Spatial Modelling of the Epithelial-to-Mesenchymal Transition in Cancer Using Geostatistical and Machine Learning Approaches

Eloise Withnell

Department of Genetics, Evolution and Environment, Division of Biosciences
University College London

Thesis submitted for the degree of Doctor of Philosophy (PhD) at
University College London

Supervisor: Dr Maria Secrier
February 2025

Declaration of Ownership

I, Eloise Withnell, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Impact statement

The epithelial-to-mesenchymal transition (EMT) is a multi-stage cellular program central to metastasis and therapeutic resistance. Despite a well-documented link between EMT and worse clinical outcomes, there remains a need to understand precisely how different EMT states are influenced by both intrinsic genomic alterations and tumour microenvironment (TME) cues.

Spatial transcriptomics has recently emerged as a powerful technology that can transform our understanding of how the TME influences key cellular programs, such as EMT. However, many approaches used to analyse spatial transcriptomics ignore important aspects of spatial data and struggle to capture the spatial relationships of cell states that lie along a continuum, highlighting the need for more flexible and robust analytical frameworks to be developed.

In this work, I have developed and applied novel geostatistical and machine learning methods, including the SpottedPy Python package, to comprehensively profile and quantify EMT across multiple biological scales. I reveal how distinct EMT states, including epithelial, hybrid, and mesenchymal phenotypes, respond differentially to TME cells such as CAFs and macrophages, and other TME processes such as hypoxia and angiogenesis.

Additionally, I introduce a new approach to quantify cell plasticity. By integrating genomic events and TME information within graph neural networks and geographically weighted regression models, I quantify the strength of the TME effect on epithelial to mesenchymal plasticity (EMP) and highlight its dominant role in EMP. I uncover heterogeneous spatial relationships between EMT states and show how intermediate phenotypes express varying degrees of plasticity.

Given the importance of EMT to cancer development, these methodologies offer valuable tools for researchers investigating the spatial dynamics of EMT. Not only do they provide a framework for quantitatively assessing EMT across tumours, but they can also be adapted to study other forms of cell plasticity. By deepening our understanding of how cancer cells traverse the EMT continuum, this work opens new

possibilities for therapeutic interventions aimed at curbing metastasis and overcoming drug resistance through the precise manipulation of microenvironmental factors.

The research output presented in this thesis has been disseminated to the scientific community through the following publications:

- Withnell, E. & Secrier, M. *SpottedPy quantifies relationships between spatial transcriptomic hotspots and uncovers environmental cues of epithelial-mesenchymal plasticity in breast cancer.* Genome Biology 25, 289 (2024).
- Malagoli Tagliazucchi, G., Wiecek, A. J., Withnell, E. & Secrier, M. *Genomic and microenvironmental heterogeneity shaping epithelial-to-mesenchymal trajectories in cancer.* Nat Commun 14, 1-20 (2023).

Additionally, whilst not described in this thesis, the methods used to assess spatial relationships at scale have also been made available to the scientific community in the follow manuscripts:

- Pan, S., Withnell, E. & Secrier, M. *Classifying Epithelial-Mesenchymal Transition States in Single Cell Cancer Data Using Large Language Models.* bioRxiv (2024).
- Cenk, C., Withnell, E., Pan, S., Chu, T., Labbadia, J. & Secrier, M. *Balancing tumour proliferation and sustained cell cycle arrest through proteostasis remodelling drives immune niche compartmentalisation in breast cancer.* bioRxiv (2024).

The cell plasticity prediction framework is in the process of being submitted for publication.

Contents

Acronyms

- **AIC -** Akaike Information Criterion
- **AUC -** Area Under the Curve
- **BCC -** Basal Cell Carcinoma
- **BIC -** Bayesian Information Criterion
- **CAF -** Cancer-Associated Fibroblast
- **CNV -** Copy Number Variation
- **CRC -** Colorectal Cancer
- **DC -** Dendritic Cell
- **DCIS -** Ductal Carcinoma In Situ
- **EM2, EM3 -** Intermediate Epithelial-to-Mesenchymal Transition States
- **EMP -** Epithelial to Mesenchymal Plasticity
- **EMT -** Epithelial to Mesenchymal Transition
- **ENM -** Ecological Niche Model
- **EPI -** Epithelial (EMT state)
- **ER -** Estrogen Receptor
- **GAT -** Graph Attention Network
- **GEE -** Generalised Estimating Equations
- **GNN -** Graph Neural Network
- **GWR -** Geographically Weighted Regression
- **hEMT -** Hybrid EMT state
- **HER2** - Human Epidermal Growth Factor Receptor 2
- **HIF -** Hypoxia-Inducible Factor
- **HMM -** Hidden Markov Model
- **IDC -** Invasive Ductal Carcinoma
- **iCAF -** Inflammatory Cancer-Associated Fibroblast
- **M1 -** Quasi-Mesenchymal EMT State (late intermediate)
- **M2 -** Fully Mesenchymal EMT State
- **MAUP -** Modifiable Areal Unit Problem
- **MES -** Mesenchymal (EMT state)
- **MGWR / msGWR -** Multiscale Geographically Weighted Regression
- **MSC -** Mesenchymal Stem Cell

  MSE - Mean Squared Error

- **myCAF -** Myofibroblastic Cancer-Associated Fibroblast
- **NK -** Natural Killer cell
- **NKT -** Natural Killer T cell
- **NMF -** Non-Negative Matrix Factorization
- **OLS -** Ordinary Least Squares
- **PC -** Principal Component
- **PCA -** Principal Component Analysis
- **PDAC -** Pancreatic Ductal Adenocarcinoma
- **PMN -** Polymorphonuclear Neutrophil
- **PR -** Progesterone Receptor
- **PVL -** Perivascular-Like (cell)
- **R^2 -** Coefficient of Determination
- **ROC -** Receiver Operating Characteristic
- **SAR -** Spatial Autoregressive Model
- **scRNA-seq -** Single-Cell RNA Sequencing
- **SDM -** Species Distribution Model
- **SEM -** Spatial Error Model
- **ST2K -** Spatial Transcriptomics 2000 (platform/array)
- **TCGA -** The Cancer Genome Atlas
- **TGFB** – Transforming growth factor beta
- **TNBC** - Triple-negative breast cancer
- **TME -** Tumour Microenvironment

Abstract

The epithelial-to-mesenchymal transition (EMT) is an important cellular process involved in tumour progression, metastasis, and therapy resistance. However, the influence of the tumour microenvironment (TME) and genomic factors on EMT, and the discrete states within this transition, remains incompletely understood. In this thesis, I develop geostatistical and machine learning methods to analyse spatial transcriptomic data, to understand the spatial relationships of cancer cells undergoing EMT.

I present a novel Python package, SpottedPy, which can identify spatial hotspots of gene signatures and cell types and assess their spatial interactions with other hotspots. Using this approach, I identified EMT niches associated with angiogenic and hypoxic regions, surrounded by CAFs and macrophages. EMT hybrid and mesenchymal hotspots followed transformation gradients, becoming increasingly immunosuppressed. Importantly, SpottedPy is a flexible package, which enables users to explore spatial relationships at different scales, from immediate neighbours to larger tissue modules, allowing for new insights into the tumour microenvironment.

Building on these spatial insights, I develop a graph neural network and geographically weighted regression framework to quantify the relative contributions of intrinsic genomic changes and extrinsic microenvironmental signals on cell plasticity programmes. The approach strengthens the evidence that targeting the TME is more important for targeting EMT as opposed to targeting genomic factors. It highlights the importance of the TME in inducing both subtle, short-term changes and stable, long-term phenotypic change, whereas genomic alterations primarily contribute to more stable, long-term changes. I showed that the mesenchymal phenotype is more deterministic, while hybrid states are less predictable and thus potentially more plastic. Additionally, I found that relationships between EMT states and particular TME populations do vary across different tissue regions, notably with myoepithelial cells.

Overall, my work provides an in-depth molecular and spatial characterisation of EMP, while highlighting novel methodological approaches for capturing and measuring cell plasticity. These insights could help inform therapeutic approaches that target the genetic and microenvironmental factors linked to cancer cell plasticity.

List of research articles

Withnell, E. & Secrier, M. SpottedPy quantifies relationships between spatial transcriptomic hotspots and uncovers environmental cues of epithelial-mesenchymal plasticity in breast cancer. *Genome Biology* **25**, 289 (2024).

Malagoli Tagliazucchi, G., Wiecek, A. J., Withnell, E. & Secrier, M. Genomic and microenvironmental heterogeneity shaping epithelial-to-mesenchymal trajectories in cancer. *Nat Commun* **14**, 1–20 (2023).

Pan, S., Withnell, E., & Secrier, M. Classifying Epithelial-Mesenchymal Transition States in Single Cell Cancer Data Using Large Language Models. *bioRxiv*. (2024)

Cenk, C., Withnell, E., Pan, S., Chu, T., Labbadia, J., & Secrier, M. Balancing tumour proliferation and sustained cell cycle arrest through proteostasis remodelling drives immune niche compartmentalisation in breast cancer. *bioRxiv*. (2024)

1 Chapter 1: Introduction

1.1 The epithelial-to-mesenchymal transition and its varied roles in normal development and cancer

The epithelial-to-mesenchymal transition (EMT) is a multi-stage cellular process in which cells lose their epithelial features and acquire mesenchymal properties, involving the disruption of cell-cell adhesion and cellular polarity[1]. EMT is linked to gene expression changes and post-translational regulation that enables the cells to gain mesenchymal traits. During EMT, epithelial markers are downregulated, notably E-cadherin, and mesenchymal markers are upregulated, with vimentin, N-cadherin and fibronectin amongst the most well characterised. A mesenchymal cell typically gains migratory abilities due to the presence of actin stress fibres, and cell-matrix adhesion remodelling[2] (**Figure 1**). The reverse process, the mesenchymal-to-epithelial transition (MET), can also occur where cells regain their epithelial traits[3].

EMT was first identified by researchers studying embryogenesis in the late 1960s and is now widely observed in embryonic development and wound healing[4]. During animal development, cells are required to migrate large distances, which EMT enables. In most cases, the cell then reverses back to an epithelial state, through MET[5]. Researchers noticed parallels between embryonic development and tumour progression, observing that morphological changes in carcinoma cells resembled EMT[6]. Studies in the 1990s provided the first experimental evidence for this link, by showing that EMT inducers such as leukocyte medium, including transforming growth factor (TGFBeta)[7] and fibroblast growth factor (FGF)[8], increase the invasiveness of cancer cell lines. Further studies also showed that Ras-transformed mammary epithelial cells induced a mesenchymal-like phenotype[9], indicating that EMT could be a mechanism driving tumour progression. By the early 2000s, molecular studies identified key transcription factors such as Snail[10], Slug[8], Twist[11], and Zeb1/Zeb2[12] as master regulators of EMT in cancer, drawing direct mechanistic parallels with key transcription factors in embryonic processes.

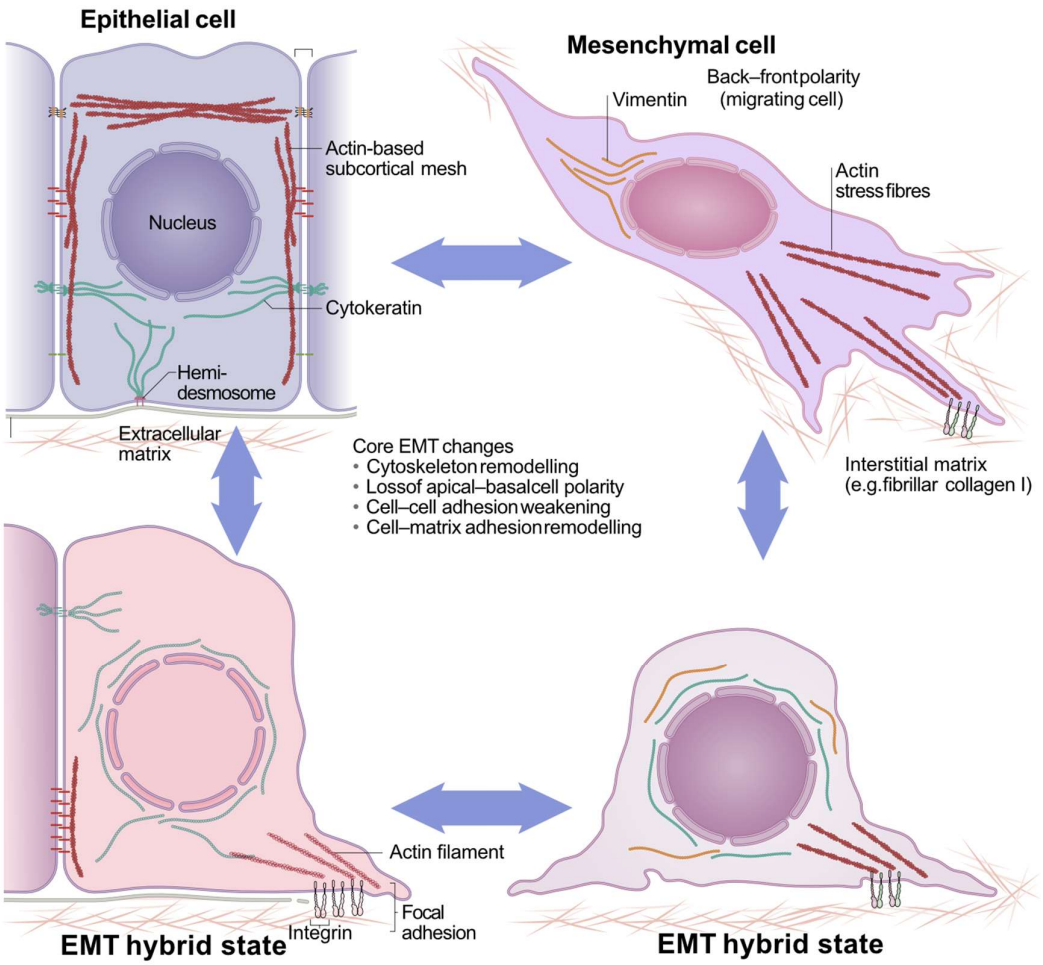

Figure 1 The key molecular and morphological changes during transitions between the EMT states. Figure adapted from Yang et al. (2020)[1].

The development of transgenic mouse models further confirmed EMT roles in tumour metastasis[13]. Mouse models were developed, such as the MMTV-PyMT model of breast cancer, which can mimic the stages of breast cancer that patients can progress through, including metastatic phases[14]. Several research groups showed that cells involved in metastasis in these models displayed EMT markers[15,16]. Later, lineage-tracing experiments allowed for more targeted tracking of metastatic cells by identifying individual cells and tracking their individual expression of markers[17]. For example, it is possible to identify the cells that express epithelial markers and track which cells then adopt mesenchymal-like phenotypes[18]. An influential paper by Li et al. tracked EMT in lung metastasis of breast cancer and identified key EMT markers

in metastasis-initiating cells[19]. The introduction of single-cell technologies has further allowed larger scale characterisation of gene expression changes to define EMT more broadly across the genome in mouse models, rather than focusing on a few key markers[20].

As growing evidence was accumulating linking EMT to metastatic properties of cancer cells, it was of growing importance to understand whether EMT was necessary for metastasis, or a correlated feature. In a landmark paper by Zheng et al, it was shown key EMT-TFs knockouts did not affect tumour progression or metastasis in a pancreatic cancer model[21]. However, it did affect chemosensitivity, which suggests why patients with more mesenchymal tumours have poorer outcomes[22]. However, it has since been shown that it is necessary depending on model system, tumour tissue and disease progression being studied[23]. It is important to note, most knock out studies have focused on a handful of EMT markers, predominantly E-cadherin, vimentin, N-cadherin, Snai1, Zeb1 and fibronectin, and there is an emphasis in the field to include a larger range of gene markers in future studies[1]. Just focusing on these handful of markers may oversimplify the complexity of EMT. Many of the EMT transcription factors used in the knockouts can function in overlapping ways complicating the assessment of individual EMT-TFs impact on EMT[24]. Many tumours exhibit context-dependent EMT signatures (**Figure 2**) involving a wide range of transcription factors, extracellular matrix components, signalling molecules and metabolic reprogramming that contribute to EMT-associated phenotypes[25]. It is therefore difficult to determine whether EMT is dispensable in a given experimental system or if there is enough redundancy in the system that means other factors can compensate for its loss.

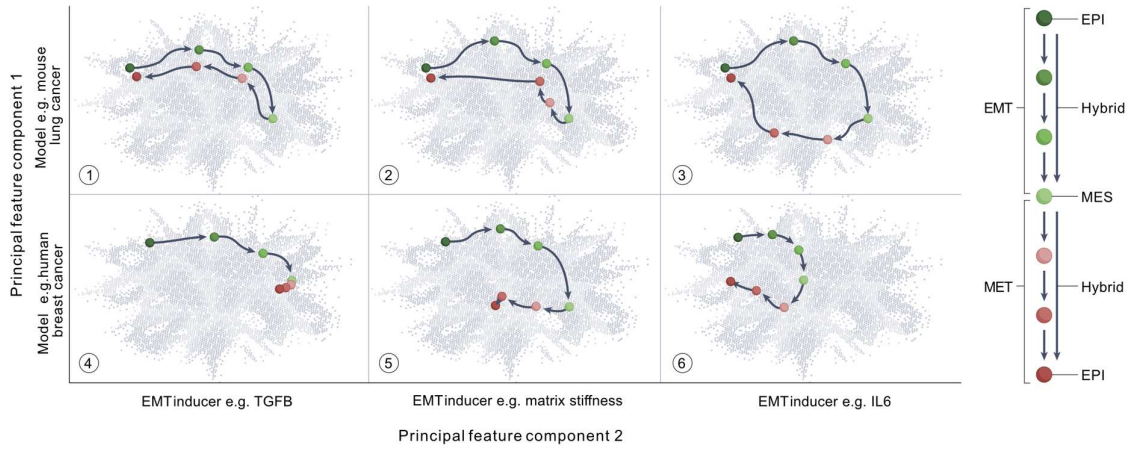

Figure 2 The various paths that epithelial tumour cells can transition through when undergoing EMT. Multiple stable states exist which can be vastly different to states within this transition found in other cancer types, model systems and under different EMT inducers. Figure adapted from Haerinck et al. (2023)[26]

In addition to metastasis, EMT has been linked to chemoresistance in several experimental contexts, including cell line experiments, mouse models, and patient studies. For example, cells undergoing EMT in response to TGF-β signalling in breast cancer cell lines become chemoresistant to drugs such as etoposide and paclitaxel[27]. EMT has been shown to increase the expression of drug efflux pumps such as ABC transporters, which in turn increase chemoresistance[28]. TGF-β stimulation in ovarian cancer cell lines including HO8910 and SKOV3 increases resistance to cisplatin treatment, a platinum-based chemotherapy commonly used to treat ovarian cancer[29]. Alongside cell line experiments, in mouse models including breast, prostate and lung cancer, EMT has been linked to a range of chemoresistance[30,31]. Additionally, similar results have been observed at the patient level using bulk transcriptomics[32,33]. For example, in NSCLC patients, a higher expression of EMT markers in tumour samples was correlated with resistance to platinum-based chemotherapy[34]. Furthermore, studies in ovarian cancer patients treated with cisplatin have also shown that the presence of EMT markers in tumour biopsies is associated with worse treatment outcomes[35].

1.2 EMT states and links with cell plasticity

Historically, EMT has been considered a binary process, with an epithelial state displaying the marker E-cadherin, and a mesenchymal state consisting of the loss of E-cadherin marker and the gain of vimentin[9]. However, this has been challenged in

recent years, with multiple states being observed along an EMT axis[36]. The states in between a fully epithelial and mesenchymal state are known as hybrid (hEMT) or partial EMT states (pEMT) (**Figure 1**). These states have been linked to different properties[37]. The hybrid state has been associated with increased invasive and migration properties[22]. This is likely due to the advantage of having both the mesenchymal invasive properties, in addition to useful epithelial cell characteristics such as adhesion to neighbouring cells. Fully developed mesenchymal cells are less likely to revert to an epithelial state which is necessary if the cell is to form a distinct colony[38]. The number of EMT states is unknown currently, with researchers debating as to whether it should be treated in a continuous or discrete phenotype[39]. Brown *et al.* (2022) identified six stable states in breast cancer, and showed that all states had variations in migration and invasion traits, with intermediate states EM2 and EM3 scoring the highest in mouse models.

Attempts have been made to develop mathematical models to quantify EMT dynamics between different states. One study used Markov models to understand how microstates and macrostates shape EMT transitions, emphasising the non-linear effects of intermediate states[40]. These findings concluded that destabilising intermediate states could be a potential therapeutic strategy to mitigate metastasis. Other attempts have used ordinary differential equations, such as through population growth models, to model the intrinsic growth rates for epithelial and mesenchymal cells and understand how EMT states related to EMT heterogeneity[41]. Using these models, the analysis highlighted the importance of considering both intrinsic cell plasticity and population-level interactions to gain a full understanding of EMT states and tumour heterogeneity. The findings demonstrate that epithelial and mesenchymal subpopulations were able to influence each other's growth through either cooperative or suppressive effects. Importantly, the models with the best fit accounted for cell-state transitions and population density-dependent growth.

There is great diversity in definitions for EMT, predominantly due to the varying phenotypic manifestations across model systems. EMT is highly context dependent, with different gene programs depending on factors such as the model system analysed, stimuli used to promote EMT or the tissue in the body[26]. For example Puram et al. showed that partial EMT signatures are different between tumours[42], and Peixoto

et al. showed that only around 10-30% of differentially expressed genes are shared in EMT responses in cell line microarray data[43]. Cook et al. presented similar findings, and showed that on average only 22% of genes are shared between EMT inducers in cell line studies[39]. In 2020, a group effort led by 'the EMT International Association' defined key terms in the field and key areas of research to focus on in the future[1]. The main recommendation on the criteria to define EMT was that "EMT cannot be assessed on the basis of one or a small number of molecular markers"[1]. In addition, they suggested defining EMT status based on changes in cellular properties. They recommend EMT to be explored beyond a traditional cell and cancer biology approach, with a focus on collaborations with systems biologists, biophysicists and mathematical modellers. Key unanswered questions also highlighted in the report included the functional implications of EMT heterogeneity and understanding the dynamic switch between E/M states in response to distinct cues from the microenvironment.

Recently, there has been a shift in focus towards reframing epithelial-mesenchymal transition as epithelial-mesenchymal plasticity (EMP), emphasising its dynamic and reversible nature, rather than viewing it solely as a unidirectional transition[26]. This builds on the view of understanding cell states in terms of the Waddington landscape, where cell states are represented as valleys in a multidimensional landscape, and transitions between states are depicted as movements across this terrain[26] (**Figure 3).** In this framework, cell can shift between epithelial, mesenchymal, and hybrid states in response to intrinsic and extrinsic cues, with certain states easier to transition into other states than others based on intrinsic properties of the state.

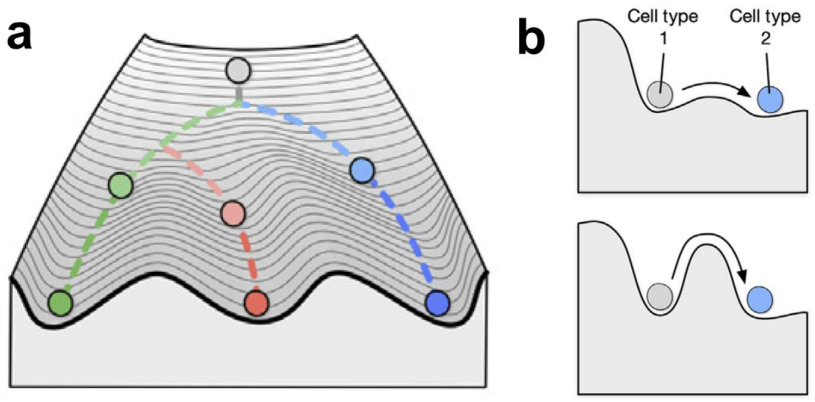

Figure 3 The Waddington landscape. The Waddington landscape represents the ease that cells transition between different cell states or types. **a.** Overview of The Waddington's landscape depicting a pluripotent cell taking different paths. **b.** Some cell state transitions are easier to traverse when the barriers between states are lower, representing a shallower hill in Waddington's landscape, whereas other transitions are more difficult to reach, representing a steeper hill. Figure adapted from Qin *et. al* (2024)[44].

1.3 Capturing EMT

Traditionally, EMT has been identified by examining the expression of a handful of well-established markers, for example, the downregulation of epithelial markers (e.g., E-cadherin) and the upregulation of mesenchymal markers (e.g., vimentin, N-cadherin)[45,46]. While these marker-based methods are straightforward, they tend to oversimplify the complex and continuum of cell states associated with EMT. Unlike discrete cell types, EMT states exist along a spectrum, necessitating methods that can capture this gradual progression[26]. scRNA-seq has enabled more sophisticated methods to capture the full spectrum of cell states. For example, scoring scRNA-seq using enrichment-based approaches, such as gene set enrichment analysis (GSEA) to score gene sets, provide quantitative measures of state transitions[47]. Additionally, extracting distinct gene modules present in the data through dimensionality reduction, such as non-negative Matrix Factorization (NMF), as used in ProjectR, and principal component analysis (PCA) are often used to assess cell states[48,49]. Other approaches involve clustering cells based on scores from continuous signatures, such as Gaussian mixture models[50]. Pseudotime approaches using tools like Monocle and Slingshot have also emerged as important tools to order cells along dynamic trajectories within scRNA-seq data, revealing the temporal progression of EMT over time[51,52].

Recent tools have expanded on these methods further. For example, CELLSTATES, built on the clustering approach, and developed an approach to capture cell states at the statistically maximum resolution[53]. This method partitioned cells based on statistically indistinguishable gene expression states, accounting for biological noise present in the data. SeaCells is another approach, that works by grouping cells into meta-cells that share similar gene expression profiles while minimizing information loss using archetypal analysis[54].

It is important to note that transcriptomic scoring of EMT can be confounded by CAF signatures as they express many of the same markers that cancer cells undergoing EMT express[55]. Approaches, such as using scRNA-seq whole transcriptome reference mapping are important to ensure the targeting of specific tumour-related EMT signatures.

Despite these advances, no single metric or method for defining a cell state or cell type is universally accepted, highlighting the need for unified definitions and standards in the field[53].

1.4 EMT and epigenetics

There is growing evidence to suggest that TME can influence the EMT state of a cancer cell through epigenetic reprogramming, in addition to the transient effect of influencing gene expression through signalling pathways[56]. Signals such as cytokines, hypoxia, and ECM interactions can act through a non-genetic, reversible route via chromatin remodelling, DNA methylation, and non-coding RNAs[56]. For example, hypoxia-induced histone modifications such as trimethylation of histone H3 at lysine 4 (H3K4me3) have been identified at promoters of EMT transcription factors such as TWIST1[57]. TGF-β has been shown to induce EMT epigenetically by the demethylation of H3K27me3 in the Snail1 promoter[58]. Aside from histone modifications, DNA methyltransferases catalyse the transfer of methyl groups to cytosine residues, primarily at CpG dinucleotides[59]. This can then lead to the transcriptional repression of epithelial genes[59]. Non-coding RNAs including microRNAs such as the *miR-200* family acts on key EMT genes, such as *ZEB1* and *ZEB2*[59]. Additionally, the non-coding RNAs *MALAT1* and *HOTAIR* have been shown to promote EMT by recruiting chromatin modifiers[60,61]. These epigenetic mechanisms enable cells to transition between epithelial and mesenchymal states dynamically, enhancing the plastic nature

of the programme. Importantly, it also allows the mesenchymal cancer cells to revert back to an epithelial phenotype through the mesenchymal-epithelial transition (MET)[59], enhancing colonisation at metastatic sites.

The lineage specification of the tumour cell can significantly influence the propensity for EMP. Epithelial subtypes, while similar in function, differ in morphology, transcriptional and chromatin landscapes, what can all influence EMP to different extents[26]. There is growing evidence that cancer cells can acquire traits by reactivating dormant developmental programs[26,62] and EMP can be viewed as cells re-traversing developmental paths[26]. Lineage-specification of the tumour cell can occur at the chromatin-level or the functional loss of certain lineage-specific transcription factors genes, such as GATA3, ELF5, FOXA1, or KLF4 which can drive EMP[26]. For example, the loss of GATA3, necessary for luminal epithelial cell commitment[63], can induce EMT in breast cancer.

1.5 Genetic constraints on EMT

Although EMT is not strongly believed to be genetically determined, as it is considered a reversible, process, accumulating evidence suggests that genetic alterations can contribute to EMT. These genetic mechanisms can range from mutations to copy number variations and chromosomal rearrangements and can act on the key EMT transcription factors sites. For example, loss of p53, a well-known mutation that promotes metastasis, can trigger an epigenetic signalling cascade acting on SNAI1, an important EMT transcription factor[64]. ZEB1 amplification can drive EMP in prostate cancer[65]. Loss of chromosome 8p has been linked to increased invasiveness in breast cancer[66] and loss of chromosome 9p and chromosome 14q have been shown to play an important role in metastatic clear-cell renal carcinoma[67]. Additionally, amplification of chromosome 11q1 has been shown to increase the expression of the actin-related protein 2/3 complex, increasing motility and invasion of cancer cells[68].

Chromosomal rearrangements can also act on key EMT regions of the genome. For example, TWIST1 is located in a well-documented unstable region of the genome that often undergoes rearrangements[69]. Gene fusions such as TMPRSS2–ERG has been linked to increased EMT[70]. These genomic mechanisms can also promote a hybrid EMT phenotype. For example, hybrid EMT states have been reported to be due to FAT1 loss, which alters the chromatin state of cells, stabilising both epithelial and

mesenchymal traits[71]. KRAS mutations have also been shown to promote both epithelial and mesenchymal gene expression[72].

1.6 EMT and the tumour microenvironment

There is growing evidence building to suggest that a large number of interactions between tumour cells and the microenvironment occur during EMT[73,74]. The TME consists of the surrounding cells, including immune cells and fibroblasts, the extracellular matrix, signalling molecules, and blood vessels that interact with the tumour. It plays a central role in the growth and invasion of the cancer cells[75]. A key component of EMT in both wound healing and cancer is the remodelling of the microenvironment. In turn, the microenvironment properties, such as the cytokines, ECM, hypoxia and growth factors can also influence the EMT states of the cancer cells. The TME is therefore a likely contributor to the stability and regulation of EMT states[74].

The relationship of regions of the tumour undergoing EMT and different components of the TME have been explored in different experimental systems[39]. Key evidence highlighting the significance of the TME emerged from studies demonstrating the crucial role of TGF-β in promoting EMT in cell lines[76]. This is a cytokine predominantly produced by immune cells and CAFs, key players within the TME[77]. Various experiments culturing cancer cells with cells found within the TME have been shown to promote EMT. For example, co-culturing cancer cell lines with fibroblasts[78,79] and macrophages[80,81] with cancer cells have promoted EMT. CAFs can also induce EMT through their roles remodelling the ECM, increasing the stiffness of the matrix[82,83]. which through mechano-transduction pathways can promote EMT. The links with the TME have been further validated in xenograft models, experimental systems in which human tissue is implanted into immunocompromised mice. These models have confirmed certain relationships observed in cell lines. For example, breast cancer xenografts injected with CAF signalling molecules had increased expression of EMT markers and promoted tumour progression[84]. Pastushenko et al. reported shifts in the composition of stromal cells as tumour cells transitioned to more mesenchymal states in a genetic mouse model of skin squamous cell carcinoma. Cells in close contact with EMT tumour cells showed significantly higher densities of CD45+ immune cells, particularly monocytes and CD68+ macrophages, as well as increased numbers of endothelial and lymphatic cells. Importantly to note this study identified EMT cells

using a specific set of markers, EpCAM, CD106 (VCAM1), and CD51 (ITGAV), and therefore may miss the complexities of different states.

Despite this knowledge, the relationship between regions of the tumour undergoing EMT and different components of the TME is not well established in human tissue samples.

1.7 EMT and druggable targets

A major goal is translating EMT research into the clinic, due to the strong links with metastasis and chemoresistance[85]. While most approaches remain experimental, some have entered clinical trials.

The majority of drugs entering clinical trials are STAT3 inhibitors, targeting a transcription factor in the signalling pathway of EMT[86]. DSP-0337, Danvatirsen and OPB-111077 are some of the inhibitors targeting STAT3 that are in phase I or phase II clinical trials[87,88]. As EMT is tightly controlled at the transcription factor level, targeting the key transcription factors such as Snail, Twist, Slug, Zeb1, and Zeb2 would have been the more intuitive target. However, as these can have overlapping functions, targeting them individually have not been successful in suppressing EMT[89]. Drugs to target key EMT inducers, such as TGF-β and TNF-α have also been developed, most notably using small molecule inhibitors. Avadomide, a small molecule inhibitor had success in human phase I study, and it is now in phase II study for advanced melanoma[90]. Monoclonal antibodies have also been developed to target these inducers, such as NIS-793 which is an anti-TGF-β monoclonal antibody[91]. Small interfering RNA (siRNA) targeting EMT transcription factors have also been a focus, and siRNAs targeting hypoxia and TGF- β signalling pathways have been tested in preclinical settings[92]. Epigenetic modulation has had recent focus, with DNA methyltransferase (DNMT) inhibitors having had success at halting EMT progression in preclinical settings[93].

As the TME plays an important role governing EMT, targeting the different TME components involved in EMT, is an emerging approach. Targeting the TME has had overwhelming success for other key aspects of cancer progression. For example, cancer immunotherapies, such as immune checkpoint inhibitors (ICIs) and chimeric antigen receptor (CAR) T-cell therapies have prolonged survival in a subset of cancer types[94,95]. CD8+ cytotoxic T-cells are critical in providing antitumour immunity, and

tumour cells can often avoid attack by producing immunosuppressive factors, which ICIs can target[95]. Gaining a deeper understanding of how EMT states interact with the TME will therefore be important in advancing this approach. TME components, such as CAFs, can be targeted by a compound that inhibits the secretion of cytokines that promote their activation, such as TGF-β inhibitors or PDGF receptor inhibitors[96]. The extracellular matrix (ECM) can be targeted by degradation enzymes[97]. Immunotherapies, are also being explored to target EMT, particularly to target the EMT states that are linked to immunosuppressive features[98].

A key challenge includes the risk that anti-EMT therapies could lead to an increase in the mesenchymal-epithelial transition and encourage disseminated tumour cells to colonise[5]. Additionally, EMT patient heterogeneity is currently poorly characterised, and therefore, it is currently unknown whether anti-EMT therapy would benefit those with early stage or late stage better, and those with or without certain cell populations[99]. Other challenges in EMT therapeutic developments include the development of screening tools to identify anti-EMT compounds[99]. 2D *in vitro* studies have been the most popular as they offer a high-throughput approach, but are limited by the fact that they cannot model the TME effect on EMT. Therefore, 3D methods, such as tumour-on-a-chip technology, are currently being developed to more successfully capture the TME[85]. An increased understanding of which TME cells are associated with different EMT states in patient tissue can greatly aid in the creation of a synthetic TME for screening purposes.

1.8 Spatial biology

Multiple studies have shown that intra- tumour heterogeneity, the distinct tumour cell populations within the same tumour, can accelerate cancer progression [100,101,102]. It is therefore vital to study spatial structures in cancer to further understand intra-tumour heterogeneity and improve therapeutic response and survival rates[103]. For example, patients can have significantly different responses to immune checkpoint inhibitors depending on the cancer lesion within the patient[104]. Additionally, spatial patterns of immune cells can also hold prognostic value. A recent study analysing the networks of over 1,000 LUAD and LUSC tumours constructed from H&E images showed that the spatial patterns of tumour-infiltrating lymphocytes (TILs) on H&E images were different between subtypes and prognostically relevant[105]. Multiple additional studies that

model tissue as cellular graphs have also identified clinically relevant features[106,107]. Determining the spatial context of EMT is therefore critical for advancing our understanding of how phenotypes develop, progress, and respond to treatment.

Spatial transcriptomics is a recent technique that enables us to profile gene expression across a tissue and is an important technique to assess tumour heterogeneity[108]. Visium, a spatial transcriptomic platform released by 10X Genomics in 2019, enables the whole transcriptome profiling of ~1,000 tissue spots (55 µm spot diameter with 100 µm centre-to-centre distance) with over 10,000 transcripts per spot[109]. Recent developments on this approach have resulted in a newer technique in 2023, Visium HD, which offers higher resolution imaging with smaller tissue spot sizes. Visium HD has much smaller, continuous spots (around 2µm) compared to the larger, spaced-out spots of the original Visium. Xenium, also developed by 10X Genomics, represents an approach offering sub-cellular resolution, but profiles a targeted set of transcripts rather than the whole transcriptome[110]. It is currently a very active research area, and multiple other companies are promising improvements on resolution and throughput in the next few years[21]. Due to the high complexity of these datasets, tailored computational methods are important for inferring conclusions[111]. A large range of methods, including spatial transcriptomic-specific normalisation[112], cellular deconvolution[113 114], and spatially varying gene identification[115 116], have been developed in recent years to address the novel questions emerging from spatial transcriptomics.

Spatial transcriptomics offers a promising avenue to confirm some of the widely established EMT spatial relationships in human tissue. Several studies have shown that EMT is a spatially located gene signature. For example, Takiet *et al*. (2021) showed that epithelial cells at the tumour edge were significantly enriched in pathways related to EMT in primary and metastatic head and neck cancer[117]. Additionally, they showed that pEMT cancer cells located at the leading edge of head and neck tumours were shown to cause invasion by interacting with cancer-associated fibroblasts. EMT-related genes were the most easily predicted in a study that attempted to predict spatial transcriptomics in breast cancer tissue, hinting at the relevance of the spatial context for EMT. Barkley *et al*. (2022) showed cells undergoing EMT correlated with fibroblasts and endothelial cells, whilst being negatively correlated with other malignant cells[118]. This study only looked at a few key EMT markers and cell type

markers, but it suggests the relationship between EMT and microenvironment cells warrants further investigation. Additionally, EMT, hypoxia and inflammation were found as the key explanatory variables for regional variations in pancreatic cancer spatial transcriptomic data[119]. Macrophages have also been linked to EMT in specific niches within breast cancer spatial transcriptomics slides[120].

1.9 Statistical and machine learning methods to analyse spatial data

Spatial transcriptomic technologies such as Visium measure expression from multiple cells within a single spot, and therefore cellular deconvolution to estimate the cellular composition of spatial transcriptomic spots is usually a necessary first step in analysis. Methods like RCTD[121], Stereoscope[122], and Cell2Location[123] have been developed to infer single-cell contributions to these spatial profiles. They use a range of techniques from non-negative matrix factorisation to probabilistic modelling approaches.

A lot of method development has focused on analysing areas within the tissue characterised by specific cellular compositions, referred to as spatial niches. These distinct cellular interactions drive phenotypic behaviours such as differentiation, migration, and response to external stimuli. These spatial niches could include invasive tumour margins, immune-infiltrated regions, or hypoxic zones. To study spatial niches, spatial clustering methods, such as SpaGCN[124], BayesSpace[125], and Giotto[126], are widely used, and have been proven to identify functional regions in tissues. These tools cluster spatial spots based on gene expression profiles and their physical location, identifying distinct regions. Additionally, niche methods can cluster at the cell type label level, such as CellCharter[127].

Many niche methods are therefore focused on detecting niches at the gene expression or cell type level. However, methods to understand the niches of tumour regions enriched for specific gene signatures are less well developed. In the case of EMT, this would be understanding the different cell types associated with regions enriched for different EMT states. Additionally, understanding how these relationships change after altering the parameters used for niche detection are less well established. For instance, adjusting the number of neighbouring cells or spots considered when defining niches, or analysing gene signatures and cell types as broader spatial regions rather than at an individual level, could influence niche characterisation. Moreover, current approaches lack analytical methods to define and compare shorter and longer-

range interactions between specific areas or cell populations of interest. They also do not have extensive functionality to determine the differential spatial relationships given two gene signatures of interest. For example, “which cells are significantly closer to high hypoxic regions compared to low hypoxic regions?”

Another focus of method development involves identifying spatially variable genes (SVGs), which involves genes whose expression levels vary significantly across different spatial regions of a tissue, reflecting genes involved in cell-cell interactions, microenvironmental heterogeneity, or tissue-specific functions. Popular methods include SpatialDE[128], which uses Gaussian process regression, and SPARK, which uses a flexible non-parametric approach[129].

Graph-based abstraction of the gene expression datasets has been an emerging way of representing spatial transcriptomic data for downstream analysis[130]. It has been used for cellular deconvolution techniques [131], inferring gene interactions[132] and neighbourhood modelling[133]. Modelling cells as graphs has been around since the early 2000s within cancer pathology, where it was shown that the graph metrics can distinguish healthy and unhealthy inflamed cells with high accuracy[134]. Recent interesting research in the field shows that graph neural networks can model tissue-level emergent phenotypes such as immune cell dispersion in colorectal tumours[135].

GNNs are particularly suited for representing cellular neighbourhoods as graph structures, where nodes represent cells or spots, and edges capture their spatial relationships. While traditional statistical and machine learning approaches treat observations as independent, GNNs use the graph structure to model the spatial dependencies and interactions between neighbouring cells[136]. Using an iterative message-passing framework, GNNs can capture information from each node to aggregate information from its local neighbourhood, enabling the model to learn hierarchical representations. With additional layers it is possible to capture both local, and global tissue organisation[137]. Graph convolutional layers, where a node’s feature representation is updated based on its own attributes and those of its neighbours, weighted by a learnable transformation function, is the most common approach[138]. Graph Attention Networks (GATs) are another approach, which use an attention mechanism that assigns different importance weights to neighbouring nodes instead of treating all neighbours equally[139]. This allows GATs to learn coefficients that focus

on the most important interactions, ensuring that the model captures the heterogeneous effects of the cells[140].

1.10 Key challenges of spatial data

It is important that when analysing spatial transcriptomics data, approaches account for the unique statistical properties of spatial data. A key property is spatial autocorrelation (**Figure 4**) where nearby observations are more likely to be similar than distant ones[141]. This violates the assumption of independence fundamental to many standard statistical models. Therefore, the use of spatially explicit models such as spatial autoregressive models or geostatistical kriging methods are required to account for these dependencies.

Another aspect of spatial analysis is the modifiable areal unit problem (MAUP), which arises when outcomes depend on the scale or boundaries of spatial aggregation[141], (**Figure 4**). MAUP occurs in two forms: the scale effect, where the level of spatial resolution alters the analysis (e.g., patterns detected at a fine scale may be masked at a coarser scale), and the zoning effect, where different ways of defining spatial boundaries or domains lead to different results. Therefore, relationships identified as spatially heterogeneous at one scale may appear locally homogeneous at another. Detecting spatially variable genes or clustering spatial domains can be highly dependent on scale and taking this into account is important to ensure biologically meaningful insights rather than artifacts of data aggregation. Multi-resolution frameworks are important to help reduce these effects. This has been widely explored in geo-statistics, but the effect within spatial biology is much less explored[142–144].

Spatial heterogeneity, also referred to as spatial non-stationarity, is another key aspect of spatial analysis[141]. It reflects the variability in relationships between variables across different spatial regions (**Figure 4**). In statistical terms, this means that the parameters of a model may vary spatially, violating assumptions of stationarity often required in standard statistical models. Techniques such as geographically weighted regression (GWR) or spatially varying coefficient models can be applied to explicitly model these spatially dependent relationships[145]. This is particularly important in biological contexts where the effects of microenvironmental influences, such as immune infiltration or stromal cell interactions, are not uniform but context dependent.

Many research studies in the field ignore the statistical considerations required for the accurate analysis of spatial data, and therefore risk losing valuable spatial context. This concern was highlighted in a recent review by Comber et al. (2024)[145], where it was noted that numerous spatial transcriptomic studies either cluster data without incorporating spatial information or propose novel approaches to address issues like spatial autocorrelation without referencing widely used geostatistical methods. For example, the SpatialDE method for spatially variable genes does not mention Moran's I, a widely used metric in geostatistics[128]. This highlights not only a lack of appropriate application of spatial statistics but also potential duplication of research efforts.

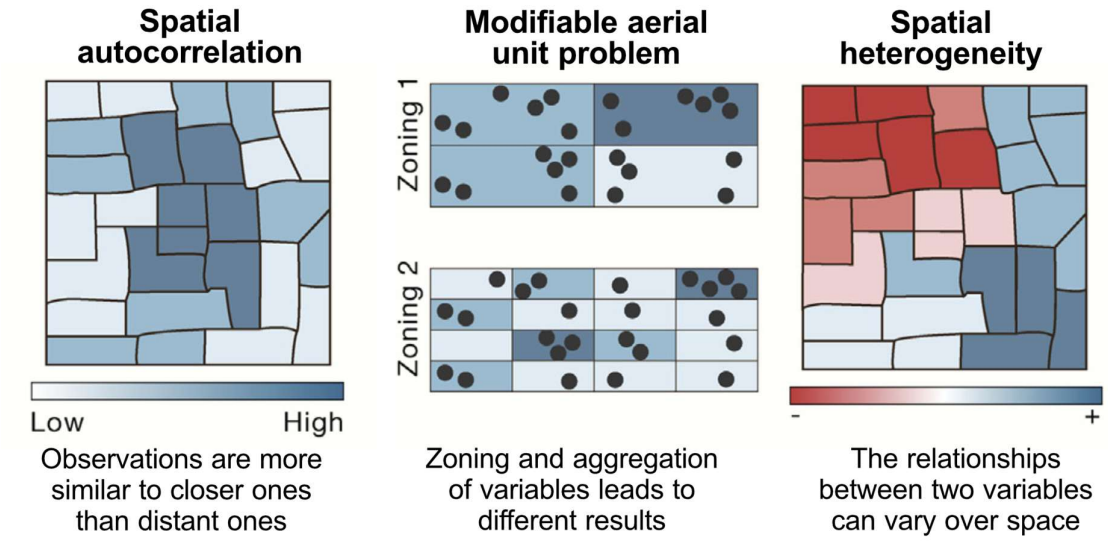

Figure 4 Key statistical properties of spatial data. From left to right, spatial autocorrelation, the modifiable aerial unit problem and spatial heterogeneity. Adapted from Comber et al. (2024)[146].

1.11 EMT in Breast Cancer

Breast cancer is the most commonly diagnosed cancer in women[147] and is a leading cause of cancer-related deaths, with over 2.3 million new cases diagnosed annually[148]. Survival rates vary widely depending on tumour subtype, stage at diagnosis, and access to medical care. Due to the advances in treatment, the five-year survival rate for localised breast cancer is over 90%, however there is a significant drop for metastatic disease, due to limited treatment options available[148].

Due to its prevalence, it provides a widely studied system with well-established clinical subtypes[147]. The predominant subtypes include hormone receptor positive

(ER+/PR+), HER2-enriched, and triple-negative breast cancer (TNBC)[147]. Additionally, there are extensive breast cancer datasets available for analysis, including bulk transcriptomics, such as from TCGA[149] and METABRIC[150], large scale scRNA-seq datasets and spatial transcriptomics datasets[109]. Given its extensive characterisation and the availability of diverse datasets, breast cancer serves as a valuable model for investigating EMT.

Additionally, in many breast cancers, well-known markers such as E-cadherin, vimentin, and GATA3 have been linked to disease progression and response to therapy[151,152]. Moreover, breast tumours often display spatial heterogeneity, with localised regions of hypoxia, immune cell infiltration, and stromal remodelling, factors known to influence EMT[120]. Therefore EMT-related spatial observations in this cancer type are likely to be clinically relevant. EMT has been well characterised in breast cancer cell lines, with the commonly used EMT inducers such as TGF-B and IL6 displaying similar effects as in other cancers, such as promoting invasiveness[27,153]. This also likely highlights the strong connection between EMT and immune cells or fibroblasts in the TME, as these stromal components actively secrete EMT-inducing factors. In breast cancer mouse models, EMT-like populations emerge within distinct spatial niches, particularly at the invasive front, where tumour-stroma interactions are most pronounced[154]. Recent studies have explored EMT in breast cancer using spatial transcriptomics, revealing interactions with CAFs, tumour-associated macrophages and hypoxia, though these efforts have been constrained by a limited number of slides and a small set of EMT markers[118,120]. Distinct EMT-related states have been detected in breast cancer, but their spatial relationships are poorly characterised[155].

1.12 Knowledge gaps and aims of the thesis

Despite the considerable progress in characterising EMT and identifying intermediate EMT states, there remains many unanswered questions. Firstly, there is a need for the development of appropriate methods that would enable a statistically rigorous approach to understand the relationships of EMT. For example, developing a method to detect statistically significant regions enriched for specific EMT states, and assessing the spatial relationships of these states. Whilst many methods detect spatial

niches, there are fewer methods designed to flexibly analyse continuous signatures, such as EMT, and compare the results using different spatial units. Additionally, there are a lack of methods that can differentially compare these relationships to other signatures, such as epithelial regions. Moreover, current approaches investigating EMT spatially focus on a limited set of marker genes, likely overlooking the complexity and heterogeneity of EMT. Additionally, they do not investigate different EMT states. Whilst many studies have focused on model systems to characterise the spatial relationships of EMT in breast cancer, studies are lacking investigating these relationships in human tissue samples, particularly statistically principled approaches to characterise EMT states and spatial relationships.

Another important question is how these diverse EMT states are influenced by both intrinsic genomic factors (e.g., copy number variants, mutations) and extrinsic microenvironmental cues (e.g., stromal cells, immune cells). Although recent studies have highlighted the importance of spatial context, showing that tumour cells undergo EMT in discrete niches, there is a lack of integrated methodologies that consider a full range of TME components, and model these variables together with genomic factors. With the wide range of models already developed in geostatistics, ecology and spatial machine learning fields to understand spatial processes on a much larger scale, I believe there is a gap in translating the advances in these fields in spatial biology.

1.13 Aims

1. Identify genomic and TME signals that impact EMT (Chapter 2)
    i. Understand genomic influences of EMT in bulk transcriptomics
    ii. Identify EMT-TME relationships in a subset of breast spatial transcriptomic data
2. Analyse the TME relationships at different biological scales to further understand the relationship with EMT in breast cancer (Chapter 3)
    i. Develop a pipeline to accurately capture EMT signals in spatial transcriptomic data using a statistically principled approach
    ii. Locate spatial clusters (hotspots) enriched for different EMT states and different cellular components within the TME

iii. Quantify the relationships between these hotspots across different spatial scales
iv. Compare these relationships to relationships identified using neighbourhood enrichment
v. Develop the pipeline into a Python package for reusability

3. Quantify the relative influence of intrinsic (genetic) and extrinsic (TME) factors on EMP in breast cancer (Chapter 4)
    i. Determine which EMT phenotypes are the least predictable by intrinsic and extrinsic factors and assess the coefficients of the models to understand the ranking of the most important genetic and microenvironmental factors
    ii. Compare the performance of the different models tested (graph neural networks compared to spatial regression models)
    iii. Assess spatial heterogeneity in EMT-TME interactions by applying geographically weighted regression (GWR) to capture intra-tumour variability in EMT-TME relationships

2 Chapter 2: Spatial heterogeneity of EMT

Despite extensive research on EMT, several key questions remain unanswered. One of the key open questions is understanding the role of hybrid EMT states, including their relationship to the TME. Additionally, much of our knowledge is based on model systems, such as cell lines and animal models. While some animal models preserve an intact TME, they come with significant limitations, including species-specific differences in immune responses, stromal composition, and tumour evolution. These models can therefore fail to fully capture the genetic and spatial heterogeneity seen in human tumours, leaving a gap in our understanding of EMT in its native context. Addressing these gaps is important for identifying therapeutic opportunities and improving the ability to predict tumour progression more accurately.

In this chapter, I begin by introducing the literature (Section 2.1) and methods (Section 2.2) before using spatial transcriptomics, specifically ST2K (first generation) and Visium, to investigate the relationships among cells undergoing EMT with other cells in the TME (Section 2.3). Notably, when this research was initiated, EMT had not yet been examined using spatial transcriptomics, and analysis using the Visium platform was in its infancy. By integrating these approaches, I build a foundation for the subsequent chapters, where more advanced spatial analyses further clarify the relationship between EMT states with the TME.

This chapter is based on material from Tagliazucchi *et al.* Nature Communications (2023), where I conducted all the spatial analyses.

2.1 Introduction

Spatial transcriptomics has significantly deepened our understanding of cancer by revealing the spatial heterogeneity within tumours, as highlighted and reviewed in Chapter 1.

A seminal study by Ståhl et al. (2016) pioneered the field by using spatially barcoded arrays to map whole-transcriptome data directly onto tissue sections, enabling histology to be linked with molecular analysis. Since then, various landmark papers have identified unique cellular niches linked with therapeutic outcomes[156–158]. For example, Berglund et al. (2018)[159] mapped prostate cancer transcriptomes and

uncovered discrete cellular niches with unique gene expression profiles, highlighting the important influence of the tumour microenvironment on prostate progression. Technological advancements, exemplified by the second-generation ST2K platform and higher-resolution methods like 10x Genomics' Visium, have enabled even more precise mapping of cell interactions. In a landmark study of HER2-positive breast cancer, Andersson et al. (2021) used ST2K combined with cellular deconvolution to generate high-resolution maps of cell type distributions and interactions in 36 tissue sections from eight HER2-positive breast cancer patients[158]. The analysis revealed that the breast tumours are highly heterogeneous, with distinct spatial domains corresponding to in situ and invasive cancer regions, immune infiltrates, and stromal compartments.

In this chapter, I build on the growing use of spatial transcriptomics in cancer research to explore the EMT continuum, its spatial organisation, and its interactions within the tumour microenvironment. By detecting EMT states in spatial transcriptomics, I begin to uncover the spatial relationships of the EMT continuum.

2.2 Methods

2.2.1 Spatial transcriptomics preprocessing

Three breast cancer patient samples were downloaded from 10x genomics (https://support.10xgenomics.com/spatial-gene-expression/datasets). Patient 1 was AJCC Stage Group I, ER positive, PR positive and HER2 negative. Patient 2 was AJCC Stage Group IIA, ER positive, PR negative and Her2 positive. Patient 3 did not have molecular details described. The analysis was conducted using data processed through the Space Ranger Visium pipeline. Normalisation was performed using the SCTransform R package, which applies a regularised negative binomial regression method. EcoTyper was used to estimate the proportions of different cell types and states for each spatial transcriptomic spot. The identified cell types included B cells, $CD4^+$ T cells, $CD8^+$ T cells, dendritic cells, endothelial cells, epithelial cells, fibroblasts, mast cells, monocytes/macrophages, NK cells, plasma cells, and neutrophils.

ST2K (ST second generation, 2000 spots/array) datasets (9 patients with 3–5 repeats each) were downloaded from https://github.com/almaan/her2st. All samples had been stained positive for HER2. The same pre-processing steps were employed as in Andersson et al.[158]. Briefly, this consisted of using SCTransform for normalisation and

Non-Negative Matrix Factorisation (NMF) for dimensionality reduction. The factors that contained consistent patterns across the tissue replicates were retained for analysis. The Stereoscope[122] (v.0.2) R package was used for cell-type deconvolution. The deconvolution data was downloaded from https://github.com/almaan/her2st. The major class consists of myeloid cells, T cells, B cells, epithelial cells, plasma cells, endothelial cells, CAFs, and perivascular-like cells (PVL cells). The minor tier contains finer partitioning of the major cell types, e.g., macrophages and CD8+ T cells. Further description of the deconvolution method is described by the authors[122].

The Seurat[160] R package was used for storing, manipulating and visualising the spatial transcriptomic data.

2.2.2 Spatial gene module scores

An EMT score was computed for each spatial transcriptomic spot by adapting the method previously used to score TCGA samples, this time leveraging scRNA-seq data from solely breast cancer cell lines. The EMT trajectory derived from single-cell data was mapped onto each spot, with the k-NN algorithm identifying the most similar single-cell samples. The mean of their pseudotime values was used to determine the EMT score. This process was performed across multiple breast cancer cell lines, and the average pseudotime across all lines was used to calculate the final EMT score.

To categorise EMT states, the pseudotime values were divided into three intervals corresponding to epithelial-like, hybrid-like, and mesenchymal-like states. The SpatialFeaturePlot function from the Seurat R package was used to visualize these scores. For correlation analysis, only spots containing epithelial cells were considered. The STUtility[161] R package was used to compute the 12 nearest neighbours for each epithelial spot, and cell type proportions were aggregated across these neighbouring spots.

2.2.3 Cluster identification from spatial transcriptomics data

The spatial transcriptomic dataset was filtered to retain only epithelial, hybrid, and mesenchymal genes. Clustering was performed using the FindClusters function in Seurat, which grouped spatial spots based on gene expression. This was achieved by computing k-NN and constructing a shared nearest neighbour graph. The EMT scores were averaged within each cluster. The results were then binned into three categories:

low, medium, and high EMT states, corresponding to epithelial (EPI), hybrid EMT (hEMT), and mesenchymal (MES) states. The cell type enrichment scores calculated per region were plotted using the enriched-region.py Python file from https://github.com/almaan/her2st.

2.2.4 Inference of interaction networks

Graph-based analysis of the Visium spatial transcriptomic slides was performed using the ScanPy[162] and SquidPy[116] Python packages, which enabled graph visualization and computation of graph metrics. The STUtility[158] package was modified to construct spatial graphs from the ST2K spatial slides.

Deconvolved spot results were used to assign node labels, while edges were established based on spot neighbourhood relationships. Further network analysis and querying were conducted using NetworkX[163].

2.3 Results

2.3.1 Tumour cell extrinsic hallmarks of EMT

To investigate associations with the TME, I analysed spatial transcriptomics data from three breast cancer slides generated using the 10x Genomics Visium platform, along with multi-region profiling of eight breast tumours using ST2K, as described by Andersson et al.[158]. This allowed me to explore the spatial heterogeneity of EMT and its links with other phenotypes within the tumour tissue. I used clustering to locate the areas within the breast cancer tissue that have homogeneous patterns of expression (see Methods) (**Figure 5**). I investigated the tumour microenvironment composition within these clusters in relation to EMT states.

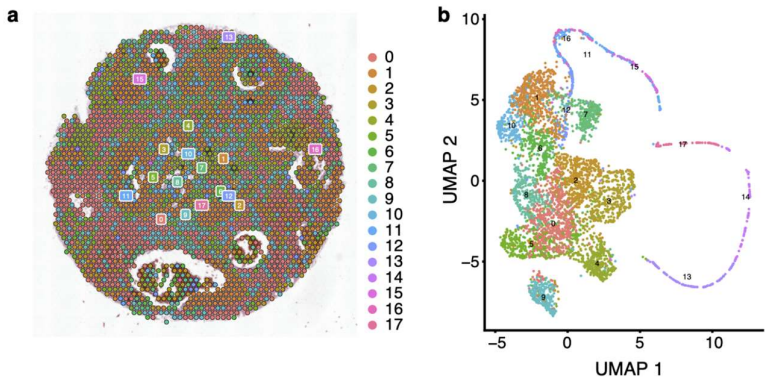

Figure 5 Spatial transcriptomic clustering. a The clusters of homogeneous expression profiles within the spatial transcriptomic spots shown for Patient 1. Each distinct cluster is represented by a unique colour, and each spot is coloured according to the cluster. **b** Expression clusters visualised using UMAP dimensionality reduction. Each dot represents a spot from the spatial transcriptomics slide and is coloured according to the cluster it was assigned to.

My analysis revealed extensive variation in EMT transformation across the tissue, with occasional clustering of EMT states within epithelial pockets (**Figure 6**). The most noticeable spatial pattern emerged in fibroblasts, which surrounded neoplastic epithelial areas in proportion to increasing EMT state, showing the strongest association with highly transformed tumour regions. Additionally, I identified links between the MES state and infiltration by $CD8^+/CD4^+$ T cells, monocytes, and macrophages (**Figure 6c, f, i**). I also found that transformed (hEMT/MES) regions were associated with dendritic cells and polymorphonuclear leukocytes (PMNs).

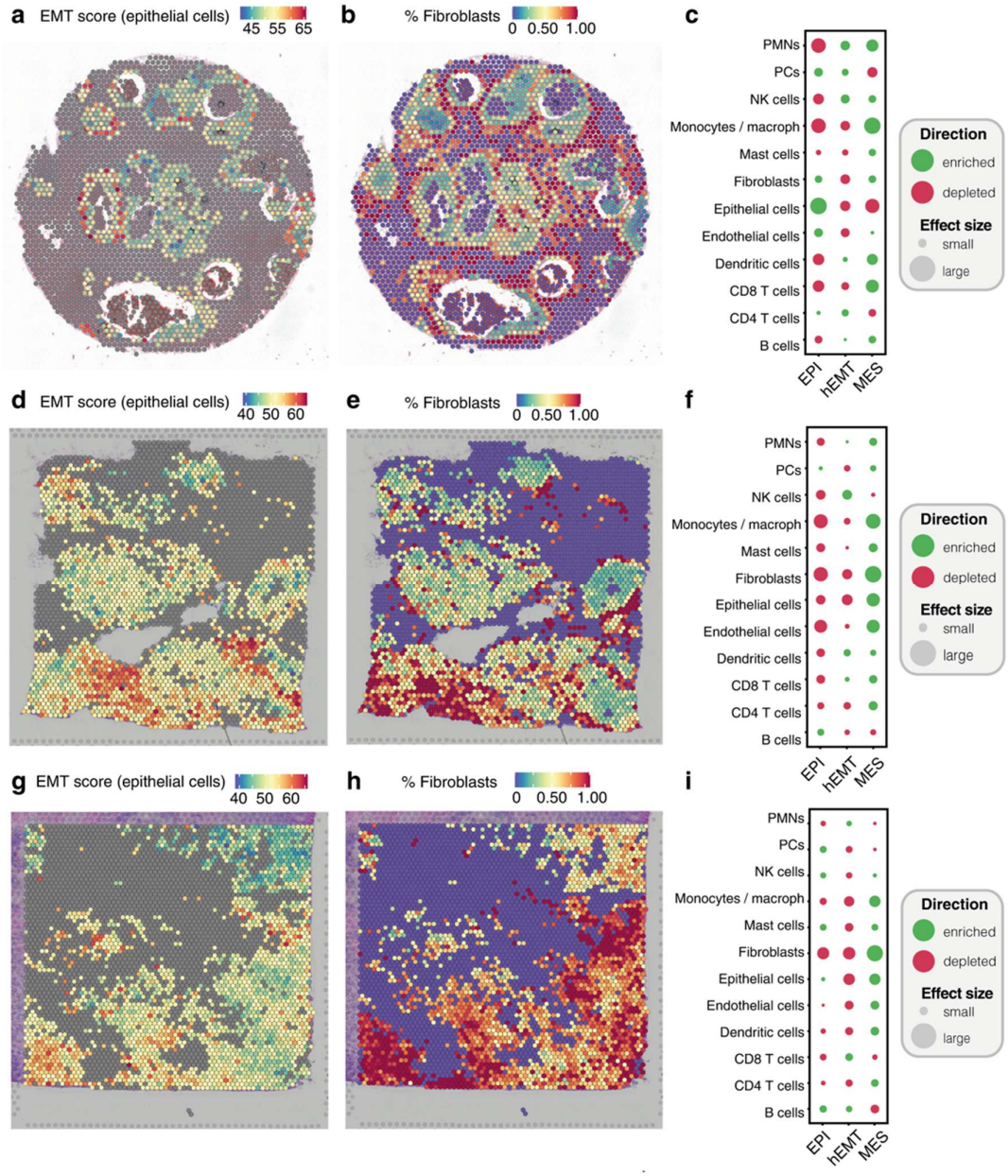

Figure 6 Spatial Patterns of EMT a- b EMT scores and the fraction of fibroblasts are visualised across within individual spots profiled across the tissue in a selected breast cancer slide, derived from spatial transcriptomics data from Patient 1 of the Visium dataset. The colour gradient reflects the expression of markers of the specific cell state (for EMT) or the fraction of cell types (for fibroblasts). **c** Enrichment and depletion of cell types in each EMT-based cluster from Patient 1. The plots represent the difference between the average cell type proportion value per region, compared to a permuted spot value (calculated 10,000 times). The plot marker size corresponds to the absolute enrichment score, and the colour represents the enrichment sign. PMN polymorphonuclear neutrophils, PC plasma cells, NK natural killer, macrophages. **d–f** The same annotations as above for a breast cancer sample from Patient 2 of the Visium dataset. **g–i** The same annotations as above for a breast cancer sample from Patient 3 of the Visium dataset

Expanding the analysis to a larger multi-region spatial transcriptomics dataset from multiple patients profiled using ST2K, I found that hEMT regions uniquely associated

with both MSC/iCAF-like and myCAF-like cells, while EPI states were linked only to MSC/iCAF-like cells, and MES states were predominantly associated with myCAF-like cells (**Figure 7**). Thus, the heterogeneity of hEMT-CAF associations may be explained by different subtypes of CAFs present in the context of hEMT and MES samples. Interestingly, natural killer (NK) cells were the only cell type exclusively associated with hEMT regions, hinting that NK cell activation strategies could be particularly effective against tumour cells in this hybrid state. Other cell types, such as endothelial cells, showed more variable associations across samples, and their spatial patterns aligned less consistently with those seen in bulk tissue analyses.

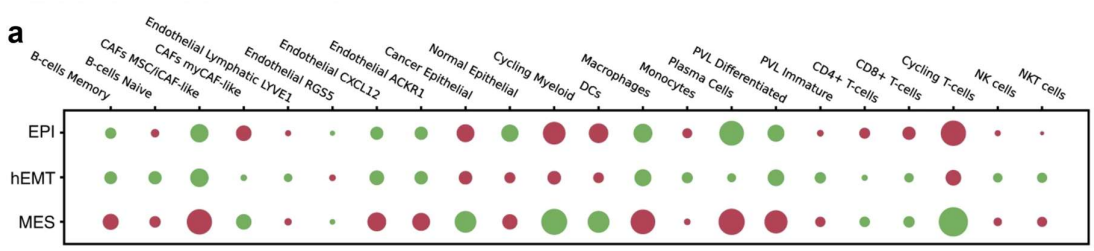

Figure 7 Spatial relationship of EMT in ST2K slides. Enrichment and depletion of cell types in EMT-based clusters derived from multi-region spatial transcriptomics slides from the ST2K cohort. CAF cancer-associated fibroblasts, myCAF myofibroblast CAF, DC dendritic cells, PVC perivascular cells, NKT natural killer T cells.

Beyond cell type enrichment, I inferred cell-cell interactions within the spatially profiled slides by analysing signal co-localisation. My findings indicate that fully mesenchymal tumour cells interact more frequently with fibroblasts, $CD8^+$, and $CD4^+$ T cells, whereas epithelial and hybrid EMT states showed no substantial differences in their interactions with immune cells (**Figure 8a, Figure 9a).** This reinforces the idea that transformed tumour cells interact with an immunogenic environment, which may, however, be suppressed by CAFs.

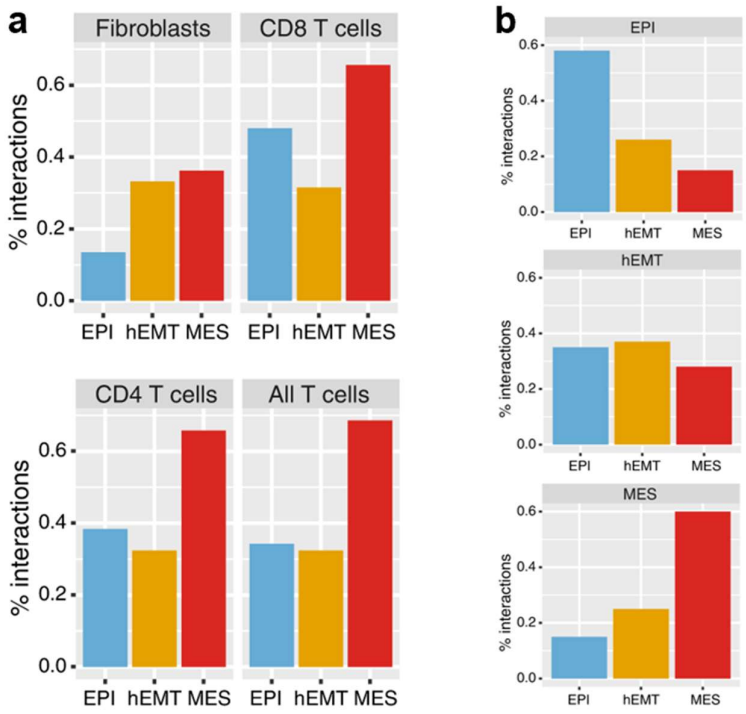

Figure 8 **Spatial interactions of EMT in Visium**. **a** Fraction of interactions established between tumour cells in the three EMT macro-states and fibroblasts or T cells in the Visium dataset. **b** Fraction of interactions established among cancer cells in different EMT macro-states in the Visium dataset.

I also examined how tumour cells interact with each other (**Figure 8b, Figure 9b**). Interestingly, cancer cells at the extremes of EMT transformation (either epithelial or fully mesenchymal) were more likely to interact with cells in the same state. In contrast, hEMT cells did not exhibit a preference for interacting with cells of any EMT state, suggesting that this hybrid phenotype may be more dynamic or more accessible from any other state, consistent with predictions from our HMM model. Notably, these patterns were highly similar across both Visium and ST2K datasets.

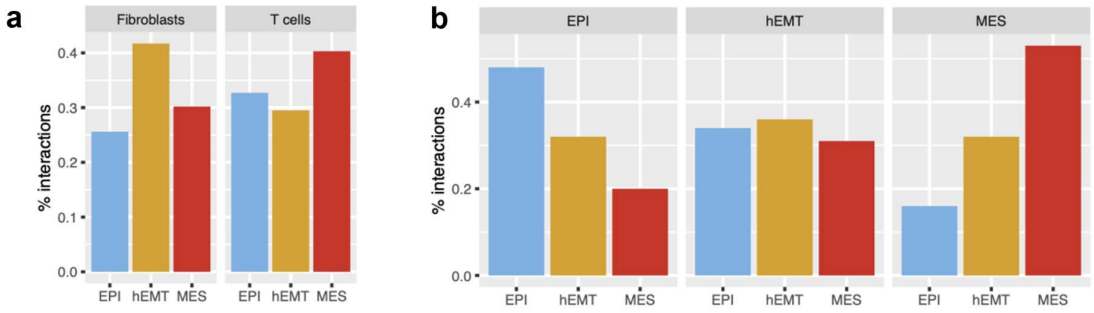

Figure 9 Spatial interactions of EMT in ST2K slides. **a** Fraction of interactions established between cells in the three EMT macro-states and fibroblasts or T cells in the ST2K dataset. **b** Fraction of interactions established among cancer cells in different EMT macro-states in the ST2K dataset.

Overall, these findings highlight a heterogeneous landscape of cell states and interactions, with both recurrent patterns and considerable spatial and patient-to-patient variability in EMT and TME composition. This suggests that local spatial effects play a key role in EMT progression. However, these associations may be partially masked by the fact that my analysis focuses on early-stage tumours, where the transition from ductal carcinoma in situ (DCIS) to invasive ductal carcinoma (IDC) is being examined. Stronger patterns may emerge in more advanced cancers, where hEMT or MES phenotypes are more prevalent, something not fully captured in this dataset.

Despite the large spatial variability, the continuum of EMT transformation is clear in spatially profiled slides, and stresses the importance of examining local effects to better understand tumour progression and responses to treatment.

2.4 Discussion

In this chapter, I present findings that reveal distinct EMT trajectories in cancer, defined by three macro-states. Fibroblasts and cytotoxic T cells often surrounded more mesenchymal neoplastic areas, and more frequent interactions with these cells were observed in this context. There was evidence for initial immune recognition as suggested by the co-localisation of MES with CD8/CD4+ T cell signals and hEMT with NK cell signals, which could be due to higher neoantigen presentation and subsequent exhaustion of T cells[164].

Cells with hybrid EMT features have been shown to give rise to daughter cells that are either mesenchymal or epithelial and are more prone to migrate[165], which could explain some of the heterogeneity observed for this state. Additionally, the hEMT state is likely composed of multiple distinct subpopulations, as highlighted by Pastushenko et al.[20], Goetz et al.[40] and Brown et al.[166].

The spatial analysis is limited by the small sample size, and larger spatial datasets will be required to further understand the more complex relationships established, particularly for intermediate EMT stages which are more heterogeneous than for the fully mesenchymal states. I analysed ST2K slides for the majority of patients, in addition to three Visium 10X genomics slides. At the time of the analysis, ST2K spatial transcriptomics was the more established platform, with Visium 10x Genomics analysis in its infancy. However, Visium 10x Genomics provides higher spatial resolution, with each spot measuring 55 μm in diameter and capturing the transcriptomic profile of approximately 1-10 cells per spot. In contrast, ST2K has larger spots, typically 100-150 μm, which means that each spot contains a larger number of cells per spot, reducing spatial resolution.

I was also limited in the ability to capture a broad spectrum along the EMT transformation as the data are only sourced from early-stage cancers. Additionally, CAFs can express similar mesenchymal markers that are also found in tumour cells undergoing EMT which can confound EMT analysis. Whilst I only focused on scoring tumour cells with EMT signatures to help ensure that signals were not mixed, deconvolving the spatial transcriptomic using a reference single-cell dataset labelled with EMT states would more accurately ensure CAF signals are not being captured.

The spatial analysis highlights the requirement for new methods to identify localised, context-specific effects within the tissue which may not generalise throughout the tumour. Additionally, techniques to identify spatial clustering of EMT signatures and their interactions with other domains of TME components will be important to understand the statistically significant relationships. Ensuring that these interactions are assessed across multiple spatial scales will provide a more comprehensive understanding of the relationships.

Despite these limitations, our analyses do serve as a proof of concept for the ability to survey EMT spatially and highlights the complex microenvironmental mechanisms that

shape EMT transformation during cancer. Intra- and inter-tumour heterogeneity are likely to create complex EMT-TME landscapes that require more extensive datasets and method development to fully understand. In Chapter 3 I address some of these limitations, including the need for more robust method development to capture spatial relationships in a statistically principled manner. Additionally, the limitations around the sample size, number of EMT states, method to distinguish EMT tumour cells from CAFs are addressed. I will further explore these relationships in a spatial single-cell resolved manner in Chapter 4.

3 Chapter 3: Multiscale spatial analysis of EMT and the TME

As discussed in Chapter 2, I identified several significant spatial relationships between cells undergoing EMT and their interactions with the TME. In this chapter, I will address some of the limitations I highlighted in Chapter 2 and investigate these relationships further.

Existing methods to interpret spatial transcriptomics focus on static units of measurement, such as fixed spot neighbourhoods or pre-defined non-spatially aware clustering methods, without systematically probing how modifying these units might alter the biological inferences drawn. Additionally, they often require discrete labels (such as cell type labels) as opposed to continuous labels (such as signature score values) for clustering. To address these challenges, I developed SpottedPy, drawing on geostatistical principles, to map biologically meaningful hotspots and assess spatial interactions.

In this chapter I expand the spatial analysis on a larger dataset of Visium breast cancer slides. I also improve the method for detecting EMT in tumour spots, to avoid potential confounding of the EMT signal with CAFs, by using Gaussian mixture modelling to assign states in scRNA-seq prior to deconvolution. Moreover, I expand on the number of intermediate EMT states investigated. I begin with a literature review of current methods, highlighting their limitations and how these gaps motivated the development of SpottedPy (Section 3.1), before describing the methodology (Section 3.2). I then provide an overview of the approach (Section 3.3.1), and show how SpottedPy can be used to investigate key cancer hallmarks (Section 3.3.2). I subsequently assess the relationship of EMT with different cell types in the TME (Section 3.3.4), before assessing the intra- and inter-patient heterogeneity. I investigate how the relationships change when increasing the size of the hotspots (Section 3.3.6), analyse the results in other cancer types (Section 3.3.8) and investigate the EMT states (Section 3.3.10). Finally, I discuss these results and the methodological approach in detail (Section 3.4).

This chapter is based on material from Withnell and Secrier. Genome Biology (2024)[167]. I have also used SpottedPy to explore gene signatures from a large

language model developed to predict EMT states, the results of which can be seen further described in Pan *et al*. BioRxiv (2024)[168] . I have also used SpottedPy to spatially characterise quiescence in breast cancer, and the results are described in detail in Celik *et al*. BioRxiv (2024)[169].

3.1 Introduction

Numerous studies have highlighted the important role of spatial transcriptomics for identifying tissue domains with distinct cell composition[170], investigating key cancer hallmarks[171], revealing immunosuppressive hubs[120,172], uncovering tumour ecotypes with divergent clinical outcomes[173] or the impact of specific drugs on inhibiting tumour progression[174]. However, isolating tissue regions specific to a biological question and examining interactions among cell populations at an appropriate scale remain significant challenges in these datasets. To identify biologically meaningful tissue subregions with spatial transcriptomics, several analytical strategies rely on unsupervised clustering of gene expression, including SpaGCN[124] and BayesSpace[125]. Other methods, such as NeST[175] or GASTON[176], go a step further by incorporating nested structures or topographical metrics to reveal hierarchically organized co-expression hotspots that mirror tissue histology.

Given that similar cells often group together[130,177], detecting statistically robust spatial cell clusters is important for validating the accuracy of cell states, especially given the inherent challenges of cell deconvolution methods in accurately identifying them. CellCharter[127] builds on this idea through Gaussian mixture models to identify stable clusters, thereby defining spatial niches that exhibit distinct shapes and cell plasticity. In addition, more targeted clustering approaches using user-defined signatures or cell types, implemented in Voyager[178] and Monkeybread[179], help refine the interpretation of cell types inferred from spatial transcriptomic deconvolution. Nevertheless, spatial clustering of continuous signatures and flexibly exploring spatial units across multiple scales remains a challenge.

Methods for evaluating the spatial proximity of different clusters typically rely on co-enrichment in the immediate neighbourhood, as highlighted by approaches available in packages such as Squidpy[116]. However, there is a shortage of methods that quantify differential spatial relationships among specific cell types or signatures (e.g., hypoxia)

and assess their spatial significance. Moreover, current strategies fall short in defining and comparing both short- and long-range interactions between specific regions or cell populations of interest. Because the spatial scale of certain cancer-related processes, such as hypoxia, remains uncertain, relying solely on neighbourhood-centric approaches may obscure more complex spatial interactions. In geostatistics, this issue is known as the modifiable areal unit problem (MAUP), where observed data patterns shift depending on the size and shape of the spatial analysis units[143]. While a growing number of methods address multi-scale analysis [116,120,126,180,181], the impact of varying spatial units has received limited attention in spatial biology[130]. Geostatistical and ecological concepts have increasingly been applied in histopathology to characterise and quantify the spatial organisation of tissue features[182]. For instance, Ripley's K, a widely used geostatistical tool for detecting random versus clustered point distributions, has been used to investigate the spatial interactions of various TME components, such as the distribution of immune cells in ovarian cancer[183]. Spatial autocorrelation metrics like Global Moran's I have been implemented to assess the overall clustering of different features within histopathology datasets[184]. In addition, hotspot analysis, widely used in areas like crime detection and epidemiology, has been used to identify immune-rich regions and to stratify patients based on breast cancer histology[185], although advanced methods to analyse hotspot relationships remain limited. Despite the increasing use of these techniques in histopathology, they are underutilised in spatial transcriptomics. Recently, Voyager was developed to provide key geostatistical tools in a format readily applicable to spatial transcriptomic data[178]. In this chapter, I build upon these geostatistical methodologies, such as those implemented in Voyager, presenting an analytical approach designed to investigate spatial relationships at multiple scales in 10X Visium transcriptomic datasets.

3.2 Methods

3.2.1 Spatial transcriptomic datasets

I combined the three datasets of breasts cancer 10X Genomics Visium spatial transcriptomic datasets into a common *anndata* Python format for analysis. Breast cancer Visium slides were obtained from Barkley et al.[118] (slides 0-2), from 10x

Genomics (slides 3-5)[109] and Wu et al.[173] (slides 6-12). Pre-processing and normalisation were conducted using the ScanPy (Single-Cell Analysis in Python) package[162]. I analysed a total of 32,845 spatially profiled spots, and retained spots if they exhibited at least 100 genes with at least 1 count in a cell, had more than 250 counts per spot and less than 20% of total counts for a cell which are mitochondrial. Pre-processed BCC slides were obtained from Gania et al.[186], PDAC slides obtained from Ma et al.[187] and CRC slides obtained from Valdeolivas et al.[188] I used the deconvolution results provided in each of the source studies.

3.2.2 Spatial data deconvolution

Due to the imperfect near-single cell resolution of current spatial transcriptomic methods, it is important to use a method to deconvolve each spot to infer the cellular populations enriched in each spot. I carried out cellular deconvolution using Cell2location[123]. Cell2location decomposes the spatial count matrix into a predefined set of reference cell signatures by modelling the spatial matrix as a negative binomial distribution, given an unobserved gene expression level rate and gene- and batch-specific over-dispersion. A scRNA-seq breast cancer dataset containing 100,064 cells from 26 patients and 21 cell types from Wu et al[173] was chosen to perform the deconvolution. Cell types in the chosen breast dataset consisted of cancer epithelial cells (basal, cycling, Her2, LumA, LumB), naïve and memory B cells, myCAF-like and iCAF-like cancer-associated fibroblasts, perivascular-like cells (PVL), including immature, cycling and differentiated, cycling T-cells, cycling myeloid cells, dendritic cells (DCs), endothelial cells expressing *ACKR1*, *CXCL12* or *RGS5*, endothelial lymphatic *LYVE1*-expressing cells, luminal progenitors and mature luminal cells, macrophages, monocytes, myoepithelial cells, natural killer (NK) cells, natural killer T (NKT) cells, plasmablasts, CD4+ T cells, and CD8+ T cells. I identified EMT states within the scRNA-seq cancer epithelial cells by scoring the cells with EPI and EMT signatures[189,190], and using Gaussian mixture modelling to assign the cells to a cluster. The optimal number of components (clusters) was determined by assessing the silhouette scores across a range of component numbers and selected the model with the highest score. This approach ensured an optimal balance between cluster separation and internal cohesion, resulting in a robust method of identifying EMT states.

The scRNA regression model was trained with 500 epochs, and the spatial transcriptomic model trained with 20,000 epochs. To delineate the tumour cells within the spatial transcriptomics dataset, I used the STARCH Python package designed to infer copy number alterations (CNAs)[191]. STARCH identifies tumour clones (setting *K*=2 clones) and non-tumour spots. It confirms identification of normal spots by clustering the first principal component into two clusters using K-means. Changing the value of *K* alters the number of identified tumour clones, but the number of cells labelled as tumour cells remains the same. This approach is based on the principle that the direction of maximum variance in the expression data typically reflects the division between non-cancerous and cancerous spots.

3.2.3 EMT state and hallmark signature scoring

To identify distinct EMT states, I employed data from Brown et al[166], consisting of seven RNA-seq sequenced cell clones, derived from SUM149PT inflammatory breast cancer cell line with 3 repeats spanning the EMT spectrum from epithelial-like (EPI), quasi-mesenchymal (M1), fully mesenchymal (M2) and three distinct intermediates (EM1, EM2, EM3) . I used these data to derive a weighted gene signature to represent the EMT states. I captured EMT gene patterns from this data using non-negative matrix factorisation (NMF) by applying the CoGAPs workflow[192]. I used ProjectR's implementation of *lmfit* R function to map the captured EMT patterns onto the spatial transcriptomic spots[49]. This transfer learning approach assumes that if datasets share common latent spaces, a feature mapping exists between them and can measure the extent of relationships between the datasets. The final states were captured with 20 patterns and 10,000 training iterations. The number of patterns were chosen based on capturing the discrete states with the highest accuracy. The EM1 state was not distinguishable from the EPI state, so I merged the two states. Thus, overall I obtained scores for one epithelial, two intermediate, a quasi-mesenchymal and a fully mesenchymal state for each spot.

Hypoxia and angiogenesis were defined based on signatures deposited at MSigDB[193]. The proliferative signature was compiled from Nielsen at al[194]. The immunosuppression signature was compiled from Wu et al[173] and Cui et al[195]. The checkpoint blockade response signature was compiled from Johnson et al[196] and Liu et al[197]. The exhaustion signature comprised classical exhaustion markers: *CTLA4,*

PDCD1,TIGIT, LAG3, HAVCR2, EOEMT, TBX21, BTLA, CD274, PTGER4, CD244 and *CD160*[198] . All these signatures were scored using *scanpy.tl.score_genes* function. EMT hotspots and coldspots were identified in the BCC, CRC and PDAC slides using the EMT hallmark signature[193].

3.2.4 Graph construction

The SquidPy[116] (Spatial Single-Cell Analysis in Python) package was used for graph construction using *sq.gr.spatial_neighbors* and slide visualisation of the Visium spatial slides. NetworkX[163] was used for further analysis of the networks derived from the spatial transcriptomic spots. The deconvolved spot results were used to assign node labels. Edges were assigned based on the spot neighbours.

3.2.5 Neighbourhood enrichment analysis

I calculated neighbours for each spot by summing the deconvolution results in a ring surrounding the spot of interest, and normalising by the number of spots assigned as a neighbour, using the adjacency matrix of the graph to calculate the interacting cells.

Two methods were developed to assess neighbourhood enrichment. *Inner outer* correlation (with the function *sp.calculate_inner_outer_correlations*) was calculated by correlating signatures across a central spot of interest and the direct neighbourhood of spots surrounding it (a ring encompassing six Visium spots), after filtering for tumour spots only. To perform the sensitivity analysis, I increased the number of rings surrounding a spatial transcriptomic spot (setting *rings_range* parameter in *sp.calculate_inner_outer_correlations* function) to consider as spot neighbours and compared the change in correlation coefficient. The first ring consists of 6 spots, and the second ring includes 18 spots (combined from the 1st and 2nd rings). Subsequent rings follow this pattern. The number of rings selected for sensitivity analysis reflects a balance between spatial coverage and resolution. Using a smaller number of rings (e.g., 1, 2, 3) allows the analysis to focus on the immediate microenvironment around the central spot, providing high resolution. As more rings are added, the spatial coverage increases, capturing broader interactions but potentially diluting local-specific signals. Correlations were calculated using Pearson's correlation coefficient.

An *all-in-one* correlation (*sp.calculate_neighbourhood_correlation* function) was calculated by correlating phenotypes with cells within a spot, and then incrementally

increasing the number of rings to correlate across progressively larger spatial units. The functions *sp.correlation_heatmap_neighbourhood* and *sp.plot_overall_change* plot the neighbourhood results.

3.2.6 Hotspot analysis

Hotspots were calculated using The Getis-Ord G* statistic as implemented using the PySAL package[199], using 10 as neighbourhood size parameter by default and a p-value of 0.05, unless otherwise stated.

The Getis-Ord G* equation is defined as follows:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{x}\sum_{j=1}^{n} w_{ij}}{s\sqrt{\frac{n\sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2}{n-1}}}$$

Where w_{ij} is the spatial weight between location i and j, $\bar{x}$ is the mean of the variable of interest across all locations, s is the standard deviation of the variable of interest across all locations and n is the total number of locations.

A high positive value at location i suggests a hotspot for the attribute, while a negative value indicates a coldspot. The significance of G^* is determined by comparing the observed G^*_{obs} to a distribution of G^* values generated under the assumption of spatial randomness. This distribution is obtained by permuting the attribute values across locations and recalculating G^* for each permutation. The p-value for a hotspot (when G^* is positive) or a coldspot (when G^* is negative) is then derived from this distribution. This approach provides a non-parametric method to evaluate the significance of spatial clusters, offering a robust measure against potential spatial randomness in the data. The significance of $G^*$is determined by comparing the observed G^*_{obs} to a distribution of G^* values generated under the assumption of spatial randomness. This distribution is obtained by permuting the attribute values across locations and recalculating G^* for each permutation. The p-value for a hotspot (when G^* is positive) or a coldspot (when G^* is negative) is then derived from this distribution. This approach provides a method to evaluate the significance of spatial clusters, offering a robust measure against potential spatial randomness in the data.

Hotspots can be identified by calling *sp.create_hotspots* function, and specifying in the *filter_columns* parameter what region within the spatial slide to calculate the hotspot from e.g. tumour cells. The *neighbourhood_parameter* can be altered here (default=10). *relative_to_batch* parameter ensures that hotspots are calculated separately for each slide, otherwise, they are calculated across multiple slides. Importantly, if multiple slides are used (highly recommended for statistical power), these should be labelled using *.obs['batch']* within the *anndata* object. Additionally, the library ID in the *.uns* data slot should be labelled with the *.obs['batch']* value. Hotspots can be plotted using *sp.plot_hotspots.*

Hotspots and coldspots for EMT states and cell proliferation were calculated after filtering for tumour cells as labelled by STARCH. All other hotspots (deconvolved cell proportion data and angiogenic and hypoxia signatures) were calculated using all the spots within the spatial transcriptomic slide.

3.2.7 Distance metrics

After calculating the hotspots and coldspots, I then assessed the distances from hotspots of interest (EPI and EMT) to other cells types and signature hotspots and coldspots. I used the ***shortest path*** approach to calculate distances between hotspots as follows:

- Let H represent the set of coordinates of spots in the hypoxia hotspot.
- Let M represent the set of coordinates of spots in the mesenchymal tumour hotspot.
- Let E represent the set of coordinates of spots in the epithelial tumour hotspot.

For a spot m in M and a spot e in E, the shortest path to any point h in H was determined:

$$d_{min}(m, H) = \min_{h \in H}(d(m, h))$$

$$d_{min}(e, H) = \min_{h \in H}(d(e, h))$$

Where $d(m, h)$ represents the Euclidean distances from a spot m in M. After obtaining the minimum distances for each spot in M and E I calculated the median (with the additional functionality to choose min or mean) to provide a summary statistic that

reflects the general proximity of each hotspot (M and E) to H. The function *sp.calculateDistances* calculates this.

To then infer the impact of cellular hotspots on distance to EMT compared to EPI hotspots, I employed Generalised Estimating Equations (GEE). This model enables me to estimate population-average effects involving repeated measurements across multiple spatial transcriptomic slides. The model estimates the coefficient (β_{mes}) for the transition from reference hotspots (E) to primary hotspot variables (M). A positive β_{mes} would indicate that mesenchymal hotspots are, on average, located further from hypoxic areas compared to epithelial hotspots, while a negative value suggests a closer proximity. *sp.plot_custom_scatter*, setting *compare_distance_metric* to *min, mean* or *median* to compare the summary statistics for each hotspot across each slide. Setting it to *None* calculates the statistical significance of all distances from each hotspot.

The ***centroid approach*** is calculated as follows. The centroid C_H of a set of spots H with coordinates x_h, y_h is the arithmetic mean of the coordinates. This point represents the centre of the mass of the points in the set H.

For set H :

$$C_H = \left(\frac{\sum_{h \in H} x_h}{|H|}, \frac{\sum_{h \in H} y_h}{|H|} \right)$$

Similar calculations are employed for M and E. I then calculated the Euclidian distance between the centroids.

3.2.8 Tumour perimeter calculation

Any spot was considered part of the tumour perimeter if it had more than one neighbouring spots (nodes in the graph) that were not classified as tumour spots. A spot $s \ \epsilon \ S$ is considered part of the tumour perimeter, P, if:

$$s \in P \iff |N(s) \cap (S \setminus T)| > 1$$

Where S denotes the set of all spots, T denotes the set of tumour spots $N(s)$ represents the neighbouring spots of spot s. This approach helped me to delineate the boundary of the tumour accurately by focusing on the transitional area where tumour

and non-tumour spots meet (called using *sp.calculate_tumour_perimeter*). To quantify the number of tumour hotspots, I calculated the number of connected components within the graph that were labelled as hotspots. This calculation was crucial for understanding the distribution and clustering of tumour cells.

3.2.9 Sensitivity analysis

The sensitivity analysis to evaluate the impact of varying hotspot sizes on the spatial relationships was achieved by incrementally adjusting the neighbourhood parameter for the Getis-Ord statistic, which directly influenced the size of identified hotspots (*sp.sensitivity_calcs)*. As I expanded the neighbourhood parameter, I compared the distances between the newly defined hotspots and other existing hotspots of interest.

To assess the robustness of the spatial relationships between cell types and gene signatures, I systematically introduced Gaussian noise into the cell type proportion data and gene signature matrix. Gaussian noise, characterised by a mean of zero and varying standard deviations, was added to mimic experimental and technical variability. This approach allows me to evaluate the stability of detected EMT hotspots under different noise conditions. I defined a range of sigma values to represent varying levels of noise intensity. To further test the robustness of the spatial relationships, I randomly shuffled the cell proportion data and gene signature values and assessed how this affected downstream analysis.

3.2.10 Statistical analysis

Groups were compared using a two-sided Student's t test. Multiple testing correction was performed where appropriate using the Bonferroni method. Graphs were generated using the seaborn and Matplotlib Python packages.

3.3 Results

3.3.1 Overview of SpottedPy methodology

Building on the EMT-TME results identified in Chapter 2, I have more robustly characterised the relationships across different spatial scales; from direct cell-cell interactions to immediate neighbourhoods to across larger modules (**Figure 10**).

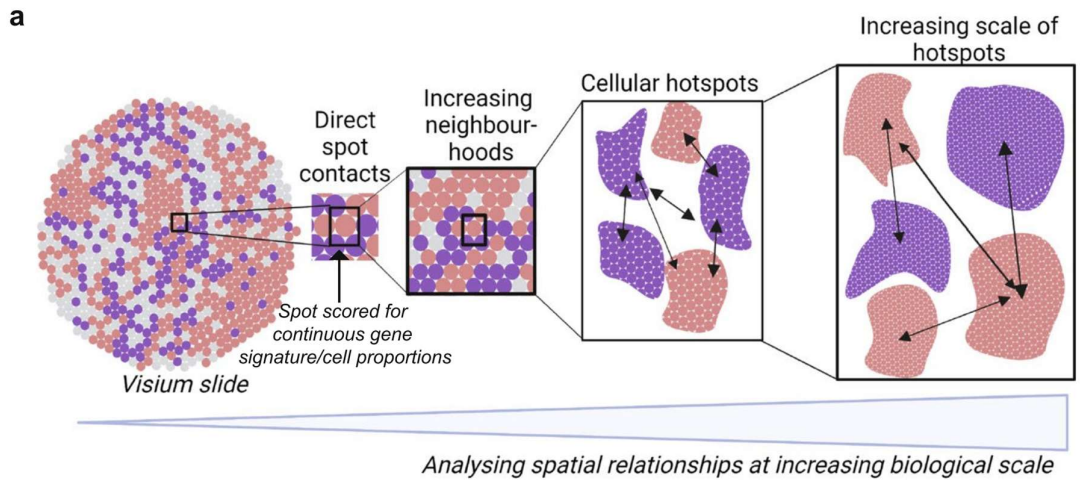

Figure 10 SpottedPy provides a multi-scale approach to analyse spatial transcriptomic relationships. Overview of spatial scales captured in the SpottedPy workflow, from direct cellular contacts to broader cellular hotspots. Figure created with BioRender.com.

Whilst neighbourhood enrichment is widely employed in the field[116,126,118], the analysis of continuous expression signatures and the influence of neighbourhood size on spatial relationships are comparatively underexplored. Additionally, the characterisation of the relationships of hotspots (spatial clusters) has yet to be addressed. This necessitated novel method development and therefore I developed SpottedPy, a Python package to allow me to probe the relationship across multiple scales in a statistically principled manner (**Figure 11**).

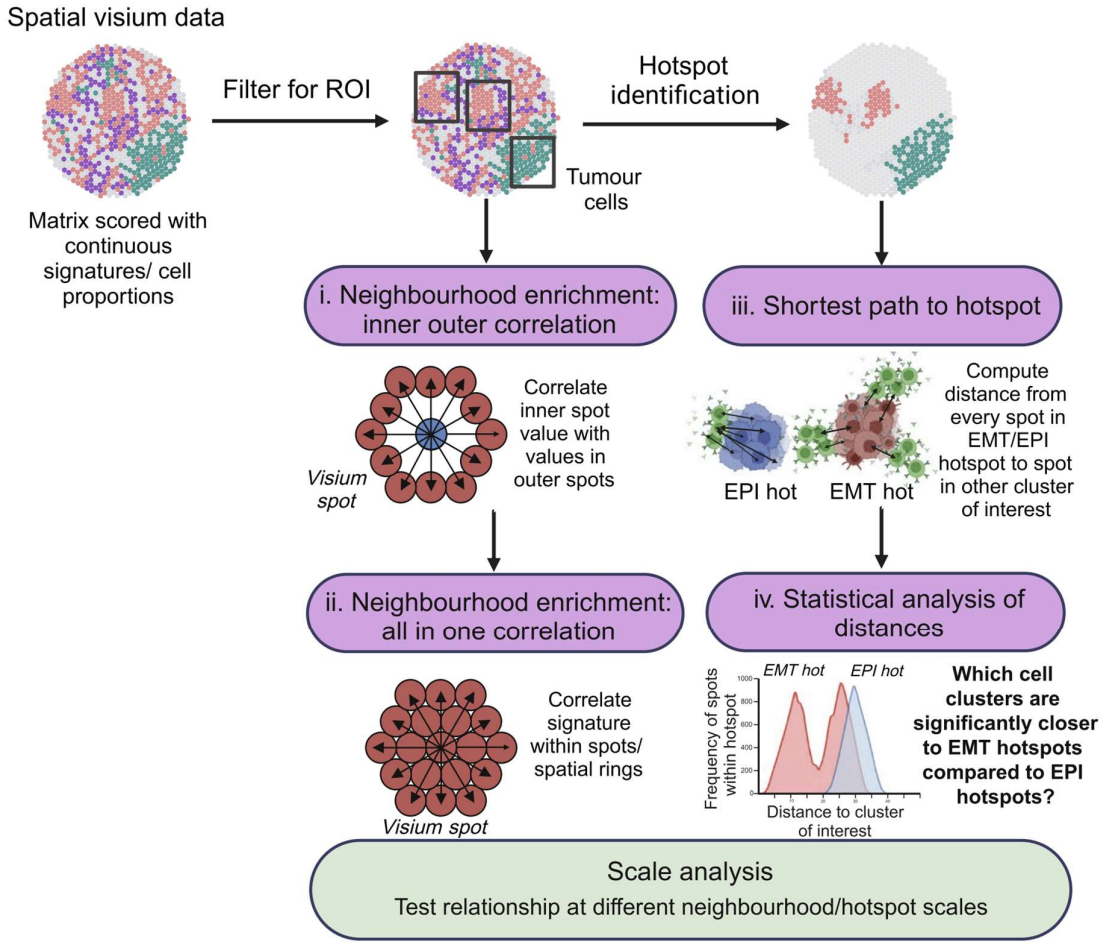

Figure 11 SpottedPy workflow overview. Visium spatial transcriptomic data is loaded as a pre-processed AnnData object where there is the option to select the region of interest (ROI) within the slide e.g., AnnData.obs column labelled with tumour cells. The default spatial analytics include: (i) Neighbourhood enrichment: inner outer correlation, which correlates cell prevalence or signatures in individual spots with their immediate neighbourhood, (ii) Neighbourhood enrichment: all in one correlation, which correlates cell prevalence of signatures within a spot or spatial unit (iii) Shortest path to hotspot, which calculates the minimum distance between each spot within a selected hotspot and the nearest spot in other hotspots, (iv) Statistical analysis of distances, which compares distances from a reference hotspot to another hotspot of interest, and assesses the statistical significance of the relationships. Scale analysis allows me to compare relationships defined at different scales in both approaches, either by increasing the number of rings included for neighbourhood enrichment or increasing the hotspot size. The outputs for the different modules include various plots to highlight the relationships. Figure created with BioRender.com.

The approach includes:

- **Neighbourhood enrichment analysis:** I develop functions to examine correlations between cell states, populations or processes within individual spatial transcriptomics spots and their immediate neighbourhood (**Figure 11i-ii**). Here, a "neighbourhood" is defined as a ring composed of six Visium spots surrounding a central spot, calculated by modelling the spots as a network. This method allows me to test how a signature affects its direct neighbourhood (inner-outer correlation) or to assess all spots within that neighbourhood collectively (all-in-one correlation).

- **Hotspot identification:** I have implemented the Getis-Ord G* statistic to identify spatial clusters of continuous gene signatures across spatial transcriptomic slides **(Figure 11).** Users can selectively focus on particular regions of the slide when generating hotspots. By comparing regions of high or low expression or cell-type signatures against a null distribution, this analysis identifies the statistically significant "hotspots" or "coldspots." Hotspots indicate areas with a concentrated presence of a specific cell type or signature (unlikely due to chance), while coldspots mark areas where the target cells or signatures are notably scarce. I also offer functionality to test whether specific gene signatures are enriched in hotspots or coldspots.

- **Distance statistics:** I provide functionality that measures and interprets the distances between detected clusters (e.g. tumour and immune hotspots). The main approach computes the *shortest path to a hotspot*, defined as the minimum distance from any point within a defined hotspot to the nearest point in another (**Figure 11iii).** Importantly, SpottedPy allows the user to compare distance distributions to key hotspots, for example, finding the hotspots that are significantly closer to mesenchymal hotspots than epithelial hotspots (or other areas that can be considered as a reference) **(Figure 11iv).** SpottedPy assigns statistical significance to these proximity measures, assessing whether observed distances are unlikely to result from random chance. To analyse relationships across multiple slides, I employ generalized estimating equations, allowing users to test the minimum, mean, or median distance from each hotspot or consider every distance from each spot within a hotspot.

- **Scale/sensitivity analysis:** I provide functionality to investigate how cell-cell relationships evolve within the tissue by varying the size of the neighbourhood or hotspot of interest. In the neighbourhood enrichment approach, this involves adjusting the number of concentric rings around the central spot. For the hotspot approach, SpottedPy recalculates the Getis-Ord G* statistic with different neighbourhood sizes, revealing clusters at multiple spatial scales. By examining how hotspot distances change with neighbourhood size, the package sheds light on how spatial relationships change or remain consistent at various scales. In addition, SpottedPy enables users to explore how cluster relationships respond to changes in the significance threshold for hotspot detection with the Getis-Ord G* statistic.

3.3.2 Spatial transcriptomic slide annotation overview

I used the SpottedPy functionality to gain deeper insights into the interactions between tumour cells undergoing EMT and the TME across 12 breast cancer 10x Genomics Visium slides. These slides were integrated from Wu et al[18], Barkley et al[24] and the 10x Genomics website[25]. To infer individual cell types within the slides, I deconvolved the slides using the Cell2location method[123] and a scRNA-seq reference of annotated breast cancer cell population profiles from 123,561 cells[13]. In the scRNA-seq dataset, I scored tumour cells with predefined epithelial (EPI) and epithelial-to-mesenchymal transition (EMT) signatures (see Methods) and employed Gaussian mixture modelling to assign a state to each tumour cell (**Figure 12)**[189,190].

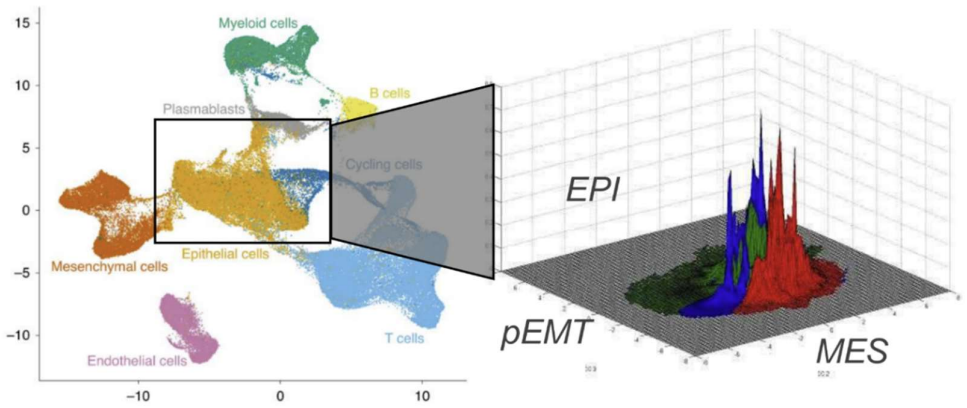

Figure 12 EMT state identification in the breast cancer scRNA-seq reference data prior to deconvolution. Adapted from Wu et al.[173].

To precisely capture the tumour cells within the spatial transcriptomic data, where expression can vary widely, I used the STARCH copy number inference tool[191]. I validated these results by comparing them with publicly available, pathologist-annotated slides[200,201] **(Figure 13).**

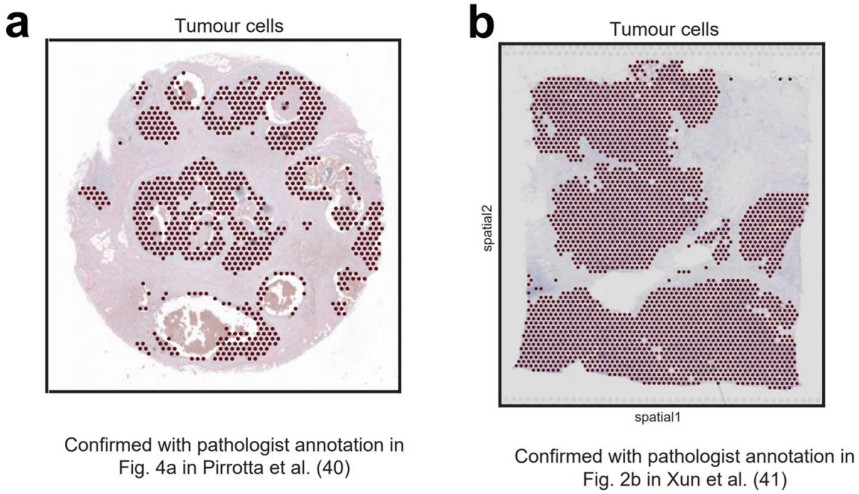

Figure 13 Validation of tumour cell identification. a Tumour cells as estimated by STARCH for slide 5, and the reference for the pathologist annotation confirming tumour cell estimation is provided. **b** Similar to (a) but for Slide 3.

Furthermore, to explore the heterogeneity of stable EMT states during the development and progression of breast cancer, I used the discrete EMT states recently defined by Brown et al[166], consisting of an epithelial phenotype, two intermediate

(hybrid) states (EM2 and EM3), a late intermediate quasi-mesenchymal state (M1) and a fully mesenchymal state (M2). A summary of my spatial slide annotation workflow is provided in **Figure 14.**

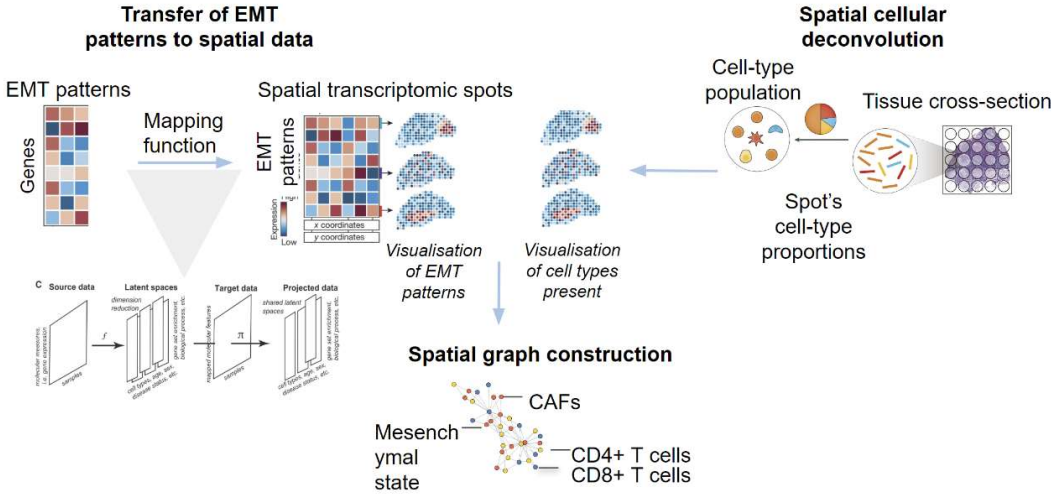

Figure 14 Preprocessing workflow prior to applying SpottedPy. Adapted from Stein-O'Brien et al[49] and Khavari et al[202]. EMT states were annotated both using cellular deconvolution and EMT pattern transfer using ProjectR. SpottedPy was applied to these various methods of detecting EMT and results compared.

3.3.3 Using SpottedPy to analyse the relationship of EMT and associated tumour hallmarks

I first focused on understanding the relationship of EMT tumour hotspots with two hallmarks of cancer known to be associated with EMT: hypoxia and angiogenesis. Hypoxia, characterised by low oxygen levels, has long been recognised as a key driver of tumourigenic processes[203]. Under hypoxic conditions, tumour cells stabilise hypoxia-inducible factors (HIFs), particularly HIF-1α, which promotes angiogenesis[204], the formation of new blood vessels from existing vasculature, to re-establish oxygen supply. Hypoxia has been shown to induce EMT and confer therapy resistance[205], highlighting the importance of understanding how these relationships develop spatially within the tissue. Such insights could facilitate the design of localised treatments that disrupt these interactions in breast cancer.

Using SpottedPy, I delineated tumour regions in each spatial transcriptomics slide and subsequently identified EMT hotspots within these areas, using the EMT state

assigned through Cell2location **(Figure 15a)**. To confirm and further explore the emergence of other cancer hallmarks emerging in the context of EMT, I defined hotspots for proliferative, hypoxic, and angiogenic gene signatures in the same slides **(Figure 15a).** Visual inspection shows that angiogenic and hypoxic hotspots frequently accompany EMT hotspots (**Figure 15a**). Quantifying hotspot distances with SpottedPy confirms that EMT hotspots generally lie closer to angiogenic and hypoxic hotspots compared with EPI hotspots, proliferative hotspots, or the overall tumour population (**Figure 15b-c)**. By contrast, proliferative hotspots are significantly nearer to EPI hotspots ($p<0.001$, **Figure 15c).**

To determine the positioning of EMT and EPI areas within the tumour, I used SpottedPy to estimate the tumour perimeter (**Figure 15d**) and calculated distances to it. EMT hotspots reside closer to the perimeter than EPI hotspots, indicating a state with significant interaction with the surrounding microenvironment (**Figure 15e**). As expected, angiogenesis hotspots appeared nearest the tumour boundary, followed by hypoxia hotspots (**Figure 15f**). The localised presence of angiogenesis near the perimeter aligns with its role in delivering nutrients and oxygen to expanding tumours[206]. Hypoxic regions developing just beyond these angiogenic zones reflect the fact that tumours often outgrow their vasculature, resulting in areas lacking sufficient oxygen[203]. I found that hypoxic coldspots occur closest to the perimeter **(Figure 15f**), where oxygen availability is higher.

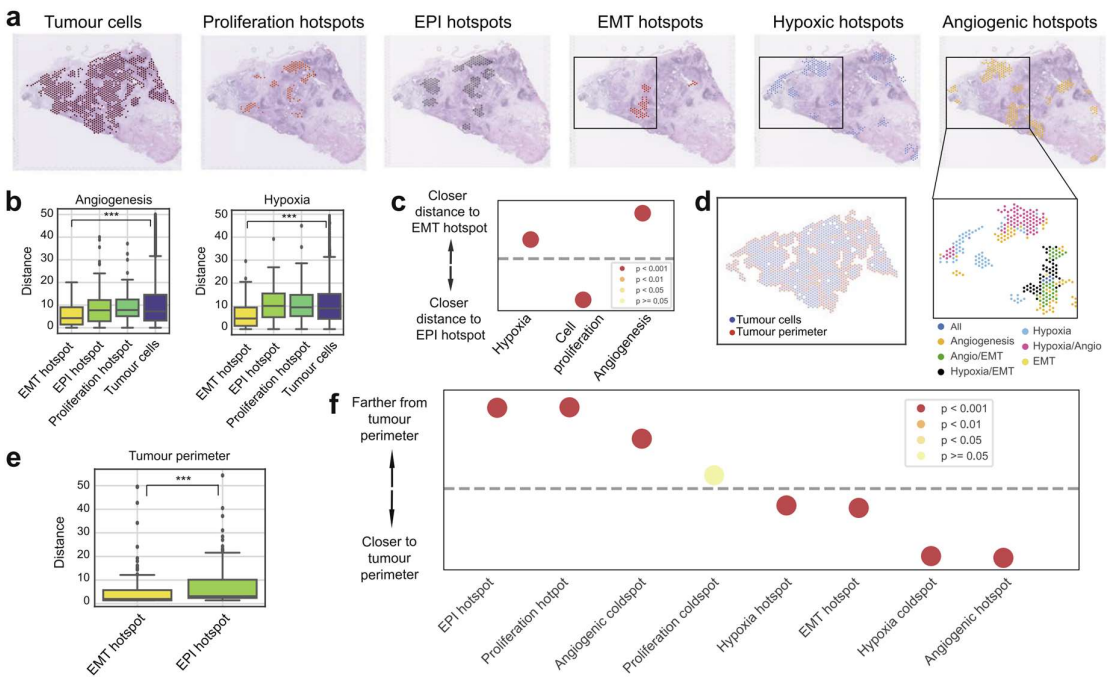

Figure 15 The spatial interplay between EMT progression and cancer hallmarks. a A spatial transcriptomics slide (slide 0) highlighting from left to right: tumour spots, proliferation hotspots, EPI hotspots, EMT hotspots, hypoxic hotspots, and angiogenic hotspots identifed by SpottedPy. The black square indicates a representative area where the close proximity of EMT, angiogenic, and hypoxic hotspots is depicted. **b** Distances from angiogenic (left) and hypoxic (right) hotspots to EMT hotspots, EPI hotspots, proliferative hotspots, and the average tumour cell, respectively, averaged across all 12 samples (*** p< 0.001). **c** Differences in proximity between EMT hotspots/EPI hotspots and hypoxic, proliferative, and angiogenic regions, summarized across the 12 slides. The dashed line represents no difference in relative distance to EMT hotspots or EPI hotspots. The dots situated above the dashed line indicate hallmarks that are significantly closer to EMT hotspots. The colors indicate the p-value ranges obtained from the Student's t-test for differences in distance to EMT hot/cold areas. **d** Spatial plot depicting the tumour perimeter in red and the tumour cells in blue. **e** Distance from the tumour perimeter to EMT hotspots and EPI hotspots, respectively (*** p < 0.001). **f** Distances from selected hotspots to the tumour perimeter, ordered by increasing proximity, across the 12 cases. The dashed line represents no significant difference. The colors depict p-value ranges obtained from Student's t-tests for differences in distance to the tumour perimeter.

In contrast, proliferative hotspots are observed farthest from the tumour edge and are spatially distinct from EMT hotspots. This pattern corroborates studies suggesting a proliferative epithelial core and a peripheral EMT population that enables cell migration and intravasation[207,208]. These spatial relationships were consistently observed across all examined breast cancer slides **(Figure 16)**.

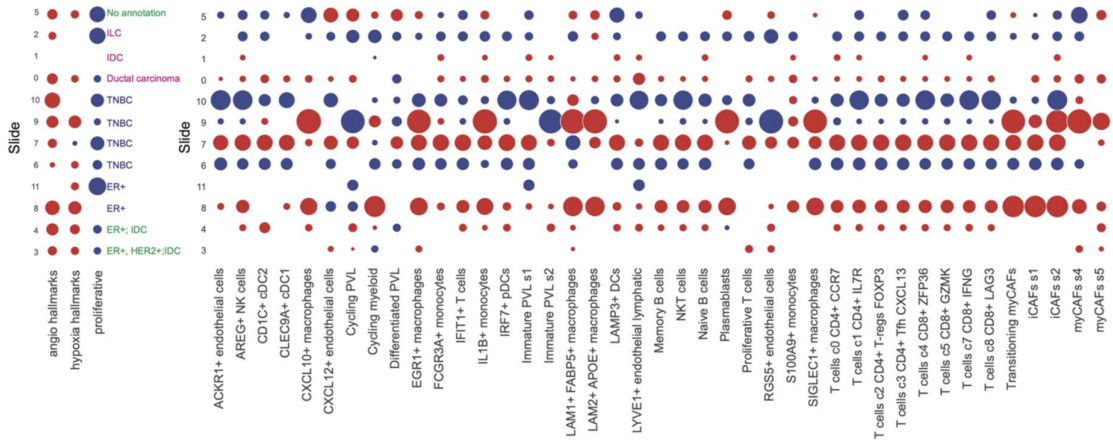

Figure 16 Bubble plot depicting distances between cancer hallmark signatures and TME classes and EMT/EPI hotspots for each slide (row). Blue depicts hallmarks that are significantly closer to EPI hotspots and red represents hallmarks that are significantly closer to EMT hotspots (Student's t test $p<0.05$), , adjusted for multiple testing using the Bonferroni correction. White indicates a non-significant relationship. Tissue annotations, if available, are included on the right-hand side for each sample, coloured by batch. IDC= Invasive Ductal Carcinoma. ILC= Invasive Lobular Carcinoma.

3.3.4 EMT hotspots exhibit immunosuppression and are shielded by myCAFs and macrophages

After validating SpottedPy's ability to capture expected spatial hallmarks of EMT in breast cancer tissue, I expanded the analysis to examine how tumour cells undergoing EMT interact with various immune and stromal cell types in the tumour microenvironment. In addition to the EMT hotspots, I identified hotspots for 41 different TME cell types, encompassing lymphocyte, myeloid, and fibroblast populations, based on the cell types as defined by Wu et al.[7] (**Figure 17a-c)**. Visual inspection of these hotspots revealed that myofibroblastic CAF (myCAF) hotspots often co-localise with EMT hotspots (**Figure 17a-c).**

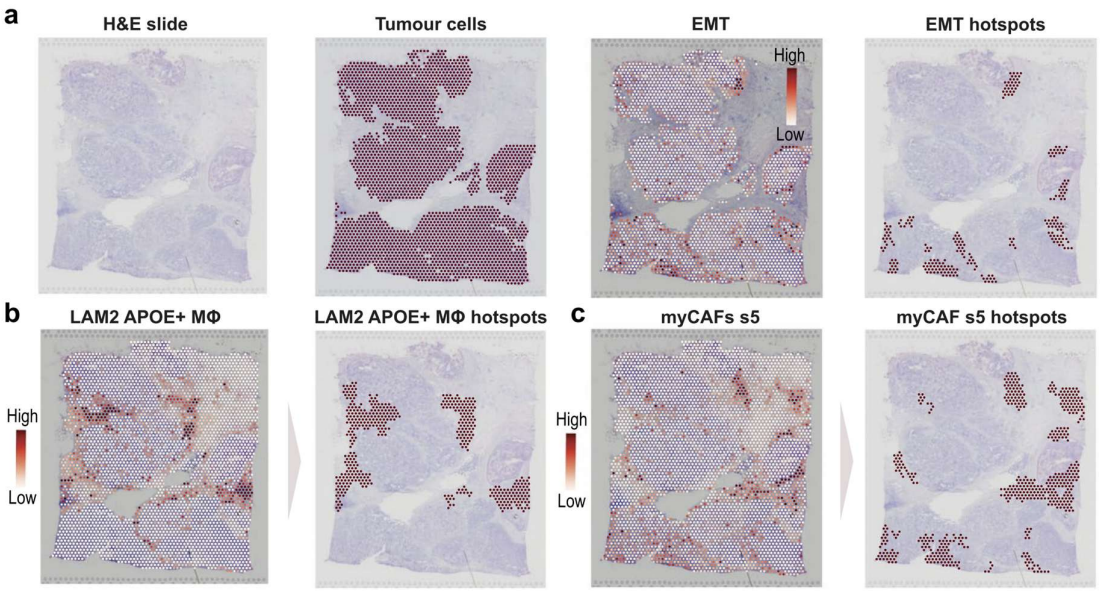

Figure 17 The spatial interplay between EMT progression and the TME. **a** Spatial transcriptomics plots highlighting tumour cell spots (left), the EMT gradient through these tumour spots (middle), and EMT hotspots identified by SpottedPy (right) in slide 5. **b** Spatial localisation of macrophage-enriched spots (left) and SpottedPy-deflned LAM2 APOE+macrophage hotspots (right) in slide 5.

Quantifying hotspot distances with SpottedPy confirmed that tumour EMT hotspots did indeed lie significantly closer to myCAF hotspots (**Figure 18).**

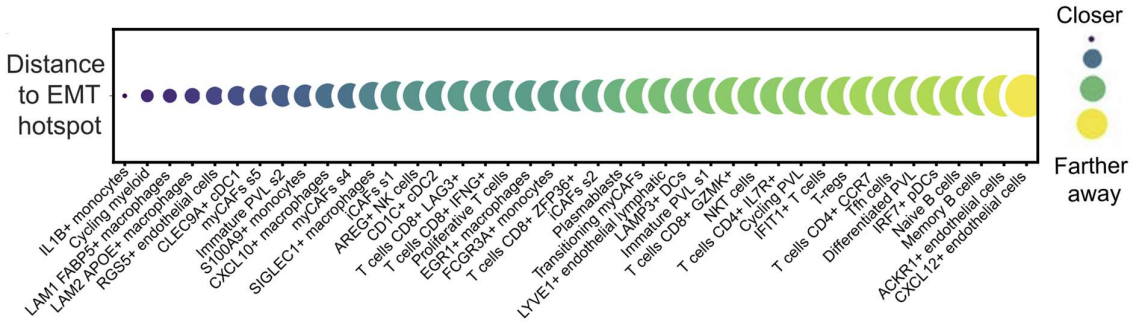

Figure 18 Distance between EMT hotspots and diferent TME cell hotspots, ranked by proximity. Smaller, darker bubbles represent shorter distances to EMT hotspots. Results are averaged over 12 slides.

The relationship is particularly highlighted when we look at the cellular niches that are significantly closer to EMT hotspots compared to EPI hotspots, revealing a predominance of various CAF subtypes **(Figure 19).** These observations align with existing literature, as myCAFs are known to secrete TGF-β, a well-recognised EMT

inducer[209], and have been linked to ECM deposition and the suppression of antitumour immunity[210,211,212,213].

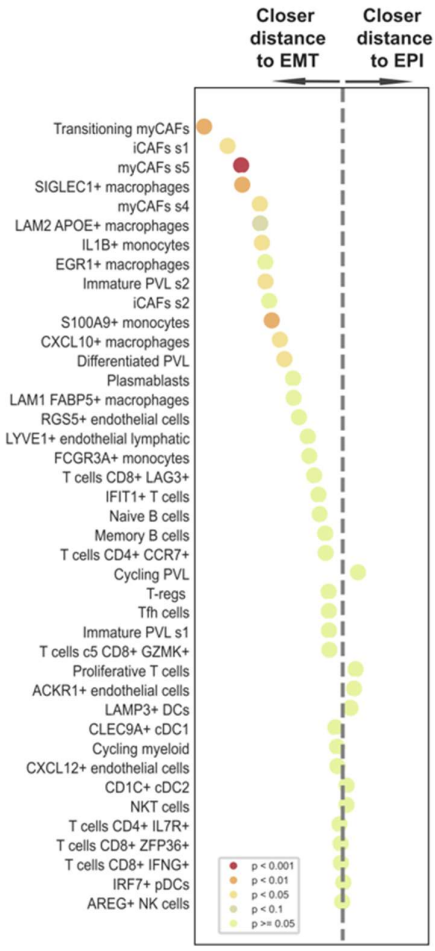

Figure 19 Distances from various cells in the TME to EMT/EPI hotspots. The dashed line represents no difference in proximity to either EMT hotspots or EPI hotspots. The dots situated to the left of the dashed line indicate cell populations that are signifcantly closer to EMT hotspots, ordered by decreasing proximity. The colors indicate the p-value ranges obtained from the GEE fit for differences in distance to EMT hot/ EPI hot areas. Results are across 12 slides.

In addition, monocytes and tumour-associated macrophages (TAMs), such as LAM2 APOE+ macrophages and SIGLEC1+ macrophages, showed a marked likelihood to cluster closer to EMT hotspots compared with EPI hotspots. Monocytes and the TAMs derived from them are understood to modulate the environment of tumour cells undergoing EMT, often by promoting immunosuppression in the TME and thereby facilitating tumour progression and metastasis[214]. By contrast, Natural Killer (NK) cells, NK T cells, and CD8+ T cells, immune cells capable of directly killing transformed cells,

were among the least closely associated with EMT hotspots, suggesting a possible mechanism of immune evasion in EMT tumour cells[215]. The T-cell subset most closely aligned with EMT hotspots relative to EPI hotspots was the LAG3+ CD8+ T-cell population, an exhausted population, and suggesting immune evasion in these EMT regions[216].

Given the close relationship between EMT hotspots and potential immunosuppressive factors, I next evaluated whether EMT hotspots are indeed immunosuppressed. I observed significantly heightened expression of immunosuppressive and exhaustion markers[173,195] in EMT hotspots compared to EPI hotspots (**Figure 20**). Notably upregulated suppressive genes included FAP, which activates regulatory T cells (Tregs) and myeloid-derived suppressor cells (MDSCs)[217,218]; INHBA, which shifts macrophage polarisation to a pro-tumour state[219]; VCAN, linked to limiting T-cell proliferation[220]; and COL6A3, which is linked to increased macrophage recruitment[221]. Key immune checkpoints, B7-H3 (CD276), OX40 (TNFRSF4), and TIM3 (HAVCR2), were also significantly upregulated ($p<0.05$) in the tumour slides (**Figure 20**). In line with these findings, EMT hotspots exhibited elevated levels of immune exhaustion (**Figure 20c-d)**, supporting the idea that prolonged immune activation in EMT hotspots results in immune cell exhaustion, which may be reversed by checkpoint blockade. To investigate this hypothesis, I examined an interferon-gamma signature previously associated with favourable responses to immunotherapy[196,197] (**Figure 20h-i**). EMT hotspots showed notably increased expression of signature genes, including HLA-A and HLA-C, which are often linked to the activation of immune responses[222], as well as HLA-F, known for its immunosuppressive properties[223]. Enhanced expression of interferon-gamma-associated genes, especially those involved in antigen presentation (e.g., HLA molecules), is generally considered a positive prognostic indicator in checkpoint blockade therapy[224]. Hence, although EMT hotspots exhibit considerable immunosuppression and T-cell exhaustion, they simultaneously retain aspects of immune activity that could be harnessed by targeted treatments, such as checkpoint inhibitors.

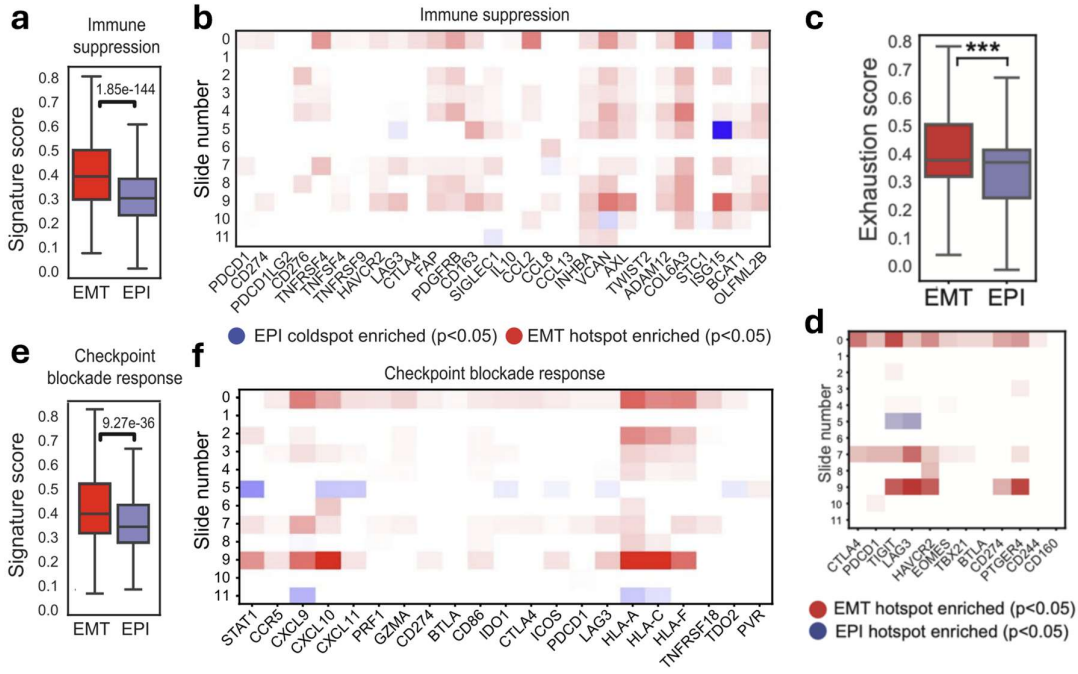

Figure 20.a. Barplots showing signature scores of immune suppression scored within EMT hotspots and EPI hotspots. **b**. Differences in the average expression of genes in the immune suppression signature between EMT and EPI hotspots for each slide (row). Red depicts genes significantly upregulated in EMT hotspots and blue indicates genes significantly upregulated in EPI hotspots (Student's t-test p<0.05, adjusted for multiple testing using the Bonferroni correction). White indicates a non-significant relationship. **c**. Similar to (a) but for immune exhaustion. **d**. Similar to (b) but focusing on the genes in the immune exhaustion signature. **e.** Similar to (a) but for checkpoint inhibitor response. f. Similar to (b) but focusing on the genes within the checkpoint inhibitor response signature

At the cohort level, the association between EMT hotspots and the myCAF s5 population persisted in individual tumours, suggesting a universal pattern of EMT transformation in breast cancer rather than a subtype-specific relationship (**Figure 16**). In contrast, substantial inter-patient heterogeneity emerged for multiple cell types, including macrophages, memory B-cells, naïve B-cells, iCAFs, NK cells, NKT cells, CD4+ T cells, and CD8+ T cells, often even within the same breast cancer subtype. However, distances to EMT hotspots were consistent across subgroups of cells (**Figure 21**), suggesting that, within individual patients, these cells share common response patterns irrespective of the broader heterogeneity observed across the patient cohort.

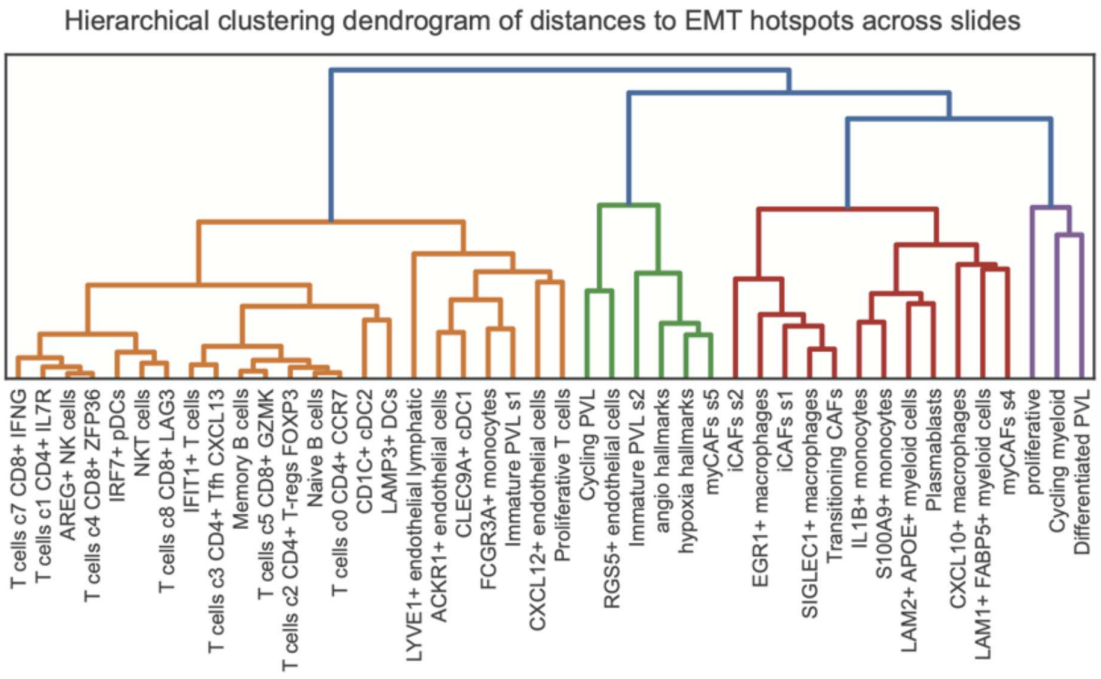

Figure 21 A dendrogram illustrating the hierarchical clustering of cells based on their proximity to EMT hotspots, as derived from the data shown in **Figure 16**.

3.3.5 EMT hotspots display intra- and inter-patient heterogeneity

I next sought to interrogate spatial relationships at a more granular level, and analysed the association of EMT hotspots with other immune and stromal areas within the same slide and across the different patient samples (**Figure 22a**). While cell types displaying the strongest relationship with EMT hotspots when averaged (such as SIGLEC+ and LAM2 APOE+ macrophages, along with CAFs) showed fairly consistent patterns across slides, some degree of heterogeneity was still evident. For instance, in slide 4, although seven EMT hotspots were closer to LAM2 APOE+ macrophages than the median EPI hotspots, two were not (**Figure 22a**). Visual inspection of these particular EMT hotspots, compared with LAM2 APOE+ macrophage hotspots, further highlights this variability (Figure 13b). As anticipated, stronger associations with myCAFs and specific macrophage subtypes were common across most EMT hotspots (**Figure 22a**). T-cells, although heterogeneous across patients, tended to cluster together, reinforcing the notion that these cells share similar responses. Notably, EMT hotspots that were closer to T-cells often showed elevated exhaustion marker expression (**Figure 22a,** right panel), suggesting ongoing immune activation. I also observed that

EMT hotspots consistently exhibited higher immunosuppressive scores than EPI hotspots (**Figure 22a**). Overall, the inter-patient heterogeneity seemed to supersede the intra-patient heterogeneity.

SpottedPy provides functionalities for examining distance distributions both within individual slides (**Figure 22c**) and across multiple slides (**Figure 22d**). Visualising these distributions illustrate that, although LAM2 APOE+ macrophages are generally positioned nearer to EMT hotspots compared with EPI hotspots, there is heterogeneity within each slide and across different slides. Overall, these results showcase the range of hotspot analyses enabled by the SpottedPy package and the potential to uncover useful biological insights.

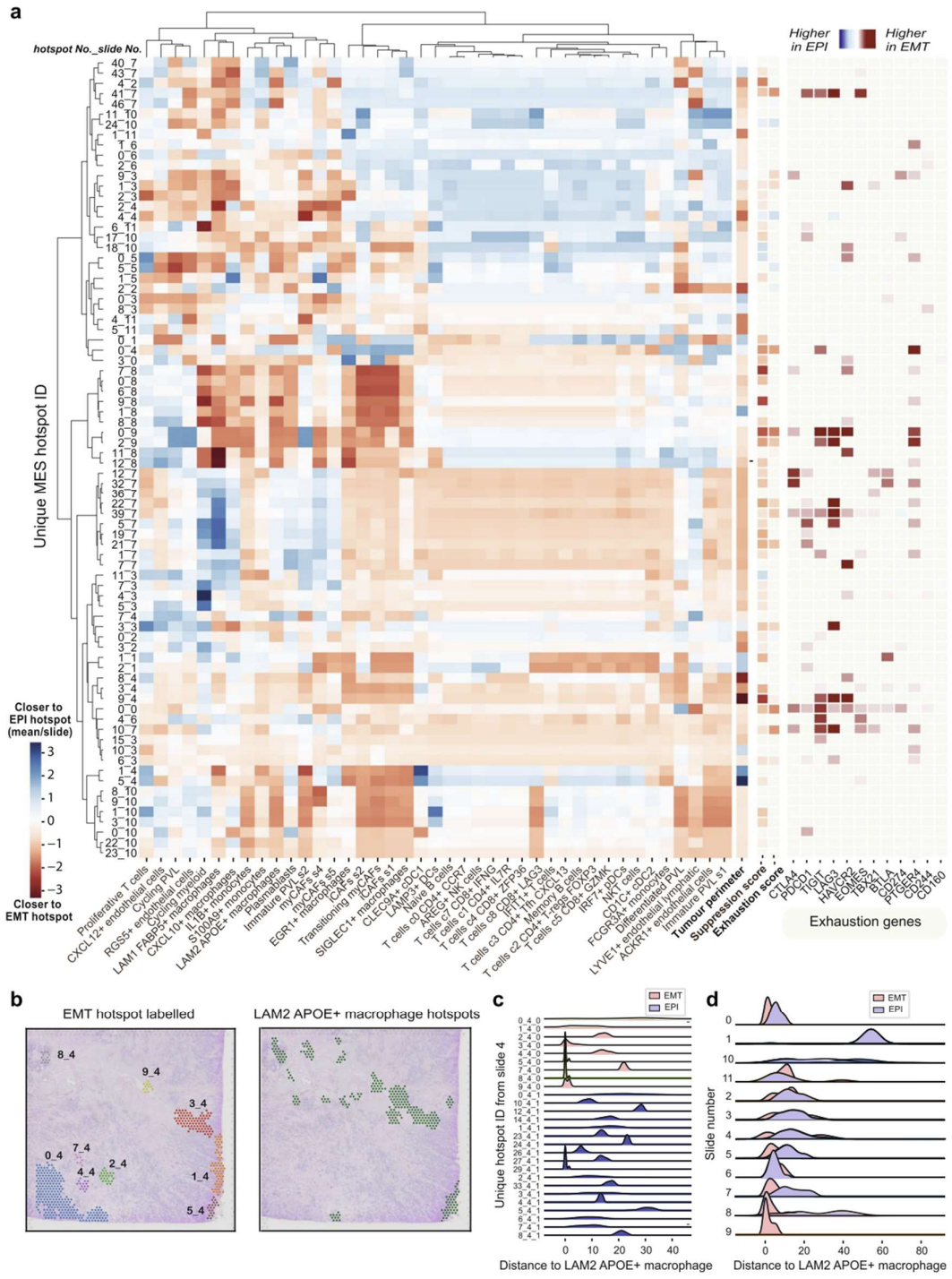

Figure 22 Inter- and intratumour heterogeneity of EMT hotspots. **a** Dendrogram highlighting the proximity of EMT and EPI hotspots to TME cell types. The dendrogram is clustered according to the distances from EMT/ EPI hotspots to regions enriched in immune/stromal cells. Red indicates that an EMT hotspot is closer to a cell type, while blue suggests that the EPI hotspots in that slide are on average closer. The x-axis displays individual EMT hotspots (label indicates hotspot number and slide number). To the right of the dendrogram, distances to the tumour perimeter, suppression, and

exhaustion signature scores are illustrated. Red indicates that the hotspot is significantly enriched in these signatures compared to the average EPI hotspot in the slide ($p< 0.05$), and blue indicates EPI hotspots are significantly enriched ($p<0.05$). Further to the right, individual genes associated with the exhaustion signature are shown, with red indicating the gene expression is higher in EMT hotspots ($p<0.05$) and blue indicating the gene expression is higher in EPI hotspots ($p<0.05$). **b** Slide 4 with individual EMT hotspots labelled (left) and LAM2 APOE+ macrophage hotspots highlighted (right). **c** Distance distributions for each EMT hotspot in slide 4 to LAM2 APOE+ macrophage hotspots. **d** Distance distributions of EMT hotspots to LAM2 APOE+ macrophages across all 12 slides in the cohort

3.3.6 Sensitivity analysis of hotspots

Determining how hotspot size, governed by the number of nearest neighbours parameter, influences spatial relationships is important for robust spatial analysis. I systematically expanded the hotspot dimensions (**Figure 23a-b**) to understand the stability and consistency of identified spatial associations. The results indicate that the spatial interplay among EMT hotspots, hypoxia, and angiogenesis, as well as their exclusivity with proliferative hotspots, remains as a resilient hallmark of the tumour microenvironment.

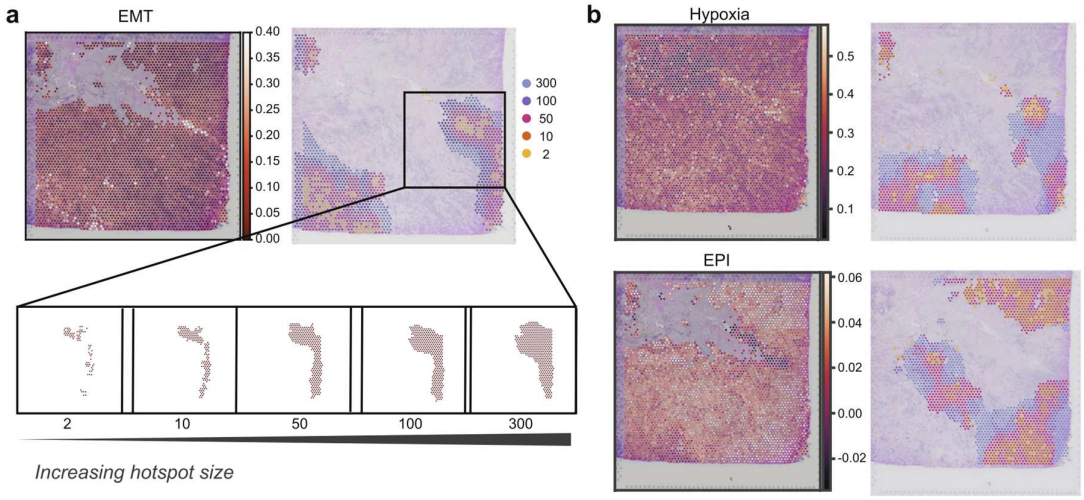

Figure 23. Sensitivity analysis of hotspot relationships. a EMT hotspot generation using a hotspot neighborhood parameter of 2, 10, 50, 100, and 300, respectively. Increasingly larger neighborhoods are highlighted in diferent colors as indicated in the legend. b Hypoxia and epithelial hotspot generation using a hotspot neighborhood parameter of 2, 10, 50, 100, and 300, respectively

These patterns remain notably stable across different hotspot sizes (**Figure 24**). Cell populations previously identified as being closest to EMT hotspots at a fixed parameter size (myCAFs, macrophages, and monocytes) also maintained this relationship when

the hotspot dimension changed. Conversely, relationships for cells situated farther away were less consistent, for example, CD8+ LAG3+ T-cells no longer retained their association at a hotspot size of 250, and naïve B-cells showed multiple shifts as the hotspot size increased (**Figure 24**).

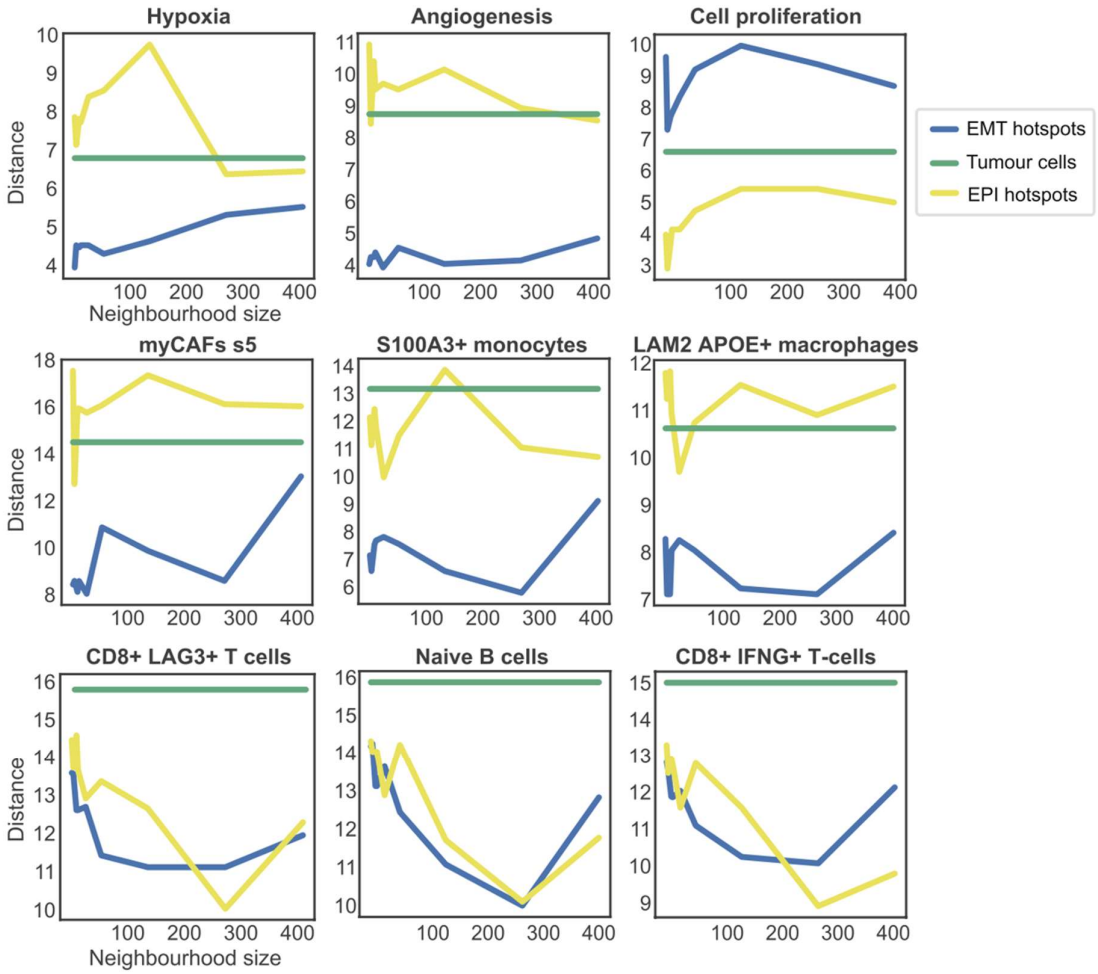

Figure 24. Sensitivity plots highlighting the distance from EMT hotspots (blue) and EPI hotspots (yellow) to regions enriched in various cancer hallmarks and TME components as the hotspot size increases. The distances to the average of all tumour cells are used as a reference (green). Distances are averaged over all 12 slides.

These findings suggest that interactions with certain cells in the TME may be more pronounced and relevant at a smaller scale. I found that proliferative hotspots were the most consistently adjacent to EPI hotspots at various hotspot sizes. Adjusting the

p-value threshold for Getis-Ord Gi* cluster detection yielded similar patterns (**Figure 25**).

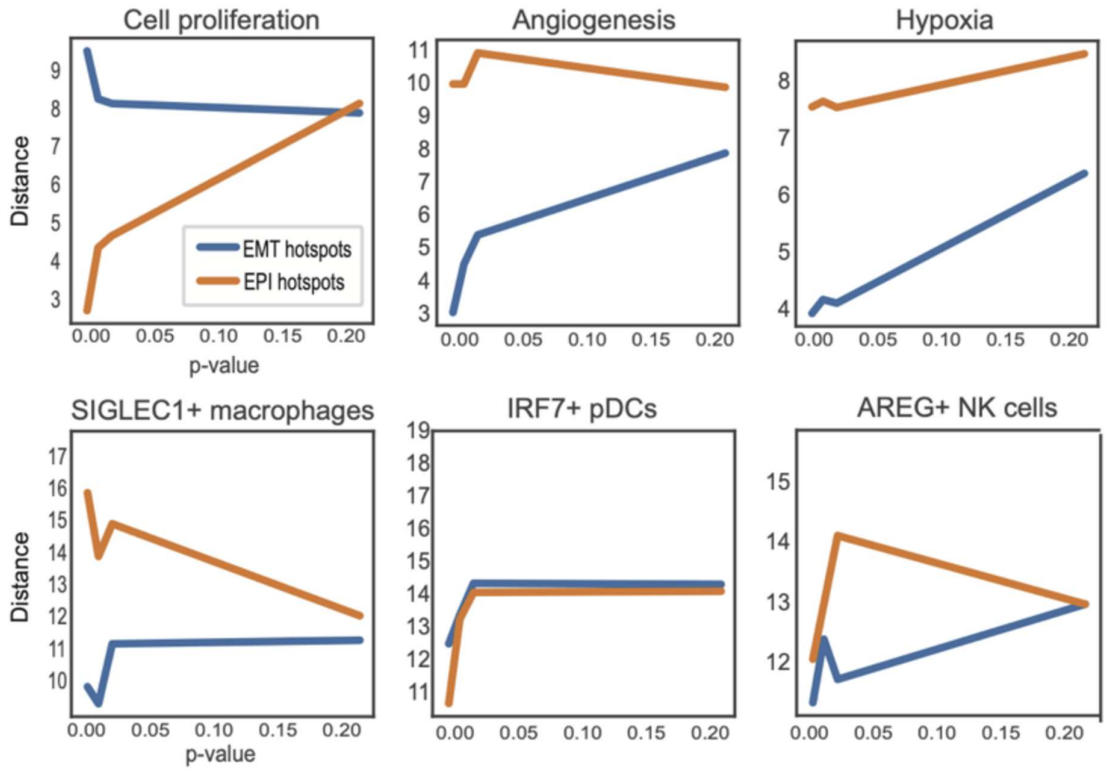

Figure 25. Sensitivity plots highlighting the distance from EMT hotspots (blue) and EPI hotspots (orange) to various TME components as the p-value parameter using to detect statistically significant hotspots increases. Distances are averaged over all 12 slides.

To further assess the robustness of these spatial relationships, I introduced Gaussian noise and performed spot reshuffling, then examined whether the identified relationships persisted (see Methods). This approach demonstrated that SpottedPy reliably distinguishes genuine biological signals from random spatial fluctuations, tolerating low noise levels effectively **(Figure 26)**.

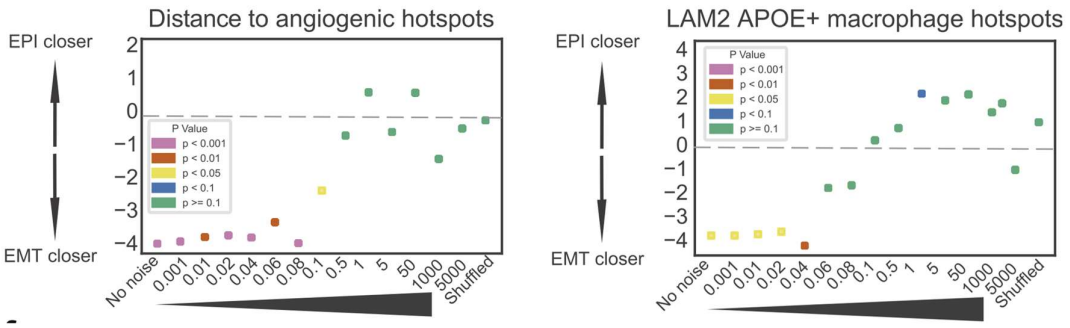

Figure 26. Evaluating the impact of increased noise or spot shuffling on the association between angiogenic hotspots (left) and LAM2 APOE macrophage hotspots (right) and EMT/EPI hotspots.

Although random noise generated through spot reshuffling can mimic certain aspects of structured data (**Figure 27**), the resulting hotspots were significantly smaller than those arising from genuine biological structure (**Figure 28)**.

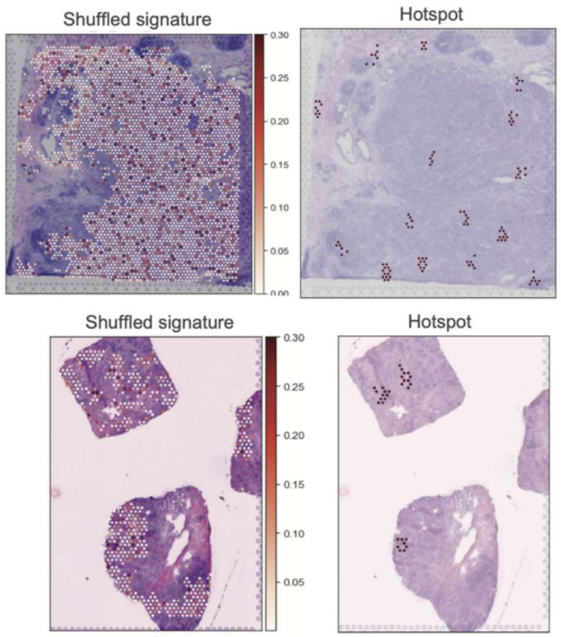

Figure 27. Spatial plots of shuffled EMT signature (left) and hotspots produced from shuffled signature (right) for Slide 6 (top) and Slide 7 (bottom).

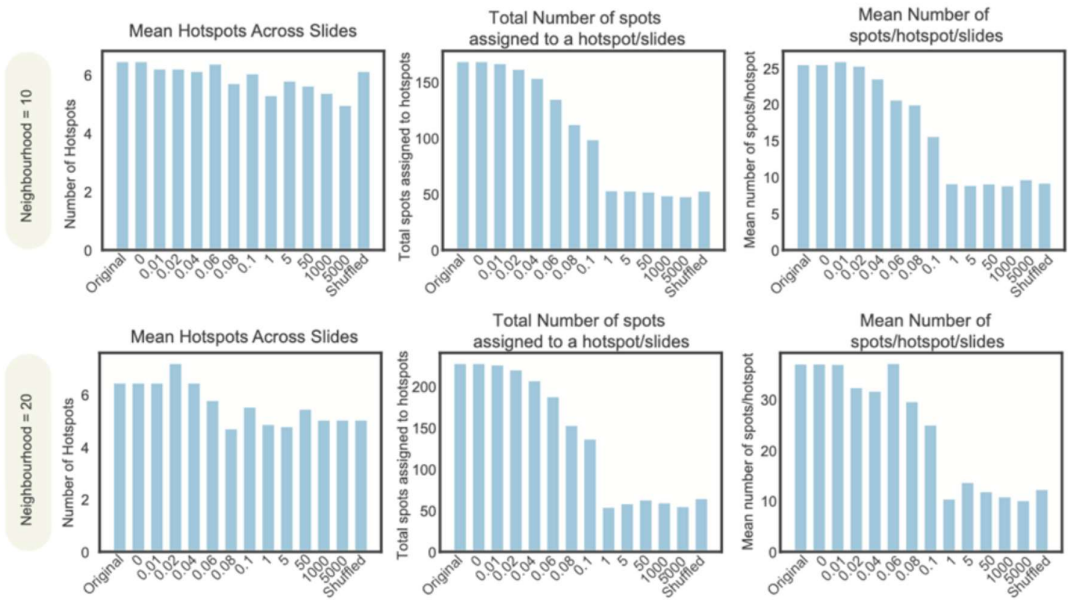

Figure 28. Frequency bar plots for neighbourhood size=10 hotspots (top) and neighbourhood size=20 hotspots (bottom), highlighting the average number of hotspots across a slide (left), total number of spots assigned to a hotspot/slide (middle) and average number of spots/hotspot/slide (right) with increased amounts of noise added to the EMT signature, or spot shuffling.

Crucially, the loss of specific associations among particular cell types when noise is introduced **(Figure 26)** helps to reduce false positives, even in instances where hotspots are identified[225].

3.3.7 Other distance metrics

Importantly to note, there are other approaches that can be used to calculate hotspot distances. For instance, the “centroid to centroid” approach offers a straightforward approximation but is inherently simplistic. As illustrated in **Figure 29**, hotspot size exerts a major influence on its centroid, meaning that larger hotspots might falsely appear farther away when measuring centroid distances, despite actually being closer at the perimeter level. Consequently, centroid-based distance metrics can overlook local variations that can be captured by the shortest-path approach. Applying a centroid-based method to the breast cancer slides might therefore fail to detect more complex spatial patterns, such as those between EMT hotspots and macrophage- or monocyte-enriched areas.

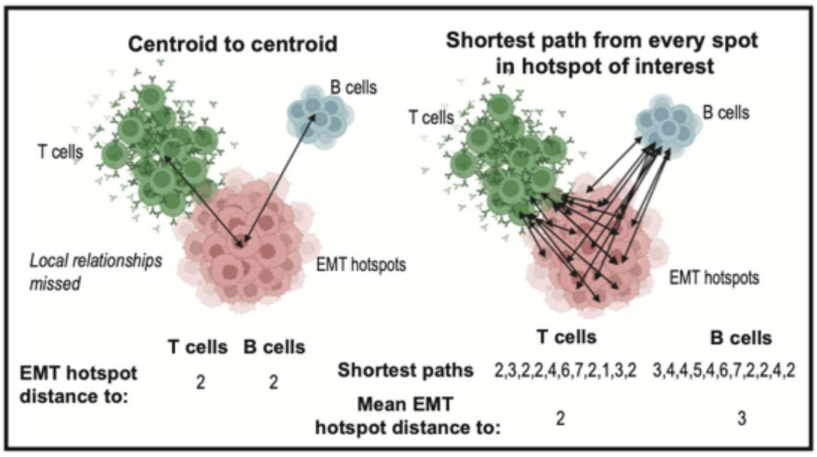

Figure 29. Diagram highlighting key differences between the centroid distance approach and the shortest path distance approach. The shortest path approach captures more local variation and therefore detects a difference between the Tcell and B-cell hotspots shown in this example (distances shown for illustration purposes). The centroid-to-centroid approach however would be unable to capture this.

3.3.8 Spatial EMT relationships in other cancer types

I then explored whether the spatial associations identified for EMT hotspots in breast cancer also manifest in other cancers. I therefore examined publicly available data from basal cell carcinoma (BCC) [186], pancreatic ductal adenocarcinoma (PDAC) [187], and colorectal cancer (CRC)[188].

In BCC, angiogenic and hypoxic hotspots were located nearer to EMT hotspots **(Figure 30a-b).** Interestingly, proliferative hotspots were also found closer to EMT hotspots, implying significantly different relationships compared to breast cancer. POSTN+ fibroblasts showed significant proximity to EMT hotspots, while T-cells and NK cells displayed no pronounced spatial relationships with EMT, mirroring observations in breast cancer.

I then investigated these relationships in one available PDAC slide **(Figure 30c)**, where I found that angiogenesis and fibroblasts were spatially correlated with EMT hotspots in a manner similar to breast cancer. In contrast, immune cells were more often found adjacent to EMT coldspots, and there was no discernible link between EMT hotspots and hypoxia, diverging from breast cancer patterns.

In CRC, myofibroblasts, angiogenesis, and hypoxia showed comparable spatial relationships to those seen in breast cancer **(Figure 30d-e).** Regulatory T-cells, T-helper cells, and NK cells were however closer to EMT hotspots, suggesting enhanced immune recognition in these regions relative to other cancer types.

Although the sample size and cell-type granularity were limited, these observations imply that tissue-specific factors may govern how EMT interacts with immune and stromal cells in the TME. Further investigation is warranted to illuminate these differences.

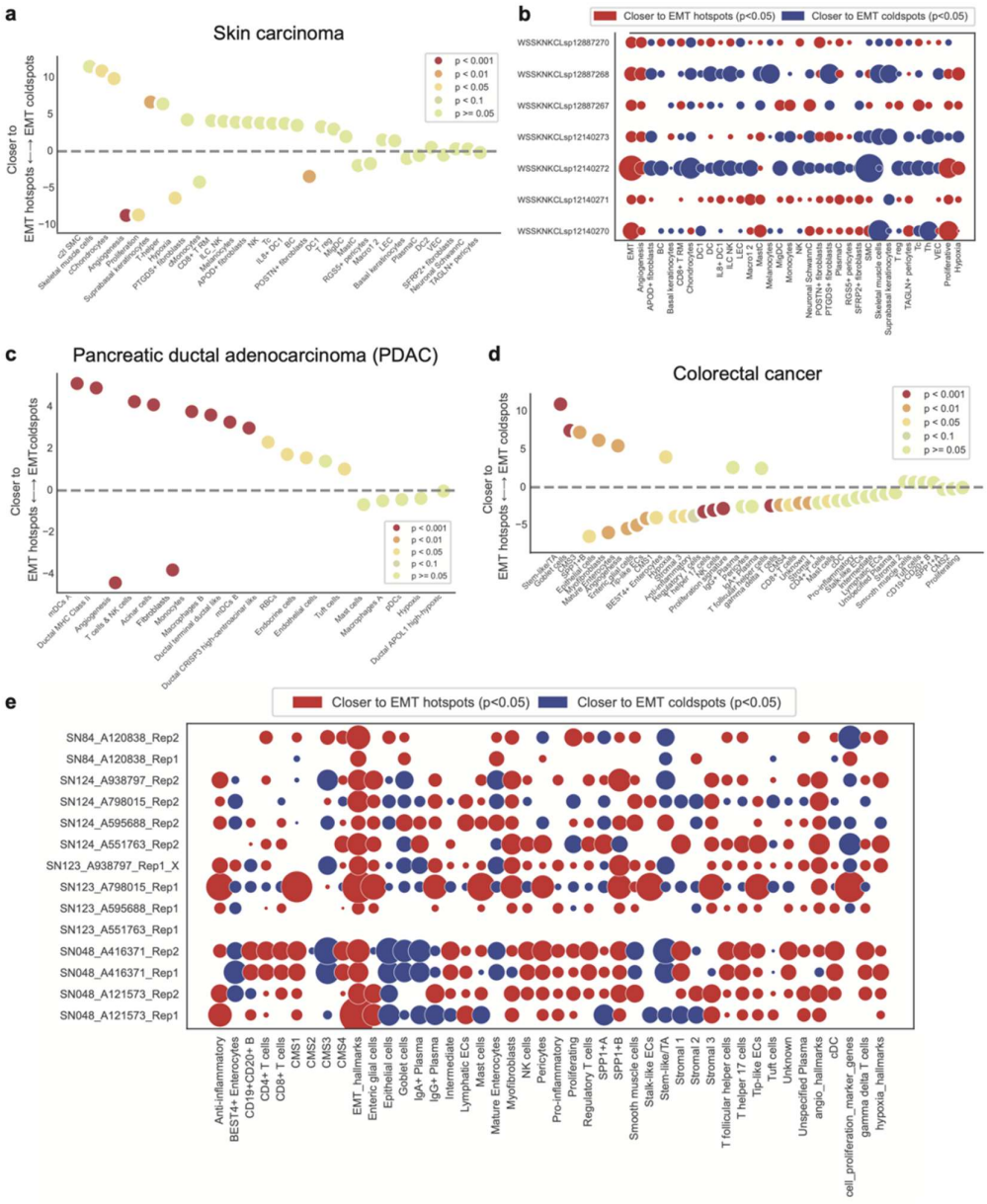

Figure 30. EMT hotspot analysis in other cancer types. a Distances from various cells in the TME to EMT hot/cold regions in basal cell skin carcinoma. The dashed line represents no difference in proximity to either EMT hotspots or EMT coldspots. The dots situated to the right of the dashed line indicate cell populations that are significantly closer to EMT hotspots, ordered by decreasing proximity. The colours indicate the p-value ranges obtained from the GEE model fit. **b** Bubble plot depicting distances between cancer hallmark signatures and TME classes and EMT hotspots/coldspots for each BCC slide (row). Blue depicts hallmarks that are significantly closer to EMT coldspots and red represents hallmarks that are significantly closer to EMT hotspots (Student's t test $p<0.05$), adjusted for multiple testing using the Bonferroni correction. White indicates a non-significant relationship. **c** Similar to (a) for one PDAC sample. **d** Similar to (a) for colorectal cancer slides. **e** Similar to (b) for the colorectal cancer slides.

3.3.9 Neighbourhood enrichment analysis

The neighbourhood enrichment technique captures more localised, shorter-range relationships with the TME. Additionally, it can assess spatial relationships of phenotypes that would be considered scattered (states that do not occur spatially clustered and therefore might be overlooked by a hotspot-based approach). I experimented with two approaches, ensuring a robust analysis that is less sensitive to the MAUP (**Figure 11bi-ii)**. I first assessed how the spatial relationships change by correlating phenotypes across a central tumour spot and the direct neighbourhood surrounding it (a ring encompassing six Visium spots). I then assessed how the phenotypes are linked within a spot and then expanding what is considered a spatial spot**.** Varying the method and the number of rings in both cases enables me to assess whether the observed hotspot relationships shift with the unit of analysis and indicates how large of an influence the EMT regions have on surrounding spots.

My analysis highlights that angiogenesis, myCAFs, macrophages, and monocytes had the strongest correlations with EMT cells ($p<0.001$) across the 12 slides (**Figure 31a**). This result corroborates the spatial interactions previously identified using the hotspot approach. By contrast, naïve B-cells, T-cells, NK cells, and NKT cells showed weaker associations, consistent with the hotspot analyses. These spatial patterns remained stable across multiple neighbourhood sizes (**Figure 31b**).

The methods show broadly similar trends, suggesting the cellular relationships observed occur both due to colocalisation in a spot as well as diffusing influence around the spot.

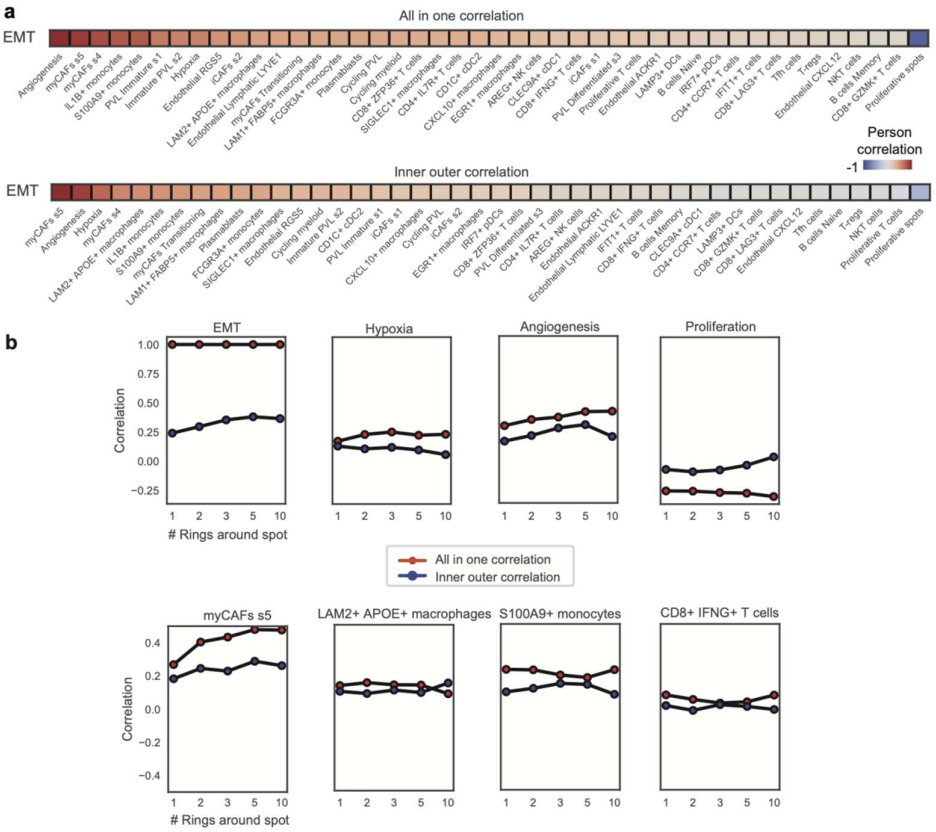

Figure 31. Neighbourhood enrichment analysis of EMT spatial dynamics. a Neighbourhood enrichment analysis results employing all in one correlation (top) and inner outer correlation (bottom) approaches. The squares display correlations between EMT levels within tumour cell spots and the abundance of various cell populations within the immediate TME (surrounding spots only). Red indicates a positive correlation, blue a negative correlation and white a non-significant correlation (Pearson p>0.05). 1 ring is used to define the neighbourhood. **** p<0.0001, ***p<0.001, **p<0.01, *p<0.05. **b** Line plots illustrating the impact of progressively expanding the number of concentric rings - from 1 to 10 - around a Visium spot on the correlation between the EMT signature and various cells in the TME. Each ring represents an incremental distance from the central spot and encompasses the surrounding spatial transcriptomic spots.

3.3.10 EMT state fluctuations shape distinct immune niches within the same tumour

Since EMT is not a binary switch but rather a spectrum of hybrid states, I sought to investigate the spatial distribution of tumour hotspots reflective of epithelial (EPI), early intermediate (EM2, EM3), late intermediate quasi-mesenchymal (M1), and fully mesenchymal (M2) states using the multi-scale framework. To identify these states, I

used Non-Negative Matrix Factorisation (NMF) via the CoGAPs workflow[192] (**Figure 32a**).

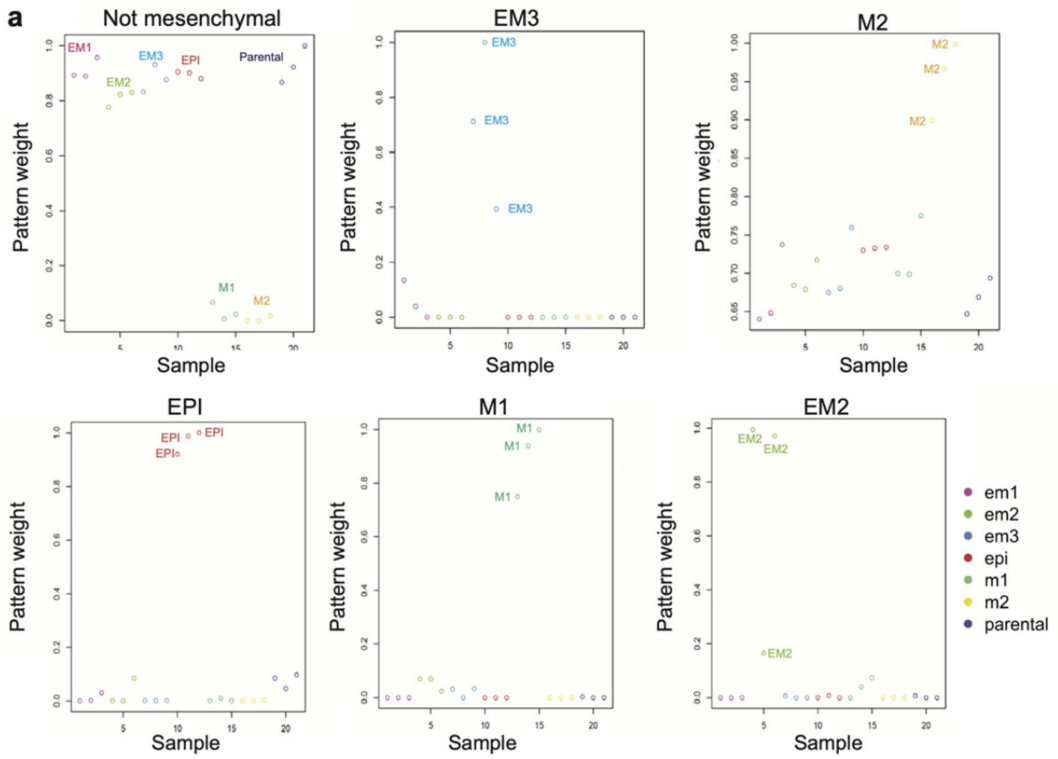

Figure 32. EMT state capture in spatial transcriptomics slides. **a** NMF patterns captured from Brown et al[166], consisting of seven RNA-seq sequenced cell clones, with three repeats spanning the EMT spectrum including epithelial-like (EPI), quasi-mesenchymal (M1), fully mesenchymal (M2) and three distinct intermediates (EM1, EM2, EM3). Each circle corresponds to one cell clone from the original dataset, and is coloured according to the assigned state. The pattern weights for each cell clone are plotted for each pattern. The patterns that were able to separate the cell clones are annotated.

The corresponding hotspots occupied distinct spatial locations within the tissue **(Figure 33**).

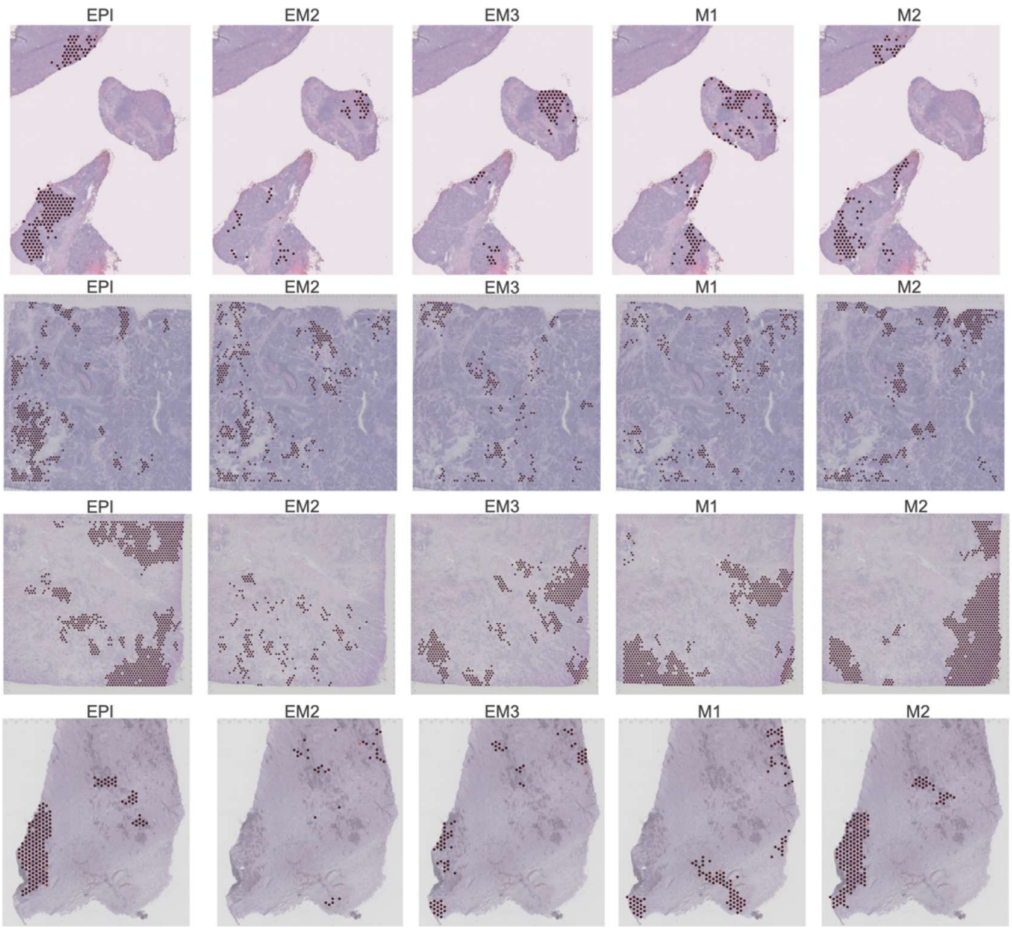

Figure 33. Spatial plots highlighting the distribution of EMT state hotspots (EPI, EM2, EM3, M1 and M2). Every row corresponds to one Visium slide.

Further visual inspection of these hotspots reveals a progressive shift in the tumour, as highlighted in Slide 4 (**Figure 34**). This transformation is characterised by a transition from EPI to the M1 state, with EM3 serving as an intermediate step. EM2 appeared more volatile in this progression, whereas M2 predominantly co-localised with the EPI state. The experimental study by Brown et al[166] had detected that M2 cell clones gained integrin β4 (a key epithelial marker) when cultured, which might have played a significant role in steering these cells towards adopting characteristics more akin to an epithelial phenotype. This would possibly explain the co-localization of these two states within the spatial transcriptomics slide.

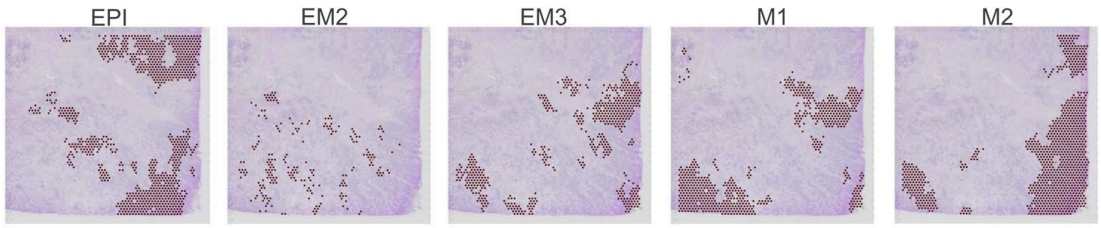

Figure 34. Spatial plots highlighting the distribution of EMT state hotspots (EPI, EM2, EM3, M1 and M2). Every row corresponds to one Visium slide.

To further investigate these states, I examined their correlations with one another and with the EMT hallmark signature **(Figure 35a**). The absence of positive correlations among EPI, EM2, EM3, M1, and M2 supports the notion that they represent distinct EMT states. The M1 state correlated strongly with the EMT hallmark signature, while the EPI state correlated negatively, as anticipated. Spatially, the EMT hallmark hotspots were nearest to the M1 hotspots and most distant from EPI hotspots (**Figure 35b**), in line with the correlation findings and confirming the hypothesised identities of these states.

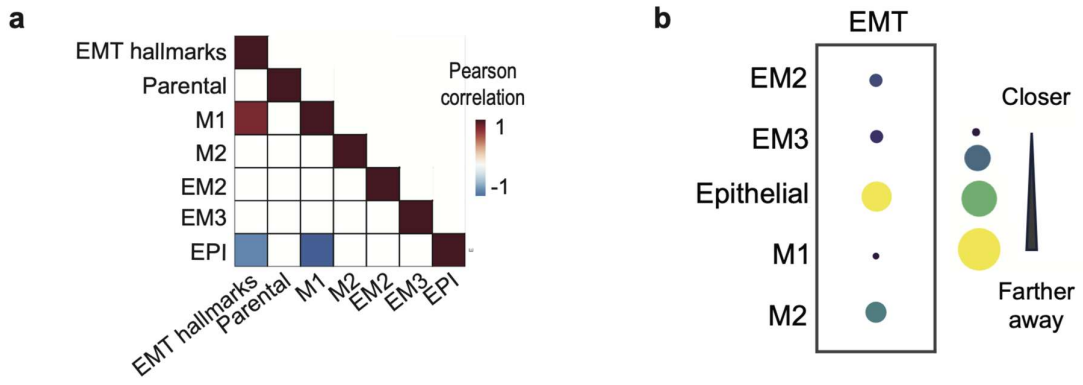

Figure 35. **a** Correlation plot of EMT state scores across all spots and slides. Red indicates positive correlation; blue indicates negative correlation. Only significant correlations ($p < 0.05$) are shown. White squares indicate non-significant correlations. **b** Bubble plot depicting the mean distance from individual EMT state hotspots to the EMT hallmark hotspots defined in the original analysis. Smaller bubbles represent shorter distances.

I then examined how tumour cells occupying these discrete EMT states relate to immune and stromal components of the TME. The analysis found that the EPI state correlated negatively with TME cells (**Figure 36a**), implying a phenotype that is not

being substantially shaped by its surroundings. Interestingly, the M1 state exhibited robust associations with many TME populations, including strong links to myCAFs, macrophages, and monocytes. The EM3 state showed moderate but still notable correlations, while the EM2 state displayed the weakest. This reduction in observed cellular interactions in the EM3 and EM2 states is in line with the idea of these states representing intermediate, more plastic states preceding the apparently more stable M1 state. Of particular interest, M1 demonstrated proximity to NK cells, a departure from the broader EMT hallmark signature. This suggests that, although M1 may resemble the general EMT state in some respects, it likely represents a specialised phenotype capable of drawing in cytotoxic immune cells.

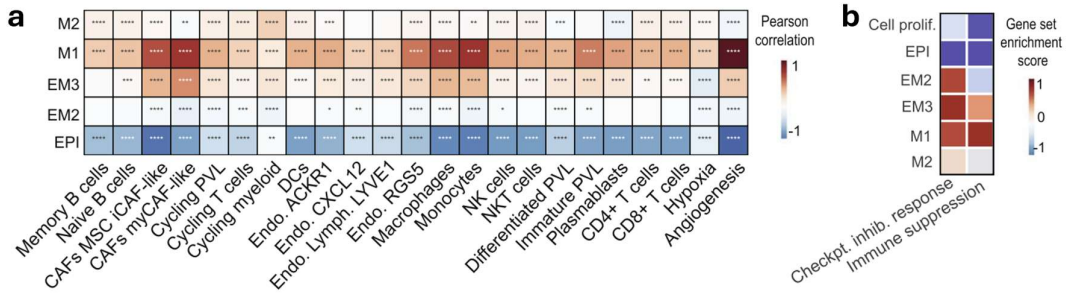

Figure 36. **a**. Neighborhood enrichment analysis depicting the association between tumour cells occupying distinct EMT states and other cells in the immediate TME, summarized across all 12 slides. Red indicates a signifcant positive correlation (Pearson, $p < 0.05$) , blue a signifcant negative correlation ($p<0.05$), and white a non-signifcant correlation ($p>0.05$). **** $p < 0.0001$, *** $p < 0.001$, **$p<0.01$, *$p<0.05$. **b**. Scaled immune suppression and immunotherapy response signature [65] scores calculated using Gene Set Enrichment Analysis (GSEA) for each EMT state hotspot and proliferative hotspot, summarized across the 12 samples.

Moreover, the quasi-mesenchymal M1 state was enriched for markers associated with immunosuppression and positive response to checkpoint inhibitors (**Figure 36b**), most notably with OX40 (TNFRSF4), TIM3 (HAVCR2), HLA-DRA, CXCL9, and CXCL10 (**Figure 37a-b**). The intermediate states (EM2, EM3) appeared partially committed to this suppressive phenotype, showing weaker yet progressively stronger associations leading towards M1. By contrast, M2 had a unique signature, exhibiting mixed positive and negative relationships within these gene sets. When I compared these findings to the proliferative signature, I found that proliferative hotspots mirrored the EPI state, suggesting that they represent a tumour phenotype not linked to immunosuppression.

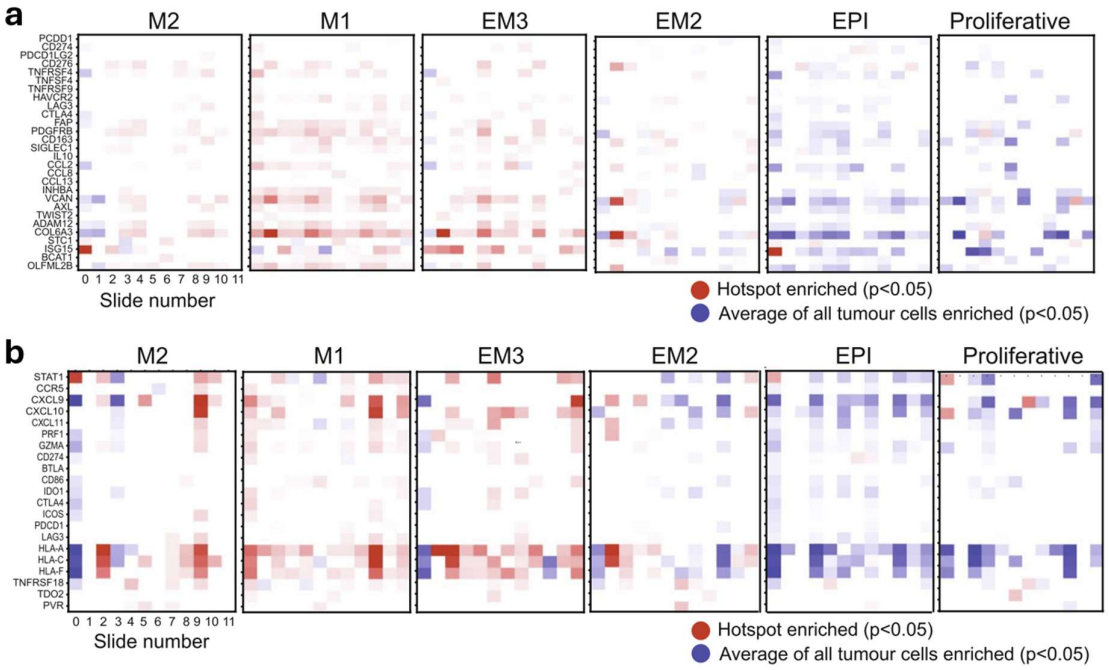

Figure 37. **a**. Enrichment and depletion of expression for genes in the immune suppression signature within EMT state hotspots for each slide (column). Red depicts genes signifcantly upregulated in EMT state hotspots compared to the average of all tumour cells and blue represents genes signifcantly downregulated in the EMT state hotspot (Student's t-test $p<0.05$). White indicates a non-signifcant relationship. P-values were adjusted for multiple testing using the Bonferroni correction. **b** Similar to (a), focusing on genes in the checkpoint inhibitor response signature

Overall, the results in this Chapter highlight the changing landscape of tumour-TME interactions during EMT progression in breast cancer, highlighting both intratumour heterogeneity as well as universal interactions that could be exploited for therapy.

3.4 Discussion

In this chapter, I introduce SpottedPy, a Python package for identifying tumour hotspots in spatial transcriptomics slides and examining their interactions with TME at multiple scales. I demonstrate that the Getis-Ord G* statistic can be applied successfully to identify cellular hotspots, yielding biologically meaningful insights into the spatial organisation of tumour tissue within its immune and stromal context. Although several recent studies have used variations of hotspot analysis on spatial transcriptomic data[226,227,178], these methods generally do not provide a way to assess the confidence level in identifying specific clusters or hotspots. By contrast, my approach assigns a p-value to each hotspot, allowing users to adjust stringency to suit their analytical needs. Furthermore, unlike most existing tools, SpottedPy conducts in-

depth distance analysis between hotspots. I also extend previous hotspot-based strategies by offering a statistically principled assessment of spatial relationships, with the added ability to anchor hotspot detection to specific regions, such as tumour versus non-tumour areas, to investigate TME dynamics more precisely. By calculating and statistically comparing distances, SpottedPy generates an interpretable and intuitive measure of spatial relationships. This approach also enables differential spatial analysis against a reference region, a capability generally absent from other packages. Moreover, by exploring how hotspot size affects spatial patterns and contrasting these findings with relationships identified through neighbourhood-based methods, SpottedPy integrates multiple layers of spatial evidence, and scale, into a single analytical framework.

By applying SpottedPy to examine tumour plasticity phenotypes in breast cancer, I uncover pronounced differences between tumour areas undergoing EMT and more epithelial regions of the tumour. My findings reveal robust spatial correlations of EMT with key cancer hallmarks, particularly hypoxia and angiogenesis, consistent with the work of He et al[120] in spatial transcriptomics of breast cancer, which detected these signatures overlapping in certain niches. As tumour cells undergo EMT in response to hypoxic stimuli, they can gain a survival advantage in nutrient-deprived conditions and migrate towards better-oxygenated regions, potentially following angiogenic gradients[228,205,208].

I also observe a strong association between EMT and myCAFs across all analysed slides. This aligns with previous bulk and spatial transcriptomic findings in which CAFs appear linked to tumour cells undergoing EMT[48], and corroborates evidence showing CAFs can induce EMT in endometrial cancer cells[72] and hepatocellular carcinoma[78]. It is worth noting that CAFs share certain genetic markers with EMT signatures, making the two difficult to distinguish, particularly in bulk tumour datasets[229]. In this work, I address that challenge by using whole transcriptome data labelled with EMT states for deconvolution and copy number aberration detection, thereby increasing the confidence of detecting EMT in tumour cells. However, these strategies do not guarantee perfect discrimination between CAFs and EMT tumour cells; future research with single-cell-resolved spatial transcriptomic platforms will help clarify this relationship further and I will address this in Chapter 4.

Beyond the observed relationship with myCAFs, my analyses show strong associations of EMT with macrophages and monocytes across a range of spatial scales. In particular, SIGLEC+ macrophages, LAM2 APOE+ macrophages, and EGR1+ macrophages, populations akin to M2-like, tumour-promoting macrophages[173], are situated closer to EMT hotspots. These macrophages secrete TGF-β, TNF-α, IL-6, and IL-8, well-established inducers of EMT[230,231]. This relationship has previously been noted in bulk transcriptomics[77], in spatial studies of murine skin carcinoma[232], and within certain breast cancer niches[120]. Although CAFs and macrophages are broadly linked to EMT in most slides, I do find instances in individual slides where some EMT hotspots do not reflect these relationships, indicating the possible influence of local factors beyond my current analyses.

I found heterogeneity in the interaction of EMT hotspots with other immune cells such as NK, NKT, and T cells. While T-cells have been reported to induce EMT in breast cancer[233,234] and this relationship has been observed in bulk transcriptomics[235] and smaller scale spatial analyses[48,236], there is also evidence of T-cell exclusion tied to the relationship of EMT with macrophages and CAFs, fostering an immune-suppressed nice,[213,237]. My analyses highlight that EMT hotspots do feature notable enrichment for immunosuppressive and checkpoint-therapy-associated signatures, consistent with previous work positing EMT as an important factor in immunotherapy strategies[85,120].

I also show that EMT hotspots form in discrete locations that are spatially separate from proliferative signatures. This aligns with Jia et al.[51], who used a more focused spatial transcriptomic dataset, Tsai et al.[85], who demonstrated that a departure from a mesenchymal-like state is a prerequisite for tumour cell proliferation in mouse models, as well as Chen et al.[86], who identified similar trends in scRNA-seq data. Such spatial characterisations at various scales were largely unexplored.

I observed that hybrid EMT states exhibit more heterogeneous and weaker associations with the TME in comparison to the quasi-mesenchymal M1 state. This may reflect the intrinsic plasticity of these transitional states[36,238], complicating the ability to delineate clear relationships, but might also suggest a directed trajectory towards an M1 state. Conversely, the M2 state had more similar distribution and TME associations to the EPI state, which may be due to the activation of integrin β4 (a key

epithelial marker) when cultured, a limitation mentioned in the original study which potentially steered the state towards a more epithelial phenotype[166].

Overall, my findings point to a highly dynamic and plastic nature of tumour cells as they engage with a complex TME. The interactions likely extend beyond a linear framework. Hypoxia, a known inducer of both angiogenesis and EMT[239], may initiate a cascade that not only intensifies these processes but also draws immunosuppressive cells like macrophages[240,241], further reinforcing angiogenesis and forming a self-sustaining loop. These results provide a further understanding of the cellular interactions and environmental factors that support tumour progression and metastasis, highlighting opportunities for targeted interventions that disrupt this cycle to achieve therapeutic benefit.

The consistency of spatial associations across different hotspot sizes and neighbourhood scales adds confidence to the robustness of these observations. The neighbourhood ring approach predominantly detects TME cells that have infiltrated the tumour, capturing immediate tumour–immune interactions. By contrast, the hotspot approach offers a wider perspective, incorporating longer-range spatial influences. Statistically defining cellular hotspots enhances the reliability of observations, particularly considering the inherent inaccuracies that can arise from identifying cell states using deconvolution algorithms applied to non-single cell transcriptomic datasets such as those from the Visium platform.

To my knowledge, there is no direct alternative to the SpottedPy method, given its unique capacity to focus on user-defined continuous signatures within discrete spatial clusters at muliple scales, conduct differential spatial relationships against a reference region, and employ downstream analyses.

Overall, this chapter confirms expected spatial effects of EMT progression in tumours, demonstrating that SpottedPy can capture complex associations between tumour cells and their microenvironment. Such insights can help unveil local effects of the TME and linked tumour cell vulnerabilities that could ultimately be exploited for therapeutic benefit. While the analyses presented here primarily illustrate insights into breast cancer tissue organisation, I note that SpottedPy can be applied to discern spatial relationships in other cancer types as well as other diseases and even within healthy tissue. For example in Pan *et al*. (2024)[168], I applied SpottedPy to characterise the

spatial relationships of gene signatures from a large language model developed to predict EMT states. In Celik *et al*. (2024)[169] I have also used SpottedPy to spatially characterise quiescence in breast cancer. SpottedPy has been developed on spatial transcriptomics data from the 10x Visium platform; however, it can be easily extended to other spatially-resolved platforms.

4 Chapter 4: Spatial predictive modelling of epithelial to mesenchymal plasticity in cancer

In earlier chapters and in many spatial transcriptomic studies, associations between the environment and cell types are often descriptive. These studies mainly focus on identifying recurring niches or calculating the proximity of specific cells to other cell states. However, a standardised, quantitative metric to measure the influence of the spatial environment on a cell type or biological process of interest has not been widely used. Such a metric could provide a statistically robust framework to assess the significance of environmental factors on cellular states or behaviours. It could also allow for direct comparisons with other variables, such as genomic features, to rank their relative contributions and importance. In this chapter, I develop a framework for quantifying spatial effects and comparing them to other cell intrinsic variables. This helps us to further understand epithelial to mesenchymal plasticity and offers a method of assessing other plastic programmes.

I will first review the literature on cell plasticity and spatial modelling and highlight the aims of the work (Section 4.1). I will then explain the preparation of the Xenium dataset used and the modelling approach (Section 4.2). Given the methods I will describe the results of the GNN approach (Section 4.3.2) before highlighting the results from the geostatistical approach (Section 4.3.3). I will discuss the results from this chapter in Section 4.4.

4.1 Introduction and Literature Review

4.1.1 Cell Plasticity

Plasticity is the ability of a cell to change its properties without it being directly due to its alterations in the genome[44]. Evolutionarily, it is an important cell trait, as it allows cells to survive under environmental stress, playing vital roles in processes such as wound healing[242]. However, it is often aberrantly activated in cancer, with growing evidence to suggest that it is a driving force in hard-to-treat cancers[243]. The ability of the cancer cell to change state means that drugs targeted for a specific cancer state can become redundant as the cell can change out of the original state[243]. In epithelial

to mesenchymal plasticity (EMP), epithelial cells can for example transition to a mesenchymal state where they evade therapy. Plasticity allows the cancer cells to adapt to the every-changing cancer microenvironment, for example nutrient scarcity, which otherwise would lead to cell death[244].

Cancer cell plasticity is an emerging field, with a lack of a widespread consensus on the definition of cell plasticity[245]. Definitions range from including cell type change to cell state change[44]. A cell type refers to a stable functional unit within an organism with established biological functions, whereas a cell state is considered a more dynamic and transient change. Most definitions emphasise that a key part of cell plasticity is the ability of a cell to change its cell type due to external stimuli. Some definitions stress that plasticity should be a quantifiable metric, measuring how easy a cell can change from its steady-state identity[44]. A recent study understanding plasticity in lung cancer defined plasticity "as the potential of a cell to manifest diverse future fates[246]". The epigenetic nature of a cell and the degree of epigenetic priming was key to this definition of cell plasticity.

Cell state changes can also be referred to as phenotypic plasticity[247]. Here, a cell's phenotype is defined as the features of a cell that treatments would target, for example uncontrolled growth or immune evasion[247]. It is important to note the distinction of phenotypic plasticity from phenotypic noise, another mechanism that can cause cells to change state that is not directly linked to alterations in the cell's genome[248]. Phenotypic noise is a difference in phenotype that occurs, despite a shared genotype and environment, from stochastic changes in transcripts. It is believed to be a more transient change, but has been shown to affect cell phenotypes[249–251]. This is suggested to be an important evolutionary trait of cell types as it can ensure variability that may confer selective advantages under different conditions[248]. This timescale differs to methods such as chromatin modification, which would be considered a more long-term method for plasticity as it can propagate through subclones[248].

Importantly, phenotypic plasticity and phenotypic noise can be enhanced through genetic changes. This can occur through a mutation increasing a cell's ability to exhibit phenotypic noise, or increasing its likelihood to shift in response to an environmental variable, such as within chromatin modifier genes allowing for more epigenetic

changes[252]. This interplay between intrinsic (genetic and stochastic) and extrinsic (environmental) factors highlights how complex it can be to define cell plasticity.

At a more fundamental level, these factors driving plasticity work by altering nucleotide sequences, epigenetic modifications, 3D DNA structure modifications, alongside regulatory mechanisms such as altering transcriptional machinery[253]. These can be experimentally measured to understand plasticity. Additionally, common approaches to assess plasticity can work on readouts of cell phenotypes using either RNA transcripts or high throughput imaging[248]. A wide range of experimental approaches have been developed to understand plasticity. These include assessing plasticity using immortalised cell lines, iPSC cell lines, organoids, model organisms, mammalian models and biopsies[44]. There is a trade-off between achieving greater experimental flexibility, as seen with systems like cell lines, which often sacrifice the accuracy of the TME representation, and maintaining the biological relevance of the TME, as in biopsies, which come at the cost of reduced experimental flexibility.

Recently, a high-throughput screen of organoid cultures was conducted to explore colorectal cancer cell plasticity. Colorectal cancer patient-derived organoids, with different mutational profiles, were co-cultured with and without CAFs and macrophages[247]. This approach showed that CAFs can enable CRC cell plasticity and induce a slow-cycling revival stem cell fate. This in turn helps cancer cells be protected from chemotherapy. However, it was limited by only including a subset of TME components. A growing number of approaches have focused on genetically engineering mouse models to carry out lineage tracing, and understand how phenotypes link to mutations within a more complex TME[254].

Often these experimental approaches are combined with an approach to quantify cell plasticity. The most common approach, although not yet widely established, correlates cell states with phylogenetic trees. A recent approach, PATH (phylogenetic analysis of trait heritability), draws on methods used in evolutionary biology to determine the extent of phenotype heritability by correlating the genetic distance using a phylogenetic tree with phenotype changes[255]. However, due to potential spatial confounders (in many cancers similar cancer clones exist in a similar spatial location[256]) this could inaccurately attribute a state to genetic heritability, whereas in fact the clones were located within the same part of the TME which caused the state change.

The spatial localisation of clones is cancer specific and it has been shown that other cancers do not have spatially located clones, for example colorectal cancer[247]. Other similar approaches include *EffectivePlasticity* metric which is also based on phylogenetic relationships to measure the distribution of transitions between cell states[254] and hidden markov models[257].

Additional approaches use chromatin measures to quantify plasticity[246]. For example, predictive models have been developed to link chromatin states to specific transcriptional phenotypes. Analysing the uncertainty in the prediction can then be used to infer high plasticity, with low uncertainty linked to low plasticity. This is as the model learns the relationship between the chromatin-derived features (e.g., ATAC-seq peaks) and the specific gene expression programs that define distinct cell states. Chromatin features mapping strongly to a single transcriptional phenotype, suggest low plasticity or a committed cell state whereas chromatin features that are difficult to predict suggest that they are compatible with multiple transcriptional states, indicating high plasticity. However, it is important to note that cells can undergo plasticity through methods beyond chromatin, such as transcriptional regulation by transcription factors and post-transcriptional modifications (e.g., RNA editing, alternative splicing), and so this approach likely picks up a narrow program of plasticity.

Overall, cell plasticity is a process that can act at the genetic, epigenetic, and environmental level to enable phenotypic adaptability. Currently definitions lack the specificity required to enable more straightforward communication of ideas about the different aspects of plasticity. For example, breaking down plasticity into capturing short term and longer term changes. I envisage that as the field matures, more precise definitions will follow.

4.1.2 Cancer as an ecological model

The relationship between environment, genotype, and phenotype is fundamental to understanding biological systems, extending beyond just cell plasticity. This relationship can be observed in many contexts: from immune cell adaptation to human responses to stressors like diet or climate and to species evolution. I can therefore leverage methods and concepts used in other areas which have extensively studied

similar relationships to enable us to model cell plasticity, and account for environmental factors.

In ecology, modelling the environment is key to understand where species are located, and how they have adapted phenotypes for their niche. In recent years, increasing analogies have been drawn between cancer cells and ecological niches, comparing the TME to an ecosystem[75,258]. An ecosystem is a natural environment where organisms interact with other organisms in addition to non-living aspects of the environment, such as the climate, water and soil. In cancer, this can be seen as cancer cells interacting with other cancer cells (of the same or different phenotypes), with other cells such as immune cells, and with other non-cellular components of the TME such as angiogenesis, hypoxia, and chemokines[75]. Additional parallels can be found in cancer evolution, where interactions between cancer cells and the TME are thought to drive evolutionary processes that enable adaptation and survival under harsh conditions. This adaptability is driven by genetic variations, including oncogenic mutations and epigenetic reprogramming, allowing cancer cells to survive within the TME.

Species distribution models and ecological niche models have emerged as a popular way of understanding the environment within ecology, and there are multiple different approaches developed using statistical and machine learning methods. Species distribution models (SDMs) focus on predicting species locations based on the environment (with variables such as temperature and soil type used), whereas ecological niche models (ENM) focus on understanding the underlying process and inference of the important variables[259–262]. These methods take into account important aspects of spatial processes, such as spatial autocorrelation, which traditional machine learning models do not. They also typically assume presence-absence data and therefore the parameters and distributions used are tailored to this[259]. This makes these models valuable as a source of inspiration but not directly applicable to our problem, which often involves continuous data (e.g., phenotype scores) or multiple discrete categories representing distinct cellular states. Nonetheless, I can adapt the concept of an ecological niche, viewing different TME components as the environmental variables that define the habitat of various cellular states.

The environment is also widely modelled by geo-scientists and economists, and a variety of spatial regression models have been developed to understand geographical phenomena[145,263]. These models are less statistically constrained compared to SDMs or ENMs, and can therefore be more easily adapted to other domains[259]. These models typically build on ordinary least squares regression (OLS), a very widely used technique to analyse the cause and effect relationship between a response variable and covariates. However, OLS is unsuitable for spatial data as it depends on the error terms in the model being independent. This is typically not the case with spatial data, due to spatial autocorrelation (observations tend to be closely related to neighbouring observations). This led to models, such as spatial autoregressive models (SAR) and spatial error models (SEM) that can account for spatially dependent variables by including a spatially lagged dependent variable as an explanatory variable (SAR) or by addressing spatial autocorrelation in the error term (SEM)[264,265]. These spatial regression models have been used to understand a range of geospatial patterns such as pollution and resource availability, and extreme weather events[266–269].

However, these models typically fail to account for spatial heterogeneity, where relationships between variables and outcomes vary across different locations. To address this limitation, geographically weighted regression (GWR) was developed to analyse how the influence of factors on a response variable changes across space[270]. A more recent development from GWR is multiscale geographically weighted regression (MGWR), where each variable can operate at its own spatial scale, recognising that spatial relationships often differ in scale[271]. Importantly though, this is a different aspect of spatial scale to the problem mentioned in Chapter 3, where scale referred to the level of aggregation. This approach calculates an optimal spatial scale (or bandwidth) for each variable by comparing models fitted across different spatial ranges using metrics such as the Akaike Information Criterion (AIC)[271]. Importantly, these methods allow for a quantification of heterogeneity of the spatial processes, rather than computing averages across the spatial landscape.

4.1.3 GeoAI to further develop statistical ecological models

Importantly, the methods described so far are fairly specific for each individual domain. The emerging field of geospatial artificial intelligence (GeoAI) aims to enhance the flexibility of existing geospatial statistical models, scale them to accommodate larger

datasets, and improve their predictive capabilities[272,273]. Due to the large feature space of spatial transcriptomic data, with unique levels of information available from gene expression to cell type information and to the importance of cell signalling networks thus driving the need to model connections between spatial variables, GeoAI approaches offer another useful framework for spatial modelling. GeoAI approaches include building in spatial autocorrelation and heterogeneity-aware methods to deep learning approaches, as well as spatial cross-validation approaches[272]. The positional encoder graph neural network is an example of a geostatistical adaptation on graph neural networks (GNNs). Here, the model is trained to predict local spatial autocorrelation of the output as an additional task, enabling the model to learn more generalisable features and predict with higher accuracy[274]. Developing deep learning methods for spatial data without incorporating the important geostatistical properties can lead to erroneous measures of accuracy of the model, and misidentify variables deemed important. However, unlike the geostatistical approaches, GeoAI variable importance is an active field of development[272].

4.1.4 Specialised methods for capturing spatial effect

Current spatial transcriptomic methods also provide valuable tools for modelling the environment. GNNs have been a powerful approach to analyse spatial transcriptomic data from cell deconvolution to ligand receptor interaction analysis[132,135]. Spatial Interaction Modelling using Variational Inference (SIMVI) recently built on these ideas to identify cell intrinsic and spatially induced latent variables in spatial transcriptomic data[275]. While SIMVI offers robust statistical guarantees for the disentanglement of these variables, it lacks interpretability, making it difficult to attribute specific contributions to individual cell types or intrinsic factors such as copy number alterations.

The understanding of how a cell changes state, whether driven by cell-intrinsic or cell-extrinsic factors, is of high importance for understanding cell plasticity and therefore developing appropriate targeted therapies[248]. In this chapter, I will walk through our approaches inspired by geo-statistics and GeoAI to quantify and explain cell-intrinsic and cell-extrinsic variables involved in cell plasticity, using EMP as an example phenotype.

4.2 Methods

4.2.1 Overview

In this work, I investigate the EMP program as a framework to showcase our methodological approach for quantifying cell plasticity. Utilising a graph neural network (GNN) model, I assess the contributions of the extrinsic and intrinsic factors to cell phenotypes. Specifically, I approximate extrinsic effects through cell type interactions and intrinsic effects through copy number alterations. By interpreting the importance of nodes and edges in the GNN predictions, I gain insights into the model's learning process.

I model EMP both as discrete states and as a continuous process, integrating geostatistical regression models to further understand EMP as a continuous phenomenon.

4.2.2 Dataset processing

- Xenium Breast Cancer Data download and processing

A Xenium breast cancer dataset consisting of 167,780 cells and 307 genes and a matched Visium dataset consisting of 4992 spots and 18,085 genes was obtained from Janesick *et al.*.[110] The sample was Stage II-B, ER + /PR − /HER2 + formalin-fixed paraffin-embedded (FFPE) breast cancer tissue. To align to Visium data to the Xenium data *SpatialData* Python package was used[276]. *SpatialData* uses landmark points in the images to transform data into a common coordinate system. Cell annotations were used as calculated in Marconato *et al.*[276]. Additional subtype annotations as labelled by a pathologist (incl. whether the region is invasive vs. DCIS) were used as described in Janesick et al.[110]

I used *Scanpy* for pre-processing, using default parameters[162]. Specifically, we filtered out genes that were in less than 5 cells, and ensured each cell had a minimum of 75 counts. For Visium I ensured the mitochondrial fraction was less than 15%, the number of genes with larger than 500 and cells has a minimum gene fraction of 0.2.

- SCEVAN clonal calculation

To perform clonal estimation in cancer epithelial cells, I utilised the matched Visium dataset due to its ability to estimate copy number amplification, currently not possible

with Xenium data because of its limited gene coverage, which limits the detection of copy number alterations. Using SCEVAN[277], I added further evidence for the tumour cells, which had been previously identified using Cell2location. Using SCEVAN, I identified subclones and estimated the chromosomal regions affected by alterations within each subclone. SCEVAN is a fast variational algorithm using multichannel segmentation. SCEVAN does not require user provided parameters as other methods such as inferCNV[278] require, and instead automatically estimates the highly confident normal cells based on count data to use as a baseline.

- Clonal PCA

To reduce the dimensions of the sub-clonal alteration matrix returned by SCEVAN (8 subclones and 160 altered regions) I ran principle component analysis (PCA) on the dataset. I obtained principal components capturing the main sources of variation within the dataset. This allowed us to derive principal components that encapsulate the key sources of variation within the data.

- EMT annotation

I used *scanpy.score_genes* to score that EMT hallmark gene signature[279] on the Xenium dataset. I ensured a high correlation between the set of EMT genes found in Xenium and the Visium datasets to ensure the limited gene coverage found in Xenium did not impact the gene signature scoring. To obtain states of epithelial, hybrid and mesenchymal cancer epithelial cells I binned the EMT hallmarks score into four quartiles, representing an epithelial state, an epithelial-hybrid state, a mesenchymal-hybrid state and a mesenchymal state.

- EMT marker genes specific to EMT tumour cells

I performed further quality control to ensure I was not mixing myCAF signatures (which display high enrichment of EMT genes) with the cancer epithelial cells labelled with a MES signature. I analysed key differentiating genes in a well annotated breast cancer atlas scRNA-seq dataset[173]. In this dataset, I filtered for only the subset of genes that were present in Xenium. I have previously annotated this dataset with EMT states. I used *sc.tl.rank_genes_groups* within the *scanpy* package to find the top marker genes between the labelled mesenchymal cancer epithelial cells and CAFs. I employed Cohen's d score to quantify the effect size and discriminative power of each candidate

marker. Cohen's d measures the standardised difference between the means of two groups, in this case, the expression levels in CAFs versus mesenchymal cancer cells and identified the number of genes required to accurately distinguish between the cell types. Using these genes I could then filter for mesenchymal cancer cells in the Xenium dataset that expressed the marker genes at an abnormally high level (higher than the 1st quartile range of LUM expression for CAFs). I had 138,780 cells after filtering.

- MERFISH dataset processing

A mouse motor cortex MERFISH dataset was downloaded from Zhang et al[280]. The data consists of 61 tissue slices and 280,000 cells. The cells were annotated using 258 genes in the original study. A total of 23 cell subclasses were identified. I used the coordinates and cell type annotations as provided by the authors

4.2.3 GNN approach

To construct the GNN prediction approach, I adapted the widely used Graph Convolutional Network (GCN) framework.

$$\boldsymbol{H^{(l+1)} = \sigma\left(\left(\widetilde{D}\right)^{-\frac{1}{2}}\widetilde{A}\left(\widetilde{D}\right)^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)}$$

Where:

- $H^{(l)}$ is the matrix of node feature representations at layer l, with $H^{(0)} = X$.
- $H^{(l+1)}$ is the matrix of node feature representations at layer $l + 1$.
- $\tilde{A}$ is the adjacency matrix with added self-loops, i.e., $\tilde{A} = A + I$.
- $\widetilde{D}$ is the degree matrix of $\tilde{A}$.
- $W^{(l)}$ is the learnable weight matrix for layer $l.\sigma$ is the activation function (e.g., ReLU).

The datasets were loaded into a PyTorch Geometric dataset, where I adapted custom classes to include various features and configurations tailored to our experiments. These configurations included incorporating cell type and/or copy number information, as well as modifying the training paradigm to be either inductive or transductive. On the MERFISH dataset, I utilised a graph neural network (GNN) with three graph convolutional layers, a learning rate of 0.01, and trained the model for 200 epochs. A

dropout rate of 0.5 was applied to regularise the model and prevent overfitting. Similarly, for the Xenium breast cancer dataset, I trained a GNN with the same architectural configuration, three layers, a 0.01 learning rate, and 200 epochs, with a dropout rate of 0.5. I also used Graph Attention Networks (GAT) but found no difference in accuracy. The *SquidPy* (Spatial Single-Cell Analysis in Python) package[116] was used for graph construction using *sq.gr.spatial_neighbors* and *NetworkX*[163] was used for further graph manipulation.

Mean squared error (MSE) loss was used for continuous training, while cross-entropy loss was used for classification tasks involving categorical labels. This enabled the implementation of both GNN regression and GNN classification models. I adjusted for spatial autocorrelation by calculating the Moran's I statistic and propagating this backwards:

$$L_{Moran} = \frac{1}{W}\sum_{i,j} w_{ij}\left(x_i - x_j\right)^2$$

L_{Moran}: Moran's loss, quantifies spatial smoothness or autocorrelation.

W: Total sum of all spatial weights

w_{ij} : Spatial weight between locations i and j, based on proximity.

x_i : Feature values (e.g., gene expression, model outputs) at locations i and j.

To mitigate the issue of class imbalance in the training data, I applied weighted loss functions, ensuring balanced representation and learning across all classes:

$$w_i = \frac{1}{f_i}$$

$$w_i^{\text{normalized}} = \frac{w_i}{\sum_{j=1}^{C} w_j}$$

f_i represents the frequency of class i in the training set. w_i is the initial weight assigned to class i, and $w_i^{\text{normalized}}$ is the weight after normalisation. The normalisation step

ensures that the sum of all weights $\sum_{i=1}^{C} w_i^{\text{normalized}}$ equals the number of classes C, maintaining balance in the loss function.

Full-graph training was conducted to use the complete structural information present. When training I used the entire graph by masking out the other non-tumour cells before loss back-propagation. Therefore the masked out non-tumour cells were used for loss calculation but they did not influence the model as they were not used in the training. This improved model training compared to a more standard approach of using subgraphs.

GNN training can be approached using either transductive or inductive learning. In transductive learning, node classification is performed within a single graph, where some nodes are masked during training and used for prediction. Inductive learning, by contrast, involves training on subsets of a graph or entirely separate graphs, aiming to generalise the learned representations to previously unseen graphs. I reported the results for both approaches.

Spatial cross-validation was conducted using 10 spatial splits. When using the Xenium dataset, inductive training was conducted by dividing the slide into 10 spatial splits, while transductive training involved randomly masking nodes, keeping 10% of the nodes within in a test set (5,776 tumour nodes). For the MERFISH dataset, which included 61 slides, inductive training was performed by splitting the data based on individual samples, whereas transductive training again involved random node masking across the entire dataset.

The models' performances were evaluated using F1 scores and ROC-AUC for classification, and mean squared error for regression.

- GNN explanation

GNNExplainer was used for edge explanation. The main goal of GNNExplainer is to find a subset of the graph that maximises the prediction probability for a given target. This helps in understanding which parts of the input graph are most relevant to the decision made by the GNN. GNNExplainer learns a mask over the edge features, and uses gradient-based optimisation to update the masks. The loss function combines the prediction loss (which ensures that the masked subgraph results in the same prediction as the original) and a regularisation term (which controls the complexity and

sparsity of the mask). I compared the explanations for each class to explanations generated from random shuffled nodes to obtain a p-value for each edge explanation.

Nodes were explained using integrated gradients, an approach which assigns an importance score to each node in a graph by measuring how adjusting that input node feature from a baseline to its actual value changes the model's prediction:

$$G_{i(x)} = \left(x_i - x_i^{'}\right) \cdot \int_{\alpha=0}^{1} \frac{\partial F\left(x^{'} + \alpha\left(x - x^{'}\right)\right)}{\partial x_i} d\alpha$$

Where:

- i is the feature
- x is the input
- $x^{'}$ is the baseline
- α is the interpolation constant to perturb features by

The definite integral is not always numerically possible so a numerical approximation is calculated instead.

4.2.4 Spatial regression modelling

Prior to spatial regression I removed variables with high VIF and autocorrelation scores. I tested for additive effects using the spatial random forest *spatialRF* R package[281]. I did not uncover significant additive effects. Unless otherwise states, I used models as implemented in the PySAL package[199].

I implemented spatial error modelling (SEM) to our datasets. SEM is an extension of linear regression that incorporates a spatially structured error term to model spatial autocorrelation in the residuals, thereby capturing spatial dependencies that would otherwise violate standard regression assumptions.

The basic form of the Spatial Error Model (SEM) can be expressed as:

$$y = X\,\beta + \lambda W\,\epsilon + \epsilon$$

Where:

- y: Represents EMT label for each observation

- X: Contains the predictors for the regression model
- β: Coefficients that represent the effect of each predictor
- W: Spatial weights matrix (spatial relationships among observations)
- λ: Spatial autoregressive coefficient for the error term (Captures the strength of the spatial dependence in the error terms)
- ϵ: Independent and identically distributed error terms; assumes a normal distribution

$$\epsilon \sim N(0, \sigma^2)$$

I applied SEM to various subsets of the Xenium data, including between ductal carcinoma in situ (DCIS) and invasive regions, to assess how spatial interactions change within these regions.

Geographically weighted regression (GWR) and multi-scale geographically weighted regression (mxGWR) were also used to further interrogate the spatial relationships. These fit local regression models for each point, and therefore answer different questions to a SEM model. SEM is important for understanding overall variable importance, and variance captured. However, it may not fit the most appropriate model for each spatial point, considering spatial heterogeneity. Comparing model fit of SEM and GWR models allows for an assessment of how much heterogeneity is present, and to visualise how relationships change over space. GWR and mxGWR were also compared for spatial fit. To evaluate model performance, R^2 was used to measure the proportion of variance explained, providing an indication of how well each model predicts EMT. The Bayesian Information Criterion (BIC) was also used to assess model complexity and fit, with lower values indicating a better balance between goodness of fit and model complexity.

GWR is a spatial analysis method that allows for the modelling of spatially varying relationships between dependent and independent variables by fitting a local regression model at each point in space. The GWR equation can be represented as:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \epsilon_i,$$

Where:

- y_i is the EMT label at location i
- $\beta_0(u_i, v_i)$ is the intercept term specific to location u_i, v_i
- $\beta_k(u_i, v_i)$ are the local regression coefficients for the k -th explanatory variable at location (u_i, v_i)
- x_{ik} are the explanatory variables at location i
- ϵ_i is the error term at location i
- p is the number of explanatory variables

mxGWR extends GWR by allowing each explanatory variable to have its own spatial bandwidth, which enables the modelling of multi-scale spatial relationships. The mxGWR model can be represented as:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i, b_k) x_{ik} + \epsilon_i$$

Where:

- b_k Unique spatial bandwidth for the k-th variable, allowing the model to adapt to varying spatial scales

GWR applies a single bandwidth across all variables, which may not capture multi-scale spatial processes accurately. In contrast, mxGWR, allows different bandwidths for each explanatory variable, provides a more flexible and specific understanding of spatial heterogeneity.

4.3 Results

In this section, I first introduce the Xenium data used as the proof of concept for our spatial predictive modelling concept (4.3.1). I then further describe the GNN approach for prediction and highlight the results in a well-annotated mouse brain dataset and the Xenium breast cancer dataset (4.3.2). Finally, I demonstrate the additional insights gained from geostatistical regression methods on the Xenium breast cancer dataset (4.3.3).

4.3.1 Characterising the Xenium breast cancer dataset

To analyse the intrinsic and extrinsic variables that shape EMT in breast cancer, I integrated annotations of the dataset from various sources. First I combined major and minor cell type annotations from a previously run cell annotation study[276] (

Figure 38). The cell types included 9 major populations: B-cells, CAFs, cancer epithelial cells, endothelial cells, myeloid cells, normal epithelial cells, PVL cells, plasmablasts and T-cells. The minor cell classes included cancer LumA stem cells, macrophages, cancer basal cells, Cancer LumB cells, luminal progenitors, PVL immature cells, mature luminal cells, cancer Her2 stem cells, myoepithelial cells, myCAF-like cells, cancer cycling cells, monocytes, cycling myeloid cells, endothelial lymphatic LYVE1 cells, endothelial CXCL12 cells, natural killer (NK) cells, cycling T cells, endothelial ACKR1 cells, plasmablasts, NKT cells, CD8+ T cells, CD4+ T cells, dendritic cells (DCs), naive B cells, memory B cells, cycling PVL cells, and PVL differentiated cells. The cell type labels make up the extrinsic variables I am modelling. In addition, I integrated pathologist labels[110], including ductal carcinoma in situ (DCIS) and invasive breast cancer subtypes, as shown in **Figure 39**.

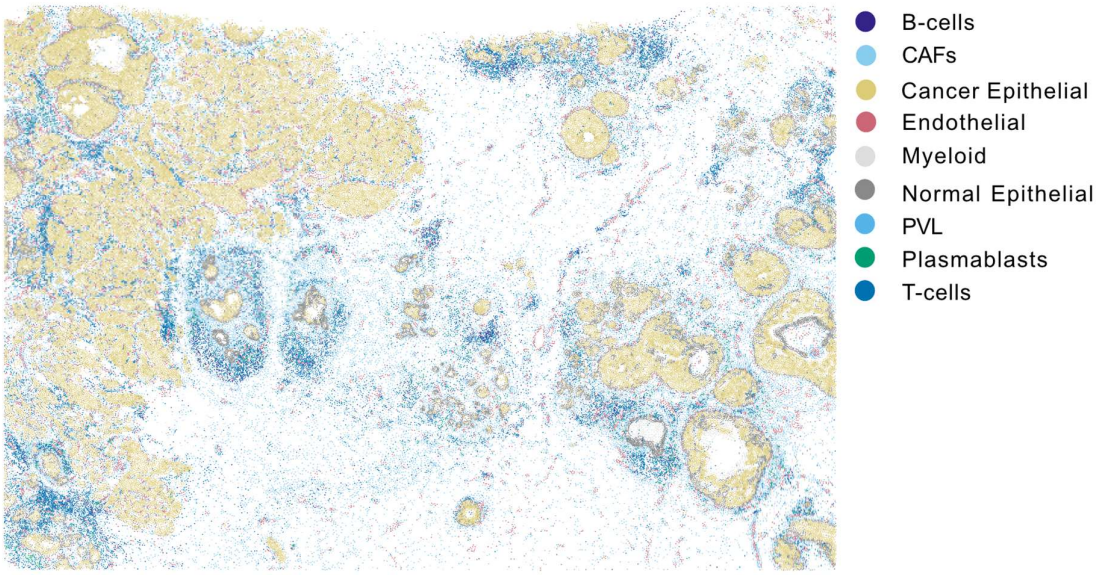

Figure 38. The major cell types annotated in the Xenium slide.

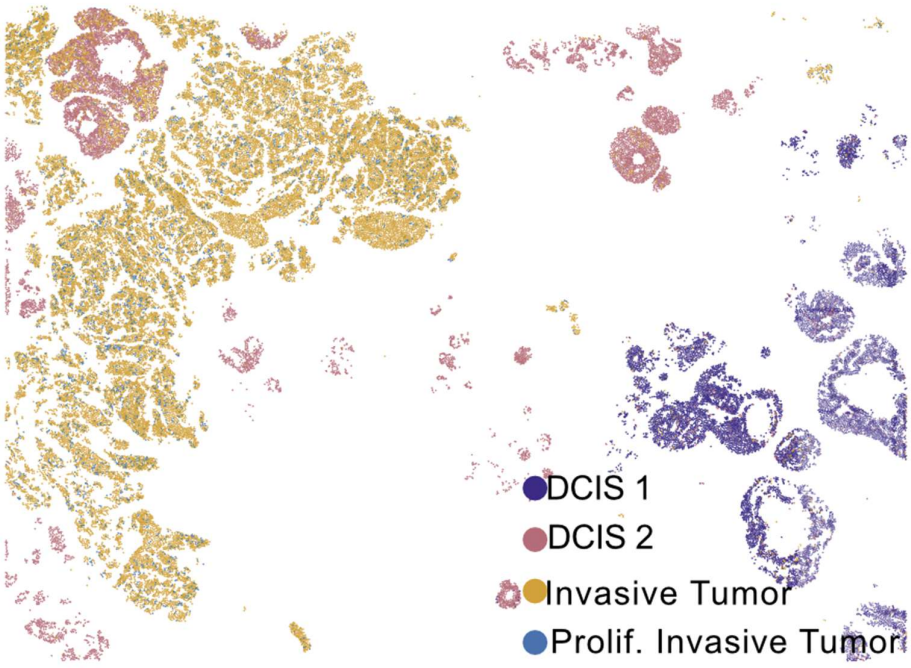

Figure 39. The pathologist labels describing the Xenium slide, including DCIS, invasive tumour and proliferative invasive regions.

To approximate the intrinsic cell variables, I estimated the copy number changes present in each cancer cell using SCEVAN[277]. This tool divides the genomic regions into bins, smoothing out noise, and uses a non-tumour cell baseline to estimate the copy number changes. For accuracy, it identifies the copy number changes in clusters (subclones). I identified 8 subclones in total sharing similar genomic alterations (**Figure 40a-c**). These subclones display a range of copy number alterations. Alterations found across the subclones include the 17q22-24 amplification, a common site of amplification in breast cancer[282], and 8q24.3 amplification which included the MYC oncogene, often amplified in aggressive breast cancer[283]. Deletions found across all subclones include chromosome 11q13.4-q25 which contains the CCND1 gene (cyclin D1) involved in cell cycle regulation[284].

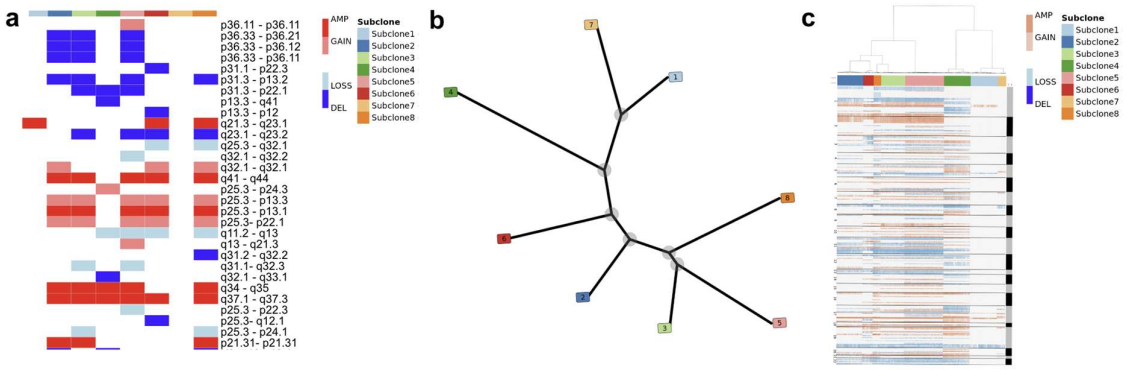

Figure 40. Copy number alterations present in Xenium slide. a. Genomic alterations present in each subclone. Heatmap represents a select number of regions. **b**. Subclone tree determined by the copy number alterations. **c**. Full heatmap highlighting genomic alterations present in each subclone.

I reduced the dimensions of the copy number alterations matrix for the subclones using PCA. This allows us to have a smaller set of uncorrelated variables capturing different portions of the variance in the data to use in the downstream GNN or spatial regression analysis. Each PC captured different aspects of the variation in the copy number alterations, with PC1 to PC7 capturing over 99% of the variation (**Figure 41**). The PCs align with genomic patterns associated with breast cancer. PC1 relates to deletion of tumour suppressor genes like BRCA2 (13q13.1-q34) and amplification of oncogenes such as CCND1 (11q13.2-q13.4). PC2 highlights important regions for breast cancer (e.g., 1p and 3q[285,286]). PC3 and PC4 mainly included oncogene amplification, such as PIK3CA (3q26.1-q26.33). PC5 and PC6 include known oncogene amplifications (e.g., CCND1 at 11p) and significant tumour suppressor gene deletions. PC7 indicates co-occurring amplifications and deletions, involving regions with genes like BRCA1 (17q12-q21.1). Supplementary Table 1 summarises the regions associated with each PC, and their associated amplification or deletion. The PCs manage to capture the trends observed in the subclone evolutionary tree, for example, it is apparent when visually inspecting the PCs, that PC1 captures chromosomal alterations found in subclone 4 to the highest extent, and then subclone 6 (**Figure 42** - **Figure 43**). On the subclone tree, I see that 6 is a subclone descended from subclone 4 (**Figure 40**).

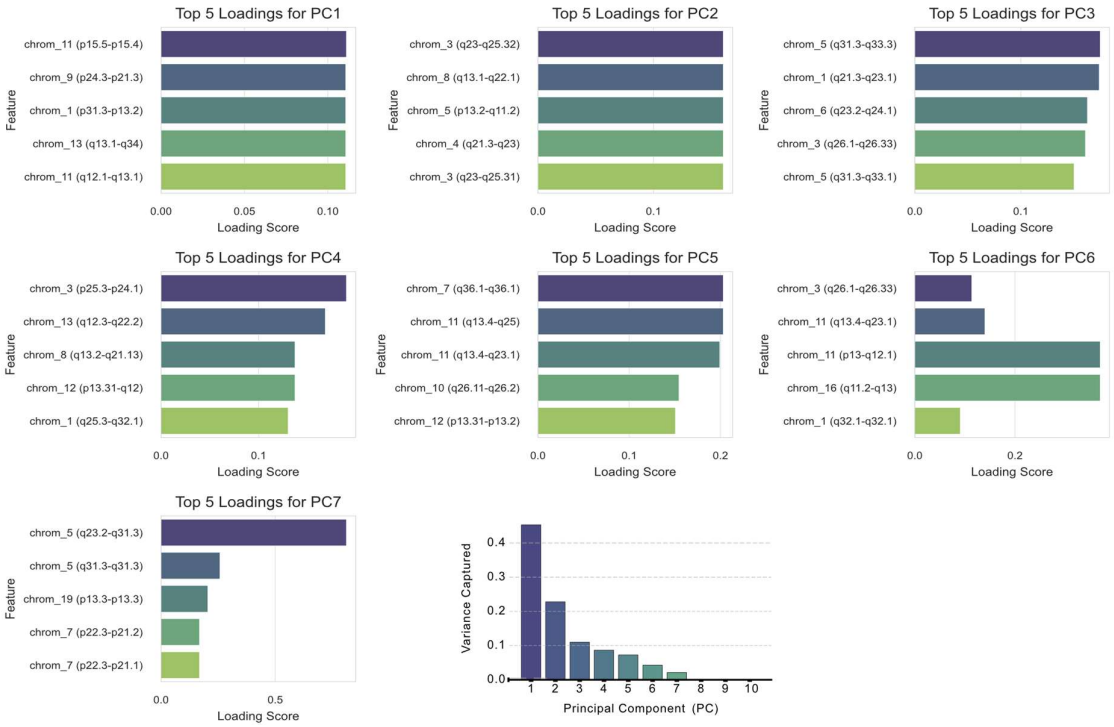

Figure 41. **The feature loadings for each PC**. The top loadings are plotted, each corresponding to different chromosomal regions and the variance explained for each PC (bottom centre).

Figure 42. The subclones annotated in the Xenium slide.

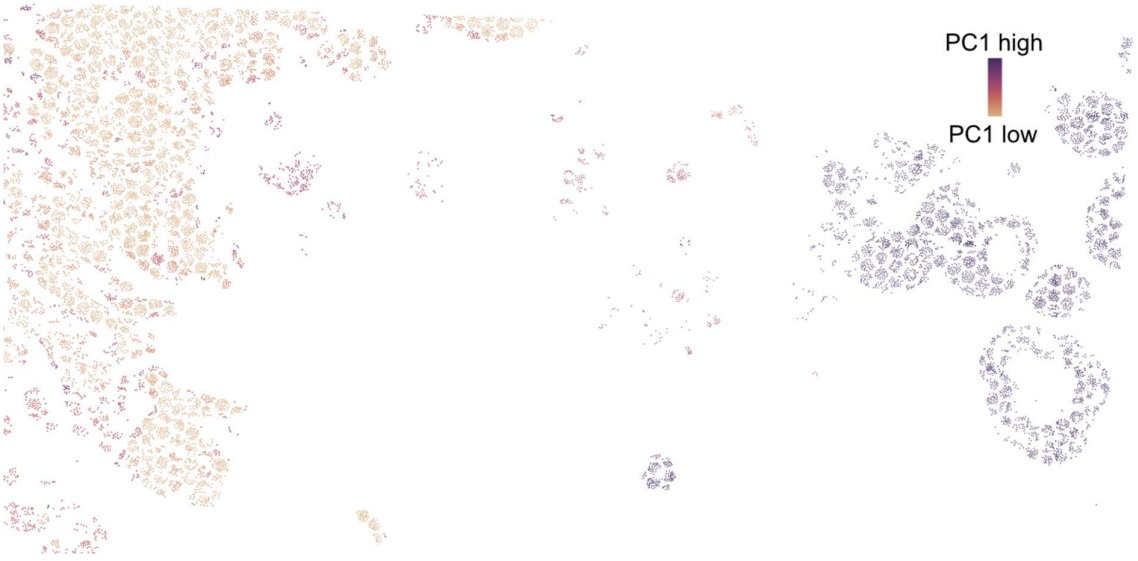

Figure 43. PC1 plotted across the Xenium slide.

I then assessed the EMT state of the cancer epithelial cells present in the Xenium data. To ensure that the reduced number of genes in Xenium did not compromise our downstream analysis, I compared the EMT signature scores derived from the reduced set of genes (20 genes that are overlapping in EMT signature and Xenium) with the scores from the full 195-gene EMT hallmark signature in the Visium data. I observed a strong correlation, confirming that the smaller gene panel still robustly captures the EMT signal (**Figure 44**). This is in line with other research that suggests that EMT hallmark gene signature can be captured more targeted gene panels[287]. I identified four EMT states by binning the signature enrichment score into four.

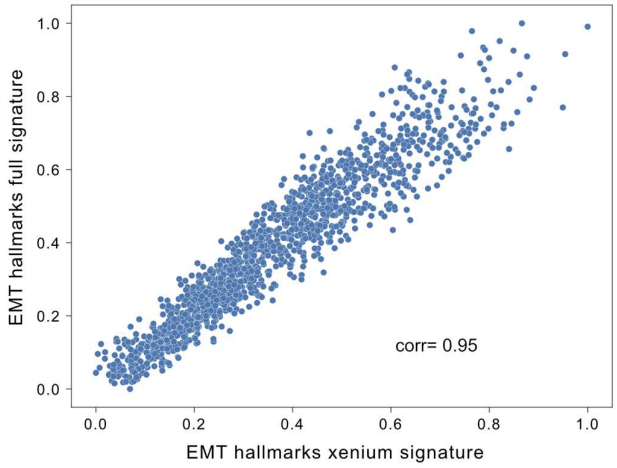

Figure 44. EMT hallmark correlation analysis. Correlation of full gene list for EMT hallmarks signature scores in Visium breast cancer dataset compared to the limited gene list of only genes present in Xenium scored in the Visium breast cancer dataset.

Given the shared transcriptional program between myCAFs and EMT cancer epithelial cells[229], I wanted to confirm that the cells labelled as EMT cancer epithelial cells were not mistakenly classified myCAFs. To do this I sought to identify uniquely expressed genes within these cells types that can be used as reliable markers. I began by identifying the top differentially expressed genes between myCAFs and cancer cells with a mesenchymal phenotype in the previously labelled scRNA-seq breast cancer dataset as described in Chapter 3 (**Figure 45)**. I then incrementally selected the top 10 differentially expressed genes (**Figure 46a),** and evaluated how well these genes could differentiate CAFs from EMT tumour cells in the scRNA-seq dataset (**Figure 46b)**. Using Cohen's d score as the metric, I found that the top differentially expressed gene, LUM, had a particularly strong discriminative power (Cohen's d = –4.26), surpassing even the separation achieved by combining the other top 2–10 genes (Cohen's d ranging from –3.45 to –3.62) **(Figure 46b)**. I further confirmed the gene expression of LUM could differentiate the myCAFs and cancer epithelial cells by assessing the distribution across the cell type subsets, (**Figure 47**-**Figure 48**). I then investigated LUM expression in the Xenium data and found a small proportion of cells labelled as MES in our dataset with high LUM expression (**Figure 49a**). I removed these cells (**Figure 49b**).

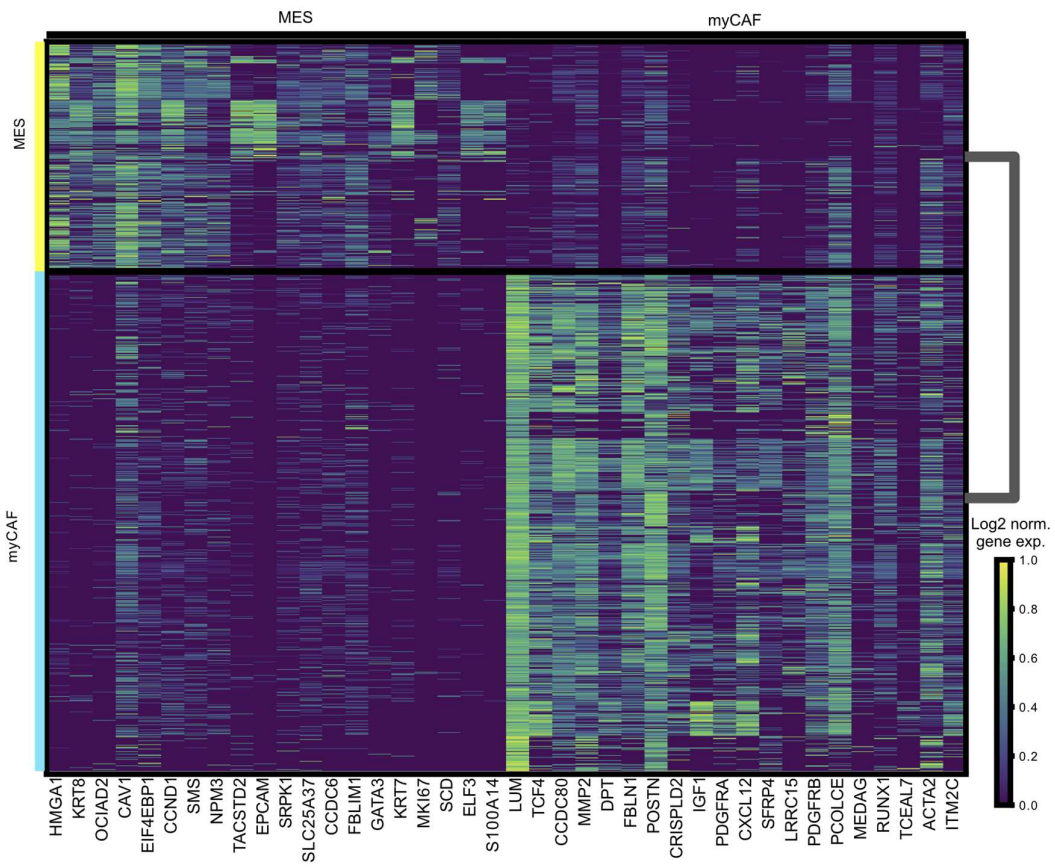

Figure 45. Heatmap illustrating the top-ranked differentially expressed genes and their expression profiles across myCAFs and mesenchymal tumour cellsin the annotated scRNA-seq dataset.

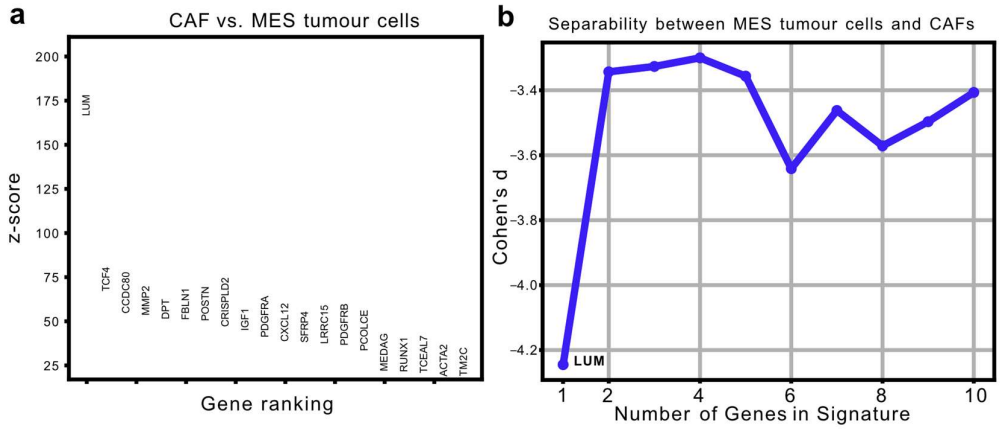

Figure 46. **a.** Top 20 differentially expressed genes distinguishing CAFs from MES tumour cells ranked by their z-score**. b.** Cohen's d metric highlighting cluster separability between CAFs and MES tumour cells using increasing number of marker genes.

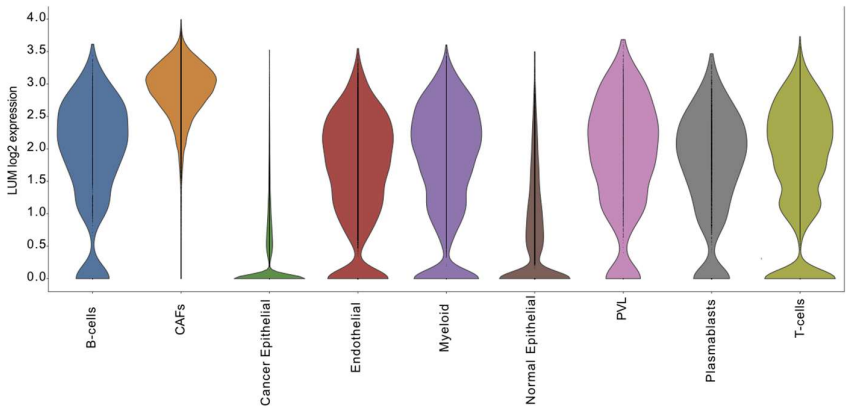

Figure 47. LUM expression across the major cell types in breast cancer scRNA-seq dataset. Values represents the log2 normalised gene expression of LUM.

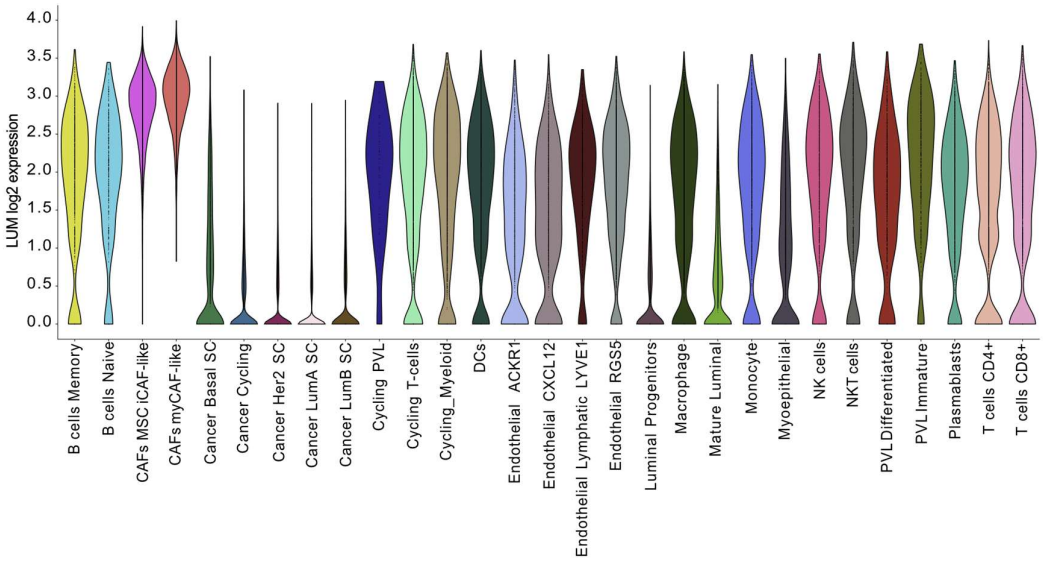

Figure 48. LUM expression across the minor cell types in the breast cancer scRNA-seq dataset. Values represents the log2 normalised gene expression of LUM.

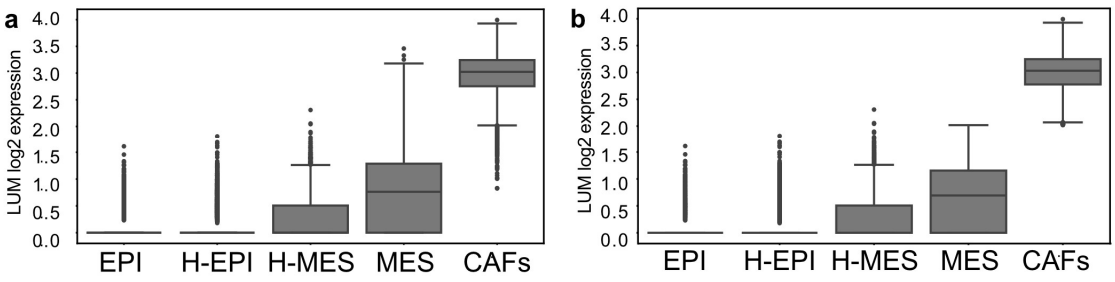

Figure 49. LUM expression across the Xenium annotated cancer cells and CAFs. a LUM expression pre filtering Xenium cells. **b** LUM expression post filtering for the 1st quartile range for CAFs. Values represents the log2 normalised gene expression of LUM.

Having defined the four states within the breast cancer Xenium dataset, I then assessed their distribution across subclones and molecular subtypes (**Figure 50**). I found that invasive regions had a higher proportion of MES cells, as expected due to the link between EMT and metastasis. Interestingly, the invasive proliferative regions contained a smaller proportion of MES cells, in line with our previous work and extensive other research[288–290] suggesting that a proliferative state acts in opposition to a mesenchymal state. The DCIS has a lower proportion of MES cells. However, all regions did have all states present within them. This was the same with the subclones as well. Interestingly, the proportion of EMT states align with their evolutionary tree. For example, the subclones with the highest MES states are closer in origin (subclone 7 and 1) and the subclones with the highest EPI states are also closer in origin (subclone 4 and 6). This hints at a genomic influence on EMT which I will explore in the modelling.

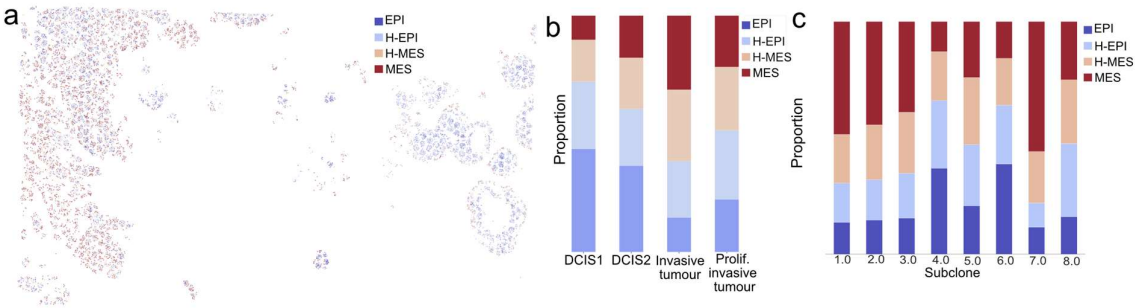

Figure 50. EMT states identified in Xenium dataset. **a** EMT state proportions visualised within the Xenium dataset. **b** EMT state proportions across the subtypes within the Xenium dataset. **c** EMT state proportions across the subclones within the Xenium dataset.

Now, I have the fully annotated dataset (**Figure 51** - **Figure 52**) that I can explore further in the next sections.

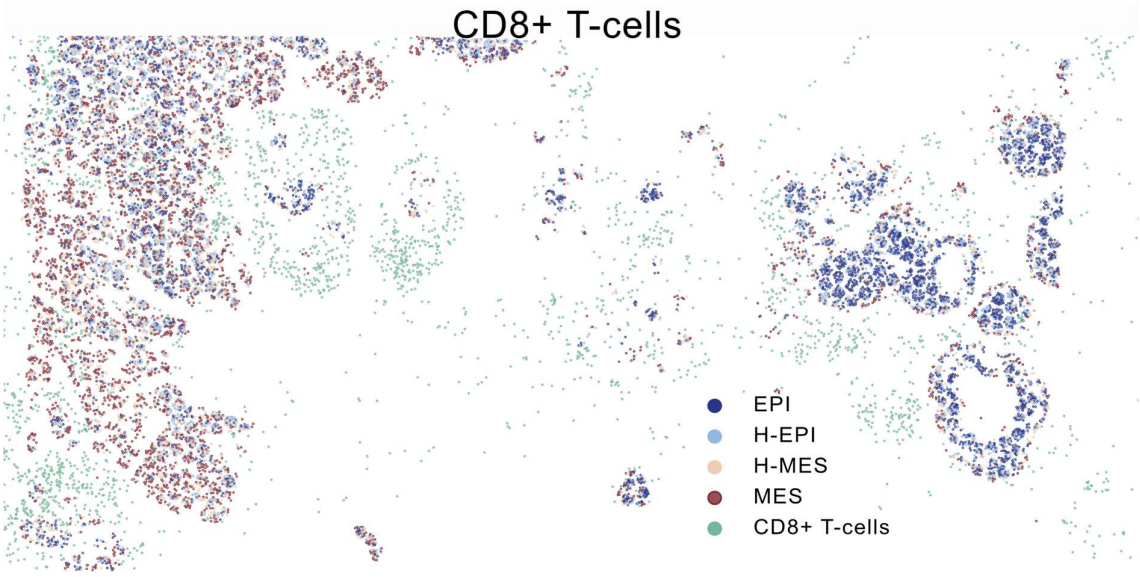

Figure 51. EMT states and CD8+ T-cell distribution in the Xenium dataset.

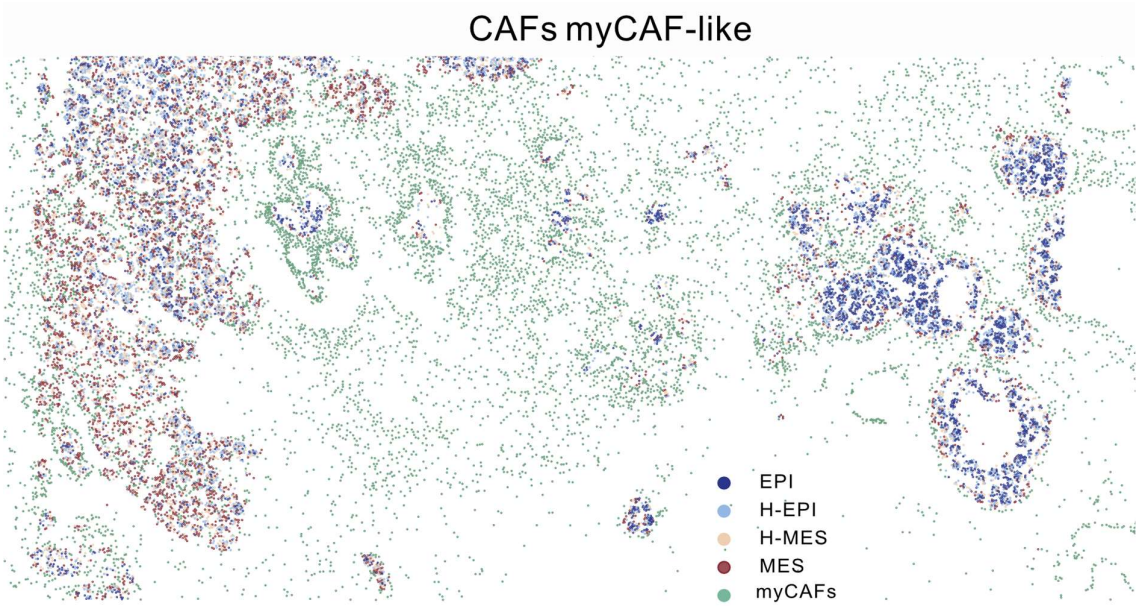

Figure 52. EMT states and myCAF cell distribution in the Xenium dataset.

4.3.2 Modelling the TME and genomic influences on EMP using graph neural networks

Rationale of GNN approach

To assess the spatial dependence of the TME on EMT I developed a graph neural network framework. If I cannot reliably predict EMT, then it suggests that the spatial positioning of cancer cells as they undergo EMT does not impact the transition, or I need to capture other levels of spatial information. I hypothesise that if the spatial component (TME) is more important for EMT compared to the intrinsic factors (genomic changes), it should be easier to predict the EMT status of the cells using this information compared to just the intrinsic information (here using copy number alterations as a proof of concept).

A GNN can capture spatial relationships between cell types in a way that other machine learning models cannot. Unlike most spatial prediction models, which typically rely on distance as a feature, GNNs can model the connections between nodes (in this case, cell types) through edges, allowing for a better representation of spatial interactions.

I can then use this framework for a more general framework of modelling which variables are most important for cell state changes. Following up from this, I can then suggest how “plastic” cell states are. I would hypothesise that states that cannot be easily predicted are flexible states exhibiting plastic potential. The states that have higher predictability are cell types likely have stable spatial relationships and specific roles in their microenvironment. This reflects biological constraints on their function. These may be the cells that are easier to treat, given their well-defined spatial relationships and may have stable therapeutic targets e.g. other immune cells. On the other hand, more plastic cells, which are harder to predict, have been shown to be more challenging to treat due to their dynamic and less constrained nature[291–293]

These can be formulated as equations:

Variables:

- S: Cell state (e.g. EMT) which can be binary or continuous
- I: Intrinsic variables e.g. copy number alterations
- E: Extrinsic variables from the TME
- $\hat{S}$: The predicted cell state from the model

Performance metrics:

- AUC: Calculates ability of the model to distinguish between classes by calculating the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one (used for classification tasks)
- R^2: Proportion of variance in cell state that is predictable from the independent variables (used for regression tasks)

A cell state e.g. EMT can be modelled as a function of intrinsic variables, extrinsic variables and randomness:

$$S = f(I, E, R)$$

i. State primarily linked to genomic and TME variables, potentially indicating genomic factors increasing cells responsive to the environment:

$$S = \alpha f(I) + \beta f(E)$$

$$AUC_I \; or \; {R_I}^2 \approx AUC_E \; or \; {R_E}^2$$

ii. State primarily linked to extrinsic and not driven by intrinsic factors

$$S = f(E)$$

$$AUC_E \; or \; {R_E}^2 > AUC_I \; or \; {R_I}^2$$

iii. State primarily linked to intrinsic variables ie. genomic variables:

$$S = f(I)$$

$$AUC_I \; or \; {R_I}^2 > AUC_E \; or \; {R_E}^2$$

Validation of GNN method

To illustrate the concept of capturing the spatial effect using the TME, I demonstrated the method using a MERFISH dataset from the mouse cortex[280]. This dataset includes both cell types that exhibit spatial dependencies (their distribution is influenced by tissue structure or interactions) and those that are more randomly distributed, lacking strong spatial organisation and therefore offers a useful test to check the method picks up the key differences in spatial dependencies (**Figure 53**).

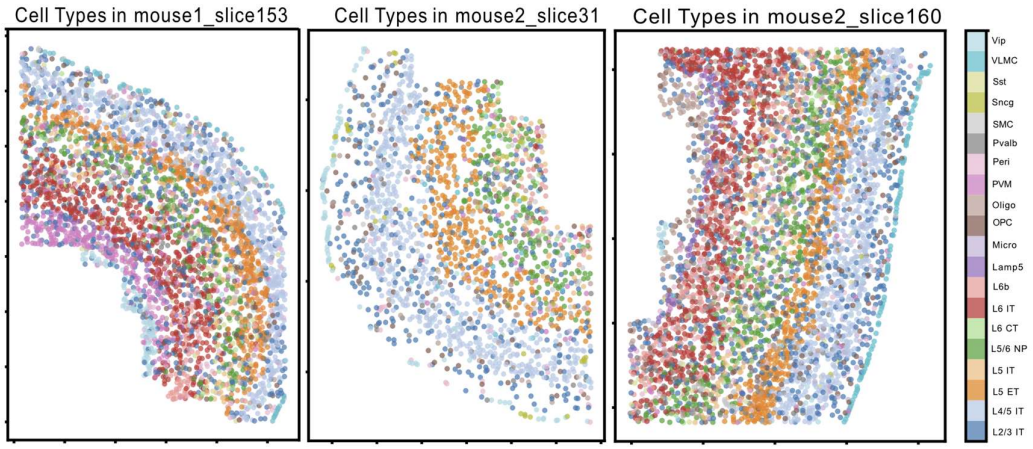

Figure 53. A representation of the cell types present in mouse motor cortex MERFISH slides.

The layered neuron cells, such as Layer 2/3 intratelencephalic neurons (L2/3 IT), Layer 4/5 IT neurons (L4/5 IT), Layer 5 extratelencephalic neurons (L5 ET), Layer 5 IT neurons (L5 IT), Layer 5/6 near-projecting neurons (L5/6 NP), Layer 6 corticothalamic neurons (L6 CT), and Layer 6b neurons (L6b), have distinct spatial organisations within the motor cortex, each with distinct roles in cortical processing. For example, L2/L3 IT cells are involved in integrating information across cortical layers and higher-order processing. In addition to the layered neurons, the vascular leptomeningeal cell type (VLMC) has a distinct spatial location at the border of the cortex (**Figure 53**).

Additional cells within the dataset include the GABAergic neuron class, which was classified into five subclasses based on marker genes; Pvalb (Parvalbumin-expressing) neurons, Sst (Somatostatin-expressing) neurons,Vip (Vasoactive Intestinal Peptide-expressing) neurons, Sncg (Synuclein-gamma-expressing) neurons, and Lamp5 (LAMP family member 5-expressing) neurons. In addition to the neuronal cell classes, the dataset includes several non-neuronal cell subclasses; astrocytes (Astro), endothelial cells (Endo), pericytes (Peri), smooth muscle cells (SMC), microglia (Micro), perivascular macrophages (PVM), oligodendrocytes (Oligo) and oligodendrocyte precursor cells (OPC).

I hypothesised that using the GNN prediction approach, the spatially constrained cells should be significantly easier to predict than the other cell types. To test this, I trained the model using the MERFISH dataset and evaluated its performance on separate

mouse MERFISH slides, ensuring no data leakage, averaging the results over 10 distinct train-test splits. For each experiment, I masked the cell type of interest and used it as the prediction label, while providing only the remaining cell types as input to the model. This process was repeated for each cell type to assess the model's predictive ability.

The known spatially constrained cells with well-defined roles do have a higher prediction accuracy (**Figure 54**). L2/3 IT, L4/5 IT, L5 ET, L5 IT, L5/6NP, L6CT and VLMC cells have an AUC above 0.85. These results are consistent regardless of the train and test split, and align with our expectation that relative to the other cell types these should have higher predictability. Oligodendrocyte precursor cells (OPCs) and pericytes (peri) display the least predictive capabilities. OPCs are considered plastic cells, as they can differentiate into oligodendrocytes depending on where myelination is needed[294]. This plasticity means that OPCs are not tied to a fixed, predefined location; instead, they may migrate or change their function depending on the local demands for myelination, making their spatial distribution more dynamic and less predictable.

Pericytes were amongst the least predictable cell types, and these are also plastic, changing their phenotype in response to many different environmental cues, such as injury, inflammation, or changes in blood flow[295,296]. Pericytes can differentiate into cells such as smooth muscle cells, fibroblasts and osteoblasts and are important in wound healing[297]. Astrocytes, and perivascular cells are also considered plastic cells, compared to cells such as VLMC and the layered neurons, and these had lower AUC values[298–300].

Plasticity in cells is inherently linked to more variability in their locations and functions, making them harder to predict accurately in spatial models. Therefore, the pattern of higher accuracy for more stable cell types and lower accuracy for more plastic ones aligns with expectations.

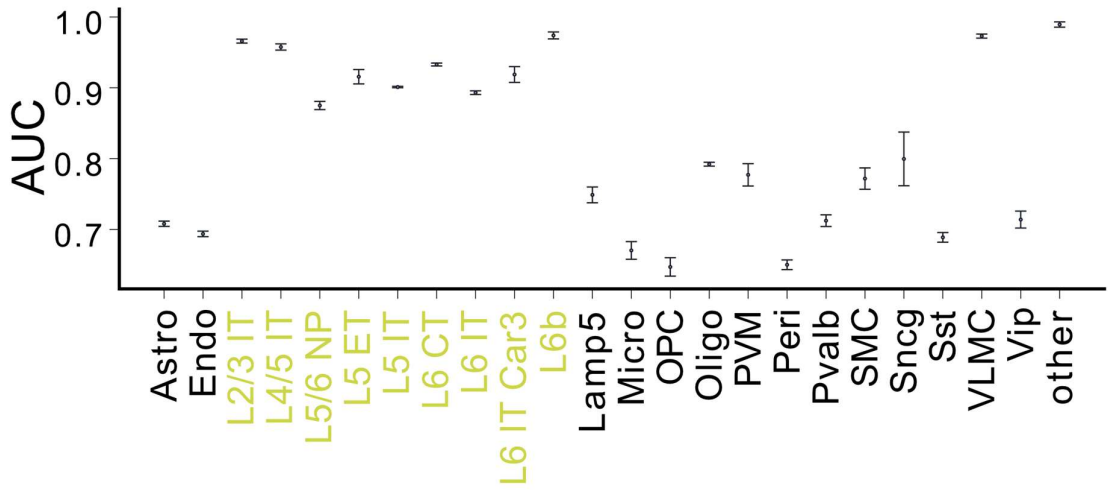

Figure 54. AUC range of predicted cell types in the mouse motor cortex MERFISH dataset. The plot displays the average AUC scores across 10 different training splits, with error bars representing the standard deviation. Layered (spatially distributed) cell types highlighted in green.

Using this model, I can analyse which cell types are most influential for each prediction (**Figure 55**). Overall, the results align with biological expectations. For example, L2/L3 neurons had a significant relationship with VIP interneurons, Pvalb interneurons, microglia and endothelial cells. These are as I would expect; Pvalb and VIP neurons and microglia are known to regulate L2/L3 neurons, and endothelial cells are important for maintaining the blood-brain barrier which supports L2/L3 IT neurons[301–303]. L6 IT neurons had a significant relationship with L6 CT neurons, which share close spatial relationships, and Sst interneurons, which target the dendrites of L6 IT neurons[304,305]. L6 IT Car3 neurons had a close relationship with L6 CT neurons, which are spatially located in a shared local circuit[306]. Oligodendrocytes had a significant relationship with OPCs, which make sense as OPCs differentiate into oligodendrocytes[307]. PVM (perivascular macrophages) interacting with pericytes also make sense, are they share a similar niche; both associated with blood vessels[308]. VIP neurons and Lamp5 neurons have been found to be enriched in the superficial (layer 1 to 3 layers), and therefore their relationship makes biological sense[309].

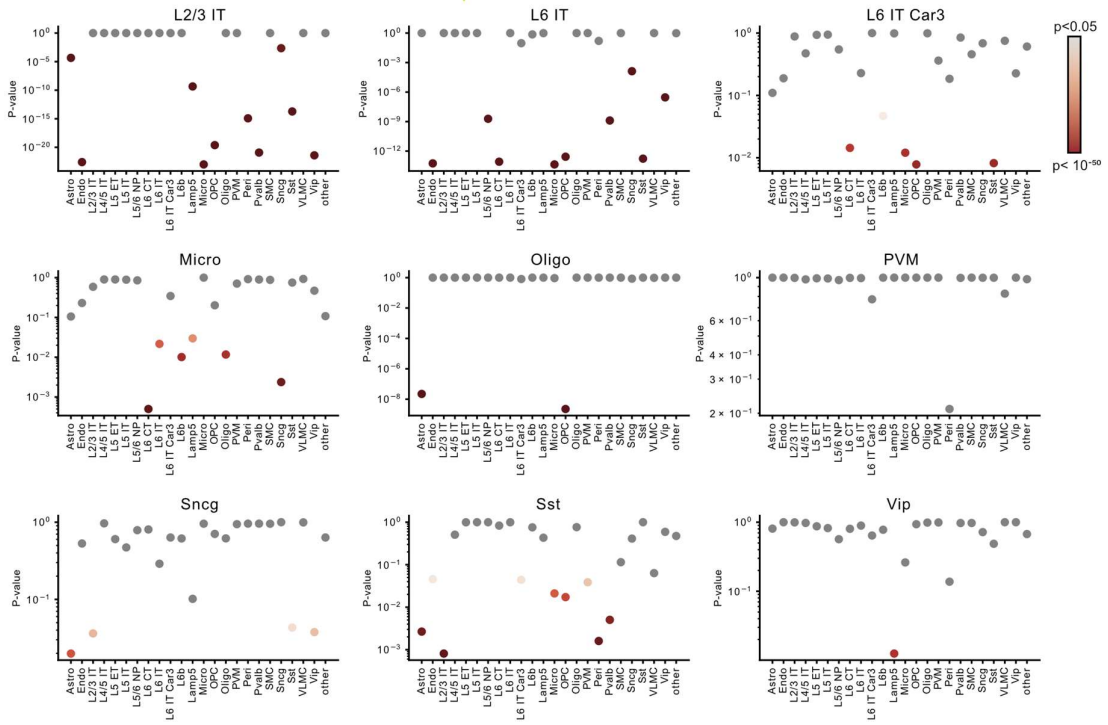

Figure 55. Cell type significance for each cell type prediction using node explanations. Coloured by p-value significance level.

I next sought to understand the edge explanations (**Figure 56**). Edge explanations can identify the important node connections, highlighting the structural dependencies in the graph. Overall, I found that the edges make biological sense. I am also likely uncovering relationships that are unknown due to the unique method of assessing the spatial relationships, in a more unsupervised strategy compared to more directed methods such as niche detection and co-localisation approaches.

VIP neurons had the greatest number of interactions with L2/3 IT neurons. This aligns with their spatial location in layer 2/3[310]. For SMCs, I found edges mainly involving VLMCs. This aligns with our biological understanding as they are both vascular-associated cells, particular at the cortical borders[306]. Astrocytes had many significant relationships when predicting PVM interactions, which makes biological sense given their role in immune modulation; astrocytes release cytokines that directly regulate PVM activity[311].

Endothelial cells demonstrated a broad range of interactions, highlighting their multiple roles in vascular support and neurovascular coupling. Endothelial interactions were particularly significant with L4/5 IT and L5 IT neurons. This is consistent with the high

vascular density of Layer 4, which supports intense sensory input processing, and the high metabolic demands of the large projection neurons in Layer 5[312].

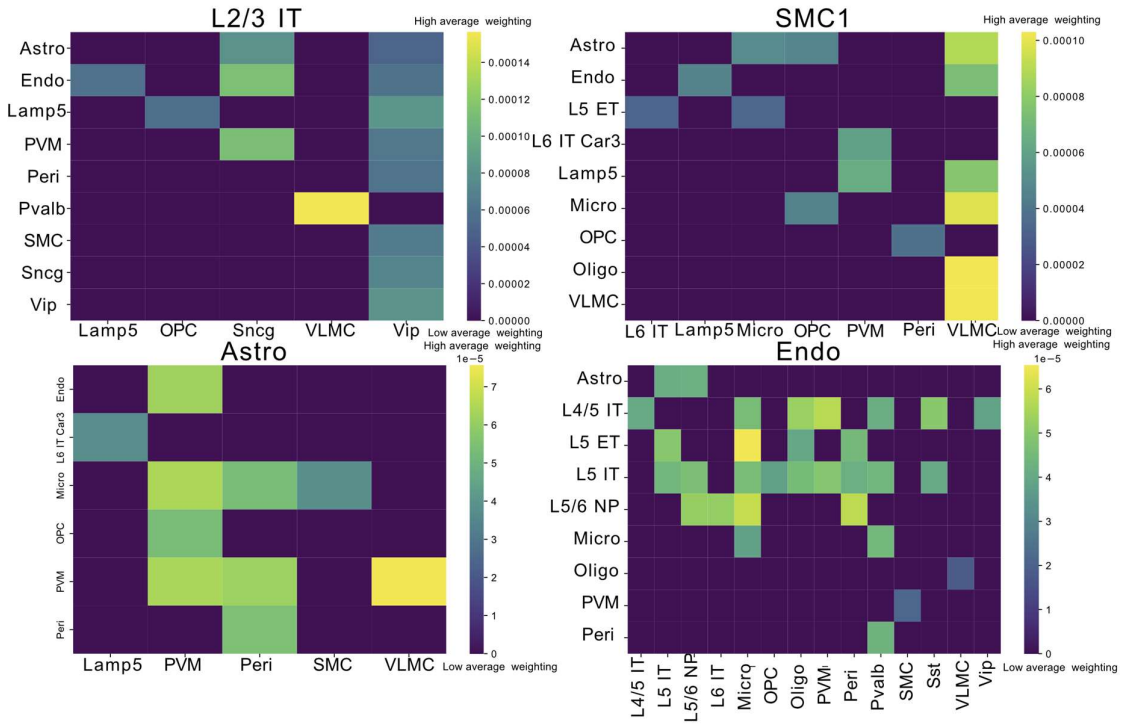

Figure 56. Heatmaps displaying the significant edges for the cell type prediction. Heatmap highlights the weighting for the significant edges.

Prediction model to understand EMT states

Having confirmed the model aligns with biological expectations, whilst potentially offering novel insights in the MERFISH dataset, I applied the method to the Xenium breast cancer dataset. The GNN approach was trained with either just the intrinsic information (copy number alterations), just the extrinsic information (the TME information) or both and the AUCs compared.

Interestingly, both genomic factors and the TME predict EMT states to a similar overall extent (**Figure 57**). The mesenchymal state was the most predictable using the TME variables, suggesting that this is the most responsive to the environment, aligning with other research[167,313,314]. The fact that genomic factors also were important suggest potentially that the genomic factors influence the responsivity of the MES state to the environment. These alterations could be selected for in MES niches, allowing these

cells to survive. These genetic changes could provide stability to their otherwise plastic state, making MES cells more predictable when combining genomic and TME features.

The hybrid states are more difficult to predict (**Figure 57**). This suggests that they are less locked into a state, and have heightened plastic potential. Genomic factors do not appear to be able to predict these hybrid states, suggestive of the idea that hybrid states are not primarily driven by intrinsic genetic alterations but rather by dynamic interactions with the TME not captured within the cell type information or stochastic fluctuations in cellular signalling and reliance on post-transcriptional or epigenetic mechanisms. This aligns with the hypothesis that hybrid states exist in a more transient and flexible state, balancing epithelial and mesenchymal characteristics, and are influenced by a broader range of non-genetic factors, such as local cytokine gradients[315].

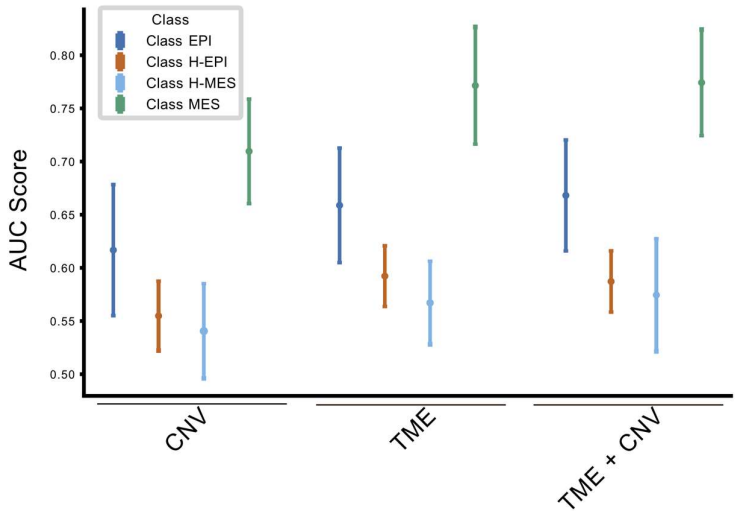

Figure 57. AUC range of predicted cell types in Xenium breast cancer dataset. The plot displays the average AUC scores across 10 different training splits, with error bars representing the standard deviation for the prediction model for each EMT state using just the CNV, just the TME and including both variables.

I then assessed the most important cells for predicting the four states (**Figure 58**). Cycling T-cells and NKT cells found within the top cells for MES state suggest an immune active state. Additionally, PVL immature cells were important, and these cells are often found in regions of active angiogenesis and have immune suppressive properties[316]. Myoepithelial cells may promote ECM remodelling, a hallmark of MES

states[317]. Interestingly the MES state had the highest number of significant nodes relating to the sub-clonal, intrinsic, information, with PC4 (which includes PIK3CA oncogene amplification in regions 3q26.1-q26.33), PC6 which includes oncogene amplification e.g. CCND1 at 11p, and PC7 which includes amplifications and deletions including regions with gene like BRCA1 (17q12-q21.1).

Hybrid MES cells appear to have more immune suppressive features, with myCAFs and macrophages forming the most important nodes. Both cell types are associated with promoting immune-suppressed environments[316,318]. Additionally, LYVE1+ endothelial lymphatic vessels were linked to this state, cells that have been associated with poor outcome in breast cancer, and strongly linked to increased metastasis[319]. A monoclonal antibody inhibiting LYVE+ has been shown to inhibit breast cancer tumour progression[319]. myCAFs, macrophages and endothelial cells have been characterised in other EMT spatial studies[110,167]. Hybrid EPI cells appear most dependent on cancer cycling cells, suggestive of a more proliferative state. Additionally, I detected a relationship with PVL Differentiated cells, which contribute to vascular structure and stability[320]. LYVE1+ endothelial lymphatic cells were also important for this state, as for hybrid MES states.

I observed that memory B-cells are most strongly associated with the EPI state. These antigen presenting cells indicate an adaptive immune response[321]. CXCL12+ and ACKR1+ endothelial cells are associated with chemokine signalling, and were significant cells liked to the EPI state[322,323]. These differ from the LYVE+ and RGS5+ endothelial cells, the endothelial cells linked to the hybrid states, which are linked to metastasis facilitation, the invasive edge and hypoxia adaption[319,324–326]. PC2 and PC5 were the more important genomic variables; PC2 is linked to preservation of critical regions (e.g., 1p and 3q), suggesting fewer genomic instabilities. PC5 also captures fewer genomic instabilities. These relationships suggest EPI states represent stable, well-differentiated epithelial cells interacting with immune and vascular systems to maintain homeostasis.

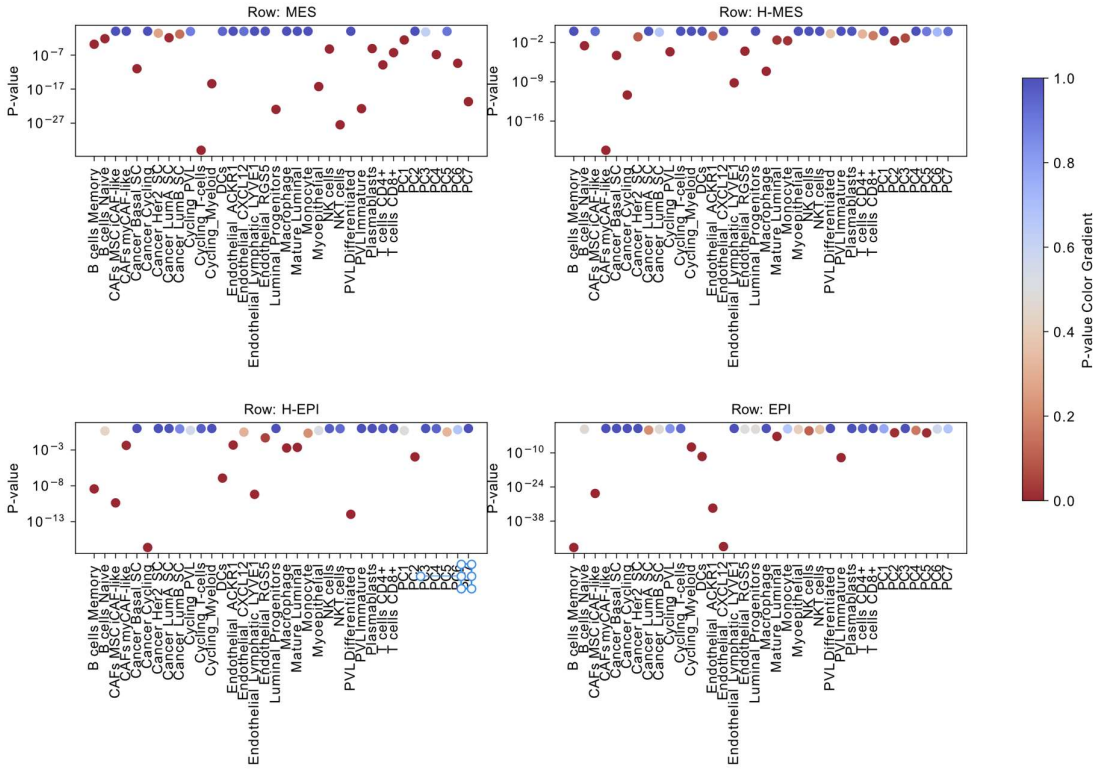

Figure 58. Cell type significance for each EMT state prediction (using the 4 state EMT model). Coloured by p-value significance level.

I also compared these explanations to a simplified 2-state EMT model, categorising cells into two bins based on their EMT signature scores (**Figure 59**). I found that the MES state explanations overlapped with those of the MES and Hybrid-MES states in the four-state model, including key cell types such as cycling T-cells, LYVE1+ endothelial cells, luminal progenitors, and myCAFs. Similarly, the EPI state explanations aligned with those of the EPI state in our model, featuring endothelial CXCL12+ cells, memory B-cells, and iCAFs. These suggest the stability of overall cell type trends, reinforcing the accuracy of our GNN explanation approach. However, it also shows that our four-state model provides additional granularity into the EMT process, revealing intermediate phenotypes and their distinct interactions within the TME.

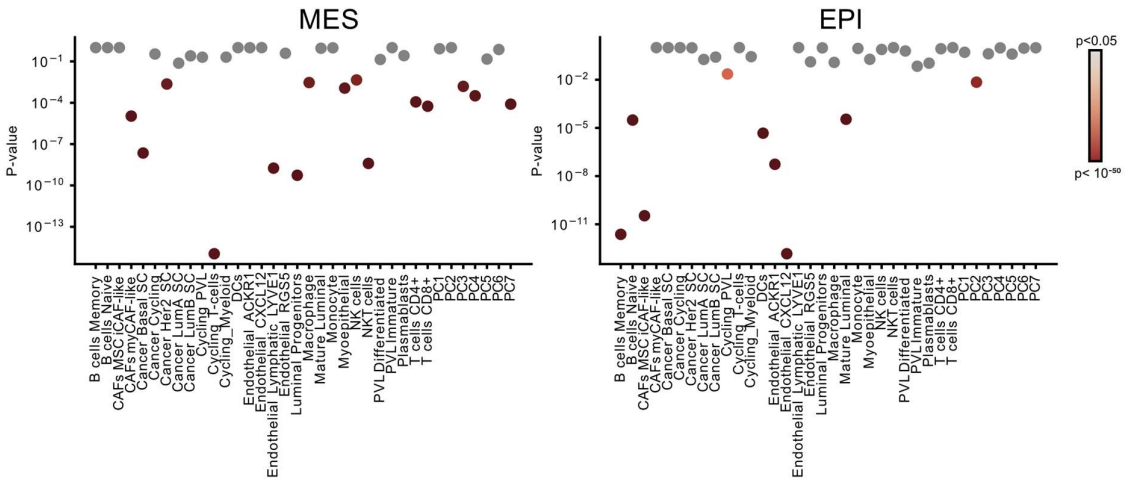

Figure 59. Cell type significance for each EMT state prediction (using the 2 state EMT model). Coloured by p-value significance level.

I then assessed the significant edge relationships (**Figure 60**). For the EPI state, interactions predominantly occur between epithelial subtypes, such as Lum A and HER2. Lum A shows the highest number of significant edges, which aligns with expectations, as it is the most epithelial-like breast cancer subtype[327]. The strong within-epithelial interactions, particularly Lum A to Lum A, make sense biologically, as these cells are likely located in the tumour core, less exposed to the TME. However, some immune infiltration is observed, including interactions with DCs and macrophages. Additionally, ACKR1+ endothelial cells form significant interactions, aiding with immune recruitment.

The hybrid EPI state, similar to the EPI state, predominantly involved interactions with epithelial cells. This suggests that hybrid EPI is closely related to the EPI state, potentially representing a transitional state still located within the tumour core. The low predictability of this state may indicate that it is a stochastic, intermediate state emerging from EPI. The hybrid MES state displays interactions with endothelial LYVE1+ cells, cancer cycling cells, PVL immature cells, and myoepithelial cells. These relationships highlight the hybrid state's increasing exposure to the TME and its intermediate nature between epithelial and mesenchymal traits.

The MES state has the largest number of significant edges, reflecting its proximity to the TME and its dependence on interactions with a broader range of cell types. This increased connectivity likely contributes to its higher predictability. Significant

interactions include those between RGS5+ endothelial cells and cancer epithelial as well as other endothelial cells. Additionally, mature luminal cells interact with cancer cells and CAFs, with myCAFs forming connections across a range of epithelial cancer cells. As expected, basal cancer epithelial cells exhibit the highest number of interactions, consistent with their invasive and mesenchymal-like characteristics[327].

These findings demonstrate how cell to cell interactions (edge explanations) uncover additional insights than solely focusing of individual cells (node importance).

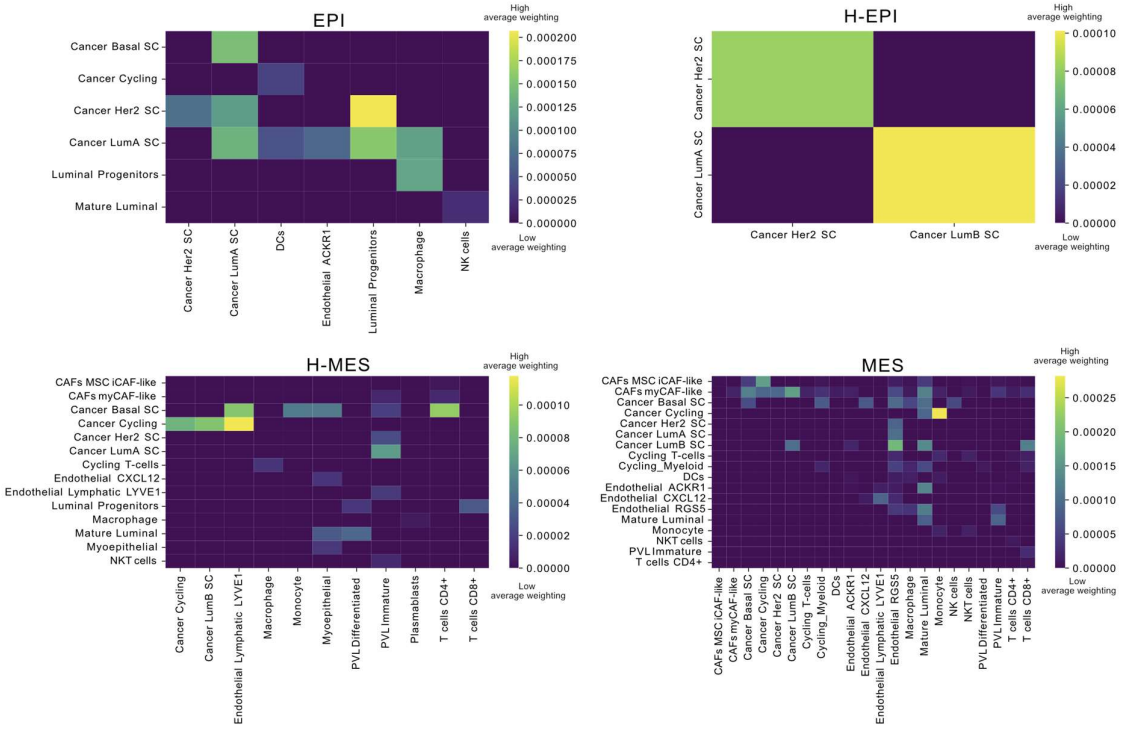

Figure 60. Cell type significance for each EMT state prediction using node explanations. Coloured by p-value significance level.

The results described above are from training and testing based on separate spatial splits; a process called inductive graph learning, which can be compared to spatial cross-validation used in geostatistics and ecological modelling[272]. However, transductive learning is another common approach in graph learning, where nodes are masked randomly within a graph (**Figure 61**). Within our problem setting, this can help gain additional clues in how much information is contained in the TME and the subclones.

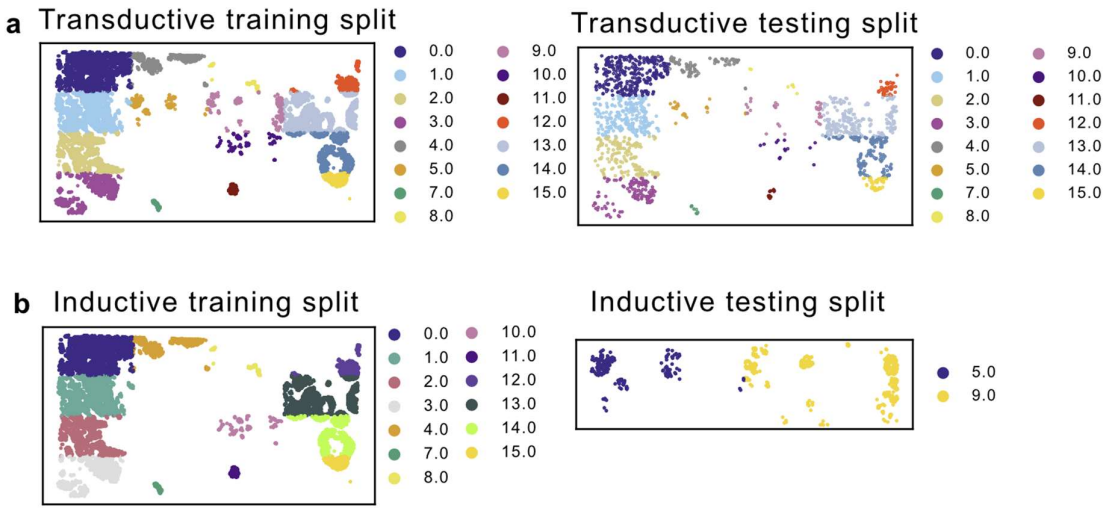

Figure 61. Different training splits used for the GNN prediction modelling task. a. Transductive training split involves randomly masking nodes (cells) across the dataset and predicting EMT state on the randomly masked nodes. Training and testing splits are coloured. **b**. Inductive training split involved spatially splitting the slide into train and test set. Training and testing splits are coloured.

I compared these two methods and found a similar overall trend in AUCs amongst the four EMT states, and minimal changes in AUC score (**Figure 62**). Inductive training as would be expected, has a larger error bar, due to more variation within the test set, suggesting some slight changes with spatial folds. However, the overall concordance suggests that there are unified common trends our GNN model is picking up that can be applied to unseen spatial regions as the inductive method of training shows.

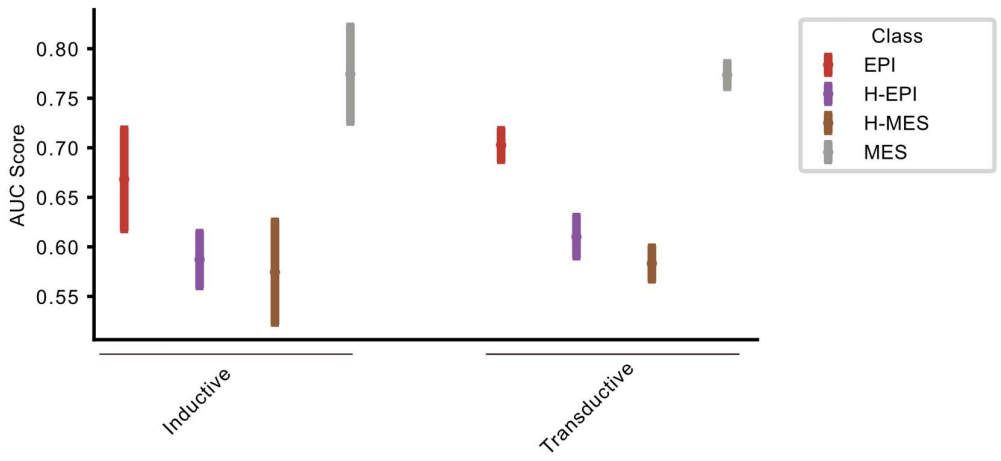

Figure 62. Comparison of inductive and transductive training split. AUC range of predicted cell types in Xenium breast cancer dataset. The plot displays the average AUC scores across 10 different training splits, with error bars representing the standard deviation for the prediction model for each EMT state using the TME and CNV.

Prediction model to understand EMT as a continuous process

I also tested the GNN approach on a continuous EMT score. Interestingly, the TME was much more important than the CNVs in predicting the EMT continuous score (**Figure 63**). This could indicate that the continuous score captures more transient or short-term stimuli from the TME, which aligns with the idea that EMT can be modulated rapidly in response to environmental conditions. In contrast, clonal information, which reflects stable, inherited genetic changes, appears to have a stronger association with large, more permanent EMT transitions (e.g., shifting between fully epithelial and fully mesenchymal states). These transitions may require more profound cellular reprogramming, which aligns with the longer timescales of clonal evolution compared to the immediate, dynamic nature of TME interactions.

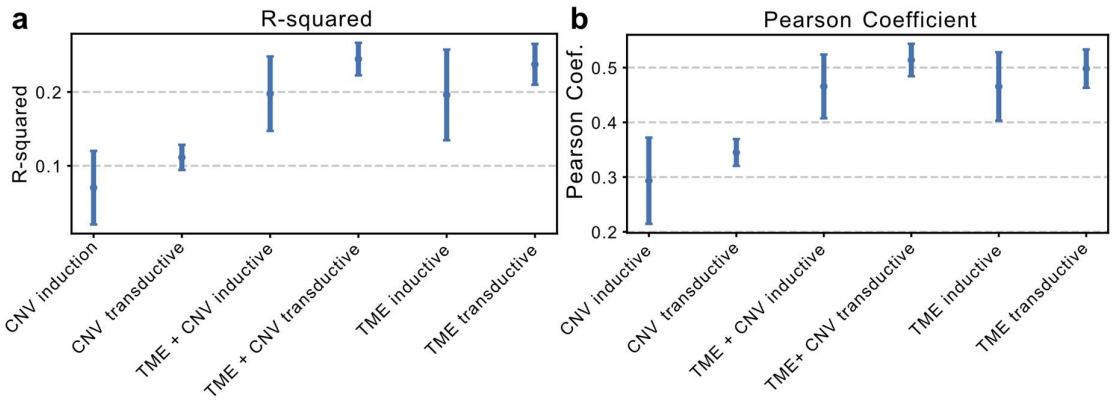

Figure 63. GNN regression performance using the different training splits and variables for prediction. a The average R^2 scores across 10 different training splits, with error bars representing the standard deviation for the continuous EMT score prediction using just the CNV, just the TME and including both variables. **b** The average correlation scores across 10 different training splits, with error bars representing the standard deviation for the continuous EMT score prediction using just the CNV, just the TME and including both variables.

4.3.3 Modelling the TME and genomic influences on EMP using spatial regression approaches

Rationale of spatial regression

While GNNs provide valuable insights into spatial structures by modelling the graph structure, enhancing our ability to capture spatial information for AUC comparisons and edge explanations, they have less statistical guarantees. For example, there are limitations in explicitly addressing confounding factors or statistically accounting for spatial effects and heterogeneity. Geostatistical models have been developed that can disentangle specific spatial influences and quantify spatial variability. For instance, spatial geostatistical regression methods can be employed to adjust for subtype-specific information, such as basal or HER2+, enabling the identification of key cell types and their contributions within these stratified spatial contexts.

Spatial error models (SEMs)

Spatial error models are a type of spatial regression that incorporate a spatial error term to account for unobserved spatial heterogeneity[89]. This allows us to estimate coefficients that represent conditional dependencies while adjusting for spatial autocorrelation.

I observe the TME variables explaining a larger portion of the variance in the regression model compared to clonal information, as observed in a higher R^2 value (**Figure 64**). This is similar to the trend observed in GNN regression. myCAFs emerge as the cell type with the greatest influence (**Figure 65**), an effect that persists even after controlling for subtype-specific effects (**Figure 66**). Furthermore, as clonal information remains significant when the effects of the TME are regressed out, this suggests that the inheritance of EMT is likely an important factor, independent of spatial colocation of similar clonal alterations. The λ (spatial error) coefficient is significant in all SEMs, highlighting the importance of accounting for spatial effects as they are a fundamental aspect of the underlying process being modelled.

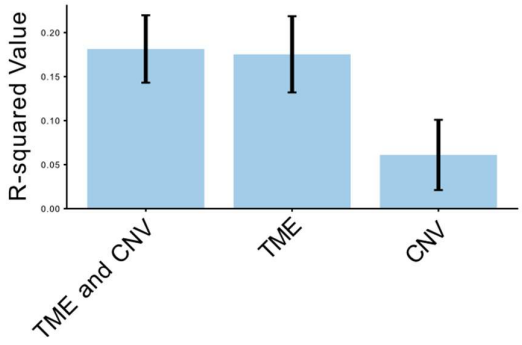

Figure 64. R^2 values for the SEM model using both extrinsic and intrinsic variables (TME and CNV), just extrinsic (TME) and just intrinsic (CNV). Error bars highlight standard deviation of the R^2 value after 10 spatial splits.

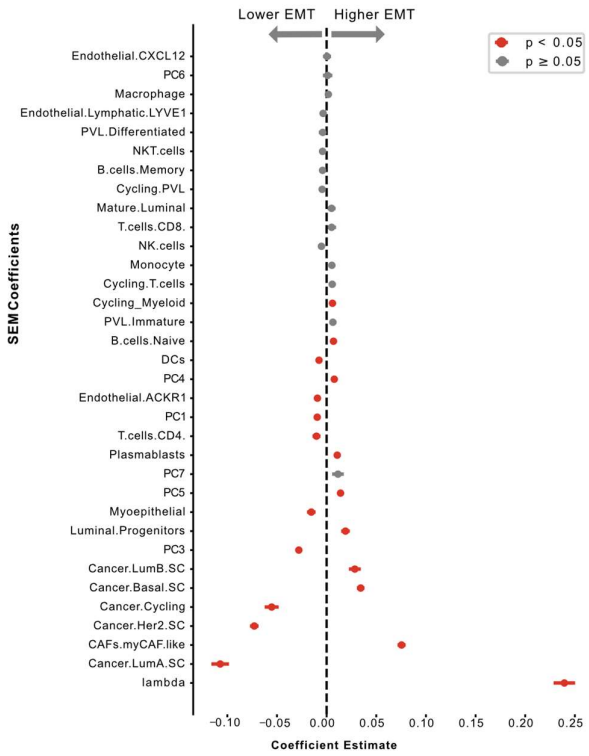

Figure 65. Spatial Error Model coefficient estimates using intrinsic and extrinsic variables, modelling across the whole slide. Red indicates a *significant coefficient*. Lambda is the spatial autoregressive coefficient for the error term (quantifies the degree of spatial dependence in the residuals).

Additionally, by regressing out the effects of DCIS versus invasive states (**Figure 66**), I find that invasive regions, as expected, are significantly closer to MES regions. Importantly, the relationship between MES cells and myCAFs remains robust despite these adjustments, reinforcing the biological significance of this interaction.

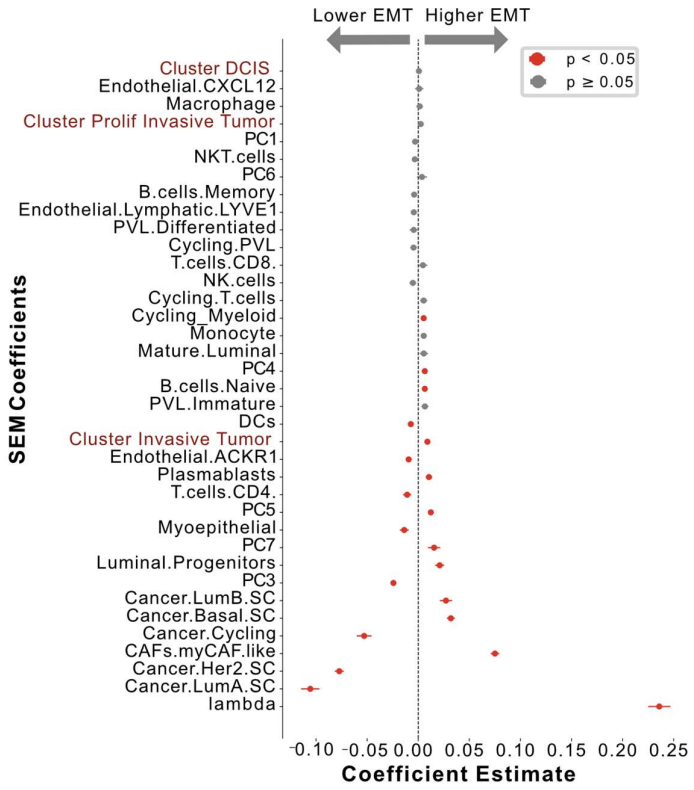

Figure 66. Spatial Error Model coefficient estimates combining intrinsic, extrinsic variables in addition to pathologist annotations of the regions within the tumour, modelling across the whole slide. Regions include proliferative invasive, invasive and DCIS (red font). Red coefficient indicates significance. Lambda is the spatial autoregressive coefficient for the error term (quantifies the degree of spatial dependence in the residuals).

To further explore these relationships, I performed separate SEMs for ductal and invasive regions (**Figure 67**). I found some similar trends, for example with myCAFs being linked with a mesenchymal phenotype in both. However, importantly, these analyses reveal some distinct relationships. For example, monocytes and macrophages appear to be linked to EMT within the DCIS regions but not the invasive regions. Interestingly, I find myoepithelial cells predominantly important in ductal regions, while their importance decreases significantly in invasive regions. Myoepithelial cells, which typically form a protective barrier in ductal regions and play a key role in maintaining epithelial integrity, are generally absent in invasive regions[328]. Their absence likely promotes tumour invasion and progression by reducing structural constraints and enabling epithelial cells to transition into more mesenchymal-like, migratory states[329].

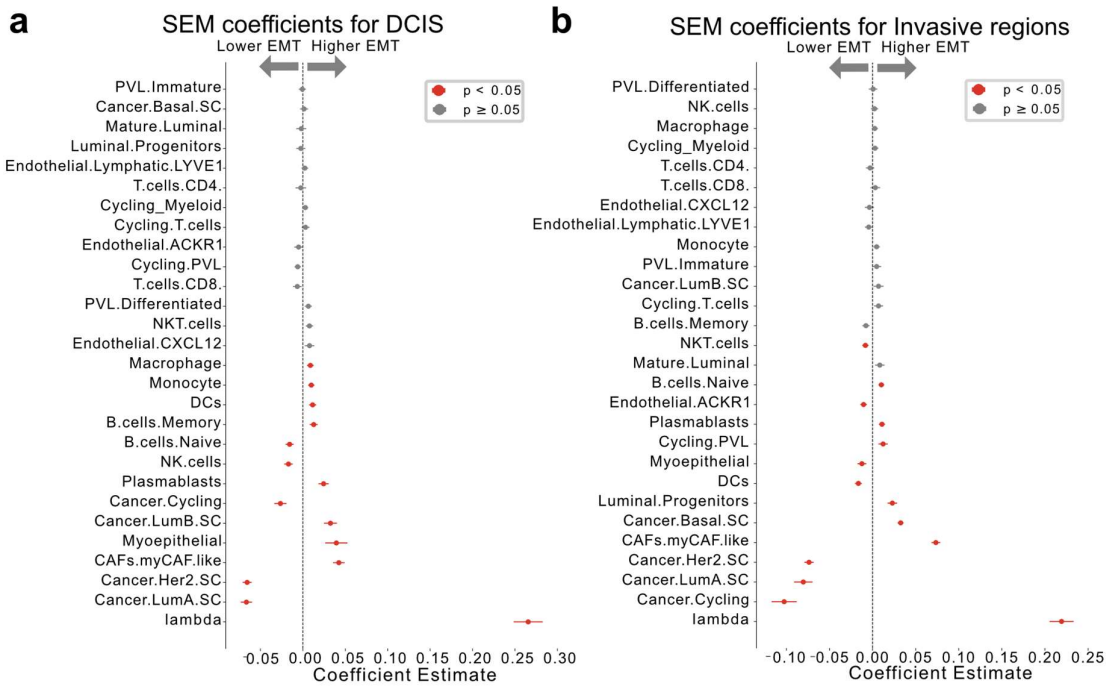

Figure 67. Spatial Error Model coefficient estimates using extrinsic variables, modelling across DCIS (a) and invasive regions (b) separately. Red indicates a significant coefficient. Lambda is the spatial autoregressive coefficient for the error term (quantifies the degree of spatial dependence in the residuals).

It is important to note that whilst they can provide additional insights that GNNs cannot, when comparing the overall regression metrics, our GNN regression approach outperforms SEM (**Figure 68**). This is as expected given that GNNs excel at capturing complex spatial relationships and multi-scale interactions (0.16 R^2 for SEM vs. 0.24 for GNN).

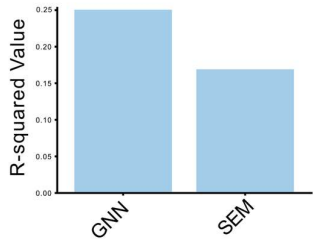

Figure 68. Comparison of GNN and SEM model fit on Xenium using R^2 value.

A comparison of Geographically Weighted Regression (GWR) and Multiscale GWR (msGWR) modelling

The example of the different weighting of myoepithelial cells depending on the region highlights spatial heterogeneity (also termed spatial nonstationary). Currently, the methods assessed so far assume that spatial relationships are uniform across locations. However, this is not how spatial relationships typically occur. Geographically weighted regression (GWR) is a technique that can assess the degree of spatial heterogeneity observed in the dataset[330]; for example, I can answer the question "do the TME cells have a different effect depending on the spatial location?"

The key advantage lies in moving beyond global averages to assess local relationships, capturing spatially varying dynamics that are important for a more detailed analysis. Traditional GWR applies a single bandwidth (spatial scale) to all variables, assuming that all relationships operate over the same spatial extent. This assumption can oversimplify spatial processes, as different variables often interact at different spatial scales. Multiscale Geographically Weighted Regression (MSGWR), a recent advance over GWR, addresses this limitation and allows each explanatory variable to have its own bandwidth, capturing relationships at the spatial scale most appropriate for that variable[271]. A larger range suggests it has an influence over a larger region within the tissue and a smaller range suggests the cell influences a smaller region within the tissue. This therefore increases the level of information I gain from the model. Understanding the bandwidth scale of cell type influence could help us understand whether localised versus systemic therapeutic strategies are important.

By assessing the R^2 of regression with and without GWR and then with and without multi-scale variable bandwidths, I can understand what processes are present in the data. I find there is a large degree of spatial heterogeneity within the tissue (**Figure 69a-b**), as GWR increases the variance captured (R^2) and lowers the Bayesian Information Criterion (BIC) compared to SEM. Interestingly, a multiscale approach does not offer an increased R^2 value, but it does lower the complexity of the model. This is reflected in the lower BIC value, which balances goodness of fit and model simplicity (fewer parameters) (**Figure 69b**). It is also reflected in the overall decreased number of parameters required for the model (**Figure 69c**). Therefore, shorter bandwidths for some variables can be equally informative, but this also can be captured in the large range too. However, the advantage is that msGWR can tell us the scale these cell operate at.

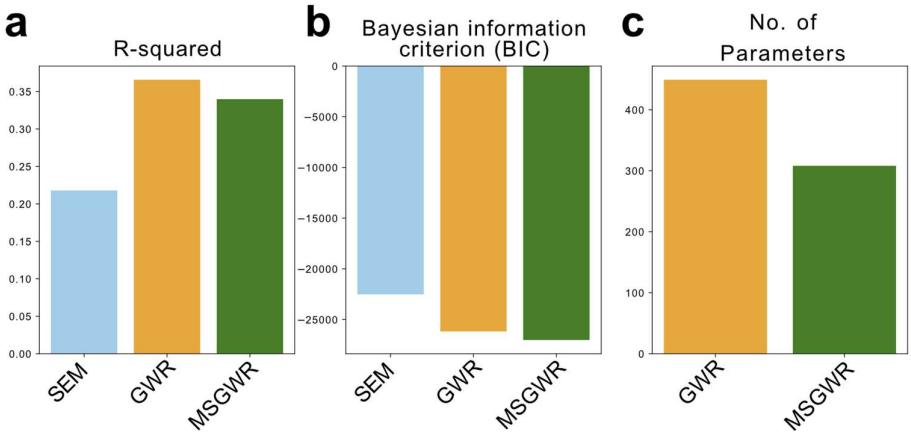

Figure 69. Metric comparisons for SEM, GWR and msGWR models fit on the DCIS regions within the Xenium slide. a. R^2 values for each model. **b.** Bayesian information criterion (BIC) values for each model**. c.** Number of parameters for GWR and msGWR. SEM is not included as it does not model spatially varying relationships.

This modelling approach confirms that myoepithelial cells have short range influence, as suggested in the SEM modelling approach, where they had a different effect depending on whether the tissue was in DCIS or invasive regions. It also highlighted that myCAFs display a longer-range influence on EMT (**Figure 70**). myCAFs contribute to ECM deposition and remodelling and therefore their influence over a larger range may potentially causes widespread stiffening of the TME, promoting EMT[331]. The subtype (e.g LumB or Basal) of neighbouring tumour cells displays local effects (small bandwidth). This smaller bandwidth suggests that the influence of the subtype is more pronounced only in certain, confined regions (niche specific effects).

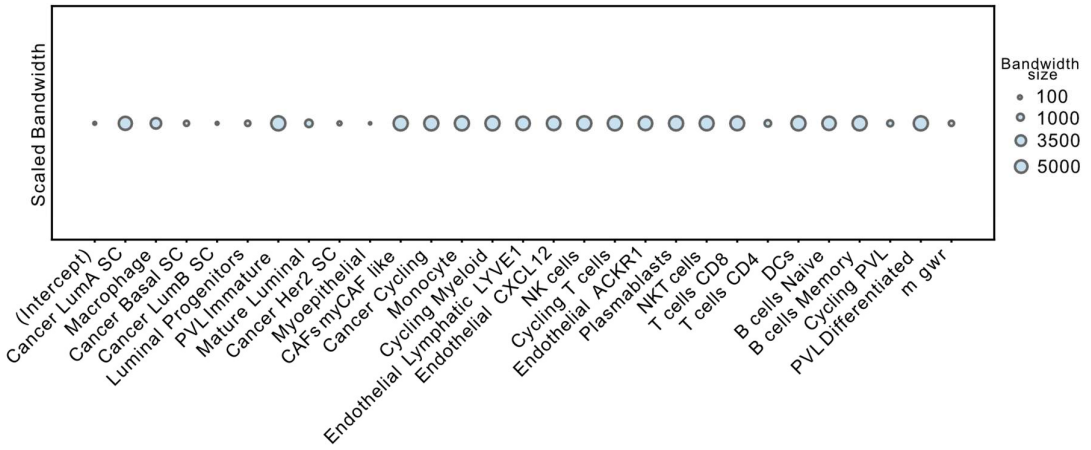

Figure 70. msGWR variable bandwidths (spatial length scale) in model used to predict EMT gradient in tumour cells. Bandwidth ranges relative to other variables in the msGWR model. A larger range suggests it has an influence over a larger region within the tissue and a smaller range suggests the cell influences a smaller region within the tissue.

The msGWR approach allows us to precisely identify and visualise the spatial localisation of important variables, such as highlighting the heterogenous influence of myCAFs, across the tissue slide (**Figure 71**). This modelling approach uncovers significant relationships in specific regions that might otherwise be missed in global regression models like SEM. For example, while CD8+ T-cell interactions were insignificant in the global SEM analysis, msGWR reveals significant localised relationships within specific regions (**Figure 72**).

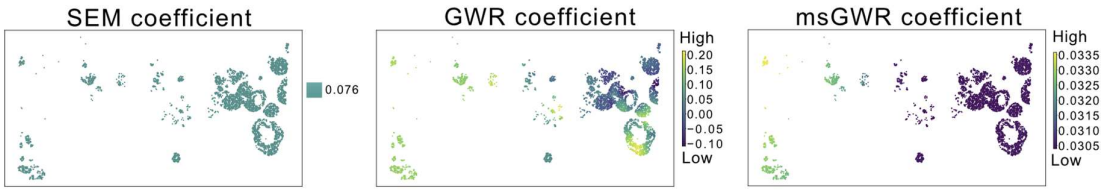

Figure 71. myCAF cell coefficients for SEM, GWR and msGWR visualised across the Xenium slide (DCIS region).

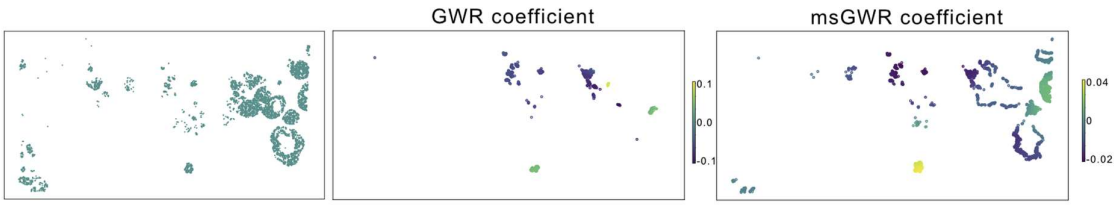

Figure 72. CD8+ T-cell coefficients for GWR and msGWR visualised across the Xenium slide. Xenium (DCIS region) for reference show in left plot.

4.4 Discussion

Many spatial studies within cancer analyse intrinsic (genomic) or extrinsic (TME) variables independently but do not integrate them into a unified framework that quantifies their relative contributions, or disentangles their effects. This gap limits our understanding of how these factors interact to drive cellular plasticity, such as EMP. Additionally, current plasticity metrics do not seek to model the environment. I have developed an approach to weigh these variables and regress out their shared contributions, to approximate which factors are closely linked to observed phenotypic changes. By addressing this challenge, we can move closer to fully understanding the mechanisms underlying EMP and other cell plasticity phenomena. I can also use this approach to rank states to understand how stable each cancer state is. This is useful from a treatment perspective as unstable, less predictable states are harder to develop drugs for. Understanding how these states could transition to more stable states could be a promising therapeutic avenue.

The MES state is the most predictable based on both the TME and genomic factors. This reflects its evolutionary adaptation and stability, indicating that it has been selectively shaped to thrive in the TME. As a result, the MES state is a more deterministic phenotype compared to others. The difference in predictability (AUC delta) between the MES and EPI states when using TME features highlights the plasticity of the EMT process. This delta could serve as a quantitative metric for plasticity potential, reflecting the extent to which cellular phenotypes adapt to TME pressures. It could suggest that CNVs are linked to an increased capacity for states to respond to environmental factors such as the TME. However, determining the direction of this relationship would require additional modelling or experimental validation through perturbation studies.

The EPI state, in contrast, appears to be less dependent on external pressures like the TME to maintain its phenotype. As a well-differentiated, proliferative state, it experiences less selective pressure from the environment. While CNVs may be present, they do not significantly alter the epithelial phenotype, allowing cells to retain their identity despite genomic variations. This aligns with the concept of the EPI state as a baseline, default state that is less adaptable or responsive to environmental changes[190,315].

Hybrid states, however, are the least predictable which suggests they are of a plastic nature. These states exist in a dynamic equilibrium, highly responsive to both intrinsic factors (e.g., stochastic gene expression, CNVs) and extrinsic cues (e.g., TME interactions). The observed dip in predictability across the EMT spectrum highlights the adaptive importance of plasticity. It suggests that retaining this variability in hybrid states is crucial for enabling cells to effectively respond to environmental challenges.

The CNV information was considerably less influential in predicting EMT continuous change compared to spatial variables. This could be due to continuous score capturing more subtle, short-term phenotypic changes, which the TME likely can induce, whereas the intrinsic variables, in this case the copy number alterations, may dictate more long-term, stable changes in cellular states

Using the AUC and R^2 metric provides an approach for evaluating whether the model effectively captures key spatial variables influencing cellular processes. For instance, future analyses could compare the inclusion of additional spatial variables, such as markers of hypoxia, to assess their impact on model performance. While cell type-level features may provide a baseline, they inherently have an upper limit in explanatory power. Capturing the full variables driving the EMT process likely requires modelling multiple layers within the tissue, such as chemokine gradients, matrix stiffness, and hypoxic gradients. I view this work as a step in building models capable of representing the diverse factors driving processes related to cell plasticity. Other metrics could additionally be used, for example entropy-based metrics in a similar manner as calculated by Burdziak et al[246] would be useful for capturing additional information returned from the GNN approach. This metric would quantify the entropy of the probability distribution across predicted classes. A high entropy value would

indicate greater plasticity, with probabilities distributed across multiple classes, suggesting the cell is less committed and more flexible.

As expected, GNNs outperform spatial regression models in prediction, and they offer valuable insights into the spatial interactions between cells, reflected in edge importance scores. While SEM offers more interpretable coefficients and robust statistical grounding, GNNs provide richer insights into complex spatial relationships. However, while GNNs can model larger spatial ranges, they are constrained by over-smoothing issues, which may be mitigated by graph transformers, a promising direction for future modelling efforts[332,333]. Combining the strengths of all approaches into one model could enhance the modelling approach going forward, for example, heterogeneity-aware deep learning spatial models[334,335]. It would also be useful to draw on these frameworks using causal modelling approaches which help assign causal direction to these relationships[336]. In a recent extension of the SIMVI framework, a metric has been introduced that effectively disentangles intrinsic and spatial effects[275]. While this method does not fully achieve the goal of producing interpretable coefficients for intrinsic and extrinsic influences, it provides a metric for capturing their overall effects. Applying this metric to our dataset could offer additional insights into the balance between intrinsic and extrinsic factors driving EMP.

I can confirm the relationship between myCAFs and EMT in a way that bulk transcriptomics and Visium cannot, due to their resolution limitations. Bulk transcriptomics averages gene expression across diverse cell types, masking cell-type-specific mesenchymal contributions. While Visium provides spatial context, its near-single-cell resolution still allows shared mesenchymal programs across different cell types to confound the analysis. In contrast, our approach, using single-cell resolved spatial transcriptomics, enables precise disentanglement of these signals, overcoming these limitations. I further validated the CAF and mesenchymal tumour cell annotations by identifying specific markers from a carefully annotated single-cell RNA sequencing dataset. However, the annotations in this dataset were computationally derived. Consequently, while these markers provide an important additional layer of evidence, they do not represent definitive ground truth. Future studies should aim to validate these markers using scRNA-seq data from experimentally sorted cell types to confirm their specificity and accuracy.

Additionally, ligand-receptor analysis would help further understand the myCAF-EMT relationship, providing a more mechanistic understanding of the interaction, and in turn potentially providing therapeutic targets to disrupt this interaction.

I identified four EMT states by binning the signature enrichment score into four groups. This straightforward binning approach was used as a proof of concept to validate the methodological framework and to provide a clear, interpretable means of distinguishing EMT states. Earlier work in this thesis involved a more sophisticated gaussian mixture modelling approach to define EMT states, and I envisage that future work, should use the more advanced GMM approach, to more accurately capture EMT states.

By framing EMT as a spatial prediction problem, our approach can also identify cells located within a typical MES-supportive niche that have not transitioned to a MES state. These cells are of particular interest because they defy the predicted spatial relationships that suggest they should exhibit a MES phenotype. Understanding the properties of these resistant cells could provide critical insights into mechanisms that inhibit EMT.

Future research should aim to validate these findings using additional Xenium datasets. While our study used a single Xenium dataset that captured a large spatial range within a tumour, including DCIS and invasive regions, this came at the expense of representing inter-patient heterogeneity. To ensure the robustness and generalisability of our conclusions, it is important to repeat these results across multiple independent datasets that encompass a diverse range of patient samples and tumour microenvironments. Additionally, confirming these relationships with experimental approaches would be highly important.

Using copy number alterations is a proof of concept for capturing the intrinsic effects and I envisage improving the model with additional mutational information and other intrinsic effects. Single cell spatial genomic information would be important for better inferences of intrinsic factors. Currently, clonal information is approximated from spot profiles at the Visium resolution, which may overlook finer clonal details and be biased by genes that are colocalised on chromosomes within certain cell populations. To handle the high dimensionality and multicollinearity of the CNV data, I used PCA for

dimensionality reduction. This method transforms the data into a smaller set of uncorrelated features that retain most of the variation. The trade-off, however, is that each principal component represents a combination of many CNVs across different genes, making it difficult to trace back to specific gene-level gains or losses. Experimenting with alternative methods for dimension reduction of clonal information, e.g. non-negative matrix factorisation, where each factor can be non-negative for enhanced interpretability, would be important. Additionally, focusing on a targeted list of genes derived from thorough literature review might further enhance interpretability while mitigating issues arising from correlated features.

Finally, it would be useful to rank different plastic and non-plastic processes to understand how they compare against each other. A useful comparison, for instance, would be to compare EMT to other cellular plasticity programmes, such as dedifferentiation, and compare to non-plastic processes like terminal differentiation, where cells commit to a fixed identity.

Conclusions

In this chapter, I have explored how to quantify and interpret the contributions of cell-intrinsic and cell-extrinsic factors to cellular phenotypes, using epithelial-to-mesenchymal plasticity as a case study. By drawing upon concepts from geostatistics, ecology, and GeoAI, I introduced a framework that moves beyond descriptive methods and enables a more integrated, quantitative understanding of the TME's role in shaping cellular states. This approach highlights the importance of the local environment in driving subtle, short-term phenotypic shifts, as well as the influence of genomic alterations that contribute to more stable, long-term changes. By applying these models, I have gained insights into how certain states, such as the mesenchymal phenotype, are more deterministic, while hybrid states are less predictable and thus potentially more adaptable. This potentially opens avenues for identifying more effective therapeutic strategies.

PC1	PC2	PC3	PC4	PC5	PC6	PC7
Chrom.11 (q13.2-q13.4) (-) MIN 0 MAX 2	Chrom.3 (q23-q25.31) (+) MIN -1 MAX 0	Chrom.7 (q22.1-q22.1) (-) MIN 0 MAX 1	Chrom.5 (q11.2-q13.2) (-) MIN 0 MAX 1	Chrom.7 (q36.1-q36.1) (+) MIN -2 MAX 0	Chrom.7 (q31.1-q31.32) (-) MIN -1 MAX 0	Chrom.5 (q23.2-q31.3) (+) MIN 0 MAX 1

Chrom.2 (p25.3-p22.1) (-) MIN 0 MAX 1	Chrom.10 (q24.31-q24.32) (-) MIN 0 MAX 2	Chrom.10 (p12.1-q11.21) (-) MIN 0 MAX 1	Chrom.3 (p25.3-p24.1) (+) MIN -1 MAX 0	Chrom.11 (q13.4-q25) (+) MIN -2 MAX 0	Chrom.2 (q31.2-q32.2) (-) MIN -2 MAX 0	Chrom.5 (q31.3-q31.3) (+) MIN -2 MAX 0
Chrom.2 (p25.3-p13.1) (-) MIN 0 MAX 2	Chrom.3 (q23-q25.32) (+) MIN -1 MAX 0	Chrom.16 (p13.3-p13.3) (-) MIN 0 MAX 2	Chrom.10 (q26.11-q26.2) (-) MIN 0 MAX 1	Chrom.7 (p11.2-q11.23) (-) MIN 0 MAX 1	Chrom.11 (p13-q12.1) (+) MIN 0 MAX 2	Chrom.1 (q21.3-q23.1) (-) MIN 0 MAX 2
Chrom.2 (p25.3-p13.3) (-) MIN 0 MAX 1	Chrom.4 (q21.3-q23) (+) MIN -2 MAX 0	Chrom.2 (q34-q35) (-) MIN 0 MAX 2	Chrom.13 (q12.3-q22.2) (+) MIN -2 MAX 0	Chrom.8 (q24.3-q24.3) (-) MIN 0 MAX 2	Chrom.16 (q11.2-q13) (+) MIN 0 MAX 1	Chrom.19 (p13.3-p13.3) (+) MIN 0 MAX 2
Chrom.1 (q41-q44) (-) MIN 0 MAX 2	Chrom.5 (p13.2-q11.2) (+) MIN -1 MAX 0	Chrom.5 (q31.3-q33.3) (+) MIN -2 MAX 0	Chrom.18 (q21.1-q21.33) (-) MIN 0 MAX 2	Chrom.10 (p15.3-p15.2) (-) MIN 0 MAX 2	Chrom.4 (q25-q28.2) (-) MIN -1 MAX 0	Chrom.7 (p22.3-p21.2) (+) MIN 0 MAX 1
Chrom.11 (p15.5-p15.4) (+) MIN -2 MAX 0	Chrom.8 (q13.1-q22.1) (+) MIN -1 MAX 0	Chrom.1 (q21.3-q23.1) (+) MIN 0 MAX 2	Chrom.12 (p13.31-q12) (+) MIN -2 MAX 0	Chrom.10 (p15.3-p15.1) (-) MIN 0 MAX 2	Chrom.18 (q12.2-q21.33) (-) MIN 0 MAX 2	Chrom.7 (p22.3-p21.1) (+) MIN 0 MAX 1
Chrom.9 (p24.3-p21.3) (+) MIN -2 MAX 0	Chrom.1 (p31.3-p22.1) (+) MIN -2 MAX 0	Chrom.5 (q31.3-q31.3) (-) MIN -2 MAX 0	Chrom.8 (q13.2-q21.13) (+) MIN -2 MAX 0	Chrom.10 (q22.2-q23.2) (-) MIN 0 MAX 1	Chrom.11 (q13.4-q23.1) (+) MIN -2 MAX 0	Chrom.17 (q12-q21.1) (-) MIN 0 MAX 2
Chrom.13 (q13.1-q34) (+) MIN -2 MAX 0	Chrom.7 (p22.3-p21.1) (-) MIN 0 MAX 1	Chrom.18 (p11.32-p11.22) (-) MIN -1 MAX 0	Chrom.3 (q26.1-q26.33) (-) MIN 0 MAX 1	Chrom.2 (q37.1-q37.3) (-) MIN 0 MAX 2	Chrom.1 (q25.3-q32.1) (-) MIN -1 MAX 0	Chrom.8 (p23.3-p21.2) (-) MIN -1 MAX 0
Chrom.17 (q25.3-q25.3) (-) MIN 0 MAX 2	Chrom.7 (p22.3-p21.2) (-) MIN 0 MAX 1	Chrom.3 (p25.3-q12.1) (-) MIN -2 MAX 0	Chrom.1 (q25.3-q32.1) (+) MIN -1 MAX 0	Chrom.18 (q12.2-q21.33) (-) MIN 0 MAX 2	Chrom.3 (q26.1-q26.33) (+) MIN 0 MAX 1	Chrom.8 (p23.3-p21.1) (-) MIN -2 MAX 0
Chrom.12 (q24.23-q24.31) (-) MIN 0 MAX 1	Chrom.12 (p13.31-p13.2) (-) MIN -2 MAX 0	Chrom.11 (q12.1-q12.3) (-) MIN -1 MAX 0	Chrom.6 (q23.2-q24.1) (-) MIN 0 MAX 2	Chrom.11 (q13.4-q23.1) (+) MIN -2 MAX 0	Chrom.6 (p22.3-p21.31) (-) MIN 0 MAX 2	Chrom.17 (p13.1-p11.2) (-) MIN -1 MAX 0
Chrom.12 (q12-q13.12) (-) MIN 0 MAX 1	Chrom.3 (q13.2-q13.33) (-) MIN -2 MAX 0	Chrom.17 (q24.3-q25.3) (-) MIN -1 MAX 0	Chrom.18 (p11.32-p11.22) (+) MIN -1 MAX 0	Chrom.3 (p25.3-p24.1) (-) MIN -1 MAX 0	Chrom.1 (q32.1-q32.1) (+) MIN 0 MAX 1	Chrom.8 (p23.3-p11.22) (-) MIN -2 MAX 0
Chrom.11 (q12.1-q13.1) (+) MIN -2 MAX 0	Chrom.17 (q12-q12) (-) MIN -2 MAX 0	Chrom.8 (q21.11-q21.2) (-) MIN -1 MAX 0	Chrom.3 (p25.3-q12.1) (+) MIN -2 MAX 0	Chrom.10 (q26.11-q26.2) (+) MIN 0 MAX 1	Chrom.3 (q13.2-q13.33) (-) MIN -2 MAX 0	Chrom.17 (q22-q24.2) (-) MIN 0 MAX 2
Chrom.6 (q23.3-q24.2) (-) MIN 0 MAX 2	Chrom.1 (q32.1-q32.1) (+) MIN 0 MAX 1	Chrom.6 (q23.2-q24.1) (+) MIN 0 MAX 2	Chrom.17 (q24.3-q25.3) (+) MIN -1 MAX 0	Chrom.12 (p13.31-p13.2) (+) MIN -2 MAX 0	Chrom.17 (q12-q12) (-) MIN -2 MAX 0	Chrom.3 (q26.1-q26.33) (+) MIN 0 MAX 1
Chrom.1 (p31.3-p13.2) (+) MIN -2 MAX 0	Chrom.9 (q34.11-q34.3) (-) MIN 0 MAX 2	Chrom.1 (p31.1-p22.3) (-) MIN -2 MAX 0	Chrom.11 (q12.1-q12.3) (+) MIN -1 MAX 0	Chrom.17 (p13.1-p11.2) (+) MIN -1 MAX 0	Chrom.2 (q34-q35) (+) MIN 0 MAX 2	Chrom.1 (q25.3-q32.1) (-) MIN -1 MAX 0
Chrom.13 (q13.1-q22.2) (+) MIN -2 MAX 0	Chrom.16 (q22.1-q22.1) (-) MIN 0 MAX 1	Chrom.1 (p13.3-p12) (-) MIN -2 MAX 0	Chrom.1 (p13.3-p12) (+) MIN -2 MAX 0	Chrom.1 (q32.1-q32.1) (-) MIN 0 MAX 1	Chrom.5 (q31.3-q33.3) (-) MIN -2 MAX 0	Chrom.8 (q21.11-q21.2) (-) MIN -1 MAX 0
Chrom.19 (p13.11-p13.11) (-) MIN 0 MAX 2	Chrom.19 (p13.3-p13.2) (-) MIN 0 MAX 2	Chrom.5 (q31.1-q31.2) (-) MIN 0 MAX 1	Chrom.8 (q21.11-q21.2)	Chrom.3 (q13.2-q13.33) (+) MIN -2 MAX 0	Chrom.3 (p13-q12.1) (-) MIN -2 MAX 0	Chrom.6 (q23.2-q24.1) (+) MIN 0 MAX 2

			(+) MIN -1 MAX 0			
Chrom.4 (p14-p13) (-) MIN 0 MAX 1	Chrom.15 (q11.2-q15.1) (+) MIN -1 MAX 0	Chrom.3 (q26.1-q26.33) (+) MIN 0 MAX 1	Chrom.1 (p31.1-p22.3) (+) MIN -2 MAX 0	Chrom.17 (q12-q12) (+) MIN -2 MAX 0	Chrom.5 (q31.3-q31.3) (+) MIN -2 MAX 0	Chrom.11 (q12.1-q12.3) (-) MIN -1 MAX 0
Chrom.1 (p36.33-p36.21) (+) MIN -2 MAX 0	Chrom.16 (q22.1-q22.2) (-) MIN 0 MAX 1	Chrom.1 (q25.3-q32.1) (-) MIN -1 MAX 0	Chrom.11 (q13.4-q25) (+) MIN -2 MAX 0	Chrom.2 (q11.2-q13) (+) MIN -1 MAX 0	Chrom.14 (q12-q12) (-) MIN -1 MAX 0	Chrom.1 (p13.3-p12) (-) MIN -2 MAX 0
Chrom.1 (p36.33-p36.11) (+) MIN -2 MAX 0	Chrom.15 (q11.2-q14) (+) MIN -1 MAX 0	Chrom.3 (p21.31-p21.31) (-) MIN 0 MAX 2	Chrom.7 (q36.1-q36.1) (+) MIN -2 MAX 0	Chrom.13 (q12.12-q12.3) (-) MIN 0 MAX 1	Chrom.5 (q31.3-q35.1) (-) MIN -2 MAX 0	Chrom.1 (p31.1-p22.3) (-) MIN -2 MAX 0
Chrom.1 (p36.33-p36.12) (+) MIN -2 MAX 0	Chrom.19 (p13.3-p13.11) (-) MIN 0 MAX 2	Chrom.5 (q31.3-q33.1) (+) MIN -2 MAX 0	Chrom.10 (q22.2-q23.2) (-) MIN 0 MAX 1	Chrom.6 (p21.32-p21.31) (-) MIN 0 MAX 1	Chrom.3 (q21.3-q22.1) (+) MIN 0 MAX 2	Chrom.18 (p11.32-p11.22) (-) MIN -1 MAX 0

Supplementary Table 1: Chromosomal alterations linked to each principal component. Chromosome and chromosomal region are annotated for each principal component, along with the direction of alteration, where negative indicates a loss, and a positive sign indicates a gain. The rank of values are also indicated.

5 Chapter 5: Discussion

5.1 Summary and conclusions

When I began my PhD in 2020, spatial transcriptomics was still an emerging field. Methodological development was in its early stages, and many of the initial analytical approaches did not fully integrate the spatial dimension of the data. For instance, spatially aware clustering was not commonly performed, with clustering often applied in a way that overlooked spatial context. Additionally, EMT had not yet been characterised in human tissues using more than a handful of markers, and no studies had systematically analysed the spatial relationships with EMT states. Spatial transcriptomics provides a powerful approach to studying biological processes in ways that traditional model systems cannot, by capturing cellular interactions within intact human tissue and its native tumour microenvironment. While it is important to recognise the limitations of snapshot-based data in providing causal or mechanistic insights, spatial transcriptomics can complement model system findings, reinforcing and contextualising experimental observations.

In this thesis, I focused on understanding how epithelial-to-mesenchymal plasticity interacts with the tumour microenvironment and how genomic alterations shape these relationships, primarily focusing on breast cancer spatial transcriptomic data. In the initial chapters, I described how EMT can be viewed not merely as a binary process but rather as a set of discrete states that allows tumour cells to flexibly transition between epithelial, hybrid, and mesenchymal states (Chapters 2 and 3). I then demonstrated that these states display unique spatial relationships with components of the TME. Notably, I found that cells undergoing EMT often colocalise with immunosuppressive niches containing myofibroblastic cancer-associated fibroblasts and macrophages (Chapters 3).

A large focus has been on developing and applying new analytical tools. I developed the SpottedPy Python package (Chapter 3), to analyse spatial hotspots of tumour and microenvironmental features. I then used SpottedPy to assess relationships at multiple scales, from the immediate cellular neighbourhood to broader tissue-wide clusters, to understand both inter- and intra-tumour heterogeneity.

The package is supported by user-friendly notebooks to promote reproducibility and to encourage the use on a wide-range of biological questions extending to other

diseases and spatial platforms. Applying SpottedPy to Visium breast cancer datasets highlighted associations between EMT hotspots, hypoxia and angiogenesis, as well as links with immunosuppressive stromal and immune cell populations (Chapters 3), substantiating the results from Chapter 2. By transferring EMT signatures representing distinct states onto spatial transcriptomic spots (Chapter 3), I highlighted the varied spatial relationships of different EMT states, providing insights into potential therapeutic vulnerabilities tied to each phenotype.

Finally, in Chapter 4, I developed a predictive modelling approach to quantitatively assess the contributions of genomic and environmental variables on EMT using a single-cell-resolution spatial breast cancer dataset. This novel approach allows us to move beyond descriptive methods and enables a more integrated, quantitative understanding of the TME's role in shaping cellular states. By drawing on geostatistical methods and graph neural networks, I compared how well each factor, copy number alterations or microenvironmental signals, could explain EMT states.

Overall, the approach strengthens the evidence that targeting the TME is more important for targeting EMT as opposed to targeting the genomic factors. It highlights the importance of the TME in inducing both subtle, short-term changes and stable, long-term phenotypic change, whereas genomic alterations primarily contribute to more stable, long-term changes. I have shown that the mesenchymal phenotype is more deterministic, while hybrid states are less predictable and thus potentially more adaptable. I suggested that EMP hybrid states may be harder to therapeutically target due to their unpredictability.

I highlighted how Geographically Weighted Regression can reveal the degree of spatial heterogeneity within tumours, and I found that relationships between EMT states and particular TME populations do vary across different tissue regions. I highlight how framing spatial relationships in this manner can help to further understand approaches to target EMP. For example, I show how EMT relationships change across DCIS and invasive regions within the tissue, most noticeably with myoepithelial cells, where they are significantly associated with EMT in DCIS, and in GWR modelling had localised effects. Therefore, this relationship would require more localised targeting. Furthermore, I highlight how it is possible to confirm the association between myCAFs and EMT in a way that is not possible with bulk transcriptomics and

Visium (near-single cell resolution), where common mesenchymal programs can confound the analysis.

Collectively, these chapters examine EMT across multiple biological resolutions, from single-cell to spatially resolved data, and highlight that EMT is governed by heritable genetic events and local environmental cues. I demonstrate the value of combining geospatial statistics, GeoAI approaches, and transcriptomics to deepen our understanding of EMP and suggest potential therapeutic insights.

5.2 Limitations + Future directions

5.2.1 Tackling EMT challenges

Whilst there are extensive studies linking EMT to metastasis and chemoresistance, it is still a controversial process[337]. This is as some lineage-tracing experiments did not find EMP to be an important part of dissemination, and epithelial features have been shown to be important for migration[20,337,338]. These findings lead to the focus on hybrid EMT states, which retain some epithelial traits, whilst also retaining the migratory traits of mesenchymal cells. I have shown that hybrid states have distinct TME interactions, hinting at potential therapeutic options to target these states. However, I also found that these states are the hardest to predict and are likely more transient and plastic.

Importantly, part of the controversies stem from the difficulty of correctly defining EMP. As shown in this thesis, and other studies by Pastushenko et al.[20], Goetz et al.[40] and Brown et al.[166], there are multiple hybrid EMT states that exist. The diversity of EMT intermediate states is only beginning to be uncovered, and a deeper understanding of their frequency and context is needed to clarify how these transitions occur. I used a range of approaches to characterise EMP and cell states, from gene signature scoring, NMF and gaussian mixture modelling. The gene signature method offers a straightforward approach whereas the gaussian mixture modelling providing a more complex, but more statistically sound approach. However, each method relies on different interpretations of the data and more statistically robust frameworks should be developed in the future. Building on frameworks such as those in CELLSTATES, a statistically principled method to capture states at the maximum likelihood could lead to better clarity[53]. In CELLSTATES, the authors show that given the known measurement noise structure of scRNA-seq data, this problem is mathematically well-

defined and they derive its unique solution from first principles. This allows for a parameter-free approach, less subject to varying interpretations for state identification.

Additionally, given the fact that many gene programs are redundant, and other levels of gene regulation extend beyond gene expression, identifying EMT at the morphological and protein level is important for more accurately identifying EMP. Enhancing the detection of EMT, for example by matched high resolution imaging with spatial transcriptomic data would enable phenotypic changes to be detected and linked with the transcriptomic changes. With the rise of spatial proteomics, which was named "Method of the Year" in 2024, this would allow us to confirm EMT at a protein level[339]. Whilst some studies have shown that gene expression and protein levels do correlate, other work has shown poor correlations with certain genes and proteins because of post-transcriptional regulation[340]. Therefore, combining these modalities offers additional confirmation and allows us to investigate how EMT genes correlate to actual protein expression in the cells.

Recently, new developments have enabled even more modalities (DNA, chromatin accessibility, and histone modification) to be captured simultaneously on the same tissue section[341]. Given that EMT is governed by epigenetic mechanisms, this multi-levelled approach, capturing the range of epigenetic mechanisms, could capture more accurate EMT states. While gene expression data provide snapshots of transcriptional activity, they often fail to distinguish between transient fluctuations and stable regulatory changes that define EMT plasticity[342]. In contrast, epigenetic modifications, such as chromatin accessibility, histone modifications, and DNA methylation, serve as more stable indicators of a cancer cell’s state, and therefore offers a more accurate approach to understand EMP.

The difficulty in accurately detecting EMT also stems from the varying EMT programmes that occur in different contexts[26], for example different tumour stages, and different tumour sites can have different EMT programmes. Whilst I have investigated a range of EMT signatures, it is important to deepen our understanding of the impact of using different EMT gene signatures on the downstream analysis. Recently, I have been involved in a collaboration developing a language model-based prediction model to detect common EMT programmes across tissue types[168]. We used a pre-trained single-cell large language model (LLM) to develop an EMT-language

model (EMT-LM), which was able to identify discrete EMT states within the EMT continuum within single-cell RNA-seq data. The model was able to classify EMT states with high accuracy (AUROC of 90%) across multiple cancer types. Further spatial analysis of these signatures would provide valuable insights into how different EMT programmes affect downstream spatial analysis results. In particular, it would important to conduct extensive benchmarking of the spatial relationships of EMT signatures using well-curated ground truth datasets that are both in situ and specific to the cancer type under investigation.

5.2.2 Spatial transcriptomic methodological challenges

There are various limitations within the spatial transcriptomic analysis. I have extensively studied spatial relationships in whole-transcriptome spatial data, but at a resolution of around 10 cells per spatial spot (Chapter 2 and Chapter 3), and I studied single-cell resolved spatial data but not at the whole transcriptome level (Chapter 4). Therefore, future work should validate these findings on single-cell spatial whole transcriptome emerging technologies. For example, Visium HD was released recently[343], offering single cell whole transcriptome spatial analysis. It would therefore be valuable to repeat our analysis with a breast cancer Visium HD dataset.

A key challenge in accurately identifying EMT in non-single cell resolved spatial transcriptomics is the overlap of mesenchymal markers between CAFs and tumour cells undergoing EMT. Throughout this thesis, I have taken steps to mitigate this issue to the best of my ability. For instance, when deconvolving the Visium dataset, I used scRNA sequencing data labelled with EMT states before deconvolution, rather than assigning EMT states to tumour cells after deconvolution. This approach allows for a more comprehensive gene expression profile to distinguish between the cells, improving the distinction between CAFs and tumour cells. Additionally, I validated these findings in the Xenium single-cell resolved data, which ensures that the gene expression signatures are attributed to individual cells. However, further validation using larger single-cell datasets will be crucial to strengthening these results.

Furthermore, incorporating ligand-receptor signalling information into the evaluation of spatial effects on cell populations will be important to increase the confidence in the identified relationships. Ligand-receptor signalling can capture the functional crosstalk

between cell populations, and can better assess whether identified spatial clusters are biologically interconnected or merely spatially co-located without functional interaction.

I inferred genomic variables (CNVs) from the transcriptomic data alone, which may introduce inaccuracies due to the reliance on expression data. For instance, if closely situated genes show correlated expression changes because of a shared regulatory process, rather than an actual copy number alteration, this would yield a false-positive CNV signal. Therefore, integrating matched spatial transcriptomic data with spatial genomic data would greatly enhance confidence in these inferred genomic relationships. I used PCA to transform the high-dimensional CNV data into a smaller set of uncorrelated features. While this retains the majority of variation in the data, each principal component represents a combination of many CNVs across different genes, making it difficult to trace back specific gains or losses at the gene level. This was chosen as the approach as the CNV data has a large number of correlated features, as many genes share similar gains/loss patterns with other genes as they are on a similar region of the chromosome. Correlated features leads to multicollinearity in regression, which can lead to unstable coefficients, reducing accurate interpretation of the coefficients. When individual CNVs are used as input features in a GNN, a highly correlated feature set also reduces the unique contribution of individual genes, making interpretation difficult. Future work could only model a very targeted list of genes based on extensive literature analysis to improve interpretability.

The potential to integrate spatial datasets at scale would be important to truly understand the patient-patient heterogeneity. Additionally, combining the data with clinical data would be an important next step. Several datasets are currently being collected that could help us link molecular and spatial insights directly to patient outcomes. For example a recent industry-academic partnership is using spatial transcriptomics to profile 7,000 tumour samples with matched clinical data[344]. This would then enable us to link the EMT-TME interactions with outcome and treatment insights, with enhanced understanding of the patient-patient heterogeneity.

From each patient, we typically have one, or at most two slides. However, given that tumour heterogeneity is a key defining property of tumours, capturing a larger spatial area of the tumour through both across the tumour and at multiple depths would be important. Such an approach could serve as a spatial normalisation procedure. For

example, for each patient being studied, a sub-selection of patients can be profiled at multiple sections across the tumour. The degree of heterogeneity within a patient observed could then direct the study design into whether more slides are needed per patient. 3D spatial transcriptomics, which profiles spatial transcriptomics at multiple depths would enable us to address how accurate a single 2D slice is at representing the spatial relationships. Morphological 3D characterisation of tissue has been well established through fluorescence microscopy[166]. However, the molecular characterisation in 3D of tissue is still in its infancy. A multi-plex, single-cell approach was recently undertaken to understand 3D spatial context of lung cancer[167]. Compared with 2D approaches, analysing 3D tissue revealed previously unidentified dendritic niches and identified the 3D extent of T-cell niches. This suggests that a 3D approach could similarly identify novel spatial relationships between cells undergoing EMT.

5.2.3 Spatial statistical modelling challenges

From a spatial modelling perspective, I envisage that future approaches should incorporate causal inference techniques to help us to infer the direction of the EMT-TME relationships identified. For instance, matching methods, which include propensity score matching, estimate causal relationships by forming comparable treatment and control datasets within the original dataset[336]. By creating treated and control groups that are balanced on the relevant observed covariates, matching methods can help approximate the conditions of a randomised experiment, even when the data are observational rather than experimental. This can then help to understand whether the treatment (for instance, the presence of certain immune cells) may be causing changes in EMT, rather than reflecting coincidental correlations. Structural Equation Models (SEM), and in particular the geo-additive SEM approach, the spatial variant, would allow us to estimate both direct and indirect causal pathways among multiple biological variables, while accounting for spatial confounding in both predictors and outcomes[345]. This could then help determine, for example, whether immune cells initiate the EMT process or accumulate in areas where EMT is already underway. These approaches are currently in the early stages of being adopted in spatial transcriptomic workflows[346].

The existing methods in spatial transcriptomics currently capture gene expression and relationships at a single time point. However, methods that could simultaneously capture both the spatial and time dependence of RNA profiles would address a wider range of research questions. This would also help to add directionality to many relationships identified, increasing the likelihood of accurately identifying causal relationships. Additionally, it would allow us to understand how much information a single time point can capture, and how much information we lose by ignoring this concept. TEMPOmap is a recent approach for spatiotemporally resolved transcriptomics that could help address these questions[347].

In modelling tissue-level relationships, it is important to capture both short and long-range interactions. Cells within a tissue do not just influence their immediate neighbours, but their secreted factors and signals can travel over longer distances, affecting the behaviour of remote cells. GNNs, as used in Chapter 4, are typically limited by the fact that they rely on message-passing schemes, which operate primarily in localised neighbourhoods and require many layers to capture global interactions. However, with too many layers, GNNs can suffer from the over-smoothing problem[332], which is when after multiple layers of message passing, the node representations in a graph become very similar. Whilst in many other neural network based approaches adding more layers increases the accuracy of the task at hand, in a GNN this can reduce the model's ability to distinguish among different nodes and therefore lowers accuracy in tasks such as node classification. Transformer-based models, which use attention-based methods, are well suited to this challenge because they use attention mechanisms that allow them to consider relationships across the full input space[348]. Therefore, they can capture longer range cellular signalling and model more complex cellular processes. Nicheformer is a recent technique that demonstrates how transformers can learn biologically relevant latent spaces from spatial transcriptomic data[349]. Alternative approaches to capture short and long range spatial relationships include multi-mesh approaches, which contains nodes with different spatial resolutions. This was recently shown to be highly successful at weather prediction[350].

5.2.4 Future methodological considerations

In spatial transcriptomics analysis, as with many other types of biological data analysis, there is an inherent trade-off between methodological rigor and efficiency.

The process of selecting a single approach and set of parameters often stems from the need for clarity and ultimately, the ability to draw conclusions that progresses the field. However, this comes at the cost of overlooking the vast landscape of possible methodological choices, each of which could lead to slightly different, yet equally valid, interpretations of the data. Parameter selection must happen at many different steps of analysis, from the type of spatial transcriptomic platform to use, the number and types of cells types to use for deconvolution, the method of choice for dimensionality reduction approaches to whether to analyse spatial neighbours or domains.

A large focus of the thesis has been to ensure there is robust analysis of parameters. For example, I have highlighted the importance of considering the modifiable areal unit problem which exemplifies how shifting parameters can alter spatial interpretations, making it clear that no single choice is inherently correct, and I have investigated different spatial platforms to analyse the spatial relationships. Future research should aim to standardise best practices for balancing rigor with efficiency, perhaps through presenting findings in more interactive formats, allowing readers to explore how parameter variations influence downstream results and biological interpretations. By focusing on transparency, we can move towards research that acknowledges methodological uncertainty while still enabling meaningful scientific progress.

Also, extending from these more obvious choices, there are biological conceptual choices. Many biological processes exist on a continuum, and there are researchers that argue even the choice of communicating about genes and cell types using discrete concepts can overlook important biological complexities. For example, there is research that suggests analysing and communicating about cell types instead through a hierarchical tree approach could help more accurately represent these biological units[351]. There is also research that suggests that gene function is limited by single ontologies and that genes should instead be treated as distributions over cellular contexts[352]. However, often the more straight-forward concept of a gene and cell type is important for analysing and communicating about biological systems. In a similar manner, identifying spatial domains are useful to help us make sense of and communicate about complex tissue organisation, even though they may overlook important aspects of the tissue such as long-range cellular interactions. As artificial intelligence systems become more advanced, these predefined classifications may

become less essential, allowing for more abstract representations of biological structures and processes.

Multi-scale tissue analysis helps us see how seemingly simple cell-to-cell interactions can add up to complex, unpredictable behaviours, which can also be referred to as emergent properties[353,354]. In cancer, such emergent properties can occur when tumour cells and various elements of the TME continuously signal and adapt to one another. These interactions can give rise to unexpected outcomes like immune evasion, metastasis, and therapy resistance. Spatial transcriptomics offers a powerful way to observe these interactions in situ, making it easier to appreciate how emergent phenomena unfold. Recognising these properties is important for determining where and how to target treatments most effectively. By focusing on tumour and TME hotspots and their interactions, I begun to see how multiple cell types and signals come together to drive larger-scale changes in tissue. Going forward, finding new ways to measure and quantify emergence, using mathematical, computational, or other approaches applied to spatial transcriptomics will be essential for understanding cancer as a dynamic ecosystem.

5.3 Concluding remarks

This thesis has explored how the TME and intrinsic genomic factors interplay with epithelial-mesenchymal plasticity, using spatial transcriptomic data as a key tool to investigate these relationships. I have identified stable EMT niches that are enriched in hypoxic and angiogenic regions and are closely associated with key microenvironmental players such as CAFs and macrophages and I have explored these relationships across multiple spatial scales. By drawing on geostatistical methods and graph neural networks I developed a new method to quantify the relative contributions of intrinsic genomic changes and extrinsic microenvironmental signals on cell plasticity programmes. The approach highlights the importance of the TME in inducing both subtle, short-term changes and stable, long-term phenotypic change, whereas genomic alterations primarily contribute to more stable, long-term changes. Overall, these findings highlight the dominant influence of the TME in shaping EMT.

6 Bibliography

1. Yang, J. *et al*. Guidelines and definitions for research on epithelial–mesenchymal transition. *Nat Rev Mol Cell Biol* **21**, 341–352 (2020).
2. Ugurlu, B. & Karaoz, E. Comparison of similar cells: Mesenchymal stromal cells and fibroblasts. *Acta Histochem* **122**, 151634 (2020).
3. Molecular mechanisms of epithelial–mesenchymal transition - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4240281/.
4. Kim, D. H. *et al*. Epithelial Mesenchymal Transition in Embryonic Development, Tissue Repair and Cancer: A Comprehensive Overview. *J Clin Med* **7**, 1 (2017).
5. Pei, D., Shu, X., Gassama-Diagne, A. & Thiery, J. P. Mesenchymal–epithelial transition in development and reprogramming. *Nat Cell Biol* **21**, 44–53 (2019).
6. Thiery, J. P. Epithelial–mesenchymal transitions in development and pathologies. *Current Opinion in Cell Biology* **15**, 740–746 (2003).
7. Oft, M., Heider, K. H. & Beug, H. TGFbeta signaling is necessary for carcinoma cell invasiveness and metastasis. *Curr Biol* **8**, 1243–1252 (1998).
8. Savagner, P., Yamada, K. M. & Thiery, J. P. The Zinc-Finger Protein Slug Causes Desmosome Dissociation, an Initial and Necessary Step for Growth Factor–induced Epithelial–Mesenchymal Transition. *Journal of Cell Biology* **137**, 1403–1419 (1997).
9. Fujimoto, K., Sheng, H., Shao, J. & Beauchamp, R. D. Transforming Growth Factor-β1 Promotes Invasiveness after Cellular Transformation with Activated Ras in Intestinal Epithelial Cells. *Experimental Cell Research* **266**, 239–249 (2001).
10. Batlle, E. *et al*. The transcription factor Snail is a repressor of E-cadherin gene expression in epithelial tumour cells. *Nat Cell Biol* **2**, 84–89 (2000).
11. Yang, J. *et al*. Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* **117**, 927–939 (2004).

12. Eger, A. *et al*. DeltaEF1 is a transcriptional repressor of E-cadherin and regulates epithelial plasticity in breast cancer cells. *Oncogene* **24**, 2375–2385 (2005).

13. Thiery, J. P. & Sleeman, J. P. Complex networks orchestrate epithelial–mesenchymal transitions. *Nat Rev Mol Cell Biol* **7**, 131–142 (2006).

14. Guy, C. T., Cardiff, R. D. & Muller, W. J. Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease. *Mol Cell Biol* **12**, 954–961 (1992).

15. Saxena, M., Kalathur, R. K. R., Neutzner, M. & Christofori, G. PyMT-1099, a versatile murine cell model for EMT in breast cancer. *Sci Rep* **8**, 12123 (2018).

16. Trimboli, A. J. *et al.* Direct evidence for epithelial-mesenchymal transitions in breast cancer. *Cancer Res* **68**, 937–945 (2008).

17. Quinn, J. J. *et al*. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **371**, eabc1944 (2021).

18. Aiello, N. M. *et al*. EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. *Dev Cell* **45**, 681-695.e4 (2018).

19. Li, Y. *et al*. Genetic Fate Mapping of Transient Cell Fate Reveals N-Cadherin Activity and Function in Tumor Metastasis. *Developmental Cell* **54**, 593-607.e5 (2020).

20. Pastushenko, I. *et al*. Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463–468 (2018).

21. Zheng, X. *et al.* Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature* **527**, 525–530 (2015).

22. Santamaria, P. G., Moreno-Bueno, G., Portillo, F. & Cano, A. EMT: Present and future in clinical oncology. *Molecular Oncology* **11**, 718–738 (2017).

23. Bakir, B., Chiarella, A. M., Pitarresi, J. R. & Rustgi, A. K. EMT, MET, Plasticity, and Tumor Metastasis. *Trends in Cell Biology* **30**, 764–776 (2020).

24. Debnath, P., Huirem, R. S., Dutta, P. & Palchaudhuri, S. Epithelial–mesenchymal transition and its transcription factors. *Biosci Rep* **42**, BSR20211754 (2021).

25. Aiello, N. M. & Kang, Y. Context-dependent EMT programs in cancer metastasis. *J Exp Med* **216**, 1016–1026 (2019).

26. Haerinck, J., Goossens, S. & Berx, G. The epithelial–mesenchymal plasticity landscape: principles of design and mechanisms of regulation. *Nat Rev Genet* **24**, 590–609 (2023).

27. Asiedu, M. K., Ingle, J. N., Behrens, M. D., Radisky, D. C. & Knutson, K. L. TGFβ/TNFα-Mediated Epithelial-Mesenchymal Transition Generates Breast Cancer Stem Cells with a Claudin-Low Phenotype. *Cancer Res* **71**, 4707–4719 (2011).

28. Jiang, Z.-S., Sun, Y.-Z., Wang, S.-M. & Ruan, J.-S. Epithelial-mesenchymal transition: potential regulator of ABC transporters in tumor progression. *J Cancer* **8**, 2319–2327 (2017).

29. Wani, T. U. *et al*. Mechanistic insights into epithelial-mesenchymal transition mediated cisplatin resistance in ovarian cancer. *Sci Rep* **15**, 3053 (2025).

30. Du, B. & Shim, J. S. Targeting Epithelial–Mesenchymal Transition (EMT) to Overcome Drug Resistance in Cancer. *Molecules* **21**, 965 (2016).

31. Lüönd, F. *et al*. Distinct contributions of partial and full EMT to breast cancer malignancy. *Developmental Cell* **56**, 3203-3221.e11 (2021).

32. Wang, C. & He, Z. Integrating bulk and single-cell RNA sequencing data reveals epithelial-mesenchymal transition molecular subtype and signature to predict prognosis, immunotherapy efficacy, and drug candidates in low-grade gliomas. *Front Pharmacol* **14**, 1276466 (2023).

33. Kaur, B., Mukhlis, Y., Natesh, J., Penta, D. & Musthapa Meeran, S. Identification of hub genes associated with EMT-induced chemoresistance in breast cancer using integrated bioinformatics analysis. *Gene* **809**, 146016 (2022).

34. Wei, L. *et al.* Overexpression of 14-3-3ζ primes disease recurrence, metastasis and resistance to chemotherapy by inducing epithelial-mesenchymal transition in NSCLC. *Aging* **14**, 5838–5854 (2022).

35. Haslehurst, A. M. *et al.* EMT transcription factors snail and slug directly contribute to cisplatin resistance in ovarian cancer. *BMC Cancer* **12**, 91 (2012).

36. Sinha, D., Saha, P., Samanta, A. & Bishayee, A. Emerging Concepts of Hybrid Epithelial-to-Mesenchymal Transition in Cancer Progression. *Biomolecules* **10**, 1561 (2020).

37. Shibue, T. & Weinberg, R. A. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nat Rev Clin Oncol* **14**, 611–629 (2017).

38. Mani, S. A. *et al.* The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell* **133**, 704–715 (2008).

39. Cook, D. P. & Vanderhyden, B. C. Context specificity of the EMT transcriptional response. *Nat Commun* **11**, 2142 (2020).

40. Goetz, H., Melendez-Alvarez, J. R., Chen, L. & Tian, X.-J. A plausible accelerating function of intermediate states in cancer metastasis. *PLOS Computational Biology* **16**, e1007682 (2020).

41. Jain, P. *et al.* Cell-state transitions and density-dependent interactions together explain the dynamics of spontaneous epithelial-mesenchymal heterogeneity. *iScience* **27**, (2024).

42. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).

43. Peixoto, P. *et al.* EMT is associated with an epigenetic signature of ECM remodeling genes. *Cell Death Dis* **10**, 1–17 (2019).

44. Qin, X. & Tape, C. J. Functional analysis of cell plasticity using single-cell technologies. *Trends in Cell Biology* **34**, 854–864 (2024).

45. Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**, 871–890 (2009).

46. Satelli, A. & Li, S. Vimentin in cancer and its potential as a molecular target for cancer therapy. *Cell Mol Life Sci* **68**, 3033–3046 (2011).

47. Subramanian, A. *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

48. Malagoli Tagliazucchi, G., Wiecek, A. J., Withnell, E. & Secrier, M. Genomic and microenvironmental heterogeneity shaping epithelial-to-mesenchymal trajectories in cancer. *Nat Commun* **14**, 1–20 (2023).

49. Stein-O'Brien, G. L. *et al*. Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Systems* **8**, 395-411.e8 (2019).

50. Yu, B. *et al*. scGMAI: a Gaussian mixture model for clustering single-cell RNA-Seq data based on deep autoencoder. *Brief Bioinform* **22**, bbaa316 (2021).

51. Van den Berge, K. *et al*. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat Commun* **11**, 1201 (2020).

52. Street, K. *et al*. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

53. Grobecker, P., Sakoparnig, T. & Nimwegen, E. van. Identifying cell states in single-cell RNA-seq data at statistically maximal resolution. *PLOS Computational Biology* **20**, e1012224 (2024).

54. Persad, S. *et al*. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat Biotechnol* **41**, 1746–1757 (2023).

55. Szabo, P. M. *et al*. Cancer-associated fibroblasts are the main contributors to epithelial-to-mesenchymal signatures in the tumor microenvironment. *Sci Rep* **13**, 3051 (2023).

56. Skrypek, N., Goossens, S., Smedt, E. D., Vandamme, N. & Berx, G. Epithelial-to-Mesenchymal Transition: Epigenetic Reprogramming Driving Cellular Plasticity. *Trends in Genetics* **33**, 943–959 (2017).

57. Malouf, G. G. *et al*. Architecture of epigenetic reprogramming following Twist1-mediated epithelial-mesenchymal transition. *Genome Biology* **14**, R144 (2013).

58. Kim, B. N. *et al*. TGF-β induced EMT and stemness characteristics are associated with epigenetic regulation in lung cancer. *Sci Rep* **10**, 10597 (2020).

59. Serrano-Gomez, S. J., Maziveyi, M. & Alahari, S. K. Regulation of epithelial-mesenchymal transition through epigenetic and post-translational modifications. *Molecular Cancer* **15**, 18 (2016).

60. Hussein, M. A., Valinezhad, K., Adel, E. & Munirathinam, G. MALAT-1 Is a Key Regulator of Epithelial–Mesenchymal Transition in Cancer: A Potential Therapeutic Target for Metastasis. *Cancers (Basel)* **16**, 234 (2024).

61. HOTAIR lncRNA promotes epithelial-mesenchymal transition by redistributing LSD1 at regulatory chromatin regions - PubMed. https://pubmed.ncbi.nlm.nih.gov/33960111/.

62. Micalizzi, D. S., Farabaugh, S. M. & Ford, H. L. Epithelial-Mesenchymal Transition in Cancer: Parallels Between Normal Development and Tumor Progression. *J Mammary Gland Biol Neoplasia* **15**, 117–134 (2010).

63. Chou, J., Provot, S. & Werb, Z. GATA3 in development and cancer differentiation: Cells GATA have it! *Journal of Cellular Physiology* **222**, 42–49 (2010).

64. Sharma, S. *et al*. Loss of p53 epigenetically modulates epithelial to mesenchymal transition in colorectal cancer. *Translational Oncology* **43**, 101848 (2024).

65. Dai, Y. *et al*. Copy number gain of ZEB1 mediates a double-negative feedback loop with miR-33a-5p that regulates EMT and bone metastasis of prostate cancer dependent on TGF-β signaling. *Theranostics* **9**, 6063–6079 (2019).

66. Cai, Y. *et al*. Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. *Cancer Cell* **29**, 751–766 (2016).

67. Turajlic, S. *et al*. Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell* **173**, 595-610.e11 (2018).

68. Rothschild, B. L. *et al*. Cortactin Overexpression Regulates Actin-Related Protein 2/3 Complex Activity, Motility, and Invasion in Carcinomas with Chromosome 11q13 Amplification. *Cancer Res* **66**, 8017–8025 (2006).

69. Role of TWIST proteins in cancer progression. https://atlasgeneticsoncology.org/deep-insight/20081/role-of-twist-proteins-in-cancer-progression#section-5.

70. Chen, X., Agustinus, A. S., Li, J., DiBona, M. & Bakhoum, S. F. Chromosomal instability as a driver of cancer progression. *Nat Rev Genet* **26**, 31–46 (2025).

71. Pastushenko, I. *et al*. Fat1 deletion promotes hybrid EMT state, tumour stemness and metastasis. *Nature* **589**, 448–455 (2021).

72. Mair, T. *et al*. Aggressive KRAS mutations direct TGF-β response towards partial EMT in patient-derived colorectal cancer tumoroids. 2024.06.25.600620 Preprint at https://doi.org/10.1101/2024.06.25.600620 (2024).

73. D'Angelo, E. *et al*. Intrinsic and Extrinsic Modulators of the Epithelial to Mesenchymal Transition: Driving the Fate of Tumor Microenvironment. *Frontiers in Oncology* **10**, (2020).

74. Interplay between tumor microenvironment and partial EMT as the driver of tumor progression - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S258900422100081X.

75. Castillo, S. P. *et al*. The tumour ecology of quiescence: Niches across scales of complexity. *Seminars in Cancer Biology* **92**, 139–149 (2023).

76. Signaling networks guiding epithelial-mesenchymal transitions during embryogenesis and cancer progression - PubMed. https://pubmed.ncbi.nlm.nih.gov/17645776/.

77. Yoon, H. *et al*. TGF-β1-mediated transition of resident fibroblasts to cancer-associated fibroblasts promotes cancer metastasis in gastrointestinal stromal tumor. *Oncogenesis* **10**, 1–12 (2021).

78. Jia, C. *et al.* Cancer-associated Fibroblasts induce epithelial-mesenchymal transition via the Transglutaminase 2-dependent IL-6/IL6R/STAT3 axis in Hepatocellular Carcinoma. *Int J Biol Sci* **16**, 2542–2558 (2020).

79. Goulet, C. R. *et al.* Cancer-associated fibroblasts induce epithelial–mesenchymal transition of bladder cancer cells through paracrine IL-6 signalling. *BMC Cancer* **19**, 137 (2019).

80. Dehai, C. *et al.* Enhanced invasion of lung adenocarcinoma cells after co-culture with THP-1-derived macrophages via the induction of EMT by IL-6. *Immunol Lett* **160**, 1–10 (2014).

81. A co-culture system of macrophages with breast cancer tumoroids to study cell interactions and therapeutic responses: Cell Reports Methods. https://www.cell.com/cell-reports-methods/fulltext/S2667-2375(24)00148-6.

82. Fattet, L. *et al.* Matrix Rigidity Controls Epithelial-Mesenchymal Plasticity and Tumor Metastasis via a Mechanoresponsive EPHA2/LYN Complex. *Developmental Cell* **54**, 302-316.e7 (2020).

83. Wei, S. C. *et al.* Matrix stiffness drives epithelial–mesenchymal transition and tumour metastasis through a TWIST1–G3BP2 mechanotransduction pathway. *Nat Cell Biol* **17**, 678–688 (2015).

84. Wang, H. *et al.* MicroRNA-181d-5p-Containing Exosomes Derived from CAFs Promote EMT by Regulating CDX2/HOXA5 in Breast Cancer. *Molecular Therapy Nucleic Acids* **19**, 654–667 (2020).

85. Terry, S. *et al.* New insights into the role of EMT in tumor immune escape. *Mol Oncol* **11**, 824–846 (2017).

86. Yang, X.-G., Zhu, L.-C., Wang, Y.-J., Li, Y.-Y. & Wang, D. Current Advance of Therapeutic Agents in Clinical Trials Potentially Targeting Tumor Plasticity. *Front Oncol* **9**, 887 (2019).

87. Drooger, J. C., van der Padt, A., Sleijfer, S. & Jager, A. Denosumab in breast cancer treatment. *Eur J Pharmacol* **717**, 12–19 (2013).

88. Tolcher, A. *et al.* A First-in-Human Phase I Study of OPB-111077, a Small-Molecule STAT3 and Oxidative Phosphorylation Inhibitor, in Patients with Advanced Cancers. *Oncologist* **23**, 658-e72 (2018).

89. Krebs, A. M. *et al.* The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. *Nat Cell Biol* **19**, 518–529 (2017).

90. Rasco, D. W. *et al.* A First-in-Human Study of Novel Cereblon Modulator Avadomide (CC-122) in Advanced Malignancies. *Clin Cancer Res* **25**, 90–98 (2019).

91. Study Details | Phase I/Ib Study of NIS793 in Combination With PDR001 in Patients With Advanced Malignancies. | ClinicalTrials.gov. https://clinicaltrials.gov/study/NCT02947165.

92. Mirzaei, S. *et al.* Pre-Clinical and Clinical Applications of Small Interfering RNAs (siRNA) and Co-Delivery Systems for Pancreatic Cancer Therapy. *Cells* **10**, 3348 (2021).

93. Dong, B., Qiu, Z. & Wu, Y. Tackle Epithelial-Mesenchymal Transition With Epigenetic Drugs in Cancer. *Front Pharmacol* **11**, 596239 (2020).

94. Sterner, R. C. & Sterner, R. M. CAR-T cell therapy: current limitations and potential strategies. *Blood Cancer J.* **11**, 1–11 (2021).

95. Meng, L. *et al.* Mechanisms of immune checkpoint inhibitors: insights into the regulation of circular RNAS involved in cancer hallmarks. *Cell Death Dis* **15**, 1–26 (2024).

96. Liu, S., Ren, J. & ten Dijke, P. Targeting TGFβ signal transduction for cancer therapy. *Sig Transduct Target Ther* **6**, 1–20 (2021).

97. Sun, H., Wang, X., Wang, X., Xu, M. & Sheng, W. The role of cancer-associated fibroblasts in tumorigenesis of gastric cancer. *Cell Death Dis* **13**, 1–9 (2022).

98. Wang, S., He, Y., Wang, J. & Luo, E. Re-exploration of immunotherapy targeting EMT of hepatocellular carcinoma: Starting from the NF-κB pathway. *Biomedicine & Pharmacotherapy* **174**, 116566 (2024).

99. Jonckheere, S. *et al.* Epithelial-Mesenchymal Transition (EMT) as a Therapeutic Target. *CTO* **211**, 157–182 (2022).

100. Chakraborty, P., George, J. T., Tripathi, S., Levine, H. & Jolly, M. K. Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum. *Frontiers in Bioengineering and Biotechnology* **8**, (2020).

101. Prasetyanti, P. R. & Medema, J. P. Intra-tumor heterogeneity from a cancer stem cell perspective. *Mol Cancer* **16**, 41 (2017).

102. Sun, X. & Yu, Q. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacol Sin* **36**, 1219–1227 (2015).

103. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* **15**, 81–94 (2018).

104. Robert, C. A decade of immune-checkpoint inhibitors in cancer therapy. *Nat Commun* **11**, 3801 (2020).

105. Image analysis reveals molecularly distinct patterns of TILs in NSCLC associated with treatment outcome | npj Precision Oncology. https://www.nature.com/articles/s41698-022-00277-5.

106. Karimi, E. *et al.* Single-cell spatial immune landscapes of primary and metastatic brain tumours. *Nature* **614**, 555–563 (2023).

107. Sorin, M. *et al.* Single-cell spatial landscapes of the lung tumour immune microenvironment. *Nature* **614**, 548–554 (2023).

108. Exploring tissue architecture using spatial transcriptomics | Nature. https://www.nature.com/articles/s41586-021-03634-9.

109. Home Page - 10x Genomics. https://www.10xgenomics.com/.

110. Janesick, A. *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun* **14**, 8353 (2023).

111. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).

112. Saiselet, M. *et al*. Transcriptional output, cell-type densities, and normalization in spatial transcriptomics. *Journal of Molecular Cell Biology* **12**, 906–908 (2020).

113. Lopez, R. *et al*. DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat Biotechnol* **40**, 1360–1369 (2022).

114. Biancalani, T. *et al*. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods* **18**, 1352–1362 (2021).

115. Zhang, K., Feng, W. & Wang, P. Identification of spatially variable genes with graph cuts. *Nat Commun* **13**, 5488 (2022).

116. Palla, G. *et al*. Squidpy: a scalable framework for spatial omics analysis. *Nat Methods* **19**, 171–178 (2022).

117. Tumor Immune Microenvironment during Epithelial–Mesenchymal Transition | Clinical Cancer Research | American Association for Cancer Research. https://aacrjournals.org/clincancerres/article/27/17/4669/671656/Tumor-Immune-Microenvironment-during-Epithelial.

118. Barkley, D. *et al*. Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat Genet* **54**, 1192–1201 (2022).

119. Frontiers | Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum. https://www.frontiersin.org/articles/10.3389/fbioe.2020.00220/full#B37.

120. He, S. *et al*. Starfysh reveals heterogeneous spatial dynamics in the breast tumor microenvironment. 2022.11.21.517420 Preprint at https://doi.org/10.1101/2022.11.21.517420 (2022).

121. Cable, D. M. *et al*. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* **40**, 517–526 (2022).

122. Andersson, A. *et al*. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol* **3**, 1–8 (2020).

123. Kleshchevnikov, V. *et al*. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* **40**, 661–671 (2022).

124. Hu, J. *et al*. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* **18**, 1342–1351 (2021).

125. Zhao, E. *et al*. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* **39**, 1375–1384 (2021).

126. Dries, R. *et al*. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology* **22**, 78 (2021).

127. Varrone, M., Tavernari, D., Santamaria-Martínez, A., Walsh, L. A. & Ciriello, G. CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nat Genet* 1–11 (2023) doi:10.1038/s41588-023-01588-4.

128. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat Methods* **15**, 343–346 (2018).

129. Zhu, J., Sun, S. & Zhou, X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology* **22**, 184 (2021).

130. Palla, G., Fischer, D. S., Regev, A. & Theis, F. J. Spatial components of molecular tissue biology. *Nat Biotechnol* **40**, 308–318 (2022).

131. Song, Q. & Su, J. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in Bioinformatics* **22**, bbaa414 (2021).

132. Yuan, Y. & Bar-Joseph, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biology* **21**, 300 (2020).

133. Li, J., Chen, S., Pan, X., Yuan, Y. & Shen, H.-B. Cell clustering for spatial transcriptomics data with graph neural networks. *Nat Comput Sci* **2**, 399–408 (2022).

134. Gunduz, C., Yener, B. & Gultekin, S. H. The cell graphs of cancer. *Bioinformatics* **20**, i145–i151 (2004).

135. Fischer, D. S., Ali, M., Richter, S., Ertürk, A. & Theis, F. *Graph Neural Networks Learn Emergent Tissue Properties from Spatial Molecular Profiles*. http://biorxiv.org/lookup/doi/10.1101/2022.12.08.519537 (2022) doi:10.1101/2022.12.08.519537.

136. Khemani, B., Patil, S., Kotecha, K. & Tanwar, S. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data* **11**, 18 (2024).

137. Eliasof, M. & Treister, E. Global-Local Graph Neural Networks for Node-Classification. Preprint at https://doi.org/10.48550/arXiv.2406.10863 (2024).

138. Jin, D. *et al.* Universal Graph Convolutional Networks. in *Advances in Neural Information Processing Systems* vol. 34 10654–10664 (Curran Associates, Inc., 2021).

139. Veličković, P. *et al.* Graph Attention Networks. Preprint at https://doi.org/10.48550/arXiv.1710.10903 (2018).

140. Knyazev, B., Taylor, G. W. & Amer, M. Understanding Attention and Generalization in Graph Neural Networks. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).

141. E, Z., R, Q., A, C. & Sj, C. Mapping the transcriptome: Realizing the full potential of spatial data analysis. *Cell* **186**, (2023).

142. Duque, J. C., Laniado, H. & Polo, A. S-maup: Statistical test to measure the sensitivity to the modifiable areal unit problem. *PLOS ONE* **13**, e0207377 (2018).

143. Parenteau, M.-P. & Sawada, M. C. The modifiable areal unit problem (MAUP) in the relationship between exposure to NO2 and respiratory health. *International Journal of Health Geographics* **10**, 58 (2011).

144. O'Connor, C. *et al.* Assessing the impact of areal unit selection and the modifiable areal unit problem on associative statistics between cases of tick-borne disease and entomological indices. *J Med Entomol* **61**, 331–344 (2024).

145. Comber, A. *et al.* A Route Map for Successful Applications of Geographically Weighted Regression. *Geographical Analysis* **55**, 155–178 (2023).

146. Comber, A., Zormpas, E., Queen, R. & Cockell, S. J. Lessons from spatial transcriptomics and computational geography in mapping the transcriptome. *AGILE: GIScience Series* **5**, 1–6 (2024).

147. Harbeck, N. *et al.* Breast cancer. *Nat Rev Dis Primers* **5**, 1–31 (2019).

148. Arnold, M. *et al.* Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast* **66**, 15–23 (2022).

149. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71 (2016).

150. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

151. Liu, F., Gu, L.-N., Shan, B.-E., Geng, C.-Z. & Sang, M.-X. Biomarkers for EMT and MET in breast cancer: An update. *Oncol Lett* **12**, 4869–4876 (2016).

152. Kallergi, G. *et al.* Epithelial to mesenchymal transition markers expressed in circulating tumour cells of early and metastatic breast cancer patients. *Breast Cancer Research* **13**, R59 (2011).

153. XIE, G. *et al.* IL-6-induced epithelial-mesenchymal transition promotes the generation of breast cancer stem-like cells analogous to mammosphere cultures. *Int J Oncol* **40**, 1171–1179 (2011).

154. Zhao, Z. *et al*. In vivo visualization and characterization of epithelial-mesenchymal transition in breast tumors. *Cancer Res* **76**, 2094–2104 (2016).

155. Brown, M. S. & Pattabiraman, D. Phenotypic heterogeneity driven by plasticity of the intermediate EMT state governs disease progression and metastasis in breast cancer. Gene Expression Omnibus https://identifiers.org/geo:GSE172613 (2022).

156. Cabrita, R. *et al*. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* **577**, 561–565 (2020).

157. Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. & Lundeberg, J. Spatially Resolved Transcriptomics Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous Malignant Melanoma. *Cancer Res* **78**, 5970–5979 (2018).

158. Andersson, A. *et al*. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat Commun* **12**, 6012 (2021).

159. Berglund, E. *et al*. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun* **9**, 2419 (2018).

160. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).

161. Bergenstråhle, J., Larsson, L. & Lundeberg, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* **21**, 482 (2020).

162. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).

163. Proceedings of the Python in Science Conference (SciPy): Exploring Network Structure, Dynamics, and Function using NetworkX. https://conference.scipy.org/proceedings/SciPy2008/paper_2/.

164. Thelen, M. *et al*. Cancer-specific immune evasion and substantial heterogeneity within cancer types provide evidence for personalized immunotherapy. *NPJ Precis Oncol* **5**, 52 (2021).

165. San Juan, B. P., Garcia-Leon, M. J., Rangel, L., Goetz, J. G. & Chaffer, C. L. The Complexities of Metastasis. *Cancers (Basel)* **11**, 1575 (2019).

166. Brown, M. S. *et al*. Phenotypic heterogeneity driven by plasticity of the intermediate EMT state governs disease progression and metastasis in breast cancer. *Science Advances* **8**, eabj8002 (2022).

167. Withnell, E. & Secrier, M. SpottedPy quantifies relationships between spatial transcriptomic hotspots and uncovers environmental cues of epithelial-mesenchymal plasticity in breast cancer. *Genome Biology* **25**, 289 (2024).

168. Pan, S., Withnell, E. & Secrier, M. Classifying epithelial-mesenchymal transition states in single cell cancer data using large language models. 2024.08.16.608311 Preprint at https://doi.org/10.1101/2024.08.16.608311 (2024).

169. Celik, C. *et al.* Balancing tumour proliferation and sustained cell cycle arrest through proteostasis remodelling drives immune niche compartmentalisation in breast cancer. 2025.01.08.632014 Preprint at https://doi.org/10.1101/2025.01.08.632014 (2025).

170. Kumar, T. *et al*. A spatially resolved single-cell genomic atlas of the adult human breast. *Nature* **620**, 181–191 (2023).

171. Sibai, M. *et al*. The spatial landscape of Cancer Hallmarks reveals patterns of tumor ecology. *bioRxiv* 2022.06.18.496114 (2023) doi:10.1101/2022.06.18.496114.

172. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell* **182**, 497-514.e22 (2020).

173. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**, 1334–1347 (2021).

174. Lengrand, J. *et al.* Pharmacological targeting of netrin-1 inhibits EMT in cancer. *Nature* **620**, 402–408 (2023).

175. Walker, B. L. & Nie, Q. NeST: nested hierarchical structure identification in spatial transcriptomic data. *Nat Commun* **14**, 6554 (2023).

176. Chitra, U. *et al.* Mapping the topography of spatial gene expression with interpretable deep learning. *bioRxiv* 2023.10.10.561757 (2023) doi:10.1101/2023.10.10.561757.

177. Seferbekova, Z., Lomakin, A., Yates, L. R. & Gerstung, M. Spatial biology of cancer evolution. *Nat Rev Genet* **24**, 295–313 (2023).

178. Moses, L. *et al.* Voyager: exploratory single-cell genomics data analysis with geospatial statistics. 2023.07.20.549945 Preprint at https://doi.org/10.1101/2023.07.20.549945 (2023).

179. Bernstein, M. N. *et al.* Monkeybread: A Python toolkit for the analysis of cellular niches in single-cell resolution spatial transcriptomics data. 2023.09.14.557736 Preprint at https://doi.org/10.1101/2023.09.14.557736 (2023).

180. Fotheringham, A. S. & Wong, D. W. S. The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environ Plan A* **23**, 1025–1044 (1991).

181. Li, Z. & Zhou, X. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biol* **23**, 168 (2022).

182. Yuan, Y. Spatial Heterogeneity in the Tumor Microenvironment. *Cold Spring Harb Perspect Med* **6**, a026583 (2016).

183. Wilson, C. *et al.* Tumor immune cell clustering and its association with survival in African American women with ovarian cancer. *PLoS Comput Biol* **18**, e1009900 (2022).

184. Geostatistical visualization of ecological interactions in tumors - PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC7198084/.

185. Nawaz, S., Heindl, A., Koelble, K. & Yuan, Y. Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. *Mod Pathol* **28**, 766–777 (2015).

186. Ganier, C. *et al.* Multiscale spatial mapping of cell populations across anatomical sites in healthy human skin and basal cell carcinoma. *Proceedings of the National Academy of Sciences* **121**, e2313326120 (2024).

187. Ma, Y. & Zhou, X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat Biotechnol* **40**, 1349–1359 (2022).

188. Valdeolivas, A. *et al.* Profiling the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. *npj Precis. Onc.* **8**, 1–16 (2024).

189. Sahoo, S. *et al.* Multi-modal transcriptomic analysis unravels enrichment of hybrid epithelial/mesenchymal state and enhanced phenotypic heterogeneity in basal breast cancer. *bioRxiv* 2023.09.30.558960 (2023) doi:10.1101/2023.09.30.558960.

190. Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol Med* **6**, 1279–1293 (2014).

191. Elyanow, R., Zeira, R., Land, M. & Raphael, B. J. STARCH: copy number and clone inference from spatial transcriptomics data. *Phys Biol* **18**, 035001 (2021).

192. Sherman, T. D., Gao, T. & Fertig, E. J. CoGAPS 3: Bayesian non-negative matrix factorization for single-cell analysis with asynchronous updates and sparse data structures. *BMC Bioinformatics* **21**, 453 (2020).

193. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).

194. Nielsen, T. O. *et al.* A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor–Positive Breast Cancer. *Clinical Cancer Research* **16**, 5222–5232 (2010).

195. Cui, C. *et al.* Ratio of the interferon-γ signature to the immunosuppression signature predicts anti-PD-1 therapy response in melanoma. *NPJ Genom Med* **6**, 7 (2021).

196. Johnson, A. M. *et al.* Cancer Cell-Intrinsic Expression of MHC Class II Regulates the Immune Microenvironment and Response to Anti-PD-1 Therapy in Lung Adenocarcinoma. *J Immunol* **204**, 2295–2307 (2020).

197. Liu, J. *et al.* A novel immune checkpoint-related gene signature for predicting overall survival and immune status in triple-negative breast cancer. *Transl Cancer Res* **11**, 181–192 (2022).

198. Wherry, E. J. & Kurachi, M. Molecular and cellular insights into T cell exhaustion. *Nature reviews. Immunology* **15**, 486 (2015).

199. Rey, S. J. & Anselin, L. PySAL: A Python library of spatial analytical methods. *The Review of regional studies* **37**, 5–27 (2007).

200. Pirrotta, S. *et al.* signifinder enables the identification of tumor cell states and cancer expression signatures in bulk, single-cell and spatial transcriptomic data. *bioRxiv* 2023.03.07.530940 (2023) doi:10.1101/2023.03.07.530940.

201. Xun, Z. *et al.* Reconstruction of the tumor spatial microenvironment along the malignant-boundary-nonmalignant axis. *Nat Commun* **14**, 1–16 (2023).

202. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat Rev Genet* **22**, 627–644 (2021).

203. Muz, B., de la Puente, P., Azab, F. & Azab, A. K. The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia (Auckl)* **3**, 83–92 (2015).

204. Weidemann, A. & Johnson, R. S. Biology of HIF-1α. *Cell Death Differ* **15**, 621–627 (2008).

205. Tam, S. Y., Wu, V. W. C. & Law, H. K. W. Hypoxia-Induced Epithelial-Mesenchymal Transition in Cancers: HIF-1α and Beyond. *Front Oncol* **10**, 486 (2020).

206. Nishida, N., Yano, H., Nishida, T., Kamura, T. & Kojiro, M. Angiogenesis in Cancer. *Vasc Health Risk Manag* **2**, 213–219 (2006).

207. Almagro, J., Messal, H. A., Elosegui-Artola, A., Rheenen, J. van & Behrens, A. Tissue architecture in tumor initiation and progression. *Trends in Cancer* **8**, 494–505 (2022).

208. Saxena, K., Jolly, M. K. & Balamurugan, K. Hypoxia, partial EMT and collective migration: Emerging culprits in metastasis. *Translational Oncology* **13**, 100845 (2020).

209. Biffi, G. *et al*. IL-1-induced JAK/STAT signaling is antagonized by TGF-β to shape CAF heterogeneity in pancreatic ductal adenocarcinoma. *Cancer discovery* **9**, 282 (2019).

210. Ghahremanifard, P., Chanda, A., Bonni, S. & Bose, P. TGF-β Mediated Immune Evasion in Cancer—Spotlight on Cancer-Associated Fibroblasts. *Cancers* **12**, 3650 (2020).

211. Principe, D. R., Timbers, K. E., Atia, L. G., Koch, R. M. & Rana, A. TGFβ Signaling in the Pancreatic Tumor Microenvironment. *Cancers (Basel)* **13**, 5086 (2021).

212. Xu, J., Lamouille, S. & Derynck, R. TGF-β-induced epithelial to mesenchymal transition. *Cell Res* **19**, 156–172 (2009).

213. myCAFs are better than yours: targeting myofibroblasts potentiates immunotherapy. *Trends in Cancer* **9**, 1–2 (2023).

214. Amer, H. T., Stein, U. & El Tayebi, H. M. The Monocyte, a Maestro in the Tumor Microenvironment (TME) of Breast Cancer. *Cancers* **14**, 5460 (2022).

215. Jf, G. *et al*. LAG-3 regulates CD8+ T cell accumulation and effector function in murine self- and tumor-tolerance systems. *The Journal of clinical investigation* **117**, (2007).

216. Grosso, J. F. *et al*. LAG-3 regulates $CD8^+$ T cell accumulation and effector function in murine self- and tumor-tolerance systems. *J Clin Invest* **117**, 3383–3392 (2007).

217. Turley, S. J., Cremasco, V. & Astarita, J. L. Immunological hallmarks of stromal cells in the tumour microenvironment. *Nat Rev Immunol* **15**, 669–682 (2015).

218. Yang, X. *et al*. FAP Promotes Immunosuppression by Cancer-Associated Fibroblasts in the Tumor Microenvironment via STAT3–CCL2 Signaling. *Cancer Res* **76**, 4124–4135 (2016).

219. Zhao, K., Yi, Y., Ma, Z. & Zhang, W. INHBA is a Prognostic Biomarker and Correlated With Immune Cell Infiltration in Cervical Cancer. *Front Genet* **12**, 705512 (2022).

220. Johnson, K. A., Emmerich, P., Matkowskyj, K. A. & Deming, D. A. Predicting CD8+ T-cell infiltration in colorectal cancer using versican proteolysis across molecular profiles. *JCO* **38**, 189–189 (2020).

221. Chen, P., Cescon, M. & Bonaldo, P. Collagen VI in cancer and its biological mechanisms. *Trends Mol Med* **19**, 410–417 (2013).

222. Hazini, A., Fisher, K. & Seymour, L. Deregulation of HLA-I in cancer and its central importance for immunotherapy. *J Immunother Cancer* **9**, e002899 (2021).

223. Wuerfel, F. M. *et al*. HLA-G and HLA-F protein isoform expression in breast cancer patients receiving neoadjuvant treatment. *Sci Rep* **10**, 15750 (2020).

224. Ayers, M. *et al*. IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. *J Clin Invest* **127**, 2930–2940 (2017).

225. Zormpas, E., Queen, R., Comber, A. & Cockell, S. J. Mapping the transcriptome: Realizing the full potential of spatial data analysis. *Cell* **186**, 5677–5689 (2023).

226. Spatial and single-nucleus transcriptomic analysis of genetic and sporadic forms of Alzheimer's Disease | bioRxiv. https://www.biorxiv.org/content/10.1101/2023.07.24.550282v1.

227. Ospina, O. E. *et al.* spatialGE: quantification and visualization of the tumor microenvironment heterogeneity using spatial transcriptomics. *Bioinformatics* **38**, 2645–2647 (2022).

228. Emami Nejad, A. *et al*. The role of hypoxia in the tumor microenvironment and development of cancer stem cell: a novel approach to developing treatment. *Cancer Cell Int* **21**, 62 (2021).

229. Tyler, M. & Tirosh, I. Decoupling epithelial-mesenchymal transitions from stromal profiles by integrative expression analysis. *Nat Commun* **12**, 2592 (2021).

230. Li, X., Chen, L., Peng, X. & Zhan, X. Progress of tumor-associated macrophages in the epithelial-mesenchymal transition of tumor. *Front Oncol* **12**, 911410 (2022).

231. Brabletz, S., Schuhwerk, H., Brabletz, T. & Stemmler, M. P. Dynamic EMT: a multi-tool for tumor progression. *The EMBO Journal* **40**, e108647 (2021).

232. Pastushenko, I. & Blanpain, C. EMT Transition States during Tumor Progression and Metastasis. *Trends in Cell Biology* **29**, 212–226 (2019).

233. Kmieciak, M., Knutson, K. L., Dumur, C. I. & Manjili, M. H. HER-2/neu antigen loss and relapse of mammary carcinoma are actively induced by T cell-mediated anti-tumor immune responses. *Eur J Immunol* **37**, 675–685 (2007).

234. Santisteban, M. *et al.* Immune-induced epithelial to mesenchymal transition in vivo generates breast cancer stem cells. *Cancer Res* **69**, 2887–2895 (2009).

235. Galsky, M. D. *et al.* 850PD - Epithelial-mesenchymal transition (EMT), T cell infiltration, and outcomes with nivolumab (nivo) in urothelial cancer (UC). *Annals of Oncology* **28**, v297 (2017).

236. Chockley, P. J. *et al.* Epithelial-mesenchymal transition leads to NK cell–mediated metastasis-specific immunosurveillance in lung cancer. *J Clin Invest* **128**, 1384–1396.

237. Datar, I. & Schalper, K. A. Epithelial–Mesenchymal Transition and Immune Evasion during Lung Cancer Progression: The Chicken or the Egg? *Clin Cancer Res* **22**, 3422–3424 (2016).

238. Subhadarshini, S., Markus, J., Sahoo, S. & Jolly, M. K. Dynamics of Epithelial–Mesenchymal Plasticity: What Have Single-Cell Investigations Elucidated So Far? *ACS Omega* **8**, 11665–11673 (2023).

239. Zhang, Y. *et al.* Hypoxia in Breast Cancer-Scientific Translation to Therapeutic and Diagnostic Clinical Applications. *Front Oncol* **11**, 652266 (2021).

240. Fu, Z., Mowday, A. M., Smaill, J. B., Hermans, I. F. & Patterson, A. V. Tumour Hypoxia-Mediated Immunosuppression: Mechanisms and Therapeutic Approaches to Improve Cancer Immunotherapy. *Cells* **10**, 1006 (2021).

241. Lee, C.-T., Mace, T. & Repasky, E. A. Hypoxia-Driven Immunosuppression: A new reason to use thermal therapy in the treatment of cancer? *Int J Hyperthermia* **26**, 232–246 (2010).

242. Dekoninck, S. & Blanpain, C. Stem cell dynamics, migration and plasticity during wound healing. *Nat Cell Biol* **21**, 18–24 (2019).

243. Pérez-González, A., Bévant, K. & Blanpain, C. Cancer cell plasticity during tumor progression, metastasis and response to therapy. *Nat Cancer* **4**, 1063–1082 (2023).

244. Palm, W. Metabolic plasticity allows cancer cells to thrive under nutrient starvation. *Proc Natl Acad Sci U S A* **118**, e2102057118 (2021).

245. Mills, J. C., Stanger, B. Z. & Sander, M. Nomenclature for cellular plasticity: are the terms as plastic as the cells themselves? *EMBO J* **38**, e103148 (2019).

246. Burdziak, C. *et al.* Epigenetic plasticity cooperates with cell-cell interactions to direct pancreatic tumorigenesis. *Science* **380**, eadd5327 (2023).

247. Househam, J. *et al.* Phenotypic plasticity and genetic control in colorectal cancer evolution. *Nature* **611**, 744–753 (2022).

248. Whiting, F. J. H., Househam, J., Baker, A.-M., Sottoriva, A. & Graham, T. A. Phenotypic noise and plasticity in cancer evolution. *Trends in Cell Biology* **34**, 451–464 (2024).

249. Dar, R. D., Hosmane, N. N., Arkin, M. R., Siliciano, R. F. & Weinberger, L. S. Screening for noise in gene expression identifies drug synergies. *Science* **344**, 1392–1396 (2014).

250. Wang, Q., Huang, L., Wen, K. & Yu, J. The mean and noise of stochastic gene transcription with cell division. *MBE* **15**, 1255–1270 (2018).

251. Soltani, M., Vargas-Garcia, C. A., Antunes, D. & Singh, A. Intercellular Variability in Protein Levels from Stochastic Expression and Noisy Cell Cycle Processes. *PLOS Computational Biology* **12**, e1004972 (2016).

252. Yang, J. *et al.* Epigenetic regulation in the tumor microenvironment: molecular mechanisms and therapeutic targets. *Sig Transduct Target Ther* **8**, 1–26 (2023).

253. Handy, D. E., Castro, R. & Loscalzo, J. Epigenetic Modifications: Basic Mechanisms and Role in Cardiovascular Disease. *Circulation* **123**, 2145–2156 (2011).

254. Yang, D. *et al.* Lineage Tracing Reveals the Phylodynamics, Plasticity and Paths of Tumor Evolution. *Cell* **185**, 1905-1923.e25 (2022).

255. Schiffman, J. S. *et al.* Defining heritability, plasticity, and transition dynamics of cellular phenotypes in somatic evolution. *Nat Genet* **56**, 2174–2184 (2024).

256. Lomakin, A. *et al.* Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature* **611**, 594–602 (2022).

257. Mohammadi, F. *et al.* A lineage tree-based hidden Markov model quantifies cellular heterogeneity and plasticity. *Commun Biol* **5**, 1–14 (2022).

258. Aguadé-Gorgorió, G., Costa, J. & Solé, R. An oncospace for human cancers. *BioEssays* **45**, 2200215 (2023).

259. Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J. & Elith, J. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs* **92**, e01486 (2022).

260. Sillero, N. *et al.* Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. *Ecological Modelling* **456**, 109671 (2021).

261. Chollet Ramampiandra, E., Scheidegger, A., Wydler, J. & Schuwirth, N. A comparison of machine learning and statistical species distribution models: Quantifying overfitting supports model interpretation. *Ecological Modelling* **481**, 110353 (2023).

262. Kearney, M. & Porter, W. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters* **12**, 334–350 (2009).

263. Okunlola, O. A. *et al.* Spatial regression and geostatistics discourse with empirical application to precipitation data in Nigeria. *Sci Rep* **11**, 16848 (2021).

264. Kazar, B. M. & Celik, M. *Spatial AutoRegression (SAR) Model: Parameter Estimation Techniques*. (Springer US, Boston, MA, 2012). doi:10.1007/978-1-4614-1842-9.

265. Elhorst, J. P. Applied Spatial Econometrics: Raising the Bar. *Spatial Economic Analysis* **5**, 9–28 (2010).

266. Kamel Boulos, M. N. & Wilson, J. P. Geospatial techniques for monitoring and mitigating climate change and its effects on human health. *Int J Health Geogr* **22**, 2 (2023).

267. Wolf-Jacobs, A., Glock-Grueneich, N. & Uchtmann, N. Mapping Out Our Future: Using Geospatial Tools and Visual Aids to Achieve Climate Empowerment in the United States. in *Storytelling to Accelerate Climate Solutions* (eds. Coren, E. & Wang, H.) 339–364 (Springer International Publishing, Cham, 2024). doi:10.1007/978-3-031-54790-4_16.

268. Park, S. & Ko, D. Spatial Regression Modeling Approach for Assessing the Spatial Variation of Air Pollutants. *Atmosphere* **12**, 785 (2021).

269. Yuan, M., Huang, Y., Shen, H. & Li, T. Effects of urban form on haze pollution in China: Spatial regression analysis based on PM2.5 remote sensing data. *Applied Geography* **98**, 215–223 (2018).

270. Wheeler, D. C. & Páez, A. Geographically Weighted Regression. in *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications* (eds. Fischer, M. M. & Getis, A.) 461–486 (Springer, Berlin, Heidelberg, 2010). doi:10.1007/978-3-642-03647-7_22.

271. Fotheringham, A. S., Yang, W. & Kang, W. Multiscale Geographically Weighted Regression (MGWR). *Annals of the American Association of Geographers* **107**, 1247–1265 (2017).

272. *Handbook of Geospatial Artificial Intelligence*. (CRC Press, Boca Raton, 2023). doi:10.1201/9781003308423.

273. Janowicz, K., Gao, S., McKenzie, G., Hu, Y. & Bhaduri, B. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science* **34**, 625–636 (2020).

274. Klemmer, K., Safir, N. S. & Neill, D. B. Positional Encoder Graph Neural Networks for Geographic Data. in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* 1379–1389 (PMLR, 2023).

275. Dong, M., Su, D., Kluger, H., Fan, R. & Kluger, Y. SIMVI reveals intrinsic and spatial-induced states in spatial omics data. *bioRxiv* 2023.08.28.554970 (2024) doi:10.1101/2023.08.28.554970.

276. Marconato, L. *et al*. SpatialData: an open and universal data framework for spatial omics. *Nat Methods* 1–5 (2024) doi:10.1038/s41592-024-02212-x.

277. De Falco, A., Caruso, F., Su, X.-D., Iavarone, A. & Ceccarelli, M. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. *Nat Commun* **14**, 1074 (2023).

278. Patel, A. P. *et al*. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

279. A, L. *et al*. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems* **1**, (2015).

280. Zhang, M. *et al*. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137–143 (2021).

281. Easy Spatial Modeling with Random Forest. https://blasbenito.github.io/spatialRF/.

282. Bärlund, M. *et al*. Increased copy number at 17q22-q24 by CGH in breast cancer is due to high-level amplification of two separate regions. *Genes Chromosomes Cancer* **20**, 372–376 (1997).

283. Grisanzio, C. & Freedman, M. L. Chromosome 8q24–Associated Cancers and MYC. *Genes Cancer* **1**, 555–559 (2010).

284. Peters, G., Fantl, V., Smith, R., Brookes, S. & Dickson, C. Chromosome 11q13 markers and D-type cyclins in breast cancer. *Breast Cancer Res Treat* **33**, 125–135 (1995).

285. Bièche, I. *et al*. Two distinct regions involved in 1p deletion in human primary breast cancer. *Cancer Res* **53**, 1990–1994 (1993).

286. Qian, J. *et al*. A 3q gene signature associated with triple negative breast cancer organ specific metastasis and response to neoadjuvant chemotherapy. *Sci Rep* **7**, 45828 (2017).