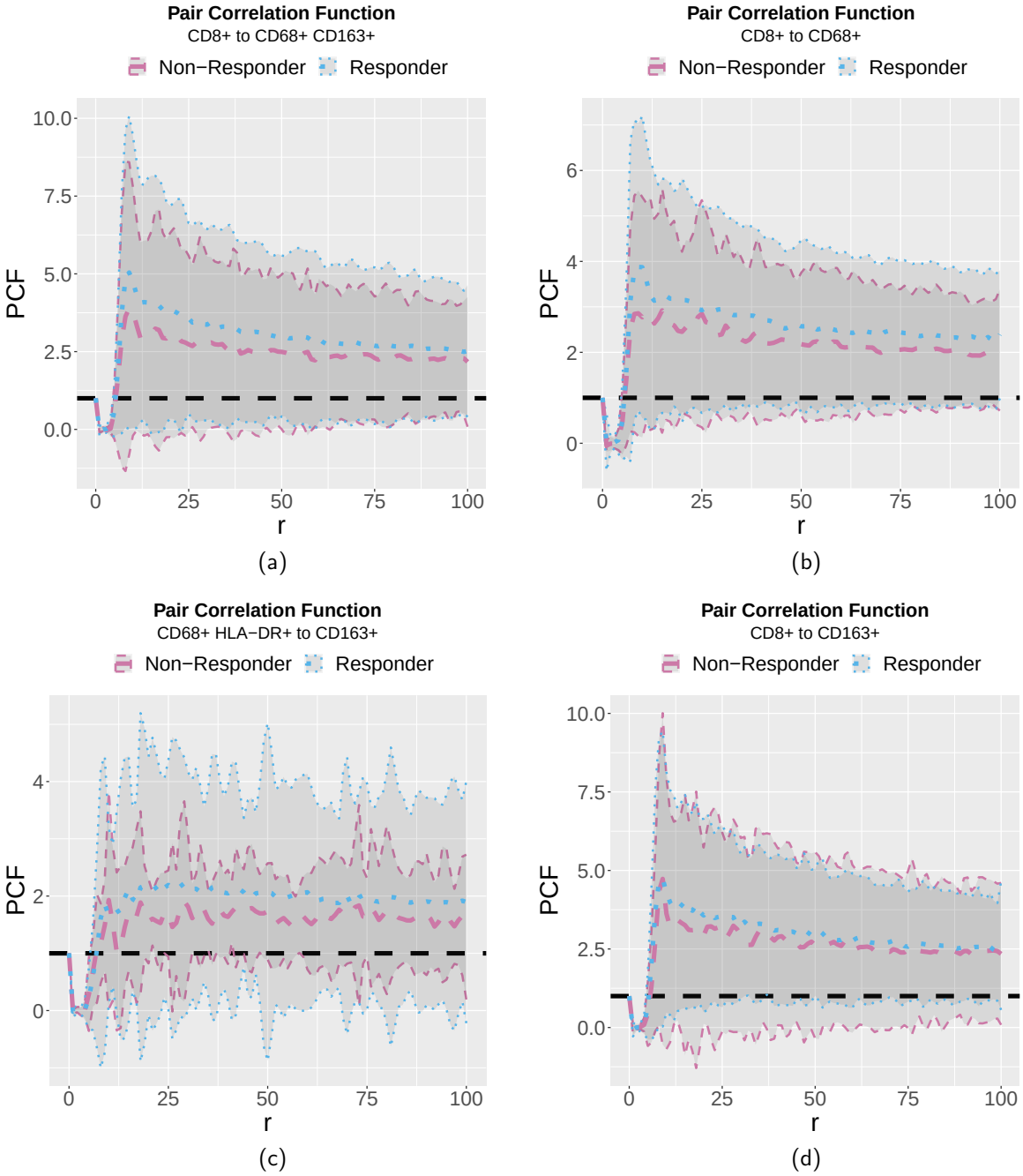


Figure 2.29: PCF plots for cell marker pairs that produced the highest weighted area differences, i.e. PCF values with larger differences to the spatial Poisson point process in Responders than Non-Responders for $r \in [0, 100]$. Here we present the pairwise comparisons for (a) CD8+ cells to CD68+ CD163+ cells (weighted area difference $= 52.3$), (b) CD8+ cells to CD68+ cells (weighted area difference $= 38.4$), (c) CD68+ HLA-DR+ cells to CD163+ cells (weighted area difference $= 32.7$), and (d) CD8+ cells to CD163+ cells (weighted area difference $= 28.2$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group.

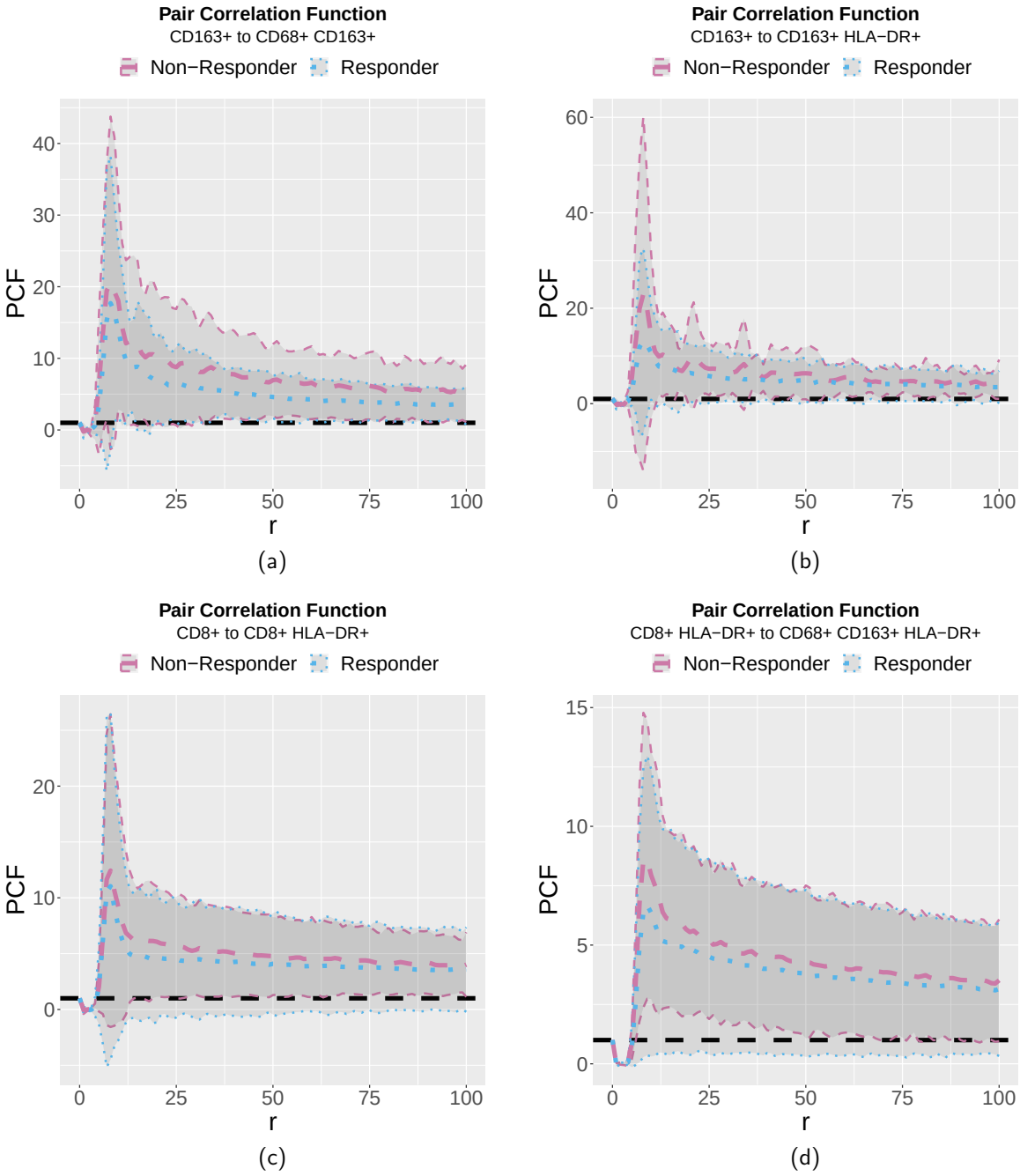


Figure 2.30: PCF plots for cell marker pairs that produced the lowest weighted area differences, i.e. PCF values with larger differences to the spatial Poisson point process in Non-Responders than Responders for $r \in [0, 100]$. Here we present the pairwise comparisons for (a) CD163+ cells to CD68+ CD163+ cells (weighted area difference $= -224.5$), (b) CD163+ cells to CD163 CD163+ HLA-DR+ cells (weighted area difference $= -144.2$, (c) CD8+ cells to CD8+ HLA-DR+ (weighted area difference $= -75.0$, and (d) CD8+ cells to CD68+ CD163+ HLA-DR+ cells (weighted area difference $= -56.3$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group.

The results in Figure 2.29, have been indicated to have the largest differences in PCF values across length scales between patient responses with a larger difference in PCF value in Responders than Non-Responders when compared to CSR. Of the highest four weighted area differences, the pairwise comparisons for the PCF contain combinations of only four cell markers: CD8+, CD68+, CD163+ and HLA-DR+. However, the combination of these markers into cell phenotypes indicates nuance in their biological mechanism. With the knowledge of this nuanced biological mechanism, the combination of these markers interprets spatial correlations that are more profound in Responders than is apparent in Non-Responders. Even though the spatial correlation is more profound on average, it is worth noting that the confidence interval at the 95% confidence level for each does contain the value of $g(r) = 1$ for all $r \in [0, 100]$ and thus the correlation is represented by a random Poisson point process instead. On the other hand, the observed mean between the two patient groups indicates that the relationship is more likely a positive spatial correlation, i.e. the two cell phenotypes are more often closer together at the observed values of r. In addition, with the knowledge from our artificial point patterns[1] we can characterise the PCFs as well-mixed high-density clusters between the two cell phenotypes, except for Figure 2.29c. That is except for the hard-shelled property of cells which produce a PCF value of zero from $0 < r < 5$ with some variation $r < 10$.

The Non-Responder distributions presented in Figure 2.30 do indicate that the average value of the PCF is likely different to the spatial Poisson point process. For Responders, we can observe that for the two most profound differences in weighted area difference (Figure 2.30a and Figure 2.30b) have confidence intervals that do not include $g(r) = 1$ for $r \in [0, 100]$. The other Responder distributions cannot exclude at the 95% confidence level that the underlying value of $g(r)$ is not equal to the spatial Poisson point process. Similar to our highest weighted area differences, all of our lowest weighted area differences indicate a positive spatial correlation between pairs of cell phenotypes. More specifically, they indicate a more profound peak in spatial correlation for $r = 10$ in all distributions. This would indicate a much higher observed value of mean cells in immediate proximity to one another. In addition, with the knowledge

[1] See Section 2.7

from our artificial point patterns[1] we can characterise the PCFs as likely to be well-mixed high-density clusters between all pairs of cell phenotype. As such, the results indicate significant clustering between all pairs of cell phenotypes in Non-Responders with a smaller effect observed in Responders.

Since we are interested in the interaction of the epithelial, the main cell phenotype of the tumour, we focus our attention on the PCF results relating the tumour to other immune cells. First, to provide context to the general distribution of the epithelial, Figure 2.31 shows the spatial correlation within CK+, between CK+ and CK+ HLA-DR+, and then within CK+ HLA-DR+ sub-populations of the epithelial. It can be observed that similar sub-populations (Figure 2.31a and Figure 2.31c) have confidence intervals which do not contain the expected value of the spatial Poisson point process, i.e. $g(r) > 1$ for $r \in [0, 100]$. In addition, we can observe that there exists a more profound peak in PCF value for radii close to the limit of the indicated hard-shell property of cells $r \approx 10$. From this result, we can describe both sub-populations as forming high-density aggregations with themselves. Similarly, Figure 2.31b presents a more pronounced peak for small r. However, variation in both samples resembles less positive spatial correlation between the sub-populations, inferred by the confidence intervals overlapping the theoretical value of CSR. This is interesting since experimental observation also indicates that HLA-DR presentation is not well-mixed in the epithelial population but instead presents in neighbourhoods of self-similar HLA-DR presentation.

The inference gained from analysing the spatial distributions of epithelial sub-populations provides context for how the immune cell population is then spatially correlated with the cells. In Figure 2.32, we have PCF plots which have the largest absolute weighted area difference between patient response groups for immune cell phenotypes to CK+ cells. Unlike previous results for spatial correlation, we can see that three of the four distributions presented indicate a negative spatial correlation between the immune cell phenotype and the epithelial. Of these, Figure 2.32a, Figure 2.32b and Figure 2.32d have confidence intervals which do not overlap $g(r) = 1$ for Responders at approximately $r < 15$, $r < 8$, and $r < 12$, respectively , and Non-

[1] See Section 2.7

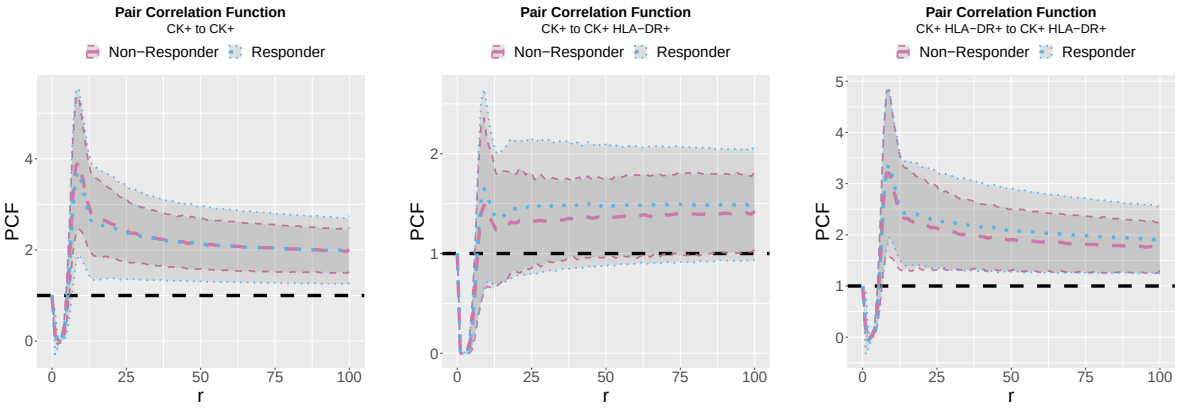


Figure 2.31: PCF plots for epithelial markers (i.e. CK+ and CK+ HLA-DR+) when comparing between patient responses. Here we present the pairwise comparisons for (a) CK+ cells to CK+ cells (weighted area difference $= -2.69$), (b) CK+ cells to CK+ HLA-DR+ cells (weighted area difference $= 11.5$), and (c) CK+ HLA-DR+ cells to CK+ HLA-DR+ cells (weighted area difference $= -15.5$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group.

Responders at approximately $r < 30$, $r < 13$, and $r < 18$, respectively. Observing this we can infer that, instead of clustering at these length scales, the distributions reflect a negative spatial correlation. In the context of our earlier observations about the distribution of CK+ within itself, this could indicate spatial inhibition between the cell types or dense clusters of CK+ cells with the respective immune cell types positively correlated but on the perimeter. To confirm this type of spatial correlation, we show the within-type comparisons in Figure 2.33. Inspecting the mean value of the PCF for each distribution, we can observe a likely positive spatial correlation within cell phenotype. However, the effect size is not large enough at the smallest of r in either Responder or Non-Responder to indicate a positive spatial correlation above $g(r) = 1$. Only for the within CD68+ HLA-DR+ comparison does the Responder confidence interval not contain the value of $g(r) = 1$ for $r \in [8, 12]$. This might suggest that there are aggregations of a few CK+ cells and CD68+ HLA-DR+ are positively spatially correlated with them. Independent of distributions for pairs of cell phenotypes, we can detect differences between the distribution of Responders and Non-Responders, specifically how some cell phenotypes in Non-Responders have a more pronounced inhibition from clusters of CK+ cells at the 95% confidence level.

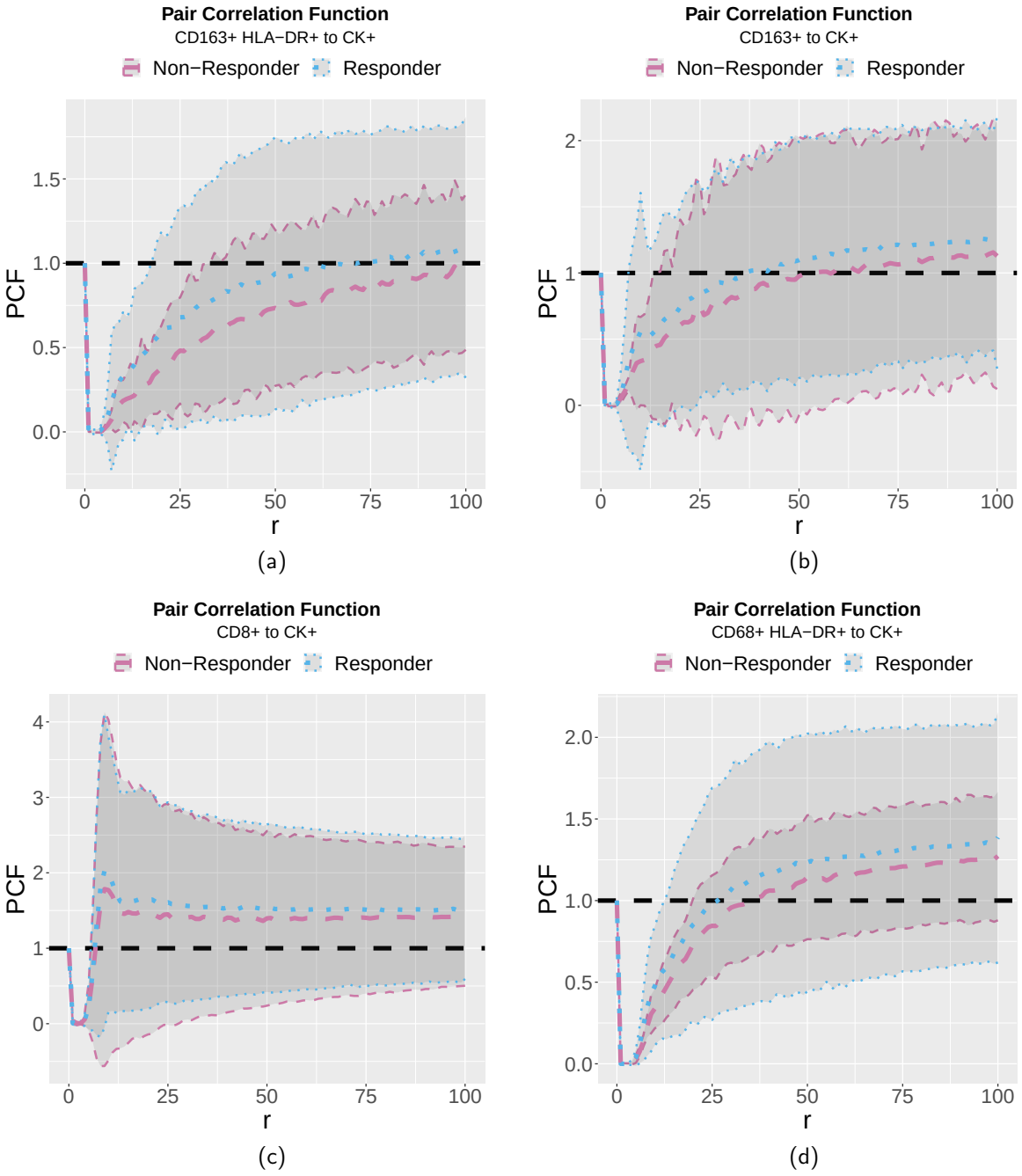


Figure 2.32: PCF plots for cell marker pairs that produced the largest absolute weighted area differences with CK+ phenotypes when comparing between patient responses. Here we present the pairwise comparison for CK+ cells to (a) CD163+ HLA-DR+ cells (weighted area difference $= 15.5$), (b) CD163+ cells (weighted area difference $= 13.0$), (c) CD8+ cells (weighted area difference $= 12.4$), and (d) CD68+ HLA-DR+ cells (weighted area difference $= 10.2$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group.

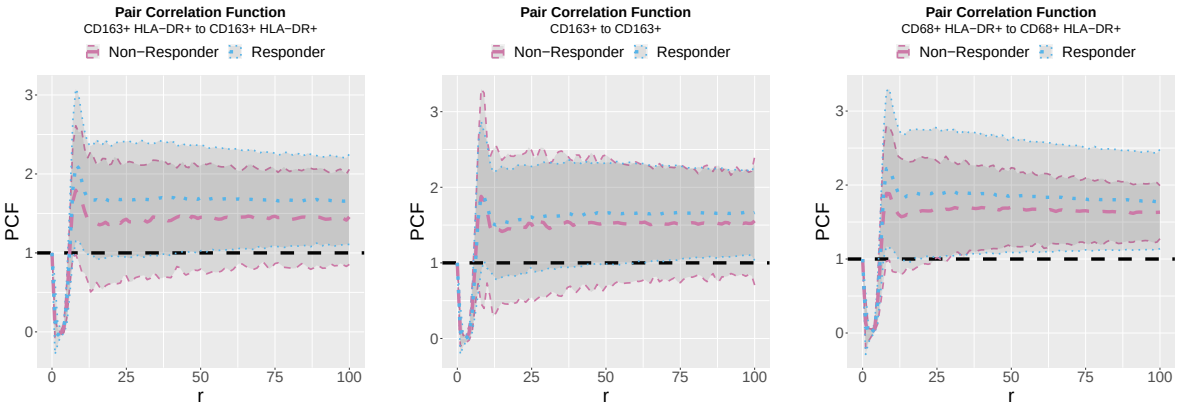


Figure 2.33: PCF plots for within cell phenotype comparison to confirm the characteristic distribution of CK+ cells to immune cells of interest from Figure 2.31. Here we present the pairwise comparison within (a) CD163 HLA-DR+ cells (weighted area difference $= 23.2$), (b) CD163+ cells (weighted area difference $= 11.1$), and (c) CD68+ HLA-DR+ cells (weighted area difference $= 18.3$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group.

We show in Figure 2.34 the PCF plots with the largest absolute weighted area difference between patient response groups for immune cell phenotypes with the CK+ HLA-DR+ epithelial sub-population. For CD68+ CD163+ HLA-DR+ cells, we have a positive spatial correlation and a more profound mean PCF value for $r \approx 10$. This, and the steady decline in PCF as r increases, suggest a well-mixed cluster of both cell phenotypes which is more pronounced in Non-Responders than in Responders. However, the specific peak has a confidence interval at the 95% confidence level which contains $g(r) = 1$ and so cannot reject the possibility that the distribution could be explained by a value which is equal to CSR. On the other hand, the distribution between CD8+ and CK+ HLA-DR+ cells does not present a similar profound peak for small r. Instead, the PCF curve represents gradual spatial clustering for $r \in [10, 25]$ to a constant positive spatial clustering, which tends to a value of 2 in both the Responder and Non-Responder samples. Although an interesting trend, only length scales greater than $r = 40$ of the Non-Responders have a confidence interval which does not contain the value for CSR. As a consequence, neither PCF confirms a significant characteristic correlation between the cell phenotypes. However, some indication for a larger effect size in Non-Responders than Responders is present between CK+ HLA-DR+ and CD68+ CD163+ HLA-DR+.

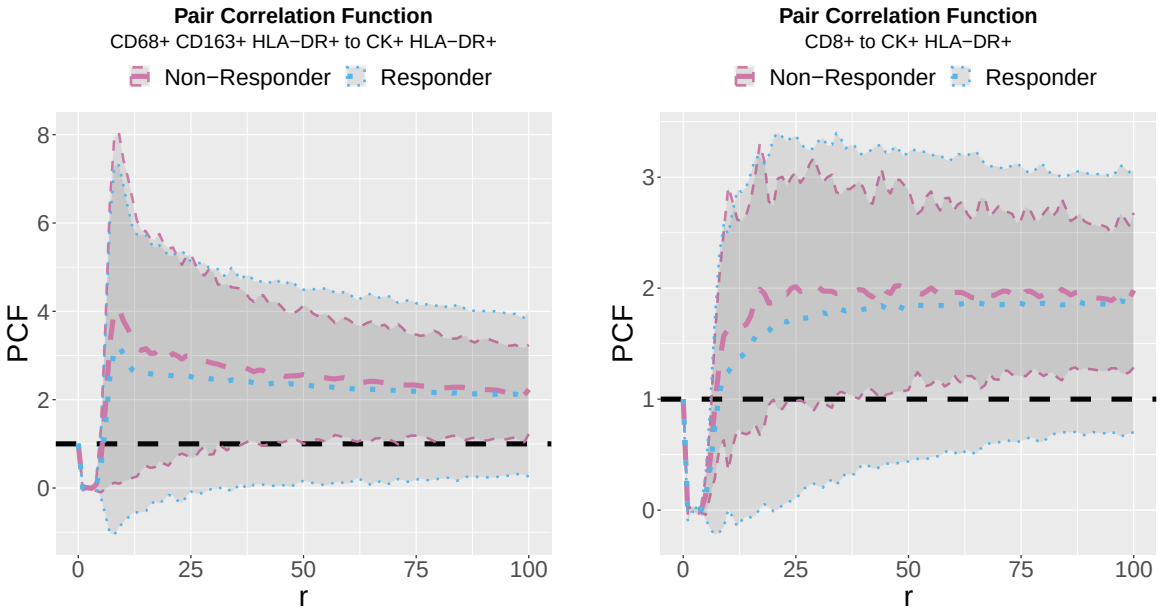


Figure 2.34: PCF plots for cell marker pairs that produced the largest absolute weighted area differences with CK+ HLA-DR+ phenotypes when comparing between patient responses. Here we present the pairwise comparisons for CK+ HLA-DR+ cells to (a) CD68+ CD163+ HLA-DR+ cells (weighted area difference $= -25.2$, and (b) CD8+ cells (weighted area difference $= -14.4$). The envelope for the respective mean PCF represents the confidence interval at the 95% confidence level for the sample of each patient response group.

2.11.6 Cellular Neighbourhoods

The Neighbourhood Analysis Score produces a single value to quantify the proportion of neighbours of a cell phenotype which surround the average cell of another phenotype. To calculate the Neighbourhood Analysis Score, one must choose an appropriate graph construction upon which to calculate the score upon. For our analysis, we have chosen the Delaunay triangulation with the maximum distance for any edge set to 35 μm. The reason for this graph construction is discussed in detail with references to other successful studies[1]. In Figure 2.37, we present the Neighbourhood Analysis Scores which have proven statistically significant, at the 95% confidence level, to reject the Null Hypothesis that differing patient responses hold similar distributions.

In Figure 2.35, we are comparing the proportion of macrophage neighbours to CK+ cells (a.k.a. epithelial cells with no HLA-DR+ marker). We focus on this set of pairs of cell phenotypes since the Null Hypothesis, at the 95% confidence level, is rejected when comparing patient responses. Specifically, the proportion of CD68+ CD163+ HLA-DR+ cell neighbours to CK+ cells. This statistically significant result is presented in Figure 2.35a. We can observe that the median Neighbourhood Analysis Score for Responders is higher than the median Neighbourhood Analysis Score for Non-Responders. Furthermore, the median of the Responders is higher than all the scores of the Non-Responders. As a result, the distribution between the two patient responses has shown to reject the Null Hypothesis, at the 95% confidence level, and so the scores are deemed to represent different distributions. This first result may not be surprising after considering the overall density of CD68+ CD163+ HLA-DR+ cells between patient responses (see Figure 2.22), however, what is now quantified is that the difference in cell phenotype density is not only higher in Responders but is also distributed among more of the epithelial sub-population. This result has provided insight beyond simply counting the cell phenotype. This result provides evidence for spatial interaction and, thus, potential biological mechanisms previously unmeasured between the phenotypes. This outcome is enhanced by the absence of significant differences between patient responses for all other macrophage

[1] See Section 2.8, Subsection 2.8.4

phenotypes with the epithelial, as referenced in Figures 2.35b, 2.35c, and 2.35d.

The previous result observed a biological mechanism occurring between the innate immune response of a patient and the tumourous epithelial. Figure 2.36 represents a relationship within the adaptive immune cell population, i.e. between T cells. The Neighbourhood Analysis Score identified a difference in distribution between patient response groups, at the 95% confidence level, for the proportion of neighbours of CD8+ HLA-DR+ cells for the average CD8+ cell. The values for Responders are observed to have a lower median value than the median value for Non-Responders. Moreover, the upper quartile for Responders is almost equivalent to the lower quartile for Non-Responders. Additionally, this relationship is not observed in the reflection of pairwise comparisons when looking at the CD8+ HLA-DR+ neighbours of CD8+. Instead, the Null Hypothesis cannot be rejected at the 95% confidence level. Therefore, the difference between patient response groups in the Neighbourhood Analysis Score is compelling evidence that the presence of CD8+ in the neighbourhood of CD8+ HLA-DR+ has a relationship with patient response.

Finally, we discuss the remaining Neighbourhood Analysis Scores for which the Null Hypothesis has been rejected at the 95% confidence level when comparing patient responses. Of these Neighbourhood Analysis Scores, the pair of cell phenotypes involved including CD163 in all of them, with one comparing CD163+ HLA-DR+ cells (see Figure 2.37e). Despite showing a statistical significance between patient responses, the interpretation of this marker interaction may not be a cell-specific presentation. When describing the phenotype in Table 2.1, it is suggested that this marker is referred to as the scavenger receptor on macrophages, as described by Onofre et al. [168]. However, since CD163+ is identified as a cell marker presented on primarily macrophages followed by dendritic cells, both markers included in the panel, it raises the question about the validity of a cell phenotype consisting of only CD163+. One explanation is that we do not monitor neutrophils, the last type of monocyte on which CD163+ could present. Yet, the interpretation of the literature suggests that the marker is almost exclusively presented on macrophages with fewer examples in dendritic cells [168]. Another explanation is that CD163+ markers are free-form in the tissue and as a conse-

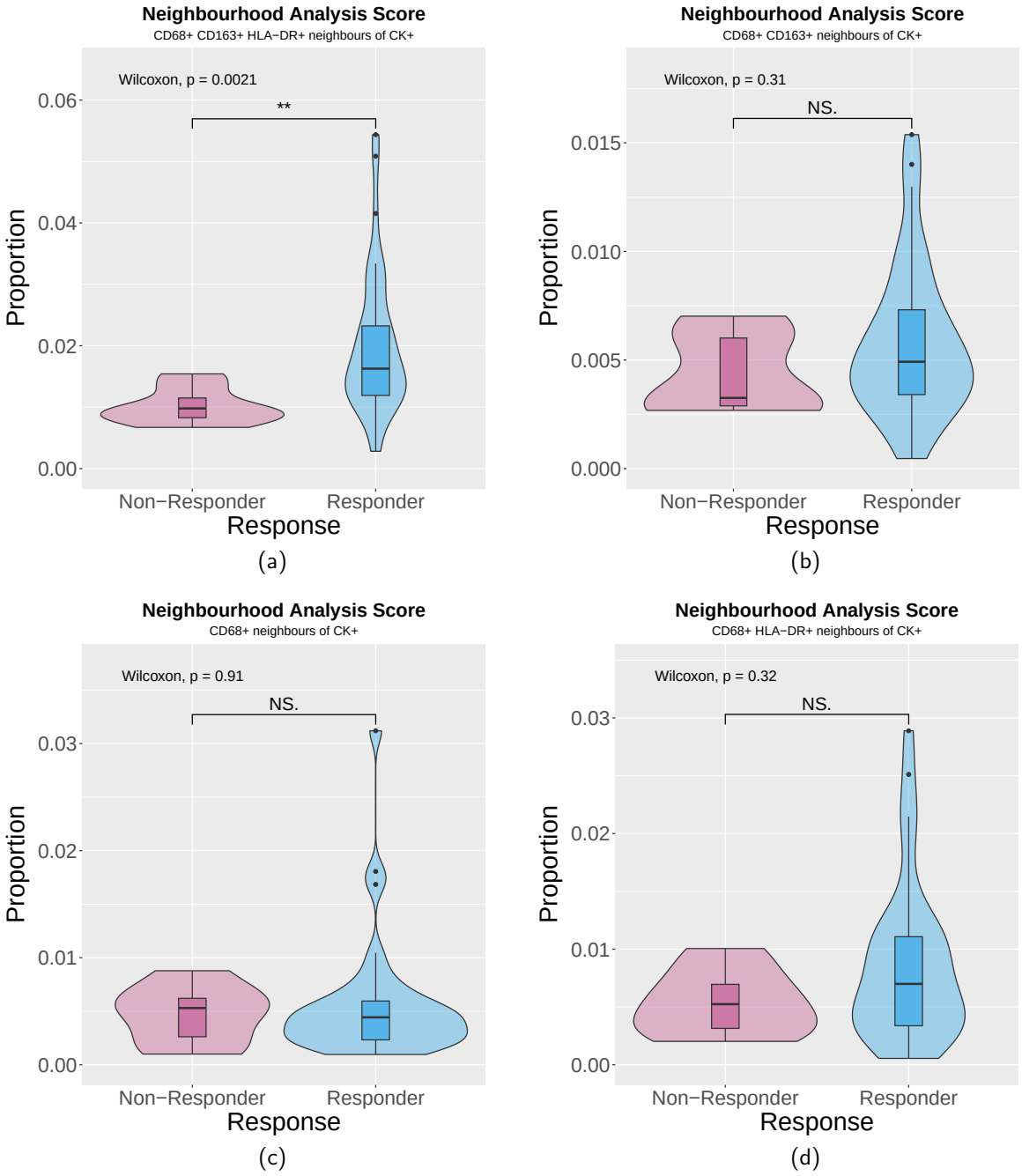


Figure 2.35: Neighbourhood Analysis Scores of macrophage neighbours to CK+ cells comparing patient response for the $n_{(\text{CT, IM})} = 43$ patients.

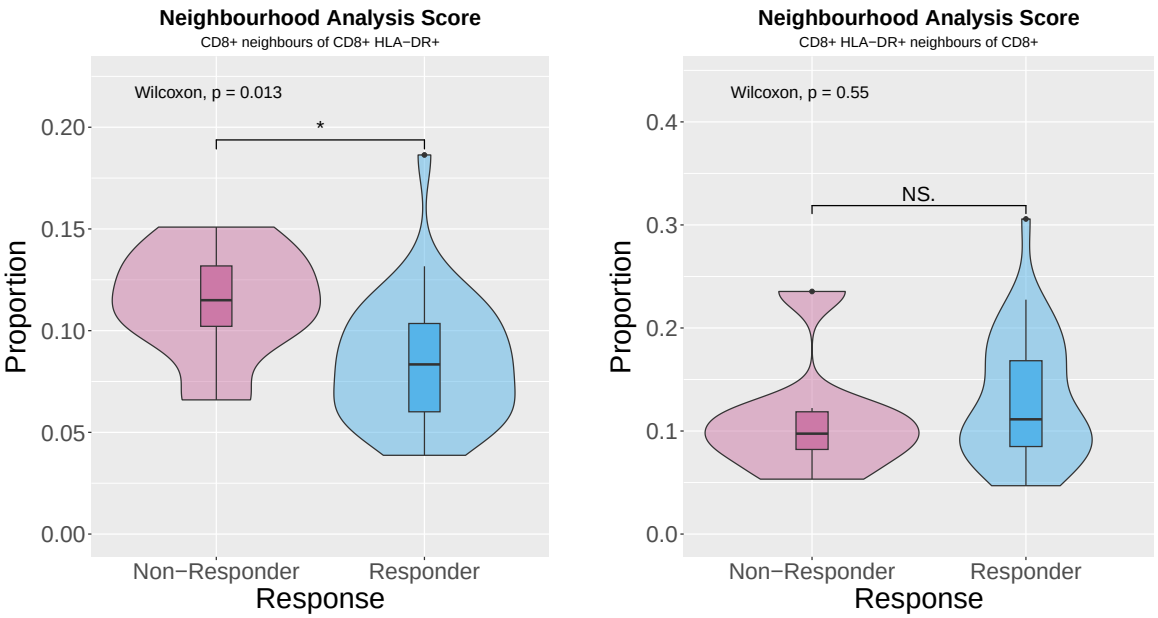


Figure 2.36: Neighbourhood Analysis Scores for between sub-populations of T cells comparing patient response for the $n_{(\text{CT, IM})} = 43$ patients.

quence co-localise with segmented cells with phenotypes which we do not monitor in the panel. CD163+ has been known to exist in soluble form within the plasma [169, 170, 171]. In this soluble form, CD163+ is known for the clearance of haemoglobin and participation as an anti-inflammation agent [168]. Either way, we can interpret the presence of CD163+ in the neighbourhood of any corresponding cell phenotype as an anti-inflammatory response located near that phenotype. The difference between patient responses could then indicate the importance of this mechanism in the survival of patients.

2.12 Summary and Outlook

2.12.1 Summary

We have analysed surgical biopsies from a cohort of CRC patients. In doing so, we have identified a series of measures which characterise the response in patients to continuing treatment. Each measure displays a different interpretation of the cellular mechanisms and tissue retal-

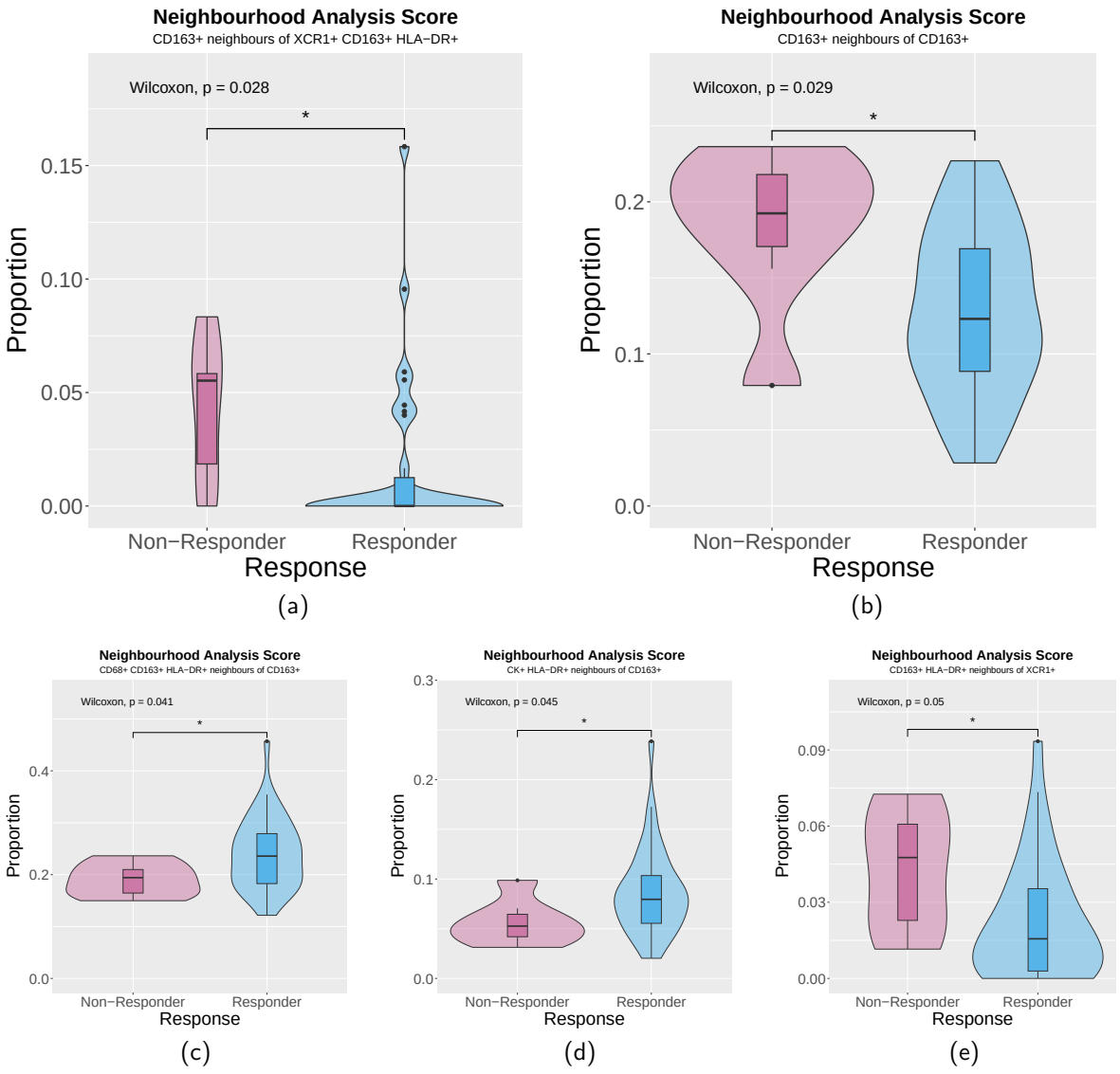


Figure 2.37: Neighbourhood Analysis Scores of the remaining statistically significant differences, at the 95% confidence level, between patient responses for the $n_{(\text{CT, IM})} = 43$ patients.

iation. Aside from implementing measures from the literature, we introduce new measures, such as the ECI curves, and extensions to metrics already available, such as the weighted area difference to distinguish interesting cell phenotype comparisons between PCF curves.

The results of the HALO1079 data set have revealed several intriguing relationships to inspire further study. The cell phenotype which has been present in the most significant results for the majority of metrics has been CD68+ CD163+ HLA-DR+, indicative of the HLA-DR positive M2 macrophages. This cell phenotype has shown to not only be more present in Responder samples of CRC but characterised in several interesting spatial distributions with the tumourous tissue. Beginning with the ECI curves, of all similar macrophage cell phenotypes, only CD68+ CD163+ HLA-DR+ cells showed persistent infiltration into the CT for Responders, despite a similar cell marker showing a comparable high density of cell phenotype in the IM. Then, when analysing the spatial correlation of cell phenotypes, we observe a PCF distribution which reflects a more prominent spatial mixing of CD68+ CD163+ HLA-DR+ cells with HLA-DR positive epithelial. This is in addition to better spatial mixing with CD8+ HLA-DR+ cells in Non-Responders than Responders. Finally, when considering only interactions with immediate neighbours, CD68+ CD163+ HLA-DR+ cells had a higher likelihood of being in the presence of HLA-DR negative epithelial cells in Responders than in Non-Responders. The summation of all these results provides compelling evidence for a biological mechanism involving the interaction of CD68+ CD163+ HLA-DR+ with the TME as an avenue of further investigation to improve the recovery of CRC patients.

2.12.2 Future work

As described in the summation of the results, CD68+ CD163+ HLA-DR+ cells require further investigation for the response of CRC patients. Not only do these results provide evidence for further biological experimentation but can provide a crucial link between clinical imaging and *in silico* modelling. Interest in M2 macrophages has already been monitored for *in silico* study, with an example of agent-based modelling [172] recently published.

In addition to *in silico* modelling, the study also suggests an extension to a continuum

of cell phenotyping specific for M1 and M2 macrophages. This would envisage PCF curves to be transformed into landscapes where a whole continuum of cell phenotype is compared for spatial correlation. We consider this approach not only intriguing for our analysis of macrophage phenotypes but also could be used when comparing the presentation of HLA-DR activation. Since HLA-DR activation can be considered as a continuum of activation in itself, this approach could prove useful. Although this approach is currently applied to spatial correlation, it can be perceived how a similar interpretation of a continuum in phenotype could also apply to another metric: the Neighbourhood Analysis Score. Instead of a single value characterising the average proportion of neighbours of a single discrete phenotype, the probability mass function of a range of continuum-like markers could be recorded.

One extension for the analysis performed here is for the heterogeneity scores calculated in this study. In their current format, the scores compute a whole slide image entropy for the proportion of different discrete cell phenotypes. An extension to these measures would be to consider sampling smaller areas of the tumour to characterise the spatial distribution of the phenotype heterogeneity. An approach is described in the ATHENA library [158], however, validation for within cohort variation and suitable comparison were not resolved in this study.

For the computation of ECI, the current comparison between patient response groups is through the mean value of the density for respective groups. However, comparing the individual samples to the equivalent spatial Poisson point process would benefit the interpretation of the size of the effect between patient responses. Then, instead of averaging between observed densities for each sample, a comparison can be made between samples that have variable global density for the cell phenotype but are locally distinct at the CT-IM boundary interface. Understanding the implications of this metric might provide a useful metric to detect the spatial correlation of a cell phenotype in this region.

In this investigation, a choice of graph construction was implemented for the optimisation of computational cost, which may compromise biological intuition. Therefore it would be prudent to validate in our own choice of graph construction. So, as the value in computational power continues to improve and resources for implementation become available, validation of each of

the graph constructions discussed would become practical. Not only this, but this avenue of investigation would provide evidence to the field which does not currently challenge this step to calculate subsequent metrics.

2.12.3 Conclusion

In conclusion, we have applied and extended metrics most useful in the application of mIF images to decipher multiple phenotypic intercellular interactions. We have highlighted the spatial interaction of one cellular phenotype in the survival of CRC, that would have been undetected in clinical scores and non-spatial metrics. The analysis performed in this study, and subsequent extensions discussed herein, allow for further hypothesis generation. Moreover, the work carried out here provides a link between the monitoring of spatial biological data and future *in silico* models. The knowledge provided by our analysis has shown the importance of exploratory analysis of multiple cell phenotypes and their spatial characterisation for the monitoring of patient survival.

CHAPTER 3

IDENTIFYING HIGHER-ORDER STRUCTURES IN 2-DIMENSIONAL TARGETED IMMUNOFLUORESCENCE IMAGING

3.1 Motivation

Quantitative pathology has previously relied on human experience to identify patterns in complex biological images. Developments in microscopy, computer science and mathematical techniques open up a new possibility for automating this work at scale. Whereas manual interpretation for hundreds of images can become tiresome, increasing the likelihood of errors, a computer can repeat the same predictable algorithm and fast. Furthermore, automation can help to quantify new patterns that are tricky by hand. This field is commonly referred to as *high-throughput microscopy*. This term encompasses the acquisition and processing of a large volume of data by automating sample preparation and data analysis techniques [173]. Combinations of advanced techniques come together to form what's called a *workflow*.

Workflows are a series of tools that have been verified individually to perform a specific task which then in combination identify the desired patterns in an image [174]. This then can be automated at scale for similar input data. As more workflows are built, improved upon and verified with new data sets, determining which is the best has become a problem in itself [175]. Carpenter et al. [176] list some criteria that such research software should meet. The criteria can be summarised into five key attributes: user-friendly, modular, developer-friendly, validated

and interoperable. As a consequence, a successfully designed workflow should consider all of these attributes.

In this chapter, we will build a workflow for the task of identifying a specific type of multi-cellular structure from human biopsies. The biopsies have been taken from patients suffering from Sjögren's syndrome. This syndrome is an autoimmune disease that affects parts of the body that produce fluids, like tears and saliva. It usually starts in people aged 40 to 60 and is much more common in women than men [177]. Secondary forms are associated with rheumatoid arthritis and primary biliary cholangitis (i.e. an autoimmune disease of the liver). In this data set, the presence of multi-cellular structures called Tertiary/Ectopic Lymphoid Structures/Organs (TLS) [178, 107] are being quantified and compared to prognosis. These higher-order structures have gained special attention due to prognostic value in several cancers [102]. Therefore, understanding their influence in a disease caused by the immune system is crucial in understanding the development of Sjögren's syndrome. In addition, TLS is currently manually identified with no automated workflows openly available. In this chapter, we develop a workflow to identify TLSs quantified by geometric and topological properties usually determined subjectively.

During the construction of our workflow, we identified several techniques suitable for quantifying properties related to TLS. To decide the most appropriate algorithms, we interpret the theory behind them and validate them on artificial data sets. In addition, we develop a new technique to quantify the density gradient of cell phenotypes from the boundary of a high-density aggregation. This proves as an accurate addition to distinguish TLS from other types of high-density aggregation. In addition, we provide a novel application of Persistent Homology (PH) to identify homological features, i.e. loops, to ascertain whether a surrounding cell type is radially symmetric around the border of the TLS. The combination of the validated and novel techniques proves to be effective in the application of TLS, with statistics evaluating performance on artificial data as well as objective validation from supplementary cell markers for TLS.

First, the chapter will summarise the biological background necessary to understand these

images. Then, we outline the approach used to process the images into data structures that we can input into our workflow, reflecting the criteria of good workflow practices. Next, we illustrate the methods employed in our analysis to objectively identify TLS using geometric and topological algorithms. We then validate each algorithm on a series of benchmarks, before validating the performance on an artificial data set. Afterwards, we test the performance of the workflow on biopsy images from Sjögren's syndrome patients and discuss their performance. Finally, we consider follow-on methods which analyse the results of our workflow.

3.2 Biological Context

3.2.1 Ectopic / Tertiary Lymphoid Structures / Organs

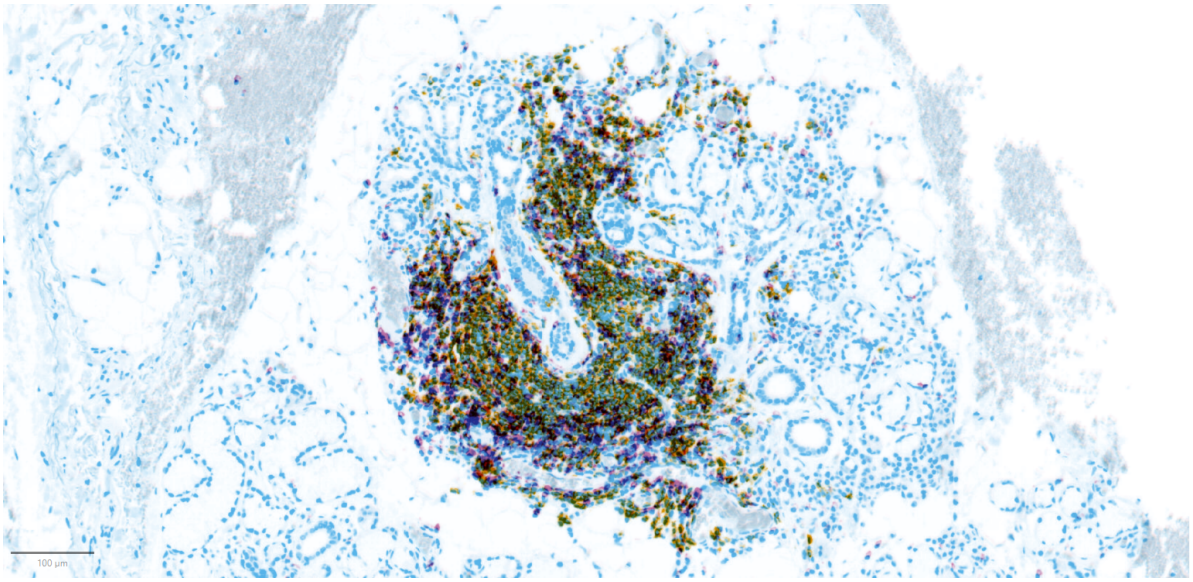
In Chapter 2, we have summarised some of the key cell types and their role in our immune system[1]. Here, we focus on a specific immune structure which is the focus of the data set. The name for TLS can be deconstructed into three constituent parts,

- Ectopic / Tertiary: Not themselves a well-defined organ, i.e. primary and secondary lymphoid organs such as the bone marrow or thymus and spleen or lymph nodes, respectively.
- Lymphoid: The main constituent cells are lymphocytes.
- Structures / Organs: Organised into immuno-aggregations which perform tasks as a collective, and somewhat similar to that of other lymphatic organs.

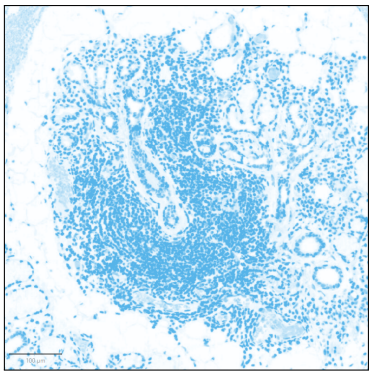
The TLS has been recognised to have prognostic value in many diseases, not least in cancers [107, 102, 131]. Aside from their standalone prognostic value, TLS presence has been indicated to improve the effectiveness of immunotherapies in a variety of cancers [132, 179, 180, 181].

The characterisation of these structures can also change depending on the organ where it presents itself [178]. However, the common markers for defining a TLS are a dense germinal

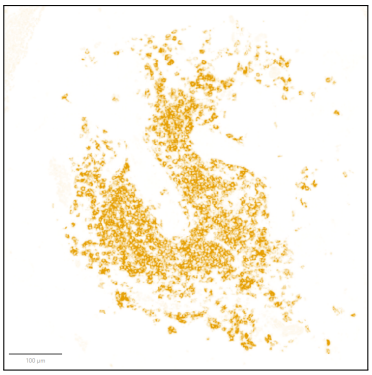
[1] See Section 2.2, Subsection 2.2.1



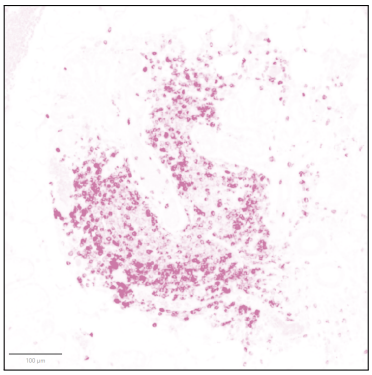
(a)



(b)



(c)



(d)

Figure 3.1: An example TLS from the TRACTISS data set. (a) A focused view on a TLS from Figure 3.2. (b) - (d) The individual markers used for the identification of TLS in the TRACTISS data set. The markers are (b) DAPI, (c) FITC, and (d) TRITC. They identify the nucleus of cells positive for CD20 and CD3, which indicate B cells and T cells, respectively.

centre of B cells with an associated ring of T cells encompassing it. Although these two statements are clear, biology often sits on a continuum between spatially distinct and well-mixed. We can see an example of this in Figure 3.1 and how the spatial organisation is hard to distinguish in real geometries of the tissue. This is where the difficulties of our research lie, and which we verify from several perspectives.

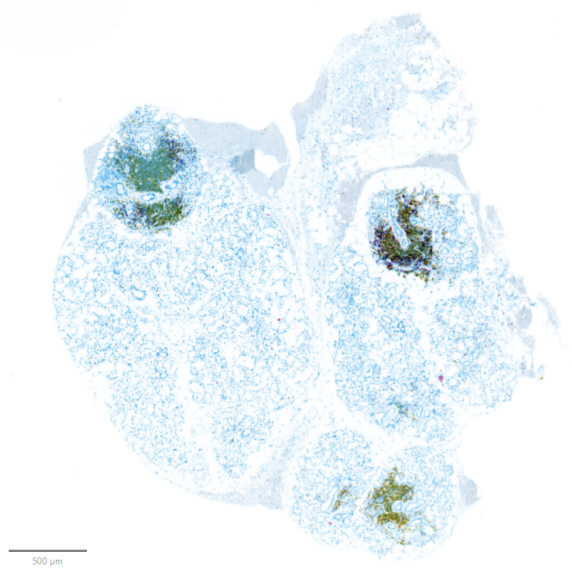


Figure 3.2: Whole slide image of a patient's salivary gland sample for the inspection of Sjögren's syndrome.

3.2.2 The TRACTISS data set

Sjögren's syndrome is a disorder of the immune system identified by its two most common symptoms: dry eyes and dry mouth. Primary Sjögren's syndrome is a chronic autoimmune disease characterized by lymphocytic infiltration and tissue destruction of exocrine glands such as salivary glands. The TRACTISS data set investigates the prognostic outcome of Sjögren's syndrome patients and compares it to the presence of TLSs. As explained in Subsection 3.2.1, the TRACTISS data set defines TLS structures by looking at the spatial organisation of B cells and T cells. As such, each human biopsy has been stained with immunofluorescence markers Cluster of Differentiation 20 (CD20) and CD3 for B cells T cells, respectively.

The study's objective is to identify TLSs for subsequent research by collaborators into their relation to disease progression. We interpret the properties of TLSs and translate them into mathematical interpretation. Then, we investigate algorithms relevant to this study and verify their accuracy in quantifying the properties of TLSs, as well as their limitations. The algorithms should be able to identify TLSs so that further investigation can analyse the number of constituent cells, area, composition, and spatial distribution to other features, such as secretory ducts or other TLSs. This information can then be compared to the prognosis of

the patient to assess the influence TLSs have. The foundation of future analyses thus relies on the accurate and automated identification of such structures.

3.3 Processing the Images

We use the following procedure to convert our images into a data format that can be used as input into our subsequent image analysis workflow. For each sample, we require the localisation of cells with respective cell phenotypes. This is similar to the format used in the previous chapter[1]. After processing the data, the application of our workflow for the automatic identification of TLSs can begin.

3.3.1 TRACTISS data acquisition

The data set was produced by Elena Pontarini and Michele Bombardieri at The Queen Mary University of London, in collaboration with Michelangelo Certo and Claudio Mauro at the University of Birmingham. The salivary gland tissue of 60 patients was imaged using mIF microscopy.

3.3.2 Cell Border Identification

First, we must identify the cell to which we will both assign a coordinate as well as a phenotype. A manual annotation for this detail in the image would be both extraordinarily time-consuming as well as subjective to an individual's experience in this type of annotation. Hence, we use artificial intelligence to replace the necessity for a trained human individual.

Two significant parameters are considered by most image processing workflows for the detection of cells. The first depends on the existence of nuclei staining. The second does not depend on this but instead on the translation of the training data to the currently applicable data set.

[1] See Chapter 2.4, Subsection 2.4.2

The first generalised method relies on the existence of nuclei staining to identify the nucleus. From here, the assumption is that the nucleus is both within a cell but also can be considered the centre of the cell. These assumptions need to be contemplated upon depending on the context of the cell types being identified. For example, some cell types of the human body can have no nucleus, like red blood cells, and others with multiple, such as osteoclasts. Once acknowledged, a threshold could be applied after a manual inspection or automatically based on the histogram of intensities, such as Otsu thresholding [182], or another artificial intelligence scheme, such as the two we introduce here. Following this validation, multiple nuclei have now been identified. The next step is to monitor the furthest extent the borders of the cell reach and distinguish that from the null space between cells.

An artificial intelligence scheme for thresholding which has proven successful for the automated detection of nuclei is called StarDist [2]. The algorithm utilises the construction of star-convex polygons radiating out from every pixel to identify the typically round cell borders. Since this is only well-defined for pixels that are associated with an object, this is used in combination with a lightweight U-Net [183] to identify whether a pixel is contained within an object to begin with.

A competing methodology can be used when the existence of nuclei staining does not exist. The majority of methods rely on convolution neural networks but are trained using different schemes. The most successful scheme used is called Cellpose [3]. The scheme is related to pre-CNN methodologies called a Watershed algorithm [184]. Instead of relying on values of pixels to create a topological map, whose basins represent the segmented regions, simulated diffusion was performed on ground-truth masks drawn by a human annotator. A neural network was then trained to predict the horizontal and vertical gradients of the topological maps, as well as a binary map that indicates if a given pixel is inside or outside of cell borders. Gradient tracking [185] then follows the vector fields and all pixels belonging to a given cell can be associated with a shared centre. Thus, by grouping pixels that converge to the same point, individual cell areas can be identified.

In our scenario, we do not have a marker for the cell membrane. As a result, the algorithms

can identify only the boundary of the nucleus with accuracy. Our input into the analyses requires only the centroid of cells, which the nucleus is a good estimate for, but a cell marker is often colocalised with a cell's cytoplasm rather than the nucleus. As an alternative, we dilate the nucleus border by a pre-defined distance to generate an approximation for the cell cytoplasm surrounding the nucleus. Then, we can determine the cell phenotype by a colocalised marker with this area.

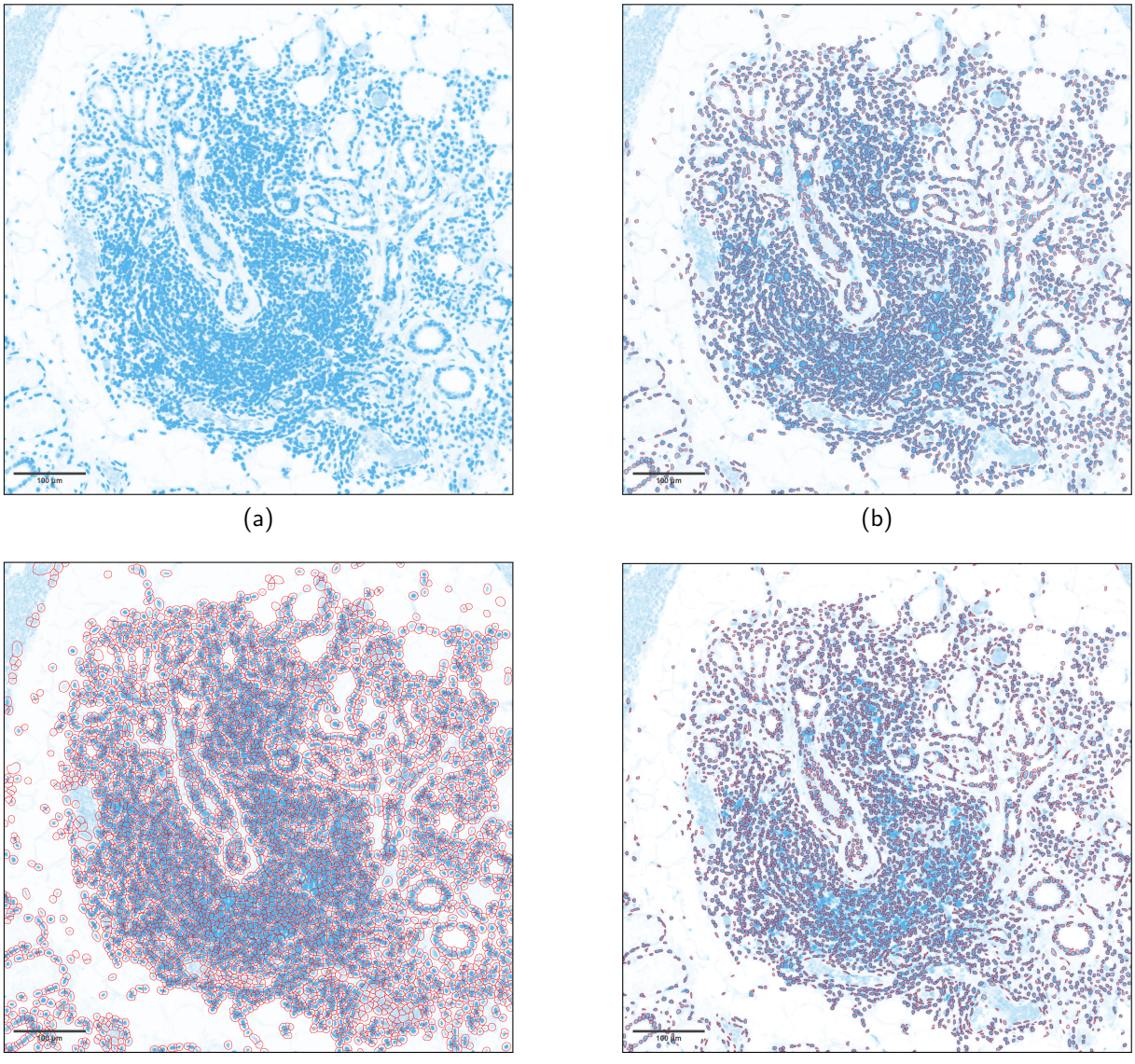
After the subjective assessment of the nucleus identification by our experimental collaborators, the Cellpose [3] algorithm was chosen for nucleus identification in our analysis. The principal reasons for this judgement were due to the preference for precision over recall, that is the identification of true nuclei at the expense of recalling all true nuclei albeit with some false positives. In addition, the boundary for nuclei defined by Cellpose appeared a closer fit than that of StarDist. However, it must be said that similar qualitative results were gained from StarDist. Yet, another reason to incorporate Cellpose was the ability to provide training images at a later date, a feature of the Cellpose algorithm, which could improve nuclei detection specific to our images.

3.3.3 Cell Centroid Approximation

Once identified, the cell centroid was obtained using the methods integrated with the QuPath [186] imaging platform. The authors consider the technique a more involved approach than simply taking the centre of the bounding box for a cell or the average of coordinates that parameterise the boundary [187]. With this guarantee, we can assume the centroid of the nuclei has been determined appropriately.

3.3.4 Cell Phenotype Identification

Incorrect cell phenotypes may be present due to low cell segmentation quality, antibody 'bleeding' from one cell to another or inadequate fluorescence signal-to-noise ratios. Despite these issues, careful experimental design can emit the majority of these shortfalls. If so, a manual



(a) (b)

(c) (d)

Figure 3.3: The application of cell border detection algorithms to the TLS in Figure 3.1. (a) The algorithms were applied to the ROI. (b) - (d) presents the outline (red) of the nucleus used as the centroid of the cells. We applied the (b) Watershed, (c) StarDist [2], and (d) Cellpose [3] algorithms.

inspection can identify the minimum fluorescence intensity values for pixels for which a cell is detected emitting a true signal of the respective phenotype. In our analysis, we have two well-defined classes of cell phenotype and so this method is sufficient. However, for the instances where there are no clear thresholds, we can refer to automated techniques that account for multiple fluorescence marker intensities to design new phenotypes or account for other features of a given detected cell, such as geometric properties. In the latter case, machine learning algorithms generally take into account properties such as cell shape, nucleus size etc., rather than a strict threshold, and therefore some false-positive cells will have high marker intensities, or false negatives will have lower intensities, yet classified under the same cell phenotype. Alternatively, new phenotypes can be designed using procedures from multi-dimensional reduction analysis. For example, two leading techniques, Uniform Manifold Approximation and Projection (UMAP) [188, 189] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [190], are utilised in the R image analysis library SPIAT [136] to discern misclassified cells.

3.4 Structured Cellular Neighbourhoods

The objective of the following is to identify appropriate algorithms in the pursuit of automatically identifying spatially organised, high-density aggregations such as TLS and which we refer to as *Structured Cellular Neighbourhoods*. One approach would be to provide images on a subset of TLS to train a supervised machine learning algorithm. However, our approach is motivated by the need to identify other kinds of spatially organised, high-density aggregations with unspecified cell markers, like those from our previous chapter. This is currently an unaddressed issue in mIF imaging with current methods considering neighbourhoods of only the nearest interacting cells[1]. The expectation is that once validated on a well-characterised higher-order structure, then some certainty can be expected in the results when applied to other cell phenotype combinations. Therefore, we expect to provide a solution which can identify structured cellular neighbourhoods on a more relevant biological scale, such as TLSs.

[1] See Chapter 2, Section 2.9

In addition, manual methods for the delineation of TLS are not well-defined. Introducing objective techniques to delineate a boundary for TLS enable systematic analyses to quantify these complex spatial structures. As a result, we expect to provide a workflow which will address this gap.

The feature of a TLS will act as a template for all types of structured cellular neighbourhoods. Therefore, our workflow will rely on algorithms which quantify the holistic features of TLS. The primary characteristics of TLS are as follows:

1. Dense aggregations of cells,
2. Germinal-like centre of a specific cell type, and
3. Surrounding boundary of a paired cell type.

In addition to these primary characteristics, the chosen algorithms need to be robust to complex geometries presented in biology. Furthermore, the variability in cell counts ranges in the early-stage formation of TLSs to that of those which have fully developed in structure, and so workflows need to account for this spectrum of TLS maturation. All this is to be interpreted from a single time point.

The remaining section identifies the core algorithms which quantify the characteristics applicable to structured cellular neighbourhoods, based on the template provided by TLS.

3.4.1 Identifying High-Density Aggregations

We begin our workflow by identifying dense aggregations in our data set. There are several algorithms available in the literature that classify a two-dimensional point cloud into constituent aggregations. A discussion of established algorithms and descriptions of the algorithms compared in this section are presented in Appendix A (Section A.5). Here, we compare two identified density-based algorithms, called Density Based Spatial Clustering And Noise (DBSCAN) and Topological Mode Analysis Tool (ToMATo), and perform a meta-comparison on a large database of benchmark problems to validate the algorithm applied in our workflow.

Studies have compared many of the aforementioned clustering algorithms. Some create artificial data directly related to their application [191, 164], while others consider benchmark data sets that test the limitation of such methods [192, 193, 149].

To compare DBSCAN and ToMATo, the algorithms are applied to a series of artificial data sets to validate spatial clustering algorithms in a wide range of contexts. The classification for a given set of parameters is then evaluated using the Adjusted Rand Index (ARI). Before explaining how to compute the ARI, we consider the Rand Index. The Rand Index, RI, is a measure of the similarity between two sets of cluster indexes and is calculated using,

$$\mathrm{RI} = \frac{a+b}{\binom{n}{2}} \tag{3.1}$$

where a is the number of agreements between the ground truth and the cluster results, and b is the number of disagreements between the ground truth and the clustering results. $\binom{n}{2}$ is the binomial coefficient describing the number of unordered pairs between sets a and b.

The ARI corrects for the random chance of points being assigned to the correct clusters. For clusterings X and Y, the common elements between the two can be summarised using a contingency table. Then, the ARI is calculated as follows,

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\Big/\binom{n}{2}} \tag{3.2}$$

where n_{ij} is the number of elements in common between $X_i \subset X$ and $Y_i \subset Y$, a_i is the sum of the contingency table for column j.

For each parameter combination, the clustering result (that is, the cluster indexes) is compared with that of the ground truth simulation, and the ARI is calculated for that parameter combination.

For an equitable comparison between clustering algorithms, we compute a parameter sweep for each clustering algorithm based on the range of parameters possible for a given point pattern. For the DBSCAN clustering algorithm, we compute a set of `Eps` equal to all Euclidean distances between nearest neighbours. Then for each value of `Eps`, the set of `MinPts` $\in$

Algorithm	ToMATo	DBSCAN	Equal
Count	35	10	27

Table 3.1: Number of data sets which either ToMATo,DBSCAN, or both produced the highest ARI score. ARI for individual data sets are presented in Table B.2

$\{m \mid 1, 2, \ldots, \lceil n/2 \rceil\}$ are the parameters for the DBSCAN clustering algorithm, where n is the number of points in the point pattern. The rationale behind these sets of parameters is that, for `Eps`, there exists at least one iteration of the clustering where every point has at least one point within their own `Eps`-neighbourhood, for `MinPts`, the minimum consideration for point density is considered as well as the maximum threshold which only one cluster is identified for the given point pattern. Similarly, the parameter set for r in the ToMATo clustering algorithm is chosen in a similar way to the `Eps` parameter set considered for the DBSCAN clustering algorithm, with the same reasoning. It has been discussed previously that the choice of persistence threshold, τ, in ToMATo is independent of parameter inputs of r. An automated scheme for the parameter τ was designed whereby persistence is considered for every sequential feature persistence value, beginning at a fixed minimum percentile. In this analysis, feature persistences in the lowest 25th percentile were not considered.

The benchmark data sets were obtained from a database of clustering benchmarks [194]. The project contains a collection of artificial and real-world problems that were found in the literature. However, only a subset of the artificial data set is used since this had a labelled ground truth associated with it. We applied the DBSCAN and ToMATo algorithms to the subset of labelled data and computed the ARI score for each.

The results for each data set are displayed in Appendix B (see Table B.2). To summarise these results, Table 3.1 presents the number of times that the ARI score was higher for which clustering algorithm and the occasions they were equivalent for both. The results of our clustering algorithm comparison show that ToMATo has a higher ARI score for the majority of data sets. This also complements other biological analyses which show that not only is ToMATo a more preferable clustering algorithm but it produces robust outcomes when analysing variations in input parameters [164]. Therefore, the ToMATo clustering method is

utilised in our workflow.

3.4.2 Identifying the Boundary of Complex Geometries

Once a set of point coordinates has been classified to be associated with a specific aggregation, the identification of the aggregation's boundary can be determined. The density-based aggregation identification algorithms previously considered assigns an index for the aggregation each point coordinate is predicted to be associated with, if any. Next, our analysis must draw a boundary to classify what can be considered as inside the aggregation and outside. This is necessary since the aggregation identification algorithms will have as input only a subset of point coordinates, specifically cells with a phenotype chosen *a priori*, and we would like to monitor all cell phenotypes contained in an aggregation. Since the output of the previous analysis has given us the point coordinates associated with a particular aggregation, we use this as our input to identify the border of the respective aggregation. To draw a boundary of best fit around an aggregation, we do so by discerning the coordinates which form the boundary of the respective polygon. The output should then produce a set of ordered boundary coordinates to draw lines between to distinguish the interior and exterior of the aggregation. This subsection will consider three methodologies to calculate the boundary for aggregations, providing insight into their computation and the benefits and limitations of each.

First, and foremost, what is meant by a boundary of *best fit*? Often, this parameter is assessed subjectively. If assessed objectively, then minimising the area for a given set of point coordinates is used. However, this criterion does not specify the number of unique boundaries so a polygon which identifies every point as associated with the boundary would maximise this metric despite identifying little to no remaining area. This would be an equivalence to the minimum spanning tree. As a consequence, the evaluation of boundary identification algorithms needs to weigh the accuracy for which they identify an area as part of the original aggregation and that which is not. As such, to properly assess whether a boundary is of the best fit we translate the evaluations of binary classifiers. In summary, the assessment considers both the area which is recovered balanced by the area which is not recovered. The evaluations

are as follows:

- True Positive (TP): correctly identifies the area as part of the original distribution.
- True Negative (TN): correctly identifies the area as not part of the original distribution.
- False Positive (FP): incorrectly identifies the area as part of the original distribution.
- False Negative (FN): incorrectly identifies the area as not part of the original distribution.

This is then normalised as a proportion of the total area covered by the convex hull, representing the maximal area detected for our considered algorithms given a set of point coordinates. The subsequent classification precision, recall and F_1 score are calculated to summarise the overall accuracy of the identified boundary.

Classification precision, p, also known as a positive predictive value, indicates the fraction of area correctly classified as the original area compared to all the area identified by the algorithm and is calculated as

$$p = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}} \tag{3.3}$$

Classification recall, r, or sensitivity, indicates the fraction of area correctly classified against the total area of the underlying polygon and is calculated as

$$r = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}} \tag{3.4}$$

The F_1 score is the harmonic mean of precision and recall, expressed as

$$F_1 = 2\left(\frac{pr}{p + r}\right) \tag{3.5}$$

The F_1 score of 0 indicates no agreement for the automated and manual analysis, while 1 indicates complete agreement. It is by these scores which we will evaluate the success of the boundary identification algorithms.

It should be noted that another metric for evaluating such areas is the Intersection of Union score, which is equivalent to only reporting the TP area as a proportion of the ground truth

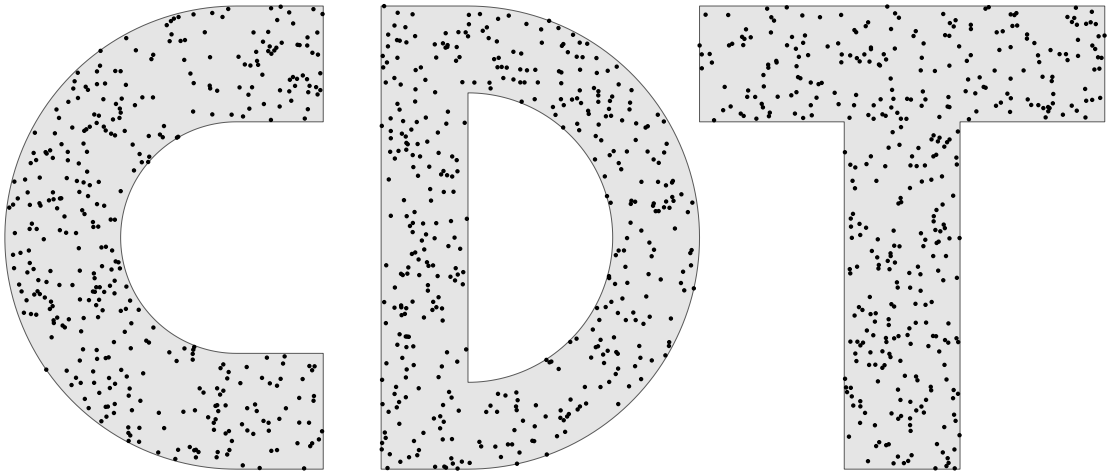
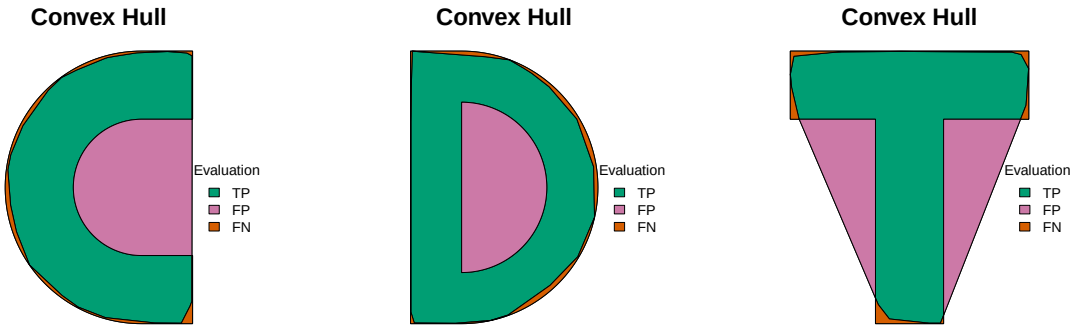


Figure 3.4: Ground truth for our benchmark dataset.

area. Therefore, there is no reported consequence for identifying too large a boundary which is guaranteed to overlap the original area. For this reason, we do not report this metric.

To assess the performance, we test the algorithms on point coordinates with known boundaries. In this case, we have a set of points sampled from the area of the letters C, D and T in Figure 3.4. These letters were chosen due to specific properties which each exemplifies. The letter C was chosen to test the identification of concave and convex boundaries. The letter D tests the ability of algorithms to identify internal boundaries. Finally, the letter T tests right-angled concave corners. To create the sample of points for each letter, points were uniformly randomly distributed across the domain, and then, the points that were within the underlying boundaries were kept. It should be noted that the area for any identified boundary based on this set of point coordinates will more than likely be an underestimate of the ground truth since they are originally defined as within the ground truth boundary. The density of points also determines the accuracy for which the boundaries can be recreated, specifically the chances that points are sampled close to the original boundary. Considering the limitations of the benchmark data set, we use these three letters to evaluate the performance of the ensuing boundary detection algorithms.

The most well-known technique for the calculation of an aggregation's boundary is called the convex hull. There are several equivalent definitions for the convex hull, which here we will



(a) (b) (c)

Figure 3.5: Evaluations of the convex hull boundaries. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively.

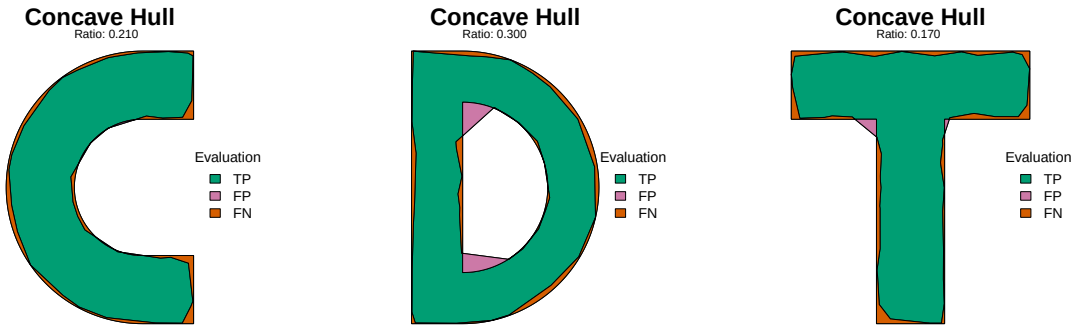
define as follows: The union of all simplices with vertices in X. Consider a set of coordinates in the 2-dimensional plane. Intuitively, one can consider the release of a rubber band around all the points; the perimeter of the final rubber band's position is defined as the convex hull for that set of points. The definition is not restricted to 2D as noted from our definition, yet in our application, we only consider a series of 2D coordinates. The implementation considered for implementation in our analysis was proposed by Eddy [195, 196]. Eddy suggests that the implementation of their algorithm CONVEX is analogous to the partitioning of points in the sorting algorithm QUICKERSORT with the additional insight that points inside a triangle formed of the convex hull vertices cannot be convex hull points themselves. Thus, the algorithm proposes ways to partition the points into such triangles. The computation of the convex hull algorithm on our benchmark data set is presented in Figure 3.5. The obvious limitation of the algorithm is the inability to properly characterise concave boundaries. Despite this, the convex hull can provide a maximal boundary to consider for our point set with no input parameter necessary. This is unlike later algorithms, which give rise to uncertainties in parameter choice for a set of point coordinates with an unknown underlying distribution.

The pivotal assumption for the convex hull is that the polygon of best fit is also convex. As shown by our benchmark data set in Figure 3.5 this becomes a bad assumption for other common boundaries. In addition, we cannot make such an assumption in the biological con-

text of aggregations either, as shown by Figure 3.1. Since we are dealing with non-convex geometries, the significant limitation of the convex hull leads us to find alternative methodologies. Therefore, we consider two algorithms that instead calculate concave geometries. An intuitive description for the concave hull is that instead of a rubber band, we are now "shrink-wrapping" a set of points. Determining the resolution for which the boundary of best fit exists now becomes the issue. Both of the subsequent algorithms include a parameter which reflects the resolution and, thus, the progressive scale of concavity. Generally, this parameter should match this intuitive understanding of concavity, which can be described in the following way:

- Simple geometric basis: this allows the user to understand the effect of the parameter and aids in determining an effective value.
- Scale-invariant: this allows a single parameter value to be effective on varying sizes of geometry, which removes an additional computational step which is efficient in batch or automated processing.
- Local (as opposed to global): A local property (such as edge length) gives the algorithm latitude to determine the concave shape of the points. A global property (such as area) over-constrains the possible result.
- Monotonic area: larger (or smaller) values produce a sequence of more concave areas
- Monotonic containment: the sequence of hulls produced are topologically nested.
- Convex-bounded: an extremal value produces the convex hull.

The first, a popular implementation, uses the Delaunay triangulation. The concave hull algorithm, as proposed by the GEOS software package [138], requires a value, called the ratio (φ), to determine the maximum threshold of the longest edge in the Delaunay triangulation. The maximum edge length is then the shortest edge length divided by the pre-defined ratio. Edges which are longer than the threshold are then removed from the triangulation. The concave hull is formed of all the remaining triangles, with the outermost edges of the Delaunay



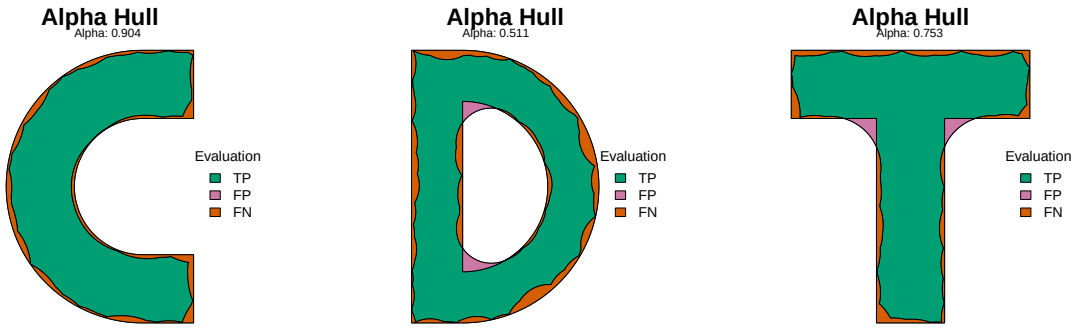
(a) (b) (c)

Figure 3.6: Evaluations of the concave hull boundaries. The respective ratio parameters found to produce the highest F_1 score were (a) $\varphi = 0.21$, (b) $\varphi = 0.30$, and (c) $\varphi = 0.17$. The evaluation scores were computed for ratio, $\varphi \in \{n\Delta x \quad | \quad n \in 1, 2, ..., 100\}$ and $\Delta x = 0.01$, with holes allowed in each case. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively.

Triangulation defining the boundary. An additional parameter considers whether holes are allowed to exist within the polygon defined by this boundary. In this case, the internal edges of the Delaunay Triangulation are removed according to the shortest edge criteria to produce voids within. The computation of the concave hull algorithm on our benchmark data set is presented in Figure 3.6. As discussed previously, this benefits from identifying concave boundaries, given a resolution of concavity.

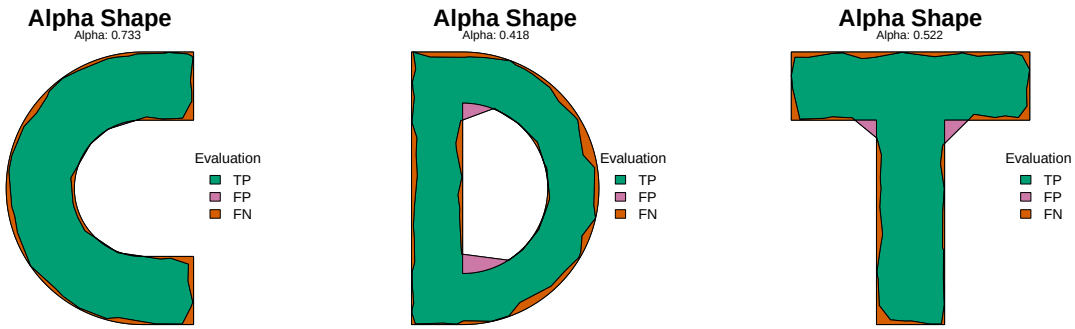
The following algorithm provides a mathematical definition of a subset of concave hull polygons. Edelsbrunner et al. [139] first describe these constructs and distinguish those that produce a convex hull as positive α-hulls and those which produce concave hulls as negative α-hulls.

The evaluation score for each algorithm given the parameter which produced the highest F_1 score is presented in Table 3.2. The summary provided by the evaluation suggests that the algorithm with the highest average F_1 score is the concave hull algorithm. This is followed by the alpha shape and alpha hull algorithms where the former has a higher precision contribution and the latter a higher recall contribution. Finally, the convex hull has the lowest F_1 but the largest recall. This supports our original choice of evaluation metric since the Intersection of



(a) (b) (c)

Figure 3.7: Evaluations of the alpha hull boundaries. The respective alpha parameters found to produce the highest F_1 score were (a) $\alpha = 0.904$, (b) $\alpha = 0.511$, and (c) $\alpha = 0.753$. The evaluation scores were computed for alpha, $\alpha \in \{\text{all edge lengths in the Delaunay triangulation of the point coordinates}\}$. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively.



(a) (b) (c)

Figure 3.8: Evaluations of the alpha shape boundaries. The respective alpha parameters were (a) $\alpha = 0.733$, (b) $\alpha = 0.418$, and (c) $\alpha = 0.522$. The evaluation scores were computed for alpha, $\alpha \in \{\text{all edge lengths in the Delaunay triangulation of the point coordinates}\}$. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively.

Union metric would have given the highest score to the convex hull for identifying more of the underlying boundary without evaluating the areas of inaccuracy, such as false positive or false negative area allocations. It is also important to note that beyond the F_1 score, the accuracy which is prioritised could persuade an individual to consider the Alpha Hull algorithm since it produces the highest precision score in two of the three underlying distributions. On the other hand, for generality, we have shown that the Concave Hull is appropriate for implementation in our workflow.

	C			D			T		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Convex Hull	0.779	0.660	**0.952**	0.824	0.726	**0.952**	0.792	0.670	**0.967**
Concave Hull	**0.962**	0.997	0.929	**0.945**	0.964	0.927	**0.953**	**0.992**	0.916
Alpha Hull	0.944	**1.000**	0.894	0.912	**0.978**	0.854	0.934	0.977	0.895
Alpha Shape	0.958	0.997	0.921	0.932	0.971	0.895	0.946	0.983	0.912

Table 3.2: The evaluation score for each algorithm and each underlying distribution to identify given the parameter which produced the highest F_1 score.

3.4.3 Density Gradients for Multitype Point Patterns

The following methods can be used to translate subjective human-made descriptions into well-defined objective metrics. The following steps assume that we now have a list of potential candidate high-density aggregations. The methods discussed here, and in the subsequent section, look to separate the high-density aggregations which meet a threshold of spatial organisation necessary to be classed as a structured cellular neighbourhood.

In our workflow, we propose splitting the determination of structure into two steps. The first determines the spatial correlation of specified cell phenotypes within the high-density aggregation and those which surround the high-density aggregation, i.e. density gradients. The second utilises persistent homology to confirm the radial symmetry in the previous observations. In this subsection, we will discuss the practicalities of measuring density gradients on resolutions similar to the scale of individual cells, with simulations performed on artificial point patterns.

Density gradients measure the density of one cell type radiating out, up to a set distance,

and calculate the rate of change in this area. In our analysis, we begin measuring from the centre, or more specifically a small region around the centre of mass of the high-density aggregation, to the boundary of the high-density aggregation. The gradient is then measured beyond the perimeter to the surrounding environment which the high-density aggregation is considered to have an influence over. As a consequence, we can measure the rate of change in a cell type from the centre of a high-density aggregation to its border, and then from the border to the periphery.

The utilisation of density gradients in our application is due to the second and third primary characteristics of a structured cellular neighbourhood. In our workflow so far, we have been able to identify high-density aggregations. Specifically for TLSs and the TRACTISS data set, these are the B cells which form the germinal centre. First, a measurement for within-aggregation density gradient will determine that there is not a higher density of cells at the boundary than compared to deeper within the aggregation, relating to primary characteristics (2) of a structured cellular neighbourhood. Second, a measurement of the density gradient from the boundary into the periphery will then determine if there is the presence of another cell phenotype in the vicinity of the high-density aggregation, relating to primary characteristics (3) of a structured cellular neighbourhood.

The density gradient is computed in a similar way to the computations of the ECI metric discussed in Chapter 2[1]. The principle is that we compute the density for binned distances from the boundary to obtain observational values which we can subsequently interpolate between. Similar to before, we utilise algorithms which offset/buffer the identified boundary of the high-density aggregation. Succeeding in these computations, the distances are then normalised between aggregations so that the comparisons can be performed between aggregations regardless of size. Finally, a threshold is then applied to determine which of the identified high-density aggregations are not considered and likely unstructured.

Next, the resolution of binned distances, thus sampled areas, is necessary to define before normalising for between-aggregation application. For the implementation of our workflow to

[1]See Chapter 2, Section 2.6.

the TRACTISS data set, we calculated the density gradient for a set of sample areas, which reflect a coarse resolution to observe the gradient density. Within the aggregation, these were the sample areas which represent the proportions of the original aggregation at binned distances [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), and [0.8, 1). The smallest area that could be sampled for the binned distance [0, 0.2) is usually too small for a realistic sample area, so would either have no cell present and a value of 0 for density, a cell present and a value of density tending to infinity, or self-intersections from the offset algorithm computation, producing boundaries which represent the "opposite" side of the polygon which they once defined which can introduce unintended holes. From these binned distances, we represent each measurement for density as the midpoint of the sample, i.e. at 0.3, 0.5, 0.7, and 0.9, respectively. For outside the aggregation, we sample areas from the boundary up to a maximum periphery distance from the aggregation boundary for which the density of cells is measured. The sampled areas then represent annuli of the surrounding environment up to this maximum periphery distance, i.e. at binned distances as a proportion of the maximum periphery distance at [0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), and [0.8, 1). From these binned distances, we represent each measurement for density as the midpoint of the sample, i.e. at 0.1, 0.3, 0.5, 0.7, and 0.9, respectively. It is here that we can then fit a simple linear regression to model the values both within and outside the aggregation, respectively. Then, the gradient of the model is an approximation of the density gradient for a cell phenotype in each region.

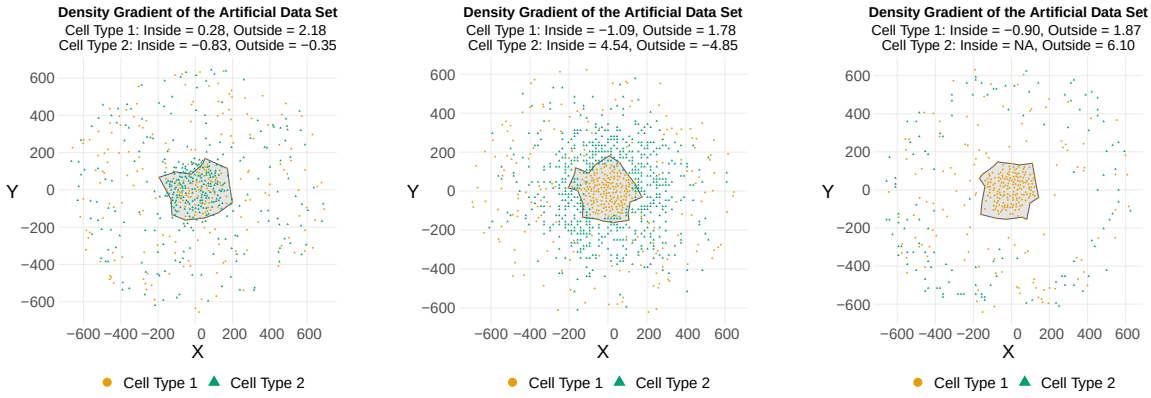
The remaining factor to assign is the threshold for which high-density aggregations are considered to have spatial organisation inside and/or near their boundary. The construction of this workflow has been purposefully designed to allow for some flexibility in this value. As mentioned in the preamble for this section, monitoring the range in structure might be another avenue of interest in determining the continuum of immature and mature TLSs. However, in our current line of analysis, there is a clear definition of structure for a given cellular neighbourhood. In this study, we are only interested in TLS aggregations which have reached full maturity. For the TRACTISS data set, this means we are interested in B cells within a core high-density aggregation. This suggests a value for the density gradient of CD20 within

the aggregation has a value which is at most a small negative value, i.e. has a higher density of cells in the centre of a core than towards the boundary, instead of a dispersion of points at the centre of the core and spatial correlation at the boundary. Next, we require the density of T cells to have a higher density of cells closer to the boundary of the aggregation, which diminishes away from the boundary and towards the periphery. As a result, the threshold for the density gradient of CD3 is expected to have a negative value above a pre-determined threshold.

To provide some perspective on the choice of thresholds, we apply the methodology to the artificial point patterns presented in Chapter 2[1]. In Figure 2.6, four artificial point patterns are presented, each constructed with unique features relevant to the HALO1079 data set discussed in Chapter 2. Nonetheless, the point patterns are relevant in the identification of structured cellular neighbourhoods since three of the four of them consider a high-density aggregation with distinct differences in structure. That is: the independent point pattern (see Figure 2.6b) has a high-density aggregation with no discernible spatial organisation between cell types; the proportional point pattern (see Figure 2.6c) has a high-density aggregation of Cell Type 1 with Cell Type 2 spatially correlated to the boundary; and, the inverse point pattern (see Figure 2.6d) has a high-density aggregation of Cell Type 1 but with spatial inhibition of Cell Type 2 from the boundary.

To identify the high-density aggregation, the ToMATo clustering algorithm was applied to the subset of Cell Type 1 points for all point patterns of the artificial data set, with parameters $r = \sigma = 50$, where σ is the value of the standard deviation for the Gaussian distribution chosen to construct Cell Type 1 in the mixed, proportional and inverse point patterns. Then, a persistence threshold of $\tau = 10$ identified exactly one high-density aggregation of Cell Type 1 in all but the Poisson point pattern. The concave hull algorithm was then applied to the identified points of Cell Type 1 in the high-density aggregation, with $\texttt{ratio} = 1/2$. Finally, the maximum periphery distance was set to 500 μm from the boundary. The results of this computation are presented in Figure 3.9. In our artificial data set, we are trying to discern the

[1]See Chapter 2, Section 2.7



(a) (b) (c)

Figure 3.9: Results of the density gradient on the artificial data set, presented in Chapter 2, Section 2.7. The plots show the density gradient results for each cell type for the high-density aggregations identified and their respective boundary (black line) in the (a) independent, (b) proportional, and (c) inverse point patterns. Only points within 500μm of the aggregation boundary are shown.

structure between high-density aggregations. Primarily, we are determining whether there are spatial differences between cell types.

For the independent point pattern, we observe that the value for gradient density (see Figure 3.9a) for both cell types has a small positive magnitude inside the identified high-density aggregation. This suggests that there is no distinctive gradient in density from the centre of the aggregation to the identified border if not slightly more in the centre. This reflects well with the subjective description of this point pattern, which is that there is one well-mixed high-density aggregation. Considering the outside estimates of gradient density, we observe that for Cell Type 2, the relationship is similar if not slightly negative this time, while the value for Cell Type 1 is a notable magnitude positive. The value for Cell Type 1 might be unexpected, however, is explained by the prior computations to identify the high-density aggregation. That is, the border has been constructed based on the clustering of Cell Type 1. Therefore, the border identified is the natural distinction between points of Cell Type 1 that are considered within a higher density to those that are in a significantly lower density. Thus, for small distances from the boundary, the density of Cell Type 1 is observed to be lower. Then, for larger distances from the boundary, the density will return to the global density

on average. It is because of the cut-off for distinguishing inside and outside a high-density aggregation that a non-significant magnitude for gradient density is observed.

In the proportional point pattern presented in Figure 3.9b we have a core high-density aggregation of Cell Type 1 surrounded by a dense annulus of Cell Type 2. This is the artificial example most related to the characteristics of our TLS, albeit for a more symmetric example. Inside the core, the value of Cell Type 1 is more prominently negative than before, suggesting the density of Cell Type 1 is more dense towards the centre of the high-density aggregation than towards the boundary. In addition, Cell Type 2 is predominantly excluded from the centre of the core reflected by a high positive value for gradient density. This is in contrast to the outside gradient density which has switched to a negative value and represents a steep gradient when considering the density close to the boundary compared to the periphery. This is complemented by a positive value in the outside gradient density of Cell Type 1, reflecting the exclusion of points close to the boundary from Cell Type 2 clustering.

The inverse point pattern produces a distinctively different high-order feature in the domain than previously considered. The inversely proportional spatial probabilistic correlation of Cell Type 2 to Cell Type 1 constructs a point pattern where Cell Type 2 is inhibited from the high-density aggregation of Cell Type 1. Once again, the density gradients summarise this effect. Principally, the spatial inhibition is illustrated by the large positive value for the outside density gradient. Then, the inside density gradient is Not Applicable (NA) since there are no observable Cell Type 2 points inside the boundary of the high-density aggregation defined by Cell Type 1. The description of the density gradient for Cell Type 1 is similar to that discussed for the proportional point pattern case.

We have observed for the three artificial point patterns where high-density aggregations have been identified that the density gradient can differentiate spatial localisation with reference to an identified boundary. The description provided by the density gradient can distinguish between no spatial correlation, a strong positive spatial correlation to high-density aggregations and an inverse spatial correlation to high-density aggregations between two different cell types. This has been determined by the magnitude and the sign of the value for the density

gradient. Where the magnitude indicates the strength of the correlation, the sign indicates the type of correlation between the cell type and the localisation to the boundary of a previously identified high-density aggregation.

So far, we have presented the theory behind this measure and the effectiveness with which it can distinguish between artificial point pattern features. When applying the measure to biological data, we need to understand the limitations of the metric. As described before, the measure averages the density within an annulus. The implicit assumption is that the density within each annulus can be estimated as radially symmetric. However, the results for density gradient can be biased by localised regions of density adjacent to but not radially symmetric around the border. In this scenario, the gradient density would produce an average which represents a spatial correlation and a segregation between the cell types. This is because the mean for each annulus would be skewed by the measurement of the localised density. Therefore, the extent to which the gradient density can identify the confirmation of a TLS is hindered by this underlying assumption. Yet, the information provided by the gradient density is still in use. In this case, we look to underwrite the implicit assumption that density around an identified high-density aggregation is radially symmetric. The answer we propose in this workflow is to utilise Persistent Homology.

3.4.4 Persistent Homology for Loop Identification

The mathematical foundations of Persistent Homology (PH) are discussed in the Appendix to this thesis[1]. We utilise PH in the application of TLS identification to confirm the assumption in measuring the density gradient calculations: that density is radially symmetric about the border of a high-density aggregation. The reasoning for using PH for our application is that our objective is to confirm whether a 1-st dimensional homology group feature encompasses our high-density aggregation. In other words, is the high-density aggregation "surrounded" by the cell type in question.

PH considers all 1-st dimensional homology group features from a point cloud. The cycles

[1]See Appendix A, Section A.2

are generated by vertices which are in turn coordinates from our point patterns. Then, we can identify the relevant points and centreline joining points of a cycle through offsetting[1] by a value of ϵ, which is equivalent to the overlapping ϵ-neighbourhoods of the vertices which construct the cycle. This is the loop which we compare to the boundary of the high-density aggregation. The cycle at birth is not used since this is the union of the boundaries of the radial balls drawn from each vertex. Therefore, this would represent a smaller version of the boundary otherwise obtained from linking between vertices. If we were to use the cycle, this would likely be smaller than the physical boundary of the high-density aggregation for which we would like to determine if the vertices of a complimentary cell phenotype surround.

Once computed, one approach would be to consider whether all 1-st dimensional homology group features in the filtration encompass the high-density aggregation through methods of mereotopology [197]. However, it would be computationally expensive to consider all features when many would be considered too small or physiologically irrelevant. Instead, understanding the intuition behind PH, we can extrapolate the physical properties of loops and threshold for features of the most relevance. The result is a quantifiable way to discard features that should not be considered and compare the remaining features which can be thought of as important. To do this, we consider two physical properties associated with each loop which we can threshold directly from PH: Birth and Persistence.

The *Birth* value of a 1-st dimensional homology group feature relates to the value of the ϵ-neighbourhood in which the feature first exists. Equivalently, the vertices of the feature are at most within distance ϵ of one another and construct a loop. The *Persistence* of a 1-st dimensional homology group feature reflects the difference in ϵ from when the feature first existed to when the feature stops to exist. That is the difference in value from when the maximum distance between points first creates a 1-st homology group feature to when the loop collapses to become a 0-th homology group feature. Another way to interpret this property is how the value relates to the diameter of a given loop. In this context, we are observing a loop constructed by the overlapping ϵ_1-neighbourhoods, where ϵ_1 is the Birth value. Then,

[1] See Chapter 2, Section 2.6

for growing ϵ-neighbourhoods where the 1-st homology group feature exists, the number of additional points continues to increase but crucially not that the cycle is subsumed into a 0-th homology group feature. Then, the feature "dies" when considering the ϵ_2-neighbourhood, where ϵ_2 is the final filtration value when the vertices of the original feature overlap with each other or points contained by the cycle. Essentially, persistence is an estimate of the void space in the centre of the loop. Extrapolating this property of the feature can inform the size of a loop and so discard loops too small to encompass the boundary of a high-density aggregation.

In Figure 3.10, we consider the procedure applied to the proportional point pattern, first presented in Chapter 2[1]. We present the computational steps to threshold identified loops and test whether the high-density aggregation is in the interior of the polygon defined by the boundary of the loop. First, we compute the PH of the point cloud that remains after removing points within the boundary of the high-density aggregation. The results of which are presented in Figure 3.10a as a persistence diagram. This identifies all homology group features constructed by the point cloud. We can appreciate the many 1-st homology group features, or cycles, which relate to our loops where often they are too small to be comparable with the high-density aggregation (see Figure 3.10b). To consider a subset of loops most relevant to the aggregation in question, we introduce thresholds on persistence and an additional threshold on the maximum vertex-vertex distance to consider only loops constructed by a more dense sequence of vertices. In Figure 3.10c, we can see that we have considered thresholds of $\tau > 20$ and $\epsilon_1 < 50$ as inequalities for Persistence and Birth, respectively. The remaining 1-st dimensional homology group feature is shown in Figure 3.10d and the high-density aggregation is confirmed to be an interior of the loop in the Euclidean plane.

The use of PH for the identification of loops which are formed by a series of close vertices provides a metric to consider when a cell phenotype is surrounding a high-density aggregation boundary. The reasoning is, that if there exists a set of vertices which are close enough together, set by the Birth value, to encompass the whole boundary, set by the Persistence, then it indicates there is no bias in gradient density due to the localised aggregations of the

[1]See Chapter 2, Section 2.7

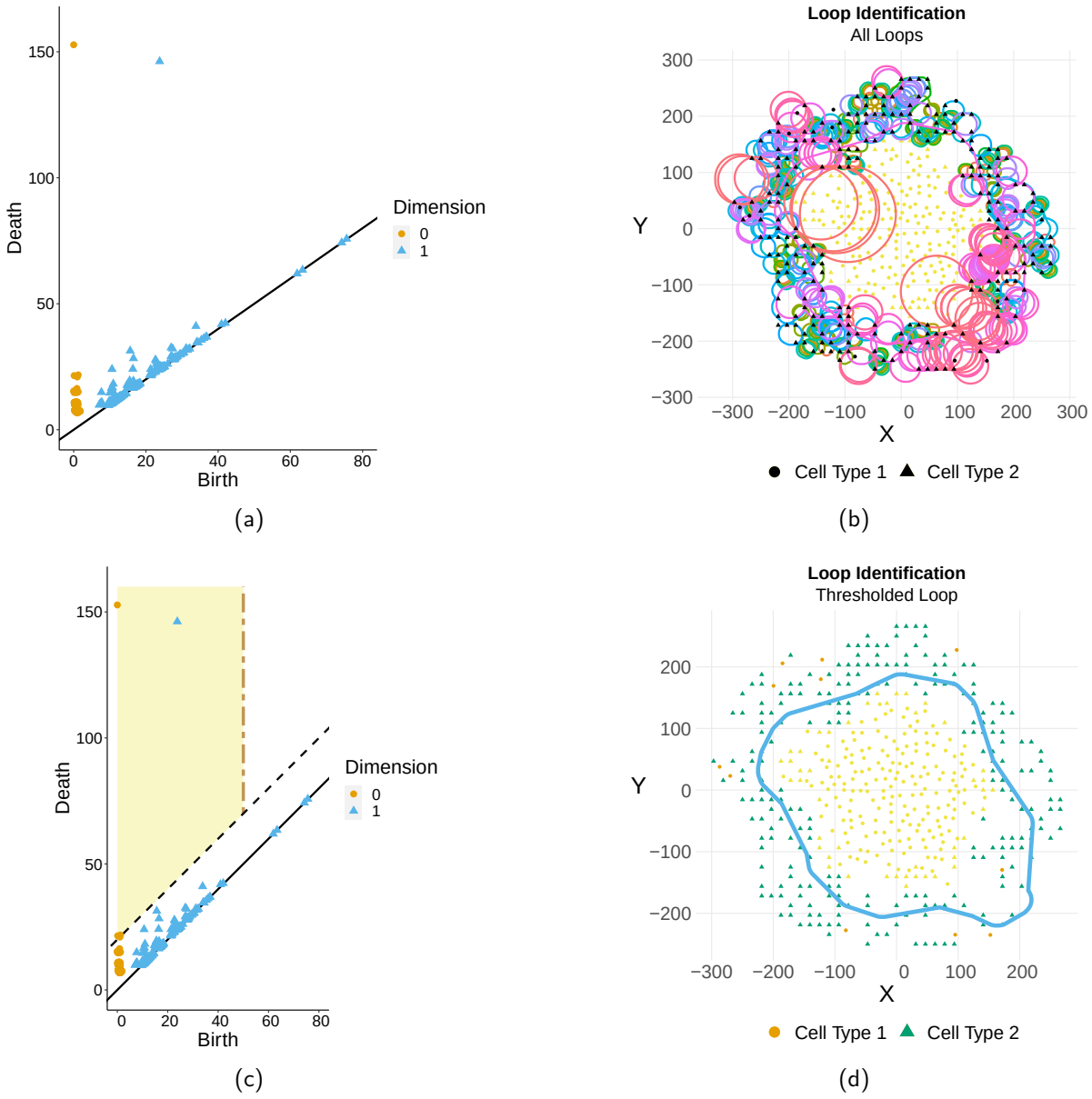


Figure 3.10: Loop identification for the proportional point pattern and the intermediary computation steps. (a) PH is computed on the area surrounding the aggregation, excluding the points within the aggregation, and identifies all the features of the 0-th dimensional homology group (orange circles) and 1-st dimensional homology group (blue triangles). (b) Excluding the points within the border of the high-density aggregation (yellow circles/triangles), we construct all loops of the Cell Type 2 point cloud which represent the midpoint of vertices that construct a cycle. (c) A threshold is applied for persistence, $\tau = 20$ (black dashed line) and a threshold is applied for the minimum vertex-vertex distance (gold dash-dot line) between points which construct the cycle, obtaining an area (yellow region) in which possible loops are considered, and (d) is the identified loop (blue line) for the set of applied thresholds, which has been confirmed to encompass the coordinates of the original high-density aggregation (yellow circles/triangles).

complimentary cell type.

We have considered the radial symmetry in the density of a cell type within the immediate perimeter of a high-density aggregation. Specifically, in Figure 3.10 we have considered the point cloud generated by the coordinates of Cell Type 2 within 100 microns of the aggregation border. Besides setting the thresholds for Birth and Persistence, we have only these parameters to fix for analysis

For the former, Cell Type 2 has been chosen since we have prior knowledge about which cell type might be inside the high-density aggregation compared to outside. This is similar to the assumptions we can make in the TRACTISS data set. However, the approach used is agnostic to cell type making it available to a wider range of contexts to identify general structured cellular neighbourhoods of many cell type relations.

The latter parameter, i.e. 100 microns, relates to the scale at which a loop can be considered "close" to the border of a high-density aggregation. Although "close" is a subjective property, this parameter can be inferred from spatial correlation metrics discussed in a previous chapter[1].

3.5 Implementation

All scripts to produce the following results have been written in the coding language R [159]. The scripts have been compiled into a library which we call *TSUNAMI*. The acronym TSUNAMI stands for "Topological and Spatial Utensils for Neighbourhood Analysis in Multiplex Images".

The core functionality of the library is a series of algorithms that compute the metrics discussed in this chapter. The library has explicit dependencies on the sf [160, 161], RSMLM [198], polyclip [163], TDA [199] and ggplot2 [165] libraries.

In our implementation of the ToMATo clustering algorithm, we utilise the RSMLM library [164] (which stands for R package for pointillist analysis of single molecule localization microscopy), in the R coding language [159]. Here, the detection density estimates are cal-

[1] See Chapter 2, Section 2.7

culated by counting the number of other detections within a specified radius. This is then implemented using the R package DBSCAN which uses the C++ library ANN to employ a k-d tree framework for efficient computation. Density modes are then calculated using a mode-seeking approach and a Vieterois-Rips graph [200].

The ARI calculation was implemented using the mclustcomp library [201].

The concave hull algorithm was implemented using the sf library [160, 161] which by extension ports the GEOS library [138].

The Density Gradient algorithm was a novel contribution as part of the TSUNAMI library, utilising the offset algorithm from the polyclip library [163] and base functions in the R coding language [159].

The Loop Identification algorithm was a novel contribution as part of the TSUNAMI library, with reliance on the TDA [199] and sf [160, 161] libraries.

The computations described were performed using the University of Birmingham's BlueBEAR High-Performance Computer service, which provides a High-Performance Computing service to the University's research community.

3.6 Results

The following sections evaluate the Structured Cellular Neighbourhood workflow. The first part will look at a sample from artificial data in an attempt to emulate the structure and complexity of TLSs. Then, the combination of algorithms, which have been tested individually so far, can be validated in combination with data in which the underlying distribution is known. Following this, we consider a sample from the TRACTISS data set, which has been validated externally by experimental collaborators.

Whilst developing the Structured Cellular Neighbourhood workflow, isolated techniques were applied to both data sets for initial consideration. Specifically, two earlier approaches were considered. The first, considered the density of T cells and B cells in combination, using a density threshold defined by a trained operator on a subsection of the data and then

reapplying the density threshold on B cells. The second applied ToMATo similarly with the threshold determined by the mid-point of the largest difference between homological features in persistence. However, both methodologies were considered inadequate by external collaborators since extraneous immuno-aggregations that were identified as germinal centres did not present surrounding T cells. This was due to neither methodology possessing a technique to verify this essential property.

3.6.1 Artificial Data Set

In this section, we will provide some context to the construction of the artificial data set before evaluating the Structured Cellular Neighbourhood workflow.

The artificial data set has been constructed based upon principles discussed in the evaluation of the most appropriate boundary identification algorithm[1]. That is, we have three underlying distributions for the structures we are identifying. Instead of individual complex geometries, for each sample, we consider the combination of these structures in a single-point pattern. Therefore, we obtain structures that combine characteristics to become more complex than they are individually. Finally, the dimensions of the artificial structures are selected from a distribution of TLS area as reported by Barmpoutis et al. [73].

Once the underlying dimensions of a structure are fixed, similar to the boundary identification examples, we sample points for each cell type in a respective area. In total there are three types of points in this data set: Cell Type 1, for the core of the artificial TLS structure; Cell Type 2, for the cell type surrounding high-density aggregations of Cell Type 1, and; Cell Type 3, all other cell coordinates which we assume follows CSR. For Cell Type 1, we sample points from a uniform distribution at a density of 0.0128 cells per micron squared within a TLS compared to 0.00032 cells per micron squared outside a TLS. This was based on the observed density of TLSs from Barmpoutis et al. [73], as well as observations by experimental collaborators for the proportion of specified cells inside versus outside a TLS. For Cell Type 2, points were sampled from a uniform distribution within the area which is 25 microns from the

[1]See Section 3.4, Subsection 3.4.2

original core TLS boundary at 0.0128 cells per micron squared, as previously established. Otherwise, Cell Type 2 points outside the TLS were sampled at 0.00032 cells per micron squared, similar to Cell Type 1. Finally, Cell Type 3 was sampled from a uniform distribution in the whole domain at 0.006 cells per micron squared to emulate the overall density of cells in the remaining domain to match approximately three times less than within the TLS as reported by Barmpoutis et al. [73].

To remain realistic to the physical realities of the biology, all cells within 5 microns of each other were removed. This affects the prescribed densities but has been measured to remain within the standard deviation observed in Barmpoutis et al. [73].

The combination of this approach is a point pattern similar in characteristics to the TRACTISS data set, where an example is shown in Figure 3.11. Here, a whole slide image, similar to that in Figure 3.2, presents the point pattern with cell type overlaid in Figure 3.11a, while the ground truth for the each TLS core is presented in Figure 3.11b. Depending on the seeding of artificial centres, up to 9 TLS cores can be assigned to the domain. If multiple cores, such as those which constitute TLS core 6, we combine their areas to construct a single TLS core which can enable more complex structures than if considered individually.

Once computed, each point pattern was analysed using the Structured Cellular Neighbourhood workflow with the same set of parameters for each stage which can be found in Appendix B (Table B.3, Table B.4, Table B.5, and Table B.6 for parameters of the respective stages of the workflow). An example of the results produced by the workflow is presented in Figure 3.12. A focused view on all the identified structured cellular neighbourhood centres is also provided in Figure 3.13.

In Figure 3.12, we can see that of the 8 artificial ground truth germinal centres the structured cellular neighbourhood workflow has identified all of them. Therefore, when comparing the identification of germinal centres at a scale of counting, we have a 100% accuracy. We continue to look at smaller scales of resolution to assess the workflow's performance and to understand the level of uncertainty when reporting on these statistics.

A post-identification stage looks at the number of cells in respective structured cellular

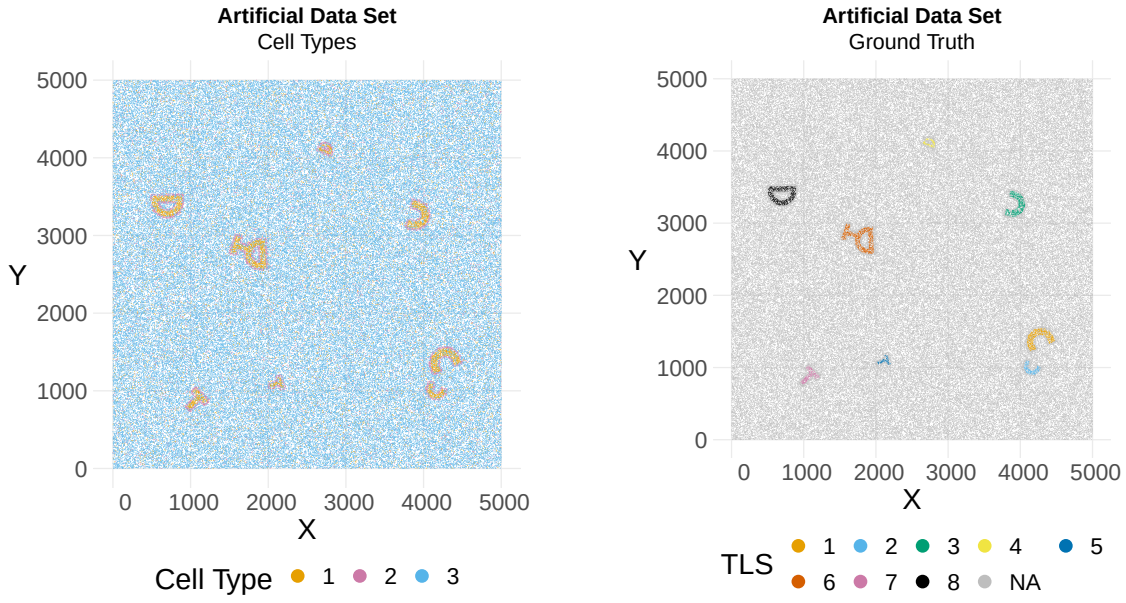


Figure 3.11: An example of the artificial data set. (a) Presents the cell types present. (b) Presents the ground truth for the cores of each TLS.

neighbourhoods. For this type of analysis, we can report on the ARI previously described as an evaluation score for clustering algorithms[1]. For correctly identifying Cell Type 1 within the centre of an artificial germinal centre the ARI score is 0.985. The minimal error in ARI score can be observed in the focused view of individual structured cellular neighbourhoods in Figure 3.13. Where there are instances of error, Cell Type 1 coordinates are more dispersed between Cell Type 3 coordinates which present as localised densities lower than the rest of the structured cellular neighbourhood. Furthermore, this appears more profound in the structured cellular neighbourhoods with the smallest underlying area, such as for indexes 1, 2 and 4.

Another resolution of analysis is in monitoring the properties of the discrete boundaries for a structured cellular neighbourhood. This can include the area, cell types within versus outside the boundary and the localisation of the structured cellular neighbourhood to other features of the microenvironment, an example in Sjörgen's disease being secretory ducts. Accordingly, we evaluate the accuracy of the borders identified by the workflow. We compute the evaluation scores applied in validating the boundary identification algorithm[2]. The average

[1]See Section 3.4, Subsection 3.4.1
[2]See Section 3.4, Subsection 3.4.2

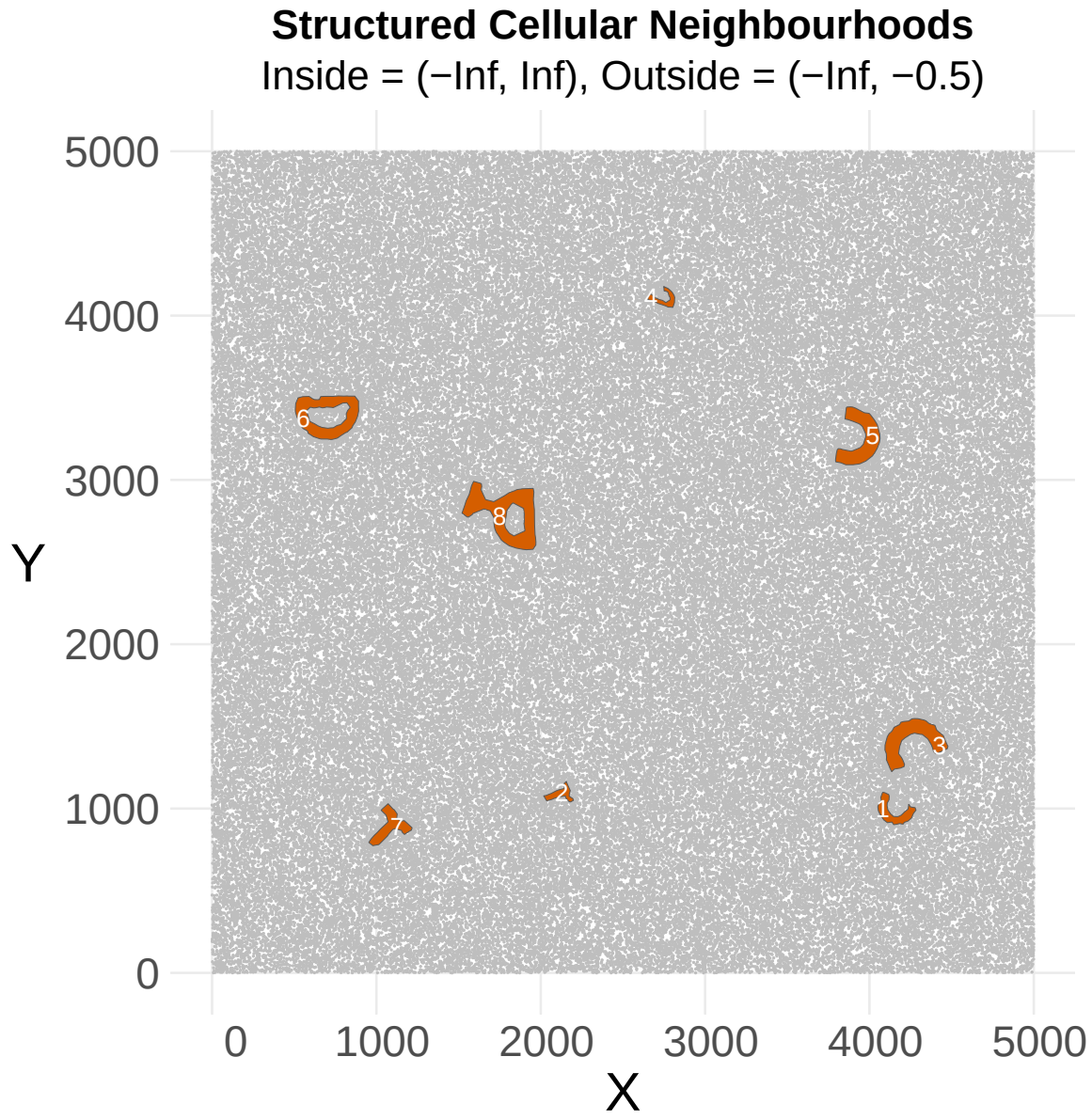


Figure 3.12: The results of the structured cellular neighbourhood workflow on a sample from the artificial data set. The core of the structured cellular neighbourhood is highlighted in orange with an index overlaid. The threshold for the inside and outside density gradients are presented, respectively.

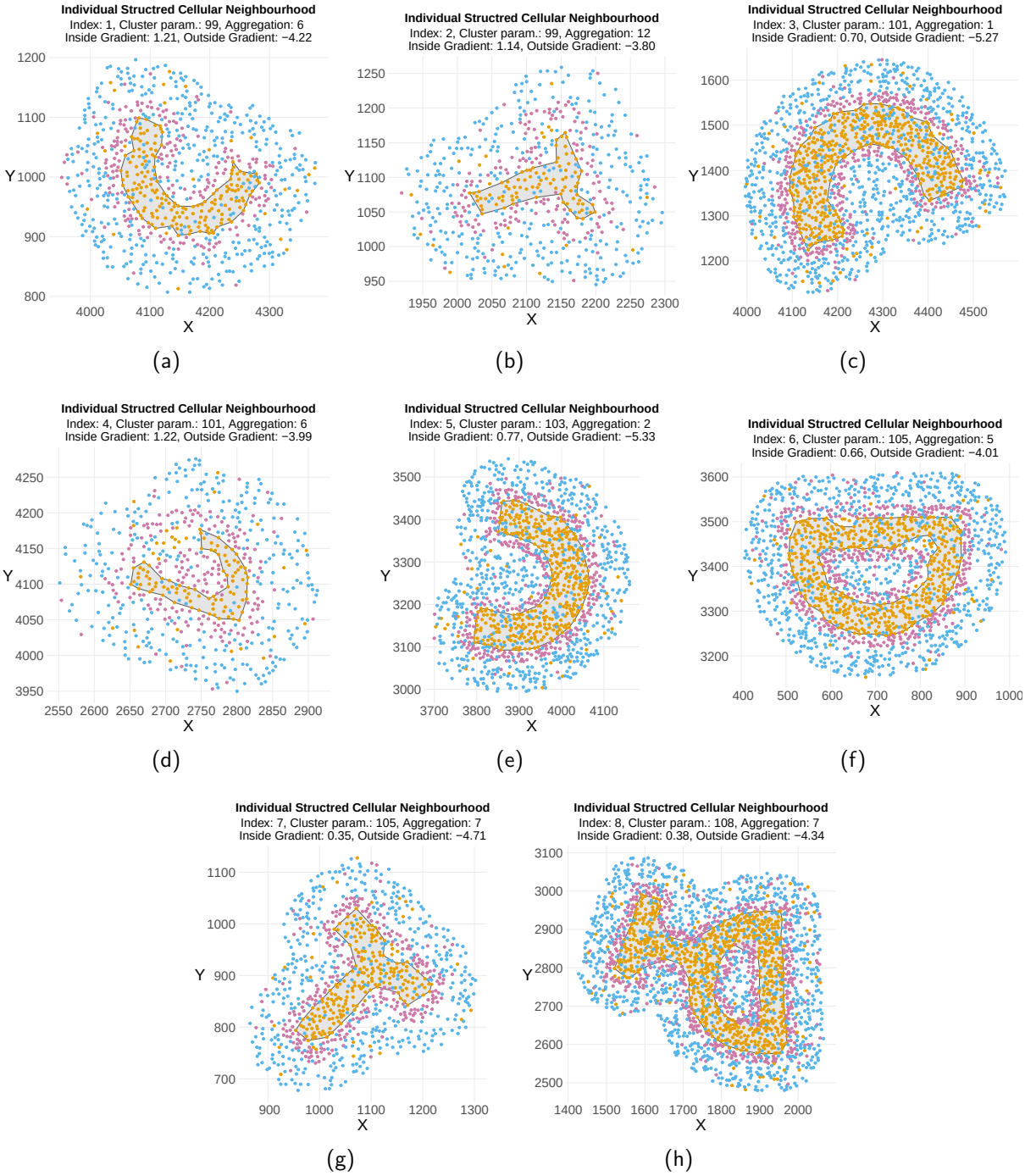


Figure 3.13: Individual structured cellular neighbourhoods identified by our workflow. The high-density aggregation boundary is outlined. Only points within 100μm of the aggregation boundary are shown. 'Index' refers to the Index given to the structured cellular neighbourhood in Figure 3.12. 'Cluster param.' refers to the set of cluster parameters used to identify the specific aggregation. Generally, higher values of 'Cluster param.' refer to larger values of the threshold followed by the radius considered for nearby points in the ToMATo clustering algorithm[4]. The 'Aggregation' refers to which aggregation identified for the given 'Cluster param.' is presented. Finally, the observed values for the density gradient for inside, i.e. for Cell Type 1, and outside, i.e. for Cell Type 2, are shown.

F_1, precision and recall scores for the individual boundary identifications are 0.891, 0.822 and 0.984, respectively. The difference in underlying distribution area and identified boundary area for individual aggregations are displayed in Figure 3.14. The individual boundary evaluation scores are presented in Appendix B (Table B.7).

The identified boundaries with the lowest boundary evaluation scores are index 4, 2 and 1, with F_1 scores of 0.740, 0.806, and 0.908, respectively. These structured cellular neighbourhoods have already been recognised for lower clustering detection when evaluated at the cellular level. As a result, this has had an impact on the delineation of the structured cellular neighbourhood's boundary. The principal reason for this deficit relates to the scale of the underlying distribution. This could be viewed as a limitation of the density-based clustering algorithm to identify aggregations at the minutia when local density is most affected by stochastic processes. Otherwise, all coordinates of the high-density aggregation could have been identified but are exaggerated by a discretisation problem, i.e. points from an underlying distribution will always underestimate the underlying distribution area, since the measurement is scaled accordingly it does not take into account the lower number of points to characterise the underlying distribution.

The instance for which the boundary identification has incorporated larger areas than the underlying distribution is in the structured cellular neighbourhood at the largest scale: index 8. The False Positive areas are located at the concave vertices of the underlying distributions. Referring to our understanding of the boundary identification algorithms we inspect Figure 3.13h for the vertices that delineate the border. Specifically, at the concave vertices there is an edge drawn between two points which cuts a portion of the area at the corner. To the human eye, it might be obvious that this is a coarse estimate of the boundary. With our understanding of the concave hull algorithm, a single parameter is defined to threshold for relevant edges in the boundary identification. This parameter is defined for all identified high-density aggregations. Therefore, a method to automatically assign the minimal edge distance for a given point pattern which generates a closed polygon would be expected to minimise this misclassification. On the other hand, the accuracy obtained from applying a single value

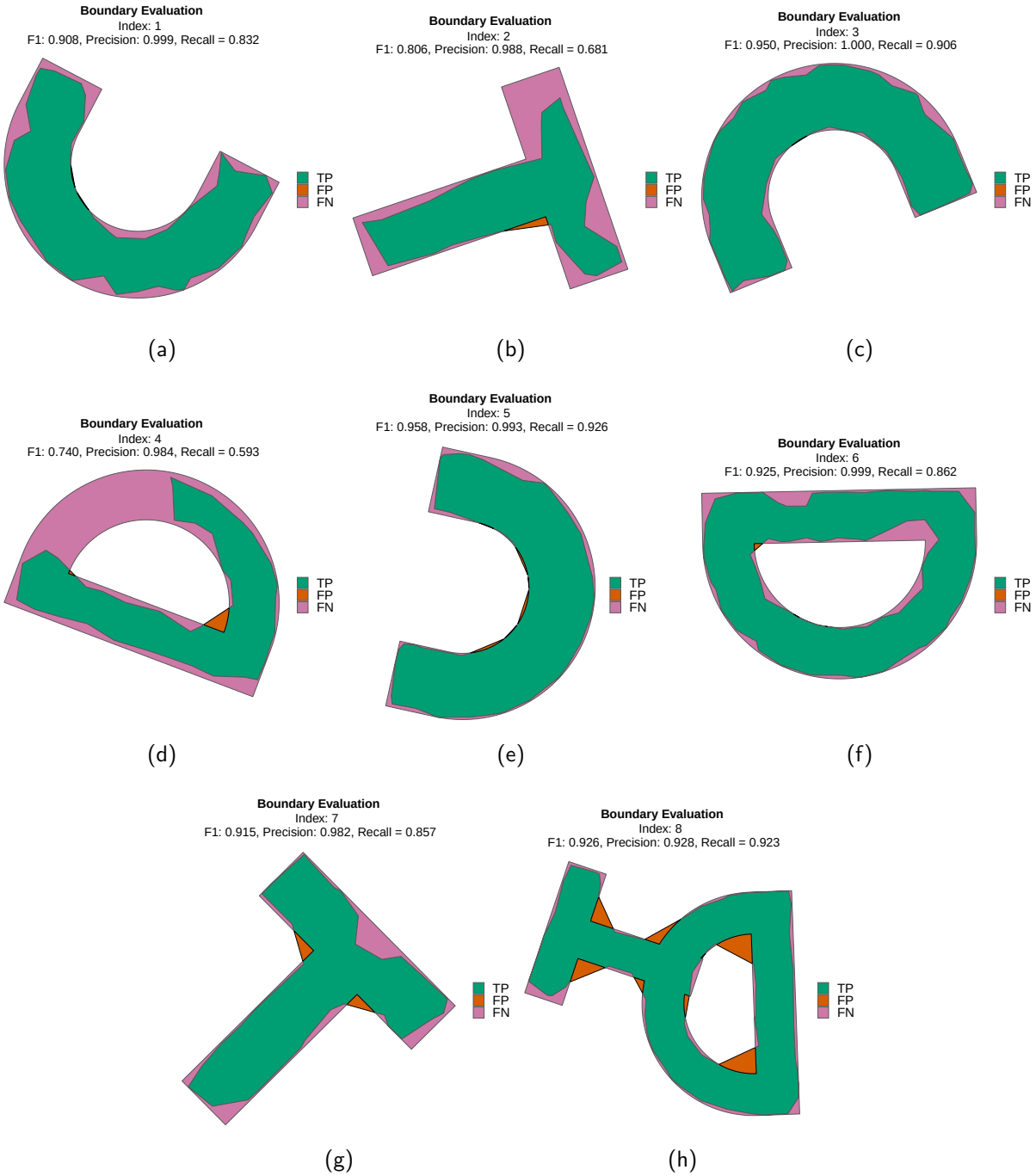


Figure 3.14: Boundary evaluations for individual identified structured cellular neighbourhoods. The F_1, precision and recall scores are presented in Appendix B (Table B.7). Index refers to the aggregation identified by the structured cellular neighbourhood workflow. TP, FP and FN stand for True Positive, False Positive and False Negative, respectively

shows the generality of the algorithm across multiple scales to still obtain accurate results.

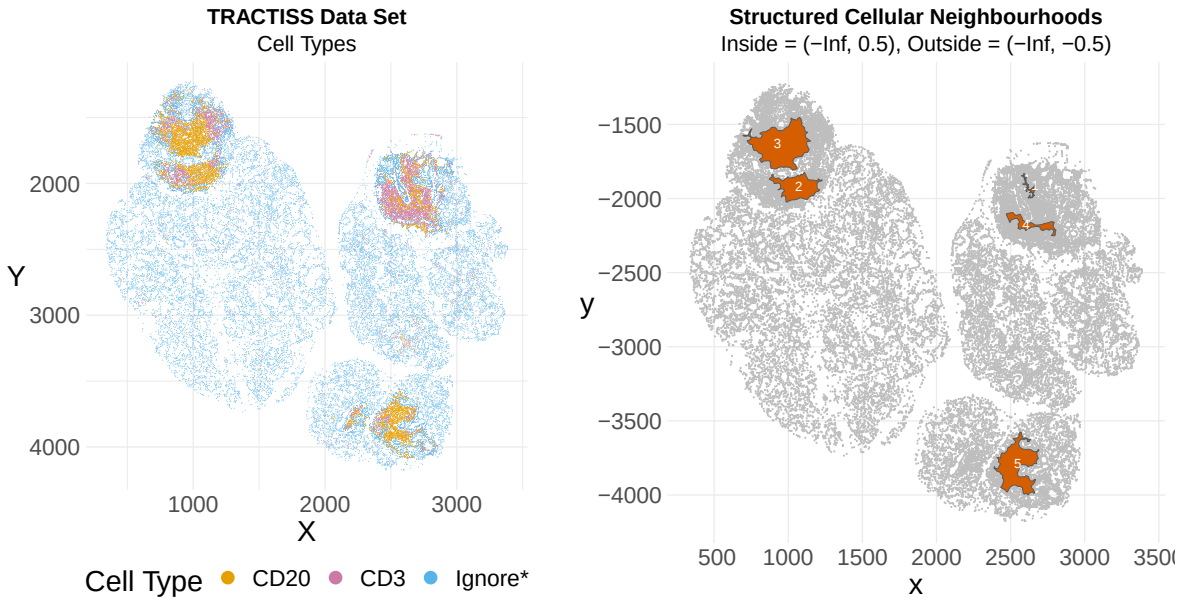
The results presented here give an indication of the workflow's accuracy and uncertainty in the identification of structured cellular neighbourhoods. To reiterate, the workflow successfully identified all structured cellular neighbourhoods in the example artificial data set, which has been repeated in other examples; the number of cells detected as associated with a structured cellular neighbourhood was also very high, and; the boundary for structured cellular neighbourhoods was established with minor misclassifications for the smallest of structured cellular neighbourhoods. Detailing the accuracy and extent of uncertainty builds confidence and appropriate scrutiny in the ability of our workflow when applied to real-world data.

3.6.2 TRACTISS Data Set

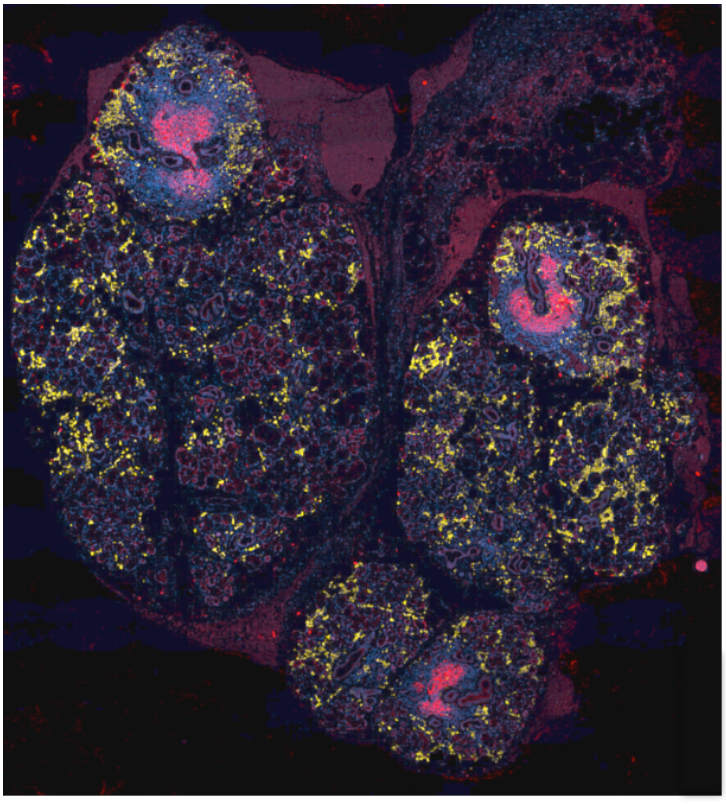
The structured cellular neighbourhood workflow has been leading up to the application in the TRACTISS data set. At the time of writing, external collaborators have been able to validate the results of the workflow for a small number of samples. As such, we discuss one of these samples in detail (namely sample N00176_001B_2) as we anticipate the validation of more samples in the future.

To begin with, we describe the validation process performed on these samples. The validation utilised confirmatory cell markers known to associate with TLS. The first is Cluster of Differentiation 21 (CD21), identifying follicular dendritic cells which are well understood to be positively spatially correlated with the germinal centres of mature TLSs [107, 178, 102, 202, 203, 179, 180, 73, 204, 132, 205]. In addition, a second marker was identified which has a positive spatial association with TLS. This marker is referred to as Cluster of Differentiation 138 (CD138) and has been observed to be on the periphery of TLSs [206, 207, 208]. The combination of these cell markers provides quantitative insight for the evaluation of the Structured Cellular Neighbourhood workflow.

The results of the structured cellular neighbourhood workflow are presented in Figure 3.15b. The results are accompanied by the point cloud in which the workflow was applied (Figure 3.15a) as well as the mIF image used to validate the results (Figure 3.15c).



(a) (b)



(c)

Figure 3.15: The results of the structured cellular neighbourhood workflow on a sample from the TRACTISS data set. Here we show the N00176_001B_2 sample. (a) The point cloud, and (b) the identified structured cellular neighbourhoods. (c) mIF image of the validation markers CD21 and CD138.

The external collaborators identify 3 TLS germinal centres in the N00176_001B_2 sample. The TLS can be observed in Figure 3.11b as in the top left of the image (TLS-1), top right of the image (TLS-2), and bottom-right of the image (TLS-3). The results of the Structured Cellular Neighbourhood workflow have identified 5 TLS germinal centres. We refer to these germinal centres with the precursor SCN, which stands for the structured cellular neighbourhood, and the index provided by the output of the workflow.

First, the identified regions in the structured cellular neighbourhood all overlap the areas which relate to a TLS germinal centre. This gives evidence that the structured cellular neighbourhood workflow is providing an automated mechanism for which structured cellular neighbourhoods are recognising true ROI. On the other hand, the results count more TLS germinal centres than confirmed by the validation cell marker. When focusing on the germinal centres (Figure 3.16) and their comparative ROI (Figure 3.17) we can begin to interpret why two of the germinal centres are split contributing to two additional germinal centres identified.

For TLS-1, the results of the structured cellular neighbourhood workflow identify two germinal centres as SCN-2 and SCN-3. For this ROI, an explanation for this occurring can be explained by the limitation in cell marker information. When looking at the area between SCN-2 and SCN-3 we see that no cells are recognised as CD20 in a gap of ≈ 50 microns between the two identified germinal centres. This is further emphasised by little to no CD3 that surround the border of the structured cellular neighbourhood, especially if the union of the germinal centres are to be considered. However, consulting the validation mIF image we see that a thin column of CD21 cells appear to connect the two identified germinal centres into one TLS. Therefore, one could conclude that based only on the markers CD20 and CD3, and without the validation markers, there would be insufficient information to indicate their association with each other.

For TLS-2, the workflow has identified two structured cellular neighbourhoods instead of a single germinal centre of a single overarching TLS. When inspecting each structured cellular neighbourhood identified, that is SCN-1 and SCN-4, there is little indication of the deficiency in the structured cellular neighbourhood and the comparative ground truth. Nonetheless, when

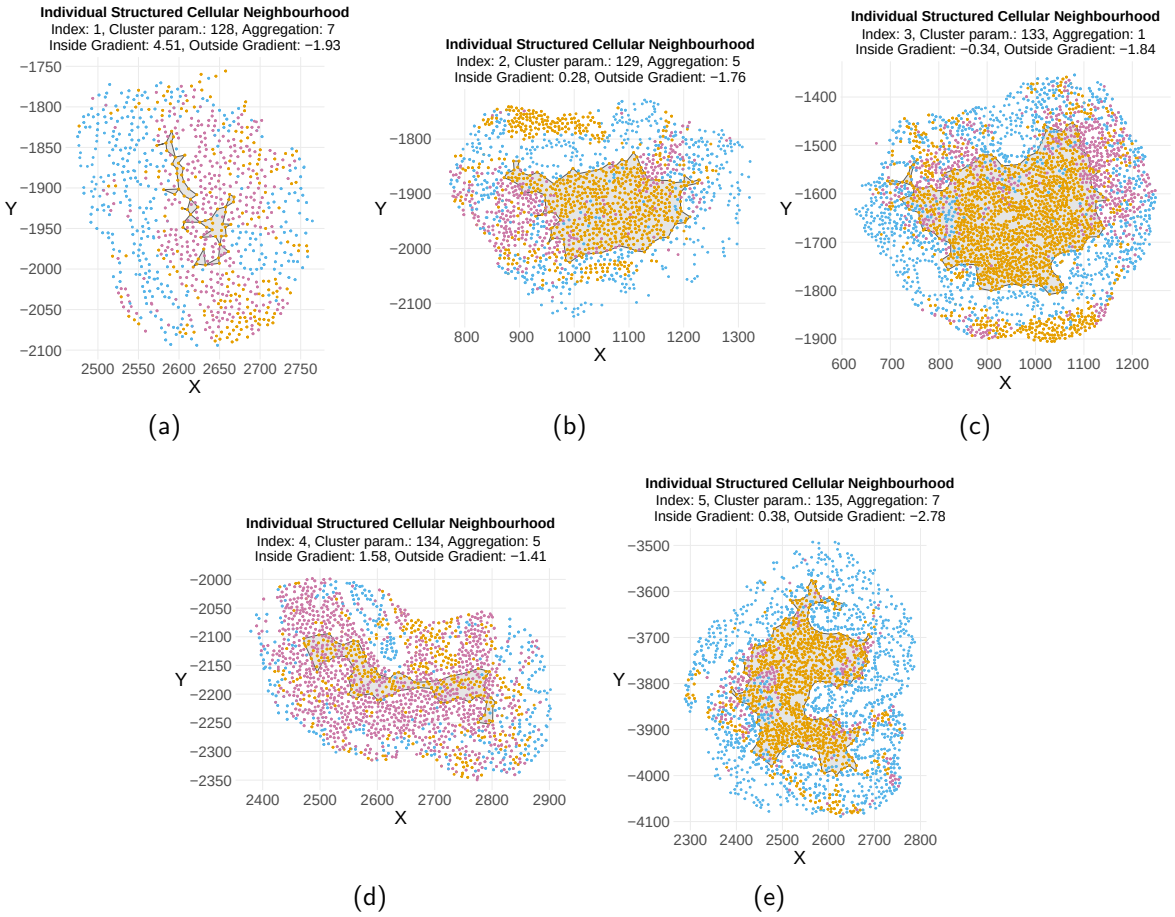
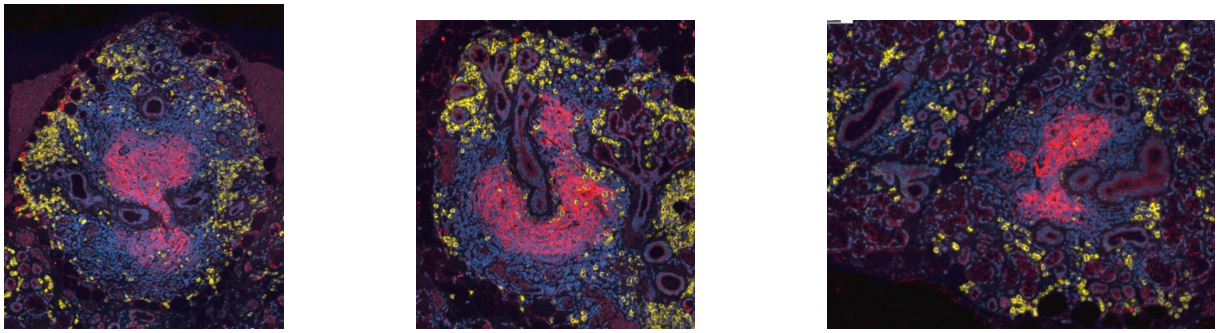


Figure 3.16: Individual structured cellular neighbourhoods identified by our workflow. The high-density aggregation boundary is outlined. Only points within 100μm of the aggregation boundary are shown. 'Index' refers to the Index given to the structured cellular neighbourhood in Figure 3.15b. 'Cluster param.' refers to the set of cluster parameters used to identify the specific aggregation. Generally, higher values of 'Cluster param.' refer to larger values of the threshold followed by the radius considered for nearby points in the ToMATo clustering algorithm[4]. The 'Aggregation' refers to which aggregation identified for the given 'Cluster param.' is presented. Finally, the observed values for the density gradient for inside, i.e. for CD20, and outside, i.e. for CD3, are shown.



(a) (b) (c)

Figure 3.17: Individual TLSs identified in sample N00176_001B_2 of the TRACTISS data set with cell markers CD21 (red) and CD138 (yellow). (a) - (c) Focused view on each TLS.

referring back to Figure 3.1, we might uncover the underlying influence image processing has on subsequent image analysis. In Figure 3.1, we are presented an example TLS. Here, the mIF image presents CD20 cells more prominently than CD3 cells. Yet, in our point cloud data, the prominence of CD20 cells is replaced by an abundance of CD3. The explanation provided is that the image processing platform used here, i.e. QuPath [186], has identified cells as both CD20 and CD3 and has kept the last tested positive association for a cell as the final cell marker associated, in this case, CD3. This is further suggested by the observations for individual image channels presented in Figure 3.1c and Figure 3.1d where CD3 is less prominent in central regions of the TLS but still present. Therefore, it is suggested here that the cell marker identification thresholds are revisited, or, a prioritisation for CD20 cells is applied when a cell is identified as both cell phenotypes.

From one sample to which the structured cellular neighbourhood workflow has been applied, we quantify the accuracy achieved and areas of uncertainty that persist in real-world data. We have provided insight into the features that can be obtained from observing two cell types and their spatial distribution to define useful structured cellular neighbourhoods. Understanding the accuracy and limitations of these results enables appropriate interpretation in downstream analysis. Further collaboration with the experimentalists will reveal the extent to which the Structured Cellular Neighbourhood workflow can identify TLS in a range of patient samples.

3.7 Summary and Outlook

3.7.1 Summary

Several techniques have been evaluated in the pursuit of constructing a workflow which can successfully characterise TLS in the TRACTISS data set. Not only this, the validation framework provided will enable others to identify the accuracy and limitations of each technique used for application in other workflows. Finally, the construction of this workflow has purposefully been designed to be agnostic to specific cell types with intuitive parameters for individuals to assign in other contexts. This means that other types of immuno-aggregations can be tested in existing mIF data sets, including the HALO1079 data set discussed in Chapter 2. As a consequence, the number, area and spatial organisation of such structures can be further analysed and compared to disease progression. For example, TLS has already been implicated for its prognostic value in various cancers. To that end, we expect the application of the structured cellular neighbourhood to aid in personalised patient care, where surgical biopsies can be analysed for structured cellular neighbourhoods of different types which indicates the need for different types of immunotherapy.

3.7.2 Future Work

We begin with an aspect of the study which is before the application of the structured cellular neighbourhood workflow: image processing. As noted previously, careful consideration for the underlying choices in the image processing platform, specifically QuPath [186] in our case, has been identified as one of the principal reasons for a TLS misclassification. Therefore, repeating this process with due diligence would be an important repeat.

As the research field advances, so do the techniques for evermore accurate nuclei segmentation. The techniques tested here are some of the most readily applied, however, researchers continue to improve on their accuracy as well as on their generality. The structured cellular neighbourhood workflow is independent of the image processing technique by considering only

the centroids of identified nuclei. Therefore, improved methodologies can easily replace the current technique with no effect on the rest of the workflow. One extension to the Structured Cellular Neighbourhood workflow would be to consider an input beyond point coordinates. improvements in cell segmentation, such as Greenwald et al. [209], suggest similar cell identification to Cellpose but with more accurate sub-cellular localisation of protein signal. With additional information such as cell boundaries, inter-cell density can be assessed taking into account differences in phenotype-specific morphology. Yet, subcellular characterisation requires cell markers for the cytoplasm, which then becomes a consideration for experimental collaborators. Accordingly, the Structured Cellular Neighbourhood workflow has the most applicability in its current form but could be advanced by the adaptation to the context of cellular morphology. Beyond image processing, we now consider initial steps for the improvement of the workflow at each stage.

The foundational stage in the structured cellular neighbourhood workflow is the use of a clustering algorithm to identify high-density aggregations. The current implementation requires the user to input a range of local density estimates, r, in our application of the ToMATo algorithm. This is a consequence of computational cost. Theoretically, the implementation would be able to consider all types of high-density aggregation and in subsequent stages discard high-density aggregation that do not meet subsequent criteria. However, a likely range of r values for which high-density aggregations are likely to exist is the current reality. Therefore, automating the choice of parameters dependent on some heuristics obtained from the point pattern would alleviate any uncertainty in choosing a range of r that misses a subset of the parameter space. A suggestion would be utilising the techniques applied in Chapter 2. In this work, exploratory tools were utilised to understand the spatial correlation between and within cell phenotypes. For example, by applying the PCF, the choice of r would be chosen to incorporate values of r for the length scales which indicate high-density aggregations within or between multiple cell phenotypes. By using this combination of complementary techniques the range of r parameters can be assigned for the workflow with more certainty.

We identify in the results of the artificial data set a limitation of assigning a single parameter

for the identification of all high-density aggregation boundaries. The analysis continues to suggest that revisiting the alpha boundary identification algorithms could take into account the scale of individual point patterns. Instead of a single parameter fixed for each high-density aggregation identified, an automated scheme could consider a range of parameters. The set of possible values is limited by the edges already apparent in the Delaunay triangulation, so a finite set of values can be considered. Then, the minimal edge length that identifies a closed polygon would be considered the 'best'. With this intuition, a natural extension in this stage of the workflow should achieve a higher accuracy in boundary identification.

The current implementation has been able to assess the number of TLSs by identifying the germinal centres. Yet, a TLS is characterised as the combination of both the germinal centre and the surrounding T cells. During this study, it has already been investigated whether the workflow could be applied to both cell types inside a single high-density aggregation. Then, only those high-density aggregations with an inside density gradient which reflects the spatial organisation of CD20 in a germinal centre and CD3 on the boundary of the aggregation will be kept. It became apparent that without a clear delineation for the germinal centre, the validation for the surrounding cell type to be radially symmetric could not be confirmed, hence the current application of the workflow. Instead, a stage could be introduced using the tools already validated in the Structured Cellular Neighbourhood workflow. This would include re-applying the clustering algorithm with the results of the density gradient aiding parameter estimation. The clustering algorithm would be employed on the cells in the vicinity of the identified border. This would then require some parameter estimation for the appropriate density and persistence of the surrounding high-density aggregation. To this effect, the results of the density gradient computation could be used to inform these parameters by using a heuristic similar to the Elbow Criterion in the k-Means algorithm but instead to identify a drop-off in the density of surrounding cells. However, instances where there appears to be a smooth transition in density from the border to the periphery would make this value difficult to ascertain. Otherwise, this could provide an estimate for identifying the whole TLS boundary, not just the germinal centre.

In the TRACTISS data set, there were two instances where split germinal centres were confirmed to be associated with the same TLS. The first was described as an absence in cell marker information. The second was attributed to prior image processing steps. Despite both instances being primarily credited to parameters outside the control of the subsequent workflow, it does raise a question in result validation. That is: if the workflow were to identify two germinal centres close to each other, then would a subsequent step consider them to be a single TLS, one as a 'true' germinal centre of the single TLS, or the germinal centres are from two very close TLSs. In the preceding text, we discussed the extension of the workflow to identify the whole TLS boundary. This could resolve the aforementioned question by identifying a shared aggregation of surrounding cell types. In this case, experimentalists would need to be consulted on the practicality of detecting two germinal centres with a shared surrounding aggregation. If both can exist as part of the same TLS then only the surrounding aggregation needs to be identified and an overlap confirmed between the two. If there can only be one germinal centre then a logical rule should be introduced to discern which one is chosen over the other. For example, the germinal centre with the largest area or the germinal centre closer to the centre of mass of the surrounding aggregation. Finally, if this should be considered two separate TLSs, as a consequence the distinction between surrounding aggregations would need to be deployed. A suggestion would be to consider an algorithm analogous to the watershed algorithm which is used in image segmentation for delineating two inseparable objects. Once the evaluation of the workflow's performance is returned from external collaborators, the necessity of this additional stage to the workflow can be revisited.

The ultimate step of our workflow uses PH to confirm the existence of a surrounding aggregation of the appropriate cell phenotypes. A recent technique to perform this more directly is referred to as Chromatic TDA [210]. In this setting, instead of considering a single phenotype, a chromatic point cloud is used as input (a point cloud with associated marks). Instead of a single persistent diagram, a sequence of diagrams is produced which refer to the homological features between those points. In this computation, features can be identified that directly reflect a 1-st dimensional homological feature surrounding a 0-th dimensional

homological feature, the feature which we intend to identify for our aggregations. When considering the methodology, it could become apparent that this technique could act as a replacement for a couple of the stages in our workflow. Yet, the features identified do not incorporate density as a criterion. Otherwise, computational expense restricts the analysis to whole slide images but restricting the domain to a localised area similar to our use of PH could prove useful. Therefore, pursuing more efficient algorithms in PH is driven by the potential of these techniques which may otherwise be restricted by computational expense.

3.7.3 Conclusion

In this chapter, we have addressed an existing challenge in biology with automated techniques from mathematics. The workflow constructed has been able to identify a template of other structured cellular neighbourhoods, namely TLS. In doing so, we have validated many techniques which relate directly to specific properties of these higher-order structures. As a result, we have accurately allocated points at a cellular resolution. This provides valuable information on the size, area and spatial distribution of these structures in disease progression. Also, the implications of this workflow go beyond the application to Sjögrens disease, with the ability to apply this workflow to other diseases as well as to other types of immuno-aggregations. The expectation is that structured cellular neighbourhoods will provide invaluable knowledge to the analysis of mIF images and subsequent treatment of patients.

CHAPTER 4

QUANTIFYING 3-DIMENSIONAL SELF-ORGANISED MICROVASCULAR MODELS AS SPATIAL GRAPHS FOR TRANSPORT SIMULATIONS

4.1 Motivation

In this study, we were given access to data from Sarah Shelton, Cynthia Hajal and Roger Kamm from the Massachusetts Institute of Technology. Their primary goal was to establish the formation of a functional network from individual endothelial cells, specifically to identify different environments that promote 'non-leaky' microvasculatures. This is where there are no holes or gaps between the endothelial cells, that form the blood vessels, to allow a perfused medium to escape. In addition, we automated their analysis to quantify the morphology of the networks under the different cell culture environments. Quantifying the morphology would provide data to inform models for the effectiveness of the microvasculature to be able to transport oxygen and nutrients to the surrounding media.

Current biological imaging techniques allow detailed reconstruction of vessel network structures. Some studies have used such data to quantify microvascular structures in a range of environments, including brain [211, 212], retina [213], heart [214], lymph node [42], placenta [215, 216] and solid tumours [217, 218, 219, 220]. Investigating oxygen and nutrient transport in these microvasculatures has been essential in understanding the evolution of diseases, such as strokes, retinal growth occlusion, atherosclerosis and subsequent cardiac arrest, immune

response, fetal growth restriction, and cancer metastasis.

Despite an active area of research, the combination of geometrical and topological techniques to characterise these blood vessel networks is sparse. In our investigation, we carry out an analysis which plans to measure all aspects of the microvasculature and then combine them into an analysis that characterises the intrinsic multi-faceted organisation of these spatial networks.

To reiterate, topology is the study of an object's properties which are invariant to continuous deformation. This tool becomes essential in understanding the microcirculation since different samples or regions might appear to have different geometrical properties, instead, they might have similar underlying principles that govern their configuration. In the following work we calculate metrics from the field of complex spatial networks and look to identify how observing differences in these properties can impact distribution and resilience in oxygen and nutrient distribution. Moreover, multi-dimensional reduction techniques are used to identify similar morphological structures from the calculated measurements and the combination of geometrical and topological metrics which identify as the best classifiers. We also consider alternative approaches, in Section 4.7, on how metrics such as the Euler characteristic and Persistent Homology have characterised the morphology from processed images.

The following chapter is presented as follows. First, a biological context for our data set is provided. Afterwards, details on the image processing techniques are explained. Then, the measurements calculated from that data set are given including the multi-dimensional reduction algorithms used in the analysis. Next, we show a mathematical derivation for flow in a network for vessels with constant cross-sectional area along each vessel. The results of the aforementioned sections are then presented before a discussion including recommendations for future work.

4.2 Biological Context

4.2.1 The Blood Vessel Microenvironment

The formation of new blood vessels, referred to as vasculogenesis [221], initiates from blood islands in a developing organism. They derive from mesodermally derived hemangioblasts which either undergo differentiation to produce angioblasts, which are the endothelial precursor cells or hematopoietic stem cells [222]. Stimulation by Vascular Endothelial Growth Factor A (VEGF) signals then fuses these blood islands to form the first primitive plexus of blood vessels in the embryonic organism [222, 223]. This primitive plexus of blood vessels undergo remodelling and stabilisation by the attachment of mural cells to the abluminal endothelial surface. In the case of capillary vessels, this involves the attachment of pericytes. Once the vessel has stabilised, the vascular network is expanded through endothelial sprouting, intussusception and collateral growth [224, 225]. Angiogenesis follows with the formation of new blood vessels from the existing vascular network, a process predominantly driven by endothelial proliferation [221, 226].

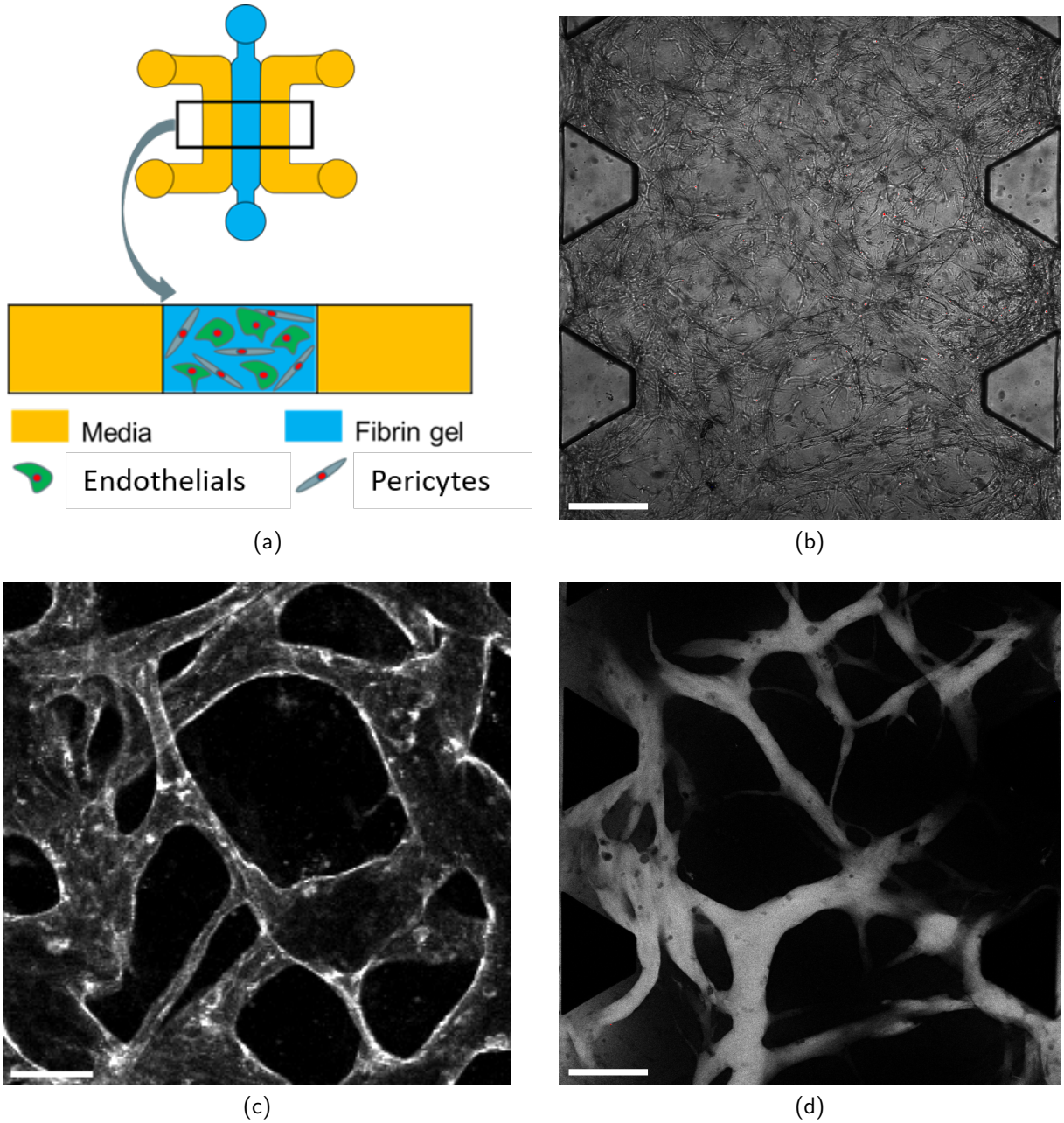
Throughout this chapter, we will also refer to a third cell type associated with the regulation of capillary vessel development, namely astrocytes. These cells are of interest specific to the microvasculature in the Blood-Brain Barrier (BBB). Astrocytes, named for their star-like shape, are understood to interact with both the microvascular system as well as neurons in the brain [227]. Loss of BBB integrity is associated with Alzheimer's disease [228, 229], Parkinson's disease [230], and multiple sclerosis [231], as well as with brain cancer [232]. Therefore, interest in the communication of astrocytes with the endothelial cells [233, 234, 235, 236] as well as their cooperation with pericytes [237, 82] has lead to interest in how to replicate these structures *in vitro* to understand the impact of treatments on BBB integrity.

4.2.2 The Self-Organised Microvascular Model Data Set

The following sections work with a data set kindly provided by Sarah Shelton, Cynthia Hajal and Roger Kamm from the Massachusetts Institute of Technology. In their experiments, microvasculatures were grown in microfluidic devices (Figure 4.1a) with different cultures of cells, including only endothelial cells; endothelial cells and astrocytes; endothelial cells and pericytes; and all three combined cell types (Figure 4.1b). This is similar to a previous study carried out by the collaborators and Campisi et al. [82]. After an incubation period, either the cells seeded were fluorescent or a fluorescent dye was perfused through the system to illuminate the blood vessel networks. The images were then captured using confocal imaging using a confocal laser scanning microscope, which images a slice of the setup at different focus depths to create a 3D volume rendering of the observed microvasculature. In this study, live imaging techniques (Figure 4.1c) were obtained at a few time points as well as perfusion experiments (Figure 4.1d).

4.3 Processing the Images

To begin with, the images described in Section 4.2.2 were cropped to only include the region between the trapezoidal structures captured in the raw images (see Figure 4.1). Once completed, image processing could follow the well-recognised stages to identify the correct pixels of an image as either part of the blood vessel or that of the surrounding fibrin gel. These consist of *pre-processing*, to increase the signal-to-noise ratio, *segmentation*, where pixels are either identified as part of the region of interest (ROI) or not, and *post-processing*, where erroneous objects are removed from the segmentation stage. All three stages are referred to as a workflow. Here we will describe two distinctive workflows that were carried out on the data. The most significant difference between the two workflows is during the segmentation stage: one uses a global threshold, and the other a machine learning software called Ilastik [238]. The first workflow was developed for fast processing of the specific images shared at the start of the investigation. The second is a workflow which is more versatile to future data



(a) (b)

(c) (d)

Figure 4.1: An illustration of the experimental set-up and data set acquired for the following section. (a) A diagram of the microfluidic set-up with an endothelial cell and pericyte co-culture. (b) Image slice captured by brightfield microscopy of the microfluidic setup from the endothelial cell and pericyte co-culture sample. (c) Image slice captured by confocal microscopy via fluorescence of the endothelium, using a lectin live staining of the luminal wall. (d) Image slice captured by confocal microscopy while the fluorescent dye is perfused through the blood vessels network. Images (b) and (d) are captured at the same time point, while (c) is captured at an earlier time point. Scale bars indicate 200 microns

sets expected to be analysed in the future. In each case, the pre-processing stages have been adapted to maximise the effectiveness of the segmentation methods. Here, we will detail the methods used in both workflows.

4.3.1 Workflow 1: Segmentation via Global Thresholding

The following workflow was carried out in Fiji is Just ImageJ 2.0.0 [239] and plugins describe coded scripts which perform specific tasks on the platform.

Pre-processing: First, we normalised the set of images. The *Enhance Contrast...* plugin was enacted whereby contrast stretching on the stack histogram was performed with a saturation level kept at the default 0.3. Due to the heterogeneity in the proportion of higher-intensity pixels between sample environments, no histogram equalisation was performed. Next, the *Despeckle* plugin was executed with the 3D kernel to remove any Salt and Pepper noise. To illustrate what Salt and Pepper noise is we consider the scenario when adding such artefacts to an image. When adding Salt and Pepper noise to an image this is defined as randomly replacing approximately 2.5% of the pixels with black pixels and approximately 2.5% with white pixels. The *Despeckle* is a median filter which replaces each pixel with the median value in its 3 x 3 x 3 neighbourhood. Median filters are a good pre-processing step to remove Salt and Pepper noise [240].

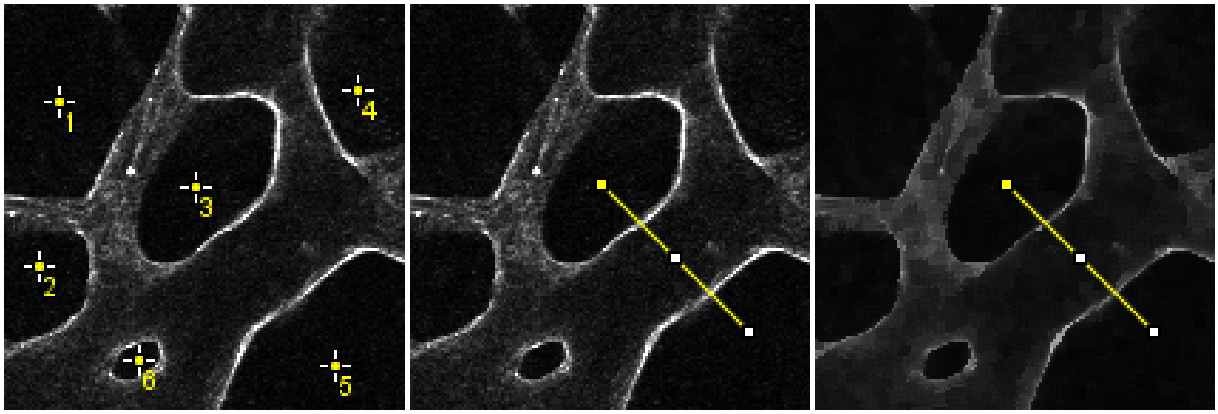
Segmentation: The *Threshold...* plugin was used to perform a global threshold on the image stack (specifically an intensity greater than 5540 for the 16-bit image type in the following data set). The quality of the segmentation was validated by a skilled operator before continuing to post-processing.

Post-processing: Two instances of the same convolution were applied for two instances of smoothing of the segmented image stack. This was carried out by the *Median 3D* plugin. The first uses the 8-pixel neighbourhood 3D kernel for the eradication of small holes, and filament elements on the boundary of blood vessels and completing connections between previously lower-pixel intensity blood vessels. The second instance considers the 64-pixel neighbourhood 3D kernel for the smoothing of whole blood vessels and to eradicate any small high-intensity

objects in the void between blood vessels. Finally, the *Analyze Particles...* plugin was used to apply a size filter on medium-sized holes in blood vessels, that have been suggested to be air bubbles, and therefore no fluorescent dye is present and only appears on specific x-y planes of the image stack.

4.3.2 Workflow 2: Segmentation via Ilastik

Ilastik is an interactive machine learning and segmentation toolkit used for batch image processing [238]. The following workflow was carried out in Fiji is Just ImageJ 2.0.0 [239] using pre-installed plugins and then additional plugins which integrate the use of the Ilastik platform.
Pre-processing: Each stage of the pre-processing discussed here is presented in Figure 4.2. Similar to **Workflow 1**, we first normalised the set of images. The *Enhance Contrast...* plugin was enacted whereby contrast stretching on the stacked histogram was performed with a saturation level kept at the default 0.3 %. Due to the heterogeneity in the proportion of higher-intensity pixels between sample environments, no histogram equalisation was performed. However, a distinction from **Workflow 1** is the use of a non-linear convolution matrix called the Kuwahara filter [241]. This is an edge-preserving smoothing filter. This was necessary specifically in the endothelial fluorescence data set, where the dye attenuated in the medium towards the lumen's centre (see Figure 4.3 for an example ROI which we present a corresponding plot profile for in Figure 4.2). Therefore, a filter was necessary to take the value of the pixels surrounding, although not blur the boundaries of the blood vessel and less so increase the intensity of pixels beyond the edge of the blood vessel. We can see that the results display a larger distinction in pixel values from the inside of the lumen compared to pixel intensities beyond the edge. Also, the filter is complementary to the features recorded by Ilastik, as discussed in the segmentation stage of this workflow. Post-application, the *Despeckle* plugin becomes redundant as what may have once been Salt and Pepper noise has been balanced out or removed by the application of the Kuwahara filter.
Segmentation: As mentioned before, we use Ilastik and respective plugins in Fiji to utilise the interactive learning and segmentation toolkit provided. The *Pixel Classification* module cal-



(a) (b) (c)

Figure 4.2: The results of pre-processing on a ROI. (a) The original ROI with indexed markers indicating voids outside the blood microvasculature. (b) The original ROI with indicated line segment. (c) The post-processed ROI with indicated line segment. The yellow line segment measures 150 microns across.

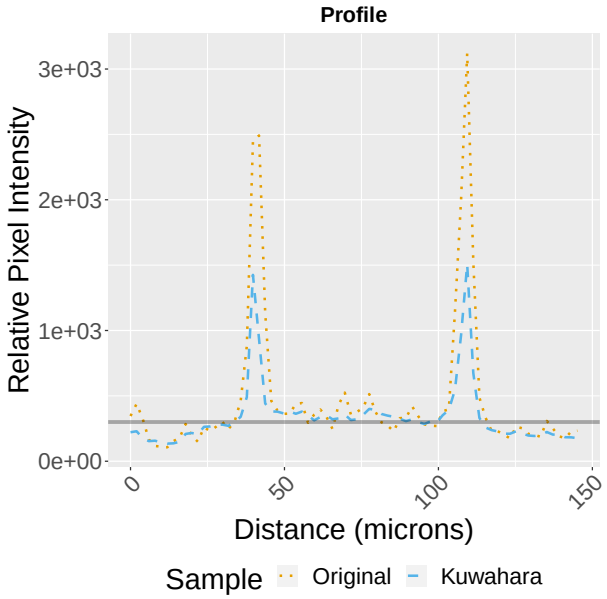


Figure 4.3: The relative pixel intensity values from the yellow line segment in Figure 4.2. Showing the original ROI and the same ROI after pre-processing. The grey horizontal line shows an example threshold where pixel intensities are separated above and below for inside and outside, respectively, after applying the Kuwahara filter but not for the line segment in the original ROI.

culates 35 classification features from an image stack to interpret information about intensity, edge, texture and orientation. We have discussed before how the Kuwahara filter has emphasised the attenuating fluorescence we wish to identify and thus increased the feature expression we would like to segment. These features help produce a classifier which is applied to whole data sets to provide a prediction for a number of ROI, in our case only one for the blood vessel or not blood vessel, for every pixel. Adjacent pixels of the same class then form our ROIs. So, unlike before, we can train the classifier to the kinds of images specifically collected for the research. In addition, the user-friendly graphical user interface (GUI) is well-suited for easy operation and limited user knowledge. Finally, there is some additional computational cost with the added complexity of the algorithm although this is still significantly faster when combined with the training annotations than normally associated with convolution neural networks, considered the next level in accuracy. As before, the quality of the segmentation was validated by a skilled operator before continuing onto post-processing.

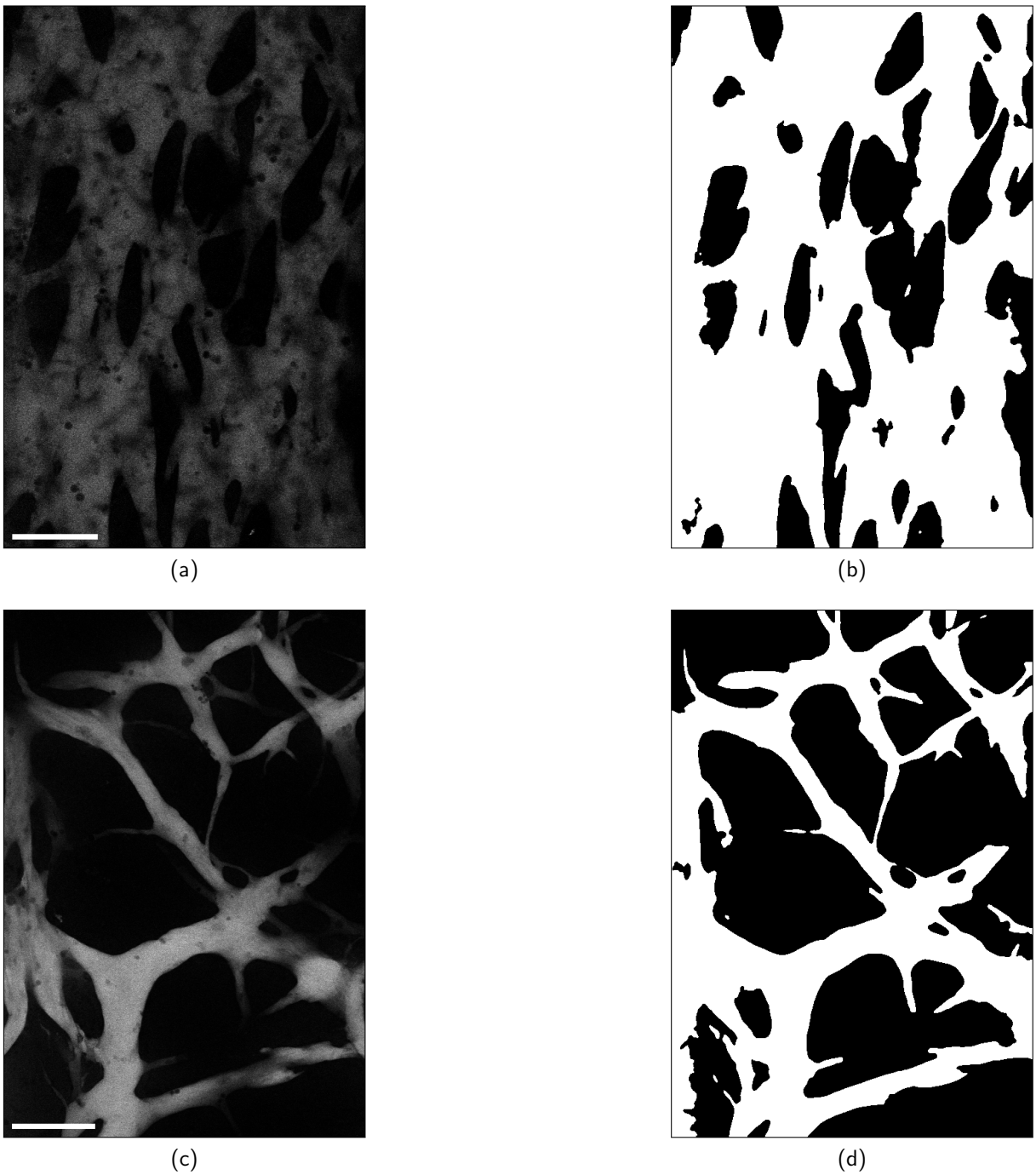
Post-processing: The same post-processing is performed as in **Workflow 1**.

4.3.3 Validation of Image Processing

Figure 4.4 displays slices taken from two samples of our data set, where we compare the raw image to the final binary image. Slices for each of the samples from the volume of the original images and the final binary images for identifying the blood vessels using the methods described in **Workflow 1** are presented in the Appendix (Figure B.3, Figure B.4, and Figure B.5 for examples slices from mono-culture, remaining co-culture and tri-culture samples). Similar results were obtained from **Workflow 2**.

4.3.4 Further transformations

The following are an explanation of further transformations of the final binary image which in combination enable the measurements for geometry and topology. The following transformations were implemented in the MATLAB coding language [242].



(a) (b)

(c) (d)

Figure 4.4: A comparison of unprocessed and processed images via Workflow 1. (a) and (c) are unprocessed 2D slices taken from the 3D volumetric image stack. (b) and (d) are the processed final binary image used in subsequent processing and analysis. White (i.e. pixel value equal to 1) is the blood vessel volume and black (i.e. pixel value equal to 0) for the surroundings. (a) - (b) EC+PC-1. (c) - (d) EC+PC-2. Scale bars indicate 200 microns.

Isotropic voxel processing: A voxel of an image is the three-dimensional version of a pixel. During image acquisition, there was a finer resolution of the microscope in the x-y plane (1.59 microns) than in the z-direction (5 microns). This was because the confocal microscope would capture a slice of the object at a fixed spacing which was larger than the resolution in the x- or y- direction. Therefore, as further measurements depended on the dimensions of each voxel being of equal size we designed a script that would transform the image stack from the original voxel dimensions to the target isotropic voxel dimensions with the `interp3` command using a linear interpolation.

Skeletonisation: This transformation describes how we obtain the centerline of a blood vessel from the binary image stack before converting it into a graph adjacency matrix. To do this we used the `bwskel` command, which uses methods from Lee et al. [243] and Kerschnitzki et al. [28].

Euclidean Distance transform: Calculates the distance from the centre of a non-zero voxel to its nearest zero voxel, and replaces the value of the zero pixel with this distance. Here the `bwdist` command was used.

Graph representation: After obtaining the skeletonisation of the ROI in the image stack, we continued to further reduce the information into a graph representation with a corresponding adjacency matrix[1]. A graph is composed of a series of nodes and edges which identify which nodes are connected. In our case, we have edges where blood vessels exist and nodes where they diverge at branching points. An adjacency matrix is simply a representation of this information with the index of the node as the row and column number, and an entry, A_{ij}, is zero if no edge connects node i to node j while either 1 means an edge joins those two edges together or a non-zero number which indicates the weight (e.g. length) that is associated with the edge between the two nodes. To construct the adjacency matrix we used the command developed in Kollmannsberger et al. [30] known as `Skel2Graph3D`. In addition to the adjacency matrix, properties of the skeletonisation were stored such as node and link structures recording pixel attribution to nodes and edges, respectively. A node has the following properties

[1] See Chapter 1, Section A.1

- idx: List of voxel indices of this node,
- links: List of links connected to this node,
- conn: List of destinations of links of this node,
- comX, comY, comZ: Respective coordinate for the center of mass in voxels of this node,
- ep: 1 if node is endpoint (degree 1), 0 otherwise.

A link has the following properties:

- n1: Node where link starts,
- n2: Node where link ends,
- point: List of voxel indices of this link.

4.4 Complex Spatial Network Analysis

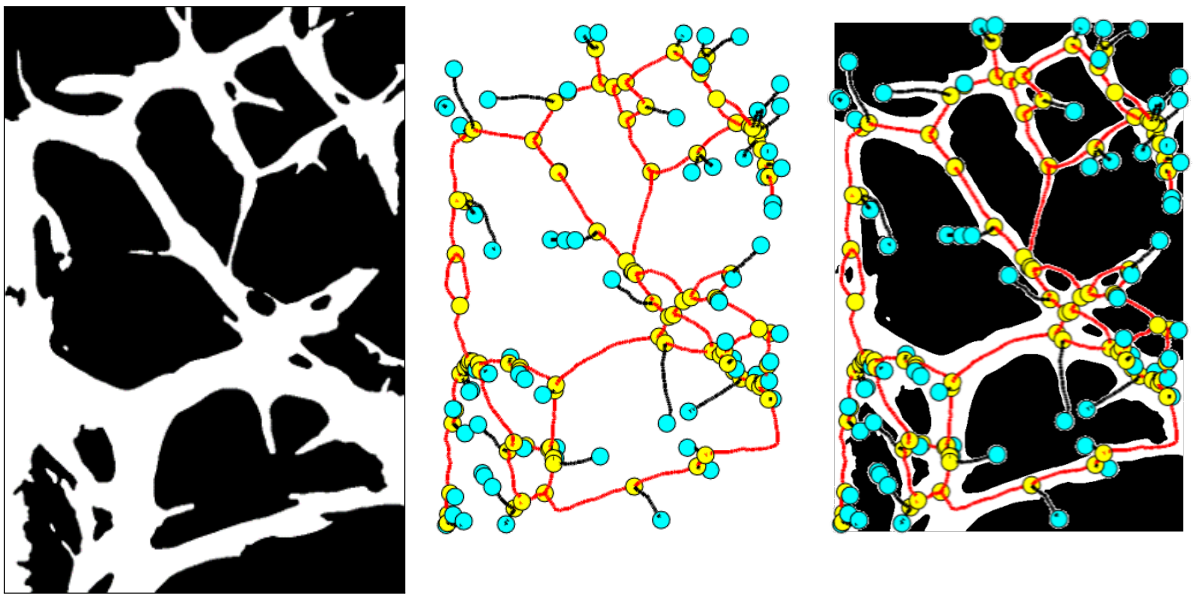
The subsequent section is a description of the measurements obtained from the processing performed in the previous section and their definitions.

4.4.1 Geometry Metrics

Here, we utilise several plugins available to us through the Fiji platform [239]. To obtain values specific to the skeleton of our respective samples, macros were created to manage the translation of plugin outputs to the centreline voxels which were then summarised.

Volume: We recorded the following measurement using the *3D Roi Manager* plugin in Fiji developed by Ollion et al. [244].

Surface area: We recorded the following measurement using the *3D Roi Manager* plugin in Fiji developed by Ollion et al. [244].



(a) (b) (c)

Figure 4.5: An example of the further transformations performed on one of the data sets, namely the Skeletonisation and annotations according to the Graph representation. (a) A slice of the original image after processing, as specified in Section 4.3. (b) The 3D skeletonisation of the whole blood vessel network with annotations according to the graph representation. Here, nodes with only one edge (coloured black) are coloured cyan while nodes with more than one edge connected are coloured yellow with respective edges coloured red. (c) An overlay of the 3D skeletonisation on a slice of the original image.

Vessel Length: The length of a vessel was calculated using the length of the link (i.e. edge) from the skeletonisation.

Vessel Radius: There were three types of cross-section considered in this study, the first two assume a circular vessel. Usually, a circular cross-section is sufficient for most data sets, however, after converting the image into isotropic voxels it became obvious that each of the samples was flatter in the z-axis than the x-y axis (see Figure 4.6). As a consequence, an estimate for the major-axis of an elliptical cross-section is also recorded. An edge is made up of many points, i.e. every pixel associated with that edge. For a circular cross-section, the minimum distance of a point from a vessel skeleton to the nearest background pixel provides a lower bound for the radius estimate.

An upper bound on the radius estimate can be provided by a metric called the *Local Thickness* [245]. Let $\Omega \subset R^3$ be the set of all points of the spatial structure under study and $\mathbf{p} \in \Omega$ an arbitrary point in the structure. We define the Local Thickness, $\tau(\mathbf{p})$, as the largest sphere which contains the point, $\mathbf{p}$, and is completely inside the structure

$$\tau(\mathbf{p}) = 2 \cdot \max\left(\{r \mid \mathbf{p} \in \text{sph}(\mathbf{x}, r) \subseteq \Omega,\, \mathbf{x} \in \Omega\}\right) \tag{4.1}$$

where $\text{sph}(\mathbf{x}, r)$ is the set of points inside a sphere with centre $\mathbf{x}$ and radius r. In the case of an ideal cylinder with infinite length, for example, all points in the structure will have the same Local Thickness. In our implementation, we only consider the Local Thickness at the points of the vessel centreline.

For the elliptical cross-section, the semi-minor axis was considered the same as the value for the circular cross-section whilst the semi-major axis was calculated as the minimum distance from every centreline pixel to the background in only the x-y plane, as the vessels are predominantly aligned in this plane. In either case, the edge is considered to have a uniform cross-section for Hagen–Poiseuille flow so the cross-section is assigned as the mean of all the points of an edge.

Vessel Tortuosity: There are multiple ways to estimate tortuosity depending on the type

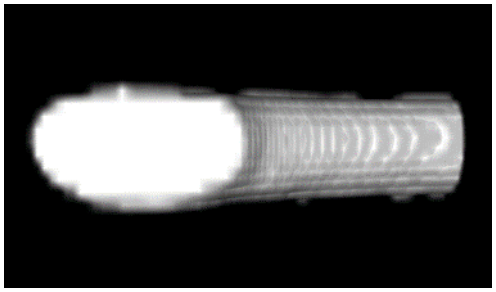


Figure 4.6: A volume rendering of a processed image (i.e. binary) where the cross-section appears elliptical rather than circular. The whitest area on the left is the cross-section and the rest of the vessel curves behind and to the right.

of curvature to be summarised, as previously discussed[1]. The simplest is referred to as the *Arc-Chord Length* or *Distance Metric*. The total length of a vessel is subsequently divided by the Euclidean distance between the endpoints of the same vessel. Therefore, a measurement of 1 characterises a straight line and a limit to infinity estimates a circle.

The Distance Metric calculates how far a path deviates from that of a straight line, its usefulness is limited in distinguishing "C"-shaped curves from a much more tortuous vessel that makes abrupt changes in direction. Bullit et al. [96] introduce a method to compute a curve's "inflection points" and combine this with the Distance Metric. For a 3D curve, an inflection point has a locus that exhibits a minimum of total curvature. In particular, the Normal and Binormal axes of the Frenet frame change orientation by close to 180 degrees as the frame passes through an inflection point. As a result, one can search for 3D inflection points by identifying large local maxima of the dot product $\Delta N \cdot \Delta N$, where N is the unit vector representing the Frenet normal axis, and N represents the difference of the normal axes associated with points $\mathbf{p}_k$ and $\mathbf{p}_{k-1}$. The Distance Metric with the number of inflections then create the Inflection Count Measure (ICM). The ICM multiplies the Distance Metric by the inflection count, using a minimum inflection count of 1. As compared to the Distance Metric, the ICM is more sensitive to oscillating curves and receives high values of tortuosity. This contrasts to a similar vessel with equal total length and distance between endpoints but instead forms a large "C"-shaped curve.

The final tortuosity measure reported in our study is the Sum of Angles Metric (SOAM).

[1]See Chapter 1, Section 1.2.

The Arc-Chord ratio nor ICM represent tight coil vessels well [96]. As an alternative, an estimate for the integral of total curvature along a curve can be computed and then normalised by path length. For any point $\mathbf{p}_k$ the vectors $\mathbf{T1}$, $\mathbf{T2}$, and $\mathbf{T3}$ are defined by:

$$\mathbf{T1} = \mathbf{p}_k - \mathbf{p}_{k-1} \tag{4.2}$$

$$\mathbf{T2} = \mathbf{p}_{k+1} - \mathbf{p}_k \tag{4.3}$$

$$\mathbf{T3} = \mathbf{p}_{k+2} - \mathbf{p}_{k+1} \tag{4.4}$$

The in-plane angle at point $\mathbf{p}_k$, α_k, and the torsional angle, β_k, are given by:

$$\alpha_k = \cos^{-1}((\mathbf{T1}/|\mathbf{T1})(\mathbf{T2}/|\mathbf{T2})) \tag{4.5}$$

$$\beta_k = \cos^{-1}((\mathbf{T1} \cdot \mathbf{T2}/|\mathbf{T1} \cdot \mathbf{T2})(\mathbf{T2} \cdot \mathbf{T3}/|\mathbf{T2} \cdot \mathbf{T3})) \tag{4.6}$$

The total angle θ_k at point $\mathbf{p}_k$ is then,

$$\theta_k = \sqrt{(\alpha_k \cdot \alpha_k) + (\beta_k \cdot \beta_k)} \tag{4.7}$$

The SOAM then considers the total curve tortuosity around an inflection point as,

$$\text{SOAM} = \left(\sum_{k=1}^{n-3} \text{CP}_k\right) \Bigg/ \left(\sum_{k=1}^{n-1} |\,\mathbf{p}_k - \mathbf{p}_{k-1}|\right) \tag{4.8}$$

4.4.2 Network Metrics

By reducing the network of blood vessels into a graph representation, metrics from network science and complex stochastic networks [100] can be utilised. During the literature review, it was revealed that the neuroscience field has developed many tools for this use. Here, we found an extensive MATLAB toolbox called Brain Connectivity Toolbox [246] which was originally

applied to brain-connectivity data sets [14], however, is transferable to our processed images. This library was used to calculate a selection of network-based measures considered relevant to transport networks both intuitively and from the literature. The following are definitions for each of the network-based measures calculated.

Degree distribution: Node degree is the number of links (i.e. edges) connected to the node.

Strength: Node strength is the sum of weights of links (i.e. edges) connected to the node.

Clustering coefficient: The clustering coefficient is the fraction of triangles around a node and equivalent to the fraction of a node's neighbours that are neighbours of each other.

Core-Periphery ratio: The optimal core/periphery subdivision is a partition of the network into two non-overlapping groups of nodes, a core group and a periphery group, in a way that maximises the number/weight of within core-group edges, and minimises the number/weight of within periphery-group edges. A version of Kernighan-Lin algorithm for graph partitioning used in community detection [247] is applied to optimise a core-structure objective described in Borgatti and Everett [248]. Coded by Rubinov, Ypma et al. [249].

Characteristic path length: The network characteristic path length is the average shortest path length between all pairs of nodes in the network. Here, length is the sum of the physical lengths of the edges.

Mean Eccentricity: The nodal eccentricity is the maximal path length between a node and any other node in the network.

Network Diameter: The diameter is the maximal eccentricity.

Maximum Normalised Betweenness: Node betweenness centrality is the fraction of all shortest paths in the network that contain a given node. Nodes with high values of betweenness centrality participate in a large number of shortest paths. Maximum Betweenness centrality may be normalised to the range [0,1] as $B/(N-1)(N-2)$, where B is the betweenness calculated and N is the number of nodes in the network. Coded by Kintali [250].

Edge Betweenness: Edge betweenness centrality is the fraction of all shortest paths in the network that contain a given edge. Edges with high values of betweenness centrality participate in a large number of shortest paths [251].

Participation coefficient: Participation coefficient is a measure of diversity of inter-modular connections of individual nodes [252].

Gateway coefficient: Gateway coefficient is a variant of participation coefficient. Similar to participation coefficient, gateway coefficient measures the diversity of inter-modular connections of individual nodes, but this is weighted by how critical these connections are to inter-modular connectivity (e.g. if a node is the only connection between it's module and another module, it will have a higher gateway coefficient) [253].

Diversity coefficient: The Shannon entropy[1] based diversity coefficient measures the diversity of inter-modular connections of individual nodes and ranges from 0 to 1 [254, 255].

Flow coefficient: Computes the flow coefficient for each node and averages over the network [256]. It is calculated as the number of actual paths of length 2 divided by the number of all possible paths of length 2 that traverse a central node. The flow coefficient is similar to betweenness centrality but works in a local neighbourhood. It is mathematically related to the clustering coefficient (cc) at each node as, $fc + cc <= 1$. Hub regions that act as bridges between different communities of nodes are likely to exhibit small clustering coefficients [257] and large flow coefficients.

Louvain Community: This metric has been defined in the Apendix[2]. The function used in the Brain Connectivity Toolbox is a fast and accurate multi-iterative generalisation of the Louvain community detection algorithm [258, 259, 260, 261, 255].

Subgraph Centrality: The subgraph centrality of a node is a weighted sum of closed walks of different lengths in the network starting and ending at the node. This function returns a vector of subgraph centralities for each node of the network [262, 263].

4.4.3 Dimensional Reduction Techniques

Here each ROI identified has 28 attributes from the intensity, geometry and topology metrics. The objective now is to combine all of this information and map each sample onto a lower-

[1] See Appendix A, Subsection A.3.2
[2] See Appendix A, Definition A.1.20

dimensional domain where we can begin to infer similar and distinct morphological features depending on the most significant metrics. For this analysis, we used two multi-dimensional reduction methods: Principal Component Analysis (PCA) and UMAP. These are explained in more detail below. The results of both analyses are displayed in the Supplementary Material, as Figure 4.8 and Figure 4.9.

PCA: PCA is a methodology to reduce the dimensionality of a data set in which there are a large number of interrelated variables. The methodology is explained in more detail by Jolliffe [264] or Goodfellow et al. [265]. The order depends on the variance in that dimension. The result is a representation whose elements have no linear correlation with each other. The axes of the representation, called principal components, provide a summary ordering the most influential factors in describing morphology. We do not reduce the dimensionality of the data, however, the technique is useful to display the most significant contributions of the variability for a user to interpret. The relationship between variance and information is that, the larger the variance carried along the axis, the larger the dispersion of the data points along it. Therefore, if the dispersion in that dimension is larger, the separation between different data and their outcome is more likely distinguishable.

UMAP: UMAP [188] uses local manifold approximations and patches together their local fuzzy simplicial set representations to construct a topological representation of the high dimensional data. Given some low dimensional representation of the data, a similar process can be used to construct an equivalent topological representation. UMAP then optimises the layout of the data representation in the low dimensional space, to minimise the cross-entropy between the two topological representations. The construction of fuzzy topological representations can be broken down into two problems: approximating a manifold on which the data is assumed to lie; and constructing a fuzzy simplicial set representation of the approximated manifold. A hyper-parameter, called number of neighbours controls how UMAP balances local versus global structure in the data. It does this by constraining the size of the local neighbourhood UMAP will look at when attempting to learn the manifold structure of the data.

4.5 Analytical Solution of Flow in a Network

The following section describes the derivation for solving Hagen–Poiseuille flow in a network of blood vessels assumed to have uniform cross-section along its length.

4.5.1 Hagen–Poiseuille flow from the Navier–Stokes equations

We first consider the fluid flow in a segment of the network. When considering fluid flow, we refer to the Navier–Stokes (NS) equations. The NS equations represent the conservation of momentum and can be understood as Newton's second law of motion for fluids. In the case of an incompressible flow,

$$\underbrace{\rho\left(\frac{\partial \rho}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right)}_{(1)} = \underbrace{-\nabla\mathbb{p}}_{(2)} + \underbrace{\mu\nabla^2\mathbf{u}}_{(3)} + \underbrace{\rho\mathbf{F}}_{(4)} \tag{4.9}$$

where $\mathbf{u}$ is the fluid velocity, $\mathbf{p}$ is the fluid pressure, ρ is the fluid density, μ is the fluid dynamic viscosity, and $\mathbb{F}$ represents body accelerations acting on the continuum. The different terms correspond to the inertial forces (1), pressure forces (2), viscous forces (3), and external forces applied to the fluid (4). The equations are solved together with the continuity equation, which for an incompressible fluid where $\frac{\partial \rho}{\partial t} = 0$, the equation of continuity becomes purely kinematic,

$$\nabla\cdot\mathbf{u} = 0 \tag{4.10}$$

which represents the conservation of mass.

In our model, the single segment is represented as a pipe with uniform cross-section with a laminar flow dependent on a given pressure gradient. This setup is widely known as Hagen–Poiseuille flow. The equation governing Hagen–Poiseuille flow can be derived directly from Eqn. 4.9 and Eqn. 4.10 by considering the system in cylindrical coordinates, with $\mathbf{u} = (u_r, u_\theta, u_z)$ as the components of flow along the (r, θ, z) axes, and the following set of assumptions

1. The flow is steady, i.e. $\frac{\partial \mathbf{u}}{\partial t} = 0$.

2. The radial and swirl components of the fluid velocity are zero, i.e. $u_r = u_\theta = 0$.

3. The flow is fully developed, i.e. $\frac{\partial u_z}{\partial z} = 0$

The radial and azimuth component of flow in Eqn. 4.9 reduces to $\frac{\partial p}{\partial r} = \frac{\partial p}{\partial \theta} = 0$, essentially the pressure p is a function of the axial coordinate z only and this component in Eqn. 4.9 reduces to

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u_z}{\partial r}\right) = \frac{1}{\mu}\frac{\partial p}{\partial z} \tag{4.11}$$

Solving for u_z, we have

$$u_z = \frac{1}{4\mu}\frac{\partial p}{\partial z}r^2 + c_1 \ln(r) + c_2. \tag{4.12}$$

where c_1 and c_2 are constants to be solved according to the boundary conditions. Since u_z needs to be finite at $r = 0$, so $c_1 = 0$. For c_2, the no-slip boundary condition at the segment wall requires that $u_z = 0$ when we consider $r = R(\theta)$ where R is the distance to the segment wall depending on the direction on the angular axis due to the different cross-sections being considered. The following will explain how we derive the Hagen–Poiseuille equation for the different cross-sections considered in our model.

4.5.2 Deriving Hagen–Poiseuille flow for circular and elliptical cross-sections

We first consider flow in a segment of pipe with a uniform circular cross-section before considering an elliptical cross-section. Derivations for alternative well-known cross-sections do exist for rectangular, equilateral triangular and isosceles triangular cross-sections, with an arbitrary result available which involves integrating a Laplace equation with a differently defined dependent variable.

Circular cross-section: The no-slip boundary condition at the segment wall requires that $u_z = 0$ when $r = R$, where R is the radius of the pipe and does not depend on θ. This yields

$c_2 = -\frac{-1}{4\mu}\frac{\partial p}{\partial z}R^2$. Thus, we have the following parabolic velocity profile,

$$u_z = -\frac{R^2}{4\mu}\frac{\partial p}{\partial z}\left(1 - \left(\frac{r}{R}\right)^2\right) \tag{4.13}$$

There are two points of importance to note from this profile. The first is that the maximal velocity occurs at the pipe centerline, $r = 0$, so

$$u_{z_{max}} = \frac{R^2}{4\mu}\left(-\frac{\partial p}{\partial z}\right) \tag{4.14}$$

and the volumetric flow rate, Q, can be obtained by integrating over the pipe cross-section,

$$\begin{aligned} Q &= \int_0^R u_z 2\pi r dr \\ &= \frac{1}{2}\pi R^2 u_{z_{max}} \end{aligned} \tag{4.15}$$

Let a circular cross-section of pipe have a specified length, L. Assuming that the pressure decreases linearly across the length of the pipe, we have $-\frac{\partial p}{\partial z} = \frac{\Delta p}{L}$ which is a constant. Substituting this and the expression from Eqn. 4.14 into the expression for Q from Eqn. 4.15, we get

$$Q = \frac{\pi R^4}{8\mu}\frac{\Delta p}{L}. \tag{4.16}$$

We rearrange for the pressure drop to give the *Hagen–Poiseuille equation*

$$\Delta p = \frac{8\mu L Q}{\pi R^4} \tag{4.17}$$

which is an analytical solution for pressure-driven flow of a liquid with kinematic viscosity μ in a single segment with length L and circular cross-section of radius R.

Elliptical cross-section: We begin by reconsidering the system in Euclidean coordinates,

where the cross-section is defined with $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. The velocity profile becomes,

$$u_z = u_{z_{max}} \left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right) \tag{4.18}$$

which satisfies the boundary condition that u_z is zero on the pipe wall and for the maximal velocity to occur at $x = y = 0$, we have

$$u_{z_{max}} = \frac{1}{2\mu} \frac{a^2 b^2}{a^2 + b^2} \frac{\partial p}{\partial z} \tag{4.19}$$

Therefore, we have the volumetric flow rate through the pipe as

$$\begin{aligned} Q &= 4u_{z_{max}} \int_0^a \int_0^{b\sqrt{1-x^2/a^2}} \left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right) \mathrm{d}y\mathrm{d}x \\ &= \frac{\pi}{2} abu_{z_{max}} \end{aligned} \tag{4.20}$$

Hence,

$$Q = \frac{\pi}{4\mu} \frac{a^3 b^3}{a^2 + b^3} \frac{\partial p}{\partial z}. \tag{4.21}$$

Combine this with the similar assumption that an elliptical cross-section of pipe has a specified length, L, and that the pressure decreases linearly across the length of the pipe, we have $-\frac{\partial p}{\partial z} = \frac{\Delta p}{L}$, a constant. Substituting this and the expression from Eqn. 4.19 into the expression for Q from Eqn. 4.21, we get

$$Q = \frac{\pi}{4\mu} \frac{a^3 b^3}{a^2 + b^2} \frac{\Delta p}{L}. \tag{4.22}$$

We rearrange for the pressure drop to give the *Hagen–Poiseuille equation* for a pipe with elliptical cross-section,

$$\Delta p = \frac{4\mu L}{\pi} \frac{a^2 + b^2}{a^3 b^3} Q \tag{4.23}$$

which is an analytical solution for pressure-driven flow of a liquid with kinematic viscosity μ in a single segment with length L and elliptical cross-section of semi-major axis length a and semi-minor axis length b.

4.5.3 Solving Hagen–Poiseuille flow in a network

The following approach is similar to a problem applied to electrical circuits discussed by Strang [266]. First, we consider the network as a set of edges meeting at nodes, with the blood vessels being the edges and the branching points (i.e. where a blood vessel splits into two or more blood vessels) as the nodes. In each instance, we can reduce the Hagen–Poiseuille equation into

$$\Delta p = C^{-1}Q \tag{4.24}$$

where C is the conductance of an edge, which for a circular cross-section equals $C = \pi R^4/8\mu L$ and for an elliptical cross-section is $C = \pi a^3 b^3/4\mu L(a^2 + b^2)$. We can then compare the understanding of blood flow in our network to a similar model of an electrical circuit, described in depth by Strang [266].

The following are the terms that will be referred to in our network and the physical relevance they have. We have a network with n nodes and m edges. Let the value of pressure for the i-th node be represented as the i-th entry in the vector $\mathbf{x} \in \mathbb{R}^n$. Also, let the value of the volumetric flow rate for the j-th edge be represented as the j-th entry in the vector $\mathbf{y} \in \mathbb{R}^m$. The conductance of an edge is stored in the square matrix $C \in \mathbb{R}^{m\times m}$, where the c_{jj} entry relates to the conductance of the j-th edge, otherwise zero. The vector $\mathbf{f} \in \mathbb{R}^n$ expresses the flow rate injected or exiting the network with the i-th entry relating to the i-th node. Finally, we have the edge-node incidence matrix, $\mathbf{A} \in \mathbb{R}^{m\times n}$, representing the connection of edges between nodes and the direction to which the flow is moving. We construct, $\mathbf{A}$, where an edge goes from node a to b, with the row associated with that edge having a -1 in column a and $+1$ in column b. Otherwise the entries are zero.

We now introduce the assumptions of the model for the flow in a network of nodes and edges with Hagen–Poiseuille flow. First, we have a conservation of mass in the system (i.e. the net current into every node is zero). We can express this as

$$\mathbf{A}^T\mathbf{y} - \mathbf{f} = 0. \tag{4.25}$$

Next, we have shown through the derivation of the Hagen–Poiseuille equation we can model the flow y_i through a segment as proportional to the pressure drop, e_j, for some constant conductance, c_{jj}, expressed as

$$y_j = c_{jj} e_j. \tag{4.26}$$

We are solving for the pressure drop across all edges, e. The multiplication $\mathbf{Ax}$ gives the pressure difference between the nodes, therefore the reversing of the sign gives the pressure drop. So, the pressure drop is $e = -\mathbf{Ax}$ across all edges which substituting into Eqn. 4.26, we have

$$\mathbf{y} = -\mathbf{CAx}. \tag{4.27}$$

Now, we have a system of linear equations that need to be solved, which in block matrix form can be expressed as

$$\begin{bmatrix} \mathbf{C}^{-1} & \mathbf{A} \\ \mathbf{A}^T & T \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{f} \end{bmatrix} \tag{4.28}$$

For block elimination the pivot is $\mathbf{C}^{-1}$. Block elimination enables us to isolate $\mathbf{x}$ from $\mathbf{y}$. To do this we multiply the top row of the system of linear equations by $\mathbf{A}^T\mathbf{C}$, here $\mathbf{C}^{-1}$ has been inverted hence it being called the pivot. So, we have

$$\begin{aligned} &\mathbf{A}^T\mathbf{C}\mathbf{C}^{-1}\mathbf{y} + \mathbf{A}^T\mathbf{CAx} = 0 \\ \iff\ &\mathbf{A}^T\mathbf{y} + \mathbf{A}^T\mathbf{CAx} = 0 \end{aligned} \tag{4.29}$$

We take Eqn. 4.29 and subtract it from the bottom row of Eqn. 4.28 to receive,

$$\begin{bmatrix} \mathbf{C}^{-1} & \mathbf{A} \\ \mathbf{A}^T - \mathbf{A}^T & -\mathbf{A}^T\mathbf{C}\mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{f} \end{bmatrix}$$
$$\iff \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{A} \\ 0 & -\mathbf{A}^T\mathbf{C}\mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{f} \end{bmatrix} \tag{4.30}$$

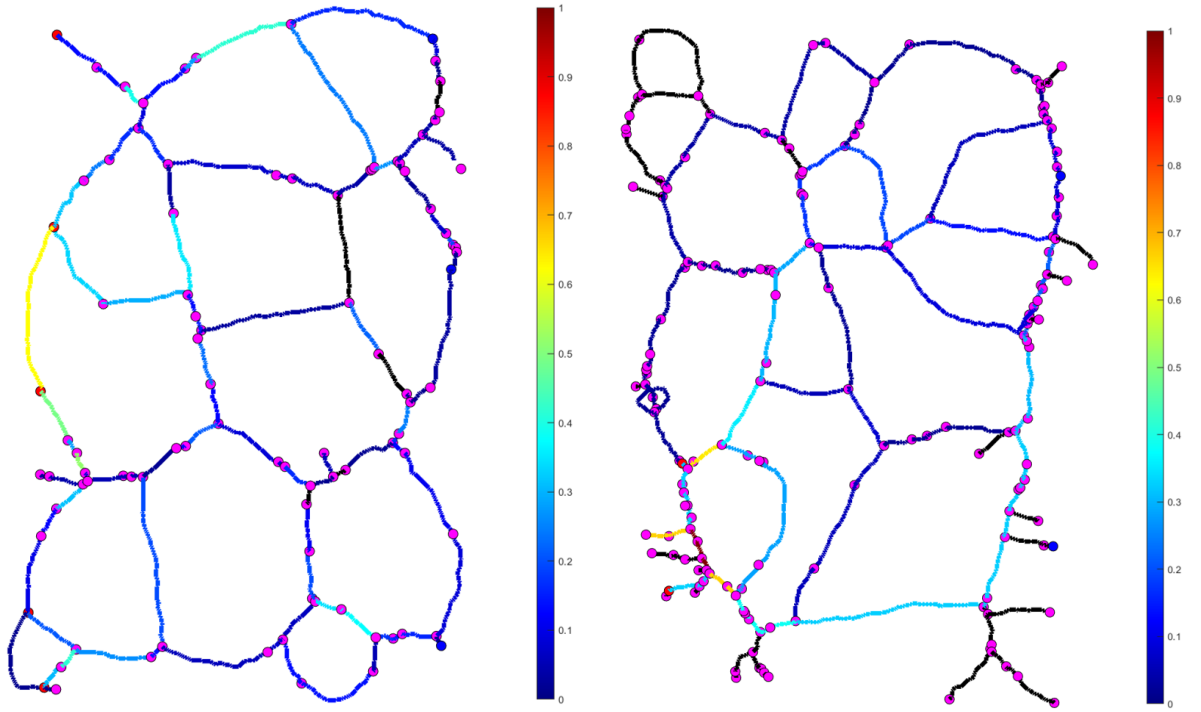
The equation for $\mathbf{x}$ alone is the bottom row, with the symmetric matrix $\mathbf{A}^T\mathbf{C}\mathbf{A}$,

$$\mathbf{A}^T\mathbf{C}\mathbf{A}\mathbf{x} = \mathbf{f} \tag{4.31}$$

We now have a means of solving for $\mathbf{x}$ only. Then, back-substitution into Eqn. 4.27 produces $\mathbf{y}$.

4.6 Results

Throughout the results of our data set, we will be referring to the samples according to their culture microenvironment followed by their index in the group. Also, since we have two co-culture groups we will distinguish the additional cell phenotype explicitly. So, we have EC, EC+AC, EC+PC and TC, where TC refers to triple culture consisting of endothelial, astrocyte and pericyte cells. Then, a suffix with a hyphen and the index refers to the specific sample in the culture group. For example, the first sample in the endothelial and astrocyte co-culture group is referred to as EC+AC-1.



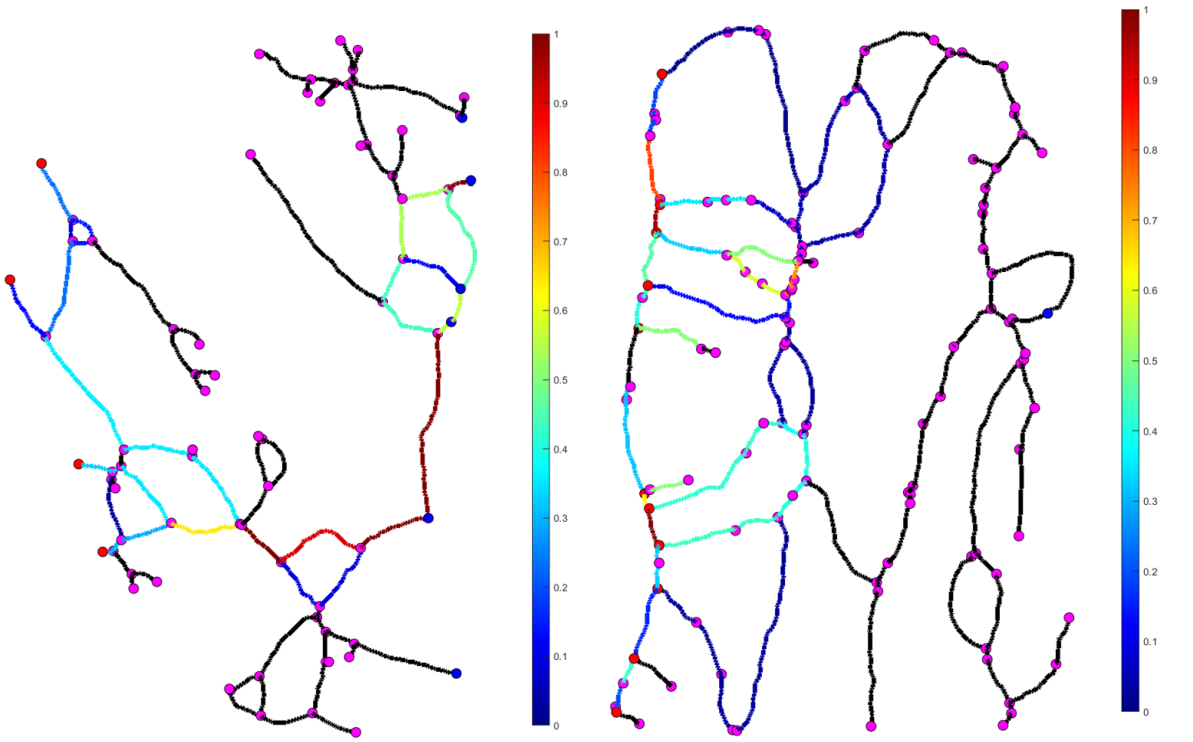
(a) (b)

4.6.1 Analytical Flow Simulations

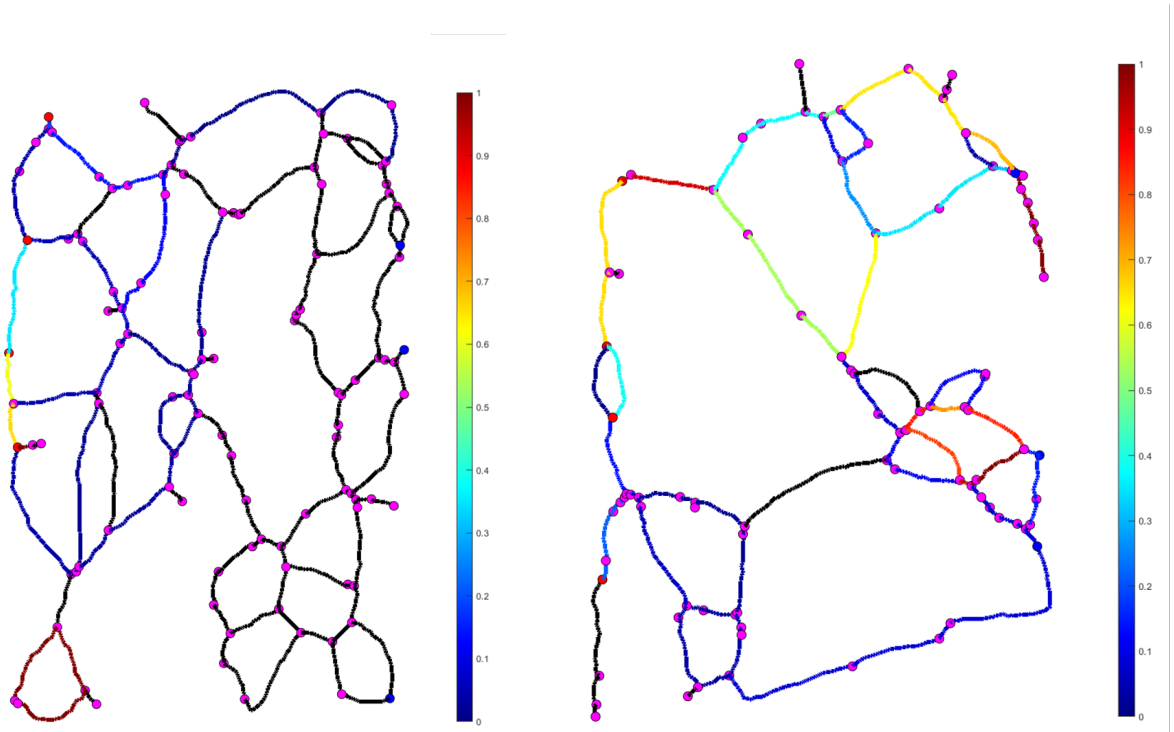
We begin with the presentation of the simulations for flow via the analytical methodology detailed in Section 4.5. For the implementation of the simulation, the conductance for every edge was set to 1. Therefore, the results presented here provide an indication of the network connectivity instead of the geometric and network properties. We expect further analyses will consider this. In this analysis, we identify inlet nodes as those which are furthest left in the network and likely correspond to vessel connections to the media compartment of the microassay, as well as outlet nodes which are furthest right in the network and likely correspond to vessel connections in the media compartment of the microassay on the opposite side.

The results of the simulation on our data set are shown in Figure 4.4 where the relative volumetric flow for each edge in our network. The distribution of the relative volumetric flow within each network is presented in Figure 4.5.

We see in all simulations that the relative volumetric flow generally has the highest relative

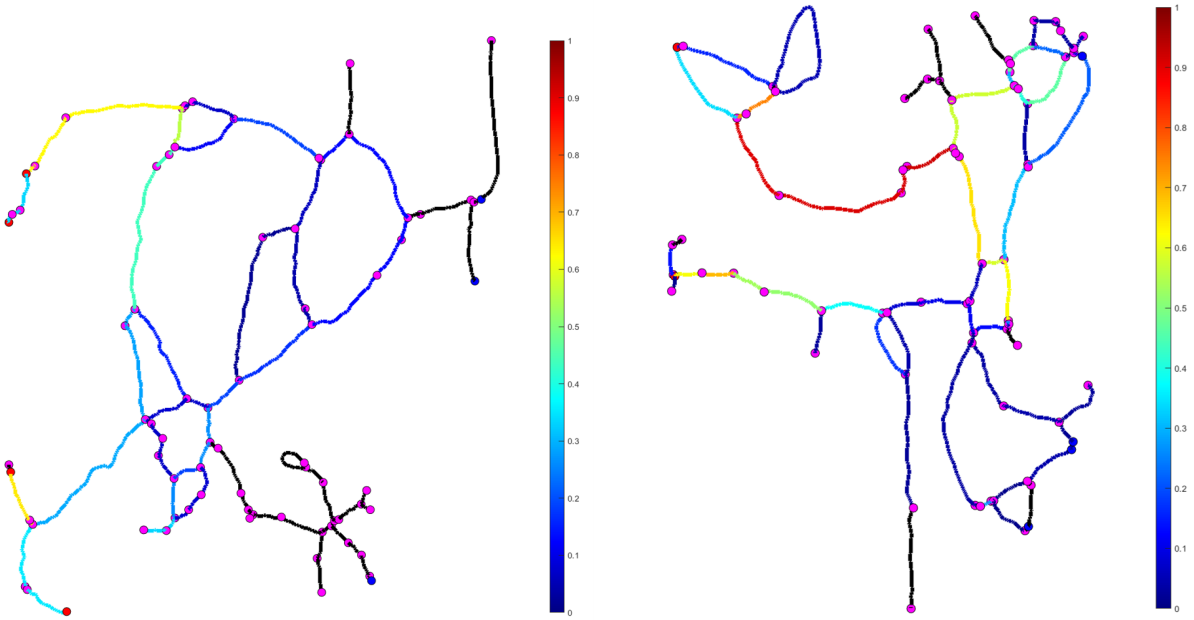


(c) (d)



(e) (f)

volumetric flow on the left of the network and a low relative volumetric flow on the right. This is a phenomenon directly related to the equal conductance of the network where a small number of edges from inlet nodes constrict the amount of flow in the system which is then proportioned at subsequent branching points. There are networks with few edges on the right half of the network, such as samples EC+AC-1, EC+PC-2, and TC-2 (Figure 4.4c, Figure 4.4f, and Figure 4.4h, respectively). In these cases, we see that the restriction of available paths from inlet to outlet results in the highest flow in the network since flow is restricted by the lack of connectivity. Reflecting on the distribution presented in Figure 4.5, we see that in all samples there is a skew towards lower values in the relative volumetric flow with few edges considered to have relative volumetric flows similar to the largest value. Looking across cell culture environments, there is no suggestion that the distribution of relative volumetric flow is indicative of specific culture environments. The most extreme example is the within-sample variation in distribution for EC+PC cultures. The distribution of EC+PC-1 has nearly all edges providing a relative volumetric flow close to zero while EC+PC-2 has a distribution which is



(g) (h)

Figure 4.4: The results of the analytical flow results on the skeleton of the network. The volumetric flow was calculated for each edge according to the analytical derivation. Each edge was then normalised according to the largest volumetric flow rate and coloured according to the `jet` colour scheme, where the warmest colours (e.g. red) being the largest flows and the coolest (e.g. blue) being the lowest with black as near zero flow rates. Nodes are coloured according to whether they are an inlet (red), outlet (blue) or neither (magenta). (a) and (b) are for the microassay cultures with only endothelial cells; (c) and (d) are for the microassay cultures with endothelial and astrocytes; (e) and (f) are for the microassay cultures with endothelial and pericytes; (g) and (h) are the microassay cultures with all three cell types (i.e. endothelial, astrocyte and pericyte).

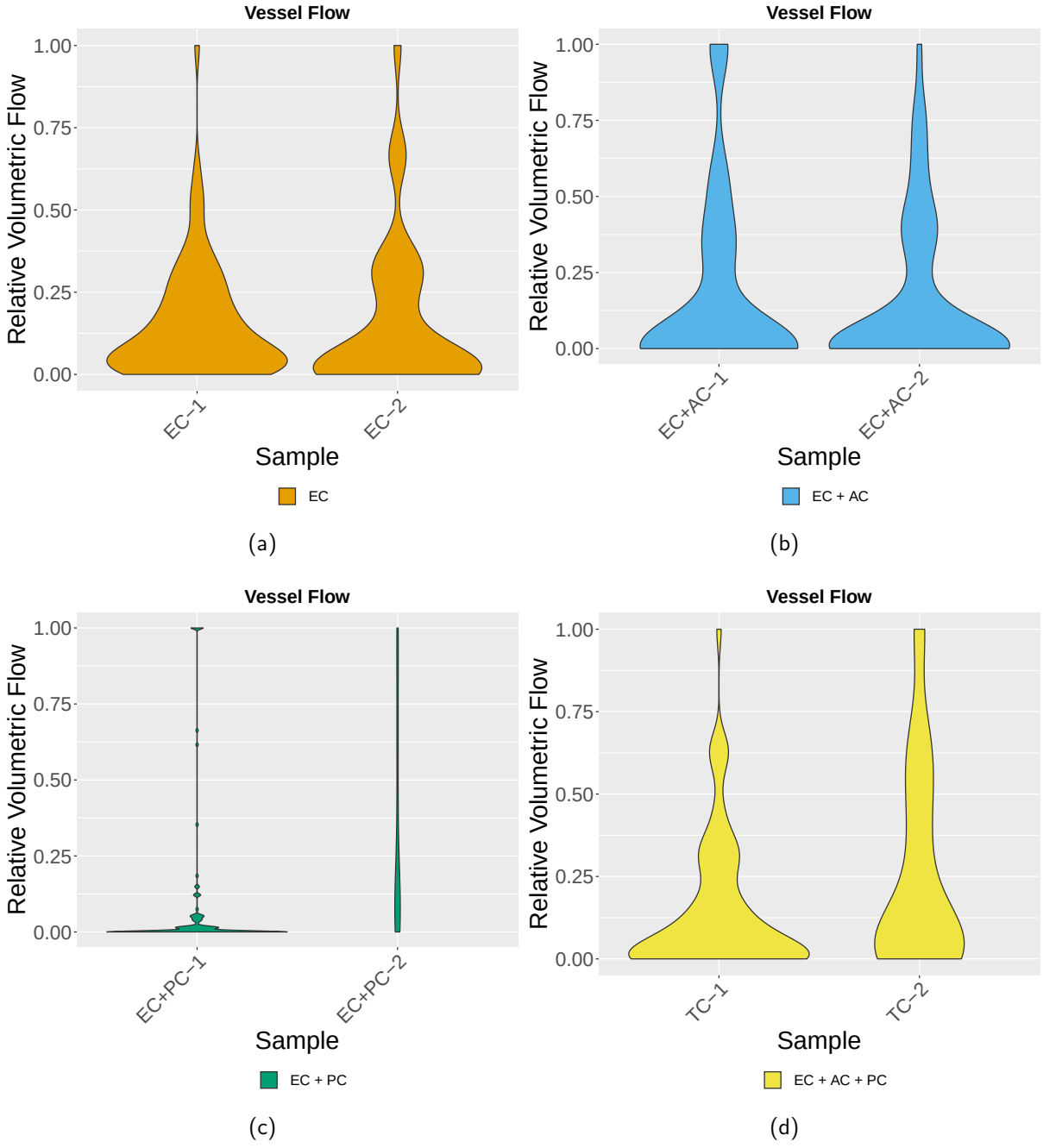


Figure 4.5: Relative volumetric flow of individual vessels defined by the skeletonisation. We have the distributions for (a) Endothelial cells only, (b) Endothelial and Astrocyte cells, (c) Endothelial and Pericyte cells, and (d) Endothelial, Astrocyte and Pericyte cells. To obtain the results as a relative measurement within each sample, all vessel flows have been divided by the maximum flow calculated for a given vessel. During this computation the conductance of every edge was set equal to 1.

less skewed to zero values. The conclusion drawn from this data set is that there is more difference within-culture microenvironments than there is difference between different culture microenvironments.

Considering the conclusion that within-group difference is greater than cross-group difference, what about samples we might consider similar irrespective of culture microenvironments? From our previous description of the data, we have already identified samples with flow dynamics which have been influenced by limited paths from inlet to outlet, i.e. EC+AC-1, EC+PC-2, and TC-2. We see that those flow distributions with bottlenecks close to the inlets are the corresponding network in each of the aforementioned groups, that is EC+AC-2, EC+PC-1, and TC-1. The final subjective class of spatial flow distributions are allocated to both mono-culture microenvironments where, similar to before, the highest values of relative volumetric flow are close to the inlets, yet edges close to the outlets present some non-zero value of relative volumetric flow. This is unlike before, where the majority of edges connected to and propagating from the outlets hold a near-zero value of relative volumetric flow. This is a relationship not apparent in the distributions of relative volumetric flow but an observation when referring to the spatial distribution of the flows. In the remaining analysis, we consider the classification of samples into the groups which match these heuristics since the culture microenvironment is either not apparent with the current number of samples or a variable which does not relate to microvascular construction.

4.6.2 Individual Metric Analysis

The next phase of the analysis investigated the possibility of a single metric to summarise the similarities or differences in flow properties. Earlier in this section, definitions have been provided for the metrics applied. They split into the geometry of the overall sample, along the skeleton of vessels, vessel length and tortuosity as well as network metrics. Tables with the summary values calculated for each of the microassay samples are presented here and in Appendix B (Table 4.1, Table B.8, Table 4.2, Table 4.3, and Table B.9, respectively). Here, we present a number of the distributions produced by all microassay samples for relevant

metrics. We illustrate with these diagrams the differences and similarities between microassay samples in determining biologically relevant properties which indicate within-culture similarities or similarities in spatial distributions of flow.

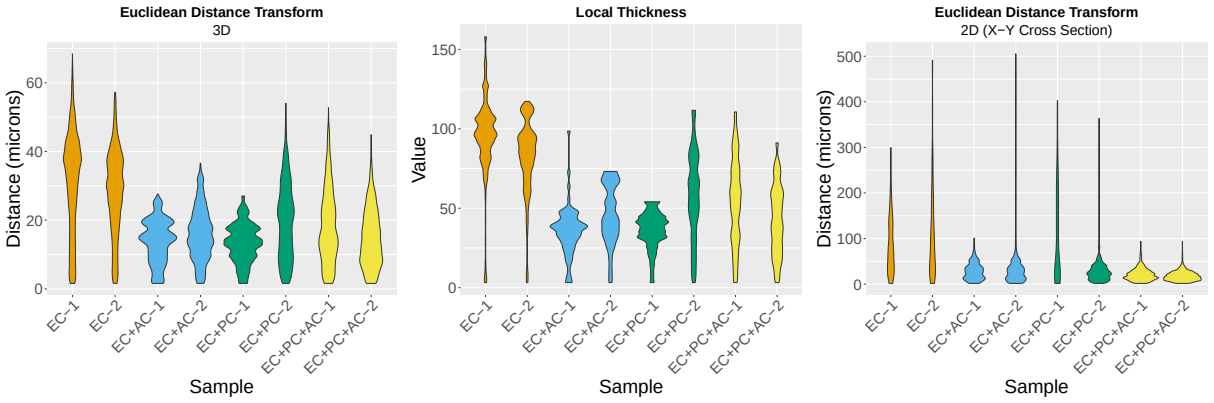
First, we consider Table 4.1 the overall geometry of the largest functional microvasculature, assumed to be the largest connected object in the microassay sample. We compare this to the remaining endothelial negative volume, also referred to as the background, to compute a ratio between the two regions of a microassay. Beginning with the volume of the microvasculature, most notably the volume of the mono-culture microenvironments produces the largest volumetric blood vessel networks. This is verified by the relatively large cross-sectional radii of blood vessels in the example slices shown in Appendix B (Figure B.3). Beyond this observation, the volume of the microvascular does not appear to relate to microenvironment similarities. The next two largest volumes are from different co-culture microenvironments. In addition, the two smallest microvasculatures are from a co-culture and a tri-culture microenvironment, respectively. On the other hand, despite the simulations of flow considering no variation in geometry, it could be suggested that the group of most similar microvasculatures result in a similar value of geometry. In this case, we see the two mono-cultures being grouped with the largest volumes followed by EC+AC-2, EC+PC-1, and TC-1 and the smallest volumes calculated for EC+AC-1, EC+PC-2 and TC-2. However, when normalised for the total volume considered, this trend does not persist. Again, the volume ratio does not appear to suggest a similar microenvironment culture produces a similar volume ratio. Yet, the indicated groupings inferred from the spatial localisation of flow from before, the groupings are less defined since EC-2 has a value within the range of the three highest volume ratios after excluding the mono-culture microassays. A similar narrative is observed between the similarity in surface area values and the samples. That is, the mono-culture microassays have the highest surface area and the same groupings thereafter but when normalised for the field of view the surface area ratio fails to distinguish similarities within microenvironment cultures or similar spatial distributions of flow. Following this analysis, the microvasculature volume or surface area suggests low classification power in either culture or the spatial distribution of flow.

Culture	Cell Phenotypes	Sample	Volume			Surface Area		
			Microvasculature	Background	Volume Ratio	Microvasculature	Background	Surface Area Ratio
MC	EC	1	5.52E+07	8.29E+07	0.67	3.02E+06	5.30E+06	0.57
MC	EC	2	4.03E+07	1.06E+08	0.38	2.71E+06	5.22E+06	0.52
CC	EC+AC	1	1.35E+07	4.20E+07	0.32	1.34E+06	2.82E+06	0.48
CC	EC+AC	2	2.91E+07	7.41E+07	0.39	2.43E+06	4.21E+06	0.58
CC	EC+PC	1	2.30E+07	5.78E+07	0.40	2.39E+06	4.38E+06	0.55
CC	EC+PC	2	1.52E+07	1.01E+08	0.15	1.65E+06	4.17E+06	0.39
TC	EC+PC+AC	1	1.62E+07	4.65E+07	0.35	1.69E+06	3.09E+06	0.55
TC	EC+PC+AC	2	1.11E+07	6.57E+07	0.17	1.51E+06	3.80E+06	0.40

Table 4.1: Table of measurements for the volume and surface area for the largest functional microvasculature, assumed to be the largest connected endothelial positive object, in comparison to the rest of the endothelial negative volume, also known as background.

The analytical simulations of flow computed assumed all vessels of the network had equal conductance. In reality, there is variation in vessel cross-sectional area. It is discussed how this information is to be incorporated into the simulations as future work[1]. Instead, we observe the similarities in radii estimates to monitor properties that the culture microenvironment might induce and/or their correlation with spatial distributions of flow. The summary statistics are presented in the Appendix B (Table B.8). In all instances, radii estimates along the skeleton of vessels are generally the highest in the mono-culture microenvironments, similar to the observation in volume and surface area from before. Both mono-culture microassays do not have the highest values of radii estimate for only one average radii estimate. Rather, EC+PC-2 has the second highest mean value in minimum distance from the vessel skeleton to a background pixel in the 2D X-Y cross-section after EC-2 but before EC-1. For all other microassay samples, the average radii estimate does not match either within-culture microenvironment similarities or spatial distributions of flow. When looking at the distributions of radii estimates in Figure 4.6, the similarities between radii estimates are subjectively visible between mono-cultures. Here, both have a higher proportion of measurements larger than all other samples with a wider range in value too. On the other hand, new similarities appear between the samples. In both the 3D Euclidean Distance Transform and the Local Thickness calculation, EC+AC-1 and EC+PC-1 appear to share a resemblance with smaller variations in measurement and maximum values which are lower than most other distributions. We also

[1]See Section 4.7, Subsection 4.7.2



(a) (b) (c)

Figure 4.6: Distributions of vessel radii estimates, including the (a) Minimum distance from the skeleton voxel to the nearest background voxel in 3D, (b) Local Thickness at the skeleton voxels, and (c) Minimum distance from the skeleton voxels to the nearest background pixel in the 2D X-Y Cross Section.

recognise a similarity in distributions for EC+PC-2 and TC-1. In contrast, these distributions in radii estimates have a larger variance than most other distributions. However, this does not persist in the 2D Euclidean Distance Transform measurements. Instead, we observe similarities within culture microenvironments than we do between, except for EC+PC-1 which has a distribution more similar to the mono-culture microassays in average value, variance and range. Once again, the average and distribution of radii estimates provide low classification power in either culture or spatial distribution of flow in different microassay samples.

Next, we consider geometry measurements performed at a whole vessel scale. This consists of distributions for vessel length and several tortuosity metrics, as shown in Figure 4.7 with summary statistics in Table 4.2 and Table 4.3. To reiterate, we are considering individual metrics and the similarities in distributions to either identify emergent properties due to culture microenvironments or correlate the spatial distribution of the flow simulated.

No pattern in vessel length appears distinguishable for either within-culture similarities or for the spatial distribution of the flow in the networks. For example, in Figure 4.7 and Table 4.2, the mono-culture microenvironments do not have average vessel lengths which are closer to one another than other microassay samples. Also, when looking at variation around the mean there is more similarity between EC-1 and TC-1 than their respective same-culture

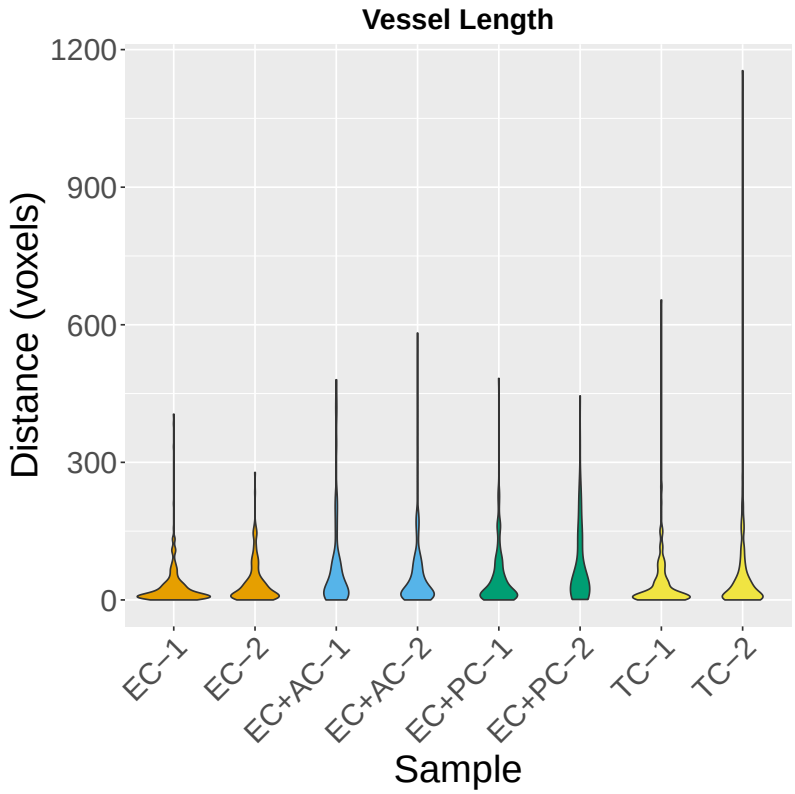


Figure 4.7: Distributions of vessel length for individual microassay samples.

Culture	Cell Phenotypes	Sample	Length (Median)	Length (Mean)	Length (Std. Dev.)
MC	EC	1	14.0	29.861	47.902
MC	EC	2	19.5	36.532	44.589
CC	EC+AC	1	31.0	69.716	104.124
CC	EC+AC	2	25.0	51.974	75.642
CC	EC+PC	1	21.0	46.475	62.567
CC	EC+PC	2	39.0	71.202	84.532
TC	EC+PC+AC	1	13.0	36.264	66.717
TC	EC+PC+AC	2	20.0	49.248	109.776

Table 4.2: Table of summary statistics for the measurement of vessel lengths, in voxels.

microassays. The co-culture microenvironments do indicate some correlation with higher vessel lengths. With similarity between EC+AC-1 and EC+PC-2 as well as EC+AC-2 and EC+PC-1 as suggested in the spatial distribution of flows. Yet, both groups do not have further similarities with the respective tri-culture microenvironments. In a similar way as before, the average or distribution of vessel length measurements provides little predictive power to either within-culture properties or the resulting spatial distribution of flow in different microassay samples.

In Table 4.3, we present the results for three measurements of tortuosity. First, we consider the Arc-Chord Ratio, otherwise known as the Distance Metric. The majority of vessels are close to a value of 1, i.e. vessels follow the most direct path between endpoints. Only a few vessels produce a centreline which has been measured an order of magnitude more than the most direct path. The highest values in average Distance Metric suggest the same group identified from our spatial analysis of flow from analytical simulations, that is EC+AC-1, EC+PC-2 and TC-2. However, the mono-culture microenvironments do not share the most similar value in the Arc-Chord Ratio, with EC-2 producing the lowest average tortuosity and EC-1 having the third lowest. In particular, from observations of the distribution in Appendix B (Figure B.6a) EC-2 has no vessels which consider paths an order of magnitude more than the most direct, so variation from the direct path between endpoints is much lower than all other microassay samples. When considering the number of inflections along the centreline of a vessel, the narrative is the same again. The group identified in the spatial distribution of flow have the highest average number of inflections, except for EC-1 ranking third. We observe from their distributions, presented in Appendix B (Figure B.6 that the bias towards larger magnitudes is apparent but with a significant skew towards the mean value. The multiplication of individual vessels' Distance Metric with their number of inflections is the definition of the ICM. We observe from the summary statistics in Table 4.3 that the grouping indicates a classification of microassay samples similar to that by our flow simulations. That is, EC+AC-1, EC+PC-2 and TC-2 have values of both the Distance Metric and the number of inflections which multiply to a higher average measure of tortuosity across the network. This is followed by the second

highest set of average ICM being EC+AC-2, EC+PC-1, and TC-1. Finally, the mono-culture microenvironments produce the lowest values of tortuosity according to this metric. This is in particular of interest since previous measures of tortuosity either did not present the same classification based on a microassay sample's average, or the values were considered very close. For example, EC-1 had previously been ranked as having the third-highest average number of inflections along vessel centrelines but now has the lowest average value of tortuosity according to the ICM metric. This would suggest that although there are many inflection points along the vessel centreline, these are likely small amplitude in an otherwise direct vessel between endpoints. Therefore, the ICM has distinguished between a type of tortuosity, as suggested by Bullitt et al. [96], which aids in the distinction between groups of differing flow distributions. The only limitation in this metric is the variation in value across the network, represented by a high standard deviation which would suggest that the occurrence of a few vessel centrelines that are magnitudes higher than the average, since values cannot be negative, might skew the mean. SOAM is the final metric utilised for the computation of tortuosity. In this case, the distinction between groups of microassay samples is less distinguishable again. The mono-culture microenvironments are not similar in average value, where EC-1 has the third highest average value and EC-2 has the lowest. The variation in the average suggests the distinction between average values is unclear. To conclude this analysis on the tortuosity of vessels in our microassay samples, we have received some insight into the type of tortuosity that might be able to distinguish between the different types of networks as reflected in our spatial distribution of flow. However, tortuosity measurements appear to have many vessels that would be considered near direct between endpoints, with only a few vessels with a significantly higher tortuosity. As a consequence, the metrics presented here give some indication of the geometry of individual vessels but are amenable to high variation, creating uncertainty in the average vessel's underlying geometry.

The final set of metrics summarise the topological characteristics of our microvascular samples, referred to as network metrics. We considered 14 of these measures which the summary is provided in Appendix B (Table B.9). In summary, no single metric was able to characterise

Culture	Cell Phenotypes	Sample	DM (Mean)	DM (Std. Dev.)	Inflections (Mean)	Inflections (Std. Dev)	ICM (Mean)	ICM (Std. Dev.)	SOAM (Mean)	SOAM (Std. Dev.)
MC	EC	1	1.613	4.115	5.314	6.390	9.381	32.901	0.286	0.136
MC	EC	2	1.205	1.065	6.835	8.683	7.394	10.190	0.329	0.125
CC	EC+AC	1	2.918	6.266	9.254	10.425	32.022	81.422	0.299	0.145
CC	EC+AC	2	1.995	3.081	6.209	6.472	11.859	25.599	0.325	0.117
CC	EC+PC	1	1.765	3.384	5.409	5.505	10.900	34.938	0.302	0.138
CC	EC+PC	2	2.089	4.418	10.438	11.657	21.440	45.771	0.335	0.119
TC	EC+PC+AC	1	1.525	2.310	5.172	5.149	9.516	19.177	0.328	0.139
TC	EC+PC+AC	2	2.537	4.905	6.394	7.066	24.191	63.105	0.332	0.152

Table 4.3: Table of summary statistics for the measurements of tortuosity.

either within-culture similarities or the similarities between culture microenvironments which resulted in comparable flow distributions from our simulations. The closest topological measures which indicated a resemblance to the classifications previously discussed included the eccentricity, core-periphery ratio, edge betweenness and flow coefficient. For the first, there is a distinction between culture microenvironments into the broadest categories of mono-culture, co-culture and tri-culture. In this case, the highest values of mean eccentricity are highest in the co-cultures, followed by the mono-cultures, with the lowest average values for the tri-culture microenvironments. However, the distinction within the co-culture microenvironment is more indicative of the spatial distributions of flow simulated where EC+AC-1 is more similar to EC+PC-2 and EC+AC-2 is more similar to EC+PC-1. Similarly, EC+AC-1, EC+PC-2 and TC-2 are more similar in value for both the core-periphery ratio and the flow coefficient. For both these metrics, the group of microassay samples presented the highest values. Both measures could be suggested to be measuring the same effect but by different methods. The first, where the ratio of core nodes to periphery nodes is computed, either there are fewer branching points which lead to no further connections or there are a high number of interconnected vessels within the network. Computing multiple network metrics allows for complimentary metrics to specify outcomes when multiple states produce the same value. In this case, we refer to the Edge Betweenness for a measure of vessel interconnectivity. The three microassay samples, EC+AC-1, EC+PC-2 and TC-2, have the lowest value of mean edge betweenness which suggests that the networks are not interconnected. This is also reflected in the flow coefficient with the same samples producing the lowest mean values. However, as summarised before neither these metrics nor any other network metrics appear to suggest a classification of the microassay samples according to their culture microenvironment or the spatial distribution

of flow in the analytical simulations.

The objective of this analysis was to understand the predictive power of individual metrics computed for either emergent properties due to microenvironment culture or characteristics indicative of observed spatial distributions of flow. To summarise, no single metric provided convincing evidence of predictive power for either account. On the other hand, there were several metrics which identified some relationship between samples albeit with small effect size. Despite not proving to establish clear classifications between the microassay samples, combining multiple measures with low accuracy might identify the networks that are more or less similar. Since individual metrics for geometry and topology do not identify this, we consider methods which combine multiple statistics and reveal the most influential trends in similarity.

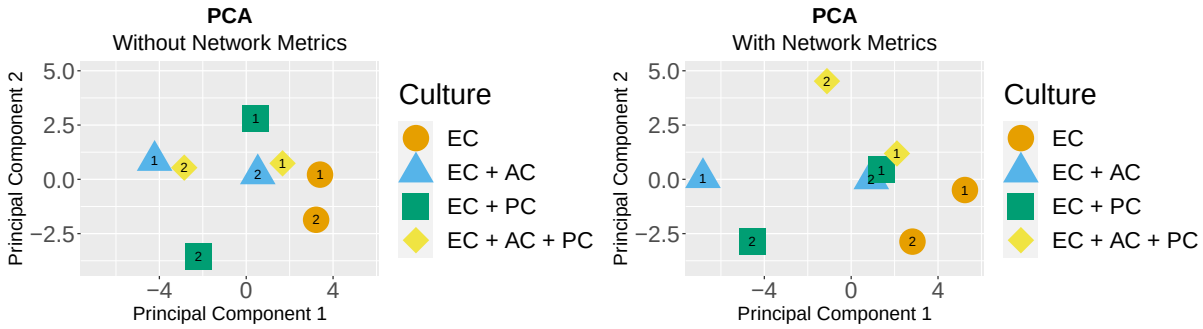
4.6.3 Dimensional Reduction Analysis

The distributions on individual metrics presented previously have been unable to discern the cultures of individual samples or identify a pattern comparable to our analytical simulations of flow. In this analysis, we will interpret the complete set of metrics computed in this analysis in unison to interpret the underlying similarity between microassay samples. To do so, we employ PCA and UMAP which provide two different perspectives to reduce high-dimensional data. The first measures similarity according to cross-sample variation, and the second models the manifold in which the data is embedded with a fuzzy topological structure. There have been two inputs used for each multi-dimensional reduction technique. This provides additional insight into using only geometry metrics versus geometry and network metrics to interpret the microassay samples.

To compare metrics between the different samples, some measures have been normalised to produce relative measures. For example, volume and surface area have been normalised by the volume of the domain imaged with the assumption that this is the maximum volume that could have been considered. Otherwise, all summary values of the individual metrics are used, that is average and standard deviation values.

We consider the results from PCA as presented in Figure 4.8. In the first analysis, only the geometry metrics are used as input. In Figure 4.8a, the first two principal components represent 72.9% of the cumulative variance (Appendix B, Table B.10). The three metrics which contribute the highest loadings to the 1-st Principal Component are the mean ICM, standard deviation for Vessel Length and mean Distance Metric (Appendix B, Table B.11). The three metrics which contribute the highest loadings to the 2-nd Principal Component are the normalised Surface Area, the number of inflections along a vessel plus one, and the standard deviation for Local Thickness (Appendix B, Table B.12). In the second analysis, the geometry metrics and network metrics are used as input. In Figure 4.8b, the first two principal components represent 67.0% of the cumulative variance (Appendix B, Table B.13). The three metrics which contribute the highest loadings to the 1-st Principal Component are the mean Vessel Length, characteristic path length and core-periphery ratio (Appendix B, Table B.14). The three metrics which contribute the highest loadings to the 2-nd Principal Component are the standard deviation for degree, the standard deviation for SOAM, and maximum normalised betweenness (Appendix B, Table B.15).

When considering the most similar microassay samples in the PCA without network metrics, we see that the mono-culture microenvironments are considered close together in the reduced plane. This remains to be so in the PCA with geometry and network metrics. However, the distinction has become more clear from other microassay samples since the PCA without network metrics depicts a more significant similarity to TC-1 (referred to as EC+AC+PC-1). An equivalent observation is made for the similarity between EC+AC-2, EC+PC-1 and TC-1. For the PCA without network metrics, the similarity between the microassays is in the same region of the plane but the EC+AC-2 and TC-1 are considered much closer to each other. In addition, EC-1 is also considered more similar to these two microassay samples than EC+PC-1. However, when we consider the PCA with networks included, we observe that all three samples are now considered very close to one another around the origin. When we consider the last three microassay samples, they are separated from the other microassay samples in the 1-st Principal Component in both PCA results. However, only two of them are considered similar



(a) (b)

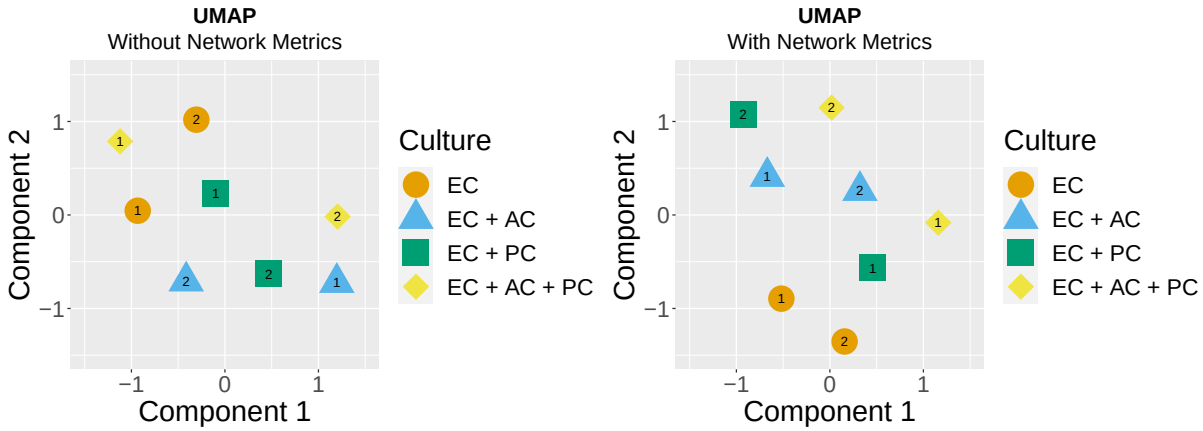
Figure 4.8: Results from PCA multi-dimensional reduction where (a) excludes the network-based metrics and (b) includes the network-based metrics.

when no network metrics are included, that is EC+AC-1 and TC-2. When we refer to the PCA with network metrics this similarity does not persist but shows more variance in the metrics which contribute most to the new 2-nd Principal Component than were observed previously. Referring to the metrics which contribute most to the principal components, we can see that in all of them, where EC+AC-2 and EC+PC-1 have the lowest levels TC-2 has the highest. All three of the metrics refer to the variation in network distribution whether directly through the standard deviation or indirectly by the maximum normalised betweenness which by its definition looks at the edge which is most extreme in the network. Also, when we refer to the analytical simulations of flow, although the distribution is not concentrated only close to the inlets there is more variation across the whole network. Therefore, the initial results of the PCA indicate that the introduction of network metrics introduces a better distinction between the culture microenvironments to represent traits we observe in the spatial distribution of flow.

We replicate the dimensionality reduction using an alternative method to see if similar results can be obtained. By using UMAP, we have only an indication for some notion of similarity between the immediate neighbours of the reduced plane. In the first analysis, where

only geometry metrics are used, the microassays we observe as similar reflect resemble only one of the relationships we have observed so far. Here, we see that EC+AC-1, EC+PC-2 and TC-2 are the only microassays that are considered similar. Otherwise, the monoculture microenvironments are not considered similar to one another as well as EC+AC-2, EC+PC-1 and TC-2. However, when we consult the UMAP representation once both geometry and network-based metrics are introduced we recognise the microassays with a similar spatial distribution of flow as the closest in the reduced manifold. That is, we have both monocultures closest to each other, followed by EC+AC-2, EC+PC-1 and TC-1 with EC+AC-1, EC+PC-2 and TC-2 in the final triplet but EC+AC-1 and EC+PC-2 closer together than to TC-2 as discussed in the equivalent PCA analysis. As a result, using a different technique for interpreting the numerous measures the similarity has persisted in the microassays recognised to have similar spatial distributions of flow. In addition, when considering only geometry measurements, the distinction was not possible. It was only by the introduction of the network metrics that the description of our flow simulations could be revealed. Also, no individual network metric was recognised to distinguish the microassay samples as they have but when considered in tandem to provide a holistic understanding of the microvasculatures the emergent properties were then present.

The results of the dimensionality reduction have provided evidence for the consultation of both geometry and network-based metrics to characterise the transport of oxygen and nutrients. Using these techniques also allowed for the unsupervised investigation of metrics which most influenced the similarity between microassay samples. For example, in PCA, the loading of metrics towards the principal components presents the metrics which microassay samples present similar variation across. Therefore, these metrics should be consulted first when trying to investigate the measures which might best summarise the heuristics we observe in the analytical flow simulations. When more samples are obtained, this approach is well-suited to more samples. It is with ease that it can be verified if the same metrics can summarise the patterns we see in the spatial distribution of flow.



(a) (b)

Figure 4.9: Results from UMAP multi-dimensional reduction where (a) excludes the network-based metrics and (b) includes the network-based metrics.

4.7 Outlook

4.7.1 Assumptions in Geometry Measurements

One of the most significant inaccuracies of the geometrical metrics measured here is in the vessel cross-section noted in 4.4.1. We provide a range for the radii assuming a circular cross-section and an estimate for the semi-major and semi-minor axes when assuming an elliptical cross-section. When assuming a circular cross-section, the assumption is that the minimum radii are representative of the radii of the vessel. From observation, we know this to not be true as the vessels appear elliptical in cross-section (Figure 4.6). Therefore, the original minimum radii is assumed to be the semi-minor axis and the semi-major axis the minimum distance from the centreline to the background in the x-y plane only. However, the assumption that the alignment of the vessels in the x-y plane might be sufficient for this data set but is not generalisable to other data sets that might align similarly. Sampling different points on the surface of the microvasculature at a tangent to the orientation of the centreline would provide the most accurate estimates, with studies implementing this [220, 267] as well as commercially

available software, such as Amira [268]. With this information, a more accurate measurement of the geometrical properties and subsequent models for blood flow can be obtained.

4.7.2 Extracting Geometry Measurements to Estimate Vessel Conductance, c_{jj}

In the derivation for our analytical flow simulations in a network, we summarise some coefficients into a term for conductance. In this term, we can relate estimates for the geometrical properties of the blood vessel. For example, if we consider a vessel with circular cross-section the conductance of a vessel is, $c_{jj} = 8\mu L/\pi R^4$. the length of an individual vessel is already measured. However, the radius of the vessel has a dominant effect in conductance so accuracy is required. In addition, radius in a vessel varies and so might need to be taken into account in the "true" conductance of a vessel. Not only this, but the tortuosity of a vessel is likely to impact the conductance of a vessel, whether caused by the change in momentum or additional effects from turbulence. Regardless, this approach would be able to consider not only the connectivity of blood vessels but also variations in geometry. The result would likely vary the results for our current estimate of flow. However, we expect connectivity in the microvasculature to have a significant influence.

4.7.3 Application of Topological Data Analysis

Stolz et al. [269] propose the utilisation of PH to characterise the emergent connectivity and tortuosity of the microvasculature. Here, the analysis used persistent homology to understand microvasculatures in whole tumours, with the different Betti numbers indicating the tortuosity, loops and voids at different depths of the tumour. In our studies, the analogy of a central point does not hold since perfusion occurs from one side of the microassay to the other. Instead, Bendich et al. [270] use PH in dimension 0 to study the tortuosity of brain arteries by considering a filtration which moves a plane along some tangent to the network. In the first instance, the full network is situated on one side of the plane. As the plane moves, it starts

intersecting the network. The authors interpret the persistence of features in dimension 0 as a measure of the size of a bend in a brain artery. However, the data set in our study is too small to maximise the utility of such analyses when it can be directly computed with minimal computational cost. With the anticipation of larger data sets, it would be prudent to be aware of the potential for PH in analysing more complex spatial structures.

Another topological technique could provide a useful tool for distinguishing blood vessel morphology. The Euler characteristic is derived from a foundational equation in topology. Most notably, the metric had been used widely in analysing geological structures [271, 272, 273, 274, 275]. In microvasculatures, the metric has been used in the context of blood vasculatures such as to identify diabetic retinopathy in retinal images [276]. Therefore, a natural next step would be to implement this analysis as an additional metric to capture primarily topological, but some geometric properties of the network too.

4.7.4 Conclusion

In this study, 3D images with challenging spatial organisation and irregular pixel intensity were accurately recovered automatically. An analytical simulation of flow was derived which incorporated the spatial organisation of connecting blood vessels. We have demonstrated the analysis of self-organised microvasculature can be carried out to obtain results indicating spatial distributions of flow. From the review of the literature, we utilised metrics on a newly acquired data set and combined them using dimensionality reduction techniques. In addition, this analysis showed that the combination of both geometry and topology was necessary to identify flow similarities between microassay samples. Several considerations for the accuracy of flow simulations have been discussed which aim to resolve uncertainty in the model and improve upon its realism. With the introduction of additional data points, we expect to identify the emergent properties that determine specific distributions of flow around the microvasculature.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Discussion

The stated objectives of this thesis were to investigate:

- How does the spatial cellular organisation of immune cells predict patient outcome for Colorectal Cancer and Sjörgens disease?
- Can we identify higher-order structures of cellular organisation that predict disease progression outcomes?
- Which geometric and/or topological metrics characterise the transport of oxygen and nutrients in networks undergoing vasculogenesis in different microenvironments?

In the present thesis, we have quantified the spatial distributions of cell-cell proximity and their resulting structures. In CRC, we computed metrics for individual cell spatial correlations over multiple length scales. This included a novel methodology for the measurement of density distribution as measured from the boundary of two discrete areas. Following this, we were able to consider the most advanced measures for multi-cellular nearest neighbour communities. In Sjögren's disease, we identified higher-order structures previously dependent on subjective demarcation. As well as validating and using the most optimal algorithms currently available in the literature, we introduced new concepts, such as density gradients within and beyond

demarcated clusters and utilised advanced mathematical techniques, namely PH, to verify abstract geometrical concepts. Not only does the workflow show promise on a subset of the data set, but the approach is agnostic to cell phenotype allowing avenues into the exploratory analysis of unknown immuno-aggregations with structure. In live blood vessel imaging, we extend the current analysis to combine geometrical and network-based metrics to quantify morphology that could be used to classify emergent properties dependent on the microenvironment within which the blood vessels are grown. In each of these cases, we demonstrate the computation of advanced computational techniques on biological data to assess statistical differences between the spatial distributions of samples.

We observe in this work the necessity to measure the spatial organisation between cells at multiple length scales. This is because cell signalling is possible across several ranges, the result of which can induce spatial correlation or anti-correlation affecting homeostasis of biological systems. In CRC, analyses would report only the cell phenotype whose density appeared to be statistically significant between two patient cohorts. It is through subsequent analysis that we were able to infer specific mechanisms by which a higher density of cell phenotype can permeate throughout the tumour as well as indicate the spatial organisation directly with tumorous cells. Specifically, how the M2 macrophage population was spatially correlated with the immune-activated and non-activated epithelial population. In Sjögren's disease, we expanded on previous metrics for communities explored in our analysis of CRC, where a neighbourhood only considered the most immediate cells as functional units. Here, we evolved the analysis to consider structures which are known to directly influence disease progression. In live blood vessel imaging, although cellular microenvironments were assumed to be homogeneous, we were able to measure the heterogeneity between vascular networks. When geometrical properties alone were unable to determine the differences in cell culture the evidence suggested that topological properties were able to characterise the effect of the microenvironment in combination with geometrical attributes. In our analyses, we have demonstrated how we have incorporated multi-scaled measurement techniques in the holistic assessment of biological systems. As a result of the characterisation of multiple forms of cell

signalling, we can provide a holistic interpretation of the biological system.

When working at the interface of disciplinary boundaries, knowledge and understanding from both areas of research are paramount to attaining scientific objectives. In CRC, we were able to describe multiple cell phenotypes and their spatial organisation not only with each other but with specific reference to the tumourous epithelium. In Sjögren's disease, part of the biology is already made apparent by investigating TLS structures directly. In this case, the high-density aggregations of one cell phenotype are surrounded by a complimentary cell phenotype. However, by understanding the nuances of the algorithms available we were able to apply those which were best suited to high-density aggregations, such as the ToMATo and DBSCAN clustering algorithms, as well as quantify the completeness of a surrounding phenotype via PH as well as new methods solely for this objective. Therefore, we have demonstrated a comprehension of both biology and applied mathematics to give direction on the relevant characterisations of biological systems as well as identify the most compatible mathematical apparatus to assess them.

When applying new computational methodologies the complexity can limit the broader application of such procedures. An example of this is in the general applicability of PH. Despite being an essential step in our workflow, the computational cost of the algorithm, even on a high-performance computer like that used throughout this work, meant it was only possible to constrain the application to a subset of the whole point cloud. Even when considering this reductive technique, this remained by far the most computationally expensive stage of the workflow. This is an assessment also echoed by Stolz-Pretzer [277], where in the discussion it is said that reductive techniques can only go so far if the application allows. Otherwise, for broader impact, further investigation into mathematical techniques is necessary to provide universality to these methods. Chromatic TDA is expected to propose similar challenges. Albeit similar in objective to our methods, the increased computational complexity would require pre-processing of the input point cloud or mathematical efficiencies to enable computation on the scale relevant to biological systems, i.e. a magnitude comparable to a million coordinates. In any case, our analysis has been able to complete the analysis of multiple

samples, in a timely fashion albeit with the usage of high-performance computing. However, in the two studies which utilised high-performance computing, the necessity of computational power reflected the number of patient samples rather than the complexity of the methods used. This differs from the aforementioned where the complexity of the method is instead the limiting factor. The results of individual samples were computed within an hour, except for the construction of the Delaunay Triangulation which took a few hours. Therefore in practice, where there are significant resource limitations for multiple sample collections, the methodologies proposed fit into a pipeline where an analysis can be completed on the time scale that it takes to obtain another sample.

The utilisation of new imaging modalities results in data sets that are often limited by either quality or size. For example, in CRC, we question the legitimacy of a cell marker, specifically CD163, without colocalised expression of a complimentary cell phenotype, i.e. macrophages. In Sjögren's disease, limitation of cell marker information and sensitivity in thresholding cell phenotypes lead to the incomplete characterisation of TLS germinal centres. In live blood vessel imaging, access was available to only a small sample due to the difficulties in biological manipulation from either toxicity from imaging or the complexity of vasculogenesis. Again, this emphasises the need to be armed with appropriate biological understanding as well as robust mathematical techniques to tackle the uncertainty in such results.

Independent of our novel methodologies, we do not expect to have presented an exhaustive list of measures or the respective data sets. Instead, we report here the most applicable from our assessment of the literature review for each of the applications. It is from here that we have been able to evaluate the best approaches to interpret the complex spatial structures hypothesised. As a consequence, we consider the techniques developed in application to each of our data sets to be the most important for the interrogation of complex spatial structures.

In summary, the spatial organisation of cells and the surrounding microenvironment has been measured at a range of length scales with emergent structures identified. We expect our methods to provide insight into biological problems beyond the applications shown here. Future analyses should look to build upon the continuum of geometrical and topological information

that we have already been able to accumulate and interpret the two in unison for greater biological understanding.

5.2 Conclusion

For each of these projects, we have been able to carry out research which investigates the complex spatial structures in biomedical images. The images contain information on the organisation of cells within a multi-element, multi-scaled and multi-regulated microenvironment. The first project characterised the TME in CRC to better understand the spatial interplay between it and a subset of the immune system with minimal assumptions on the biology at play. The second project automatically identified immuno-aggregations already recognised as key agents in autoimmune diseases and immune responses to cancer alike. The third project investigated how endothelial cells self-assembled into blood vessel networks, from low-resolution live cell images, according to different culture environments. The analysis gave initial insight into the effect of these different microenvironments on the transport of oxygen and nutrients to a hypothetical surrounding tissue.

To conclude, the biological context, mathematical techniques and analysis developed in this thesis have proven valuable to the discovery of novel biological organisation through the application of forefront and often novel techniques, where similar discoveries are expected in future analogous systems.

APPENDIX A

USEFUL MATHEMATICAL DEFINITIONS

A.1 Complex Spatial Networks

Let us decompose the phrase complex spatial networks into its constituent parts to better define it. The *spatial* attribute refers to elements being embedded in a space, a boundless, three-dimensional extent in which objects and events occur and have relative position and direction [278]. A *network* can be described as distinct elements represented by *vertices* (or *edges*) and the connections between the elements as *edges* (or *links*). The addition of the word *complex* signifies that we refer to a continuum of networks which are neither trivial nor uniform. So, complex spatial networks exist in a measurable space and yet follow connections between elements that are not simple.

The area of complex networks, coined by Barabási and Réka [279], has proven useful in application for the mathematical construct of *graphs* to many problems. In this section, we present the foundational definitions in mathematics which are commonly referred to in the following theses and provide the backdrop for which we discuss our research.

First, we define a graph, as understood in mathematics.

Definition A.1.1. Graph, Subgraph

We define a *graph* as

$$\mathbb{G} = (V, E)$$

where V denotes the set of *elements*, called vertices (or nodes), in the network. $E \subseteq V \times V$ is the set of *connections*, called edges (or links). The number of vertices in a network is given by:

$$n = |V|$$

and the number of edges by:

$$m = |E|.$$

A *subgraph*, $G' = (V', E')$, where both $V' \subseteq V$ and $E' \subseteq E$.

Graphs can be categorised as *undirected*, where an edge has no directionality assigned to them, or *directed*, with an attribute of directionality assigned to a connection.

Definition A.1.2. Undirected, Directed, Di-graphs

A graph is considered *undirected* if for two nodes $u, v \in E$ connected by the edge $(u, v) \in E \implies (v, u) \in E$, while a graph is *directed* if the equivalent edge $(u, v) \in E \implies (v, u) \in E$. Directed graphs are sometimes called *di-graphs*.

Frequently, graphs are represented in a form readily applicable to linear algebra, known as an *adjacency matrix*, which summarises the connections between edges.

Definition A.1.3. Adjacency Matrix

An *adjacency matrix*, $\mathbf{A}$, is an $n \times n$ square matrix, where the non-zero elements of the matrix represent the existence of an edge between two pairs of vertices. Specifically,

$$\mathbf{A}_{u,v} \neq 0 \iff (u, v) \in E$$

Within our context of complex spatial networks, we are considering networks with additional properties that refer to an element's location in space. Beyond location, vertices and edges can be annotated with supplementary data. This additional information is called a network's *attributes*. For example, in social networks, this could be the proportion of followers on a social media platform who are likely to adopt the same linguistic style [280]; the migration

rate of species of bird [281]; or the streamflow data of monitoring stations along rivers [282]. Considering the attributes of a network, we can incorporate this additional information. The *weight* of an edge represents the association between vertices based on a metric determined by an attribute [283]. Sometimes considered separately, the value of weight can take negative values and then be referred to as a *signed network*. This can be particularly useful when considering signalling networks [284], where an edge reflects whether a relationship is activatory with a positive sign or inhibitory with a negative one. Both extensions of graphs are coherent with many concepts. To illustrate, the adjacency matrix becomes a *weighted adjacency matrix*, $\mathbf{W}$, where each element in $\mathbf{W}$ contains the weight of the edge between vertices.

We now introduce some basic network concepts. We separate the concepts into those that relate to the microscopic properties of the network, then how elements and connections relate to each other and, finally, concepts that summarise the global properties of a network. We do not state here how these concepts are then quantified statistically. Instead, we leave this to the specific application. However, the terminology presented here relates to specific chapters more specifically. For example, in Chapter 2, we consider different constructions of networks from spatial point patterns, looking at the microscopic properties followed by the relationship between pairs of cell phenotypes. In Chapter 3, we quantify collections of cells to understand the mesoscopic features that appear in a network. Finally, Chapter 4 considers the distribution of microscopic features of a network as well as computing specific measures of the mesoscopic and macroscopic properties to differentiate networks.

A.1.1 Microscopic Network Features

The previous set of definitions has enabled us to construct a network. Here, we define the properties of the network which can be directly inferred from its construction. We differentiate these from metrics which measure properties that statistically quantify the network. In this case, statistical extensions often refer to the distribution of such microscopic network features. Instead, we present common terminology used in the field of complex spatial networks. For microscopic network features, this refers to properties of individual elements: their vertices or

edges.

Definition A.1.4. Vertex degree

The *degree* of a vertex, $u \in V$, is the number of edges that contain elements made up of the vertex u.

Definition A.1.5. Vertex Strength

The *strength* of a vertex is the sum of the weights of edges that connect to vertex u.

Definition A.1.6. Self-loop

We define a *self-loop* as an edge which connects a vertex to itself. Therefore, we can define the set of self-loops as

$$E_{\text{self}} = \{(u, v) \in E | u = v\}.$$

When a graph, G, has neither self-loops nor has more than two edges defined for any pair of vertices, we categorise this as a *Simple Graph*.

Definition A.1.7. A *branching point* are vertices with a degree equal to 3. Therefore, we can define the set of branching points as

$$E_{\text{branching}} = \{u | \text{degree}(u) = 3\}.$$

A.1.2 Mesoscopic Network Features

Unlike microscopic terms, mesoscopic network features refer to properties which relate to a pair or collection of elements. First, whether a chain of connections exists between elements. Then, we can quantify the distance between elements in several ways as well as identify sets of vertices that are connected similarly.

Definition A.1.8. Reachability

We say that a vertex v is *reachable* from another vertex u if there exists a path that begins at u and finishes at v.

Definition A.1.9. Path, Path Length

A *path* is the set of sequential vertices, $v_1, \ldots, v_l \in P \subseteq V$, where $\forall i \in [1, l-1], (v_i, v_{i+1}) \in E$.

The *path length* is the cardinality of the corresponding edge set for a given path.

Definition A.1.10. Random Walk

A *random walk* is a given path from a vertex, $u \in V$, to another reachable vertex $v \in V$, where subsequent edges of the path are determined using a stochastic process.

We differentiate between a path and a random walk as they reveal different properties about a network. In the first, we are interrogating the pre-determined set of edges between two vertices. For a random walk, the path is undetermined. This can emulate processes within a system in specific ways in which objects travelling along connections have little to no knowledge about the graph construction itself. Instead, the sequence of vertices is chosen by considering the information available at a vertex or from previous vertices. Additionally, the process often takes into account some element of uncertainty in each decision. As such, the stochastic process can take into account factors, such as first-order connections [285], higher-order connections [286] or node and/or edge attributes [287].

Considering vertices as the building blocks of a graph, they can be summarised by the combination of connections to their *neighbours*, and neighbours thereof.

Definition A.1.11. k-th Neighbours

The k-th neighbours of a vertex, u is the set of vertices, $v_1, \ldots, v_l \in N_k$ where each vertex, $v_i \in N_k$, has a path from u to v_i and of all paths the minimum path length is equal to k.

Definition A.1.12. Clique, Maximal Clique, Maximum Clique

A *clique*, C, in an undirected graph $\mathbb{G} = (V, E)$ is a subset of the vertices, $C \subseteq V$, such that every two distinct vertices, $u, v \in V$ are adjacent. This is equivalent to all vertices in C being the 1-st neighbour to every other vertex in C.

We refer to cliques based on the cardinality of the set C, i.e. if $|C| = 3$, then C would be referred to as a 3-clique, and so on.

A *maximal clique*, is a clique that cannot be extended by including one more adjacent vertex, that is, a clique which does not exist exclusively within the vertex set of a larger clique.

A *maximum clique*, C_m, of the graph, $\mathbb{G}$, is such that there are no other cliques with cardinality greater than the C_m. Therefore, there can be many maximum cliques with the same number of vertices.

Definition A.1.13. Network Motifs

A *network motif* [288] can be defined as a pattern of interconnections occurring in complex networks at numbers that are significantly higher than those in randomised Erdős-Renyi-Gilbert models [289].

Definition A.1.14. Cycle / Circuit / Feedback Loop / Loop

When a path begins and ends at the same vertex, this can be referred to as a *cycle*, *circuit*, *feedback loop*, or simply a *loop*.

When considering only undirected networks, motifs can be divided into a-cyclic (linear or tree-like) and cyclic, i.e., those with loops. For directed networks, the loops can be further divided into feed-forward and feedback loops [290]. Feedback loops have been a useful description for phenomena in neural networks [291] and introducing hierarchy in directed networks [292].

A.1.3 Macroscopic Network Features

The following set of terminology considers the properties of the whole network. These can be binary, discrete or numeric values assigned to describe features concerning all elements of the network. We do not describe statistical measures of microscopic features, such as average degree, as macroscopic features but rather new terminologies that require the information of the whole network.

Definition A.1.15. Network density

The *network density*, ρ, is the ratio of the number of edges in a network to the total possible number of edges. Mathematically, we have

$$\rho = \frac{2m}{n(n-1)}$$

for undirected networks, and

$$\rho = \frac{m}{n(n-1)}$$

for directed networks, where n and m are quantities from Definition A.1.1.

Definition A.1.16. Connected, Strongly Connected, Weakly Connected

An undirected network is said to be *connected* if every vertex is reachable from every other vertex.

A directed network is said to be *strongly connected* if there exists a path from every vertex to every other vertex.

A directed network is said to be *weakly connected* if the undirected equivalent of the network is connected.

Definition A.1.17. Eccentricity, Network Radius, Network Diameter

The nodal eccentricity is the maximal path length between a node and any other node in the network.

The *network radius* is the minimal eccentricity

The *network diameter* is the maximal eccentricity.

Definition A.1.18. Component

A *component* of a graph is a subgraph where all undirected vertices are reachable to every other vertex.

The largest component is colloquially known as the *giant component* [293]. In undirected networks, this can be relevant to understanding the criteria for which connections might exist [289]. In directed networks, the size of the giant component can be estimated by the number

of backward edges in tropic levels [294]. However, networks are often investigated for their structure within connected components. For this, we refer to definitions of hierarchy.

Definition A.1.19. Hierarchical Organisation

Hierarchical organisation refers to nodes in a complex network dividing into groups that "further subdivide into groups of groups, and so forth over multiple scales" [295].

Definition A.1.20. Louvain community The optimal *Louvain community* structure is a subdivision of the network into non-overlapping groups of nodes which maximises the number of within-group edges and minimises the number of between-group edges.

Definition A.1.21. Core-Periphery Structure

The optimal *core-periphery* subdivision is a partition of the network into two non-overlapping groups of nodes, a core group and a periphery group, in a way that maximises the number (or weight) of within core-group edges and minimises the number (or weight) of within periphery-group edges.

The preceding terminology describes the essential properties described in the thesis. When quantifying some of these features, it is useful to compare them to well-understood constructions of complex networks, such as Erdős-Renyi-Gilbert models. These are artificial networks which follow connection criteria emphasising properties observed in many real networks. Therefore, quantifying the prior traits can provide insight into the underlying mechanisms which determine connections in a network as well as their collective effect.

A.2 Topological Data Analysis

Considering the first printed reference to the field of 'Topologie' was by Johann Benedict Listing in 1847 [296], Topological Data Analysis (TDA) has only reached the forefront of data science in the last two decades. At the time of writing, PH is the most widely used TDA methodology. Previous research existed which could be categorised as TDA, such as Muldoon in dynamical systems [297] or skeletonisation of networks in image analysis by Zhang

et al. [298]. Yet, TDA is often referred to as undergoing a rejuvenation in 1994; this year in particular after Barannikov [299] described an efficient algorithm for the computation of the canonical forms of filtered complexes, the foundational elements of PH.

Poincaré's influential papers [300] established the mathematical field of homology. Homology is a procedure to associate a sequence of abelian groups with a given mathematical object such as a topological space or a group[1]. Persistence provides a value in which homological structures are more likely to reflect underlying signals rather than contributions from noise. By giving mathematical preciseness, the combination of persistence and homology provides a process to identify the homological structures in a data set which is robust against small perturbations of noise. In addition, the abstract notion of homology allows a generalisation for higher dimensional data sets. Therefore, once combined with computational power and 'Big Data', the ability to detect homological structures in large and high-dimensional data sets is possible.

The meaning of such homological structures then becomes a problem for interpreting the results of PH. Even though unseen structures can now be identified, the abstract definitions which define them have made their translation into tangible meaning difficult. Referring to the intuitive meaning of homology groups can provide some insight such as the 0-th Betti number relating to connectedness and the 1-st Betti number relating to loops[2]. However, their meaning in higher-dimensional spaces loses such physical meaning. In addition, although stable to small perturbations of noise, the differentiation between noise and underlying signal relies on parameters set by the investigator. Usually, the stability of the process provides an obvious threshold between the persistent characteristics and the large volume of noisy data points but for a rigid process and full automation, other research areas such as stochastic analysis and machine learning algorithms will provide solutions. We will identify some of these limitations as we make similar analyses of the results in Section A.2.6.

First, the proceeding section will cover the essential mathematics needed to develop such results. The areas come under the branch of algebraic topology, alongside some of the more

[1] See Section A.2, Subsection A.2.5
[2] See Section A.2, Subsection A.2.5

general terms used throughout topology. We then finish the section with what defines PH, what the outcome of such an analysis achieves and some references to open-source libraries to carry out the analysis in various coding languages.

A.2.1 Motivation

Here we describe the notable definitions and theorems which lead up to the application of PH as a viable technique in TDA. Much of this work is inspired by the following works. For a thorough introduction into the mathematics leading up to the application of PH, refer to Edelsbrunner and Harer [301]; on a focused look into the validity of PH and its interpretation, refer to Chazal and Michel [302], and for a brief overview of both refer to Ghrist [303].

A.2.2 Complexes

In PH simplicial complexes are foundational structures which represent topological spaces derived from a set of data points. We first consider the elemental part of a simplicial complex: a simplex. In anticipation of this definition, we consider the definitions for affinely independent and convex hull.

Definition A.2.1. Affine Combination, Affine Hull, Affinely Independent

Let $u_0, u_1, \ldots, u_k$ be points in $\mathbb{R}^d$. A point $x = \sum_{i=0}^{k} \lambda_i u_i$, with each $\lambda_i \in \mathbb{R}$, is an *affine combination* of the the u_i if the λ_i sum to 1. The *affine hull* is the set of affine combinations. For two affine combinations, $x = \sum \lambda_i u_i$ and $y = \sum \mu_i u_i$, we say they are *affinely independent* if they are the same iff $\lambda_i = \mu_i$ for all i.

By definition, the $k + 1$ points are affinely independent iff the k vectors $u_i - u_0$, for $1 \leq i \leq k$, are linearly independent.

Definition A.2.2. Convex Combination, Convex Hull

An affine combination, $x = \sum \lambda_i u_i$, is a *convex combination* if all λ_i are non-negative. The *convex hull* is the set of convex combinations.

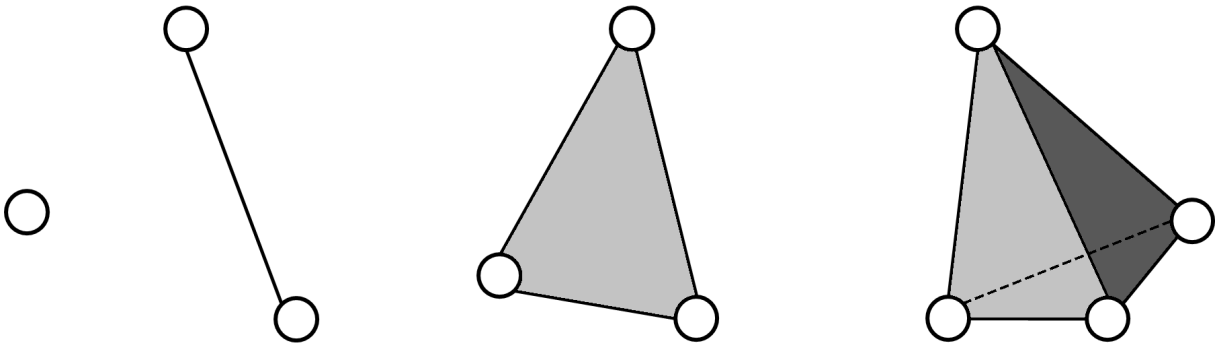


Figure A.1: From left to right: a vertex, an edge, a triangle, and a tetrahedron; which represent the 0-, 1-, 2- and 3-dimensional simplex, respectively.

Now, we can define what a simplex is with this terminology.

Definition A.2.3. k-Simplex

A k-*simplex* is the convex hull of $k+1$ affinely independent points, $\sigma = \text{conv}\{u_0, u_1, \ldots, u_k\}$. Its dimension is $\dim(\sigma) = k$.

The first few dimensions, as seen in Figure A.1, are very intuitive to geometries we usually come across and are often called by those names: *vertex* for 0-simplex, *edge* for 1-simplex, *triangle* for 2-simplex, and *tetrahedron* for 3-simplex. When talking about these elements abstractly we usually drop the dimension and refer to them as either a simplex or many simplices.

Having now defined a simplex, we are closer to defining what a simplicial complex is. However, we need to first understand the relationship between simplices, especially when talking across different dimensions.

Definition A.2.4. Face, Proper Face

A *face* of σ is the convex hull of a non-empty subset of the set of u_i, and it is a *proper face* if the subset is not the entire set.

In Figure A.1, we can observe the first few faces from each simplex: an edge has two vertices, a triangle has three edges, and a tetrahedron has four triangles as faces. We sometimes write $\tau \leq \sigma$ if τ is a face and $\tau < \sigma$ if it is a proper face of σ. Since a set of size $k+1$

has 2^{k+1} subsets, including the empty set, σ has $2^{k+1} - 1$ faces, all of which are proper faces except for σ itself. Here we define a simplicial complex.

Definition A.2.5. Simplicial Complex

A *simplicial complex* is a finite collection of simplices K such that $\sigma \in K$ and $\tau \leq \sigma$ implies $\tau \in K$, and $\sigma, \sigma_0 \in K$ implies $\sigma \cap \sigma_0$ is either empty or a face of both.

It is often easier to construct a complex abstractly and consider the physical interpretation after. The equivalent meaning of a simplicial complex is thus called an abstract simplicial complex.

Definition A.2.6. Abstract Simplicial Complex

An *abstract simplicial complex* is a finite collection of sets A such that $\alpha \in A$ and $\beta \subseteq \alpha$ implies $\beta \in A$.

The terms we have described before, from a geometric perspective, all have analogous meanings in the abstract definition. Then, we can use the following theorem to guarantee a geometric equivalent to our abstract notion of a simplicial complex.

Theorem A.2.1. *Geometric Realisation Theorem*

Every abstract simplicial complex of dimension d has a geometric realisation in $\mathbb{R}^{2d+1}$.

This section has gone through the founding mathematical concepts which all introductions to PH would introduce someone to. Next, we consider what form our data needs to be and how we can then translate the information into a format that can be easily manipulated by other concepts from topology.

A.2.3 Building Simplicial Complexes from Point Clouds

Data points are a collection of datum with d attributes associated with each. The number of attributes defines the dimensional space, $\mathbb{R}^d$, in which the data resides. The space this is related to is often a Euclidean space to relate structures to direct physical concepts and as such is referred to as a point cloud data set.

Definition A.2.7. Point Cloud

We denote by a *point cloud* data set a collection of points in Euclidean d-dimensional space $\mathbb{E}^d$. Also referred to as a *point pattern*.

From Theorem A.2.1, we were able to relate abstract simplicial complexes as topological spaces which could be related to a geometric realisation. Given a data set, there exist many ways to build simplicial complexes. Here are the most commonly considered: the Čech complex and the Vietoris-Rips complex.

Definition A.2.8. Čech complex

Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the *Čech complex*, $\mathcal{C}_\epsilon$, is the abstract simplicial complex whose k-simplices are determined by unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ whose closed $\epsilon/2$-ball neighbourhoods have a point of common intersection, i.e. a non-empty intersection for ball.

Definition A.2.9. Vietoris-Rips complex

Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the *Vietoris-Rips complex*, $\mathcal{R}_\epsilon$, is the abstract simplicial complex whose k-simplices correspond to unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ which are pairwise within distance ϵ.

The distinction between the two definitions is subtle but results in profound differences in the geometric realisation of the data, like in Figure A.2, and the mathematical rigour they establish for later theorems. To analyse this distinction, we need to consider terminology from algebraic topology and other mathematical branches. The final result, called the Nerve Theorem, provides support for which construction should be considered.

A.2.4 The Nerve Theorem

A fundamental theorem in algebraic topology is called the Nerve Theorem. Here, we provide definitions which build constituent parts of the theorem.

Definition A.2.10. Cover, Nerve of $\mathcal{U}$

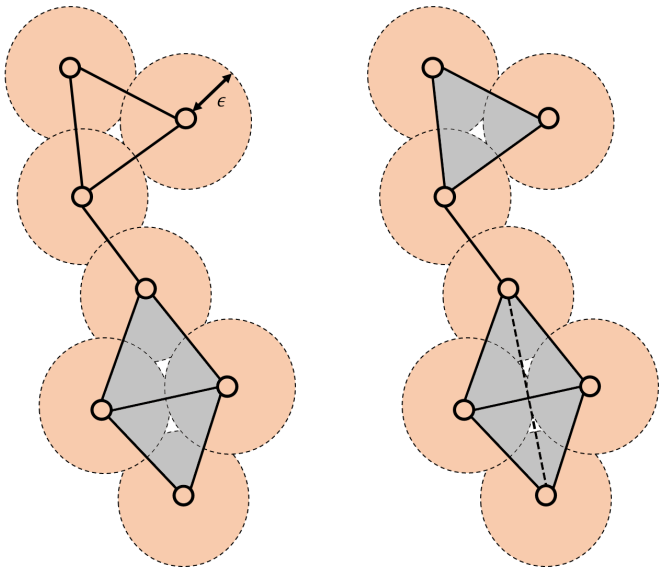


Figure A.2: The (a) Čech complex, $\mathcal{C}_\epsilon$, and (b) Vieterois-Rips complex, $\mathcal{R}_\epsilon$, for the same point cloud data set for the same ϵ. Note the difference in dimension for the two simplicial complexes formed under the two different definitions. For $\mathcal{C}_\epsilon$, we have four edges and two triangles but for $\mathcal{R}_\epsilon$ we have one edge, one triangle and a tetrahedron.

Given a *cover* $\mathcal{U} = (U_i)_{i \in I}$ of $\mathbb{M}$, where $\mathbb{M}$ is the family of sets U_i such that $\mathbb{M} = \cup_{i \in I} U_i$, the *nerve of* $\mathcal{U}$ is the abstract simplicial complex $\mathcal{C}(\mathcal{U})$ whose vertices are the U_i's and such that,

$$\sigma = [U_{i_0}, \ldots, U_{i_k}] \in \mathcal{C}(\mathcal{U}) \quad \text{iff} \quad \cap_{j=0}^{k} U_{i_j} \neq \emptyset. \tag{A.1}$$

Two topological spaces X and Y are usually considered as being the same from a topological point of view if they are homeomorphic.

Definition A.2.11. Homeomorphic

We say our topological spaces X and Y are *homeomorphic* if there exists two continuous bijective maps $f : X \to Y$ and $g : Y \to X$ such that $f \circ g$ and $g \circ f$ are the identity map of Y and X respectively.

In many cases, asking X and Y to be homeomorphic turns out to be a too strong requirement to ensure that X and Y share the same topological features of interest for TDA. That is where the weaker term, homotopic, comes in.

Definition A.2.12. Homotopy, Homotopic

A *homotopy* between $f : X \to Y$ and $g : X \to Y$ is another continuous map $H : X \times [0,1] \to Y$ that agrees with f for $t = 0$ and with g for $t = 1$. Noting that this defines an equivalence relation, we write $f \simeq g$ and call f and g *homotopic* if there is a homotopy between them.

This notion can be used to relate spaces, beginning with a special case.

Definition A.2.13. Retract, Retraction, Deformation retract/retraction/retracts

Let $Y \subseteq X$. We say it is a *retract* of X if there is a continuous map $r : X \to Y$ with $r(y) = y$ for all $y \in Y$. The map r is called a *retraction*. We call Y a *deformation retract* and r a *deformation retraction* if there is a homotopy between r and the identity on X, $\iota \circ r \simeq \mathbb{I}_X$, where $\iota : Y \to X$ is an inclusion mapping. We also say that X *deformation retracts* onto Y.

Every deformation retract is a retract but not the other way around. We can also consider maps in both directions.

Definition A.2.14. Homotopy equivalent/type/equivalence/inverse

We call two not necessarily nested topological spaces, X and Y, *homotopy equivalent* if there are continuous maps $f : X \to Y$ and $g : Y \to X$ such that $g \circ f \simeq \mathbb{I}_X$ and $f \circ g \simeq \mathbb{I}_Y$. This gives an equivalence relation, and we write $X \simeq Y$ and say they have the same *homotopy type* if they are homotopy equivalent. Each of the maps f and g is called a *homotopy equivalence*, and g is said to be a *homotopy inverse* of f and vice versa.

The notion of homotopy equivalence is weaker than the notion of homeomorphism: if X and Y are homeomorphic then they are also homotopy equivalent, the converse is not true. However, spaces that are homotopy equivalent still share many topological invariants, in particular, they have the same homology[1].

Definition A.2.15. Contractible

A space is said to be *contractible* if it is homotopy equivalent to a point.

[1] See Section A.2, Subsection A.2.5

Basic examples of contractible spaces are the balls and, more generally, the convex sets in $\mathbb{R}^d$. Open covers whose all elements and their intersection are contractible have the remarkable following property.

Theorem A.2.2. *Nerve Theorem*

Let $\mathcal{U} = (U_i)_{i \in I}$ be a cover of a topological space X by open sets such that the intersection of any subcollection of the U_i's is either empty or contractible. Then, X and the nerve $\mathcal{C}(\mathcal{U})$ are homotopy equivalent.

In particular, if a metric space $\mathbb{X}$ is a set of points in $\mathbb{R}^d$, then the Čech complex $\mathcal{C}_\epsilon$ is homotopy equivalent. The Nerve Theorem plays a fundamental role in TDA: it provides a way to encode the topology of continuous spaces into abstract combinatorial structures that are well-suited for the design of effective data structures and algorithms.

A.2.5 Homology

Definition A.2.16. k-Chain, Coefficients, Group of k-Chains

Let K be a simplicial complex and k a dimension. A *k-chain* is a formal sum of k-simplices in K. The standard notation for this is $c = \sum a_i \sigma_i$, where the σ_i are the k-simplices and the a_i are the *coefficients*, where we consider only modulo 2 coefficients. The k-chains on K whose elements are the formal (finite) sums of k-simplices of K are described as the *group of k-chains* denoted as $(C_k, +)$, or simply $C_k = C_k(K)$.

Associativity follows from associativity of addition modulo 2. The neutral element is $0 = \sum 0\sigma_i$. The inverse of c is $-c = c$ since $c + c = 0$. Then, C_p is abelian because addition modulo 2 is abelian.

Definition A.2.17. Boundary

To relate these groups, we define the *boundary* of a k-simplex as the sum of its $(k-1)$-dimensional faces. Writing $\sigma = [u_0, u_1, \ldots, u_k]$ for the simplex spanned by the listed vertices,

its boundary is

$$\partial_k \sigma = \sum_{j=0}^{k} [u_0, \ldots, \hat{u}_j, \ldots, u_k], \tag{A.2}$$

where the hat indicates that u_j is omitted.

We distinguish two particular types of chains and use them to define homology groups.

Definition A.2.18. k-Cycle, Group of k-Cycles

A *k-cycle* is a k-chain with empty boundary, $\partial c = 0$. Since ∂ commutes with addition, we have a *group of k-cycles*, denoted as $\mathsf{Z}_k = \mathsf{Z}_k(\mathsf{K})$, which is a subgroup of the group of k-chains. In other words, the group of k-cycles is the kernel of the k-th boundary homomorphism, $\mathsf{Z}_k = \ker \partial_k$.

Definition A.2.19. k-Boundary

A *k-boundary* is a chain that is the boundary of a $(k+1)$-chain, $c = \partial d$ with $d \in \mathsf{C}_{k+1}$. Since ∂ commutes with addition, we have a *group of p-boundaries*, denoted by $\mathsf{B}_k = \mathsf{B}_k(\mathsf{K})$, which is again a subgroup of the k-chains. In other words, the group of k-boundaries is the image of the $(k+1)$-st boundary homomorphism, $\mathsf{B}_k = \mathsf{B}_k(\mathsf{K})$, which is again a subgroup of the k-chains.

The fundamental property that makes homology work is that the boundary of a boundary is necessarily zero.

Theorem A.2.3. *Fundamental Lemma of Homology*

$\partial_k \partial_{k+1} d = 0$ for every integer k and every $(k+1)$-chain d.

Finally, we can define a homology group and the respective Betti number.

Definition A.2.20. Homology Group

The *k-th homology group* is the k-th cycle group modulo the k-th boundary group, $\mathsf{H}_k = \mathsf{Z}_k \,/\, \mathsf{B}_k$.

Definition A.2.21. Betti number

The k-th *Betti number* is the rank of the k-th homology group, $\beta_k = \operatorname{rank} \mathsf{H}_k$.

These terms, specifically the Betti number, can summarise the topology of our simplicial complexes. Simply put, β_0 counts the number of "connected" components, β_1 counts the number of "holes" (for example, the unshaded triangle in Figure A.2) and β_2 the number of "voids". Beyond two, we have to interpret the meaning of the Betti number abstractly, not geometrically. However, when constructing the simplicial complexes in Figure A.1, we fixed the value of ϵ. Without knowing the correct value for ϵ, PH considers the set of simplicial complexes generated by monotonically increasing $\epsilon \in (0, \infty)$ in a continuous way. Then, we can assess true topological features as those that persist longer, and disregard the less persistent as noise.

A.2.6 Persistent Homology

To begin considering the simplicial complexes constructed by continuously varying ϵ, we introduce the definition for such a construct as the filtration of f.

Definition A.2.22. Filtration of a Simplicial Complex

A *filtration of a simplicial complex* K is a nested family of subcomplexes $(K_{\epsilon'})_{\epsilon' \in T}$, where $T \subseteq \mathbb{R}$, such that for any $\epsilon, \epsilon' \in T$, if $\epsilon' \leq \epsilon$ then $K_{\epsilon'} \subseteq K_{\epsilon}$, and $K = \cup_{\epsilon' \in T} K_{\epsilon'}$.

More than in the sequence of complexes, we are interested in the topological evolution, as expressed by the corresponding sequence of homology groups. For every, $i \leq j$ we have an inclusion map from the underlying space of K_i to that of K_j and therefore an induced homomorphism, f_p^{ij} :$\mathsf{H}_p(K_i)$ for each dimension p. The filtration thus corresponds to a sequence of homology groups connected by homomorphisms,

$$0 = \mathsf{H}_p(K_0) \to \mathsf{H}_p(K_1) \to \cdots \to \mathsf{H}_p(K_n) = \mathsf{H}_p(K) \tag{A.3}$$

again one for each dimension p. As we go from K_{i-1} to K_i, we gain new homology classes and we lose some when they become trivial or merge. We collect the classes that are "born" at or before a given threshold and "die" after another threshold in groups.

Definition A.2.23. Persistent Homology Groups, Persistent Betti Numbers

The *p-th persistent homology groups* are the images of the homomorphisms induced by inclusion, $\mathsf{H}_p^{ij} = \mathsf{im} f_p^{i,j}$, for $0 \leq i \leq j \leq n$. The corresponding *p-th persistent Betti numbers* are the ranks of these groups, $\beta_p^{i,j} = \mathsf{rank}\ \mathsf{H}_p^{i,j}$.

A.2.7 Visualising Persistence

Now, given a point cloud data set, we can apply our changing ϵ and record the filtration. With this information, we can now display our results in many ways.

When Edelsbrunner et al. [304] published their work which made the computation for PH accessible, they visualised the results in a way which can be described as an amalgamation of the following diagrams.

The first method to visualise the results is through a barcode. For each k-th Betti number being considered, a line is drawn for when the topological feature is recorded to exist. Where the topological feature is to begin existing for the value of ϵ, it is called the *birth*, and at the end, the *death*. Figure A.3e displays an example of such a visualisation for an arbitrary point cloud data set.

In a similar way to a barcode, an alternative method of visualisation is to plot the birth against the death. Figure A.3f is such a visualisation, called a persistence diagram. We note a few of the significant features. First, in addition to the axes, there is a black diagonal line indicating the values for which birth is equivalent to death. This is because a topological feature cannot "die" before it has existed, thus we expect no points to exist in the lower segment. To compile the information for the different elements of our homology group into one figure, we have indicated the points for the 0-th dimension in orange and those for the 1-st dimension in blue.

To understand the need for further methods of visualisation, let us analyse the example case given in Figure A.3. The point cloud data set is made up of 18 data points. The Čech complex is displayed at several values of ϵ. Despite not knowing the specific context of the data, we are still able to identify topological features. Consider the elements for the 0-th

homology group. A point is a connected component in its own right, so we begin with all 18 recorded with birth when $\epsilon = 1$. When $\epsilon = 5.0$ we see that some of the data points now form an edge with one another, resulting in the death of one of the two connected components and a continuation of the other, whichever one is chosen to die or continue is insignificant. At $\epsilon = 7.0$, we can observe an element in the 1-st homology group and the data points have now formed four distinct connected components, noting that the far-right component is still a vertex in the simplicial complex. Finally, at $\epsilon = 9.5$, we have one connected component and two holes within that. Beyond this value of ϵ, more data points will form higher dimensional simplices until there is no change in the complexity, and hence on the barcode, we will always have the one connected component persisting.

Having analysed the filtration of an exemplar point cloud data set, the key principle of persistence in PH comes in. Here is where the current problem in the literature lies, what are the topological features that truly represent the data? Here, we might be able to identify two distinct structures that form in the data. One can observe when $\epsilon = 7.0$ (see Figure A.3c) and the other when $\epsilon = 9.5$ (see Figure A.3d). For the former, by eye, we can assume this structure accurately describes the data with the connected components appearing quite distinct. However, when referring to the 1-st dimensional homology group, the most persistent feature is beyond this value of ϵ and can be observed in the simplicial complex at $\epsilon = 9.5$ as the larger of the two holes. On the barcode, the difference in the significant connected components is hard to distinguish although if we can distinguish which are the true 1-st dimensional topological features, then we can assume that the same value of ϵ is also true for the connected components. On the other hand, we are not trying to establish the true value of ϵ in this analysis. Recognising the persistent features at other values of ϵ can identify sub-structures or meta-structures in the data beyond which is the most significant. Comparing features on a barcode is not as simple as on a persistent diagram, hence the popularity of them. When looking at a persistent diagram, the most persistent features are the points with the largest perpendicular distance away from the diagonal. Therefore, defining the parallel line to distinguish between the persistent features and those considered as noise created in

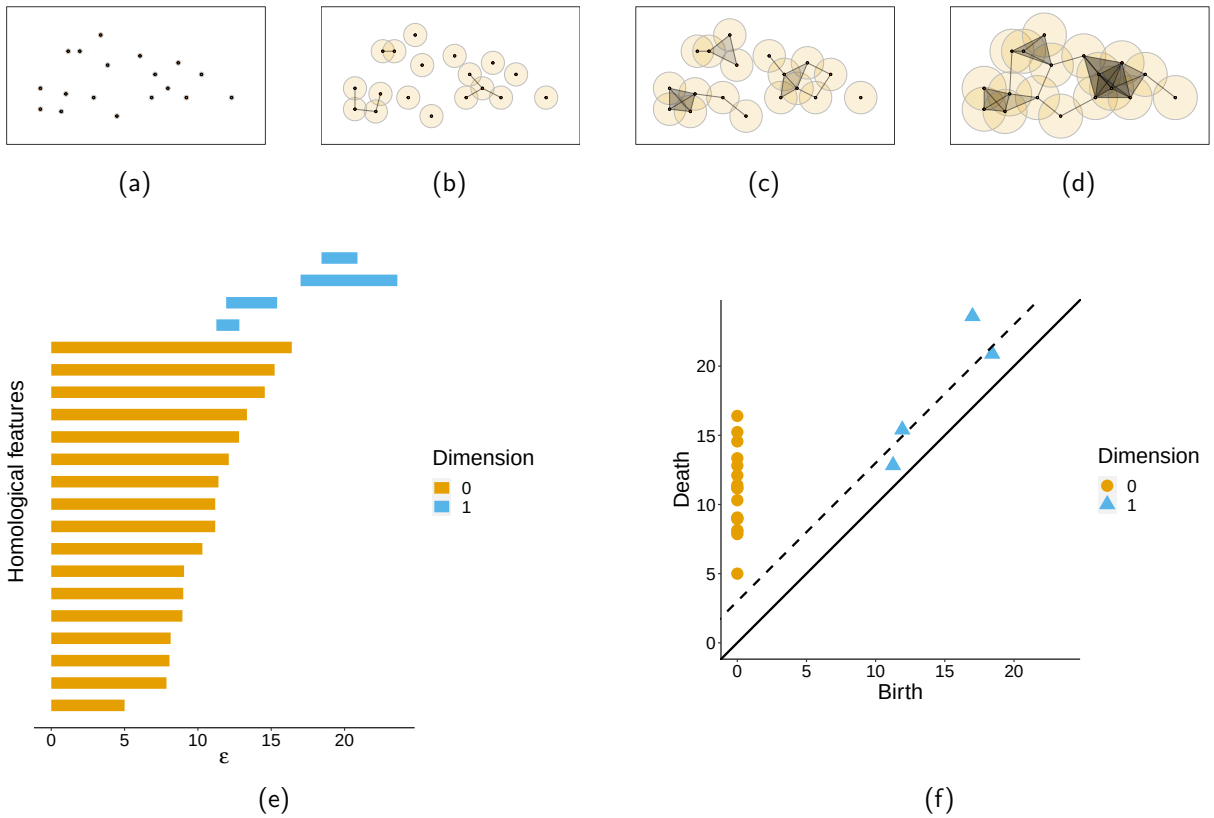


Figure A.3: The different ways the results of PH can be visualised and fed into further analysis of the topological features. (a) - (d) Depiction of the Čech complex is visualised for significant values of $\epsilon = \{1, 5, 7, 9.5\}$, respectively. Inspired by the example from [5]. (e) A barcode for the point cloud data set is shown in (a) - (d). (f) A persistence diagram for the point cloud data set is shown in (a) - (d), with a threshold at $t = 3$ (black dashed line). In (e) and (f) we show features of the 0-th dimensional homology group (orange bars / orange circles) and features of the 1-st dimensional homology group (blue bars / blue triangles).

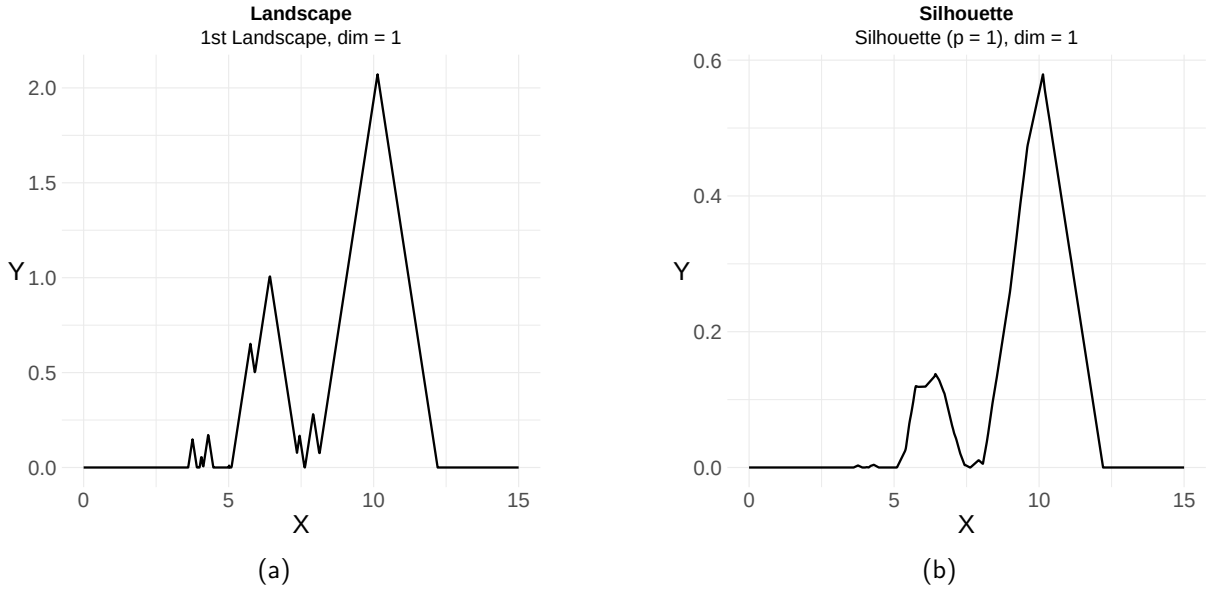


Figure A.4: Landscape and Silhouette of the 1-st dimensional homology group features of the example.

the data is crucial. Especially in an example similar to ours, where the current line drawn for features persisting over $t = 3$ can leave topological features balance between holding valuable information of our data or being disregarded altogether.

Further visualisations build upon a persistent diagram and make efforts towards distinguishing between significant and insignificant features. In Figure A.4 we present the Landscape and Silhouette introduced and studied by Bubenik [305], Chazal [306] and Chazal, Fasy, Lecci, Michel, Rinaldo, and Wasserman [307]. A persistent landscape converts the information from a persistent diagram and re-scales the most persistent with ranks before using statistical analysis to try and identify the significant topological features.

There now exists a plethora of coding libraries, readily available via open-source programs to implement and compute PH. For example, in C++ alone we have the GUDHI [308], Dionysus [309], PHAT [310] libraries, with implementations of the work by Fasy et al. [311] and Chazal et al. [307] in the TDA package [199] for the coding language R and in proceeding work we refer to RSMLM [164] as a library that implements work by Chazal et al. [4], also in R.

A.3 Heterogeneity Scores

Heterogeneity Scores are a set of metrics which aim to quantify the diversity of species in an ecosystem. As such, the application of such metrics has already been addressed to problems outside of cancer research such as ecology [312], and sociology [313]. The mathematical formulation of these metrics was adopted from principles in information theory, where analogous concepts would help discern uncertainty in communication systems [254].

The adoption of ecological terminology in the TME has provided an intuitive understanding, albeit at a smaller scale, of this type of ecosystem. As a result, there has been a prolific body of literature investigating the effects of heterogeneity on patient survival, the reader is referred to Kashyap et al. [314] for an in-depth review. For example, Graf and Zavodsky [315] apply a variation of the Shannon Index, which they call the *Cell Family Entropy*, to categorise samples of Stage 3 colorectal cancer into Grades as well as a prediction of tumour recurrence.

A.3.1 Richness

The most basic and intuitive heterogeneity score is *richness*,S, which simply counts the number of observed cell phenotype sub-populations within a tumour sample, independently of their relative abundance.

A.3.2 Shannon Index

Shannon Index, H, takes into consideration not only the number of cell sub-populations S present, but also their relative proportions. Assuming that $p(x_i)$ represents the probability of observing a cell sub-population x_i, then it is defined as:

$$H = -\sum_{i=1}^{S} p(x_i)\log(p(x_i)) \tag{A.4}$$

Another way to view the Shannon Index is as a quantification of the degree of 'surprise', in other words, how likely we are to guess the phenotype of a randomly observed cell from

knowing the relative proportions they exist in a tumour sample. Shannon Index increases with richness and evenness and reaches its maximal value when the cell sub-population distribution is uniform.

A.3.3 Gini-Simpson Index

Similarly, the *Gini-Simpson Index* considers both richness and relative abundance and is defined as:

$$G = \sum_{i=1}^{S} p(x_i)^2 \tag{A.5}$$

The Gini-Simpson Index describes the probability of sampling the same phenotype twice from the tumour. Unlike the Shannon Index, the Gini-Simpson Index decreases with increasing diversity. It is to be noted, that the Gini-Simpson has a bias to the most abundant cell phenotype, and so is regarded as a measure of dominance concentration [316].

A.3.4 Hill numbers and the Diversity Profile

The *Hill numbers*, ${}^{\alpha}D$, [316] constitute a family of heterogeneity scores under a unified mathematical expression:

$${}^{\alpha}D = \left(\sum_{i=1}^{S} p(x_i)^{\alpha} \right)^{1/(1-\alpha)} \tag{A.6}$$

It can be shown that richness S, Shannon Index H, and Gini-Simpson Index G are all special cases of ${}^{\alpha}D$:

$${}^{0}D = S \tag{A.7}$$

$${}^{1}D = \exp(H) \tag{A.8}$$

$${}^{2}D = 1/G \tag{A.9}$$

where choosing lower values of α give emphasis to rarer phenotypes [316].

An interesting application of the Hill numbers in the biological domain can be found in immunology, where immune repertoires from different patients that are not directly comparable were mapped in a common diversity index space to predict immune status called the *Diversity profile* [166]. The technique was interested in how diversity and evenness profiles can predict the immunological status of immune repertoires, primarily at what values of α their Diversity profiles intersect.

A.3.5 Renyi entropy

Similarly to the Hill numbers, *Renyi entropy* [317] is a generalisation of multiple entropic measures. In the context of tumour heterogeneity, Renyi entropy has been previously used to quantify the diversity of the tumour lung T cell receptor repertoires [318]. Renyi entropy is defined as:

$$H = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{S} p(x_i)^{\alpha} \right) \tag{A.10}$$

Similarly to the Hill numbers, it can be shown that, for different values of α, Renyi entropy is equivalent to other entropic expressions:

$$H_0 = \log R \tag{A.11}$$

$$H_1 = S \tag{A.12}$$

$$H_2 = -\log \sum_{x \in X} p(x)^2 = -\log D \tag{A.13}$$

$$H_\infty = -\log \max_x p(x) \tag{A.14}$$

where R is the Hartley entropy, S is the Shannon Index, D is the Collision entropy, and Eq. A.14 is the Min-entropy.

A.3.6 Rao's quadratic entropy

The indices and quantification methods discussed so far consider both the richness and relative abundance of phenotypes but ignore the similarity of these phenotypes. *Rao's quadratic entropy* [319] accounts for that by incorporating a distance metric $d(x_i, x_j)$ and is defined as:

$$Q = \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} d(x_i, x_j) p(x_i) p(x_j) \tag{A.15}$$

Depending on the *a priori* knowledge of the similarity between different cell markers, a weighted matrix can be constructed to represent this. This is then normalised so that the determinant is equal to 1 and we have a scale which is comparable between samples.

A.4 Alternative Spatial Correlation Metrics for Multitype Point Patterns

Here we present alternative metrics for measuring spatial correlation in multitype point patterns. We illustrate their application on the artificial point patterns presented in Figure 2.7.

A.4.1 Nearest Neighbour Distribution with the G-function

The *Nearest Neighbour Distribution*, commonly referred to as the G-function [320], computes the distance between all points and its nearest neighbour and then compares the distribution of distances to the homogeneous spatial Poisson point process. Instead of a global value that describes the spatial organisation across all length scales, the G-function provides a value across a range of length scales. This can also be compared to a theoretical value which is also computed for the same range. As a consequence, the observed G-function describes for which radii the distribution of cells deviates from Null Distribution and whether this difference is statistically significant. The theoretical G-function for the homogeneous spatial Poisson point process can be determined analytically as,

$$G_P(r) = 1 - e^{-\lambda \pi r^2} \tag{A.16}$$

where λ is the mean density for the spatial Poisson point process.

The estimated G-function for a univariate point pattern can then be calculated as,

$$\hat{G}(r) = \frac{1}{n} \sum_{i=1}^{n} I \left(\min_{x_j \in X_j} d(x_i, x_j) \leq r \right) e_{i,j} \tag{A.17}$$

where $d(x_i, x_j)$ is the distance between points $x_i, x_j \in X$, often measured using the Euclidean distance, and $I \left(\arg\min_{\forall x_j \in X_j} d(x_i, x_j) \leq r \right)$ is the indicator that equals 1 if a pair of points have a distance less than or equal to r. The term $e_{i,j}$ is the edge correction weight which depends on the choice of edge correction [64].

In our analysis, we are trying to discern the spatial correlation between point types within the point pattern. As such, the G-function can be adapted to analyse the Nearest Neighbour Distribution for a multitype point pattern,

$$G_{ij}(r) = \frac{1}{n_i} \sum_{k=1}^{n_i} I \left(\min_{x_j \in X_j} d(x_{i_k}, x_j) \leq r \right) e_{i_k,j} \tag{A.18}$$

where we have a similar equation as Eqn. A.17, but now we have n_i as the cardinality of the point set $X_i \subset X$, and $x_{i_k} \in X_i$ and $x_j \in X_j$ as the points in respective point type subsets of the multitype point pattern, i.e. $X_i, X_j \subset X$. The term $e_{i_k,j}$ is the edge correction weight which depends on the choice of edge correction [64].

The estimate of the G-function, $\hat{G}(r)$, is compared to the analytically derived value of the G-function for the spatial Poisson point process. In addition to this, simulating point sets for the same density for each cell type provides a *pointwise* significance interval to perform hypothesis testing. That is, for a specific value of r we can compare the observed value to the Null Hypothesis with a significance level dependent on the number of simulations in which the minimum and maximum values of the interval are computed for [167]. The equation for such significance intervals is,

$$\alpha = k/(1 + n_{\text{sim}}). \tag{A.19}$$

where α is the significance level, n_{sim} is the number of simulations, and k is the rank of the envelope value amongst the n_{sim} simulated values. So, a rank of 1 means that the minimum and maximum simulated values for that r will be used as bounds of the significance interval.

Presented in Figure A.5 are the multi-type G-distributions for the three artificial point patterns produced in Figure 2.6. Since the relationship for the nearest neighbour is asymmetric, i.e. the nearest neighbour of one point is not guaranteed to be the nearest neighbour of its nearest neighbour, we compute the permutations of pairwise cell type comparisons for each point pattern. In this case, we have two cell types, so the number of ordered comparisons is four. This is why we have four G-distributions for each artificial point pattern. In each plot, we have three pieces of information. First, we have the analytical G-distribution for a spatial Poisson point process. In addition, we have computed the significance interval for the comparative spatial Poisson process at the 5% significance level. Finally, we have computed the observed values of the G-function for $r \in \{2n | n \in 1, 2, \ldots, 125\}$.

In all of the G-function plots, we observe the hard-shelled property of the coordinates and see no contribution of cells less than $r \approx 10$. It is for values of r above 10 that we can critically look at the interpretation of the observed G-functions.

For our first artificial point pattern, our intuition of the point pattern's construction would indicate that for the within-cell type comparisons, the G-function would have a higher value than the spatial Poisson point process for small values of r, i.e. representing the points much closer together within the cluster. In addition, the G-function would be larger for smaller values of r between cell types and then slightly lower proportions of cells for all other values of r to compensate in the rest of the domain the sparser distribution of points to have a similar global density of points.

On first inspection, the intuition described previously is not immediately present. For Cell Type 1 to Cell Type 1 comparison the G-function does not reject the Null Hypothesis at the 5% significance level for small values of r except for radii already known to be related to

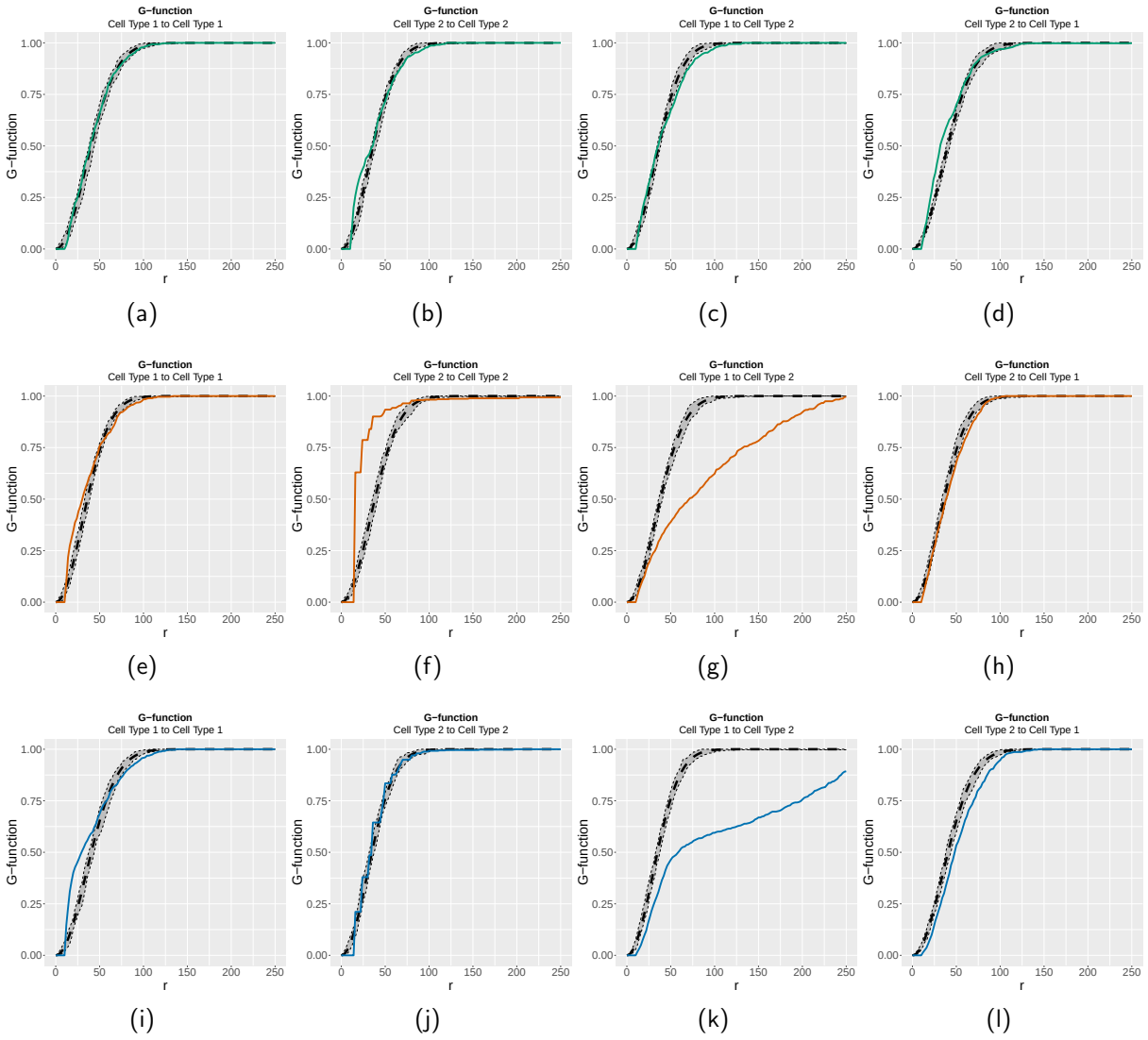


Figure A.5: G-function plots for (a) - (d) independent, (e) - (h) proportional, and (i) - (l) inverse spatial point patterns. Since pairwise comparisons are asymmetric, we have the G-function plot for each point pattern for Cell Type 1 to Cell Type 1, Cell Type 2 to Cell Type 2, Cell Type 1 to Cell Type 2, and Cell Type 2 to Cell Type 1. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level.

the hard-shelled property of the points. However, this signal is more apparent for the within Cell Type 2 comparison, with the Null Hypothesis rejected at the 5% significance level for $r \in [12, 28]$. Then, for between cell type comparisons, we see that the cumulative distribution function increases in both pairwise comparisons to just reach higher than the significance interval at $r \in [16, 26]$ and $r \in [16, 50]$ in Cell Type 1 to Cell Type 2 and Cell Type 2 to Cell Type 1 comparisons, respectively. The effect size is small but significant at the 5% significance level. For our second prediction, the Cell Type 1 to Cell Type 2 comparison is then significant for larger values, rejecting the Null Hypothesis at the 5% significance level for $r \in [44, 100]$, although, this is not as consistent for the Cell Type 2 to Cell Type 1 comparison with only $r = \{70, 74, 98, 104, 106, 108, 116\}$ considered significant at the 5% significance level. Therefore, a conclusion from this analysis of the G-function could be that Cell Type 1 is likely distributed as a spatial Poisson point process, while Cell Type 2 is clustered to itself for smaller values of r considered. Furthermore, one could conclude that both cell types are positively spatially correlated for small values of r which is then compensated in an inhibition for larger values of nearest neighbour distances for larger r.

Next, we provide intuition for the next Nearest Neighbour Distribution to be analysed which is the proportional point pattern, where proportional refers to the proportional probabilistic spatial placement of Cell Type 2 according to the local density of Cell Type 1. Here, we suggested that the Cell Type 1 to Cell Type 1 comparison would present higher values of $G(r)$ for small values of r than the Null Hypothesis since Cell Type 1 is aggregated together. Then, the Cell Type 2 to Cell Type 2 comparison would present a similar relationship since they would be near other Cell Type 2 points that are also spatially correlated with Cell Type 1. Therefore, the G-function for Cell Type 1 to Cell Type 2 might be inhibited by self-clustering before switching to values above the Null Hypothesis at some midpoint before flattening to represent the smaller chance of finding larger distances between Cell Type 1 and its nearest Cell Type 2. On the contrary, Cell Type 2 to Cell Type 1 would be expected to produce higher values of $G(r)$ for smaller values of r to then a flattening and then inhibition at larger r to compensate for the smaller chance of finding larger distances between Cell Type 2 and its

nearest Cell Type 1.

When analysing the G-function for the proportional point pattern, the first prediction is supported with the within Cell Type 1 comparison presenting a $G(r)$ value which rejects the Null Hypothesis at the 5% significance level for $r \in [16, 28]$ even if with a small observed effect size. Moreover, the Cell Type 2 comparison has a much larger effect size and for a large range of values of r rejects the Null Hypothesis for $r \in [16, 56]$. It is at this point that our predictions are less apparent. First, for the Cell Type 1 to Cell Type 2 comparison, a consistent inhibition is observed in the G-function. This is in contradiction to our prediction that the nearest Cell Type 2 for Cell Type 1 is the average distance within the cluster to the perimeter, which can be predicted analytically as $r/3 \approx 41.667$ for the cluster presented with a radius of about 125. On a second inspection of the spatial point pattern, it might become apparent that the spatial correlation of Cell Type 2 is so high that few Cell Type 2 points exist in the rest of the domain for Cell Type 1 points to have a Cell Type 2 nearest neighbour at small values of r. As such the nearest neighbour relationship is overshadowed by lower values which are less than the significance interval at the 5% significance level for values of $r \leq 250$. We proceed with the G-function for Cell Type 2 to Cell Type 1. Again, the intuition we provided does not necessarily match with the properties of the G-function. Instead, we observe a G-function which is indistinguishable from the spatial Poisson point process at the 5% significance level for most r, explicitly the Null Hypothesis is rejected for $r = \{58, 60, 62, 68, 84, 86, 88\}$. Therefore, one might infer from this series of G-function plots that the point pattern that it represents has some clustering within each of the cell types, however, Cell Type 1 is inhibited from Cell Type 2 and little difference is seen from CSR between Cell Type 2 and Cell Type 1 nearest neighbours. Hence, the G-function is limited in its interpretative power on the spatial correlation between cell types.

For our last point pattern, we provide intuition for the next Nearest Neighbour Distribution to be analysed which is the inverse point pattern, where inverse refers to the inversely proportional probabilistic spatial placement of Cell Type 2 according to the local density of Cell Type 1. As before, the within Cell Type 1 G-function would be expected to represent the same as

that already discussed in the proportional point pattern. The difference in this point pattern is in the spatial arrangement of Cell Type 2. Here, we do not expect Cell Type 2 to spatially correlate with itself, only that it is more likely to be positioned away from the high-density aggregation of Cell Type 1. As such, within cell type comparison would be expected to look like that of CSR. On the contrary, for between cell type comparisons, the distribution is expected to be lower than that for CSR. This should be more prominent again for the Cell Type 1 to Cell Type 2 comparison as the high-density aggregation of cells will contribute larger distances to their nearest Cell Type 2 neighbour depending on the coefficient of inverse proportionality.

In summary, the intuition for the expected observed Nearest Neighbour Distributions for our inverse point pattern is as expected. We have discussed the characteristic curve already of cell types which are clustering, the same as that in Figure A.5i. Also, Figure A.5j presents a characteristic curve for the spatial Poisson point process, albeit for a more step-like function due to the coarser grid which Cell Type 2 can exist on due to the artificial set-up. Finally, both between cell type comparisons express inhibition that is statistically significant at the 5% significance level from CSR. Even the Cell Type 1 to Cell Type 2 comparison has a slower ascent for increasing values of $r \geq 60$, representing the inhibition of Cell Type 2 points from the high-density aggregation of Cell Type 1 points in the centre.

In our validation of the technique, we have emphasised the identification of structures considered significant in the HALO1079 data set. Specifically, when the correlation of cell phenotypes is related to density. However, some cell phenotypes can have such low density that their presence alone is significant. In these scenarios, the utilisation of the G-function provides a framework to quantify this since we are now only interested in the nearest point and not local calculations of density. Specifically, if only a few cells of a rare phenotype are present the G-function can begin to interpret if a point is nearer to a target cell phenotype than if it were arranged by CSR.

In conclusion, the G-function can provide limited insight into the types of spatial organisation we expect to identify in our HALO1079 data set. Specifically, the features we expect are not present. Except for an absence of significantly lower values between Cell Type 2 to

Cell Type 1 in the observed value of $G(r)$, the measurable difference between the proportional point pattern and inverse point pattern is almost indistinguishable. For this reason, we continue to discuss other multitype spatial correlation metrics which will be able to reflect these important differences.

A.4.2 J-function and F, the Empty Space Distribution

The J-function [321] builds upon the computation of the G-function discussed previously. The multitype point pattern distribution is defined as

$$J_{ij}(r) = (1 - G_{ij})/(1 - F_j), \tag{A.20}$$

where G_{ij} is the G-function between points $x_i \in X_i \subset X$ and points $x_j \in X_j \subset X$, and F_j is the F-function for point set $X_j \subset X$.

The F-function referred to here, often called the *Empty Space Distribution*, gives the probability of finding a point within distance r of an arbitrary location in the domain. The F-function is estimated for a sample of points $P_k \subset P$, where $P \subset \mathbb{R}^2$ and covers the domain of X, and given point pattern $X_j \subset X$ is then equal to,

$$\hat{F}_j(r) = \frac{1}{n} \sum_{k=1}^{n} I \left(\arg \min_{\forall x_j \in X_j} d(x_j, p_k) \leq r \right) e_{i,j} \tag{A.21}$$

where $p_k \in P_k$, $d(x_j, p_k)$ is the distance between a point in X_j and the fixed point p_k, often using the Euclidean distance, and $I\left(\arg\min_{\forall x_j \in X_j} d(x_j, p) \leq r\right)$ is the indicator function that equals 1 if the distance between $p_k \in P_k$ and any $x_j \in X_j$ is less than or equal to r. The term $e_{i,j}$ is the edge correction weight which depends on the choice of edge correction. Therefore, the J-function is the ratio of the probability of finding no point within distance r of any other point to the probability of finding no point within distance r of any location in the domain.

This function is identically equal to 1 if $X_i \subset X$ and $X_j \subset X$ are (marginally) independent. However, having a J-function taking value 1 everywhere should not be seen as a characterisation of independence. In particular, $J_{ii}(r) \equiv 1, \forall r$ is not a sufficient condition for

X_i to be a spatial Poisson point process [322]. Intuitively J_{ij} is a comparison between the distributions of the distances to the nearest type j point, measured from (a) an arbitrary fixed point in $\mathbb{R}^d$, and (b) a typical type i point. The denominator of Eqn. A.21 is the unconditional probability of the event that there is no type j point within a distance r of an arbitrary fixed point. The numerator is the "conditional probability" of the same event given that there is a type i point at the arbitrary fixed point. As in the univariate case [321] the value 1 is obtained when there is no spatial correlation: if X_i and X_j are independent processes, then standard calculations give $J_{ij}(r) \equiv 1$, $\forall r$. Values $J_{ij}(r) > 1$ can be interpreted as indicating inhibition (of type j points to type i points) since this is equivalent to $G_{i,j} < F_j$, i.e. the presence of a type i point decreases the probability of finding a type j point nearby. Similarly, values less than 1 suggest a positive spatial correlation.

The J-function results for our artificial point patterns are presented in Figure A.6. Similarly to the Nearest Neighbour Distribution plots (see Figure A.5) we have the significance interval at the 5% significance level, obtained in the same way as the previously described Nearest Neighbour Distribution. Note that the definition of J_{ij} is not symmetric in i and j. Therefore, we have to consider all permutations of the different cell types, resulting in 4 J-function plots for each artificial point pattern.

For our first multitype point pattern we have both cell types aggregated in the middle independently of the other cell type, with a remaining percentage of points distributed with CSR. As a consequence, we expect the observed J-function to present a positive spatial correlation for small values of r within cell types. For between cell types, the J-function would be expected to measure stochastic independence between the cell types.

Analysing the observed J-function to the equivalent value for stochastic (marginal) independence at the 5% significance level, we observe no positive spatial correlation within Cell Type 1. Still, we do for within Cell Type 2 at $r \in [14, 50]$. Therefore, the effect size is not large enough in the first instance for the J-function to reject the Null Hypothesis at the 5% significance level to match our knowledge of the point pattern's construction. When looking at the between-type comparisons, a similar set of unexpected observations is inferred. In both

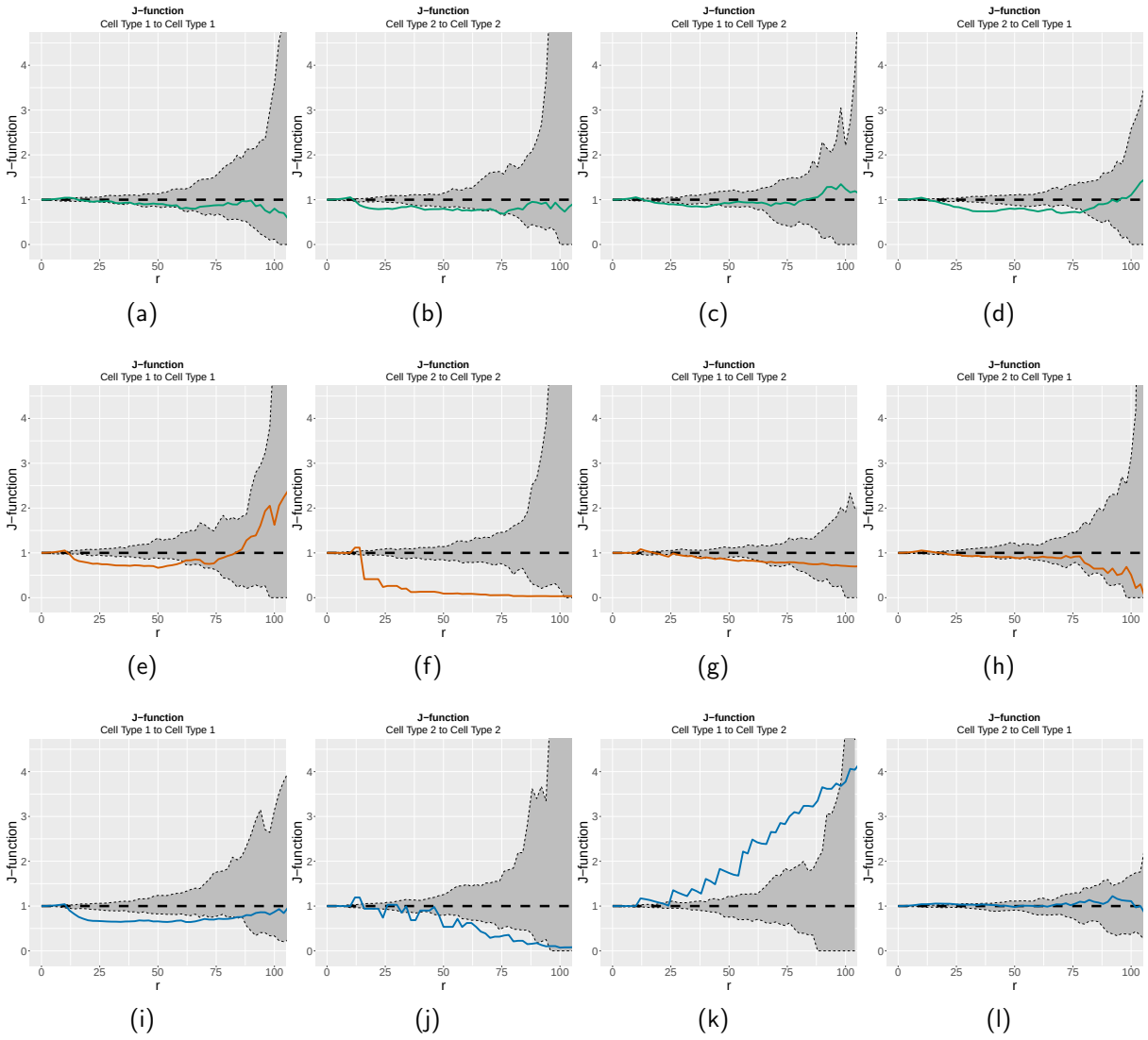


Figure A.6: J-function plots for (a) - (d) independent, (e) - (h) proportional, and (i) - (l) inverse spatial point patterns. Since pairwise comparisons are asymmetric, we have the J-function plot for each point pattern for Cell Type 1 to Cell Type 1, Cell Type 2 to Cell Type 2, Cell Type 1 to Cell Type 2, and Cell Type 2 to Cell Type 1. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level.

observations, the observed J-functions indicate a positive spatial correlation between the cell types at the 5% significance level. For the Cell Type 1 to Cell Type 2 and Cell Type 2 to Cell Type 1 comparisons we have pointwise significance at $r \in [14, 42]$ and $r \in [16, 78]$. Thus, our prior knowledge of the artificial point pattern is not discernible from the observations of the J-function or, at worst, suggests the opposite conclusion to reality.

We now refer to predicting the likely inference to be gained from the J-function for the proportional point pattern. In this scenario, we expect both within-type J-functions to represent a positive spatial correlation for small values of r. This is because Cell Type 1 is distributed in such a way and Cell Type 2 is constructed to be spatially correlated with Cell Type 1 and so other Cell Type 2 points will likely be in the vicinity too. Then, for between-type comparisons, the J-function would be expected to show a positive spatial correlation for both permutations, more likely for larger values of r since the level of self-clustering which might appear as an inhibition for smaller values of r initially.

When moving to the analysis of the proportional point pattern, the expectation and observation are mostly matched. First, the observations for the within-type comparisons are as expected, both show a positive spatial correlation for most values of r. The relationship is more pronounced for Cell Type 2 to Cell Type 2 with almost all values of r in the computed range considered to be positively correlated at the 5% significance level. When looking at the between-type comparisons, the relationship matches more directly with the theoretical distribution discussed for the Nearest Neighbour Distribution. That is, the nearest Cell Type 2 from a Cell Type 1 point is the average distance within the cluster to the perimeter, which analytically is $r/3 \approx 41.667$ for a cluster of radius 125, estimated from viewing the final boundary of the aggregation. Despite a small effect size, this prediction is borne out with the significant values of r for Cell Type 1 to Cell Type 2 being $r \in [32, 44] \setminus (38, 42)$ at the 5% significance level, respectively. Less prominently, this is reflected in the Cell Type 2 to Cell Type 1 J-function with only $r = 38, 42, 44$ presenting as values of r that reject the Null Hypothesis at the 5% significance level.

Finally, we come to the analysis of the inverse point pattern. In this instance, we posit

that the J-function would consider a positive spatial correlation within Cell Type 1 but not necessarily for the within Cell Type 2 comparison. On the other hand, a significant inhibition behaviour is expected for the spatial association of Cell Type 1 to Cell Type 2. The Cell Type 2 to Cell Type 1 J-function might see some inhibition characteristics but the effect size could be small. This is because those Cell Type 1 points not part of the aggregation form an otherwise random distribution measured from a typical Cell Type 2 point would present as no statistical correlation.

The observed J-functions for the inverse point pattern have characteristics that can be considered to match our intuition. First, the within Cell Type 1 comparison produces values of $J(r)$ that represent a positive correlation for values of $r \in [12, 70]$ at the 5% significance level, as predicted. For the within Cell Type 2 comparison, the observed values of $J(r)$ indicate a positive correlation at the 5% significance level for various values of r. On reflection, this could be a result of two effects at play. One has previously mentioned the finite placement of Cell Type 2 points for the construction of the point pattern, i.e. local density was computed for smaller windows which are the given probability for a Cell Type 2 point to exist within. The second effect that appears to be at play is the bias towards spatial correlation and this can be interpreted as the exclusion of Cell Type 2 from the higher-density region of Cell Type 1 points in a finite domain. Although edge correction has been applied in each computation, for the length-scale being calculated, $r \geq 50$, the exclusion of Cell Type 1 points to the periphery appears as a positive correlation within the subset. Next, we consider the between-type comparisons. The observed J-function for Cell Type 1 to Cell Type 2 presents a strong inhibition between the points for almost all r computed. Therefore, the J-function can identify the key property of the inverse point pattern. In the Cell Type 2 to Cell Type 1 comparison, the relationship is considered stochastically independent. Thus, our prediction for the distribution of noise for Cell Type 1 plays an overriding effect when comparing the two cell types.

In conclusion, we have shown the capabilities of the J-function for distinguishing the key characteristics of the three artificial point patterns. The J-function was able to determine differences between the proportional and inverse point patterns for between-type comparisons

where the G-function was less than obvious. However, the characterisation of the independent point pattern was unclear from the J-function. In addition, the effect size was measured to be small in almost all cases except for the positive correlation within Cell Type 2 for the proportional point pattern (see Figure A.6f) and the between Cell Type 1 to Cell Type 2 comparison (see Figure A.6k). The lack of improvement or comparable interpretive power of the J-function to the G-function is unsurprising since the former is reliant on the latter. It is because of this that we consider the following that provides an alternative metric for spatial correlation within multitype point patterns.

A.4.3 Ripley's K-function

The *Ripley's K-function* [63, 64] of a stationary point process X is is the second-moment property of a point pattern normalised by the density. As a consequence, we have the mean number of points within distance r of a point in our point set as,

$$\bar{x} = \lambda K(r), \tag{A.22}$$

where λ is the density of point process X in the domain whole domain and $K(r)$ is the respective Ripley's K-function for the point pattern at a given r.

We can estimate $K(r)$ for a given univariate point pattern as,

$$\hat{K}(r) = \frac{A}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} I(d_{i,j} \leq r) e_{i,j} \tag{A.23}$$

where A is the area of the domain considered, n is the number of points in X, and the sum is taken over all ordered pairs of points x_i and x_j in X. Here $d_{i,j}$ is the distance between the two points, often the Euclidean distance, and $I(d_{i,j} \leq r)$ is the indicator function that equals 1 if the distance is less than or equal to r. The term $e_{i,j}$ is the edge correction weight which depends on the choice of edge correction [64].

$\hat{K}(r)$ describes the density from a typical point i to points in X_j of the point process across length-scales. The observed K-function is then compared to the value of $K(r)$ for a spatial

Poisson point process, which has K-function equal to $K(r) = \pi r^2$. The estimate $\hat{K}(r)$ and subsequent deviations between this and the values of $K(r)$ for the Null Distribution values are then interpreted to suggest spatial clustering (when $\hat{K}(r) > K(r)$), spatial dispersion (when $\hat{K}(r) < K(r)$) or spatial regularity (when $\hat{K}(r) \equiv K(r)$) [145].

The equivalent K-function for considering pairs of point types in a multitype point pattern is,

$$\hat{K}_{ij}(r) = \frac{A}{nm} \sum_{k=1}^{n} \sum_{l=1}^{m} I(d(x_{i_k}, x_{j_l}) \leq r) e_{i_k, j_l} \tag{A.24}$$

where A is the area of the domain considered, n and m are the cardinalities of the respective $X_i, X_j \subset X$ point sets with $X_i \not\subset X_j$ and $x_i \in X_i$ and $x_j \in X_j$, repsectively. $I(d(x_i, x_j) \leq r)$ is the indicator function that equals 1 if the distance is less than or equal to r. The term $e_{i,j}$ is the edge correction weight, which depends on the choice of edge correction [64].

The K-function results for our artificial point patterns are presented in Figure A.7. Similarly to the previous spatial correlation metrics, we have the significance interval at the 5% significance level obtained in the same way as the aforementioned Nearest Neighbour Distribution[1]. Note that the definition of K_{ij} is symmetric in i and j. Therefore, we only need to consider all combinations of the different cell types, resulting in 3 K-function plots for each artificial point pattern.

We begin our analysis with the independently aggregated point pattern. Before analysing the Figures, we hypothesise the characteristics of the K-function which reflect our prior knowledge of the point pattern's construction. In this case, for the within-type comparison, we expect a positive spatial correlation to be detected. In addition, for the between-type comparison, Ripley's K-function would reflect the colocalisation of the individual point type aggregations as a positive spatial correlation between the cell types.

The analysis of Ripley's K-function for the type-independent, Gaussian distributed point pattern reflects the intuition very accurately. In all three plots of the K-function, the observed value of $K(r)$ rejects the Null Hypothesis at the 5% significance level. This effect is most

[1] See Section A.4, Subsection A.4.1

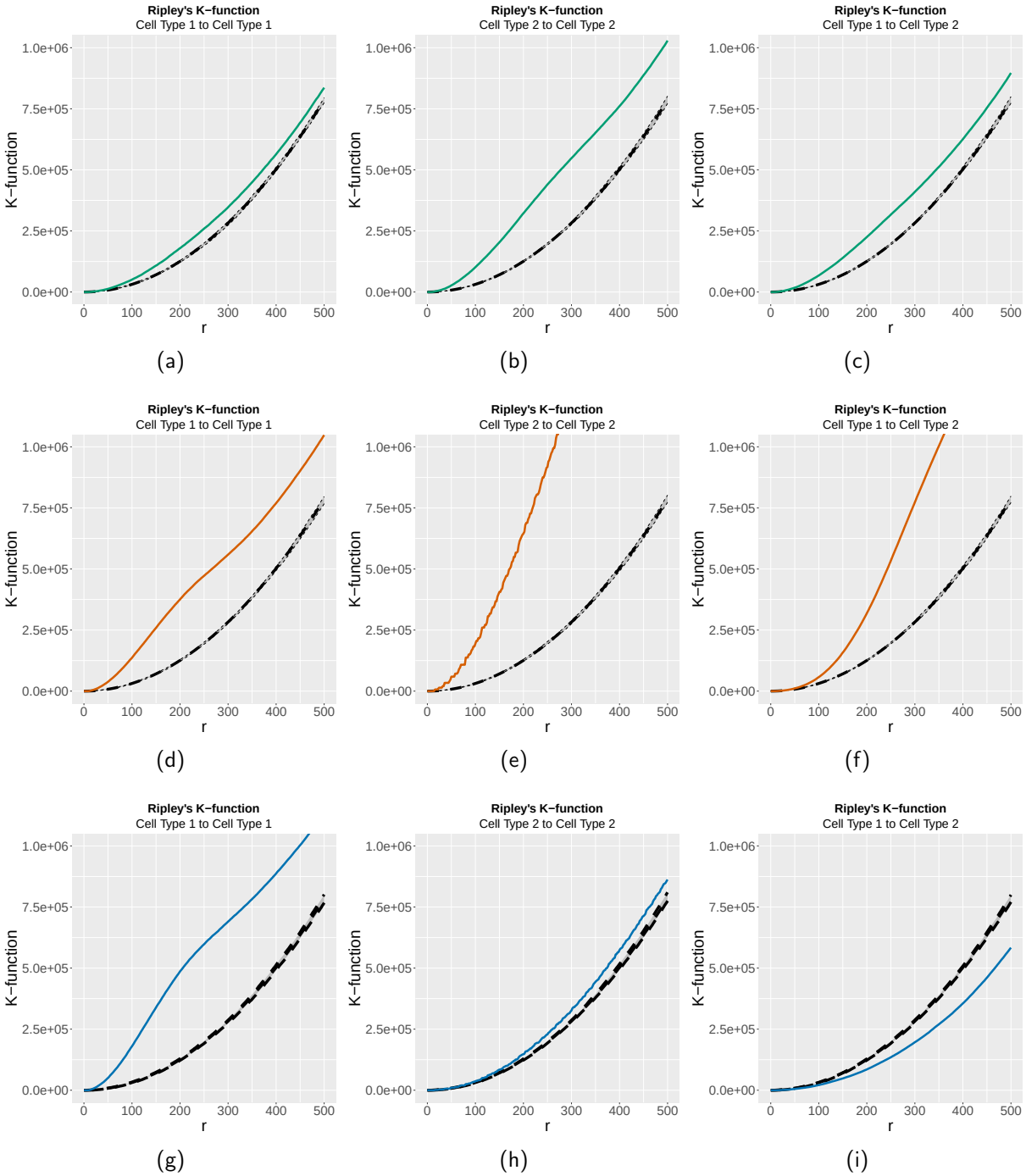


Figure A.7: Ripley's K-function plots for (a) - (c) independent, (d) - (f) proportional, and (g) - (i) inverse spatial point patterns. Since pairwise comparisons are symmetric, we have the K-function plot for each point pattern for Cell Type 1 to Cell Type 1, Cell Type 2 to Cell Type 2, and Cell Type 1 to Cell Type 2. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level.

pronounced in the Cell Type 2 to Cell Type 2 comparison, where the value increases at a higher gradient for smaller values of r, representing the higher density of points within the aggregation, and then the gradient decreases slightly beyond this. The same effect should be visible for the within Cell Type 1 comparison but as we have seen in previous spatial metrics, the effect size is much smaller, yet still significant for Ripley's K-function.

Next, we consider the K-function for the proportional point pattern. Here, we would expect the within Cell Type 1 K-function to reflect the characteristics discussed for a high-density aggregation. Then, the within Cell Type 2 comparison would reflect the characteristics of a high-density aggregation with a steeper gradient than that of Cell Type 1 since all points will likely surround the same aggregation of Cell Type 1 meaning it will reach its maximum value of $K(r)$ for r similar to the radius of the aggregation including the Cell Type 2 annulus around it. Finally, the between-type comparison should reflect some inhibition of points at the smallest scales of r, to reflect the self-clustering between Cell Types. Then, a steep increase will characterise the strong spatial correlation at the interfaces of both high-density aggregations.

We see from the observational values of the K-function that our intuition is matched almost exactly. Albeit difficult to observe at the scale of the y-axis in Figure A.7f, the final comparison does see a transition from an extended inhibition before transitioning to a strong positive correlation at the 5% significance level. Unlike before with the within Cell Type 1 comparison with only the hard-shelled property of our points presenting a rejection in the Null Hypothesis for $r \in [6, 10]$ and above the significant interval of the comparative spatial Poisson point process, the between-type comparison has a value of $K(r)$ considered significant at the 5% significance level for $r \in [4, 16]$. This then transitions for $r \in [18, 32)$ before rejecting the Null Hypothesis again at the 5% significance level for $r \in [32, 500]$. Hence, Ripley's K-function has been able to identify some nuance in the point pattern beyond global clustering or inhibition but the length scales with which this is affected.

The last artificial point pattern to consider is the inverse point pattern. The within Cell Type 1 comparison should follow the same distribution as described for the proportional point

pattern. The within Cell Type 2 comparison would be expected to follow largely a spatial Poisson point process except for a positive correlation with a small effect size due to the inhibition from the spaces of high-density Cell Type 1 aggregation. This characteristic combined with the negative correlation of Cell Type 1 to Cell Type 2 will reveal the construction of the point pattern.

The observations of the K-function for the inverse point pattern reflect our intuition exactly. That is, the within Cell Type 1 comparison represents a high-density aggregation as described for other point patterns. Next, the within Cell Type 2 comparison is positively correlated between points but with less of an effect size than is expected for a high-density aggregation unlike in the proportional point pattern. Lastly, the between-type comparison reflects an inhibition of points which is significant for most values of r at the 5% significance level.

In conclusion, Ripley's K-function has differentiated the three artificial point patterns and indicated the way the point patterns have been constructed with estimates for radii of high-density aggregations as well as nuances of within-type spatial correlation, such as in the analysis of the proportional point pattern. However, the visualisation can be difficult to quantify when interpreting increasing or decreasing gradient values since the value for the Null Hypothesis is increasing with $K(r) \propto r^2$. There are two transformations of Ripley's K-function that aid in this endeavour, first Besag's L-function which accounts for the proportional increase in the Null Distribution, and the PCF which focuses more specifically on the changes in gradient.

A transformation often considered is Besag's L-function which visualises the value for the Null Distribution as $y = x$. Then, an analysis can more easily observe values of the Ripley's K-function as deviating from this line. We provide details of Besag's L-function in Appendix A (Section A.4.4). On the other hand, the interpretation of these results is similar to the one provided for Ripley's K-function. Even slight changes in gradient can differentiate individual cell type distributions. To better characterise the differences in gradient and the effects of self-clustering, we consider a final spatial correlation metric: PCF.

A.4.4 Besag's L-function

Besag's L-function [323] is a linear transformation of Ripley's K-function,

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \tag{A.25}$$

where $K(r)$ is Ripley's K-function.

The reasoning behind this transformation is to make comparison easier to the Null Distribution. Whereas before the spatial Point process followed a value proportional to r^2 which makes visual interpretation difficult, Besag's L-function represents a comparison to a linear form of the Null Distribution. Subsequently, comparing with a spatial Poisson point process gives, $L(r) < r$ indicates a negative spatial correlation, $L(r) > r$ positive spatial correlation, and $L(r) \equiv 1$ similar to CSR.

The L-function results for our artificial point patterns are presented in Figure A.8. Similarly to the previous spatial correlation metrics, we have the significance interval at the 5% significance level obtained in the same way as the aforementioned Nearest Neighbour Distribution[1]. Note that the definition of L_{ij} is symmetric in i and j, as it is based on the K-function. Therefore, we only need to consider all combinations of the different cell types, resulting in 3 L-function plots for each artificial point pattern.

A.5 Algorithms for Identifying High-Density Aggregations

The following section discusses the different approaches for identifying high-density aggregations. Also, details for the implementation of the algorithms chosen for comparison in Chapter 3 are described.

Some methods rely on graph constructions, whereby each node is a cell coordinate and follows a rule or set of rules to be considered connected, via an edge, to another cell in the domain. A common rule is to state that the distance to another cell is within a pre-specified

[1] See Section A.4, Subsection A.4.1

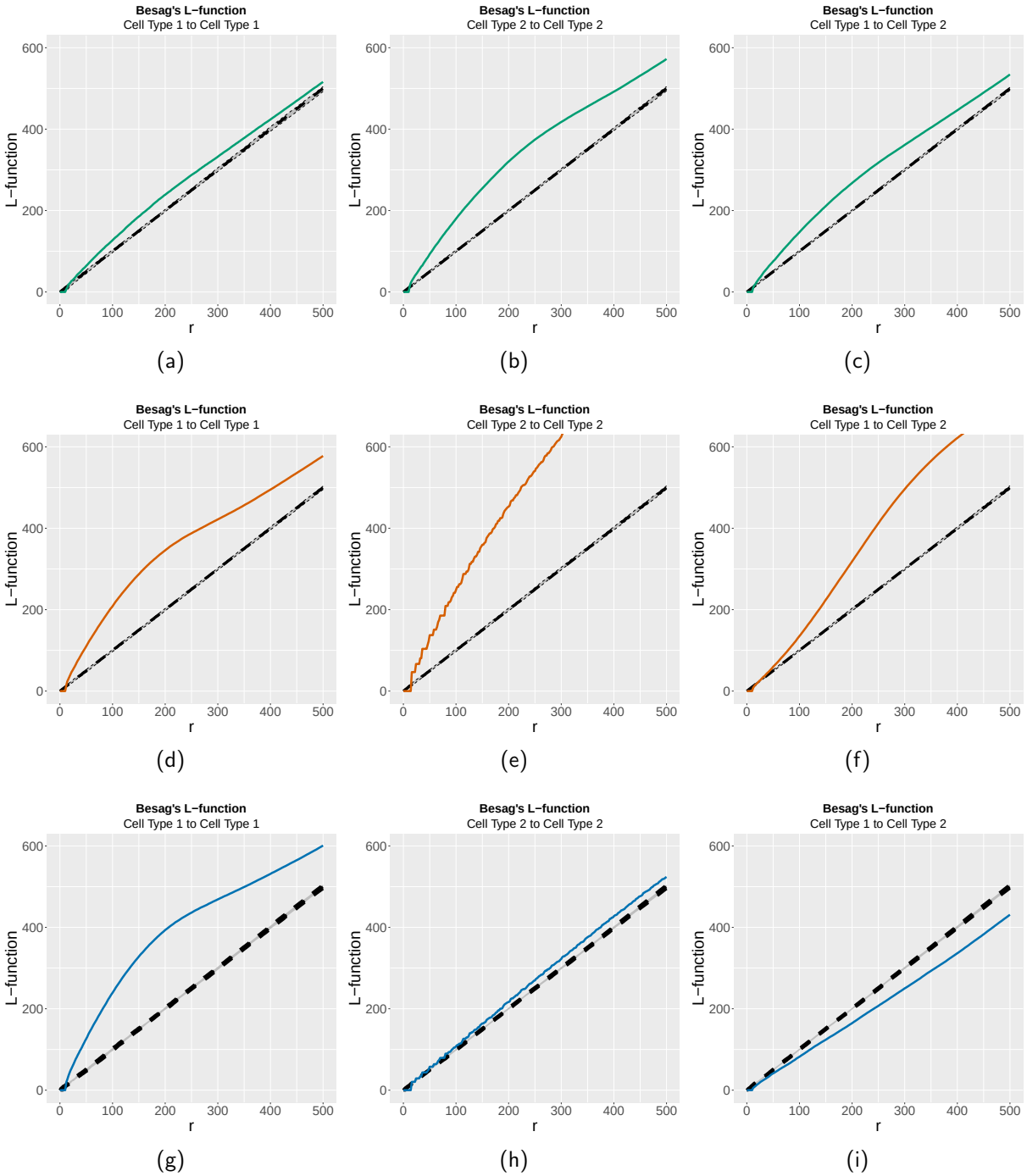


Figure A.8: Besag's L-function plots for (a) - (c) independent, (d) - (f) proportional, and (g) - (i) inverse spatial point patterns. Since pairwise comparisons are symmetric, for each point pattern we have the L-function plot for Cell Type 1 to Cell Type 1, Cell Type 2 to Cell Type 2, and Cell Type 1 to Cell Type 2. The solid coloured line represents the observed value while the dashed black line represents the theoretical value for the spatial Poisson point process for the same densities of cell types. The envelope for the spatial Poisson point process represents the two-sided significance interval at the 5% significance level.

distance of another cell. We have previously discussed[1] the arbitrary nature of which 'relevant' cell-cell distances are chosen. For example, Failmezger et al. [66], consider tumour cells to be interacting if the two cells are within 35 microns of one another. This threshold was determined by calculating the distance of cancer cells to their stromal nearest neighbour in regions of the tumour–stroma interface. Then, the mean value of this distribution was assigned as the interaction distance threshold. Using this graph construction, they then introduce the idea of a "cancer supernode". This is when a node has more than 50 connected edges to other cancer cells. The idea is that if there are a large number of cancer cells within the pre-specified distance, this can be summarised as a collective element in further graph-theoretic measures. A popular methodology but it simplifies the actions that cells and their microenvironment go through when interacting, i.e. they do not consider structures with multiple cell types. In addition, there is a bias for a "cancer supernode" to be identified when considering smaller cell types. This is because, for the same subgraph, the distance between cells is less due to a smaller radius of the cell type. Therefore, cells would be considered interacting even if more cells physically sit between them over the same distance threshold.

Another approach is the k-*Means* [324] algorithm. The algorithm works by initialising k different centres $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ by picking k points from X, where X is our set of points. Afterwards, an initial assignment of points to each cluster occurs with every point allocated to the cluster in which the point has a minimal distance value, usually the Euclidean distance, between itself and the respective cluster centroid. Translated into mathematical notation, the cluster is initially defined as

$$S^{(i)} = \{\mathbf{x} : \|\mathbf{x} - c_i\|_2 \leq \|\mathbf{x} - c_j\|_2, \quad \forall\, 1 \leq j \leq k\} \tag{A.26}$$

where each x is assigned to exactly one $S^{(i)}$, for $1 \leq i \leq k$, even if it could be assigned to two or more of them. In this case, the point is associated with the first cluster considered in the list.

Then, each centroid $\mathbf{c}_i$ is updated to the mean of all training examples assigned to cluster

[1] See Chapter 1, Section 1.1.

i, such that

$$c_i = \frac{1}{|S^{(i)}|} \sum_{\mathbf{x}_p \in S^{(i)}} \mathbf{x}. \tag{A.27}$$

The algorithm alternates between the two steps until convergence or the maximum number of iterations is achieved. The nature of the problem is NP-hard. Therefore, the k-Means algorithm assigns a centre dependent on local minimums and is not guaranteed to obtain a global minimum as desired. In addition, results are usually sub-optimal with highly non-convex clusters.

Algorithm 1: k-Means Algorithm

Input:

$X = \{x_1, x_2, \ldots, x_n\}$	Set of points,
k	Number of clusters,
$d : D \times D \to \mathbb{R}$	A distance function,
maxIter	Maximum number of iterations.

Output:

$S = \{S^{(i)} \mid i = 1, 2, \ldots, k\}$	Partition of points,
$C = \{c_i \mid i = 1, 2, \ldots, k\}$	Set of cluster centres.

$C_0 \leftarrow \{k$ random points in $X\}$;
convergence $\leftarrow$ FALSE ;
iter $\leftarrow 1$;
while $\neg$ convergence AND iter $\leq$ maxIter **do**
  compute S_{iter} from $C_{\text{iter}-1}$;
  compute C_{iter} ;
  if $C_{\text{iter}} = C_{\text{iter}-1}$ **then**
    convergence $\leftarrow$ TRUE ;
  end
  iter := iter $+ 1$;
end
$S \leftarrow S_{\text{iter}}$;
$C \leftarrow C_{\text{iter}}$;

Here, the number of clusters, k, in a data set is assigned by the user or via a heuristic such as the Silhouette Coefficient Score or Within-Cluster Sum of Squares combined with the Elbow Criterion, see Figure A.9.

Instead of defining the number of clusters that you expect, methodologies could consider the typical number of cells associated with smaller components that assemble into higher-

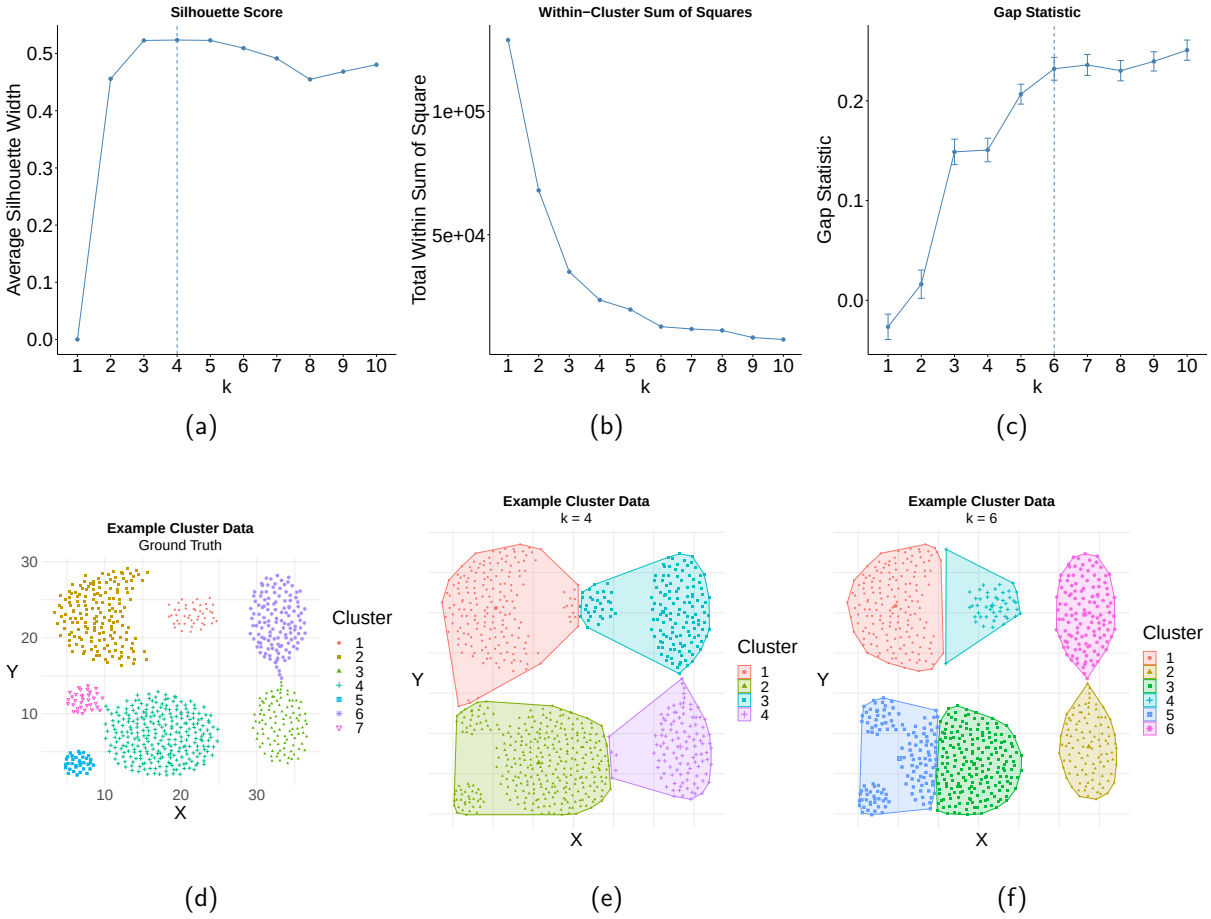


Figure A.9: The heuristics used to determine k for the k-Means clustering algorithm where (a) presents the Total Within-Cluster Sum of Squares value, (b) presents the Gap Statistic, (c) presents the Silhouette Score for $k \in \{1, ..., 10\}$. (d) Ground truth clustering of the 'aggregation' point pattern [6] with the results of the k-Means algorithm for (e) $k = 4$, and (f) $k = 6$, determined by the Gap Statistic and Silhouette Score, respectively.

order structures. One such method is the mutual kNN clustering method [325]. This has been shown in some biological contexts to provide some robust results for cell community detection [326] and is similar to that used to construct the elemental components in a metric considered from the previous chapter, namely for identifying Coordinated Cellular Neighbourhoods[1]. The mutual kNN clustering algorithm takes an input k as before but this instead relates to the degree of nearest neighbours considered around each point. Then, two points are considered part of the same cluster if they are contained in each other's k^{th}-nearest neighbour set. The value of k can be considered as an estimate for density, with a low value of k considering only the most immediate neighbours and as k increases, points which are further away can be considered part of the same cluster. Choosing k is then a crucial step in the application of the algorithm. Implementations consider "stable" values as the most appropriate, which refers to some incremental increase or decrease in k producing the same or very similar result. The mutual kNN also has a limitation in the relevance of some clustering assignments. Specifically, the relevance of cell-cell interactions is not considered. Because of this, cells that are considered to be in each other's nearest neighbour set could be at distances not biologically relevant. Therefore, we do not expect the kNN clustering algorithm to be relevant in the context of our application.

Ripley's K-function[2] and the normalisation of Besag's L-function[3] used in the context of clustering has two parameters to set: a search radius, and a threshold density. The search radius, r, is used to calculate the K- and L-functions. Values of $K(r)$ or $L(r)$ are generated for each point before interpolating between points to create a field on the domain. Applying the threshold density produces a binary cluster map of points with values above the threshold as clustered, which are subsequently connected to other points contained on the same density peak. This method has successfully been applied to identify clusters in several biological applications [327, 198].

Voronoi tessellation-based segmentation is another popular method. Tesselation-based

[1] See Chapter 2, Subsection 2.9.2.
[2] See Appendix A, Subsection A.4.3 for definition.
[3] See Appendix A, Subsection A.4.4 for definition.

clustering is implemented by removing all tiles where the normalised detection density is below a specified threshold. Density can be defined as the inverse of tile area (Voronoi zero), or the inverse of mean first-rank tile area (Voronoi 1st). The mean first-rank tile area is defined as the mean area of the specified and all adjacent tiles. Density estimates are normalised by dividing the mean density across the domain. After removing tiles from the tessellation, clusters are formed from the connected components of the dual graph, known as the Delaunay Triangulation. Again this has had successful applications in the application to biomedical images [328, 329, 198].

Density-based techniques allow for only significant features of the data set to be identified. The basis for these methods is the assumption that the data points are drawn from some unknown density function f. The most common approach is to threshold the density at some fixed level, d. Then, those points that exist in the same density regions form a super level-set $F^d = f^{-1}, [d, +\infty)$, which colloquially are clusters, and the rest of the data is noise. As the true density f is unknown, techniques to approximate for the super level-set F^d are necessary. The limitation of such an approach becomes apparent due to the dependency on the provided fixed density threshold d, where the methods do not respond well to hierarchical data sets, where multi-scale aggregation formation may occur.

An algorithm which begins to overcome the limitation in defining a single density value is the DBSCAN algorithm [192]. To do this, the user is required to provide parameters for graph-based heuristics instead of fixed density values. The first input is a minimum distance, usually the Euclidean distance, that an edge can be drawn between points, `Eps`. The second is the minimum number of neighbours to be considered a 'core' point, `MinPts`. Once all points have been classified as either 'core' or not, then those points in which a path exists between a 'core' point and any other point, where at least one point associated with each edge in the path is a 'core' point, are then identified as part of the same cluster. The pseudocode for the algorithm is presented in Algorithm 2. We also present the computation steps to reach the final cluster outcomes, where Figure A.10a depicts the Point Density Estimator for a chosen `Eps` value, and Figure A.10b depicts the points which are considered. In addition, the method

has been extended to apply to any notion of neighbourhood [330] and consider the hierarchy of clusters for a set of `MinPts` values[331].

Algorithm 2: DBSCAN algorithm

Input:
$X = \{x_1, x_2, \ldots, x_n\}$ Set of points,
Eps Neighbourhood distance threshold,
MinPts Minimum number of points threshold.

Output:
$P = \{P^{(i)} | \, i = 1, 2, \ldots, k\}$ Partition of points.

```
i ← 1 ;
for x ∈ X do
    if x ∉ V then
        Add x to V ;                                   // The set of points visited
        NeighbourPts ← regionQuery(x, X, Eps) ;
        if size.of(NeighbourPts < MinPts) then
            Add x to P^(0) ;                           // The noise partition
        else
            i := i + 1 ;
            P^(i) = expandCluster(x, NeighbourPts, i, Eps, MinPts, V) ;
            Add P^(i) to P
        end
    end
end
```

DBSCAN is not always deterministic. That is, border points that are reachable from more than one cluster can be part of either cluster and depend on the order the data is processed. Although a low probability scenario, this goes contrary to consistent statistical interpretation of density-connected components. A variation, called DBSCAN* [193] does treat border points as noise, which achieves a fully deterministic result.

Using DBSCAN does not come without some limitations. It has been shown that DBSCAN cannot cluster data sets well with large differences in densities [332]. The combination of `minPts` and `Eps` does not allow for some cluster-dependent variation in a domain which has local variations in density, such as in structures of biological organs.

The quality of DBSCAN depends on the distance measure used in the function regionQuery(x', X, `Eps`). The most common distance metric used is the Euclidean distance. Due to the "Curse of dimensionality" this metric makes it increasingly difficult to find an appropriate value of

Algorithm 3: regionQuery Algorithm (DBSCAN)

Input:

$x_i \in X$	Point,
$X = \{x_1, x_2, \ldots, x_n\}$	Set of points,
Eps	Neighbourhood distance threshold,

Output:

$X' = \{x'_1, x'_2, \ldots, x'_m\}$	Points within the eps-neighbourhood of x.

$X' \leftarrow \varnothing$;
for $x_j \in X$ **do**
  dist := $d(x_i, x_j)$;
  if dist $\leq$ Eps **then**
    Add x_j to X' ;
  end
end

Algorithm 4: expandCluster Algorithm (DBSCAN)

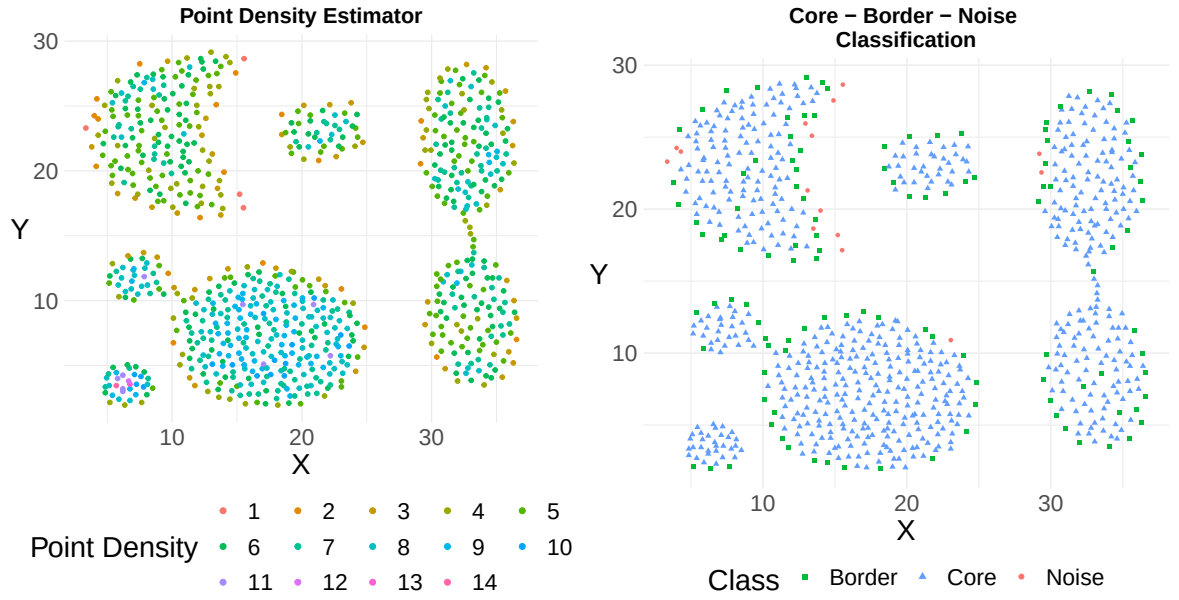
Input:

$x_j \in X$	Point,
NeighbourPts	Index for points in X,
i	Cluster index,
Eps	Neighbourhood distance threshold,
MinPts	Minimum number of points threshold,
V	Visited points.

Output:

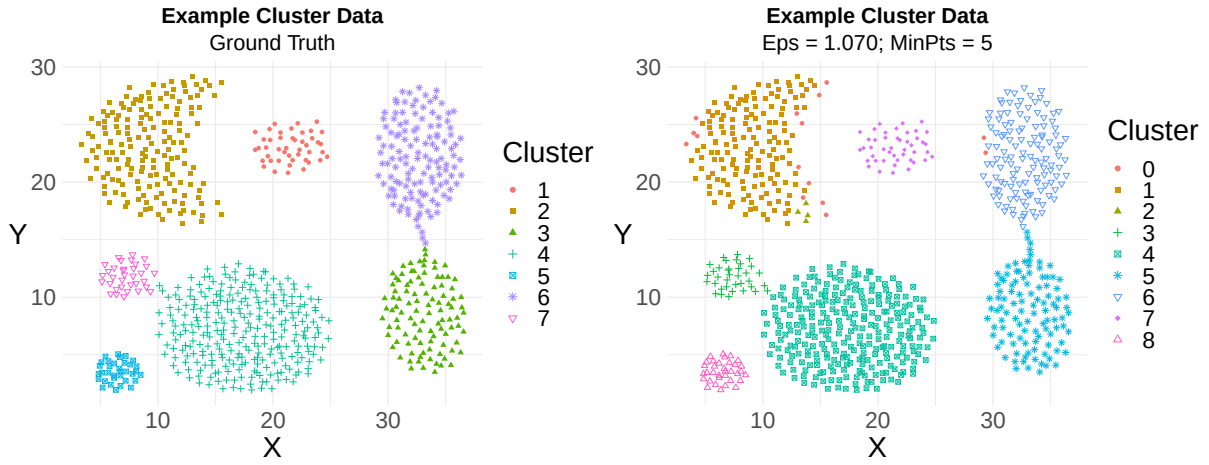
$P^{(i)}$	Points in partition i.

Add x_j to $P^{(i)}$;
for $x' \in$ NeighbourPts **do**
  if $x' \notin V$ **then**
    Add x' to V ;
    NeighbourPts' = regionQuery(x', X, Eps) ;
    if size.of(NeighbourPts') $\geq$ MinPts **then**
      NeighbourPts := NeighbourPts $\cup$ NeighbourPts' ;
    end
    if $x' \notin P^{(k)}, 1 \leq k \leq 1$ **then**
      Add x' to $P^{(i)}$;
    else
      Add
    end
  end
end



(a) (b)

Figure A.10: Computation steps in the DBSCAN clustering algorithm on the 'aggregation' point pattern [6]. (a) Detection density estimates are calculated by counting the number of other points within Eps = 1.00. (b) Classification of core, reachable and noise points for MinPts = 5.



(a) (b)

Figure A.11: Results of the DBSCAN algorithm on an artificial point pattern. (a) Ground truth clustering of the 'aggregation' point pattern [6]. (b) Clusters identified by the DBSCAN clustering algorithm, for Eps = 1.070047 and MinPts = 5, (ARI = 0.959).

Eps. This effect, however, is also present in any other algorithm based on Euclidean distance.

The methodology employed by our analysis is called ToMATo [4]. This method resolves the issues raised with the other aforementioned methodologies. It can identify the number of clusters, allocate points to clusters across scales of density thresholds, and all with reasonable computational power. The basis of this algorithm combines the original graph-based hill-climbing algorithm of Koontz et al. [200] with a cluster merging step guided by persistence, where persistence is the same as that considered in PH[1].

The input of the ToMATo algorithm is an unweighted graph, $\mathbb{G}$, whose vertex set represents the data points and whose edges connect the points according to some user-defined proximity rule. In our implementation, we use the δ-Rips graph as the basis of our proximity rule, which draws an edge between two vertices within distance δ of each other. The selection of a particular value of δ corresponds to the scale for which to inspect the data. Therefore, we reiterate ToMATo for a range of values, obtaining multiple topological features. As a consequence, we ascertain features for each which could be considered significant for further investigation in a separate step. We explain the different approaches to set a reasonable threshold later in this section. Additional graph constructions can be considered as the input for ToMATo, however, we refer the reader to Chazal et al. [4] for their discussion on the limitations of other commonly used graph constructions.

Computing the persistence diagram of a density estimator, $\hat{f}$, enables us to quantify the prominence of the peaks outputted from the hill-climbing method. Ideally, this can distinguish those that correspond to peaks of the true density f from those that are inconsequential. The algorithm is agnostic to the choice of density estimator. The density estimator used in our analysis is similar to that used in DBSCAN, the detection density estimates are calculated by counting the number of other detections within a specified radius. Alternative density estimators include the *truncated Gaussian kernel* estimator [4] and the *distance to a measure* estimator proposed by Biau et al. [333].

Finally, the clusters are then filtered to the core components via the removal of late-

[1]See Appendix A, Subsection A.2.6.

appearing connected components to the respective clusters. This is referred to as the merging phase [4]. In the first run, τ is set to $+\infty$. This value of τ produces an equivalent persistence diagram as to that of the 0-dimensional simplicies for the scalar field $\hat{f}$ over the graph $\mathbb{G}$, similar to that of the standard PH algorithm [304]. The persistence diagram gives us a precise understanding of the relationship between the choices of τ and the number of obtained clusters. Given a threshold τ, the identified clusters, m_p, of prominence less than τ are merged into its root cluster, $r(m_q)$, depending on the persistence hierarchy. The final output is then a cluster index with meaningful outputs. Either, a point has a value greater than 0, with the index referring to a unique parent cluster with persistence greater than τ. Alternatively, a point has a value equal to 0, to indicate a unique parent cluster less persistent than the assigned τ, and thus considered as noise.

Identifying the correct value for the threshold has been discussed in Section A.2.6. However, in a typical sample from our data set, we intend to identify dense aggregations across multiple scales and thus multiple parameters of density difference. This quickly becomes unfeasible to subjectively discern the most significant aggregations since we are now considering a series of persistence diagrams possibly with hundreds of features. Chazal et al. [4] suggest in the original formulation that one could consider the largest difference in sequential feature threshold values, similar to heuristics performed in Spectral Clustering to find a gap in a Laplacian spectrum [334].

Algorithm 5: ToMATo Algorithm

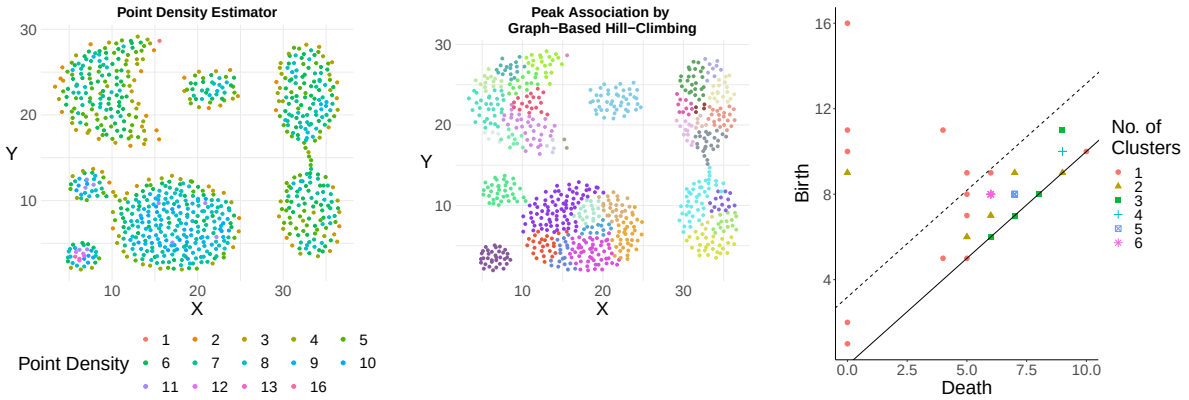
Input:

$G = (V, E)$	Unweighted Simple Graph,
$\hat{f}$	Density estimator,
τ	Persistence threshold.

Output:

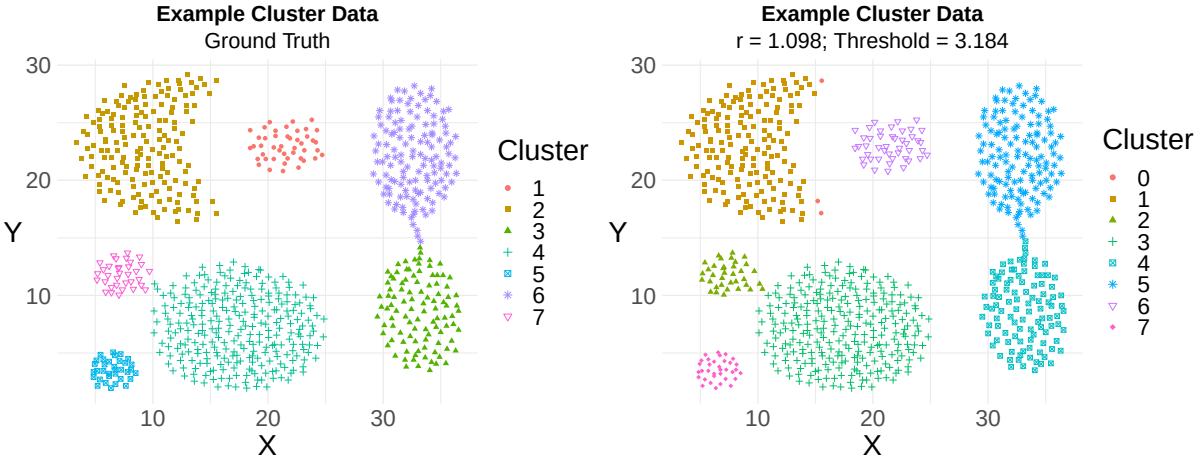
$\mathcal{U} = \{e | \hat{f}(r(e)) \geq \tau\}$ Partition of points.

Sort the vertex indices $\{1, 2, \ldots, n\}$ such that $\hat{f}(1) \geq \hat{f}(2) \geq \cdots \geq \hat{f}(n)$;
Initialise a union-find data structure $\mathcal{U}$ and two vectors g, r of size n ;
for $i = 1$ to n **do**
  Let $\mathcal{N}$ be the set of neighbours of i in G that have indices lower than i ;
  if $\mathcal{N} = \varnothing$ **then**
    // vertex i is a peak of $\hat{f}$ within G
    Create a new entry e in $\mathcal{U}$ and attach vertex i to it ;
    $r(e) \leftarrow i$; // $r(e)$ stores the root vertex associated with the entry e
  else
    // vertex i is not a peak of $\hat{f}$ within G
    $g(i) \leftarrow \text{argmax}_{j \in \mathcal{N}} \hat{f}(j)$;
    $e_i \leftarrow \mathcal{U}.\texttt{find}(g(i))$;
    Attach vertex i to the entry e_i ;
    for $j \in \mathcal{N}$ **do**
      $e \leftarrow \mathcal{U}.\texttt{find}(j)$;
      if $e \neq e_i$ AND $\min\{\hat{f}(r(e)), \hat{f}(r(e_i))\} \leq \hat{f}(i) + \tau$ **then**
        $\mathcal{U}.\texttt{union}(e, e_i)$;
        $r(e \cup e_i) \leftarrow \arg\max_{\{r(e), r(e_i)\}} \hat{f}$;
        $e_i \leftarrow e \cup e_i$;
      end
    end
  end
end



(a) (b) (c)

Figure A.12: Computation steps in the ToMATo clustering algorithm on the 'aggregation' point pattern [6]. (a) Detection density estimates are calculated by counting the number of other points within $r = 1.097725$. (b) Peaks are identified via the graph-based hill-climbing step. (c) Persistence diagram, where we indicate the number of clusters with the same birth-death values.



(a) (b)

Figure A.13: Results of the ToMATo algorithm on an artificial point pattern. (a) Ground truth clustering of the 'aggregation' point pattern [6]. (b) Clusters identified by the ToMATo clustering algorithm, for r = 1.097725 and threshold = 3.183673, (ARI = 0.987).

APPENDIX B

ADDITIONAL INFORMATION AND RESULTS

B.1 Exploring Multiple Phenotypic Intercellular Interactions in 2D Multiplex Immunofluorescence Imaging

We present Table B.1, the matrix of phenotype similarity used in the implementation of the Rao quadratic entropy on the HALO1079 data set. In addition, we present the additional results of the ECI analysis in Figure B.1 and Figure B.2.

CK+	Inf	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CK+ HLA-DR+	0.8	Inf	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CD8+	0.1	0.1	Inf	0.8	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
CD8+ HLA-DR+	0.1	0.1	0.8	Inf	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
CD68+	0.1	0.1	0.25	0.25	Inf	0.8	0.5	0.5	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
CD68+ HLA-DR+	0.1	0.1	0.25	0.25	0.8	Inf	0.5	0.5	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
CD68+ CD163+	0.1	0.1	0.25	0.25	0.5	0.5	Inf	0.8	0.5	0.5	0.25	0.25	0.25	0.25	0.25	0.25
CD68+ CD163+ HLA-DR+	0.1	0.1	0.25	0.25	0.5	0.5	0.8	Inf	0.5	0.5	0.25	0.25	0.25	0.25	0.25	0.25
CD163+	0.1	0.1	0.25	0.25	0.25	0.25	0.5	0.5	Inf	0.8	0.25	0.25	0.25	0.25	0.25	0.25
CD163+ HLA-DR+	0.1	0.1	0.25	0.25	0.25	0.25	0.5	0.5	0.8	Inf	0.25	0.25	0.25	0.25	0.25	0.25
DC-LAMP+ HLA-DR+	0.1	0.1	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	Inf	0.8	0.25	0.25	0.25	0.25
DC-LAMP+ XCR1+ HLA-DR+	0.1	0.1	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	Inf	0.25	0.25	0.25	0.25
XCR1+	0.1	0.1	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	Inf	0.8	0.8	0.8
XCR1+ HLA-DR+	0.1	0.1	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.8	Inf	Inf	0.8
XCR1+ CD163+ HLA-DR+	0.1	0.1	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.8	0.8	0.8	Inf

Table B.1: Table of phenotype similarity used to calculate Rao's entropy for the HALO1079 data set.

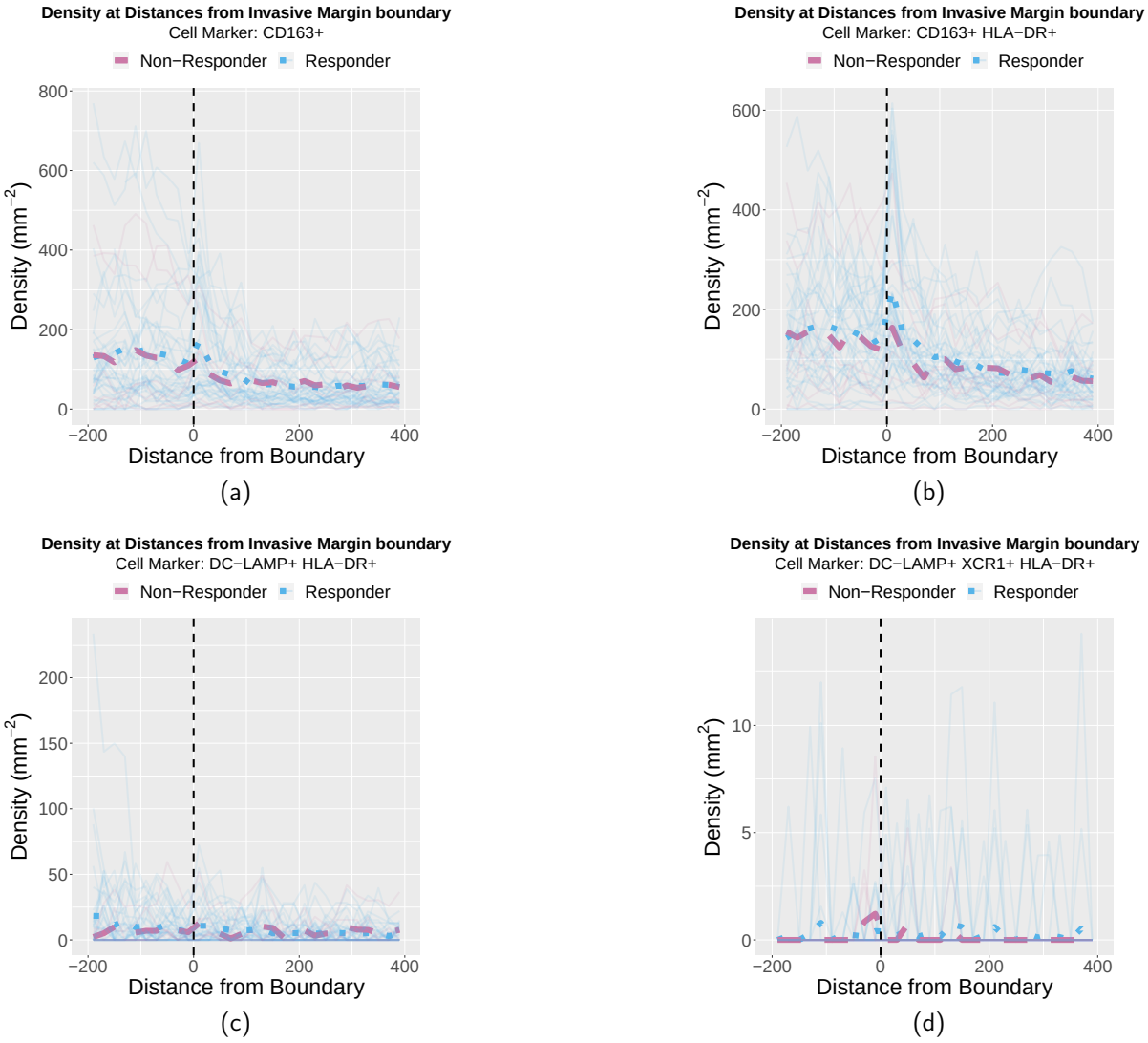


Figure B.1: Results from the ECI analysis. (a) CD163+. (b) CD163+ HLA-DR+. (c) DC-LAMP+ HLA-DR+ (d) DC-LAMP+ XCR1+ HLA-DR+.

B.2 Identifying Higher-Order Structures in 2D Targeted Immunofluorescence Imaging

B.2.1 Clustering Evaluation Scores on Benchmark Data Set

Table B.2 refers to the evaluation of the DBSCAN and ToMATo clustering algorithms on the clustering benchmark data set collated by Barton and Bruna [194].

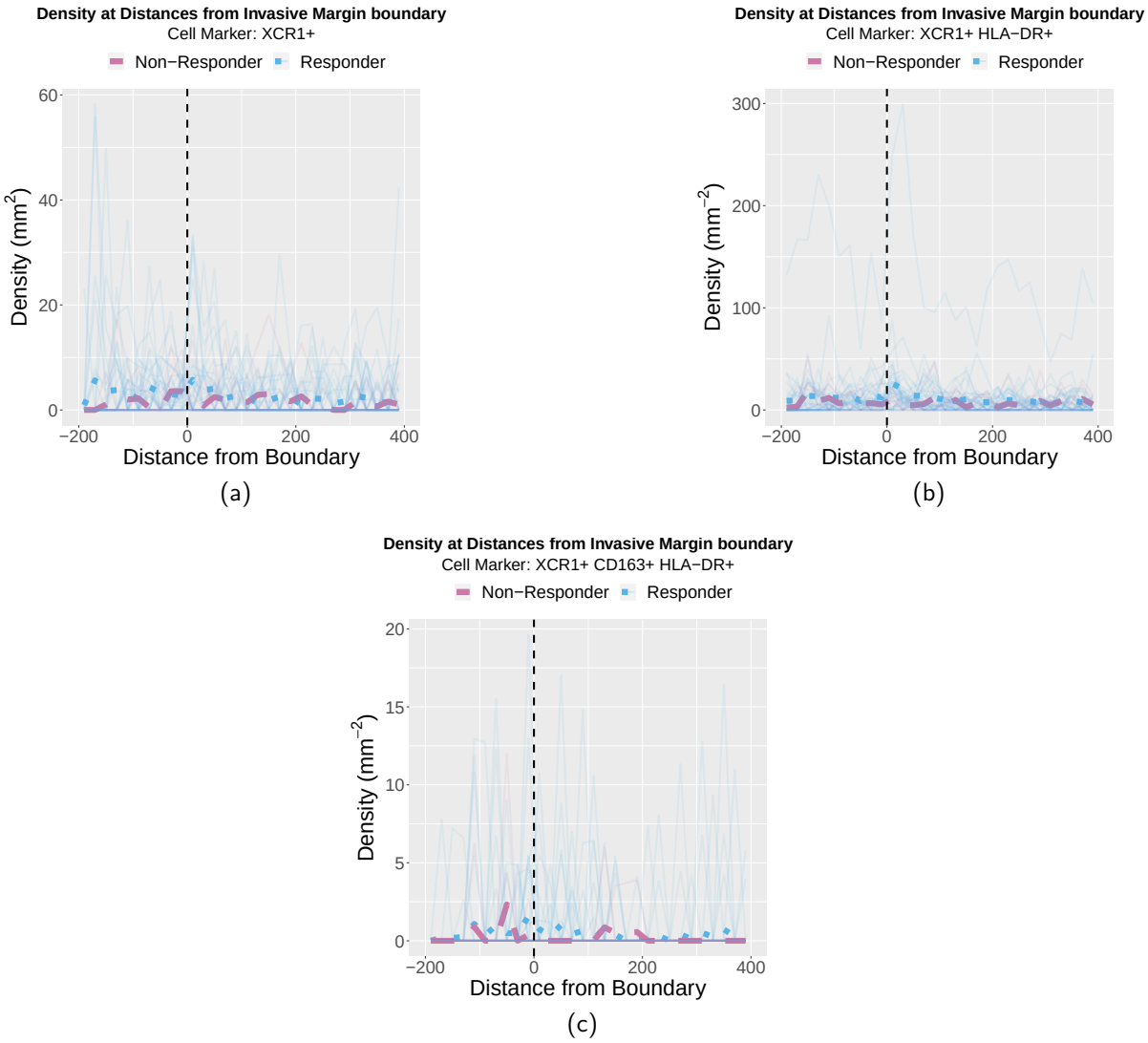


Figure B.2: Results from the ECI analysis. (a) XCR1+. (b) XCR1+ HLA-DR+. (c) XCR1+ CD163+ HLA-DR+.

	Datum	ToMATo	DBSCAN		Datum	ToMATo	DBSCAN
1	2d-3c-no123	**0.927201**	0.9232615	37	longsquare	**0.985392**	0.9567203
2	2d-4c	**1**	0.9986453	38	lsun	0.997347	0.9973471
3	2d-4c-no4	**0.982107**	0.9750629	39	pathbased	0.850102	**0.909471**
4	2d-4c-no9	**0.951199**	0.9115035	40	R15	**0.985559**	0.9767565
5	2dnormals	**0.712048**	0.1770888	41	rings	**0.75948**	0.7246664
6	3-spiral	0.995154	0.9951543	42	shapes	1	1
7	3MC	**1**	0.9676073	43	sizes1	**0.956881**	0.9138272
8	aggregation	**0.987175**	0.9590756	44	sizes2	**0.969609**	0.9310293
9	aml28	0.988679	**0.9906134**	45	sizes3	**0.962114**	0.9354643
10	blobs	**0.903006**	0.784585	46	sizes4	**0.967438**	0.9503809
11	cassini	0.998264	0.9982643	47	sizes5	**0.970017**	0.9529271
12	circle	0.165944	**0.1968945**	48	smile1	1	1
13	compound	0.978152	0.9781515	49	smile2	1	1
14	curves1	0.019374	0.0193737	50	smile3	1	1
15	curves2	0.019322	**0.0193857**	51	spherical_5_2	**0.82771**	0.7348055
16	dartboard1	1	1	52	spherical_6_2	1	1
17	dartboard2	1	1	53	spiral	1	1
18	disk-1000n	**0.638722**	0.4080037	54	spiralsquare	**0.998522**	0.9958745
19	donut1	1	1	55	square1	**0.951229**	0.9340578
20	donut2	**0.996**	0.988024	56	square2	**0.929252**	0.8935417
21	donut3	1	1	57	square3	**0.871313**	0.8024227
22	donutcurves	**1**	0.9973307	58	square4	**0.775922**	0.6682084
23	dpc	0.954936	**0.9652719**	59	st900	**0.812217**	0.5948762
24	DS-577	**0.994851**	0.9721802	60	target	0.970209	0.9702087
25	DS-850	**0.993518**	0.9856675	61	threenorm	0.396296	**0.4323959**
26	ds2c2sc13	**0.908957**	0.8112295	62	triangle1	1	1
27	ds3c3sc6	**0.839895**	0.6976319	63	triangle2	**0.989086**	0.972921
28	ds4c2sc8	**0.814081**	0.7234317	64	twenty	1	1
29	elliptical_10_2	**0.8215**	0.8066883	65	twodiamonds	**0.992509**	0.9975031
30	flame	**0.965848**	0.9550369	66	wingnut	1	1
31	fourty	0.998974	**1**	67	zelnik1	1	1
32	gaussians1	0.980196	0.980196	68	zelnik2	1	1
33	jain	1	1	69	zelnik3	1	1
34	long1	0.984096	**0.9841118**	70	zelnik4	1	1
35	long2	0.977771	**0.9777752**	71	zelnik5	1	1
36	long3	**1**	0.9964	72	zelnik6	1	1

Table B.2: Comparison of ARI for the ToMATo and DBSCAN clustering algorithms.

B.2.2 Parameters used in the Application of the Structured Cellular Neighbourhood Workflow

We have recorded the parameters used in the Structured Cellular Neighbourhood workflow to obtain the results presented on the artificial and TRACTISS data set. Each table of parameters refers to a stage in the structured cellular neighbourhood workflow. Table B.3, Table B.4, Table B.5, and Table B.6 refer to the parameters used to identify the high-density aggregations, identify the boundary of complex geometries, density gradients for multitype point patterns and application of PH for loop identification.

	Value
min. ToMATo radius	10
max. ToMATo radius	15
resolution of ToMATo radius	1
min. presistence percentile	25
no. of percentiles considered	0.25
min. cell count	45

Table B.3: Table of parameters specific to the identification of high-density aggregations.

	Value
ratio	0.1
allow_holes	TRUE

Table B.4: Table of parameters specific to the identification of boundaries of high-density aggregations.

B.2.3 Boundary Evaluations of the Individual Structured Cellular Neighbourhoods identified in the example Artificial Data Set

Evaluation results for the identification of high-density aggregation boundaries from the application of the Structured Cellular Neighbourhood workflow to the artificial data set are

	Value
min. aggregation proportion	0.2
max. aggregation influence	100
inside density gradient threshold	[-Inf, -Inf)
outside gradient density threshold	[-Inf, -0.5]

Table B.5: Table of parameters specific for the quantification of gradient densities from the individual boundaries of high-density aggregations.

	Value
max. aggregation influence	100
threshold	20
loop localisation	outside
max. vertex-vertex distance	50

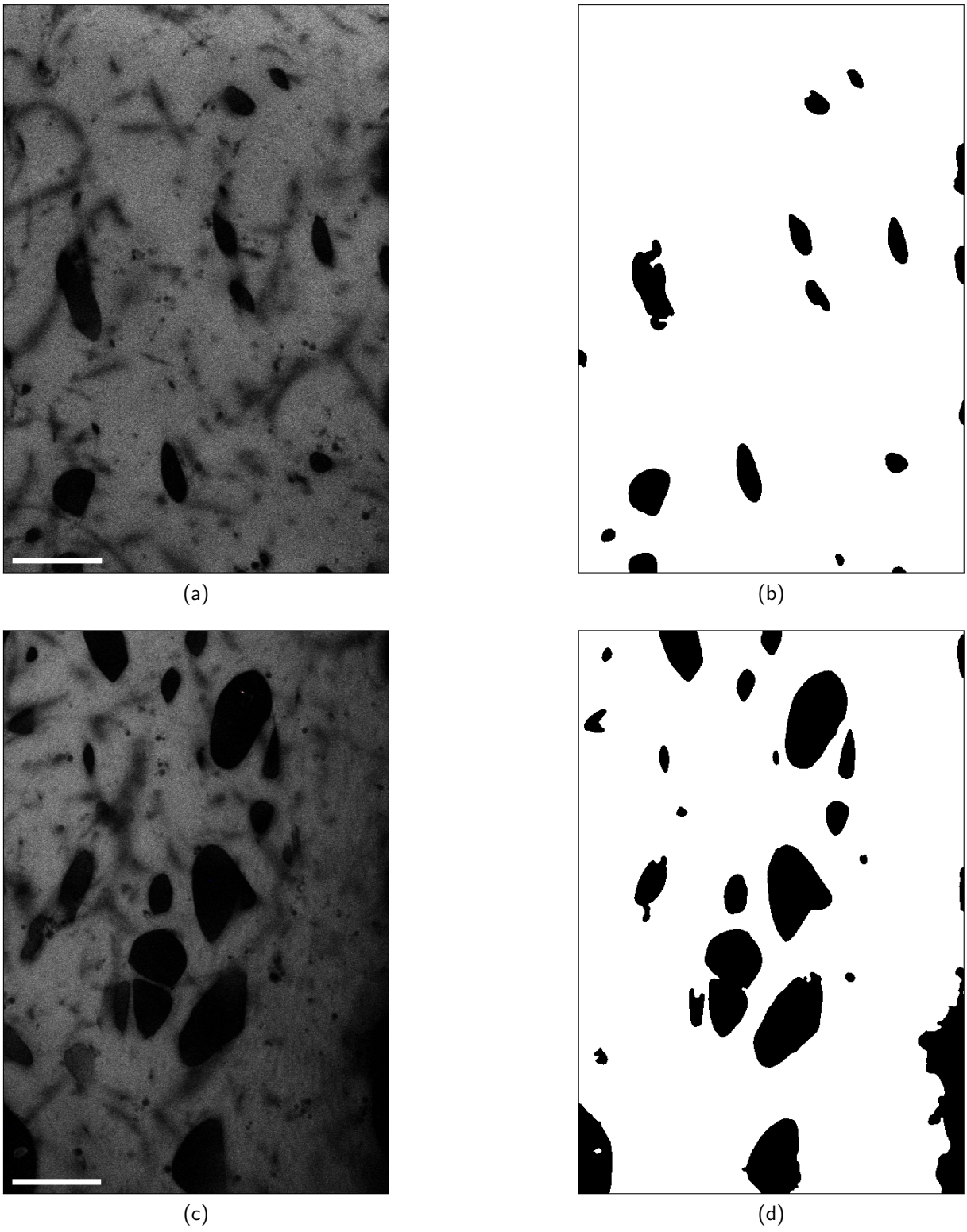
Table B.6: Table of parameters specific to identify a loop around high-density aggregations to validate a surrounding cell phenotype.

presented in Table B.7.

B.3 Quantifying 3-Dimensional Self-Organised Microvascular Models as Spatial Graphs for Transport Simulations

B.3.1 Image Processing Validation

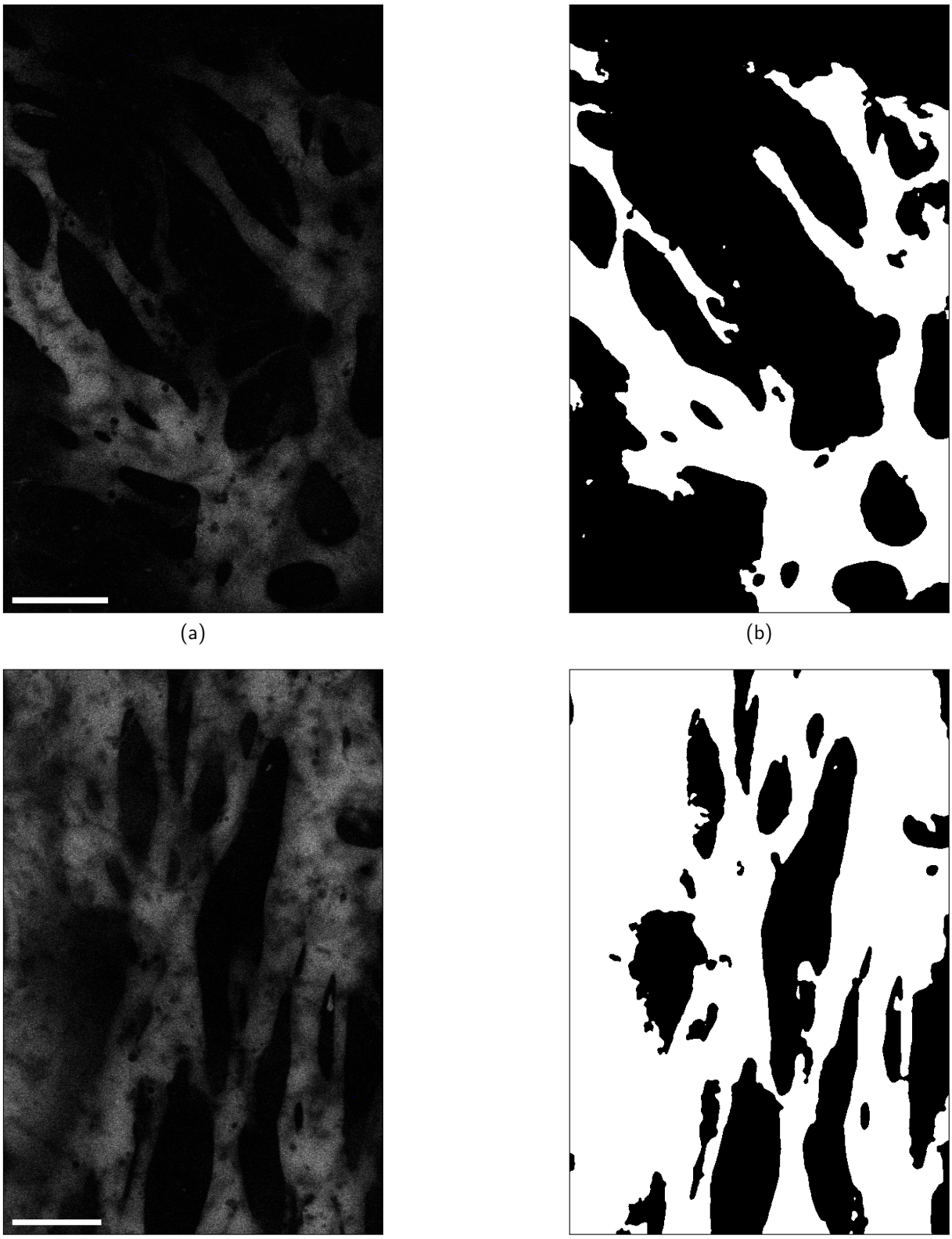
Example slices from each image obtained in the Blood Vessel data set are presented to give an indication of the processing obtained using the workflow described in Chapter 4, Section 4.3. We present Figure B.3, Figure B.4, and Figure B.5 for the mono-culture, remaining co-culture and tri-culture microassays, respectively.



(a) (b)

(c) (d)

Figure B.3: Results from the image processing. (a) - (b) EC-1. (c) - (d) EC-2. Scale bars indicate 200 microns.



(a) (b)

(c) (d)

Figure B.4: Results from the image processing. (a) - (b) EC+AC-1. (c) - (d) EC+AC-2. Scale bars indicate 200 microns.

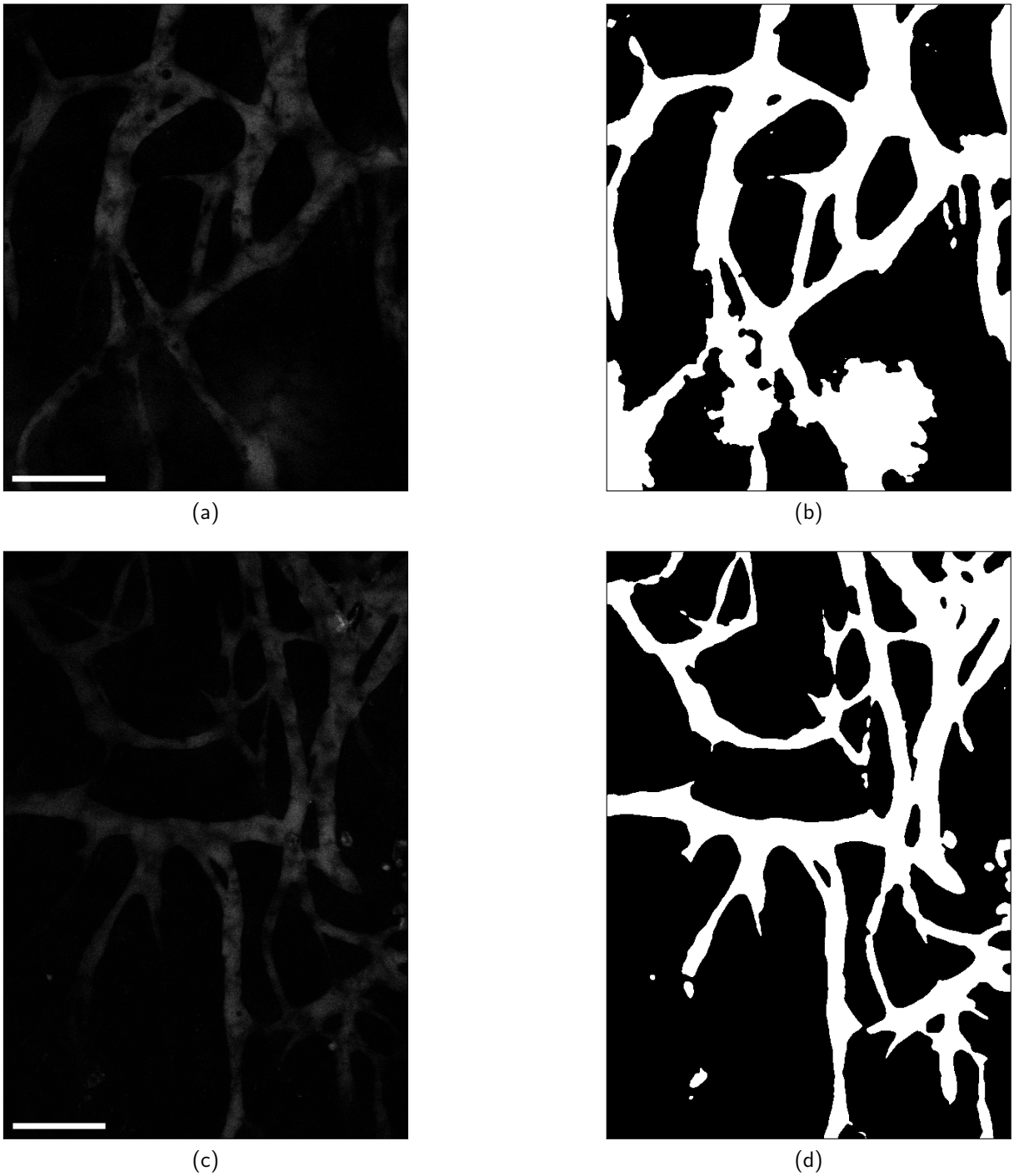


Figure B.5: Results from the image processing. (a) - (b) TC-1. (c) - (d) TC-2. Scale bars indicate 200 microns.

	F1	Precision	Recall	TP	FN	FP
1	0.908	0.999	0.832	0.561	0.113	0.00066
2	0.806	0.989	0.681	0.524	0.246	0.00610
3	0.950	0.999	0.906	0.607	0.063	0.00191
4	0.740	0.984	0.593	0.436	0.300	0.00726
5	0.958	0.993	0.926	0.620	0.050	0.00414
6	0.925	0.999	0.862	0.634	0.101	0.00073
7	0.915	0.982	0.857	0.595	0.099	0.01100
8	0.926	0.928	0.923	0.563	0.047	0.04370

Table B.7: The values of the boundary evaluations of the individual structured cellular neighbourhoods identified on the example artificial data set. TP, FP, and FN stand for True Positive, False Positive and False Negative, respectively.

B.3.2 Individual Measurements

The summary statistics for radii estimations and network-based metrics are presented in Table B.8 and B.9, respectively. In addition, the individual microassay distributions for the tortuosity metrics and specific network metrics are presented in Figure B.6 and Figure B.7, respectively.

Culture	Cell Phenotype	Sample	Min. 3D Skeleton to Background (Median)	Min. 3D Skeleton to Background (Mean)	Min. 3D Skeleton to Background (Std. Dev.)	Local Thickness (Median)	Local Thickness (Mean)	Local Thickness (Std. Dev.)	Min. 2D Skeleton to Background (Median)	Min. 2D Skeleton to Background (Mean)	Min. 2D Skeleton to Background (Std. Dev.)
MC	EC	1	33.541	30.972	14.370	96.717	95.504	24.920	70.410	88.439	65.417
MC	EC	2	29.916	28.096	12.119	89.860	82.524	22.773	92.974	129.781	119.262
CC	EC+AC	1	14.330	13.871	6.168	37.761	36.044	15.390	27.030	29.274	19.018
CC	EC+AC	2	15.821	16.006	7.598	44.520	45.803	18.221	16.401	66.495	109.273
CC	EC+PC	1	13.304	13.306	5.574	37.761	36.006	11.474	62.961	105.078	101.801
CC	EC+PC	2	19.923	19.839	11.079	59.237	58.402	27.104	24.322	26.990	17.392
TC	EC+PC+AC	1	16.600	18.298	10.677	54.803	54.498	26.751	19.147	21.654	14.169
TC	EC+PC+AC	2	13.992	14.888	8.354	41.462	42.768	21.588	15.900	16.452	8.756

Table B.8: Table of measurements for the radii estimations of the microassay samples.

Culture	Cell Phenotype	Sample	Degree (Mean)	Degree (Std. Dev.)	Strength	Clustering Coefficient	Core - Periphery	Characteristic Path Length	Eccentricity (Mean)	Network Diameter	Maximum Normalised Betweenness	Edge Betweenness	Participation Coefficient	Gateway Coefficient	Diversity Coefficient	Subgraph	Flow Coefficient
MC	EC	1	2.082	1.041	30.104	0.118	2.091	0.052	20.864	172.000	0.137	7.216	0.022	0.035	0.007	7.93E+84	0.245
MC	EC	2	2.188	1.009	32.393	0.374	2.650	0.074	20.587	199.000	0.085	4.721	0.025	0.037	0.007	2.07E+98	0.226
CC	EC+AC	1	2.183	1.017	39.437	1.994	3.575	0.213	23.454	228.000	0.103	2.672	0.032	0.063	0.010	1.37E+145	0.180
CC	EC+AC	2	2.143	0.993	33.968	0.165	2.917	0.091	22.269	249.000	0.136	5.370	0.023	0.042	0.007	2.74E+124	0.215
CC	EC+PC	1	2.226	0.980	33.945	0.405	2.613	0.085	21.733	176.000	0.123	3.977	0.023	0.034	0.007	6.25E+88	0.239
CC	EC+PC	2	2.203	0.979	39.793	0.505	3.102	0.165	24.930	218.000	0.104	2.740	0.032	0.054	0.010	1.77E+129	0.191
TC	EC+PC+AC	1	2.071	1.030	26.922	0.365	2.526	0.080	17.251	133.000	0.146	4.723	0.028	0.052	0.009	8.80E+69	0.242
TC	EC+PC+AC	2	2.061	1.115	28.661	0.000	3.165	0.081	19.695	283.000	0.139	3.884	0.023	0.034	0.007	2.70E+120	0.186

Table B.9: Table of measurements for the topology of the microassay samples.

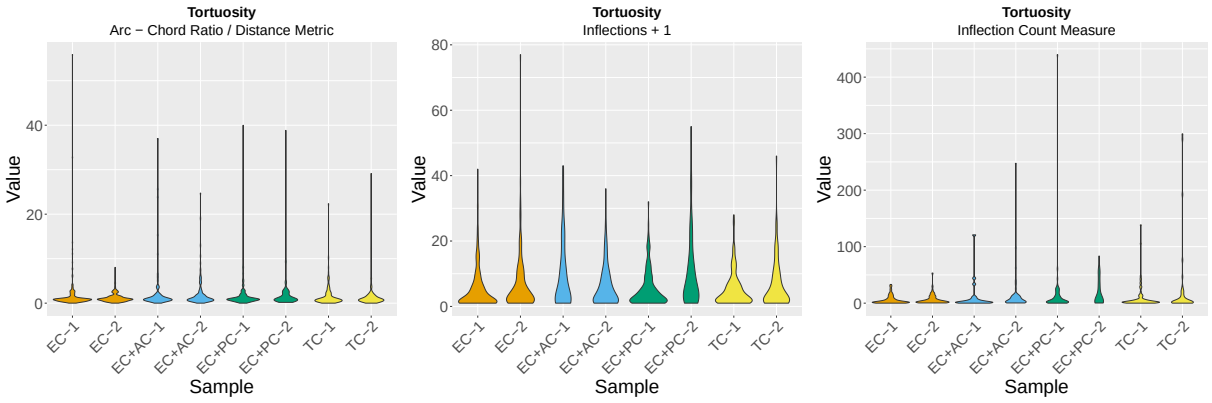


Figure B.6: Distributions of tortuosity for individual samples, for three different measures of tortuosity, (a) Arc-Chord ratio / Distance Measure, (b) Number of Inflections plus one, and (c) ICM.

B.3.3 Multi-Dimensional Reduction analysis

The results of the PCA without network-based metrics, with Table B.10 showing the standard deviation of each Principal Component with Table B.11 and Table B.12 presenting the loadings for the 1-st and 2-nd Principal Components, respectively.

The results of the PCA with network-based metrics, with Table B.13 showing the standard deviation of each Principal Component with Table B.14 and Table B.15 presenting the loadings for the 1-st and 2-nd Principal Components, respectively.

Pricipal Component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard Deviation	2.828	1.916	1.520	1.146	0.580	0.499	0.352	1.06E-16
Proportion of Variance	0.500	0.230	0.144	0.082	0.021	0.016	0.008	0.00E+00
Cumulative Proportion	0.500	0.729	0.874	0.956	0.977	0.992	1.00E+00	1.00E+00

Table B.10: Table of the standard deviations, proportion of variance and cumulative proportion of the Principal Components in the PCA without network metrics. Standard deviation is the square roots of the eigenvalues of the covariance/correlation matrix.

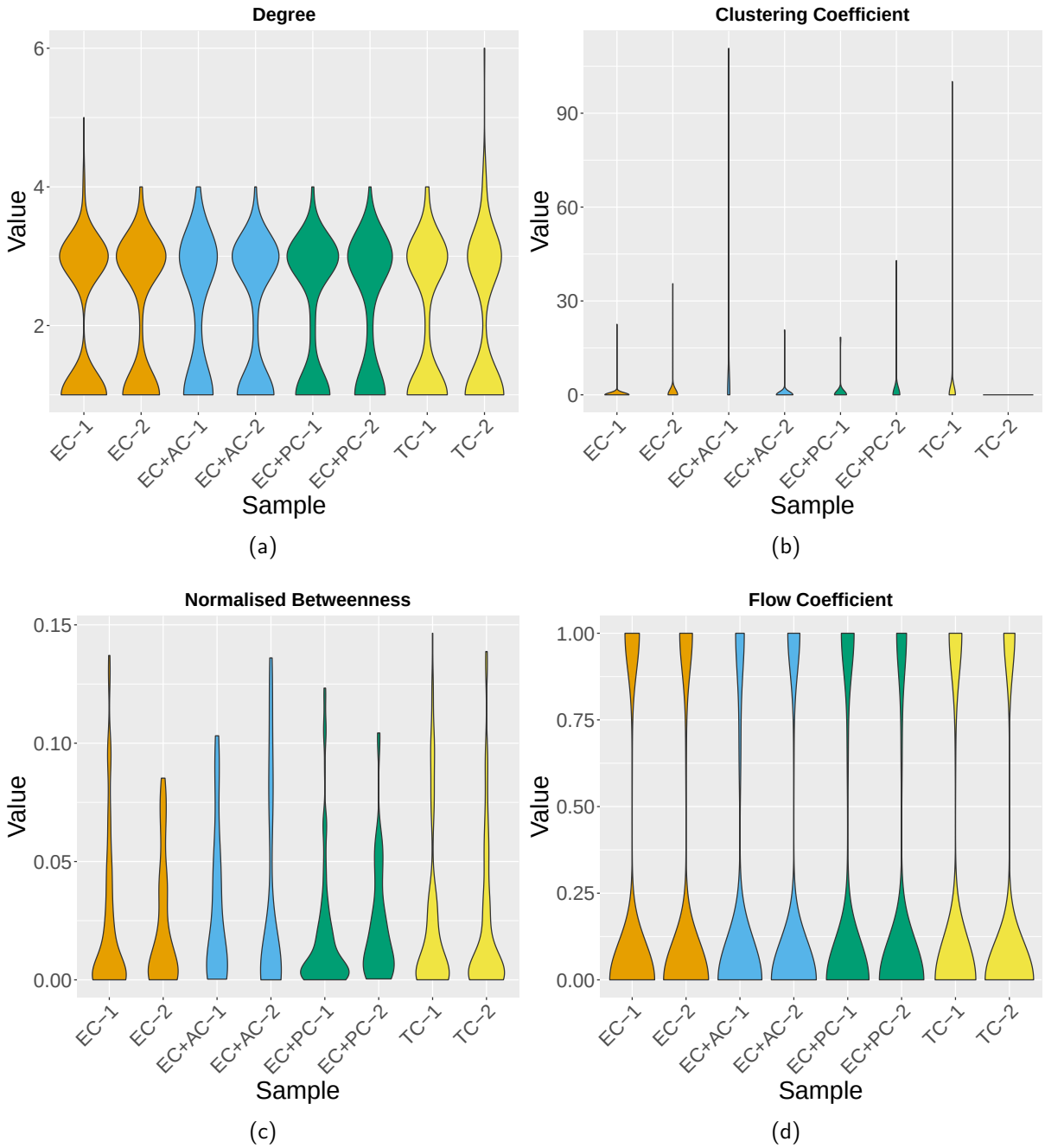


Figure B.7: Distributions of network metrics for individual samples, including the (a) Degree, (b) Clustering coefficient, (c) Normalised Betweenness, and (d) Flow coefficient.

Metric	PC1 weighting
icm_tortuosity_MEAN	0.338
length_STDDEV	0.337
dm_tortuosity_MEAN	0.332
icm_tortuosity_STDDEV	0.317
length_MEAN	0.309
dm_tortuosity_STDDEV	0.286
thickness_AVE	0.260
thickness_MEDIAN	0.258
volume_normalised	0.258
surfaceCorr_normalised	0.255
inflections_tortuosity_MEAN	0.237
inflections_tortuosity_STDDEV	0.194
soam_tortuosity_STDDEV	0.123
thickness_STDDEV	0.066
surfaceArea_normalised	0.056
soam_tortuosity_MEAN	0.056

Table B.11: Table of the 1-st Principal Component loadings for the PCA without network metrics.

Metric	PC2 weighting
surfaceArea_normalised	0.480
inflections_tortuosity_STDDEV	0.371
thickness_STDDEV	0.355
inflections_tortuosity_MEAN	0.333
soam_tortuosity_STDDEV	0.318
soam_tortuosity_MEAN	0.280
thickness_MEDIAN	0.242
thickness_AVE	0.238
volume_normalised	0.208
length_MEAN	0.147
surfaceCorr_normalised	0.119
icm_tortuosity_STDDEV	0.080
dm_tortuosity_MEAN	0.080
dm_tortuosity_STDDEV	0.078
length_STDDEV	0.032
icm_tortuosity_MEAN	0.029

Table B.12: Table of the 2-nd Principal Component loadings for the PCA without network metrics.

Principal Component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard Deviation	3.987	2.354	1.881	1.698	1.603	1.003	0.751	4.35E-16
Proportion of Variance	0.497	0.173	0.111	0.090	0.080	0.031	0.018	0.00E+00
Cumulative Proportion	0.497	0.670	0.781	0.871	0.951	0.982	1.00E+00	1.00E+00

Table B.13: Table of the standard deviations, proportion of variance and cumulative proportion of the Principal Components in the PCA with network metrics. Standard deviation is the square roots of the eigenvalues of the covariance/correlation matrix.

Metric	PC1 weighting
length_MEAN	0.240
characteristic_path_length	0.238
core_periphery_ratio	0.234
icm_tortuosity_MEAN	0.229
louvain_community	0.225
number_of_components	0.223
flow_coef	0.223
edge_betweenness	0.223
inflections_tortuosity_MEAN	0.215
dm_tortuosity_MEAN	0.207
length_STDDEV	0.203
diversity_coef	0.199
ice_tortuosity_STDDEV	0.199
participation_coef	0.195
inflections_tortuosity_STDDEV	0.189
strength_MEAN	0.187
gateway_coef	0.185
clustering_coef	0.182
volume_normalised	0.175
dm_tortuosity_STDDEV	0.175
thickness_MEAN	0.156
eccentricity_MEAN	0.156
thickness_MEDIAN	0.152
surfaceCorr_MEAN	0.139
network_diameter	0.122
max_normalised_betweenness	0.096
degree_MEAN	0.094
surfaceArea_normalised	0.064
soam__tortuosity_MEAN	0.046
soam_tortuosity_STDDEV	0.029
thickness_STDDEV	0.028
degree_STDDEV	0.025

Table B.14: Table of the 1-st Principal Component loadings for the PCA with network metrics.

Metric	PC2 weighting
degree_STDDEV	0.329
soam_tortuosity_STDDEV	0.328
max_normalised_betweenness	0.299
degree_MEAN	0.272
surfaceCorr_normalised	0.262
length_STDDEV	0.234
strength_MEAN	0.233
tinflections_tortuosity_STDDEV	0.221
eccentricity_MEAN	0.218
thickness_MEDIAN	0.214
thickness_MEAN	0.204
dm_tortuosity_MEAN	0.186
inflections_tortuosity_MEAN	0.186
icm_tortuosity_STDDEV	0.167
surfaceArea_normalised	0.165
dm_tortuosity_STDDEV	0.146
participation_coef	0.144
network_diameter	0.137
icm_tortuosity_MEAN	0.117
diversity_coef	0.103
clustering_coefficient	0.097
gateway_coef	0.093
characteristic_path_length	0.093
volume_normalised	0.079
core_periphery	0.068
flow_coef	0.067
thickness_STDDEV	0.066
length_MEAN	0.055
number_of_components	0.016
louvain_community	0.010
soam_tortuosity_mean	0.005
edge_betweenness	0.004

Table B.15: Table of the 2-nd Principal Component loadings for the PCA with network metrics.

LIST OF REFERENCES

[1] B. Grigoryan, S. J. Paulsen, D. C. Corbett, D. W. Sazer, C. L. Fortin, A. J. Zaita, P. T. Greenfield, N. J. Calafat, J. P. Gounley, A. H. Ta, F. Johansson, A. Randles, J. E. Rosenkrantz, J. D. Louis-Rosenberg, P. A. Galie, K. R. Stevens, and J. S. Miller, "Multivascular networks and functional intravascular topologies within biocompatible hydrogels," *Science*, vol. 364, pp. 458–464, May 2019.

[2] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell Detection with Star-Convex Polygons," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds.), Lecture Notes in Computer Science, (Cham), pp. 265–273, Springer International Publishing, 2018.

[3] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, pp. 100–106, Jan. 2021.

[4] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba, "Persistence-Based clustering in riemannian manifolds," *Journal of the ACM*, vol. 60, no. 6, pp. 1–38, 2013.

[5] C. M. Topaz, L. Ziegelmeier, and T. Halverson, "Topological Data Analysis of Biological Aggregation Models," *PLOS ONE*, vol. 10, p. e0126383, May 2015.

[6] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, pp. 4–es, Mar. 2007.

[7] American Cancer Society, "Colorectal Cancer Stages | Rectal Cancer Staging | Colon Cancer Staging." Accessed 2022-02-21 15:44:50.

[8] M. R. Weiser, "AJCC 8th Edition: Colorectal Cancer," *Annals of Surgical Oncology*, vol. 25, pp. 1454–1455, June 2018.

[9] E. Gavagnin and C. A. Yates, "Chapter 2 - Stochastic and Deterministic Modeling of Cell Migration," in *Handbook of Statistics* (A. S. R. Srinivasa Rao and C. R. Rao, eds.),

vol. 39 of *Integrated Population Biology and Modeling, Part A*, pp. 37–91, Elsevier, Jan. 2018.

[10] A. D. Srikanth and T. Schroepfer, "Network Science-based Analysis of Urban Green Spaces in Singapore," *International Journal on Smart and Sustainable Cities*, p. 2340004, June 2023.

[11] J. Snow, *On the mode of communication of cholera*. London : John Churchill, 1855.

[12] Christopher J. Banks, Ewan Colman, Thomas Doherty, Oliver Tearne, Mark Arnold, Katherine E. Atkins, Daniel Balaz, Gaël Beaunée, Paul R. Bessell, Jessica Enright, Adam Kleczkowski, Gianluigi Rossi, and Anne-Sophie Ruget, "SCoVMod – a spatially explicit mobility and deprivation adjusted model of first wave COVID-19 transmission dynamics," *Wellcome Open Res 2022*, vol. 7, no. 161, 2022.

[13] J. C. Paetzold, J. McGinnis, S. Shit, I. Ezhov, P. Büschl, C. Prabhakar, M. I. Todorov, A. Sekuboyina, G. Kaissis, A. Ertürk, S. Günnemann, and B. H. Menze, "Whole Brain Vessel Graphs: A Dataset and Benchmark for Graph Learning and Neuroscience (VesselGraph)," *arXiv:2108.13233 [cs, q-bio]*, Aug. 2021.

[14] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, pp. 1059–1069, Sept. 2010.

[15] W.-L. Wang, Y.-L. Li, M.-X. Zheng, X.-Y. Hua, J.-J. Wu, F.-F. Yang, N. Yang, X. He, L.-J. Ao, and J.-G. Xu, "Altered Topological Properties of Grey Matter Structural Covariance Networks in Complete Thoracic Spinal Cord Injury Patients: A Graph Theoretical Network Analysis," *Neural Plasticity*, Feb. 2021.

[16] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, pp. 186–198, Mar. 2009.

[17] A. Zalesky, A. Fornito, and E. T. Bullmore, "Network-based statistic: Identifying differences in brain networks," *NeuroImage*, vol. 53, pp. 1197–1207, Dec. 2010.

[18] X. Lin, Y. Chen, M. Wang, C. Song, B. Lin, X. Yuan, Q. Liu, H. Yang, and N. Jiang, "Altered Topological Patterns of Gray Matter Networks in Tinnitus: A Graph-Theoretical-Based Study," *Frontiers in Neuroscience*, vol. 14, 2020.

[19] C. Liu, D. Yu, X. Ma, S. Xie, and H. Zhang, "Neural evidence for image quality perception based on algebraic topology," *PLOS ONE*, vol. 16, p. e0261223, Dec. 2021.

[20] S. Das, S. A. Haque, and M. I. Tanveer, "Persistence Homology of TEDtalk: Do Sentence Embeddings Have a Topological Shape?," *arXiv:2103.14131 [cs]*, Mar. 2021.

[21] H. Masoomy, B. Askari, S. Tajik, A. K. Rizi, and G. R. Jafari, "Topological Analysis of Interaction Patterns in Cancer-Specific Gene Regulatory Network: Persistent Homology Approach," preprint, In Review, Feb. 2021.

[22] A. S. Blevins, J. Z. Kim, and D. S. Bassett, "Variability in higher order structure of noise added to weighted networks," *arXiv:2101.03933 [math, q-bio]*, Jan. 2021.

[23] T. Songdechakraiwut and M. K. Chung, "Topological Learning for Brain Networks," *arXiv:2012.00675 [cs, q-bio]*, Nov. 2020.

[24] M. Hajij, K. Istvan, and G. Zamzami, "Cell Complex Neural Networks," *arXiv:2010.00743 [cs, math, stat]*, Oct. 2020.

[25] K. Hess, "Topological Adventures in Neuroscience," in *Topological Data Analysis* (N. A. Baas, G. E. Carlsson, G. Quick, M. Szymik, and M. Thaule, eds.), Abel Symposia, (Cham), pp. 277–305, Springer International Publishing, 2020.

[26] L. Kanari, H. Dictus, A. Chalimourda, A. Arnaudon, W. V. Geit, B. Coste, J. Shillcock, K. Hess, and H. Markram, "Computational synthesis of cortical dendritic morphologies," *Cell Reports*, vol. 39, Apr. 2022.

[27] S. L. Mason, L. Junges, W. Woldman, E. R. Facer-Childs, B. M. d. Campos, A. P. Bagshaw, and J. R. Terry, "Classification of human chronotype based on fMRI network-based statistics," *Frontiers in Neuroscience*, vol. 17, 2023.

[28] M. Kerschnitzki, P. Kollmannsberger, M. Burghammer, G. N. Duda, R. Weinkamer, W. Wagermaier, and P. Fratzl, "Architecture of the osteocyte network correlates with bone material quality," *Journal of Bone and Mineral Research*, vol. 28, no. 8, pp. 1837–1845, 2013.

[29] B. Yu, A. Pacureanu, C. Olivier, P. Cloetens, and F. Peyrin, "Quantification of the bone lacunocanalicular network from 3D X-ray phase nanotomography images," *Journal of Microscopy*, vol. 282, no. 1, pp. 30–44, 2021.

[30] P. Kollmannsberger, M. Kerschnitzki, F. Repp, W. Wagermaier, R. Weinkamer, and P. Fratzl, "The small world of osteocytes: connectomics of the lacuno-canalicular network in bone," *New Journal of Physics*, vol. 19, p. 073019, July 2017.

[31] C. Q. Li, J. F. Magland, C. S. Rajapakse, X. E. Guo, X. H. Zhang, B. Vasilic, and F. W. Wehrli, "Implications of resolution and noise for in vivo micro-MRI of trabecular bone," *Medical Physics*, vol. 35, no. 12, pp. 5584–5594, 2008.

[32] J. F. Magland and F. W. Wehrli, "Trabecular Bone Structure Analysis in the Limited Spatial Resolution Regime of In Vivo MRI," *Academic Radiology*, vol. 15, pp. 1482–1493, Dec. 2008.

[33] L. Pothuaud, A. Laib, P. Levitz, C. L. Benhamou, and S. Majumdar, "Three-Dimensional-Line Skeleton Graph Analysis of High-Resolution Magnetic Resonance Images: A Validation Study From 34-um-Resolution Microcomputed Tomography," *Journal of Bone and Mineral Research*, vol. 17, no. 10, pp. 1883–1895, 2002.

[34] G. Chang, S. K. Pakin, M. E. Schweitzer, P. K. Saha, and R. R. Regatte, "Adaptations in trabecular bone microarchitecture in Olympic athletes determined by 7T MRI," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 5, pp. 1089–1095, 2008.

[35] P. K. Saha, Y. Xu, H. Duan, A. Heiner, and G. Liang, "Volumetric Topological Analysis: A Novel Approach for Trabecular Bone Classification on the Continuum Between Plates and Rods," *IEEE Transactions on Medical Imaging*, vol. 29, pp. 1821–1838, Nov. 2010.

[36] F. W. Wehrli, B. R. Gomberg, P. K. Saha, H. K. Song, S. N. Hwang, and P. J. Snyder, "Digital Topological Analysis of In Vivo Magnetic Resonance Microimages of Trabecular Bone Reveals Structural Implications of Osteoporosis," *Journal of Bone and Mineral Research*, vol. 16, no. 8, pp. 1520–1531, 2001.

[37] A. C. Jones, C. H. Arns, A. P. Sheppard, D. W. Hutmacher, B. K. Milthorpe, and M. A. Knackstedt, "Assessment of bone ingrowth into porous biomaterials using MICRO-CT," *Biomaterials*, vol. 28, pp. 2491–2504, May 2007.

[38] L. M. Griffin, S. Honig, C. Chen, P. K. Saha, R. Regatte, and G. Chang, "7T MRI of distal radius trabecular bone microarchitecture: How trabecular bone quality varies depending on distance from end-of-bone: 7T MRI of Distal Radius," *Journal of Magnetic Resonance Imaging*, vol. 45, pp. 872–878, Mar. 2017.

[39] P. Schneider, T. Krucker, E. Meyer, A. Ulmann-Schuler, B. Weber, M. Stampanoni, and R. Müller, "Simultaneous 3D visualization and quantification of murine bone and

bone vasculature using micro-computed tomography and vascular replica," *Microscopy Research and Technique*, vol. 72, no. 9, pp. 690–701, 2009.

[40] Y. Jung, Z. Zhi, and R. K. Wang, "Three-dimensional optical imaging of microvascular networks within intact lymph node in-vivo," *Journal of Biomedical Optics*, vol. 15, p. 050501, Sept. 2010.

[41] M. B. Vickerman, P. A. Keith, T. L. McKay, D. J. Gedeon, M. Watanabe, M. Montano, G. Karunamuni, P. K. Kaiser, J. E. Sears, Q. Ebrahem, D. Ribita, A. G. Hylton, and P. Parsons-Wingerter, "VESGEN 2D: Automated, User-Interactive Software for Quantification and Mapping of Angiogenic and Lymphangiogenic Trees and Networks," *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology*, vol. 292, pp. 320–332, Mar. 2009.

[42] I. D. Kelch, G. Bogle, G. B. Sands, A. R. J. Phillips, I. J. LeGrice, and P. Rod Dunbar, "Organ-wide 3D-imaging and topological analysis of the continuous microvascular network in a murine lymph node," *Scientific Reports*, vol. 5, p. 16534, Nov. 2015.

[43] M. Kazmierski and B. Haibe-Kains, "Lymph Node Graph Neural Networks for Cancer Metastasis Prediction," *arXiv:2106.01711 [cs, eess, q-bio]*, June 2021.

[44] F. Bruyère, L. Melen-Lamalle, S. Blacher, G. Roland, M. Thiry, L. Moons, F. Frankenne, P. Carmeliet, K. Alitalo, C. Libert, J. P. Sleeman, J.-M. Foidart, and A. Noël, "Modeling lymphangiogenesis in a three-dimensional culture system," *Nature Methods*, vol. 5, pp. 431–437, May 2008.

[45] R. Coico and G. Sunshine, *Immunology: A Short Course*. New York, UNITED KINGDOM: John Wiley & Sons, Incorporated, 2015.

[46] C. M. Schürch, S. S. Bhate, G. L. Barlow, D. J. Phillips, L. Noti, I. Zlobec, P. Chu, S. Black, J. Demeter, D. R. McIlwain, S. Kinoshita, N. Samusik, Y. Goltsev, and G. P. Nolan, "Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front," *Cell*, vol. 182, pp. 1341–1359.e19, Sept. 2020.

[47] H. Mori, J. Bolen, L. Schuetter, P. Massion, C. C. Hoyt, S. VandenBerg, L. Esserman, A. D. Borowsky, and M. J. Campbell, "Characterizing the Tumor Immune Microenvironment with Tyramide-Based Multiplex Immunofluorescence," *Journal of Mammary Gland Biology and Neoplasia*, Feb. 2021.

[48] C. C. Hoyt, "Multiplex Immunofluorescence and Multispectral Imaging: Forming the Basis of a Clinical Test Platform for Immuno-Oncology," *Frontiers in Molecular Biosciences*, vol. 8, 2021.

[49] V. H. Koelzer, K. Sirinukunwattana, J. Rittscher, and K. D. Mertz, "Precision immunoprofiling by image analysis and artificial intelligence," *Virchows Archiv*, vol. 474, pp. 511–522, Apr. 2019.

[50] K. E. Blise, S. Sivagnanam, G. L. Banik, L. M. Coussens, and J. Goecks, "Single-Cell Spatial Proteomics Analyses of Head and Neck Squamous Cell Carcinoma Reveal Tumor Heterogeneity and Immune Architectures Associated with Clinical Outcome," *bioRxiv*, p. 2021.03.10.434649, Mar. 2021.

[51] A. Behanova, A. Klemm, and C. Wählby, "Spatial Statistics for Understanding Tissue Organization," *Frontiers in Physiology*, vol. 13, 2022.

[52] T. K. M. Leong, W. S. Lo, W. E. Z. Lee, B. Tan, X. Z. Lee, L. W. J. N. Lee, J.-Y. J. Lee, N. Suresh, L.-H. Loo, E. Szu, and J. Yeong, "Leveraging advances in immunopathology and artificial intelligence to analyze in vitro tumor models in composition and space," *Advanced Drug Delivery Reviews*, vol. 177, p. 113959, Oct. 2021.

[53] H. Shimizu and K. I. Nakayama, "Artificial intelligence in oncology," *Cancer Science*, vol. 111, no. 5, pp. 1452–1460, 2020.

[54] S. Robertson, H. Azizpour, K. Smith, and J. Hartman, "Digital image analysis in breast pathology—from image processing techniques to artificial intelligence," *Translational Research*, vol. 194, pp. 19–35, Apr. 2018.

[55] C. Reichling, J. Taieb, V. Derangere, Q. Klopfenstein, K. L. Malicot, J.-M. Gornet, H. Becheur, F. Fein, O. Cojocarasu, M. C. Kaminsky, J. P. Lagasse, D. Luet, S. Nguyen, P.-L. Etienne, M. Gasmi, A. Vanoli, H. Perrier, P.-L. Puig, J.-F. Emile, C. Lepage, and F. Ghiringhelli, "Artificial intelligence-guided tissue analysis combined with immune infiltrate assessment predicts stage III colon cancer outcomes in PETACC08 study," *Gut*, vol. 69, pp. 681–690, Apr. 2020.

[56] J. Caldara, J. S. Krueger, E. Ergon, S. Kearney, and B. Vennapusa, "Abstract 4923: Analysis of companion diagnostic potentials for multifaceted PD-L1 assays," *Cancer Research*, vol. 79, p. 4923, July 2019.

[57] B. J. Ainscough, E. K. Barnell, P. Ronning, K. M. Campbell, A. H. Wagner, T. A. Fehniger, G. P. Dunn, R. Uppaluri, R. Govindan, T. E. Rohan, M. Griffith, E. R. Mardis,

S. J. Swamidass, and O. L. Griffith, "A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data," *Nature Genetics*, vol. 50, pp. 1735–1743, Dec. 2018.

[58] A. B. Sanghvi, E. Z. Allen, K. M. Callenberg, and L. Pantanowitz, "Performance of an artificial intelligence algorithm for reporting urine cytopathology," *Cancer Cytopathology*, vol. 127, no. 10, pp. 658–666, 2019.

[59] A. Heindl, S. Nawaz, and Y. Yuan, "Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology," *Laboratory Investigation*, vol. 95, pp. 377–384, Apr. 2015.

[60] R. Ehsani and F. Drabløs, "Robust Distance Measures for kNN Classification of Cancer Data," *Cancer Informatics*, vol. 19, Jan. 2020.

[61] M. Rabbani, J. Kanevsky, K. Kafi, F. Chandelier, and F. J. Giles, "Role of artificial intelligence in the care of patients with nonsmall cell lung cancer," *European Journal of Clinical Investigation*, vol. 48, no. 4, p. e12901, 2018.

[62] M. Bortolomeazzi, L. Montorsi, D. Temelkovski, M. R. Keddar, A. Acha-Sagredo, M. J. Pitcher, G. Basso, L. Laghi, M. Rodriguez-Justo, J. Spencer, and F. D. Ciccarelli, "A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially-resolved tissue phenotyping at single-cell resolution," *Nature Communications*, vol. 13, p. 781, Feb. 2022.

[63] B. D. Ripley, "Modelling Spatial Patterns," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 2, pp. 172–192, 1977.

[64] B. D. Ripley, *Statistical inference for spatial processes*. Cambridge: University Press, 1988.

[65] J.-R. Lin, S. Wang, S. Coy, M. Tyler, C. Yapp, Y.-A. Chen, C. N. Heiser, K. S. Lau, S. Santagata, and P. K. Sorger, "Multiplexed 3D atlas of state transitions and immune interactions in colorectal cancer," *bioRxiv*, p. 2021.03.31.437984, Apr. 2021.

[66] H. Failmezger, S. Muralidhar, A. Rullan, C. E. d. Andrea, E. Sahai, and Y. Yuan, "Topological Tumor Graphs: A Graph-Based Spatial Model to Infer Stromal Recruitment for Immunosuppression in Melanoma Histology," *Cancer Research*, vol. 80, pp. 1199–1209, Mar. 2020.

[67] E. Küçükköse, M. J. D. Baars, M. Amini, S. J. Schraa, E. Floor, G. M. Bol, I. H. M. Borel Rinkes, J. M. L. Roodhart, M. Koopman, J. Laoukili, O. Kranenburg, and Y. Vercoulen, "Stromal localization of inactive CD8+ T cells in metastatic mismatch repair deficient colorectal cancer," *British Journal of Cancer*, pp. 1–11, Dec. 2023. Publisher: Nature Publishing Group.

[68] A. Feichtenbeiner, M. Haas, M. Büttner, G. G. Grabenbauer, R. Fietkau, and L. V. Distel, "Critical role of spatial interaction between CD8+ and Foxp3+ cells in human gastric cancer: the distance matters," *Cancer Immunology, Immunotherapy*, vol. 63, pp. 111–119, Feb. 2014.

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

[70] N. Zamanitajeddin, M. Jahanifar, and N. Rajpoot, "Cells are Actors: Social Network Analysis with Classical ML for SOTA Histology Image Classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, eds.), Lecture Notes in Computer Science, (Cham), pp. 288–298, Springer International Publishing, 2021.

[71] L. O. Schwen, E. Andersson, K. Korski, N. Weiss, S. Haase, F. Gaire, H. K. Hahn, A. Homeyer, and O. Grimm, "Data-Driven Discovery of Immune Contexture Biomarkers," *Frontiers in Oncology*, vol. 8, 2018.

[72] E. R. Parra, "Methods to Determine and Analyze the Cellular Spatial Distribution Extracted From Multiplex Immunofluorescence Data to Understand the Tumor Microenvironment," *Frontiers in Molecular Biosciences*, vol. 8, 2021.

[73] P. Barmpoutis, M. D. Capite, H. Kayhanian, W. Waddingham, D. C. Alexander, M. Jansen, and F. N. K. Kwong, "Tertiary lymphoid structures (TLS) identification and density assessment on H&E-stained digital slides of lung cancer," *PLOS ONE*, vol. 16, p. e0256907, Sept. 2021.

[74] S. Javed, A. Mahmood, M. M. Fraz, N. A. Koohbanani, K. Benes, Y.-W. Tsang, K. Hewitt, D. Epstein, D. Snead, and N. Rajpoot, "Cellular community detection for tissue phenotyping in colorectal cancer histology images," *Medical Image Analysis*, vol. 63, p. 101696, July 2020.

[75] A. Mahmood, M. Small, S. A. Al-Maadeed, and N. Rajpoot, "Using Geodesic Space Density Gradients for Network Community Detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, pp. 921–935, Apr. 2017.

[76] D. Phillips, M. Matusiak, B. R. Gutierrez, S. S. Bhate, G. L. Barlow, S. Jiang, J. Demeter, K. S. Smythe, R. H. Pierce, S. P. Fling, N. Ramchurren, M. A. Cheever, Y. Goltsev, R. B. West, M. S. Khodadoust, Y. H. Kim, C. M. Schürch, and G. P. Nolan, "Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma," *Nature Communications*, vol. 12, p. 6726, Nov. 2021.

[77] P. Chen, M. Aminu, S. E. Hussein, J. D. Khoury, and J. Wu, "Hierarchical Phenotyping and Graph Modeling of Spatial Architecture in Lymphoid Neoplasms," *arXiv:2106.16174 [cs, eess, q-bio]*, Sept. 2021.

[78] P. J. Lawson, *Persistent Homology for the Quantification of Prostate Cancer Morphology in Two and Three-Dimensional Histology*. Ph.D., Tulane University School of Science and Engineering, United States – Louisiana, 2020. ISBN: 9798569914364.

[79] F. Wang, S. Kapse, S. Liu, P. Prasanna, and C. Chen, "TopoTxR: A Topological Biomarker for Predicting Treatment Response in Breast Cancer," *arXiv:2105.06049 [cs, eess, q-bio]*, May 2021.

[80] O. Vipond, J. A. Bull, P. S. Macklin, U. Tillmann, C. W. Pugh, H. M. Byrne, and H. A. Harrington, "Multiparameter persistent homology landscapes identify immune cell spatial patterns in tumors," *Proceedings of the National Academy of Sciences*, vol. 118, Oct. 2021.

[81] T. Gavrilchenko and E. Katifori, "Distribution networks achieve uniform perfusion through geometric self-organization," *arXiv:2009.04375 [nlin, physics:physics]*, Sept. 2020.

[82] M. Campisi, Y. Shin, T. Osaki, C. Hajal, V. Chiono, and R. D. Kamm, "3D self-organized microvascular model of the human blood-brain barrier with endothelial cells, pericytes and astrocytes," *Biomaterials*, vol. 180, pp. 117–129, Oct. 2018.

[83] P. R. Barber, S. M. Ameer-Beg, B. Vojnovic, R. J. Hodgkiss, G. M. Tozer, J. Wilson, and V. E. Prise, "3D Imaging and Quantification of Complex Vascular Networks," in *Confocal, Multiphoton, and Nonlinear Microscopic Imaging*, (Munich, Germany), p. 5139_67, OSA, 2003.

[84] E. M. Brey, T. W. King, C. Johnston, L. V. McIntire, G. P. Reece, and C. W. Patrick, "A Technique for Quantitative Three-Dimensional Analysis of Microvascular Structure," *Microvascular Research*, vol. 63, pp. 279–294, May 2002.

[85] M. Sun, C. Li, N. Chen, H. Zhao, L. Ma, C. Liu, Y. Shen, R. Lin, and X. Gong, "Full three-dimensional segmentation and quantification of tumor vessels for photoacoustic images," *Photoacoustics*, vol. 20, p. 100212, Dec. 2020.

[86] S. Heinzer, G. Kuhn, T. Krucker, E. Meyer, A. Ulmann-Schuler, M. Stampanoni, M. Gassmann, H. H. Marti, R. Müller, and J. Vogel, "Novel three-dimensional analysis tool for vascular trees indicates complete micro-networks, not single capillaries, as the angiogenic endpoint in mice overexpressing human VEGF165 in the brain," *NeuroImage*, vol. 39, pp. 1549–1558, Feb. 2008.

[87] P. Blinder, P. S. Tsai, J. P. Kaufhold, P. M. Knutsen, H. Suhl, and D. Kleinfeld, "The cortical angiome: an interconnected vascular network with noncolumnar patterns of blood flow," *Nature Neuroscience*, vol. 16, pp. 889–897, July 2013.

[88] J. A. Rytlewski, L. R. Geuss, C. I. Anyaeji, E. W. Lewis, and L. J. Suggs, "Three-Dimensional Image Quantification as a New Morphometry Method for Tissue Engineering," *Tissue Engineering Part C: Methods*, vol. 18, pp. 507–516, Jan. 2012.

[89] K. M. Meiburger, S. Y. Nam, E. Chung, L. J. Suggs, S. Y. Emelianov, and F. Molinari, "Skeletonization algorithm-based blood vessel quantification usingin vivo3D photoacoustic imaging," *Physics in Medicine and Biology*, vol. 61, pp. 7994–8009, Oct. 2016.

[90] A. Giacomini, M. Ackermann, M. Belleri, D. Coltrini, B. Nico, D. Ribatti, M. A. Konerding, M. Presta, and M. Righi, "Brain angioarchitecture and intussusceptive microvascular growth in a murine model of Krabbe disease," *Angiogenesis*, vol. 18, pp. 499–510, Oct. 2015.

[91] M. Righi, A. Giacomini, L. Cleris, and C. Carlo-Stella, "D Quantification of Tumor Vasculature in Lymphoma Xenografts in NOD/SCID Mice Allows to Detect Differences among Vascular-Targeted Therapies," *PLOS ONE*, vol. 8, p. e59691, Mar. 2013.

[92] M. Righi, S. L. Locatelli, C. Carlo-Stella, M. Presta, and A. Giacomini, "Vascular amounts and dispersion of caliber-classified vessels as key parameters to quantitate 3D micro-angioarchitectures in multiple myeloma experimental tumors," *Scientific Reports*, vol. 8, p. 17520, Nov. 2018.

[93] M. Righi, M. Belleri, M. Presta, and A. Giacomini, "Quantification of 3D Brain Microangioarchitectures in an Animal Model of Krabbe Disease," *International Journal of Molecular Sciences*, vol. 20, p. 2384, Jan. 2019.

[94] A. Freiman, Y. Shandalov, D. Rozenfeld, E. Shor, S. Segal, D. Ben-David, S. Meretzki, D. Egozi, and S. Levenberg, "Adipose-derived endothelial and mesenchymal stem cells enhance vascular network formation on three-dimensional constructs in vitro," *Stem Cell Research & Therapy*, vol. 7, p. 5, Jan. 2016.

[95] J. P. Kaufhold, P. S. Tsai, P. Blinder, and D. Kleinfeld, "Vectorization of optically sectioned brain microvasculature: Learning aids completion of vascular graphs by connecting gaps and deleting open-ended segments," *Medical Image Analysis*, vol. 16, pp. 1241–1258, Aug. 2012.

[96] E. Bullitt, G. Gerig, S. Pizer, Weili Lin, and S. Aylward, "Measuring tortuosity of the intracerebral vasculature from MRA images," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 1163–1171, Sept. 2003.

[97] D. Guidolin, A. Vacca, G. G. Nussdorfer, and D. Ribatti, "A new image analysis method based on topological and fractal parameters to evaluate the angiostatic activity of docetaxel by using the Matrigel assay in vitro," *Microvascular Research*, vol. 67, pp. 117–124, Mar. 2004.

[98] K. J. Falconer, *Fractal geometry : mathematical foundations and applications*. Chichester, England : Wiley, 2003.

[99] T. G. Smith, G. D. Lange, and W. B. Marks, "Fractal methods and results in cellular morphology — dimensions, lacunarity and multifractals," *Journal of Neuroscience Methods*, vol. 69, pp. 123–136, Nov. 1996.

[100] R. Albert and A. L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[101] C. J. Nelson and S. Bonner, "Neuronal graphs: a graph theory primer for microscopic, functional networks of neurons recorded by Calcium imaging," *arXiv:2010.09601 [q-bio]*, Oct. 2020.

[102] C. Sautès-Fridman, M. Lawand, N. A. Giraldo, H. Kaplon, C. Germain, W. H. Fridman, and M.-C. Dieu-Nosjean, "Tertiary Lymphoid Structures in Cancers: Prognostic Value, Regulation, and Manipulation for Therapeutic Intervention," *Frontiers in Immunology*, vol. 0, 2016.

[103] D. Harper, "capillary | Etymology, origin and meaning of capillary by etymonline." Accessed 2023-11-04 15:01:18.

[104] Medicine LibreTexts, "18.2F: Capillaries." Accessed 2023-11-04 15:11:48.

[105] H. O. Alsaab, S. Sau, R. Alzhrani, K. Tatiparti, K. Bhise, S. K. Kashaw, and A. K. Iyer, "PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome," *Frontiers in Pharmacology*, vol. 8, 2017.

[106] I. Vitale, E. Shema, S. Loi, and L. Galluzzi, "Intratumoral heterogeneity in cancer progression and response to immunotherapy," *Nature Medicine*, vol. 27, pp. 212–224, Feb. 2021.

[107] T. N. Schumacher and D. S. Thommen, "Tertiary lymphoid structures in cancer," *Science*, vol. 375, p. eabf9419, Jan. 2022.

[108] W. H. Fridman, L. Zitvogel, C. Sautès–Fridman, and G. Kroemer, "The immune contexture in cancer prognosis and treatment," *Nature Reviews Clinical Oncology*, vol. 14, pp. 717–734, Dec. 2017.

[109] M. Moehler, M. Delic, K. Goepfert, D. Aust, H. I. Grabsch, N. Halama, B. Heinrich, C. Julie, F. Lordick, M. P. Lutz, M. Mauer, M. Alsina Maqueda, H. Schild, C. C. Schimanski, A.-D. Wagner, A. Roth, and M. Ducreux, "Immunotherapy in gastrointestinal cancer: Recent results, current studies and future perspectives," *European Journal of Cancer*, vol. 59, pp. 160–170, May 2016.

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

[111] Biology LibreTexts, "11.1A: Cells and Organs of the Immune System," June 2018.

[112] C. D. Mills, K. Kincaid, J. M. Alt, M. J. Heilman, and A. M. Hill, "M-1/M-2 Macrophages and the Th1/Th2 Paradigm1," *The Journal of Immunology*, vol. 164, pp. 6166–6173, June 2000.

[113] D. Li, *Active modules identification in multilayer intracellular networks*. Ph.D., University of Birmingham, July 2018.

[114] M. Rakaee, L.-T. R. Busund, S. Jamaly, E.-E. Paulsen, E. Richardsen, S. Andersen, S. Al-Saad, R. M. Bremnes, T. Donnem, and T. K. Kilvaer, "Prognostic Value of Macrophage Phenotypes in Resectable Non–Small Cell Lung Cancer Assessed by Multiplex Immunohistochemistry," *Neoplasia*, vol. 21, pp. 282–293, Mar. 2019.

[115] M. L. Pinto, E. Rios, C. Durães, R. Ribeiro, J. C. Machado, A. Mantovani, M. A. Barbosa, F. Carneiro, and M. J. Oliveira, "The Two Faces of Tumor-Associated Macrophages and Their Clinical Significance in Colorectal Cancer," *Frontiers in Immunology*, vol. 10, 2019.

[116] A. López-Janeiro, C. Padilla-Ansala, C. E. de Andrea, D. Hardisson, and I. Melero, "Prognostic value of macrophage polarization markers in epithelial neoplasms and melanoma. A systematic review and meta-analysis," *Modern Pathology*, vol. 33, pp. 1458–1465, Aug. 2020.

[117] J. Jackute, M. Zemaitis, D. Pranys, B. Sitkauskiene, S. Miliauskas, S. Vaitkiene, and R. Sakalauskas, "Distribution of M1 and M2 macrophages in tumor islets and stroma in relation to prognosis of non-small cell lung cancer," *BMC Immunology*, vol. 19, p. 3, Jan. 2018.

[118] Valeria Barresi, Luca Reggiani Bonetti, Antonio Leni, Rosario Alberto Caruso, and Giovanni Tuccari, "Histological grading in colorectal cancer: new insights and perspectives," *Histology and Histopathology*, vol. 30, pp. 1059–1067, Aug. 2015.

[119] S. R. Hamilton, F. T. Bosman, P. Boffetta, E. Riboli, and L. H. Sobin, "Carcinoma of the colon and rectum," *WHO classification of tumors of the digestive system*, pp. 134–146, 2010.

[120] V. Barresi, L. Reggiani Bonetti, G. Branca, C. Di Gregorio, M. Ponz de Leon, and G. Tuccari, "Colorectal carcinoma grading by quantifying poorly differentiated cell clusters is more reproducible and provides more robust prognostic information than conventional grading," *Virchows Archiv*, vol. 461, pp. 621–628, Dec. 2012.

[121] D. M. Hari, A. M. Leung, J.-H. Lee, M.-S. Sim, B. Vuong, C. G. Chiu, and A. J. Bilchik, "AJCC Cancer Staging Manual 7th Edition Criteria for Colon Cancer: Do the Complex Modifications Improve Prognostic Assessment?," *Journal of the American College of Surgeons*, vol. 217, pp. 181–190, Aug. 2013.

[122] K. Ganesh, Z. K. Stadler, A. Cercek, R. B. Mendelsohn, J. Shia, N. H. Segal, and L. A. Diaz, "Immunotherapy in colorectal cancer: rationale, challenges and potential," *Nature Reviews Gastroenterology & Hepatology*, vol. 16, pp. 361–375, June 2019.

[123] Y. Li, Y. Li, Z. Xia, D. Zhang, X. Chen, X. Wang, J. Liao, W. Yi, and J. Chen, "Identification of a novel immune signature for optimizing prognosis and treatment prediction in colorectal cancer," *Aging (Albany NY)*, vol. 13, pp. 25518–25549, Dec. 2021.

[124] F. Yu, A. Makrigiorgos, K. W. Leong, and G. M. Makrigiorgos, "Sensitive detection of microsatellite instability in tissues and liquid biopsies: Recent developments and updates," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 4931–4940, Jan. 2021.

[125] R. Yamashita, J. Long, T. Longacre, L. Peng, G. Berry, B. Martin, J. Higgins, D. L. Rubin, and J. Shen, "Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study," *The Lancet Oncology*, vol. 22, pp. 132–141, Jan. 2021.

[126] C. Rosty, E. J. Williamson, M. Clendenning, R. J. Walters, A. K. Win, M. A. Jenkins, J. L. Hopper, I. M. Winship, M. C. Southey, G. G. Giles, D. R. English, and D. D. Buchanan, "Should the grading of colorectal adenocarcinoma include microsatellite instability status?," *Human Pathology*, vol. 45, pp. 2077–2084, Oct. 2014.

[127] J. Galon, B. Mlecnik, G. Bindea, H. K. Angell, A. Berger, C. Lagorce, A. Lugli, I. Zlobec, A. Hartmann, C. Bifulco, I. D. Nagtegaal, R. Palmqvist, G. V. Masucci, G. Botti, F. Tatangelo, P. Delrio, M. Maio, L. Laghi, F. Grizzi, M. Asslaber, C. D'Arrigo, F. Vidal-Vanaclocha, E. Zavadova, L. Chouchane, P. S. Ohashi, S. Hafezi-Bakhtiari, B. G. Wouters, M. Roehrl, L. Nguyen, Y. Kawakami, S. Hazama, K. Okuno, S. Ogino, P. Gibbs, P. Waring, N. Sato, T. Torigoe, K. Itoh, P. S. Patel, S. N. Shukla, Y. Wang, S. Kopetz, F. A. Sinicrope, V. Scripcariu, P. A. Ascierto, F. M. Marincola, B. A. Fox, and F. Pagès, "Towards the introduction of the 'Immunoscore' in the classification of malignant tumours," *The Journal of Pathology*, vol. 232, no. 2, pp. 199–209, 2014.

[128] J. Galon, A. Costes, F. Sanchez-Cabo, A. Kirilovsky, B. Mlecnik, C. Lagorce-Pagès, M. Tosolini, M. Camus, A. Berger, P. Wind, F. Zinzindohoué, P. Bruneval, P.-H. Cugnenc, Z. Trajanoski, W.-H. Fridman, and F. Pagès, "Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome," *Science*, vol. 313, pp. 1960–1964, Sept. 2006.

[129] N. Halama, S. Michel, M. Kloor, I. Zoernig, A. Benner, A. Spille, T. Pommerencke, D. M. von Knebel, G. Folprecht, B. Luber, N. Feyen, U. M. Martens, P. Beckhove, S. Gnjatic, P. Schirmacher, E. Herpel, J. Weitz, N. Grabe, and D. Jaeger, "Localization and Density of Immune Cells in the Invasive Margin of Human Colorectal Cancer Liver Metastases Are Prognostic for Response to Chemotherapy," *Cancer Research*, vol. 71, pp. 5670–5677, Aug. 2011.

[130] C. F. Kuper, C. Ruehl-Fehlert, S. A. Elmore, and G. A. Parker, "Chapter 49 - Immune System," in *Haschek and Rousseaux's Handbook of Toxicologic Pathology (Third Edition)* (W. M. Haschek, C. G. Rousseaux, and M. A. Wallig, eds.), pp. 1795–1862, Boston: Academic Press, Jan. 2013.

[131] F. Posch, K. Silina, U. Petrausch, S. Leibl, A. Muendlein, H. Moch, A. R. Siebenhuener, B. C. Pestalozzi, J. M. Riedl, M. Stotz, J. Szkandera, H. Stoger, M. Pichler, R. Stupp, P. Schraml, A. Gerger, and T. Winder, "The maturation stage of tumoral tertiary lymphoid structures to predict recurrence risk in localized colorectal cancer.," *Journal of Clinical Oncology*, vol. 35, pp. e15083–e15083, May 2017.

[132] A. Maoz, M. Dennis, and J. K. Greenson, "The Crohn's-Like Lymphoid Reaction to Colorectal Cancer-Tertiary Lymphoid Structures With Immunologic and Potentially Therapeutic Relevance in Colorectal Cancer," *Frontiers in Immunology*, vol. 10, 2019.

[133] Akoya Biosciences, "Vectra Polaris."

[134] Akoya Biosciences, "PhenoChart Whole Slide Viewer."

[135] Akoya Biosciences, "HALO."

[136] T. Yang, V. Ozcoban, A. Pasam, N. Kocovski, A. Pizzolla, Y.-K. Huang, G. Bass, S. P. Keam, P. J. Neeson, S. K. Sandhu, D. L. Goode, and A. S. Trigos, "SPIAT: An R package for the Spatial Image Analysis of Cells in Tissues," May 2020.

[137] D. Righelli, L. M. Weber, H. L. Crowell, B. Pardo, L. Collado-Torres, S. Ghazanfar, A. T. L. Lun, S. C. Hicks, and D. Risso, "SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor," *Bioinformatics*, vol. 38, pp. 3128–3131, June 2022.

[138] GEOS contributors, *GEOS coordinate transformation software library*. Open Source Geospatial Foundation, 2021.

[139] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Transactions on Information Theory*, vol. 29, pp. 551–559, July 1983.

[140] C. C. Maley, K. Koelble, R. Natrajan, A. Aktipis, and Y. Yuan, "An ecological measure of immune-cancer colocalization as a prognostic factor for breast cancer," *Breast Cancer Research*, vol. 17, p. 131, Sept. 2015.

[141] F. Pagès, B. Mlecnik, F. Marliot, G. Bindea, F.-S. Ou, C. Bifulco, A. Lugli, I. Zlobec, T. T. Rau, M. D. Berger, I. D. Nagtegaal, E. Vink-Börger, A. Hartmann, C. Geppert, J. Kolwelter, S. Merkel, R. Grützmann, M. Van den Eynde, A. Jouret-Mourin, A. Kartheuser, D. Léonard, C. Remue, J. Y. Wang, P. Bavi, M. H. A. Roehrl, P. S. Ohashi, L. T. Nguyen, S. Han, H. L. MacGregor, S. Hafezi-Bakhtiari, B. G. Wouters, G. V. Masucci, E. K. Andersson, E. Zavadova, M. Vocka, J. Spacek, L. Petruzelka, B. Konopasek, P. Dundr, H. Skalova, K. Nemejcova, G. Botti, F. Tatangelo, P. Delrio, G. Ciliberto, M. Maio, L. Laghi, F. Grizzi, T. Fredriksen, B. Buttard, M. Angelova, A. Vasaturo, P. Maby, S. E. Church, H. K. Angell, L. Lafontaine, D. Bruni, C. El Sissy, N. Haicheur, A. Kirilovsky, A. Berger, C. Lagorce, J. P. Meyers, C. Paustian, Z. Feng, C. Ballesteros-Merino, J. Dijkstra, C. van de Water, S. van Lent-van Vliet, N. Knijn, A.-M. Mușină, D.-V. Scripcariu, B. Popivanova, M. Xu, T. Fujita, S. Hazama, N. Suzuki, H. Nagano, K. Okuno, T. Torigoe, N. Sato, T. Furuhata, I. Takemasa, K. Itoh, P. S. Patel, H. H. Vora, B. Shah, J. B. Patel, K. N. Rajvik, S. J. Pandya, S. N. Shukla, Y. Wang, G. Zhang, Y. Kawakami, F. M. Marincola, P. A. Ascierto, D. J. Sargent, B. A. Fox, and J. Galon, "International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study," *The Lancet*, vol. 391, pp. 2128–2139, May 2018.

[142] B. R. Vatti, "A generic solution to polygon clipping," *Communications of the ACM*, vol. 35, pp. 56–63, July 1992.

[143] X. Chen and S. McMains, "Polygon Offsetting by Computing Winding Numbers," in *Volume 2: 31st Design Automation Conference, Parts A and B*, (Long Beach, California, USA), pp. 565–575, ASMEDC, Jan. 2005.

[144] M. K. Agoston, *Computer Graphics and Geometric Modelling*. Springer Science & Business Media, Jan. 2005.

[145] Dietrich Stoyan and Helga Stoyan, *Fractals, random shapes and point fields: Methods of geometrical statistics*, vol. 302. John Wiley & Sons Incorporated, 1994.

[146] E. Gavagnin, J. P. Owen, and C. A. Yates, "Pair correlation functions for identifying spatial correlation in discrete domains," *Physical Review E*, vol. 97, p. 062104, June 2018.

[147] Y. Yoshihara, T. Ohkuro, B. Buuveibaatar, U. Jamsran, and K. Takeuchi, "Spatial pattern of grazing affects influence of herbivores on spatial heterogeneity of plants and soils," *Oecologia*, vol. 162, pp. 427–434, Feb. 2010.

[148] S. Getzin, T. Wiegand, K. Wiegand, and F. He, "Heterogeneity influences spatial patterns and demographics in forest stands," *Journal of Ecology*, vol. 96, no. 4, pp. 807–820, 2008.

[149] S. Barua, P. Fang, A. Sharma, J. Fujimoto, I. Wistuba, A. U. K. Rao, and S. H. Lin, "Spatial interaction of tumor cells and regulatory T cells correlates with survival in non-small cell lung cancer," *Lung Cancer*, vol. 117, pp. 73–79, Mar. 2018.

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

[151] A. Fick, "Ueber Diffusion," *Annalen der Physik*, vol. 170, no. 1, pp. 59–86, 1855.

[152] K. Francis and B. O. Palsson, "Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion," *Proceedings of the National Academy of Sciences*, vol. 94, pp. 12258–12262, Nov. 1997.

[153] P. Pati, G. Jaume, A. Foncubierta-Rodríguez, F. Feroce, A. M. Anniciello, G. Scognamiglio, N. Brancati, M. Fiche, E. Dubruc, D. Riccio, M. Di Bonito, G. De Pietro, G. Botti, J.-P. Thiran, M. Frucci, O. Goksel, and M. Gabrani, "Hierarchical graph representations in digital pathology," *Medical Image Analysis*, vol. 75, p. 102264, Jan. 2022.

[154] J.-R. Lin, S. Wang, S. Coy, Y.-A. Chen, C. Yapp, M. Tyler, M. K. Nariya, C. N. Heiser, K. S. Lau, S. Santagata, and P. K. Sorger, "Multiplexed 3D atlas of state transitions and immune interaction in colorectal cancer," *Cell*, vol. 186, pp. 363–381.e19, Jan. 2023. Publisher: Elsevier.

[155] C. Cable, K. F. Jones, J. R. Lundgren, and S. Seager, "Niche graphs," *Discrete Applied Mathematics*, vol. 23, pp. 231–241, June 1989.

[156] S. Lange, R. Mogwitz, D. Hünniger, and A. Voss-Böhme, "Modeling age-specific incidence of colon cancer via niche competition," July 2022.

[157] C. Greggio, F. De Franceschi, M. Figueiredo-Larsen, S. Gobaa, A. Ranga, H. Semb, M. Lutolf, and A. Grapin-Botton, "Artificial three-dimensional niches deconstruct pancreas development in vitro," *Development (Cambridge)*, vol. 140, no. 21, pp. 4452–4462, 2013.

[158] A. L. Martinelli and M. A. Rapsomaniki, "ATHENA: analysis of tumor heterogeneity from spatial omics measurements," *Bioinformatics*, vol. 38, pp. 3151–3153, May 2022.

[159] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[160] E. Pebesma, "Simple Features for R: Standardized Support for Spatial Vector Data," *The R Journal*, vol. 10, no. 1, pp. 439–446, 2018.

[161] E. P. Bivand, Roger, *Spatial Data Science: With Applications in R*. New York: Chapman and Hall/CRC, May 2023.

[162] G. Csárdi and T. Nepusz, "The igraph software package for complex network research," 2006.

[163] A. Johnson, "angusj.com."

[164] J. A. Pike, A. O. Khan, C. Pallini, S. G. Thomas, M. Mund, J. Ries, N. S. Poulter, and I. B. Styles, "Topological data analysis quantifies biological nano-structure from single molecule localization microscopy," *Bioinformatics*, vol. 36, pp. 1614–1621, Mar. 2020.

[165] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Use R!, New York: Springer-Verlag, 2 ed., 2016.

[166] V. Greiff, P. Bhat, S. C. Cook, U. Menzel, W. Kang, and S. T. Reddy, "A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status," *Genome Medicine*, vol. 7, p. 49, May 2015.

[167] B. D. Ripley, *Spatial statistics*. Wiley series in probability and mathematical statistics, New York ; Chichester: Wiley, 1981.

[168] G. Onofre, M. Koláčková, K. Jankovičová, and J. Krejsek, "Scavenger Receptor CD163 and Its Biological Functions," *Acta Medica (Hradec Kralove, Czech Republic)*, vol. 52, no. 2, pp. 57–61, 2009.

[169] J. H. Graversen, M. Madsen, and S. K. Moestrup, "CD163: a signal receptor scavenging haptoglobin–hemoglobin complexes from plasma," *The International Journal of Biochemistry & Cell Biology*, vol. 34, pp. 309–314, Apr. 2002.

[170] B. O. Fabriek, C. D. Dijkstra, and T. K. van den Berg, "The macrophage scavenger receptor CD163," *Immunobiology*, vol. 210, pp. 153–160, Aug. 2005.

[171] B. H. Davis and P. V. Zarev, "Human monocyte CD163 expression inversely correlates with soluble CD163 plasma levels," *Cytometry Part B: Clinical Cytometry*, vol. 63B, no. 1, pp. 16–22, 2005.

[172] J. A. Bull, E. J. Mulholland, S. J. Leedham, and H. M. Byrne, "Extended correlation functions for spatial analysis of multiplex imaging data," June 2023.

[173] M. Mattiazzi Usaj, E. B. Styles, A. J. Verster, H. Friesen, C. Boone, and B. J. Andrews, *High-Content Screening for Quantitative Cell Biology*, vol. 24. Elsevier Ltd, Aug. 2016.

[174] P. M. A. Antony, C. Trefois, A. Stojanovic, A. S. Baumuratov, and K. Kozak, "Light microscopy applications in systems biology: opportunities and challenges," *Cell Communication and Signaling*, vol. 11, p. 24, Apr. 2013.

[175] A. Reinke, M. D. Tizabi, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A. E. Kavur, T. Rädsch, C. H. Sudre, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M. Blaschko, F. Buettner, M. J. Cardoso, V. Cheplygina, J. Chen, E. Christodoulou, B. A. Cimini, G. S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, B. Glocker, P. Godau, R. Haase, D. A. Hashimoto, M. M. Hoffman, M. Huisman, F. Isensee, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, H. Kenngott, J. Kleesiek, F. Kofler, T. Kooi, A. Kopp-Schneider, M. Kozubek, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, S. M. Rafelski, N. Rajpoot, M. Reyes, M. A. Riegler, N. Rieke, J. Saez-Rodriguez, C. I. Sánchez, S. Shetty, M. van Smeden, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. Van Calster, G. Varoquaux, M. Wiesenfarth, Z. R. Yaniv, P. F. Jäger, and L. Maier-Hein, "Understanding metric-related pitfalls in image analysis validation," Sept. 2023.

[176] A. E. Carpenter, L. Kamentsky, and K. W. Eliceiri, "A call for bioimaging software usability," *Nature Methods*, vol. 9, pp. 666–670, July 2012.

[177] nhs.uk, "Sjögren's syndrome."

[178] E. J. Colbeck, A. Ager, A. Gallimore, and G. W. Jones, "Tertiary Lymphoid Structures in Cancer: Drivers of Antitumor Immunity, Immunosuppression, or Bystander Sentinels in Disease?," *Frontiers in Immunology*, vol. 0, 2017.

[179] C. Sautès-Fridman, F. Petitprez, J. Calderaro, and W. H. Fridman, "Tertiary lymphoid structures in the era of cancer immunotherapy," *Nature Reviews Cancer*, vol. 19, pp. 307–325, June 2019.

[180] B. A. Helmink, S. M. Reddy, J. Gao, S. Zhang, R. Basar, R. Thakur, K. Yizhak, M. Sade-Feldman, J. Blando, G. Han, V. Gopalakrishnan, Y. Xi, H. Zhao, R. N. Amaria, H. A. Tawbi, A. P. Cogdill, W. Liu, V. S. LeBleu, F. G. Kugeratski, S. Patel, M. A. Davies, P. Hwu, J. E. Lee, J. E. Gershenwald, A. Lucci, R. Arora, S. Woodman, E. Z. Keung, P.-O. Gaudreau, A. Reuben, C. N. Spencer, E. M. Burton, L. E. Haydu, A. J. Lazar, R. Zapassodi, C. W. Hudgens, D. A. Ledesma, S. Ong, M. Bailey, S. Warren, D. Rao, O. Krijgsman, E. A. Rozeman, D. Peeper, C. U. Blank, T. N. Schumacher, L. H. Butterfield, M. A. Zelazowska, K. M. McBride, R. Kalluri, J. Allison, F. Petitprez, W. H. Fridman, C. Sautès-Fridman, N. Hacohen, K. Rezvani, P. Sharma, M. T. Tetzlaff, L. Wang, and J. A. Wargo, "B cells and tertiary lymphoid structures promote immunotherapy response," *Nature*, vol. 577, pp. 549–555, Jan. 2020.

[181] L. Vanhersecke, M. Brunet, J.-P. Guégan, C. Rey, A. Bougouin, S. Cousin, S. L. Moulec, B. Besse, Y. Loriot, M. Larroquette, I. Soubeyran, M. Toulmonde, G. Roubaud, S. Pernot, M. Cabart, F. Chomy, C. Lefevre, K. Bourcier, M. Kind, I. Giglioli, C. Sautès-Fridman, V. Velasco, F. Courgeon, E. Oflazoglu, A. Savina, A. Marabelle, J.-C. Soria, C. Bellera, C. Sofeu, A. Bessede, W. H. Fridman, F. L. Loarer, and A. Italiano, "Mature tertiary lymphoid structures predict immune checkpoint inhibitor efficacy in solid tumors independently of PD-L1 expression," *Nature cancer*, vol. 2, pp. 794–802, Aug. 2021.

[182] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, pp. 62–66, Jan. 1979.

[183] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), Lecture Notes in Computer Science, (Cham), pp. 234–241, Springer International Publishing, 2015.

[184] C. Lantuejoul and S. Beucher, "Quelques considérations sur la distance géodésique.," *Note interne du Centre de Morphologie mathématique.*, 1979.

[185] G. Li, T. Liu, J. Nie, L. Guo, J. Chen, J. Zhu, W. Xia, A. Mara, S. Holley, and S. Wong, "Segmentation of touching cell nuclei using gradient flow tracking," *Journal of Microscopy*, vol. 231, no. 1, pp. 47–58, 2008.

[186] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman, J. A. James, M. Salto-

Tellez, and P. W. Hamilton, "QuPath: Open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, p. 16878, Dec. 2017.

[187] P. Bankhead, "How QuPath calculate the centroid coordinate for each cell? - Usage & Issues," June 2020.

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

[189] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnology*, vol. 37, pp. 38–44, Jan. 2019.

[190] L. V. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[191] D. J. Nieves, J. A. Pike, F. Levet, D. J. Williamson, M. Baragilly, S. Oloketuyi, A. de Marco, J. Griffié, D. Sage, E. A. K. Cohen, J.-B. Sibarita, M. Heilemann, and D. M. Owen, "A framework for evaluating the performance of SMLM cluster analysis algorithms," *Nature Methods*, vol. 20, pp. 259–267, Feb. 2023.

[192] M. Ester, H.-P. Kriegel, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *KDD-96 Proceedings*, 1996.

[193] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Advances in Knowledge Discovery and Data Mining* (J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 160–172, Springer, 2013.

[194] T. Barton and T. Bruna, "Clustering benchmarks," July 2023.

[195] W. F. Eddy, "A New Convex Hull Algorithm for Planar Sets," *ACM Transactions on Mathematical Software*, vol. 3, pp. 398–403, Dec. 1977.

[196] W. F. Eddy, "Algorithm 523: CONVEX, A New Convex Hull Algorithm for Planar Sets [Z]," *ACM Transactions on Mathematical Software*, vol. 3, pp. 411–412, Dec. 1977.

[197] G. Landini, A. Galton, D. Randell, and S. Fouad, "Novel applications of discrete mereotopology to mathematical morphology," *Signal Processing: Image Communication*, vol. 76, pp. 109–117, Aug. 2019.

[198] J. A. Pike, *Quantitative microscopy workflows for the study of cellular receptor trafficking*. Ph.D., University of Birmingham, July 2017.

[199] B. T. Fasy, J. Kim, F. Lecci, C. Maria, D. L. Millman, and V. Rouvreau, "TDA: Statistical Tools for Topological Data Analysis," Feb. 2023.

[200] W. L. G. Koontz, P. Narendra, and K. Fukunaga, "A Graph-Theoretic Approach to Nonparametric Cluster Analysis," *IEEE Transactions on Computers*, vol. C-25, pp. 936–944, Sept. 1976.

[201] K. You, "mclustcomp: Measures for Comparing Clusters," June 2021.

[202] S. Fleig, T. Kapanadze, J. Bernier-Latmani, J. K. Lill, T. Wyss, J. Gamrekelashvili, D. Kijas, B. Liu, A. M. Hüsing, E. Bovay, A. C. Jirmo, S. Halle, M. Ricke-Hoch, R. H. Adams, D. R. Engel, S. von Vietinghoff, R. Förster, D. Hilfiker-Kleiner, H. Haller, T. V. Petrova, and F. P. Limbourg, "Loss of vascular endothelial notch signaling promotes spontaneous formation of tertiary lymphoid structures," *Nature Communications*, vol. 13, p. 2022, Apr. 2022.

[203] A. Ahmed, S. Köhler, R. Klotz, N. Giese, T. Hackert, C. Springfeld, I. Zörnig, D. Jäger, and N. Halama, "Tertiary lymphoid structures and their association to immune phenotypes and circulatory IL2 levels in pancreatic ductal adenocarcinoma," *OncoImmunology*, vol. 11, p. 2027148, Dec. 2022.

[204] S. Nayar, *Lymphoid like stromal cells in a model of tertiary lymphoid organ formation*. Ph.D., University of Birmingham, Dec. 2014.

[205] N. J. Krautler, V. Kana, J. Kranich, Y. Tian, D. Perera, D. Lemm, P. Schwarz, A. Armulik, J. L. Browning, M. Tallquist, T. Buch, J. B. Oliveira-Martins, C. Zhu, M. Hermann, U. Wagner, R. Brink, M. Heikenwalder, and A. Aguzzi, "Follicular Dendritic Cells Emerge from Ubiquitous Perivascular Precursors," *Cell*, vol. 150, pp. 194–206, July 2012.

[206] F. Humby, M. Bombardieri, A. Manzo, S. Kelly, M. C. Blades, B. Kirkham, J. Spencer, and C. Pitzalis, "Ectopic Lymphoid Structures Support Ongoing Production of Class-Switched Autoantibodies in Rheumatoid Synovium," *PLOS Medicine*, vol. 6, p. e1, Jan. 2009.

[207] E. Corsiero, M. Bombardieri, A. Manzo, S. Bugatti, M. Uguccioni, and C. Pitzalis, "Role of lymphoid chemokines in the development of functional ectopic lymphoid structures in rheumatic autoimmune diseases," *Immunology Letters*, vol. 145, pp. 62–67, July 2012.

[208] D. R. Kroeger, K. Milne, and B. H. Nelson, "Tumor-Infiltrating Plasma Cells Are Associated with Tertiary Lymphoid Structures, Cytolytic T-Cell Responses, and Superior Prognosis in Ovarian Cancer," *Clinical Cancer Research*, vol. 22, pp. 3005–3015, June 2016.

[209] N. F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, T. Dougherty, C. C. Fullaway, B. J. McIntosh, K. X. Leow, M. S. Schwartz, C. Pavelchek, S. Cui, I. Camplisson, O. Bar-Tal, J. Singh, M. Fong, G. Chaudhry, Z. Abraham, J. Moseley, S. Warshawsky, E. Soon, S. Greenbaum, T. Risom, T. Hollmann, S. C. Bendall, L. Keren, W. Graf, M. Angelo, and D. Van Valen, "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," *Nature Biotechnology*, vol. 40, pp. 555–565, Apr. 2022.

[210] O. Draganov and M. Mahini, "chromatic-tda," Nov. 2023. Accessed 2023-11-28 10:41:55.

[211] F. Cassot, F. Lauwers, C. Fouard, S. Prohaska, and V. Lauwers-Cances, "A Novel Three-Dimensional Computer-Assisted Method for a Quantitative Study of Microvascular Networks of the Human Cerebral Cortex," *Microcirculation*, vol. 13, pp. 1–18, Jan. 2006.

[212] D. Mayerich, J. Kwon, C. Sung, L. Abbott, J. Keyser, and Y. Choe, "Fast macro-scale transmission imaging of microvascular networks using KESM," *Biomedical Optics Express*, vol. 2, pp. 2888–2896, Oct. 2011.

[213] G. Chan, C. Balaratnasingam, P. K. Yu, W. H. Morgan, I. L. McAllister, S. J. Cringle, and D.-Y. Yu, "Quantitative Morphometry of Perifoveal Capillary Networks in the Human Retina," *Investigative Ophthalmology & Visual Science*, vol. 53, pp. 5502–5514, Aug. 2012.

[214] J. P. H. M. van den Wijngaard, J. C. V. Schwarz, P. van Horssen, M. G. J. T. B. van Lier, J. G. G. Dobbe, J. A. E. Spaan, and M. Siebes, "3D Imaging of vascular networks for biophysical modeling of perfusion distribution within the heart," *Journal of Biomechanics*, vol. 46, pp. 229–239, Jan. 2013.

[215] R. Plitman Mayo, D. S. Charnock-Jones, G. J. Burton, and M. L. Oyen, "Three-dimensional modeling of human placental terminal villi," *Placenta*, vol. 43, pp. 54–60, July 2016.

[216] P. Pearce, P. Brownbill, J. Janáček, M. Jirkovská, L. Kubínová, I. L. Chernyavsky, and O. E. Jensen, "Image-Based Modeling of Blood Flow and Oxygen Transfer in Feto-Placental Capillaries," *PLOS ONE*, vol. 11, p. e0165369, Oct. 2016.

[217] M. A. Konerding, W. Malkusch, B. Klapthor, C. v. Ackern, E. Fait, S. A. Hill, C. Parkins, D. J. Chaplin, M. Presta, and J. Denekamp, "Evidence for characteristic vascular patterns in solid tumours: quantitative studies using corrosion casts," *British Journal of Cancer*, vol. 80, pp. 724–732, May 1999.

[218] M. A. Konerding, E. Fait, and A. Gaumann, "3D microvascular architecture of precancerous lesions and invasive carcinomas of the colon," *British Journal of Cancer*, vol. 84, pp. 1354–1362, May 2001.

[219] A. A. Folarin, M. A. Konerding, J. Timonen, S. Nagl, and R. B. Pedley, "Three-dimensional analysis of tumour vascular corrosion casts using stereoimaging and microcomputed tomography," *Microvascular Research*, vol. 80, pp. 89–98, July 2010.

[220] A. d'Esposito, P. W. Sweeney, M. Ali, M. Saleh, R. Ramasawmy, T. A. Roberts, G. Agliardi, A. Desjardins, M. F. Lythgoe, R. B. Pedley, R. Shipley, and S. Walker-Samuel, "Computational fluid dynamics with imaging of cleared tissue and of in vivo perfusion predicts drug uptake and treatment responses in tumours," *Nature Biomedical Engineering*, vol. 2, pp. 773–787, Oct. 2018.

[221] W. Risau, H. Sariola, H.-G. Zerwes, J. Sasse, P. Ekblom, R. Kemler, and T. Doetschman, "Vasculogenesis and angiogenesis in embryonic-stem-cell-derived embryoid bodies," *Development*, vol. 102, pp. 471–478, Mar. 1988.

[222] L. Lamalice, F. Le Boeuf, and J. Hout, "Endothelial Cell Migration During Angiogenesis," *Circulation Research*, vol. 100, pp. 782–794, Mar. 2007.

[223] X. Yang and C. L. Cepko, "Flk-1, a Receptor for Vascular Endothelial Growth Factor (VEGF), Is Expressed by Retinal Progenitor Cells," *Journal of Neuroscience*, vol. 16, pp. 6089–6099, Oct. 1996.

[224] P. Carmeliet, "Angiogenesis in health and disease," *Nature Medicine*, vol. 9, pp. 653–660, June 2003.

[225] W. Risau and I. Flamme, "Vasculogenesis," *Annual review of cell and developmental biology*, vol. 11, no. 1, pp. 73–91, 1995.

[226] C. J. Drake, "Embryonic and adult vasculogenesis," *Birth Defects Research Part C: Embryo Today: Reviews*, vol. 69, no. 1, pp. 73–82, 2003.

[227] H. K. Kimelberg and M. D. Norenberg, "Astrocytes," *Scientific American*, vol. 260, no. 4, pp. 66–77, 1989.

[228] B. V. Zlokovic, "Neurovascular mechanisms of Alzheimer's neurodegeneration," *Trends in Neurosciences*, vol. 28, pp. 202–208, Apr. 2005.

[229] R. D. Bell and B. V. Zlokovic, "Neurovascular mechanisms and blood–brain barrier disorder in Alzheimer's disease," *Acta Neuropathologica*, vol. 118, pp. 103–113, July 2009.

[230] R. Kortekaas, K. L. Leenders, J. C. H. van Oostrom, W. Vaalburg, J. Bart, A. T. M. Willemsen, and N. H. Hendrikse, "Blood–brain barrier dysfunction in parkinsonian midbrain in vivo," *Annals of Neurology*, vol. 57, no. 2, pp. 176–179, 2005.

[231] E. Waubant, "Biomarkers indicative of blood-brain barrier disruption in multiple sclerosis," *Disease Markers*, vol. 22, pp. 235–244, Jan. 2006.

[232] R. K. Jain, E. di Tomaso, D. G. Duda, J. S. Loeffler, A. G. Sorensen, and T. T. Batchelor, "Angiogenesis in brain tumours," *Nature Reviews Neuroscience*, vol. 8, pp. 610–622, Aug. 2007.

[233] B. Dehouck, M.-P. Dehouck, J.-C. Fruchart, and R. Cecchelli, "Upregulation of the Low Density Lipoprotein Receptor at the Blood-Brain Barrier: Intercommunications between Brain Capillary Endothelial Cells and Astrocytes," *Review & Expositor*, vol. 84, pp. 465–473, Feb. 1987.

[234] R. C. Janzer and M. C. Raff, "Astrocytes induce blood–brain barrier properties in endothelial cells," *Nature*, vol. 325, pp. 253–257, Jan. 1987.

[235] J. Laterra, C. Guerin, and G. W. Goldstein, "Astrocytes induce neural microvascular endothelial cells to form capillary-like structures in vitro," *Journal of Cellular Physiology*, vol. 144, no. 2, pp. 204–215, 1990.

[236] G. R. J. Gordon, S. J. Mulligan, and B. A. MacVicar, "Astrocyte control of the cerebrovasculature," *Glia*, vol. 55, no. 12, pp. 1214–1221, 2007.

[237] A. Mishra, J. P. Reynolds, Y. Chen, A. V. Gourine, D. A. Rusakov, and D. Attwell, "Astrocytes mediate neurovascular signaling to capillary pericytes but not to arterioles," *Nature Neuroscience*, vol. 19, pp. 1619–1627, Dec. 2016.

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

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

[240] R. Chan, C.-W. Ho, and M. Nikolova, "Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization," *IEEE Transactions on Image Processing*, vol. 14, pp. 1479–1485, Oct. 2005.

[241] M. Fujita, S. Sasayama, C. Kawai, S. Eiho, and M. Kuwahara, "Automatic processing of cineventriculograms for analysis of regional myocardial function.," *Circulation*, vol. 63, pp. 1065–1074, May 1981.

[242] MathWorks, "MATLAB," 2020.

[243] T. C. Lee, R. L. Kashyap, and C. N. Chu, "Building Skeleton Models via 3-D Medial Surface Axis Thinning Algorithms," *CVGIP: Graphical Models and Image Processing*, vol. 56, pp. 462–478, Nov. 1994.

[244] J. Ollion, J. Cochennec, F. Loll, C. Escudé, and T. Boudier, "TANGO: A generic tool for high-throughput 3D image analysis for studying nuclear organization," *Bioinformatics*, vol. 29, no. 14, pp. 1840–1841, 2013.

[245] T. Hildebrand and P. Rüegsegger, "A new method for the model-independent assessment of thickness in three-dimensional images," *Journal of Microscopy*, vol. 185, no. 1, pp. 67–75, 1997.

[246] M. Rubinov and O. Sporns, "Brain Connectivity Toolbox." Accessed 2023-12-13 16:21:39.

[247] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, p. 8577, June 2006.

[248] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social Networks*, vol. 21, pp. 375–395, Oct. 2000.

[249] M. Rubinov, R. J. F. Ypma, C. Watson, and E. T. Bullmore, "Wiring cost and topological participation of the mouse brain connectome," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 10032–10037, Aug. 2015.

[250] S. Kintali, "Betweenness Centrality : Algorithms and Lower Bounds," *arXiv:0809.1906 [cs]*, Oct. 2008.

[251] U. Brandes, "A faster algorithm for betweenness centrality," *The Journal of Mathematical Sociology*, vol. 25, pp. 163–177, June 2001.

[252] R. Guimerà and L. A. Nunes Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, pp. 895–900, Feb. 2005.

[253] E. Ruiz Vargas and L. M. Wahl, "The gateway coefficient: a novel metric for identifying critical connections in modular networks," *The European Physical Journal B*, vol. 87, p. 161, July 2014.

[254] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[255] M. Rubinov and O. Sporns, "Weight-conserving characterization of complex functional brain networks," *NeuroImage*, vol. 56, pp. 2068–2079, June 2011.

[256] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns, "Network structure of cerebral cortex shapes functional connectivity on multiple time scales," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 10240–10245, June 2007.

[257] O. Sporns and J. D. Zwi, "The small world of the cerebral cortex," *Neuroinformatics*, vol. 2, pp. 145–162, June 2004.

[258] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, Oct. 2008.

[259] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, p. 016110, July 2006.

[260] P. Ronhovde and Z. Nussinov, "Multiresolution community detection for megascale networks by information-based replica correlations," *Physical Review E*, vol. 80, p. 016109, July 2009.

[261] Y. Sun, B. Danila, K. Josić, and K. E. Bassler, "Improved community structure detection using a modified fine-tuning strategy," *EPL (Europhysics Letters)*, vol. 86, p. 28004, Apr. 2009.

[262] E. Estrada and J. A. Rodríguez-Velázquez, "Subgraph centrality in complex networks," *Physical Review E*, vol. 71, p. 056103, May 2005.

[263] E. Estrada and D. J. Higham, "Network Properties Revealed through Matrix Functions," *SIAM Review*, vol. 52, pp. 696–714, Jan. 2010.

[264] I. T. Jolliffe, "Principal Components in Regression Analysis," in *Principal Component Analysis* (I. T. Jolliffe, ed.), Springer Series in Statistics, pp. 129–155, New York, NY: Springer, 1986.

[265] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[266] G. Strang, *Linear Algebra and Its Applications*. Thomson Brooks/Cole, 4 ed., 2006.

[267] P. W. Sweeney, A. d'Esposito, S. Walker-Samuel, and R. J. Shipley, "Modelling the transport of fluid through heterogeneous, whole tumours in silico," *PLOS Computational Biology*, vol. 15, p. e1006751, June 2019.

[268] Thermo Fisher Scientific, "Amira," 2020.

[269] B. J. Stolz, J. Kaeppler, B. Markelc, F. Braun, F. Lipsmeier, R. J. Muschel, H. M. Byrne, and H. A. Harrington, "Multiscale topology characterizes dynamic tumor vascular networks," *Science Advances*, vol. 8, p. eabm2456, June 2022.

[270] P. Bendich, J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer, "Persistent homology analysis of brain artery trees," *Annals of Applied Statistics*, vol. 10, no. 1, pp. 198–218, 2016.

[271] R. T. Armstrong, J. E. McClure, V. Robins, Z. Liu, C. H. Arns, S. Schlüter, and S. Berg, "Porous Media Characterization Using Minkowski Functionals: Theories, Applications and Future Directions," *Transport in Porous Media*, vol. 130, pp. 305–335, Oct. 2019.

[272] K. Hu, M. Ziehmer, K. Wang, and E. T. Lilleodden, "Nanoporous gold: 3D structural analyses of representative volumes and their implications on scaling relations of mechanical behaviour," *Philosophical Magazine*, vol. 96, pp. 3322–3335, Dec. 2016.

[273] L. T. Choong, P. Yi, and G. C. Rutledge, "Three-dimensional imaging of electrospun fiber mats using confocal laser scanning microscopy and digital image analysis," *Journal of Materials Science*, vol. 50, pp. 3014–3030, Apr. 2015.

[274] Z. Jiang, K. Wu, G. D. Couples, and J. Ma, "The Impact of Pore Size and Pore Connectivity on Single-Phase Fluid Flow in Porous Media," *Advanced Engineering Materials*, vol. 13, no. 3, pp. 208–215, 2011.

[275] P. Levitz, V. Tariel, M. Stampanoni, and E. Gallucci, "Topology of evolving pore networks," *The European Physical Journal Applied Physics*, vol. 60, p. 24202, Nov. 2012.

[276] G. Beltramo, R. Andreeva, Y. Giarratano, M. O. Bernabeu, R. Sarkar, and P. Skraba, "Euler Characteristic Surfaces," *arXiv:2102.08260 [cs, math]*, Feb. 2021.

[277] B. Stolz-Pretzer, *Global and local persistent homology for the shape and classification of biological data*. Ph.D., University of Oxford, Jan. 2019.

[278] Britannica, "Space." Accessed 2023-08-15 10:23:21.

[279] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, pp. 509–512, Oct. 1999.

[280] M. B. Paradowski and L. Jonak, "Diffusion of Linguistic Innovation as Social Coordination," *Psychology of Language and Communication*, vol. 16, pp. 131–142, Dec. 2012.

[281] A. Kölzsch, E. Kleyheeg, H. Kruckenberg, M. Kaatz, and B. Blasius, "A periodic Markov model to formalize animal migration on a network," *Royal Society Open Science*, vol. 5, p. 180438, June 2018.

[282] K. Fang, B. Sivakumar, and F. M. Woldemeskel, "Complex networks, community structure, and catchment classification in a large-scale river basin," *Journal of Hydrology*, vol. 545, pp. 478–493, Feb. 2017.

[283] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 3747–3752, Mar. 2004.

[284] P. Fernández and R. V. Solé, "Neutral fitness landscapes in signalling networks," *Journal of The Royal Society Interface*, vol. 4, pp. 41–47, Aug. 2006.

[285] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '14, (New York, NY, USA), pp. 701–710, Association for Computing Machinery, Aug. 2014.

[286] Aditya Grover and Jure Leskovec, "node2vec: Scalable Feature Learning for Networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, Aug. 2016.

[287] C. Hou, S. He, and K. Tang, "RoSANE: Robust and scalable attributed network embedding for sparse networks," *Neurocomputing*, vol. 409, pp. 231–243, Oct. 2020.

[288] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," *Science*, vol. 298, pp. 824–827, Oct. 2002.

[289] P. Erdős and A. Rényi, "On the Evolution of Random Graphs," *Public*, vol. 5, no. 1, pp. 17–60, 1960.

[290] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of Evolved and Designed Networks," *Science*, vol. 303, pp. 1538–1542, Mar. 2004.

[291] N. Rodgers, P. Tiňo, and S. Johnson, "Network hierarchy and pattern recovery in directed sparse Hopfield networks," *Physical Review E*, vol. 105, p. 064304, June 2022.

[292] N. Rodgers, P. Tiňo, and S. Johnson, "Influence and influenceability: global directionality in directed complex networks," *Royal Society Open Science*, vol. 10, p. 221380, Aug. 2023.

[293] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Physical Review E*, vol. 64, p. 026118, July 2001.

[294] N. Rodgers, P. Tiňo, and S. Johnson, "Strong connectivity in real directed networks," *Proceedings of the National Academy of Sciences*, vol. 120, p. e2215752120, Mar. 2023.

[295] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, pp. 98–101, May 2008.

[296] Johann Benedikt Listing, *Vorstudien zur Topologie*. Vandenhoeck und Ruprecht, 1848.

[297] M. R. Muldoon, R. S. MacKay, J. P. Huke, and D. S. Broomhead, "Topology from time series," *Physica D: Nonlinear Phenomena*, vol. 65, no. 1-2, pp. 1–16, 1993.

[298] T. Y. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," *Communications of the ACM*, vol. 27, pp. 236–239, Mar. 1984.

[299] S. Barannikov, "The Framed Morse complex and its invariants," *Advances in Soviet Mathematics*, vol. 21, pp. 93–116, Apr. 1994.

[300] H. Poincaré, "Nature Du Raisonnement Mathématique," *Revue de Métaphysique et de Morale*, vol. 2, no. 4, pp. 371–384, 1894.

[301] H. Edelsbrunner and J. L. Harer, *Computational Topology: An Introduction*. American Mathematical Society, Jan. 2022.

[302] F. Chazal and B. Michel, "An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists," *Frontiers in Artificial Intelligence*, vol. 4, p. 667963, Sept. 2021.

[303] R. Ghrist, "Barcodes: The persistent topology of data," *Bulletin of the American Mathematical Society*, vol. 45, no. 1, pp. 61–75, 2008.

[304] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification," *Discrete and Computational Geometry*, vol. 28, no. 4, pp. 511–533, 2002.

[305] P. Bubenik, "Statistical topological data analysis using persistence landscapes," Jan. 2015.

[306] F. Chazal, B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman, "Subsampling Methods for Persistent Homology," June 2014.

[307] F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman, "Stochastic Convergence of Persistence Landscapes and Silhouettes," in *Proceedings of the thirtieth annual symposium on Computational geometry*, SOCG'14, (New York, NY, USA), pp. 474–483, Association for Computing Machinery, June 2014.

[308] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec, "The Gudhi Library: Simplicial Complexes and Persistent Homology," in *Mathematical Software – ICMS 2014* (H. Hong and C. Yap, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 167–174, Springer, 2014.

[309] D. Morozov, "Dionysus 2," Sept. 2023. Accessed 2023-11-04 12:50:54.

[310] U. Bauer, M. Kerber, and J. Reininghaus, "PHAT (Persistent Homology Algorithm Toolbox), v1.4.0," July 2023.

[311] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh, "Confidence sets for persistence diagrams," *The Annals of Statistics*, vol. 42, pp. 2301–2339, Dec. 2014.

[312] R. K. Peet, "The Measurement of Species Diversity," *Annual Review of Ecology and Systematics*, vol. 5, pp. 285–307, 1974.

[313] R. Gonzalès and L. Parrott, "Network Theory in the Assessment of the Sustainability of Social–Ecological Systems," *Geography Compass*, vol. 6, no. 2, pp. 76–88, 2012.

[314] A. Kashyap, M. A. Rapsomaniki, V. Barros, A. Fomitcheva-Khartchenko, A. L. Martinelli, A. F. Rodriguez, M. Gabrani, M. Rosen-Zvi, and G. Kaigala, "Quantification of tumor heterogeneity: from data acquisition to metric generation," *Trends in Biotechnology*, vol. 40, pp. 647–676, Nov. 2021.

[315] J. F. Graf and M. I. Zavodszky, "Characterizing the heterogeneity of tumor tissues from spatially resolved molecular measures," *PLOS ONE*, vol. 12, p. e0188878, Nov. 2017.

[316] M. O. Hill, "Diversity and Evenness: A Unifying Notation and Its Consequences," *Ecology*, vol. 54, no. 2, pp. 427–432, 1973.

[317] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, vol. 4, pp. 547–562, University of California Press, 1961.

[318] K. Joshi, M. R. de Massy, M. Ismail, J. L. Reading, I. Uddin, A. Woolston, E. Hatipoglu, T. Oakes, R. Rosenthal, T. Peacock, T. Ronel, M. Noursadeghi, V. Turati, A. J. S. Furness, A. Georgiou, Y. N. S. Wong, A. Ben Aissa, M. W. Sunderland, M. Jamal-Hanjani, S. Veeriah, N. J. Birkbak, G. A. Wilson, C. T. Hiley, E. Ghorani, J. A. Guerra-Assunção,

J. Herrero, T. Enver, S. R. Hadrup, A. Hackshaw, K. S. Peggs, N. McGranahan, C. Swanton, S. A. Quezada, and B. Chain, "Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer," *Nature Medicine*, vol. 25, pp. 1549–1559, Oct. 2019.

[319] C. R. Rao, "Diversity and dissimilarity coefficients: A unified approach," *Theoretical Population Biology*, vol. 21, pp. 24–43, Feb. 1982.

[320] P. Diggle, *Statistical analysis of spatial point patterns / Peter J. Diggle.* Mathematics in biology, London: Academic, 1983.

[321] M. N. M. van Lieshout and A. J. Baddeley, "A nonparametric measure of spatial interaction in point patterns," *Statistica Neerlandica*, vol. 50, no. 3, pp. 344–361, 1996.

[322] T. Bedford and J. V. D. Berg, "A Remark on the Van Lieshout and Baddeley J-Function for Point Processes," *Advances in Applied Probability*, vol. 29, pp. 19–25, Mar. 1997.

[323] J. Besag, "Contribution to the discussion on dr ripley's paper," *JR Stat Soc B*, vol. 39, pp. 193–195, 1977.

[324] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, Mar. 1982.

[325] M. R. Brito, E. L. Chávez, A. J. Quiroz, and J. E. Yukich, "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection," *Statistics & Probability Letters*, vol. 35, pp. 33–42, Aug. 1997.

[326] J. M. Geusebroek, A. W. M. Smeulders, and F. Cornelissen, "Segmentation of Cell Clusters by Nearest Neighbour Graphs," in *Proceedings of the third annual conference of the Advanced School for Computing and Imaging*, Sept. 2002.

[327] D. M. Owen, C. Rentero, J. Rossy, A. Magenau, D. Williamson, M. Rodriguez, and K. Gaus, "PALM imaging and cluster analysis of protein heterogeneity at the cell surface," *Journal of Biophotonics*, vol. 3, no. 7, pp. 446–454, 2010.

[328] F. Levet, E. Hosy, A. Kechkar, C. Butler, A. Beghin, D. Choquet, and J.-B. Sibarita, "SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data," *Nature Methods*, vol. 12, pp. 1065–1071, Nov. 2015.

[329] L. Andronov, I. Orlov, Y. Lutz, J.-L. Vonesch, and B. P. Klaholz, "ClusterViSu, a method for clustering of protein complexes by Voronoi tessellation in super-resolution microscopy," *Scientific Reports*, vol. 6, p. 24084, Apr. 2016.

[330] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications," *Data Mining and Knowledge Discovery*, vol. 2, pp. 169–194, June 1998.

[331] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 10, pp. 5:1–5:51, July 2015.

[332] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[333] G. Biau, F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodríguez, "A weighted k-nearest neighbor density estimate for geometric inference," *Electronic Journal of Statistics*, vol. 5, pp. 204–237, Jan. 2011.

[334] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, Dec. 2007.