THE UNIVERSITY OF
SYDNEY

INTERPRETABLE ANALYTICAL METHODS FOR CHARACTERIZING DISEASE-ASSOCIATED PHENOTYPIC MODULATION OF CELLS

ELIJAH SAAH WILLIE

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

School of Mathematics and Statistics
Faculty of Science
The University of Sydney

05 January 2026

To Aisha

whose boundless curiosity, quiet strength, and fierce love for life continue to guide and inspire me.
Your laughter still echoes in my mind, your encouragement still fuels my resolve, and your memory lives in every page of this work.

May these efforts stand as a small tribute to the joy you brought into the world and to the dreams we once imagined together.

Forever in my heart, always in my science.

STATEMENT OF ORIGINALITY

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

The research in this thesis was supported by the University of Sydney Postgraduate Excellence Award and Research Training Program scholarship awarded to the PhD Candidate.

During the preparation of the thesis I used Claude for the purposes of text enhancement. The use of this generative AI tool includes improving sentence clarity and grammatical structure. For example, refining verbose passages such as "The cells which were measured using flow cytometry and they showed significant variation" to "Cells measured using flow cytometry showed significant variation," and ensuring consistency in technical terminology throughout the document. I confirm that where text was modified by generative AI, the content was reviewed for possible errors, inaccuracies, and bias. I take full responsibility for the submitted thesis and ensures the work is my own and I have used generative AI within the parameters of use.

Signature:

Elijah Saah Willie, 05 January 2026

ABSTRACT

Realizing the analytical potential of single-cell technologies requires computational frameworks that address fundamental challenges in characterizing cellular heterogeneity and disease-associated phenotypes. High-throughput cytometry and spatial omics platforms have advanced substantially in their resolution and throughput, enabling simultaneous measurement of dozens of parameters across millions of cells. However, these technologies present analytical obstacles including technical artifacts, batch effects, and the complexity of spatial interactions that limit biological discovery and clinical application. Addressing challenges such as robust cell type identification, cross-cohort transferability, and spatial variance decomposition necessitates development of innovative computational approaches. To this end, this thesis contributes to single-cell computational biology by (1) establishing a multiview framework that harmonizes correlation and distance metrics through ensemble learning, overcoming similarity metric limitations in imaging cytometry to enable robust identification of cellular phenotypes, (2) developing transferable deep learning approaches combining batch-agnostic normalization with hierarchically-structured feature selection to enable cross-institutional validation of cytometry-based biomarkers, and (3) creating variance decomposition methods for spatial transcriptomics that quantify cell type-specific interactions while correcting for lateral spillover, revealing how proximity modulates gene expression programs in breast cancer and melanoma. The frameworks outlined in this thesis provide validated computational tools advancing single-cell analysis and establish foundations for future development of interpretable methods characterizing disease-associated cellular modulation.

PUBLICATIONS

The majority of the methods, concepts, analyses and results contained in this thesis have appeared previously in publications and pre-prints listed below:

1. **Elijah Willie**, Pengyi Yang, Ellis Patrick, The impact of similarity metrics on cell-type clustering in highly multiplexed in situ imaging cytometry data, Bioinformatics Advances, Volume 3, Issue 1, 2023, vbad141, https://doi.org/10.1093/bioadv/vbad141

2. **Elijah Willie**, Shreya Rao, Gemma Figtree, Jean Yang, Barbara Fazekas de St Groth, Helen McGuire, Ellis Patrick, dioscRi enables transferable prediction of clinical outcomes in multi-parameter cytometry data, bioRxiv 2025.09.08.675022; doi: https://doi.org/10.1101/2025.09.08.675022

The following publications, not included in this thesis, are the result of collaborative research that I contributed to throughout my candidature. They all align with the topics addressed in this thesis

1. Lin, Y., Cao, Y., **Willie, E**. et al. Atlas-scale single-cell multi-sample multi-condition data integration using scMerge2. Nat Commun 14, 4272 (2023). https://doi.org/10.1038/s41467-023-39923-2

2. A.L. Ferguson, T. Beddow, E. Patrick, **E. Willie**, M.S. Elliott, T.H. Low, J. Wykes, M.H. Hui, C.E. Palme, M. Boyer, J.R. Clark, J.H. Lee, U. Palendira, R. Gupta, Spatial mapping reveals unique cellular interactions and enhanced tertiary lymphoid structures in responders to anti-PD-1 therapy in mucosal head and neck cancers, doi: https://doi.org/10.1101/2024.04.18.590189

AUTHORSHIP ATTRIBUTION STATEMENT

Chapter 2 of this thesis is published as Willie *et al.* (2023). I collected the data, performed the experiments, and interpreted the results. I also wrote the majority of the manuscript under the supervision of A/Prof. Ellis Patrick and A/Prof. Pengyi Yang.

Chapter 3 is available on bioRxiv as Willie *et al.* (2025) and has been submitted to *Nature Machine Intelligence*. I collected the data, developed the deep-learning package, conducted the experiments, interpreted the results, and drafted the manuscript with input from A/Prof. Ellis Patrick, A/Prof. Helen McGuire, Prof. Gemma Figtree, Shreya Rao, and Prof. Jean Yang.

Chapter 4 is currently being drafted for review. I collected the data, conducted the experiments, interpreted the results, and drafted the manuscript with input from A/Prof. Ellis Patrick and Shreya Rao.

I acknowledge that the publisher's policy grants permission for the use of published material within this thesis.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Signature:

Elijah Saah Willie, 05 January 2026

As primary supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Signature:

Ellis Patrick, 05 January 2026

As secondary supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Signature:

Pengyi Yang, 05 January 2026

ACKNOWLEDGMENTS

While pursuing a doctorate may appear to be a solitary endeavor, this achievement would not have been possible without the support of many individuals. I extend my heartfelt gratitude to the following.

First, I thank my Lord and Savior Jesus Christ for His provision, unconditional love, and saving grace, through whom all things are possible.

I am deeply grateful to my supervisor, Dr. Ellis Patrick, for his unwavering support, particularly during challenging periods of my PhD. His belief in me and encouragement helped push me beyond my limits. His ability to constructively challenge my ideas was instrumental in stimulating intellectual growth and shaping me as both an individual and academic.

To my family, my lovely mother Sianeh, my beautiful sisters Selina and Sylvia, and my wonderful nieces and nephews, thank you for your unconditional love, nurturing spirit, and constant encouragement. You all are the foundation upon which all my strength is built. None of this would have been possible without you.

I extend my gratitude to the Uduman family for their continuous support through both triumphs and hardships. Thank you Karl for entertaining my ideas and providing guidance; Lalangi for your caring nature and ensuring I was always well-fed; Dmitry and Viran for taking on the role of older brothers and tolerating me with patience. Your guidance and zest for life have been truly inspirational.

I thank Dr. Leonid Chindelevitch, who cultivated my passion for research during my undergraduate studies. Your commitment to my development opened countless opportunities, and your brilliance and love for science ignited the spark that brought me here.

To my colleagues and friends at the Precision Data Science Lab in Sydney, including Farhan, Max, Daniel, Jackson, Kristy, Rojashree, Martin, Rajan, Nick, Sean, Lijia, Harry, Jamie, Andy, and Shreya, thank you for making this journey enjoyable and enriching. I learned immensely from each of you through our discussions, collaborations, and social gatherings.

Finally, I thank Reighen, who joined this journey in its final stages yet provided invaluable support during the thesis writing process. Thank you for your patience and encouragement.

CONTENTS

1

INTRODUCTION

1.1 CORE BIOLOGICAL CONCEPTS

In recent years, cellular resolution genomics and imaging have transformed how we study disease. Traditional bulk assays average signals across many cells, which can mask rare subpopulations that shape therapeutic response and drive relapse. By measuring thousands of features per cell, and increasingly within intact tissue, single-cell and spatial technologies reveal heterogeneous cell states, lineage programs, and microenvironmental interactions that remain invisible in tissue aggregates. This shift from tissue averages to cell-level readouts provides more accurate mechanistic insight and supports the development of personalised, context-aware treatment strategies.

Understanding cellular behaviour, therefore, requires embracing dynamism. Cells transition through states under developmental programmes, microenvironmental cues, and external signals, and their gene and protein profiles are measurable readouts of these trajectories. Within tissues, spatial proximity and neighbourhood composition further modulate these states, making *both identity and location* essential for explaining physiology and pathology. Such complexity, in turn, demands complementary analytical approaches. Two technological families now enable tissue studies at single-cell resolution. Multi-parameter cytometry profiles thousands to millions of cells across dozens to thousands of markers, providing statistical power for cohort-scale inference. Spatially resolved assays measure molecular targets *in situ*, preserving cellular architecture

and evidence of cell-cell relationships. Together, they turn static cell catalogues into dynamic maps of tissue organisation.

A fundamental distinction exists between protein and RNA measurements that shapes analytical approaches throughout this thesis. Cytometry technologies (flow cytometry, mass cytometry, and imaging mass cytometry) measure protein abundance, typically quantified as transformed fluorescence or ion intensities. After appropriate transformations such as arcsinh or biexponential scaling, these protein measurements approximate continuous distributions that are well-modelled by Gaussian assumptions, enabling the use of standard statistical methods including linear models and correlation-based clustering metrics. In contrast, spatial transcriptomics and single-cell RNA sequencing measure discrete transcript counts, which exhibit characteristic overdispersion where variance exceeds the mean. These count data follow negative binomial distributions, requiring specialised statistical frameworks such as generalised linear models with appropriate link functions. This distinction between continuous protein measurements and discrete RNA counts has direct implications for method development: approaches that work well for cytometry data cannot be naively applied to transcriptomic data, and vice versa. Each chapter of this thesis explicitly addresses this distinction by developing methods appropriate to the underlying data distribution.

1.1.1 *The evolution of single-cell technologies*

The transition to single-cell biology has unfolded in waves, each addressing limitations of the previous generation while introducing new analytical challenges.

Flow cytometry, introduced in the 1970s, pioneered single-cell analysis (Bonner et al., 1972). Early instruments measured only two fluorescent proteins; however, this technology enabled unprecedented resolution of cellular heterogene-

ity. By the 2000s, clinical cytometers routinely quantified 15–20 markers, establishing immunophenotyping as standard for haematological malignancies.

The constraint of fluorescence spectral overlap prompted the next leap. Mass cytometry (CyTOF) replaced fluorophores with metal isotopes detected by mass spectrometry (Bendall et al., 2011). Eliminating spectral overlap enabled the measurement of 40–50 proteins per cell, reframing immune differentiation as a continuum rather than discrete bins. Greater resolution arrived with computational costs; millions of cells in dozens of dimensions demanded new analytical frameworks.

Suspension-based technologies, however, discard spatial context. Such limitations spurred imaging platforms. Imaging Mass Cytometry (IMC) and Multiplexed Ion Beam Imaging (MIBI-TOF) are coupled with CyTOF-level multiplexing, providing subcellular spatial resolution and revealing features such as tertiary lymphoid structures, stem-cell niches, and architectural disruptions accompanying metastasis (Giesen et al., 2014a; Angelo et al., 2014). Cyclic immunofluorescence (e.g., CODEX) pushed multiplexing further, surpassing 60 markers through iterative staining and imaging (Goltsev et al., 2018).

In parallel, transcriptomic profiling of tissues using technologies that can measure thousands of mRNA molecules has undergone its own evolution to achieve single-cell resolution. Early single-cell RNA sequencing required manual cell isolation and profiled only a few hundred cells (Tang et al., 2009). Droplet methods (Drop-seq, inDrop) and commercial platforms (10x Genomics) have increased throughput by orders of magnitude (Macosko et al., 2015; Klein et al., 2015; Zheng et al., 2017), enabling the comprehensive mapping of thousands of cells across organs. The Human Cell Atlas and related efforts now profile millions of cells, introducing integration challenges across laboratories, platforms, and conditions (Regev et al., 2017; Rozenblatt-Rosen et al., 2017; Luecken and Theis, 2019a; Tran et al., 2020b).

Recent developments in spatial transcriptomics now capture transcriptomes while preserving spatial context. The original barcoded-array platform measured genome-wide expression directly from tissue sections (Ståhl et al., 2016). Subsequent innovations improved resolution and scale. Slide-seq and HDST approached near-single-cell resolution (Rodriques et al., 2019; Vickovic et al., 2019), while *in situ* hybridisation (MERFISH, seqFISH) provided subcellular localisation for targeted panels (Chen et al., 2015; Eng et al., 2019a). Commercial systems have accelerated adoption by bridging molecular profiling with tissue architecture. Instruments such as CosMx and Xenium reveal how spatial context shapes expression patterns across biological systems, from neuronal circuit organisation in brain regions (Wang et al., 2025; Yao et al., 2021) to tumour–immune cell interfaces that determine therapeutic response (Hunter et al., 2021; Arora et al., 2023a; Oliveira et al., 2025).

1.2 FROM MEASUREMENTS TO QUESTIONS

These technologies have transformed capabilities for measuring proteins and mRNA, yet they suffer from technical artefacts that can obscure the very biology they promise to reveal (Table 1.1). Technical effects propagate through pipelines from preprocessing to cell identification and spatial analysis. Defining, modelling, and mitigating these effects frames the present frontier of computational single-cell biology.

1.2.1 *Data characteristics and preprocessing challenges*

Modern single-cell and spatial profiling technologies generate data at a scale that exceeds the capabilities of conventional analysis workflows. For example, mass cytometry (CyTOF) can routinely profile millions of cells with 40-50 features each (Spitzer and Nolan, 2016). At the same time, spatial transcriptomics

Platform	Scale	Type	Artefacts	Challenges
Flow cytometry	10^{4-6} cells 15–30 markers	Protein	Spectral overlap, autofluorescence	Subjective gating, standardisation
Mass cytometry	$\sim 10^{6}$ cells 40–50 markers	Protein	Channel spillover, oxides, drift	Batch integration, hierarchy clustering
Imaging (IMC/MIBI)	10^{4-5} cells/img 40–60 markers	Protein	Lateral spillover, hot pixels	Segmentation, neighbourhoods
scRNA-seq	10^{4-6} cells 20K+ genes	RNA	Ambient RNA, doublets, dropout	Normalisation, batch integration
Spatial transcriptomics	10^{3-5} spots or 10^{5-6} molecules	RNA	RNA diffusion, optical crowding	Deconvolution, spatial patterns

Table 1.1: Comparison of high-dimensional single-cell and spatial omics technologies. Each platform presents distinct data scales, molecular modalities, technical artefacts, and analytical challenges that require specialised computational approaches.

captures the expression of hundreds to thousands of genes across hundreds of thousands of spatially resolved cells (Moses and Pachter, 2022). These massive datasets often exceed standard memory and computational limits, necessitating the use of specialised data structures and streaming algorithms (Hao et al., 2021a).

Meeting computational demands alone is insufficient for reliable inference, as each technology introduces specific artefacts that can confound naïve analyses. In CyTOF, for instance, signal spillover between neighbouring mass-to-charge channels causes 3-5% contamination, which is less severe than fluorescence cross-talk, but still capable of generating artificial cell populations (Chevrier et al., 2018). While compensation methods can mitigate these effects, optimal computational strategies remain a topic of debate (Schuyler et al., 2019).

Technical artefacts can vary between batches of processed samples, introducing batch effects that can further confound or mask the biological signal. Batch effects are typically systematic, non-biological variations introduced by differ-

ences in instrument calibration, reagent batches, experimental protocols, operators, or research centres (Nowicka et al., 2019; Takahashi et al., 2021; Gagnon, 2025). They can lead to shifts in marker intensities, detection efficiencies, and even cell-type composition across experiments. The scale of such effects is well documented. Multi-centre studies have reported non trivial differences in specific marker intensities (Leipold et al., 2018), with coefficients of variation exceeding 30% in standardised panels across labs (Finak et al., 2016). In such settings, simple normalisation approaches like global z-scoring risk distorting true biological signals by misaligning cell-type distributions or obscuring real heterogeneity (Butler et al., 2018).

Spatial assays introduce additional complications driven by both the underlying physics of measurement and the challenge of attributing molecules to individual cells. For instance, in imaging mass cytometry, lateral spillover can occur when ablated material contaminates neighbouring pixels, creating halo artefacts around highly expressing cells (Damond et al., 2019a). Similarly, *in situ* transcriptomics experiments typically suffer from RNA diffusion, where, when tissues are fixed, permeabilised, or otherwise processed, RNA molecules can leak out of the cells. Furthermore, imaging-based *in situ* transcriptomics can also suffer from optical crowding in densely labelled regions, which leads to systematic undercounting of transcripts (Moffitt et al., 2022). If uncorrected, these biases can distort downstream analyses such as neighbourhood quantification, cell-cell interaction inference, and spatial differential expression.

Effective preprocessing must therefore carefully balance noise reduction with signal preservation. Over-correction risks eliminating meaningful biological variation, while under-correction allows artefacts to dominate downstream interpretation. These challenges collectively underscore the need for analytical frameworks that explicitly model both technical noise and biological variability (Luecken and Theis, 2019b).

1.2.2 *Cell type identification in high-dimensional Space*

Manual gating has long served as the gold standard for defining cell types based on marker gene expression in flow and mass cytometry, offering unparalleled interpretability through expert-defined sequential biaxial plots that construct parent-child hierarchies (Perfetto et al., 2004; Maecker and Trotter, 2006). In flow cytometry, cell types are traditionally defined by the presence or absence of specific surface markers, with manual gating providing transparent decision boundaries that are explicit, traceable, and grounded in biological knowledge. However, as the number of features increases from dozens in conventional flow cytometry to hundreds or thousands in spectral cytometry and imaging platforms, manual gating faces increasing challenges. The curse of dimensionality becomes particularly acute, as pairwise distances converge and traditional similarity metrics degrade (Kiselev et al., 2019). Beyond the combinatorial explosion of possible bivariate plots when examining 40+ markers, manual strategies suffer from operator bias, poor cross-site reproducibility, and inability to capture multivariate relationships (Aghaeepour et al., 2013). Most critically, manual gating cannot be generalised to new datasets without complete re-analysis, limiting its utility for extensive cohort studies or clinical deployment, where consistent, transferable definitions are essential (Weber and Robinson, 2016).

Automated clustering algorithms promise to address these scalability challenges while uncovering cell populations that may be invisible to bivariate analysis. Graph-based methods, such as Phenograph (Louvain) and Leiden, have emerged as popular choices, constructing k-nearest neighbour graphs that naturally handle high-dimensional data (Levine et al., 2015b; Traag et al., 2019). Nevertheless, these approaches introduce their own interpretability challenges. Clusters lack the clear parent-child relationships of manual gating, require sub-

stantial effort to annotate biologically, and remain sensitive to technical artefacts that can fragment or merge true populations (Freytag et al., 2018).

A fundamental but often overlooked aspect of clustering is the choice of similarity metric. The work done by (Kim et al., 2019a) demonstrated that correlation-based metrics (Pearson, Spearman) consistently outperform Euclidean distance for single-cell RNA-seq clustering, finding that correlation metrics better capture co-expression patterns while remaining robust to technical variation (Kim et al., 2019a). Such insights are particularly relevant for cytometry. FlowSOM, despite its widespread use, defaults to Euclidean distance. This choice may be suboptimal given the scale differences between markers and batch-specific intensity shifts common in cytometry data (Van Gassen et al., 2015a). The superiority of correlation metrics extends beyond transcriptomics. In protein-based assays, where marker expression spans multiple orders of magnitude and absolute intensities vary between instruments, correlation-based approaches can preserve biological relationships while mitigating technical artefacts.

Modern clustering algorithms must not only be accurate but also memory-efficient and parallelisable. Self-organising maps offer one solution, reducing millions of cells to thousands of representative nodes that preserve topological relationships (Van Gassen et al., 2015a). Alternatively, subsampling strategies and mini-batch optimisation enable graph-based methods to scale, though at the potential cost of missing rare populations.

While unsupervised clustering can process high-dimensional data at scale, the resulting populations often lack clear biological meaning without extensive post-hoc annotation. Supervised cell type prediction methods address this gap by directly mapping cells to established biological categories; however, their accuracy depends critically on the quality of the reference datasets. Prediction performance degrades when test data originates from different platforms or patient populations than the training reference (Abdelaal et al., 2019), highlighting that reference datasets must themselves be robust to technical variation while

capturing biological diversity. Methods like scPred (Alquicira-Hernandez et al., 2019) and Seurat's reference mapping (Hao et al., 2021a) incorporate uncertainty quantification and batch correction. However, the fundamental challenge remains that robust cell annotation requires robust references, emphasising that the same principles of cross-platform invariance apply throughout the analysis pipeline.

1.2.3 *Interpretable and cross cohort biomarkers*

Effective cell annotation enables the detection of population shifts and cellular changes associated with disease states or therapeutic interventions. Once cells are accurately classified, we can extract meaningful features from simple cell type proportions that capture immune infiltration patterns to complex expression changes within specific populations that reveal cellular dysfunction. These features form the foundation for predictive modelling in clinical cytometry. Recent work has systematically evaluated feature representations for single-cell data, with scFeatures providing a comprehensive framework that generates six distinct feature types, including cell type proportions, gene expression aggregates, and pathway activities (Cao et al., 2022b). This multi-perspective approach acknowledges that different biological questions require different feature representations. Cell type proportions may capture immune shifts in one disease while marker expression changes within stable populations drive another. The key insight is that robust features must capture biological signal while remaining invariant to technical artefacts.

However, the reproducibility of these inferences and the transferability of predictive models remains challenging. Models that perform well within a single study often fail when applied to external datasets, a phenomenon that extends beyond classical overfitting to encompass systematic technical and biological differences between cohorts (Arvaniti and Claassen, 2017a). This gen-

eralisation challenge is particularly evident in cytometry, where instrument settings, reagent batches, and patient populations vary substantially across sites and time points. Recent clinical applications has shown the importance of feature robustness. In (Robertson et al., 2024), the authors demonstrated that carefully engineered transcriptomic features can predict allograft dysfunction across multiple organ types, provided that these features are specifically designed to be platform-agnostic and batch-robust. Their work revealed that simple proportion-based features often outperformed complex expression signatures in external validation, suggesting that interpretable features may inherently possess greater transferability. This principle extends across modalities. CPOP demonstrated that ratio-based genomic features eliminate platform-specific biases while maintaining predictive power across microarray, RNA-sequencing, and NanoString platforms (Wang et al., 2022a), while treekoR showed that measuring cell proportions relative to parent populations rather than total populations dramatically improves clinical associations in cytometry data (Chan et al., 2021). Together, these findings challenge the assumption that model complexity necessarily improves generalisation, instead suggesting that relationally defined features, capturing fundamental biological relationships, provide superior cross-platform robustness.

Traditional batch-alignment strategies attempt to remove technical variation in mass and flow cytometry data through data transformation. Harmony, initially developed for scRNA-seq but widely adopted in CyTOF analysis, performs iterative clustering and correction while preserving cell-type distinctions (Korsunsky et al., 2019). Scanorama aligns flow cytometry panels with different marker sets by identifying mutual nearest neighbours to define shared latent spaces (Hie et al., 2019), while CyCombine, explicitly designed for CyTOF data, employs hierarchical modelling to account for both technical and biological variation (Trussart et al., 2020). Although effective for exploratory analyses, these embedding-based approaches fundamentally alter the data space. The resulting transformations can distort quantitative relationships between features and

outcomes, limiting their utility for downstream prediction tasks where maintaining biological effect sizes is critical (Tran et al., 2020a).

Deep learning methods offer an alternative paradigm where models learn to be invariant to batch effects without explicit data correction. Variational autoencoders can discover latent representations that capture biological structure while ignoring technical artefacts (Lopez et al., 2018). The scVI framework, primarily applied to single-cell RNA sequencing data, extends this approach by explicitly modelling technical parameters alongside biological variation, enabling cross-platform integration between technologies like 10X Genomics and Smart-seq2 (Gayoso et al., 2022). Methods employing maximum mean discrepancy regularisation learn batch-invariant representations through adversarial training (Dincer et al., 2020). Nevertheless, this sophistication creates further problems. While deep models may achieve superior predictive accuracy, they function as black boxes that obscure the specific features driving predictions. This opacity limits biological insight, complicates clinical interpretation, and prevents the feature-level validation essential for translational applications (Elmarakeby et al., 2021).

The limitations of both batch correction and black-box models highlight the value of interpretable approaches that maintain transferability. When a model's decision process is transparent, practitioners can assess whether learned patterns reflect biological reality or technical artefacts. Interpretable models also facilitate knowledge transfer between studies since researchers can directly compare which features drive predictions across cohorts. This transparency becomes essential in clinical settings where regulatory approval and physician adoption require a clear understanding of the mechanistic basis.

Hybrid architectures (Burkhardt et al., 2022) offer a pragmatic compromise by combining the representational power of deep learning with interpretable prediction layers. For example, CellTypist combines deep learning feature extraction with interpretable linear classifiers for each cell type (Domínguez Conde

et al., 2022b), while scArches uses transfer learning with explicit cell type anchors (Lohoff et al., 2022). Attention mechanisms applied over annotated cell types can highlight which populations drive predictions while maintaining end-to-end learning (Burkhardt et al., 2022). Similarly, structured regularisation that incorporates biological knowledge, such as cell type hierarchies or pathway relationships, can guide models toward biologically meaningful solutions while preserving the flexibility to discover novel patterns. These approaches suggest that the dichotomy between performance and interpretability may be false when models are designed to leverage rather than ignore biological structure.

1.2.4 *Spatial analysis and cellular interactions*

Spatial transcriptomic technologies shift the analytical focus from simply identifying which cell types are present to understanding how their spatial positioning modulates cellular behaviour and gene expression programmes. Additionally, these technologies can detect cell types in fragile tissues or those poorly captured by dissociative methods, as they preserve tissue architecture during measurement (Palla et al., 2022). This paradigm shift reveals that transcriptomic changes observed in disease contexts often reflect not only intrinsic cellular programmes but also extrinsic spatial cues. For example, a tumour cell's gene expression profile is influenced by its proximity to vasculature, immune infiltrates, and stromal barriers. Recent studies demonstrate that extracellular matrix (ECM) remodelling, particularly involving collagen and fibronectin, can create transient stromal barriers that restrict immune cell access to melanoma lesions, thereby contributing to therapy resistance (Hsu et al., 2025).

Recent spatial transcriptomics studies have demonstrated that hundreds to thousands of genes exhibit spatial patterning within tissues, with effect sizes comparable to those associated with disease (Svensson et al., 2018). Gaussian process-based methods such as SpatialDE identify these broad spatial expres-

sion patterns by modelling gene expression as a function of spatial coordinates (Svensson et al., 2018). Similarly, SPARK employs generalised linear spatial models to detect genes with spatial expression patterns while accounting for technical noise and overdispersion common in spatial data (Sun et al., 2020). However, these univariate (singe gene) approaches often overlook the multivariate (multi-gene) nature of spatial regulation and fail to account for cell-type-level changes, where coordinated changes across gene modules define functional microenvironments.

Spatial biology generates complex data requiring specialised analytical approaches to handle simultaneous measurements across multiple genes or markers and cell types. A critical challenge in these analyses involves distinguishing actual biological patterns such as cell-cell interactions, signalling gradients, and tissue architecture from technical artefacts inherent to the measurement platform. Cell segmentation algorithms must delineate boundaries between densely packed cells, yet struggle with overlapping nuclei, irregular cell shapes, and incomplete membrane staining, leading to merged or fragmented cell representations that distort downstream analyses (Stringer et al., 2021; Greenwald et al., 2022; Pachitariu and Stringer, 2022; Bannon et al., 2021). Even with accurate segmentation, lateral spillover remains problematic in imaging mass cytometry, where ablated material from highly expressing cells contaminates neighbouring pixels (Bai et al., 2021b). Similarly, in spatial transcriptomics, RNA diffusion from cells during tissue processing can create artificial signals in adjacent capture spots, obscuring actual spatial gene expression patterns (Moses and Pachter, 2022; Williams et al., 2022). In array-based methods like Visium, lateral diffusion of RNA during permeabilisation and transfer to the capture array limits spatial resolution (Eng et al., 2019a). Without proper compensation for both segmentation errors and spillover artefacts, these technical limitations can lead to false discoveries of cell-cell interactions or misidentification of rare cell types at tissue interfaces. Recent methods, such as expansion spatial transcriptomics, address these limitations by physically expanding tissue prior to RNA capture,

thereby reducing cellular overlap and improving effective resolution while facilitating more accurate cell segmentation (Fan et al., 2023b).

Computational frameworks for spatial analysis reveal how cellular neighbourhoods influence cell function and phenotype through local signalling networks. Spatial variance component analysis provides one such framework, decomposing expression variance into cell-intrinsic and spatially-induced components (Arnol et al., 2019). This approach reveals which genes respond to spatial context versus those that maintain stable expression within cells regardless of location. More recent methods model spatial communication networks, inferring ligand-receptor interactions and their downstream effects on gene expression (Fischer et al., 2023). These approaches can identify whether T cell exhaustion correlates with proximity to regulatory T cells, or how macrophage neighbourhoods create immunosuppressive microenvironments. These analyses address key questions about how tissue organisation drives cellular function.

1.3 THESIS OUTLINE AND CONTRIBUTIONS

This thesis addresses challenges in single-cell and spatial omics data analysis through three methodological contributions that advance both computational rigour and biological interpretability.

Chapter 2 tackles the fundamental problem of cell type identification in highly multiplexed imaging cytometry, where existing methods developed for suspension technologies fail to account for imaging-specific artefacts. It proposes FuseSOM, a multiview ensemble clustering framework that integrates complementary similarity metrics through self-organising maps, and systematically benchmarks distance versus correlation metrics across multiple datasets from various imaging platforms.

Chapter 3 presents dioscRi, a deep learning framework that enables clinical predictions from cytometry data to generalise across patient cohorts by address-

ing the batch effects that typically prevent model transferability. The framework combines a transferable normalisation scheme with interpretable disease-relevant feature selection, improving both prediction accuracy and biological insight for precision medicine applications.

Chapter 4 develops PACE, a statistical framework that reveals how cells alter their gene expression programs based on the identity of their spatial neighbours. By decomposing expression variance into cell type identity, pairwise spatial interactions, and technical artifacts, PACE quantifies which specific cell type pairs drive spatial signals and identifies the gene programs mediating these proximity-driven changes, enabling spatial transcriptomics to reveal not just where cells are, but how proximity to specific neighbors reshapes their molecular identity.

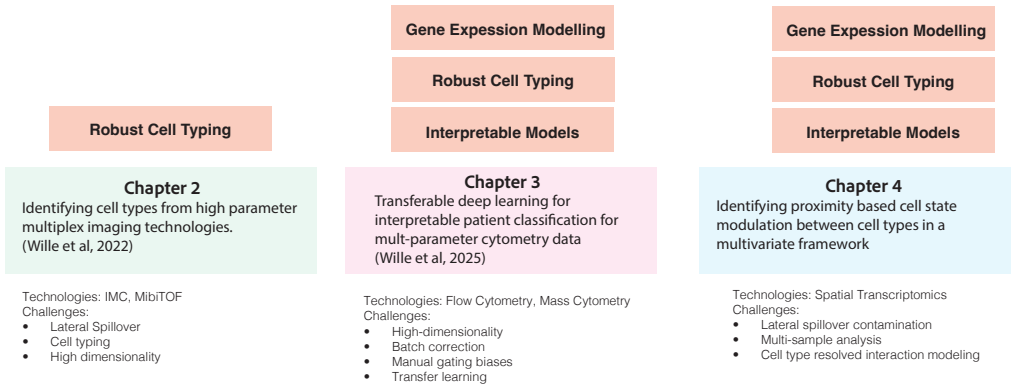

Figure 1.1: Overview of thesis contributions addressing computational challenges in single-cell and spatial omics analysis. Each chapter develops methods that progressively build capabilities in robust cell typing, interpretable models, and gene expression modelling across different single-cell technologies.

Together, these chapters tackle interconnected challenges in high-dimensional cellular analysis, progressing from robust cell type annotation to spatial interaction modelling (Figure 1.1). Cell type identification forms the foundation for all downstream analyses. However, technical artefacts from each measurement platform can obscure true biological signals, whether through spectral overlap in flow cytometry, lateral spillover in imaging mass cytometry, or RNA diffusion in spatial transcriptomics. Building on accurate cell typing, the thesis then addresses how to extract interpretable predictive features that generalise across cohorts despite batch effects, and finally, how to quantify the spatial dependencies that shape cellular behaviour in intact tissues. This progression highlights the need for robust computational frameworks that can handle technical noise while preserving biological variation simultaneously.

In summary, this thesis explores computational strategies for robust cell type identification in imaging cytometry, develops transferable deep learning frameworks for clinical prediction from cytometry data, and introduces statistical methods for quantifying cell type-specific spatial interactions in transcriptomic data. Each contribution addresses distinct analytical challenges including handling platform-specific artifacts in high-dimensional clustering, achieving cross-cohort generalisation despite batch effects, and separating biological proxim-

ity signals from technical contamination. These approaches balance computational sophistication with biological interpretability, recognizing that methods must be both statistically rigorous and practically accessible to the research community. The publicly available R packages developed in this thesis provide validated implementations of these methods for multiview clustering and interpretable prediction. These contributions will benefit researchers working with high-dimensional cellular data and inspire future methodological developments in single-cell and spatial omics analysis.

2

THE IMPACT OF SIMILARITY METRICS ON CELL TYPE CLUSTERING IN HIGHLY MULTIPLEXED IN SITU IMAGING CYTOMETRY DATA

Highly multiplexed *in situ* imaging cytometry assays have enabled researchers to scrutinise cellular systems at an unprecedented level. With the capability of these assays to simultaneously profile the spatial distribution and molecular features of many cells, unsupervised machine learning, and in particular clustering algorithms, have become indispensable for identifying cell types and subsets based on these molecular features. The most widely used clustering approaches applied to these novel technologies were developed for cell suspension technologies. To date, there have been no systematic evaluations of the properties of these methods that are optimal for *in situ* imaging assays.

In this chapter, we assess how the choice of similarity metric affects cell type clustering. We systematically examine how metric choice shapes neighbourhood structure and downstream partitions in high-parameter *in situ* imaging cytometry. Building on these observations, we introduce *FuseSOM*, an ensemble framework that integrates complementary similarity measures via hierarchical multi-view learning atop self-organising maps, yielding topology-preserving and hierarchy-aware cell-type definitions. The accompanying analysis workflow employs stratified subsampling to provide practical guidance on metric selection and resolution setting across panels and tissues, illustrating how multiview integration stabilises clustering decisions in imaging-based studies.

This chapter consists of work published in *Bioinformatics Advances* (2023) (Willie et al., 2023). As first author, I led the method development in collaboration with A/Prof. Ellis Patrick, curated the benchmarking datasets, implemented the full analysis workflow, and developed the accompanying R package released on Bioconductor.

2.1 INTRODUCTION

Technological advancements over the past decade have provided researchers the capability to simultaneously measure multiple molecular features in tissue at subcellular resolution (Lewis et al., 2021). Key technologies that are pioneering a new era for spatially resolved proteomics include imaging mass cytometry (IMC) (Giesen et al., 2014c), multiplexed ion beam imaging by time of flight (MIBI-TOF) (Keren et al., 2019), co-Detection by indEXing (CODEX) (Black et al., 2021) and its successor phenocycler. These technologies can measure approximately 50-100 features with high throughput, enabling researchers to address complex questions about the spatial distribution and interaction of various types of cells *in situ* (Baharlou et al., 2019). A ubiquitous analytical step when analysing highly multiplexed imaging data is defining functionally distinct cell groupings. While there have been recent developments in spatial analysis approaches that simultaneously phenotype cells by their cellular environment and molecular features (Lee et al., 2023; Liu et al., 2023), the most commonly used phenotyping approaches only use molecular features and are not intentionally biased by cellular interactions.

Unsupervised clustering algorithms are valuable tools for discovering both known and novel cell types in highly multiplexed data, even in cases where prior knowledge of the cell types present in an experiment is lacking (Karim et al., 2020). Here we use the terminology *"cell type"* liberally, with clusters also potentially representing distinct known or novel cell states. In our review of

the literature, over 70 percent of manuscripts employing highly multiplexed imaging data for analysis utilised one of three clustering algorithms. IMC and MIBI-TOF data was predominately clustered using either Phenograph (Levine et al., 2015a), a graph-based Louvain community detection method, or FlowSOM (Van Gassen et al., 2015b), a self-organising map approach, while CODEX data was predominately clustered using X-shift, a KNN algorithm accessible through the Vortex GUI (Samusik et al., 2016). In the remaining manuscripts, other Louvain and Leiden graph-based community detection algorithms, hierarchical clustering and K-means clustering were used.

Despite their popularity in imaging modalities, Phenograph, FlowSOM, and X-shift were developed in 2015 and 2016 for suspension cytometry technologies which do not share all of the same technical limitations and noise profiles with tissue-based imaging technologies. Technical artefacts present in most imaging technologies include nonspecific binding (Batth et al., 2020) and lateral marker spillover (Bai et al., 2021a). Additionally, in practice, these methods are often employed to generate a large set of candidate clusters which using expert domain knowledge are then manually clustered, refined, and annotated based on biological features such as key marker expression, and cell localisation. Following this, there exist multiple avenues for further exploration of how clustering algorithms could be tailored for multiplexed imaging data.

Choosing an appropriate similarity metric is crucial for clustering algorithms as it determines how points in a dataset, in our case cells, are partitioned into clusters. Different similarity metrics, often referred to as distance metrics, can yield different clusters. While the Euclidean distance is commonly used in many clustering algorithms, recent studies have shown that correlation-based metrics such as Pearson or Spearman correlation perform better when clustering in other multiplexed single cell technologies (Kim et al., 2019b; Watson et al., 2022). Evaluating the performance of different similarity metrics for defining cell types in multiplexed imaging data may guide the improvement or devel-

opment of new clustering algorithms which are optimal for these exciting technologies.

In this study, we systematically assess the performance of various distance- and correlation-based metrics in 15 imaging datasets, using multiple performance metrics such as the Adjusted Rand Index (ARI), the Normalised Mutual Information (NMI), the FMeasure and the Fowlkes–Mallows Index (FM-Index). We also compare the performance of best-practice clustering methods that currently employ different similarity metrics. Based on our assessment, which highlights the benefits of combining information from multiple similarity metrics, we introduce a new clustering algorithm called FuseSOM. FuseSOM utilises self-organising maps (SOM) and combines multiple similarity metrics through multi-view ensemble learning and hierarchical clustering. This algorithm aims to accurately and robustly identify cell types in multiplexed in situ imaging cytometry assays. Overall, our work demonstrates the impact of similarity metrics on clustering cells in multiplexed imaging cytometry data and proposes FuseSOM as a promising method for the analysis of such data.

2.2 METHODS

Datasets

To benchmark the performance of FuseSOM on imaging datasets from various technologies, we curated a set of 15 datasets generated using different imaging technologies. We selected datasets with human intervention in manually gating cell populations or merging biologically similar clusters. The intention of selecting datasets with manual intervention in defining cell types is to reduce bias towards the original clustering method when evaluating clustering performance. Using these types of datasets also provides higher confidence in the quality of clusters since expert domain knowledge has been applied to scruti-

nise the clusters further. Datasets were sourced from major databases, including Zenodo, Figshare, and Mendeley. When available, we used the version data that had been processed as described in the original manuscript. We also used the same markers and the same final number of clusters for clustering as described in the manuscript. The imaging technologies used included Co-detection by indexing(CODEX) (Black et al., 2021) (four datasets), Imaging Mass Cytometry(IMC) (Giesen et al., 2014b) (six datasets), Multiplexed ion beam imaging by the time of flight (MIBI-TOF) (Keren et al., 2019) (four datasets), and sequential Fluorescence In Situ Hybridisation (seqFISH) (one dataset) (Eng et al., 2019b). See (Table 2.1) for a detailed description of the datasets used.

Evaluation metrics

To evaluate clustering performance for clustering solutions generated across methods, we used a set of methods; the Adjusted Rand Index (ARI), Normalised Mutual Information (NMI), Fowlkes-Mallows index (FM-Index), and the F-Measure (Steinley, 2004; Kvålseth, 2017; Fowlkes and Mallows, 1983; Hripcsak, 2005). The ARI measures the similarity between two data clusterings, adjusting for chance

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

where n_{ij} is the number of pairs of elements that are in the same set in both clusterings, a_i is the total number of pairs in the same set for the first clustering, b_j is the total number of pairs in the same set for the second clustering, and n is the total number of elements.

NMI is a normalisation of the Mutual Information (MI) score to scale the results between 0 (no mutual information) and 1 (perfect correlation) and is defined as follows:

$$NMI(X, Y) = \frac{2 * I(X, Y)}{H(X) + H(Y)} \tag{2}$$

where $I(X, Y)$ is the mutual information between clusters X and Y, and $H(X)$ and $H(Y)$ are the entropies of clusters X and Y respectively.

The Fowlkes-Mallows Index (FM-Index) is the geometric mean of precision and recall, and it is defined as follows,

$$FM = \sqrt{\frac{TP}{TP + FP} * \frac{TP}{TP + FN}} \tag{3}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

The F-Measure which is the harmonic mean of the precision and recall is defined as follows:

$$F = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

where

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

and

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

All these metrics take values between 0 and 1, with 0 being no similarity and 1 being perfect similarity.

Distance metrics

Six types of metrics across two classes that are predominantly used across machine learning clustering literature were used in this study. The two classes include correlation-based and distance-based. The distance-based metrics were Euclidean, Manhattan, and Maximum distance, while the correlation-based metrics included Pearson correlation, Spearman correlation, and Cosine similarity. More formally, let x_{im} and x_{jm} denote the expression of a marker $m = 1, ..., M$ in cell $i = 1, ..., N$ and cell $j = 1, ..., N$, where G and N are the total number of markers and cells, respectively. Led $D = d_{ij}$ be a distance matrix where d_{ij} represents the distance between $cell_i$ and $cell_j$. We can then define the distance-based metrics as follows:

Euclidean distance,

$$d_{ij} = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2} \tag{7}$$

Manhattan distance,

$$d_{ij} = \sum_{m=1}^{M} |x_{im} - x_{jm}| \tag{8}$$

Maximum distance,

$$d_{ij} = \max_{m} |x_{im} - x_{jm}| \tag{9}$$

Similarly, the correlation-based metrics can be defined as follows:

Pearson distance,

$$d_{ij} = \sqrt{2\left(1 - \frac{\sum_{m=1}^{M}(x_{im} - \bar{x}_i)(x_{jm} - \bar{x}_j)}{\sqrt{\sum_{m=1}^{M}(x_{im} - \bar{x}_i)^2}\sqrt{\sum_{m=1}^{M}(x_{jm} - \bar{x}_j)^2}}\right)} \tag{10}$$

Spearman distance,

$$d_{ij} = \sqrt{2\left(1 - \frac{\sum_{m=1}^{M}(r_{im} - \bar{r}_i)(r_{jm} - \bar{r}_j)}{\sqrt{\sum_{m=1}^{M}(r_{im} - \bar{r}_i)^2}\sqrt{\sum_{m=1}^{M}(r_{jm} - \bar{r}_j)^2}}\right)} \tag{11}$$

Cosine distance

$$d_{ij} = \sqrt{2\left(1 - \frac{\sum_{m=1}^{M} x_{im}x_{jm}}{\sqrt{\sum_{m=1}^{M} x_{im}^2}\sqrt{\sum_{m=1}^{M} x_{jm}^2}}\right)} \tag{12}$$

where r_{ij} is the rank of marker m in $cell_i$, $\bar{x_i}$ is the mean expression of $cell_i$, $\bar{x_j}$ is the mean expression of $cell_j$, $\bar{r_i}$ is the mean expression rank of $cell_i$, and $\bar{r_j}$ is the mean expression rank of $cell_j$,

Clustering algorithms

For this work, a few clustering algorithms were used for comparing the effects of distance metrics on clustering outcomes. These algorithms include hierarchical clustering, FlowSOM (Van Gassen et al., 2015b), K-means, and Phenograph (Levine et al., 2015a).

The *hierarchical clustering* builds a hierarchy of clusters by either merging smaller clusters into larger ones (agglomerative) or dividing a large cluster into smaller ones (divisive) using a linkage function. The process continues iteratively, resulting in a tree-like diagram called a dendrogram that represents the nested clusters. The agglomerative version was used with the average linkage function (Nielsen, 2016).

The *FlowSOM* utilises self-organising maps (SOMs) and hierarchical clustering to analyse and visualise complex datasets, particularly in flow cytometry. It groups cells into nodes on a grid based on similarity, providing insights into data structures. This technique is especially valuable in identifying and understanding cell populations. The FlowSOM algorithm (version 2.8.0) was obtained from Bioconductor.

The *K-means* clustering partitions data into 'k' clusters by repeatedly assigning data points to the nearest centroid and recalculating the centroids. The process

continues iteratively until the centroids stabilise. The *Base R* implementation of the K-means algorithm was used.

The *Phenograph* is a clustering method that constructs a k-nearest neighbour (kNN) graph from data, usually applied to single-cell data analysis. Community detection is performed on this graph using the Louvain method to identify clusters or communities of similar nodes. The *phenograph* function from the *ReductionWrappers version 2.5.4* (Smith, 2023) R package was used.

FuseSOM

Here the FuseSOM algorithm is described. The algorithm starts by taking in an m by n matrix where m is the number of cells and n is the number of markers. Next, FuseSOM uses this matrix to generate a Self Organising Map. A Self Organising Map (SOM) is a type of dimensionality reduction algorithm that maps points in a high dimensional space $(d > 2)$ to a lower dimensional space $(d = 2)$. The SOM architecture preserves the topological relationships between points when reduced to a lower dimension. The SOM also provides a set of points called prototypes which are representations of points in the higher dimensional space. The SOM architecture was chosen due to its ability to preserve topological structures of the input data, and its ability to represent complex non-linear relationships in the data, allowing them to capture more intricate patterns and dependencies. Like cluster centres in the k-means algorithm, many points can be mapped to a single prototype. For a more thorough treatment of SOMs, see (Miljkovic, 2017). The *YASOMI* package (version 0.3) was obtained and modified to implement the self-organising map used in FuseSOM (Rossi, 2012).

In this work, we generate a SOM, and the prototypes are used for clustering. After clustering, the clusters are projected back to the original data to classify the original data points. The SOM algorithm requires a 2-d grid(x,y) size, which

determines the number of prototypes. Grid sizes of varying shapes are allowed. However, square grids are typically used. To estimate the size of the grid for a dataset, we use the method described in (Patterson et al., 2006). This method computes the number of eigenvalues of a covariance matrix significantly different from the Tracy-Widom distribution (Tracy and Widom, 1994).

Next, multiview integration combines the Pearson correlation distance, Cosine distance, Spearman correlation, and Euclidean distance between the prototypes to generate a final distance matrix for clustering. Multiview ensemble learning is a machine learning strategy that employs multiple diverse views or perspectives of the same data to enhance predictive modelling. In a typical multiview ensemble learning setup, each view represents a unique set of features or a unique preprocessing or transformation of the data. These different views capture various aspects of the data, which when combined, offer a more comprehensive and potentially more accurate representation. In this work, the different views are represented by the various distances computed between the cells. To combine these views, we adopted a multiview integration (Fuse) method to combine the four transformed matrices (Melssen et al., 2006). Formally, the multiview integration can be defined as follows:

$$D_{fused}[i,k] = \sum w_i * D[ijk] \tag{13}$$

where $D_{fused}[i,j]$ is the combined dissimilarity between samples j and k, w_i is the weight assigned to the i^{th} dissimilarity matrix, and $D[ijk]$ is the dissimilarity between j and k for the i^{th} dissimilarity matrix. All distances are weighted equally. We evaluated several weighting methods, and equal weighting (Fuse) and Dunn2 emerged as the top performers with comparable results across metrics; we selected equal weighting (Fuse) for FuseSOM due to its simplicity and interpretability. See (Supplementary Figure S5).

The *analogue* package (version 0.17-6) is used to perform the multiview integration (Simpson and Oksanen, 2021). The *psych* package (version 2.3.3) transforms

the similarity matrices into distance matrices (Revelle, 2022). Correlations and cosines are transformed into distances by using the formula:

$$D = \sqrt{2(1 - r_{ij})} \tag{14}$$

where r_{ij} is the value of the correlation or cosine between feature i and j. Finally, to generate final cluster labels using the integrated distance matrices, FuseSOM takes in a parameter k which is the number of desired clusters. Next, hierarchical clustering using the average linkage function is used to generate the final clustering solution. The *FCPS* package (version 1.3.1) was used for hierarchical clustering (Thrun and Stier, 2021). An overview of the FuseSOM algorithm is shown in Figure 2.1.

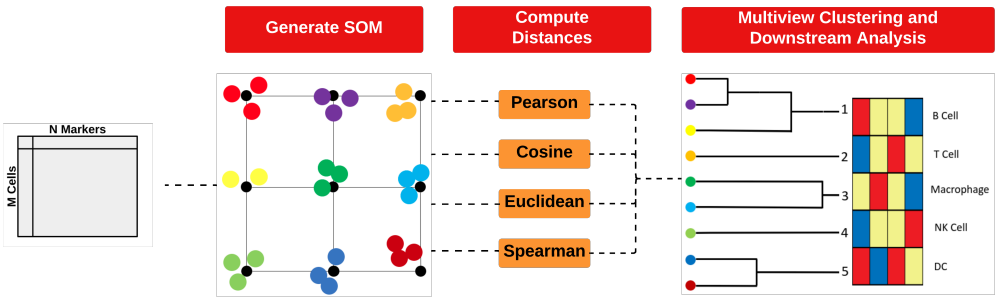

Figure 2.1: Overview of FuseSOM: This new scalable algorithm uses (i) A Self Organising Map to reduce the dimension of the data while preserving its topological structure; (ii) A multiview integration of various similarity metrics to capture all relevant signals; and (iii) Hierarchical clustering to generate a clustering solution for further downstream analysis.

Clustering framework

For consistency when comparing distance metrics across various datasets, each dataset was sampled five times to obtain 20K cells. After this, we executed each clustering algorithm and recorded the scores based on different evaluation metrics. To compare with FuseSOM, we used a substratification framework for each dataset. This framework accepts a dataset and produces five stratified samples. The purpose of substratification is to account for potential variability

in clustering outcomes. Through this framework, when a dataset is inputted, it yields five stratified datasets. Stratification involves selecting 50% of cells from every annotated class (Kim et al., 2019b).

Cluster size estimation

For most clustering algorithms, the number of clusters k is an important hyper-parameter that must be set. To this end, many methods have been developed to help practitioners choose an appropriate number for their dataset. We have included well-known methods for estimating the number of clusters as part of the FuseSOM package. These methods include the Gap statistic, the Slope statistic, the Jump statistic, the Silhouette statistic, and the within-cluster distance (WCD) (Tibshirani et al., 2001; Fujita et al., 2014; Sugar and James, 2003; Rousseeuw, 1987).

The *Gap statistic* compares the change in the within-cluster sum of squares (WSS) from the observed data to that of a random clustering. A large gap value indicates the observed data has a more pronounced clustering structure than expected under a random scenario.

The *Silhouette statistic* quantifies how close each data point in one cluster is to data points in neighbouring clusters, with values ranging from -1 to 1; higher values indicate better-defined clusters.

The *Jump statistic* evaluates the rate of increase in the WSS as a function of the number of clusters, with large jumps indicating the possible presence of distinct groups.

The *slope statistic* identifies an "elbow" or bend in the WSS plot; the point before the stabilisation or decline in the slope can suggest an optimal number of clusters.

The within-cluster distance (WCD) measures the compactness of clusters. A smaller value indicates tighter, more well-defined clusters. Each of these statistics offers unique insights and their combined interpretation aids in selecting an appropriate number of clusters.

We also implemented a *Discriminant* method for estimating the number of clusters based on the projection pursuit of the discriminant maximum clusterability. To accomplish this, we couple hierarchical clustering with discriminant analysis and multimodality testing to estimate the number of clusters (Etemad and Chellappa, 1997; Mokari et al., 2018; Samadani et al., 2013; Ameijeiras-Alonso et al., 2019; Hartigan and Hartigan, 1985; Silverman, 1981). First, we generate a dendrogram using hierarchical clustering with average linkage. Next, for each node in the resulting tree, we project the two classes onto a line such that both classes are well separated. See (Supplementary Figure S6). The dip test for multimodality testing is then applied to the distribution of the points along this line (Hartigan and Hartigan, 1985). The family-wise error rate (FWER) is controlled using the method described in (Meinshausen, 2008). Finally, the number of nodes with significant p-values is returned as the number of clusters.

2.3 RESULTS

2.3.1 *Evaluating the impact of similarity metrics*

To assess the impact of similarity metrics on clustering performance, we performed hierarchical clustering on a MIBI-TOF dataset using correlation-based metrics (Pearson, Spearman, and Cosine) or distance-based metrics (Euclidean, Manhattan, and Maximum) (McCaffrey et al., 2022). To assess performance, we compared the hierarchical clusters with the manually curated cell type labels identified in the manuscript (Figure 2.2). On average, correlation-based metrics

outperform distance-based metrics by 8.0% for ARI, 10.7% for NMI, 1.70% for FM-index, and 8.0% for F-Measure.

To provide a comprehensive assessment of the performance of similarity metrics, we quantified the clustering performance of the metrics on 15 multiplexed in situ imaging cytometry datasets (Table 2.1). These datasets were chosen as each had some manual intervention when cell type labels were defined. Each dataset was randomly subsampled to 20K cells five times, and each subset was clustered using hierarchical clustering with all the similarity metrics. Finally, the average was taken across the five subsets. Across the 15 datasets, correlation-based metrics consistently outperformed distance-based metrics (Figure 2.3, Supplementary Figure S1), more accurately recapitulating the manually curated cell-type labels from their original publications. These results show the efficacy of correlation-based metrics in hierarchical clustering.

Dataset	Technology	Num Markers	Num Cells	Num Celltypes	Disease	Tissue	Cell Annotation Method
Schurch et al (Schürch et al., 2020)	CODEX	49	258,385	29	Colorectal Cancer	Colon	X-shift then supervised merging
Phillips et al (Phillips et al., 2021)	CODEX	52	117,170	21	Lymphoma	Skin	X-shift then supervised merging
Brbic et al (Brbić et al., 2021)	CODEX	48	248,285	21	None	Small intestine/colon	Manually annotated
Brbic et al (Brbić et al., 2021)	CODEX	44	219,926	13	Barrett's esophagus	Tonsil	Manually annotated
Moldoveanu et al (Moldoveanu et al., 2022)	IMC	12	227,592	10	Melanoma	Skin	PhenoGraph + Kmeans
Van Maldegem et al (van Maldegem et al., 2021)	IMC	17	282,837	16	Lung Cancer	Lung	PhenoGraph + manual splitting
Hoch et al (Hoch et al., 2022)	IMC	41	864,263	10	Melanoma	Skin	Manual gating
Rendeiro et al (Rendeiro et al., 2021)	IMC	38	515,791	17	Covid-19	Lung	Leiden + manual merging
Damond et al (Damond et al., 2019b)	IMC	36	252,059	16	Type 1 Diabetes	Pancreas	Supervised classifier
Bortolomeazzi et al (Bortolomeazzi et al., 2021)	IMC	30	218,615	9	Colorectal Cancer	Colon	Seurat + DBSCAN
Risom et al (Risom et al., 2022)	MIBI-TOF	22	69,672	23	Breast Cancer	Breast	FlowSOM + manual merging
Keren et al (Keren et al., 2019)	MIBI-TOF	16	201,656	6	TN Breast Cancer	Various	FlowSOM + hierarchical merging
McCaffrey et al (McCaffrey et al., 2022)	MIBI-TOF	37	30,943	16	Tuberculosis	Various	Iterative FlowSOM
Liu et al (Liu et al., 2022)	MIBI-TOF	12	345,490	8	Various Cancers	Various	FlowSOM + hierarchical merging
Lohoff et al (Lohoff et al., 2022)	seqFISH	50	57,536	24	None	Mouse embryos	Louvain on top 50 PCs

Table 2.1: Imaging Datasets Used

Next, we assessed if the performance differences between correlation and Euclidean distance are consistent across multiple clustering methods. We reconfigured PhenoGraph, FlowSOM, and K-means clustering to use correlation instead of Euclidean (Levine et al., 2015a; Van Gassen et al., 2015b). As previously, the 15 datasets were randomly subsetted to 20K cells five times and the performance scores were calculated for each subset. While there does not appear to be a benefit in clustering performance for PhenoGraph ($p > 0.05$), the overall PhenoGraph performance for both distance measures is worse when compared to the other methods using correlation-based distances (Figure 2.4, Supplementary Figure S2). For K-means, hierarchical clustering, and FlowSOM, we observe

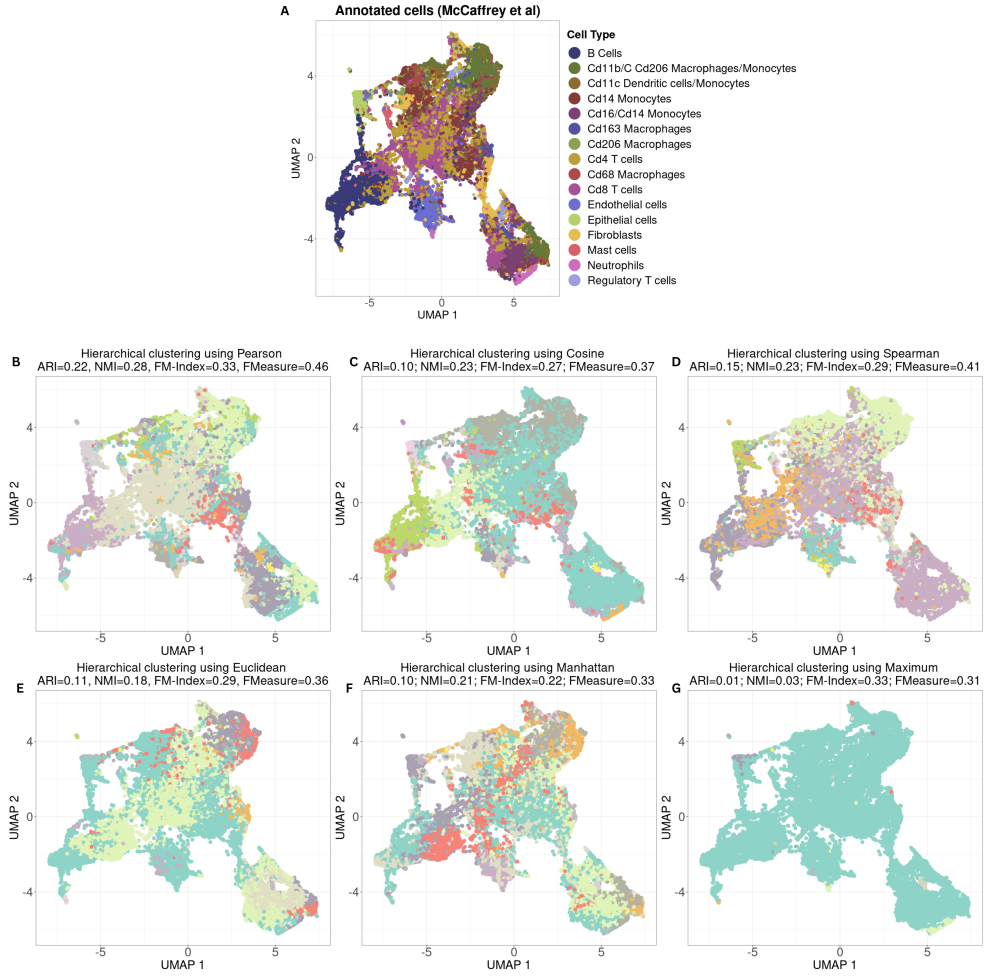

Figure 2.2: UMAP visualisation of cells from a sample imaging dataset. (A) Cells coloured by annotations from original study (McCaffrey et al., 2022). (B) Hierarchical clustering using Pearson's Correlation and concordance quantified by ARI, NMI, FM-Index, and FMeasure. (C) Hierarchical clustering using Cosine's Distance (D) Hierarchical clustering using Spearman's Correlation. (E) Hierarchical clustering using Euclidean distance. (F) Hierarchical clustering using Manhattan distance. (G) Hierarchical clustering using Maximum distance.

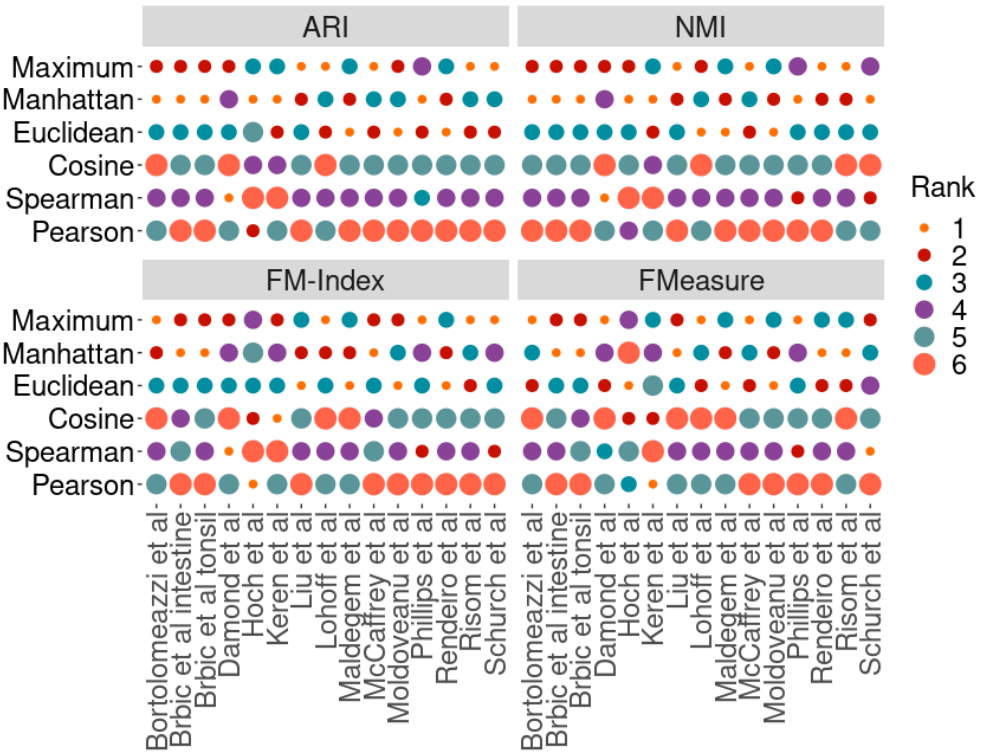

Figure 2.3: Benchmarking similarity metrics on agglomerative clustering of 15 multiplexed imaging datasets. Each dataset was subsetted to 20K cells five times, and the average clustering score was recorded. Results were ranked in descending order across all similarity metrics and datasets by each evaluation metric. A larger circle size indicates better performance. Correlation-based metrics are consistently ranked higher than distance-based metrics across most datasets.

differences in the scores between Pearson correlation compared to Euclidean distance across all evaluation metrics(Figure 2.4, Supplementary Figure S2).

2.3.2 *Combining similarity metrics is beneficial*

Given the performance differences between the similarity metrics, we next assessed whether combining multiple metrics using strategies such as multiview ensemble learning would further improve performance (Cao et al., 2020). To evaluate the efficacy of combining multiple distance metrics for clustering, we performed a comparison study combining various combinations of Pearson, Spearman, Cosine, and the Euclidean distance. All possible combinations of

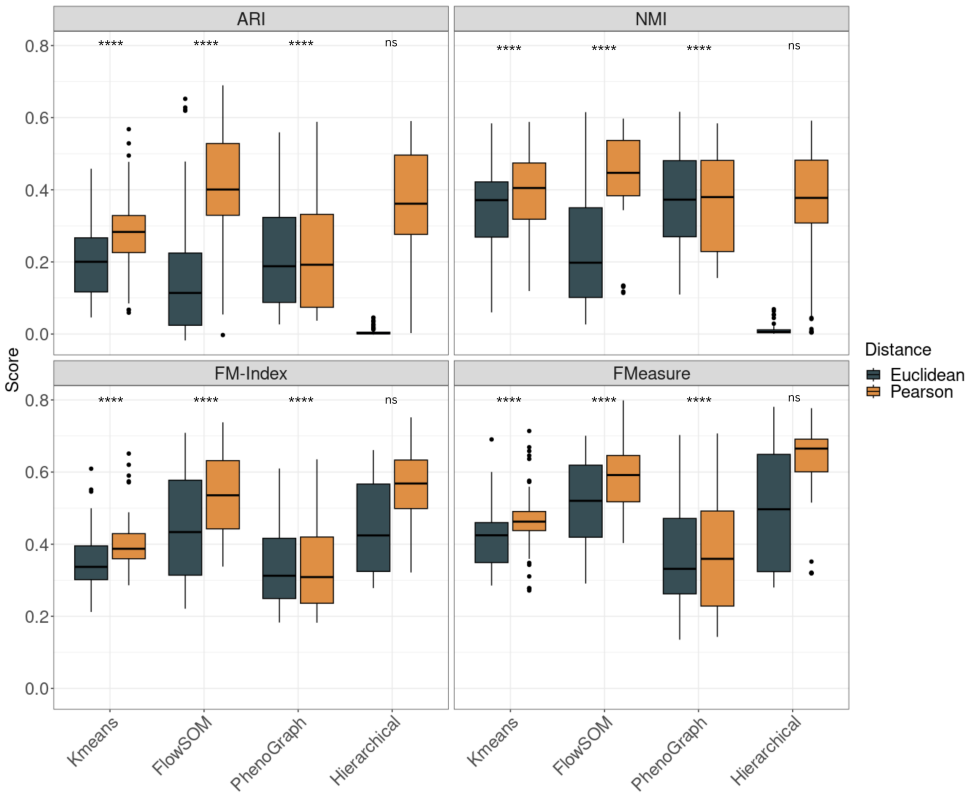

Figure 2.4: Boxplots of clustering performance of four clustering methods using Pearson correlation and Euclidean across four evaluation metrics (ARI, NMI, FM-Index, and FMeasure). For FlowSOM, Hierarchical clustering, and Kmeans, there is a statistically significant difference (****: $p < 0.0001$) in performance between Pearson and Euclidean using the Wilcoxon rank-sum test. This is not evident for PhenoGraph (ns: $p > 0.05$).

metrics were used to group the prototypes generated by the SOM algorithm and then the final scores were averaged across all datasets. The combination of Pearson, Spearman, and Euclidean, as well as the combination of all four metrics (Pearson, Spearman, Cosine, and Euclidean), consistently performed among the best across all evaluation metrics (Figure 2.5). The marginal differences between these top-performing combinations suggest that incorporating multiple correlation-based metrics provides robust clustering performance, while the addition of Cosine distance to the three-metric combination does not markedly improve overall performance.

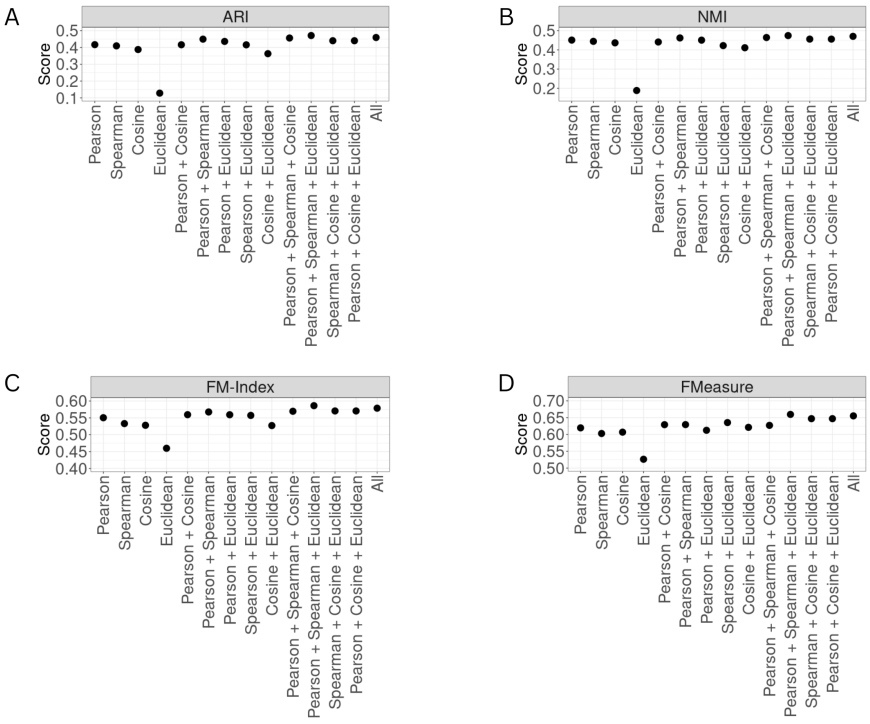

Figure 2.5: Scatter plots of average clustering performance of all combinations of Pearson, Cosine, Spearman, and Euclidean. Across all performance metrics, combinations incorporating multiple correlation-based metrics (particularly Pearson, Spearman, and Euclidean, or all four metrics) provide robust signals for clustering, with marginal differences between the top-performing combinations.

FuseSOM combines self-organising maps with multi-view ensemble learning of similarity metrics

Here, we introduce FuseSOM for the clustering of highly multiplexed imaging data. FuseSOM leverages all the ideas already discussed by combining similarity metrics with a Self Organising Map and multiview hierarchical clustering to define cell types robustly (Figure 2.1). Compared to FlowSOM, which uses Euclidean distance by default, FuseSOM has superior performance in our stratified subsampling analysis framework (Figure 2.6, $p < 0.05$, and Supplementary Figure S3) which demonstrates that a multiview ensemble of similarity metrics provides a more robust clustering. The performance gain is particularly evident

when looking at ARI and NMI, with average differences in scores being 32% for ARI, 27% for NMI, 10% for FM-Index, and 9.0% for FMeasure.

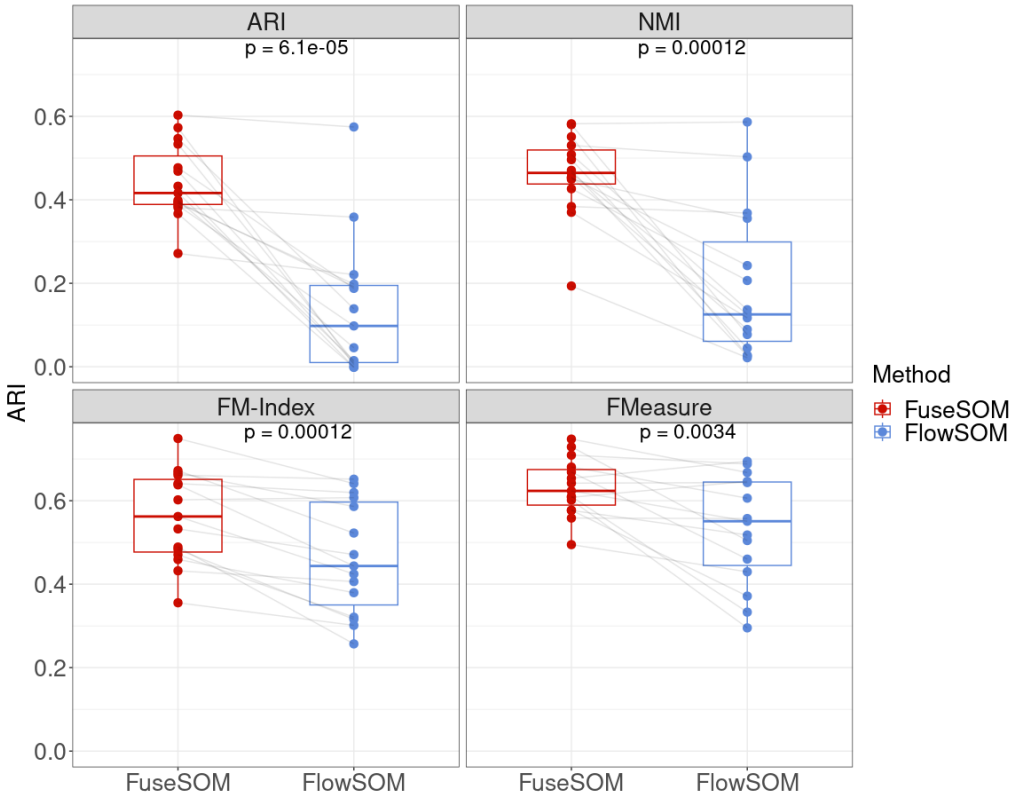

Figure 2.6: Paired boxplots for the average clustering performance across all datasets. For all four evaluation metrics, Differences in performance are evaluated using the Wilcoxon rank-sum test.

Several methods for estimating the number of clusters have been implemented in the FuseSOM R package. The relative error (RE) between the predicted number of clusters and the number of clusters used in the corresponding manuscripts was used to assess the accuracy of the cluster estimation methods. The Slope method tends to estimate cluster numbers closest to the true value, with predictions straddling zero relative error, while the Jump and Discriminant methods tend to slightly overestimate and other methods tend to underestimate the actual number of clusters (Figure 2.7A). Next, FuseSOM was used to group cells in data sets using the number of groups estimated by each method. After estimating the number of clusters and then clustering, the relative errors and

clustering scores were averaged in all data sets. When used for choosing the number of clusters, the Slope, Jump, and Discriminant methods achieve comparable performance on ARI and NMI metrics, with Slope showing strong results particularly in NMI (Figure 2.7B). While the Slope method provides the most accurate cluster number estimation, the Jump and Discriminant methods may be preferred in applications where slight overestimation is acceptable, as overestimation allows for subsequent manual merging of clusters based on domain expertise. This demonstrates the complexity in evaluating the best metric for selecting the number of clusters, as the choice depends on whether accurate estimation or downstream clustering performance is prioritised.

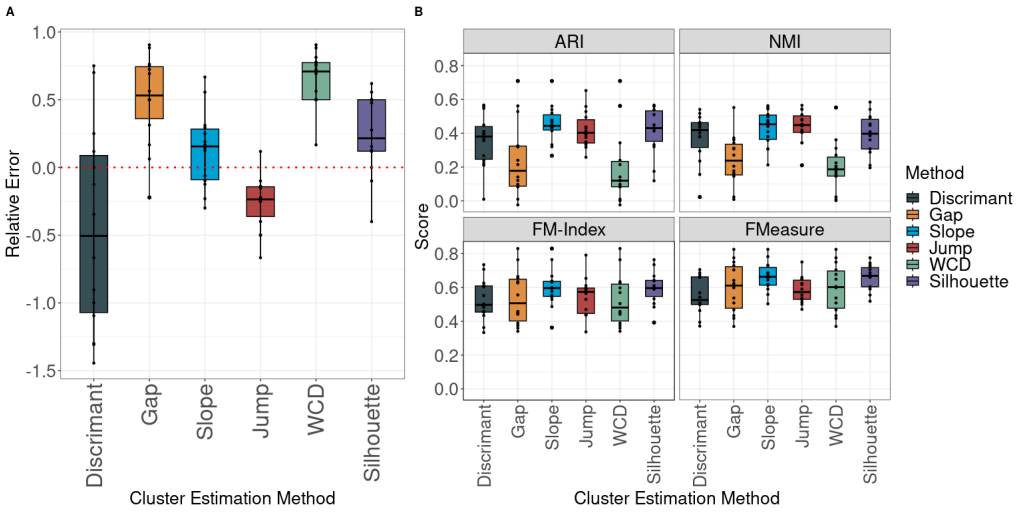

Figure 2.7: (A) Boxplots of cluster estimation performance for methods included in FuseSOM across all 15 datasets. The relative error metric was used to gauge performance. The Slope method most consistently estimates the correct number of clusters, with values straddling zero. The Jump and Discriminant methods tend to slightly overestimate. The dashed line represents a relative error of zero. Values above this line indicate an underestimation of the true number of clusters, and values below indicate an overestimation. (B) Boxplots of clustering performance based on the estimated number of clusters for methods included in FuseSOM. The Slope, Jump, and Discriminant methods show comparable performance across evaluation metrics.

To investigate the running time and memory usage of FuseSOM, we applied FuseSOM to the Hoch dataset (Hoch et al., 2022). This dataset contains 800K cells across 41 markers. Next, clustering was performed in increments of 50k cells starting from 100k cells to 400k cells (8 clustering solutions in total) to

gauge how memory and running are affected by an increasing number of cells. Clustering was performed on an 11th Gen Intel® Core™ i7-1165G7 @ 2.80GHz with four cores and 32 GB of memory. FuseSOM performance was compared against FlowSOM and PhenoGraph. In terms of running time, FuseSOM scales well and is comparable to that of FlowSOM while being faster than PhenoGraph (Supplmentary Figure S4A). For memory usage, FuseSOM is more demanding than PhenoGraph, but less demanding than FlowSOM (Supplementary Figure S4B). Doubling the size of the data requires twice as much memory for FuseSOM, while doubling the size of the data will require more than double the amount of memory for FlowSOM. The results show that FuseSOM provides a good balance between speed and memory consumption.

2.4 DISCUSSION

In this work, we performed a comparative analysis of the performance of various similarity metrics for clustering highly multiplexed *in situ* imaging cytometry assays. Using multiple clustering methods across multiple similarity metrics, we demonstrate that the choice of similarity metric affects the clustering performance of highly multiplexed cytometry *in situ* imaging data with correlation-based metrics on average outperforming distance-based metrics. We then leveraged these findings to develop a novel multi-view clustering algorithm called FuseSOM and demonstrated its ability to recover semi-supervised cell-type annotations across various datasets from differing imaging technologies with reasonable accuracy. Notably, FuseSOM showed substantial improvements on challenging datasets such as the McCaffrey MIBI-TOF tuberculosis dataset (Supplementary Figure S1), where the combination of multiple similarity metrics helped resolve cell populations that were difficult to distinguish using single metrics alone. Our results comprehensively demonstrate the impact of similarity metric choice on cell type clustering in highly multiplexed

imaging cytometry data and highlight the need to develop new best-practice clustering algorithms for these technologies.

While we have demonstrated that correlation metrics are often superior to distance metrics for multiplexed imaging data, we have not shown why this is the case. We do however demonstrate that this phenomena is consistent across the imaging platforms. Our hypothesis is that distance-based metrics such as Euclidean and Manhattan are sensitive to the scaling of the data and therefore are susceptible to changes in the expression of markers across images or even different regions of the tissue imaged. However, correlation-based metrics such as Pearson and Spearman are scale-invariant and, therefore, could be less susceptible to changes in the expression of markers driven by technical artefacts. As correlation-based metrics only consider relative expression between markers, we suspect that this makes them more robust and, therefore, more accurate in capturing cell type specific expression trends in highly multiplexed in situ imaging cytometry data. Notably, the superior performance of correlation-based metrics in imaging platforms may partly reflect their robustness to lateral spillover artefacts, as these metrics capture relative marker co-expression patterns rather than absolute intensities that are susceptible to contamination from neighbouring cells.

There are many analytical decisions and data properties that can impact the phenotyping of cells. In this manuscript, we have focused solely on the choice of distance metric used for clustering. To maintain this focus in our benchmarking study, for each dataset, we used the same cell segmentation, marker quantification, cross-image marker normalisation, marker selection, and number of clusters that were used in the original manuscripts. It should be expected that each of these components would impact the clustering of cells. We hope that our collection of data sets will assist in future benchmarks of each of these components.

Choosing the number of clusters to use when clustering remains as much an art as a science. As such, selecting a suitable number of clusters should always be viewed in the context of the application. For example, to identify rare cell types in biological data, one might need to deeply cluster the data to find smaller populations of cells. When using quantitative approaches to select the number, like those we have implemented in FuseSOM, our results highlight that some methods tend to identify more clusters on average and others less. Furthermore, while many clustering algorithms require the number of clusters to be chosen before executing the algorithm, there are others, such as graph-based and density-based methods, that can estimate the number of clusters as part of the algorithm. However, often other parameters which do need to be chosen, such as the size of the neighbourhood in density approaches, can inadvertently affect the number of clusters. Ultimately, there is no golden rule when selecting the number of clusters. Therefore, we encourage a user to employ their domain expertise and to use a variety of the methods we have implemented to arrive at a sensible choice for the number of clusters.

3

DIOSCRI ENABLES TRANSFERABLE PREDICTION OF CLINICAL OUTCOMES IN MULTI-PARAMETER CYTOMETRY DATA

Multi-parameter cytometry enables high-dimensional characterisation of immune populations at single-cell resolution. As dataset size and complexity have grown, deep learning has become an attractive option for modelling these measurements; yet practical deployment is often hindered by technical variability, batch effects, and difficulty aligning model outputs with biologically meaningful cell populations. These issues limit transfer across studies and constrain interpretability.

In this chapter, we present *dioscRi*, a transferable, interpretable framework for cytometry analysis. The approach couples a maximum mean discrepancy regularised variational encoder for normalisation and denoising with a hierarchy-aware overlapping group penalty that structures predictors by biologically or empirically derived cell type relationships. Within this setup, cell-type compositions (handled under compositional constraints) and per cell type marker summaries are modelled jointly, and the resulting coefficients map directly onto the cell-type hierarchy. The accompanying workflow provides guidance on representation learning, feature construction, and model regularisation, with diagnostics that make batch sensitivity and granularity choices explicit. Taken together, *dioscRi* offers a route to cross-dataset compatibility while maintaining the clarity needed for downstream interpretation and reporting.

This chapter is based on work shared in a preprint on *bioRxiv* (2025) (Willie et al., 2025). As first author, I led the method development in collaboration with A/Prof. Ellis Patrick and A/Prof. Helen McGuire, curated the benchmarking datasets, implemented the full analysis workflow, and developed the accompanying R package released on Github. Prof. Jean Yang, A/Prof Ellis Patrick, A/Prof. Helen McGuire, Prof. Barbara Fazekas de St Groth, Prof. Gemma Figtree, and Shreya Rao provided feedback on the final manuscript.

3.1 INTRODUCTION

Multi-parameter flow and mass cytometry technologies enable high-dimensional profiling of millions of cells in a single patient sample, a capability that can support a range of clinical applications including diagnostics, prognostics, and therapeutic monitoring (Han et al., 2015; Bailur et al., 2020; Zhang et al., 2020). These platforms can measure dozens of markers across millions of cells per sample, providing deep insights into immune states, tumour microenvironments, and signalling dynamics across diseases (Atkuri et al., 2014; Bendall et al., 2011). However, translating these insights into clinical practice requires computational methods that can scale to large datasets, remain stable in the presence of technical variability, and generalise across cohorts without retraining. To be clinically actionable, such methods must also produce outputs that are interpretable to researchers and clinicians.

Deep learning encompasses a range of methods that have shown promise for cytometry analysis, owing to their ability to model the complexity of cellular systems. Architectures such as autoencoders (AE), variational autoencoders (VAE), and convolutional neural networks (CNN) are commonly used to extract meaningful representations from high-dimensional data while remaining relatively robust to noise and variability (Hu et al., 2022). For example, MoE-SimVAE (Kopf et al., 2021) combines VAEs with Gaussian mixture models to enable

unsupervised identification of cell types, implicitly separating biological variation from technical artefacts. DGCyTOF (Cheng et al., 2022) and DeepCNN (Hu et al., 2020a) apply autoencoders for supervised cell-type classification, with DeepCNN incorporating a calibration step to address batch effects across datasets. However, such calibration requires prior knowledge of batch structure, limiting its applicability in clinical settings where classification of a sample should be done independently of other samples. For patient classification, models such as CellCNN (Arvaniti and Claassen, 2017a) and DeepCNN use CNNs to learn complex relationships directly from single-cell measurements. Recognising the importance of interpretability, DeepCNN also employs post-hoc permutation tests and decision trees to identify marker-level associations. Because these interpretability steps are applied after model training, they do not constrain the model during learning and may not reflect the true decision boundaries. As a result, existing approaches are not designed to generalise to unseen datasets and typically lack mechanisms for structured interpretation, making them difficult to apply confidently across cohorts or translate into clinical insight.

A longstanding solution to the challenges of interpretability and technical variability in cytometry has been the use of cell-type hierarchies (Herzenberg et al., 2006; Liu et al., 2024). These hierarchies are typically constructed through manual gating, a conventional technique in fluorescence and mass cytometry that defines cell populations by drawing boundaries on a series of two-dimensional marker expression plots which, taken together, encompass a complex phenotype including expression of multiple markers. Analysts begin by identifying broad immune categories, such as $CD4$ and CD8 T cells, NK cells and B cells, and progressively refine these into more specific subsets such as regulatory versus conventional $CD4$ T cells, naïve versus memory conventional $CD4$ T cells and so on. This hierarchical structure not only facilitates interpretation but also provides robustness to technical variation, as comparisons between child and parent populations help normalise differences across samples. Manual gat-

ing remains widely used because it reliably identifies cell subpopulations with known biological functions while effectively accounting for technical variability. However, it requires considerable expertise, and is very time-consuming, especially when sample-to-sample variation from biological or technical sources requires that multiple gates be adjusted for each sample. Hence the search for computational methods that can replicate and extend the complex cell subset structures within data derived from increasingly high-plex technologies (Maecker et al., 2012; Brestoff and Frater, 2022). Recent approaches aim to connect cell populations to biological or clinical outcomes by leveraging hierarchical relationships in ways that preserve interpretability and improve scalability (Chan et al., 2021; Verschoor et al., 2015) Embedding hierarchical relationships into deep learning models could improve their resilience to technical variation while preserving biologically meaningful structure.

To address the need for models that are both interpretable and transferable across batches and cohorts, we present dioscRi, a deep learning framework for predicting clinical outcomes from multi-parameter cytometry data. DioscRi combines a Maximum Mean Discrepancy variational autoencoder (MMD-VAE) to learn a transferable normalisation scheme, with an overlapping group LASSO model that incorporates biologically or empirically derived cell-type hierarchies. This design enables the model to generalise across datasets without requiring batch labels, while producing structured outputs that reflect known immune relationships. We demonstrate the utility of dioscRi through its application to a coronary artery disease cohort, where it recapitulates a subset of previously published immune associations (Kott et al., 2023). We further benchmark its performance across multiple public datasets, showing that dioscRi consistently outperforms existing deep learning approaches. Together, these results establish dioscRi as a robust and generalisable framework for clinically focused cytometry analysis.

3.2 METHODS

3.2.1 *Overview of dioscRi*

dioscRi consists of various steps: (i) Transferable data normalisation and denoising using a Maximum Mean Discrepancy Variational Autoencoder (MMD-VAE), (ii) Generation of cell types through unsupervised clustering, (iii) Extraction of cell type and sample-level features, (iv) Hierarchical grouping of cell type-level features, (v) Clinical outcome prediction using overlap group lasso, and (vi) Model visualisation and interpretation.

3.2.1.1 *Data pre-processing*

For all datasets, we applied an *arcsinh* transform ($y = \mathrm{arcsinh}(x/5)$). Following transformation, each data set was sub-sampled to 10,000 cells from each sample and combined for analysis.

Before training the MMD-VAE model, a Min-Max scaling to $(0, 1)$ was applied.

$$y = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{15}$$

Min-max scaling is commonly employed to improve the training efficiency and convergence of deep learning models.

3.2.1.2 *Maximum mean discrepancy variation auto encoder architecture*

The dioscRi architecture integrates an encoder and decoder within a Variational Autoencoder (VAE) framework. It is enhanced by the Maximum Mean Discrepancy (MMD) to better match the latent space distribution to the prior, improving adaptability to non-Gaussian data distributions (Kingma and Welling, 2013; Zhao et al., 2017). The Maximum Mean Discrepancy (MMD) penalty was implemented to enforce distributional similarity between the encoded latent variables

and a prior distribution, following the method described in (Tolstikhin et al., 2017). An unbiased U-statistic estimator of the MMD was computed using an Inverse Multi-Quadratic (IMQ) kernel. We summed IMQ kernels computed at varying scales to capture multi-scale relationships. The encoder compresses the input data into a lower-dimensional representation, thereby retaining only the most pertinent information by filtering out noise. The decoder then reconstructs this encoded data to its original form by minimising the difference between original input and the decoded output.

The model can be formulated as:

$$z = f_{\text{encoder}}(x) \tag{16}$$

$$x' = f_{\text{decoder}}(z) \tag{17}$$

$$\mathcal{L}_{\text{BCE}} = -\mathbb{E}\left[x \log(x') + (1-x)\log(1-x')\right] \tag{18}$$

$$\mathcal{L}_{\text{MMD}} = \mathbb{E}_{z,z'\sim q_z}[k(z,z')] + \mathbb{E}_{z,z'\sim p_z}[k(z,z')] - 2\mathbb{E}_{z\sim q_z, z'\sim p_z}[k(z,z')], \tag{19}$$

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{MMD}} \tag{20}$$

Here, f_{encoder} is the encoder function that compresses the input data x into a lower-dimensional latent representation z. The decoder function f_{decoder} reconstructs the input data from z back to its original form x'. The Binary Cross Entropy loss $\mathcal{L}_{\text{BCE}}$ measures the difference between the input and the reconstructed output. For the MMD loss, $k(\cdot,\cdot)$ is the IMQ kernel. This penalty en-

courages the encoded latent representation to align with the desired prior distribution while allowing flexibility for modelling complex data distributions. The overall loss $\mathcal{L}_{\text{VAE}}$ combines the Binary Cross-Entropy loss $\mathcal{L}_{\text{BCE}}$ with the MMD regularisation term $\mathcal{L}_{\text{MMD}}$, balanced by a hyper-parameter λ that controls the influence of the MMD term.

3.2.1.3 *MMD-VAE implementation*

The model architecture includes an encoder that reduces the dimensionality of the input data through dense layers with Rectified Linear Unit (ReLU) and Gaussian Error Linear unit (GELU) activations, followed by a decoder that reconstructs the input using sigmoid activation. The training objective optimises a total loss, combining binary cross-entropy for reconstruction and Maximum Mean Discrepancy (MMD) to align the latent space distribution with a standard normal prior.

The MMD-VAE model is implemented in R using the Keras library (Kalinowski et al., 2024), with RMSprop (Tieleman and Hinton, 2012) as the optimiser and a default learning rate of 0.001. The MMD-VAE is trained for 100 epochs with a batch size of 32, and the latent dimension is conservatively set to balance information retention. The model is validated using a separate dataset during training. The MMD-VAE comprises three hidden layers with layer sizes differing for each data depending on the number of features. Table S1 lists the hyper-parameters used for each dataset.

3.2.1.4 *Selection of representative sample for MMD-VAE training*

To reduce the total size of the training data while retaining core biological information, we select representative samples for training the MMD-VAE model. We ensure the model captures key biological variations without redundancy by selecting samples with the most representative characteristics. To achieve this, we use the method described in (Li et al., 2017), where each sample is evaluated based on its average covariance Frobenius norm between it and other samples.

The Frobenius norm quantifies the difference between the covariance matrices of pairs of samples, effectively measuring how similar or different each sample is to others. Formally, the Frobenius norm between samples i and j is defined as:

$$FN_{ij} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(Cov_i - Cov_j)^2}, \tag{21}$$

where Cov_i and Cov_j are the covariance matrices for samples i and j respectively. The stable sample is selected by returning the sample that satisfies:

$$\arg\min_i \left(\frac{1}{n} \sum_{j=1}^{n} (FN_{ij}) \right) \tag{22}$$

Using this criterion, we select the top K samples with the minimum average Frobenius norms as the set of reference samples, with K set to 2 in all experiments.

3.2.1.5 *Unsupervised clustering and Cell type Classification*

To generate cell types from the normalised training data, we use the FuseSOM algorithm (Willie et al., 2023), which combines multiple similarity metrics through multiview ensemble learning and hierarchical clustering to create cell types.

For the testing data, cell types were generated by training a classifier on the cell types identified in the training data. A Linear Discriminant Analysis (LDA) model was used for all cell type classifications. The embeddings from the VAE model served as features for unsupervised clustering and manual gating. Cell type classification models were fit using the *Caret* R package.

3.2.1.6 *Feature Engineering*

After cell type identification, we generate a set of features, including the logit of the cluster proportions in each sample and the mean marker expression per cell

type in each sample. The logit proportion for a cluster k in sample i is defined as:

$$\pi_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \log\left(\frac{\frac{c_{ij}}{k}}{1-\frac{c_{ij}}{k}}\right), \tag{23}$$

where n is the number of samples, k is the number of cells in each sample, m is the number of clusters, and c_{ij} is the count of cells in sample i for $i = 1, 2, ..., n$ that belong to cluster j for $j = 1, 2, ..., m$.

The mean marker expression per cell type across all samples is defined as:

$$\bar{x}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} x_{ijk}, \tag{24}$$

where x_{ijk} is the marker expression value of the k^{th} cell belonging to cell-type j and sample i. The valuen_{ijk} is the total number of cells of belonging to cell-type j in sample i.

3.2.1.7 *Incorporating empirically or biologically informed grouping structure*

Before classifying samples, we integrate cell type hierarchies into the proportions, using either biologically informed or empirically derived groupings. To empirically derive groupings we construct a hierarchical clustering tree based on Pearson correlation distance and Ward linkage, generating all possible groupings by cutting the tree at various heights. Alternatively, biologically derived hierarchies are predefined using prior knowledge of cell-type relationships by respective domain experts. These hierarchies, whether empirically or biologically informed, are combined with the marker means for each cell type to create the final set of features used for classification.

3.2.1.8 *Group logistic lasso with overlapping groups*

The Lasso (Least Absolute Shrinkage and Selection Operator) is a regularisation method used for feature selection and model regularisation (Tibshirani, 1996).

Given a matrix X of predictors with n samples and p features, and a binary response vector y, the Lasso minimises the following objective function:

$$Q(\beta|X,y) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}\left\|\beta_j\right\|_1 \tag{25}$$

where β represents the coefficient vector, $\sum_{j=1}^{p}\left\|\beta_j\right\|_1$ is the ℓ_1-norm penalty, and λ is a regularisation parameter that controls the level of sparsity in the solution. By penalising the absolute values of the coefficients, Lasso encourages sparsity, driving some coefficients to zero and excluding irrelevant features. This approach is particularly effective in high-dimensional settings, where it reduces overfitting and yields interpretable models.

The Group Lasso extends this framework by promoting sparsity at the group level, enabling feature selection based on predefined groups of predictors (Yuan and Lin, 2005). Assume that the predictors are partitioned into J non-overlapping groups $G_1, G_2, \ldots, G_J$, with each group containing a subset of features. The Group Lasso objective is:

$$Q(\beta|X,y) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{J}\sqrt{K^j}\left\|\beta_{G_j}\right\|_2 \tag{26}$$

where $\left\|\beta_{G_j}\right\|_2$ denotes the ℓ_2-norm of the coefficients within group G_j, and K^j is the number of elements in group j.

We adapt the Group Lasso with a logistic loss function for binary outcomes to handle classification. Given a binary response vector y with entries $y_i \in \{0,1\}$, the logistic loss is defined as:

$$\mathcal{L}(\beta|X,y) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i\log(p_i) + (1-y_i)\log(1-p_i)\right] \tag{27}$$

where

$$p_i = \frac{1}{1+\exp(-X_i\beta)} \tag{28}$$

represents the probability of $y_i = 1$ given X_i. The Group Lasso with logistic loss thus becomes:

$$Q(\beta|X, y) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] + \lambda \sum_{j=1}^{J} \sqrt{K^j} \left\|\beta_{G_j}\right\|_2$$

This formulation encourages sparsity in the group structure, making it well-suited for binary classification tasks involving predefined groups of features.

The Overlapping Group Lasso further generalises this approach by allowing predictors to belong to multiple groups (Zeng and Breheny, 2016). This extension is particularly useful in cases where variables naturally fall into overlapping categories, such as biological pathways where genes may belong to multiple pathways, or immune cell features that are shared across related cell types within a hierarchy. To accommodate overlapping groups, we introduce a latent structure by decomposing the coefficient vector β into group-specific latent vectors γ_j, each associated with a group G_j. Let $\tilde{X}$ be the expanded design matrix with duplicated columns for overlapping variables. The objective for the Overlapping Group Lasso is:

$$Q(\gamma|\tilde{X}, y) = L(y, \tilde{X}\gamma) + \lambda \sum_{j=1}^{J} \sqrt{K^j} \left\|\gamma_j\right\|_2 \tag{29}$$

where $L(y, \tilde{X}\gamma)$ denotes the logistic loss, $\gamma = (\gamma_1, \ldots, \gamma_J)$ represents the concatenated vector of group-specific coefficients, and $\|\gamma_j\|_2$ applies the ℓ_2-norm penalty within each group. The overlapping group structure enables shared predictors to have multiple group-specific effects, enhancing model flexibility. After fitting the model, the original coefficients β are reconstructed as:

$$\beta = \sum_{j=1}^{J} \gamma_j \tag{30}$$

To further regularise overlapping cases, we add a ridge penalty term, defining the final objective as:

$$Q(\gamma|\tilde{X}, y) = L(y, \tilde{X}\gamma) + \lambda \sum_{j=1}^{J} \sqrt{K^j} \left\|\gamma_j\right\|_2 + \frac{\lambda_2}{2} \|\beta\|_2^2 \tag{31}$$

where $\lambda_1 = \alpha\lambda$ and $\lambda_2 = (1-\alpha)\lambda$. To select the optimal α and λ, we fit models with different α values over the range [0,1], λ values over the range [0.05, 1] and compute the deviance criterion for each model. The best α and λ values are chosen using the elbow method applied to the deviances. We implemented the overlapping group lasso using the grpregOverlap R package (Zeng and Breheny, 2016), modified to support logistic loss for binary classification. This approach provides a flexible and interpretable solution for feature selection when variables belong to multiple groups.

In the overlapping group lasso model, each coefficient vector γ_j represents the contribution of a specific group G_j to the overall model, allowing for a unique interpretation of overlapping features. Since $\beta = \sum_{j=1}^{J} \gamma_j$, the original coefficient vector β is constructed by summing the group-specific latent effects γ_j. This decomposition means that each predictor can be influenced by multiple groups simultaneously, with each γ_j capturing the effect of group G_j independently. Consequently, the magnitude and direction of γ_j can be interpreted as the specific impact of group G_j on the predictor, highlighting how particular groupings contribute to the model's predictive power. When interpreting the model, non-zero γ_j values indicate that the group G_j plays a significant role, while the summed γ coefficients in β provide the overall effect of each predictor. This approach improves interpretability by revealing which features are selected and which groups drive their inclusion, offering insights into overlapping pathways or shared structures across groups.

3.2.2 *Implementation of other approaches for benchmarking comparisons*

3.2.2.1 *Training of cyCombine and iMUBAC*

We benchmarked the performance of the normalisation performance of dioscRi against other normalisation methods that explicitly consider batch information. These methods included cyCombine (Pedersen et al., 2022) and iMUBAC (Ogishi et al., 2021). For analysis, both cyCombine and iMUBAC were run with default parameters.

3.2.2.2 *Training of CellCNN and DeepCNN*

The predictive performance of dioscRi was benchmarked against CellCNN (Arvaniti and Claassen, 2017a) and DeepCNN (Hu et al., 2020a). Since both methods work directly on FCS files, after splitting each dataset into training and testing, the resulting cells were written back to FCS files for predictive modelling using these methods.

CellCNN was run using an optimal set of hyper-parameters (filter number = 5, l_2 coefficient = $1e-4$). The Adam optimiser was used for loss optimisation, with a learning rate of 0.01 over 500 epochs. After training, the model was evaluated using the testing set.

DeepCNN was run using the optimal set of hyper-parameters used in the original paper (number of filters = 3, dropout rate equals 0.2). The Adam optimiser was used for loss optimisation, with a learning rate of 0.01 over 500 epochs. After training, the model is evaluated using the testing set.

3.2.3 *Model evaluation*

Various evaluation metrics were used to evaluate components of dioscRi.

3.2.3.1 *Evaluation of normalisation by clustering*

After data normalisation, the Adjusted Rand Index (ARI) was used to assess clustering performance. The ARI metric assesses the overlap between two sets and takes on values ranging from 0 to 1, with 0 indicating no overlap and 1 indicating perfect overlap. The ARI score can be defined as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}} \tag{32}$$

where n_{ij} is the number of pairs of elements that are in the same set in both clusterings, a_i is the total number of pairs in the same set for the first clustering, b_j is the total number of pairs in the same set for the second clustering, and n is the total number of elements.
Similarly, the NMI score can be defined as:

$$NMI(X, Y) = \frac{2 * I(X, Y)}{H(X) + H(Y)} \tag{33}$$

where $I(X, Y)$ is the mutual information between clusters X and Y, and $H(X)$ and $H(Y)$ are the entropies of clusters X and Y respectively.

3.2.3.2 *Evaluation of outcome prediction*

The ROC-AUC metric was used to assess the performance of dioscRi and competing methods for predicting clinical outcomes. The metrics are defined using the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.

The ROC-AUC measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC score is a single scalar that

summarises the performance across all possible classification thresholds, with 1 indicating perfect classification and a value of 0.5 indicating no better than chance. The formula for TPR and FPR are:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

3.2.4 *dioscRi Framework Overview*

Here we present dioscRi, an interpretable framework for analysing multi-parameter cytometry data. This framework integrates an MMD-VAE with a hierarchical group LASSO tailored for the prediction of clinical outcomes (Figure 3.1). The input to the framework consists of raw gene or protein (markers) expression data of all samples, where rows represent individual cells and columns correspond to markers. The model output can include normalised expression data, unsupervised cell clusters, sample-level clinical outcome prediction, and the associations of cell-type proportions and/or markers with clinical outcomes. The MMD-VAE provides a transferable normalisation scheme, ensuring it can be applied to unseen data. The overlapping group LASSO uses a biologically or empirically derived grouping of cell-type proportions and marker means per cell-type to accurately predict clinical outcomes and uncover associations. The model outputs a tree structure that can assess the associations of these proportions and marker means within cell-types with predictions. In addition, the framework allows for the inclusion of additional clinical information, such as age and gender, which can enhance the precision of clinical outcome predictions.

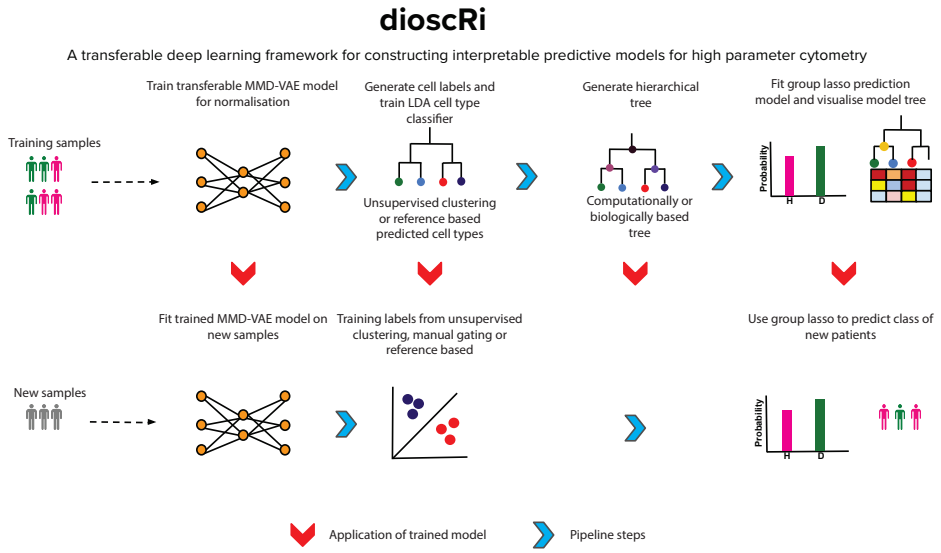

Figure 3.1: Overview of dioscRi: this new framework uses: (I) a robust MMD-VAE to train a normalisation filter scheme that can applied to new data; (ii) linear discriminant analysis trained using either manually gated cell populations or unsupervised clustering to classify cell types; (iii) a hierarchical group LASSO that uses either a biologically-defined or computationally derived tree to combine predicted cell type or cluster proportions and marker means in each sample to predict clinical outcomes; and (iv) a visualisation of the associations between cluster or predicted cell type features and clinical predictions using heatmaps and the tree structures.

3.2.5 *Datasets*

We employed a variety of datasets across different technologies and diseases. Brief descriptions of each dataset are provided below, with more details outlined in Table S2.

3.2.5.1 *BioHEART-CT*

To evaluate the dioscRi framework, we used mass cytometry data generated from the BioHEART-CT study (Kott et al., 2019, 2023). This prospective longitudinal cohort study to identify immune cell populations associated with coronary artery disease (CAD). Cell profiling was performed using mass cytometry with 41 tagged antibodies, enabling a detailed analysis of cellular characteristics. The major clinical outcome used in the previously published analysis (Kott

et al., 2023) was the Gensini score, a continuous measure of coronary artery disease (CAD) severity. For prediction tasks, we binarised this outcome: samples with a Gensini score of 0 were labelled as CAD-, and those with a score greater than 0 were labelled as CAD+. The original study (Kott et al., 2023) included discovery and validation cohorts that underwent expert manual gating to identify more than 100 subpopulations. The discovery cohort of 111 samples across 6 batches comprising a total of 41,350,484 cells was then analysed using a logistic regression model to identify 18 subpopulations whose size was identified as most different between CAD+ and CAD-. The sizes of the same 18 subpopulations were manually gated in the validation cohort consisting of 58 samples across 3 batches. Importantly, the validation cohort was acquired using a mass cytometry platform that introduced significant time-dependent changes in multiple signals (Lee et al., 2019), so the expert gating included preliminary time gates to exclude artefactual data. For the current study, the discovery cohort was downsampled to include 1,106, 443 cells. The validation cohort was downsampled to 580,000 cells without applying any time gates. For the discovery cohort, 11 manually gated cell populations, identified as detailed in Supplementary Figure 4 of the original manually gated analysis (Kott et al., 2023) were used as the basis for cell type prediction. The 11 populations were conventional CD4+ T cells (Tconv, mean 38%), CD4+ regulatory T cells (Treg, mean 1.9%), CD8+ T cells (CD8hi mean 17%), CD8lo-neg T cells (CD8lo mean 6.2%), B cells (mean 10%), NK cells (mean 10.2%), CD14+ monocytes (mean 12%), CD16+ monocytes (mean 2.1%), plasmacytoid dendritic cells (pDC mean 0.36%), CD1c+ dendritic cells (CD1c DC mean 0.65%) and CD141+ dendritic cells (CD141 DC mean 0.044%). Of the 41 metal-tagged antibodies used to assess marker expression in the original mass cytometry data, 27 were included as markers in the discovery and validation datasets.

3.2.5.2 *Breast cancer tumour*

This dataset comprises 144 human breast tumour samples and 50 non-tumour tissue samples measured with mass cytometry(Wagner et al., 2019). The primary objective was to characterise the features of breast cancer ecosystems and their correlations with clinical data. The study evaluated 73 proteins in approximately 26 million cells, utilising tumour-focused and immune cell-centric antibody panels. This extensive profiling helps understand the protein expression patterns and cellular interactions within breast cancer and adjacent non-cancerous tissues. We analysed 194 samples, predicting each as tumour vs. non-tumour. The dataset was split into 70% training and 30% testing.

3.2.5.3 *Cytomegalovirus*

This dataset encompasses data from nine human immunology studies (SDY112, SDY113, SDY305, SDY311, SDY315, SDY472, SDY478, SDY515, and SDY519) featuring 596 peripheral blood mononuclear cell (PBMC) samples from 313 subjects across 472 samples (Bhattacharya et al., 2018; Kronstad et al., 2018; Miron et al., 2018; Alpert et al., 2019). Diagnosing latent Cytomegalovirus (CMV) is challenging since it is usually asymptomatic with limited observable effect on the immune system. These studies aimed to identify immune responses to latent CMV infections. Our analysis of this dataset followed that used in the DeepCNN manuscript (Hu et al., 2020a), which used studies SDY519 and SDY515 as testing and validating datasets and the rest for model training. Clinical response is individuals having a positive or negative cytomegalovirus (CMV) infection. A set of 23 markers that were common to all nine studies.

3.2.5.4 *COVID-19 PBMC CD8+ non-naive T cells*

This dataset is derived from a high-dimensional flow mass cytometry analysis of COVID-19 samples, alongside comparisons with recovered individuals and healthy controls (Mathew et al., 2020). The study involved an integrated anal-

ysis of approximately 200 immune features. These immunological data were combined with about 50 clinical features, providing a comprehensive overview that facilitates understanding of the relationships between the immunology of SARS-CoV-2 infection and various clinical outcomes, including disease patterns, severity, and progression. We restricted analysis to CD8+ non-naive T cells and used COVID-19 recovery as the clinical response of interest. The dataset was randomly split into 70% training and 30% testing sets.

3.3 RESULTS

3.3.1 *dioscRi effectively normalises unseen Samples*

To assess the effectiveness of dioscRi's normalisation, we applied it to a mass cytometry study of coronary artery disease and evaluated its ability to reduce technical variability across samples. A discovery cohort of 111 samples was used to train the normalisation Variational Autoencoder, which was then applied to a validation cohort of 58 samples. Once trained, the model was applied consistently across both cohorts to ensure a unified normalisation strategy. The impact of normalisation was first evaluated by examining co-expression patterns using biaxial scatter plots of CD3 and HLA-DR for samples 87 and 88 (Figures 3.2A–D). In the raw data, broad and overlapping distributions indicated sample-to-sample inconsistency. Following normalisation, these distributions became more compact and well separated, supporting improved consistency in population structure and enhancing comparability across samples for downstream analyses. Next, density plots of CD3 (Figures 3.2E and 3.2F) revealed substantial improvement following normalisation. In the raw data (Figure 3.2E), distributions varied notably between samples suggesting technical variability. After applying dioscRi (Figure 3.2F), marker distributions became more closely aligned, reducing non-biological variation and preserving meaningful signal.

3.3.2 *dioscRi normalisation reduces variance and improves cell annotation*

We next assessed the effect of normalisation across all markers. We first computed the R^2 from linear models of marker expression against predicted cell type (Figure 3.2G). DioscRi consistently produced the highest R^2 values relative to raw data and other methods, indicating effective noise reduction while preserving biological signal. Next, comparisons of relative coefficient of variation (CV) (Figure 3.2H) showed that dioscRi modestly reduced variability (median = 0.92) while maintaining a tight distribution across markers; cyCombine showed no change (median = 1.00) and iMUBAC achieved the largest reduction (median = 0.45), which when interpreted in the context of the R^2 results suggest this is driven by loss of meaningful biological heterogeneity. We then confirmed this balance at a more granular level by performing marker-level analysis of variability and signal intensity across samples (Supplementary Figure S1). In the raw data, markers such as CD3, CD4, and HLA-DR exhibit higher variance, consistent with both biological reality and potential technical artefacts. In combination, these results indicate that dioscRi removes sufficient technical noise to stabilise CV without over-compressing true variation.

To assess the impact of normalisation on cell type identification, we then performed clustering on the data that had been normalised by the different approaches. Models were also trained with either unnormalised or normalised data using the 11 manually gated cell populations in the discovery cohort to predict cell types in the discovery and validation cohorts. We assessed the concordance of the clustering and predicted cell types using the adjusted Rand index (ARI; Figure 3.2I). DioscRi achieved an ARI of 0.83, outperforming cyCombine (ARI = 0.62), iMUBAC (ARI = 0.65), and the raw data (ARI = 0.62). The Rand index value of 0.83 for dioscRi indicates that the gated and clustered cell types would have biologically significant differences. However, in comparison to cyCombine and iMUBAC, the results underscore dioscRi's capacity to

preserve biologically meaningful structure even without explicit batch labels by producing clustering results that align with known cell-type assignments.

3.3.3 *dioscRi normalisation improves patient classification*

To evaluate the impact of normalisation on predictive performance, we applied dioscRi to a coronary artery disease dataset and compared outcomes using normalised and unnormalised data. Classifiers to predict coronary artery disease (CAD) were constructed that used predicted cell type proportions and marker means as input features, as well as the biologically-informed hierarchy of the manually gated cell populations. After training models on the discovery cohort, normalisation led to substantial improvements in predictive performance with AUC in the validation cohort improving from 0.68 to 0.79 (Figure 3.2J). To demonstrate that the normalisation performance benefits of dioscRi are not solely dependent on expertly derived hierarchical relationships, we then derived 11 cell populations from unsupervised clustering with FuseSOM (Willie et al., 2023) and an empirical hierarchical tree as described in the methods. For this clustering-based model, normalisation increased the AUC from 0.67 to 0.83 (Figure 3.2K). Finally, we compared the benefit of dioscRi's transferable normalisation with the two established batch correction methods, cyCombine and iMUBAC. Both cyCombine and iMUBAC used information from the discovery and validation sets for normalisation, while dioscRi only used the discovery data to train its normalisation. Following normalisation with each method, the resulting data were processed through the full dioscRi pipeline, and predictive performance was assessed using AUC scores for CAD status. Our dioscRi (AUC = 0.79) outperformed cyCombine (AUC = 0.63) and iMUBAC (AUC = 0.55) (Figure 3.2K). In combination, these results underscore the strength of dioscRi's normalisation strategy, which outperforms alternatives in downstream subject level classification without requiring explicit batch labels.

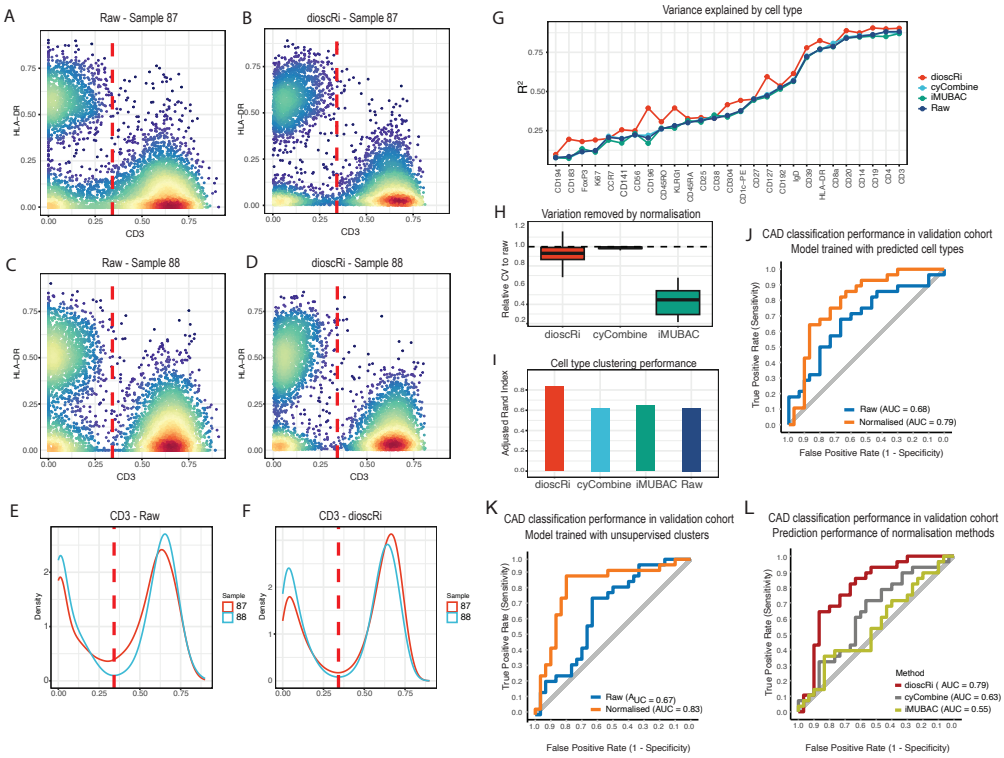

Figure 3.2: (A-D) Biaxial scatter and (E-F) density plots illustrate the effect of normalisation on marker expression and gating consistency in samples 87 (validation cohort) and 88 (discovery). In the raw data (A, C, E), marker distributions are broad, clusters are poorly defined, and CD3 alignment is inconsistent, particularly in the validation cohort sample. After dioscRi normalisation (B, D, F), distributions become tighter, cell populations separate more clearly, and CD3 expression aligns across samples. Red dashed lines indicate gating regions for specific subsets, highlighting reduced variability and preservation of biologically relevant features post-normalisation. (G) R2 values from the relationship between predicted cell types and each marker for three normalisation strategies and the raw data. (H) Coefficients of variation for each marker before and after normalisation. (I) Adjusted Rand Index is used to evaluate if clustering on the different normalised datasets recovers populations similar to the manually gated cell types. Finally, AUC of models classifying CAD using dioscRi improves patient classification in (J) models trained using cell types predicted from manually gating and (K) models trained with clustered cell populations. (L) Comparison of CAD classification reveals a higher AUC for dioscRi normalisation compared with cyCombine or iMUBAC normalisation.

3.3.4 *Cell type and marker associations provide complementary biological insights*

To assess the interpretability of dioscRi, we first examined the coefficients from the overlapping group LASSO model using different feature sets of predicted cell types from models trained on expert gating. Fitting separate models us-

ing only proportions or only marker means provided complementary views: the proportions model identified differential abundance of B cells, pDCs, and CD4+ T cells (Figure 3.3A), of which only pDCs were present differentially in the original gated data (Kott et al., 2023) while marker means highlighted expression-level differences within other cell populations, including increased CD192 and CD194 in Tregs (Fig 3.3B). However, the majority of the differences highlighted in the marker means analysis were not present in the original gated data (Supplementary Figure S2A). For example, CD4+ Treg cells do not express CD14, and additionally CD141+ DCs do not express KLRG1. The combined model revealed some associations consistent with prior findings in coronary artery disease (CAD) (Kott et al., 2023), with increases in CD4+ Treg expression of CD194, CD45RO and Ki67 (Figure 3.3C). Conventional CD4+ T cells (Tconvs) were negatively associated with CAD through markers such as CD3, CD4, and CD127 (Figure 3.3C). This reflects the previously documented changes in Tconv (Kott et al., 2023), since the expression of CD3, CD4 and CD127 differs significantly between naive T conv (reduced in CAD) and effector memory Tconv (increased in CAD). If Tconv had been divided into two populations (naïve and memory) in the original 11 cell type designation, then additional cell type proportions rather than marker means would likely have been identified by the model. This example illustrates how the marker means can, to some extent, compensate for unbalanced cell type choices in the accuracy of the subject-level prediction, although they require more subsequent analysis to determine the true cell types that differ between CAD+ and CAD- patients.

The prediction of CAD from models trained using manually gated cell types is likely limited by the ability of dioscRi to predict the hierarchically defined cell types. Heatmaps of the original gated cell types (Supplementary Figure S3A) and the predicted cell types for the discovery and validation cohorts (Supplementary Figures S3B-C) illustrate these differences. For example, in both the discovery and validation cohorts, expression of CD14, CD183, CD194, CD38, CD39 and CD56 in the three small DC populations is different from that in the

manually gated data. The concordance between the size and marker expression of the original manual gated cell types and the predicted cell types is highest for the major cell types such as CD4+ Tconv and CD8hi T cells. Initial choice of 11 gated cell types that were more evenly balanced in size would likely have improved the accuracy of the cell type prediction.

We next trained models that used unsupervised clustering in the discovery cohort to define cell types. When applied to the models built using the unsupervised clusters, dioscRi uncovered additional associations not evident from the annotation-guided analysis (Supplementary Figure S4C). Cluster 7, which had high expression of CD19 and CD20 but also expressed CD1c, CD3, CD4, CD56 and KLRG1 and therefore may comprise a mixture of B cells, DCs and multiple other cell types (Supplementary Figure S4D) showed increased CD25 expression (Supplementary Figure S4B) in CAD+ samples. It is worth noting that Cluster 7 shows elevated signal across many markers (Supplementary Figure S4D), which may indicate this cluster represents a technical artefact (such as cells with elevated autofluorescence or debris) rather than a distinct biological population. Users should interpret clusters with uniformly elevated marker expression with caution. It is not clear what this represents in biological terms, since cluster 5, whose pattern of marker expression corresponded much better with a pure population of B cells, did not show the same effect of CD25 expression on model performance. Clusters 1 and 4 were characterised by expression of CD56, a marker of NK cells, although their overall pattern of marker expression did not correspond with that of either gated or predicted NK cells (Supplementary Figure S3). Cluster 4 contributes additional information from 3 markers, but the biological relevance is unclear, since NK cells do not express CD14, while expression of KLRG1 is generally low. This pattern of large apparent fold-changes in markers with low baseline expression is analogous to observations in bulk transcriptomics, where genes at the lower end of detection often show inflated fold-changes that may not reflect true biological differences. Thus, while we will next demonstrate that the grouped LASSO approach pro-

vides excellent patient-level distinctions, it is difficult to interpret in biological terms when based on preliminary, uncurated unsupervised clustering.

Together, these results illustrate how cell type proportions capture shifts in the abundance of broadly defined populations, while marker means provide finer resolution by revealing differences in cell state within those populations. In this way, dioscRi offers a structured starting point for interpreting complex datasets, using coarse annotations to frame the analysis and marker-level features to uncover additional layers of phenotypic variation, with the understanding that some associations will require further validation.

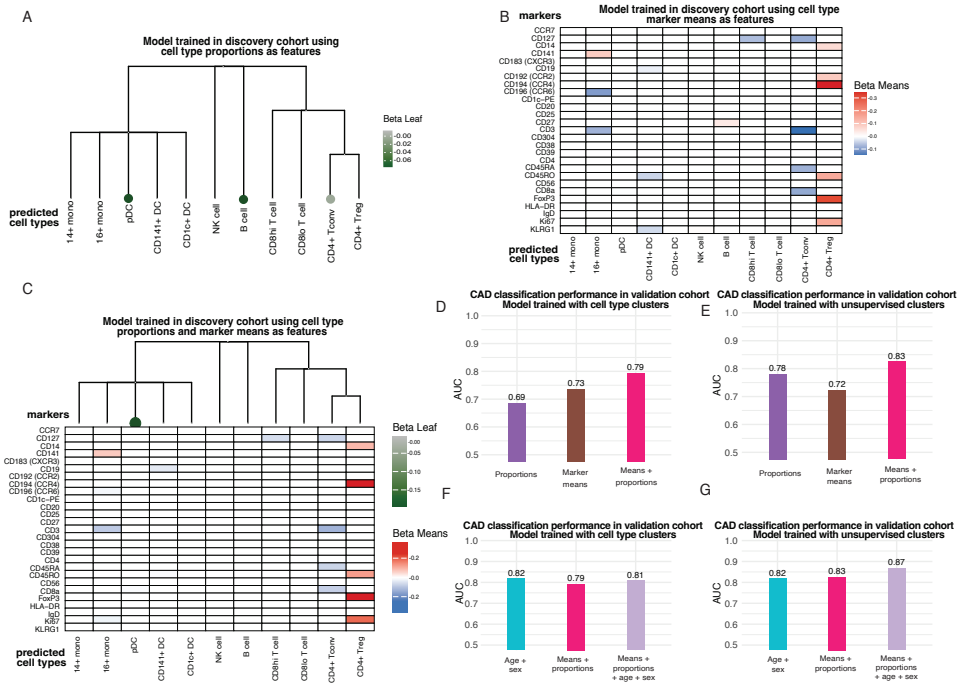

Figure 3.3: Overlapping group LASSO for the validation cohort identifies cell type associations with CAD status for the predicted cell type model trained using the discovery cohort. (A) Each coefficient corresponds to a predicted cell type proportion: negative values indicate higher abundance in CAD– samples, whereas positive values indicate enrichment in CAD+ samples. Marker-level coefficients are shown in (B) and combined cell type proportions and marker means in (C). Combining cell type proportions and marker means enhances patient classification for (D) models trained using predicted cell types and (E) models trained using cell type clusters. Incorporating clinical covariates (age and sex) further improves prediction for (F) models trained using predicted cell types and (G) models trained using cell type clusters.

3.3.5 *Evaluating predictive features and immune associations*

The dioscRi model that is used for patient classification integrates multiple immune features, including cell type proportions and marker means per cell type, which are structured within a biologically or empirically informed hierarchy using a grouped LASSO model. To evaluate the contribution of these components to patient classification performance, we compared three versions of the model: one using only proportions, one using only marker means, and the full model combining both. For models trained using predicted cell types based on manual gates and a biologically derived tree structure from the training cohort, the full model combining proportions and marker means achieved the highest AUC of 0.79 for the validation cohort (Figure 3.3D) in comparison to the proportions-only model (AUC of 0.69) and the means-only model (AUC of 0.73). As detailed above, the validation cohort data was of poor quality and prediction of cell types without first applying time-gate filtering to remove significant machine artefacts likely compromised the estimation of proportions. Indeed, marker means alone generated a higher AUC than proportions, and inclusion of both further improved model performance. In general, cell type proportions from manually gated data have proven superior in understanding patient-level distinctions. Indeed, in the original published analysis of the BioHEART-CT study (Kott et al., 2023), there were no significant differences in marker means between the corresponding gated populations in CAD+ and CAD- populations. These results suggest that proportions and marker means provide complementary information that is particularly useful in analysis of datasets that contain substantial technical artefacts.

We next evaluated the patient classification performance for the models that used clustered cell types to annotate the cells. For these models, which also used an empirically derived tree structure, a different pattern emerged. The proportions-only model performed nearly as well as the full model, with AUCs

of 0.78 and 0.83, respectively (Figure 3.3E), while the means-only model had a smaller AUC of 0.72. This stronger performance of the proportions-only model suggests that the clustering process captures biologically meaningful variation in population abundances. These proportions may reflect underlying immune shifts more directly than marker intensities alone, particularly when clusters align well with disease-related immune states. Together, these findings indicate that combining proportions and marker means improves prediction under both annotation strategies.

We further evaluated the added value of dioscRi by comparing its performance to models using only clinical variables. In the original BioHEART-CT study that was designed to describe immune contributions to CAD, independent of known risk factors, SVM models based on cytometry data alone (AUC = 0.65) were inferior to those using only age and sex (AUC = 0.82). In contrast, the normalised clustering-based model in dioscRi is comparable to the age and sex model, achieving an AUC of 0.83 (Figure 3.3F). It is worth noting that while the cytometry-based models alone (AUC = 0.79 for proportions, 0.72 for means) did not exceed the performance of age and sex alone (AUC = 0.82), this pattern is not uncommon in biomedical datasets where strong demographic predictors exist. Importantly, performance improved further when age and sex were included in conjunction with cytometry characteristics, reaching an AUC of 0.87 (Figure 3.3G). This improvement demonstrates that properly normalised cytometry data provides complementary predictive information when combined with clinical covariates, which is especially important for unsupervised clustering.

3.3.6 *dioscRi outperforms state of the art deep learning approaches*

We next evaluated the performance of dioscRi in comparison to other deep learning methods, including CellCNN and DeepCNN. We performed this comparison across four cytometry datasets: the BioHEART-CT cohorts, the CMV

study SYD519 dataset, the Breast Cancer dataset, and the COVID-19 dataset, each featuring distinct clinical outcomes. For BioHEART-CT, we trained on the discovery cohort and tested on the validation cohort. The COVID-19 and Breast Cancer datasets were split 70–30 into training and testing sets. For the CMV dataset, we followed the split previously described (Hu et al., 2020a), using SYD519 for testing and the remaining studies for training. Across all datasets, we applied a consistent configuration for dioscRi using 11 clusters, while CellCNN and DeepCNN were used with their respective recommended settings. DioscRi outperformed both CellCNN and DeepCNN in all datasets except CMV SDY519 (Figure 3.4). In the BioHEART-CT validation cohort, dioscRi achieved an AUC of 0.83, well above CellCNN (0.60) and DeepCNN (0.62). Similarly, dioscRi performed strongly on the Breast Cancer dataset (AUC = 0.91 vs. 0.74 and 0.78) and the COVID-19 dataset (AUC = 0.98 vs. 0.70 and 0.57). For the CMV dataset, both dioscRi and DeepCNN achieved high AUCs (0.94), outperforming CellCNN (0.80). Both dioscRi and DeepCNN achieve an AUC of 0.94, and upon inspection, both methods correctly classify the same patients. These results underscore dioscRi's robustness and generalisability, particularly in clinical settings where other models show reduced accuracy. In addition to high performance, dioscRi's interpretable structure facilitates biological insight, a key advantage over more opaque deep learning approaches in which expression of cell markers cannot be readily mapped back to individual cells.

3.3.7 *Factors affecting dioscRi's performance*

The modular nature of dioscRi offers flexibility across datasets, but like many models, its performance can be influenced by several tunable components. We systematically explored the impact of key parameters including the number of clusters, bottleneck layer size, regularisation strength (Lambda), and the number of reference samples used for normalisation using AUC and total loss as metrics (Supplementary Figure S5). The number of clusters was a strong driver

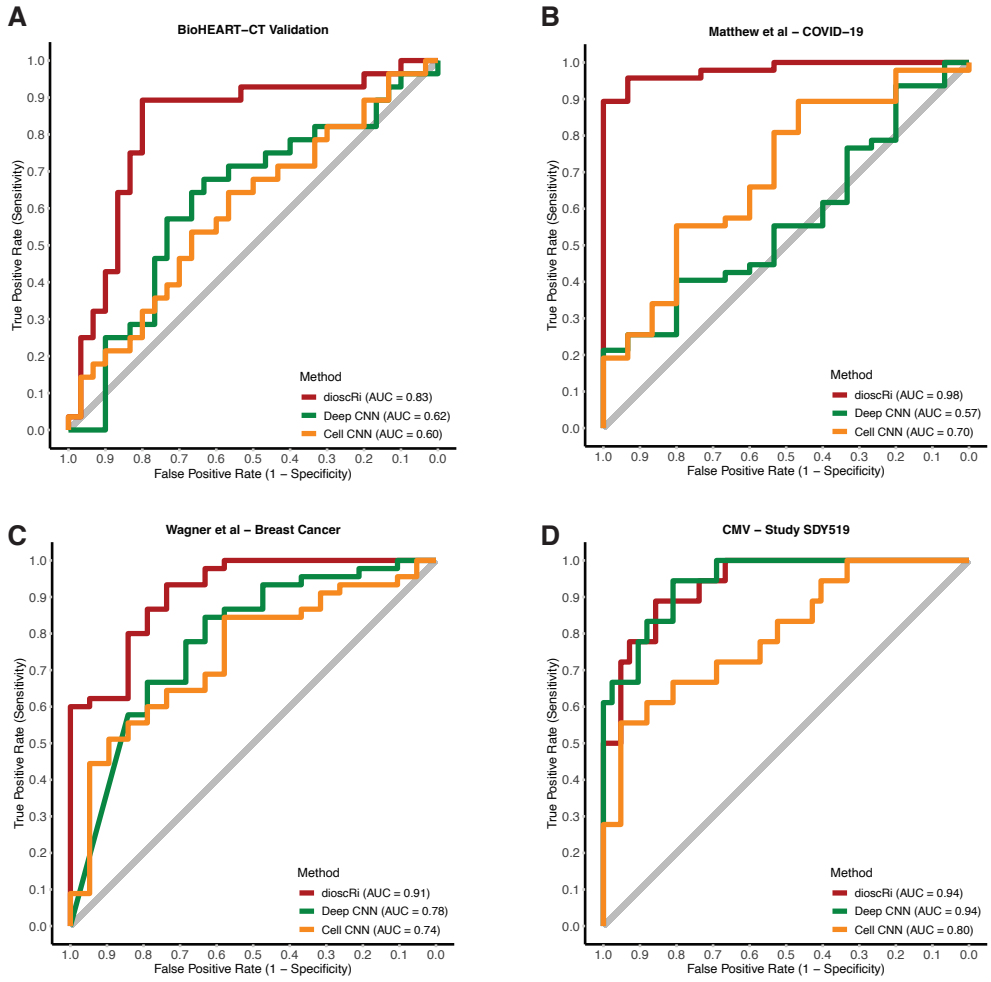

Figure 3.4: DioscRi outperforms both Deep CNN and Cell CNN across all cohorts, achieving the highest AUCs. When applied to (A) the BioHEART-CT validation study and three external datasets, (B) COVID-19, (C) Breast Cancer, and (D) CMV, the DioscRi normalisation-and-prediction pipeline consistently delivers superior classification performance. For the Breast Cancer and COVID-19 datasets, each cohort was split 70% for training and 30% for testing; for the CMV dataset, study SDY519 was reserved entirely for validation.

of predictive performance. Across datasets such as BioHEART-CT, CMV, and Matthew et al., we observed that AUC generally peaks at intermediate values, typically around 11 or 12 clusters, before plateauing or slightly declining (Supplementary Figure S5A). This suggests that dioscRi can capture underlying biological structure without overfitting when cluster number is chosen carefully, with cross-validation or heuristic-based approaches likely to help in new datasets. The bottleneck layer size determined how much information

was retained in the normalised representation, with moderate sizes (around 15) consistently yielding strong performance (Supplementary Figure S5B). Overly small bottlenecks risked losing critical information, while larger ones tended to overfit. This was reflected in a slight upward trend in validation loss at larger sizes despite total loss remaining relatively stable (Supplementary Figure S5C). Lambda, the regularisation parameter in the overlapping group LASSO, shaped the trade-off between model sparsity and performance. AUC showed considerable variability across lambda values, with strong performance observed at lambda = 0.01, 0.03, 0.04, and 0.08 (Supplementary Figure S5D). We selected lambda = 0.05 as it provides a moderate regularisation setting that balances model sparsity with predictive performance, though users may wish to tune this parameter for their specific datasets. Total loss remained stable across this range (Supplementary Figure S5E), indicating that tuning Lambda can improve generalisation without affecting model stability. Finally, the number of reference samples used in MMD-VAE normalisation influenced both performance and consistency. AUC stabilised with around 6 to 7 representative samples, and validation loss improved modestly with larger training subsets (Supplementary Figures S5F-G). These results indicate that dioscRi can achieve strong normalisation performance with a relatively small, well-chosen training set, supporting its use even in studies with limited sample sizes. Together, these analyses highlight the adaptability of dioscRi and provide practical guidance for parameter tuning, enabling users to tailor the framework to diverse cytometry applications while maintaining robust predictive performance.

3.3.8 *Discussion*

To achieve interpretable and transferable insights from cytometry data, we created dioscRi, a framework combining deep learning normalisation with biologically informed feature selection. DioscRi integrates an MMD-VAE for harmonising multi-batch data and an overlapping group LASSO that leverages

cell-type hierarchies to model immune variation. Applied to a coronary artery disease dataset, dioscRi achieved strong predictive performance, and validated a subset of the known associations previously revealed by expert manual gating (Kott et al., 2023). Across multiple datasets, dioscRi outperformed existing deep learning approaches while preserving interpretability, demonstrating its value for both accurate prediction and biological discovery.

A major strength of dioscRi lies in its use of hierarchical group structures to model both cell type proportions and marker expression. This allows it to exploit immune system organisation to improve interpretability relative to models based on unsupervised clustering. Unlike black-box deep learning models, where feature contributions often require post-hoc explanation tools like SHAP, dioscRi yields coefficients that may represent associations between immune features and clinical outcomes. While the incorporation of biologically defined hierarchies enhances interpretability by maintaining a degree of alignment with known cell type relationships, our results also support previous findings (Chan et al., 2021) showing that empirically derived hierarchies can be similarly effective. This balance between biological grounding and data-driven flexibility enables dioscRi to provide biological insights without compromising predictive performance.

Our analysis of the BioHEART-CT cohort demonstrates how dioscRi can enhance both prediction and biological insight from cytometry data. In the original BioHEART-CT study (Kott et al., 2019), SVM models based solely on immune features showed modest predictive performance and did not improve upon simple clinical variables such as age and sex. Using overlapping group LASSO models, we observed more stable and accurate predictions from LDA models trained on either the original manually gated dataset or unsupervised clustering approaches. These improvements were supported by the framework's normalisation strategy and its structured modelling of cell type proportions and marker expression. DioscRi reproduced some previously reported associa-

tions, such as decreased pDCs in CAD, and identified marker-level signals that had been obscured by the initial choice of cell types. This included features related to the previously known differences in the population sizes of naïve and memory CD4+ Tconvs. Importantly, integrating cytometry features with the clinical variables age and sex further improved predictive performance, illustrating the added value of immune profiling when supported by effective normalisation and feature selection.

Through the complementary use of predicted cell type proportions and marker means dioscRi captures immune variation at different resolutions. The proportions reflect differences in cell abundance, while the marker means capture changes in phenotype within the predicted cell types. In practice, the line between these two views is shaped by clustering resolution, where higher-resolution clustering can convert marker-level differences into abundance shifts by dividing broader populations into distinct subtypes. Interestingly, the proportion model performed particularly well with unsupervised clusters, likely because the clustering emphasised sample-level heterogeneity in dominant populations. These findings suggest that proportions can serve as strong predictors when clustering effectively captures meaningful immune variation. Although the combined model ultimately performs best, the strong performance of the predicted cell type model highlights its interpretability and value in workflows driven by data-derived cluster structures.

Despite its strengths, dioscRi has some limitations. While the VAE-based normalisation reduces technical variability and enables transfer across batches, it does not explicitly model batch differences. In cases with substantial differences, such as shifts in instrument settings or sample handling, residual batch effects may persist and moderately influence downstream analyses. In addition, subtle panel variations, including different antibody clones or minor protocol changes, can introduce technical variability in marker expression that is difficult to distinguish from the biological signal. Although dioscRi's latent representa-

tion is robust to many such differences, these sources of variation could impact performance when applying the model across datasets with substantial experimental heterogeneity. In practice, careful study design and harmonisation of upstream protocols can mitigate many of these effects.

In summary, dioscRi offers a robust and interpretable framework for extracting biologically relevant insights from cytometry data and applying them to clinically important prediction tasks. By integrating transferable deep learning–based normalisation with structured modelling of cell type features, it overcomes key limitations of existing approaches and improves both performance and interpretability. Across diverse datasets, dioscRi demonstrates consistent predictive strength. These capabilities position dioscRi as a powerful tool for advancing immune profiling in translational research and clinical applications.

3.4 DATA AVAILABILITY

Publicly available data were used for all evaluations. All data were downloaded as described in the originating manuscripts. The raw and processed data used in the manuscript have been uploaded to Zenodo.

3.5 CODE AVAILABILITY

The *dioscRi* R package is available on Github and is available under the GPL-3 license. All code for analysis done can be found at Github.

4

PACE, PROXIMITY ASSOCIATED CHANGES IN EXPRESSION

Spatial transcriptomics technologies enable measurement of gene expression while preserving tissue architecture, yet quantifying how specific spatial cell-cell interactions drive transcriptional changes remains challenging. Current computational methods either aggregate spatial effects across all neighbor types, analyze genes individually, or fail to account for technical contamination from transcript diffusion between adjacent cells. These limitations obscure which cell type pairs drive spatial signals and how much variance each interaction explains.

In this chapter, we present PACE (Proximity-Associated Changes in Expression), a hierarchical mixed-modelling framework that quantifies cell type-specific spatial interactions in transcriptomic data. PACE decomposes gene expression variance into interpretable components associated with cell type identity, pairwise spatial interactions, and technical spillover. Through hierarchical generalized linear mixed models, the framework partitions expression variance at both single-sample and multi-sample levels, with the latter capturing treatment-specific spatial responses. The Multi-Component Spectral Decomposition (MCSD) scoring identifies coordinated gene programs driving each variance component, revealing spatial regulatory modules rather than isolated gene responses.

Application to breast cancer Xenium data revealed that myoepithelial cells at tumor interfaces undergo progressive loss of basal identity markers, with substantial downregulation of canonical myoepithelial keratins KRT5 and KRT14

quantifying proximity-driven dedifferentiation. Analysis of melanoma CosMx data distinguished progressive from stable disease through differential spatial programs: FN1+ cancer-associated fibroblasts showed opposite proximity responses between outcomes, while SPP1+ tumor-associated macrophages maintained elevated expression throughout progressive disease tissues. By providing quantitative measures of cell type-specific spatial interactions while accounting for technical noise, PACE enables researchers to identify biologically meaningful proximity effects that would be missed by bulk or single-cell analyses lacking spatial context.

This chapter presents work currently being prepared for submission. The research was conducted in collaboration with A/Prof. Ellis Patrick, Shreya Rao, and Prof. John Ormerod. As first author, I led the method development, implemented the PACE framework in R, curated and processed the spatial transcriptomics datasets, performed all computational analyses, interpreted the results, and drafted the manuscript. A/Prof. Ellis Patrick supervised the project and provided guidance on the statistical framework. Shreya Rao contributed to data interpretation and manuscript revision. Prof. John Ormerod provided input on the hierarchical mixed model formulation.

4.1 INTRODUCTION

Cells continuously respond to signals from their tissue microenvironment through proximity-dependent mechanisms that govern immune surveillance, wound healing, and cancer progression (Domínguez Conde et al., 2022a; Liu et al., 2025b; Arora et al., 2023b). In tumours, malignant cells remodel the surrounding stroma through the activation of cancer-associated fibroblasts (Ma et al., 2023). Immune populations modulate tumour behaviour via spatial arrangements that predict therapeutic responses (Wang et al., 2022b). Additionally, epithelial barriers lose integrity through the disruption of tight junction sig-

nalling networks (Zihni et al., 2016). Understanding how specific cell types influence gene expression programs in their neighbours is essential for decoding tissue pathology and identifying therapeutic targets. Recent advances in spatially resolved omics technologies, including Xenium (Janesick et al., 2023) and CosMx Spatial Molecular Imaging (SMI) (He et al., 2022) now enable single-cell resolution mapping of RNA and proteins in intact tissues, revealing immune infiltration patterns in cancer (Jackson et al., 2020), morphogen gradients during development (Wu et al., 2021), and region-specific glial programs (Chen et al., 2020). Datasets generated by these state-of-the-art technologies promise mechanistic insights into cell-cell communication, but realising this potential requires computational frameworks that can quantify cell-type-specific spatial influences while accounting for technical artifacts inherent to these platforms.

Computational methods for analyzing spatial patterns in transcriptomics data have progressed from single to multi gene approaches, yet gaps remain in cell type-specific analysis. SVCA (Arnol et al., 2019), one of the earlier spatial variance decomposition methods, partitions each gene's expression into intrinsic (cell-autonomous) and extrinsic (spatially-induced) components using Gaussian process regression. While SVCA pioneered the concept of separating spatial from non-spatial variation, its gene-by-gene analysis cannot identify which neighboring cell types drive the spatial signals observed. MISTY (Tanevski et al., 2022) advanced to multi gene modelling, using random forests to predict gene expression from multiple spatial views that capture intrinsic expression, juxtacrine signals (direct cell-cell contact effects within approximately 10 micrometers), and paracrine signals (secreted factors that diffuse across broader tissue regions up to hundreds of micrometers). While MISTY successfully identifies spatially regulated genes, it aggregates all neighborhood expression without distinguishing whether effects arise from proximity to immune cells versus stromal cells. SpatioMark (Iyengar et al., 2024) differs by explicitly modelling pairwise relationships, analyzing how each gene's expression correlates with proximity to specific cell types rather than aggregate neighborhoods; however,

it remains at the single gene level and cannot capture coordinated transcriptional programs across multiple genes. SIMVI (Dong et al., 2025a) employs variational autoencoders to learn latent spatial representations that capture how cellular neighborhoods influence expression patterns, but these representations aggregate all spatial influences into a single embedding without cell type resolution. While these methods address some of these challenges individually, the fail to simultaneously quantify cell type-specific interactions, identify coordinated gene programs across multiple genes, support multi-sample analyses for treatment comparisons, and model how the identity of both focal and neighboring cells shapes expression responses.

A further methodological consideration is the choice of distributional assumptions. Earlier spatial variance decomposition methods such as SVCA and Statial (Arnol et al., 2019; Iyengar et al., 2024) employ Gaussian process regression or linear models that assume normally distributed residuals. However, single-cell expression counts exhibit overdispersion that violates Gaussian assumptions, with variance typically exceeding the mean due to biological heterogeneity and technical noise (Hafemeister and Satija, 2019). Moving from Gaussian to negative binomial (NB) models introduces additional complexity: the NB distribution requires estimation of a dispersion parameter that governs the mean-variance relationship, and combining NB likelihoods with Gaussian random effects in a hierarchical framework necessitates careful consideration of how variance partitioning operates across the link function. Despite these challenges, NB-based generalised linear mixed models provide more appropriate inference for count data and enable principled variance decomposition on the link scale (Nakagawa et al., 2017).

Compounding these analytical limitations, transcript misassignment between adjacent cells creates pervasive technical artifacts through multiple mechanisms. Lateral signal spillover occurs when transcripts cross cellular boundaries due to imperfect segmentation algorithms that struggle with tightly packed cells,

RNA diffusion during permeabilisation steps (Saiselet et al., 2020), and Z-axis projection artifacts where overlapping cells in three-dimensional space appear merged in two-dimensional imaging planes (Pentimalli et al., 2025). These phenomena generate false positive correlations between adjacent cells, with contamination rates that vary substantially depending on tissue density and processing conditions (Fan et al., 2023a). The problem is particularly severe at tumour-stroma interfaces and immune synapses where cells are tightly apposed, creating spurious co-expression patterns that can be misinterpreted as biological interactions (Zahedi et al., 2024; NPJ Precision Oncology Editorial Team, 2025). While improved segmentation algorithms reduce error rates (Stringer et al., 2021; Greenwald et al., 2022), they cannot eliminate molecular cross-talk in dense tissues or prevent RNA molecules from diffusing across porous membranes during fixation and permeabilisation (Ergen et al., 2025). Without explicit correction for spillover contamination and diffusion artifacts, analyses risk conflating technical noise with biological signal, a critical confound when studying phenomena such as myoepithelial barrier function at tumour interfaces or macrophage activation programs in the tumour microenvironment.

To address these gaps, we present Proximity Associated Changes in Expression (PACE), a statistical framework that quantifies how cells change their gene expression when positioned near specific neighboring cell types. The core biological question PACE addresses is whether cells adopt different expression profiles based on their spatial neighbors, such as fibroblasts near tumor cells versus those in stromal regions. PACE models pairwise cell type interactions while separating biological proximity effects from technical spillover contamination, and identifies coordinated gene programs responding to spatial context. For multi-sample studies, PACE distinguishes treatment-specific spatial responses from baseline effects, revealing how therapies alter cellular communication. By quantifying cell type-specific spatial interactions while accounting for technical noise, PACE enables researchers to identify biologically meaningful proximity effects that would be missed by single-cell analyses lacking spatial context.

4.2 METHODS

4.2.1 *Spatial neighbourhood quantification*

Here we outline our proposed approach, Proximity Associated Changes in Expression (PACE), for identifying spatially associated differences in cell state of one cell type when it is close in proximity to another cell type (Figure 4.1). PACE quantifies multivariate changes in gene expression, $\mathbf{Y}$, relative to a cell type proximity matrix, $\mathbf{X}$, correcting for lateral spillover, $\mathbf{S}$, using hierarchical mixed-models. We describe this in detail as follows.

4.2.1.1 *Neighbourhood abundance matrix*

We quantified the cellular microenvironment of each cell by constructing a neighbourhood abundance matrix $\mathbf{X}$. For each focal cell i at spatial coordinates $(\mathbf{x}_i, \mathbf{y}_i)$, we computed the abundance of each cell type k within radius r:

$$X_{ik} = \sum_{j \in \mathcal{N}(i),\, c_j = k} K(d_{ij}) \tag{34}$$

where d_{ij} denotes the Euclidean distance between cells i and j, $\mathcal{N}(i) = \{j : d_{ij} \leqslant r\}$ denotes neighbours within the specified radius, $c_j \in \{1, \ldots, K\}$ indicates cell type, and $K(d) = \mathbb{1}_{\{d \leqslant r\}}$ is a uniform kernel. This yields an $n \times K$ matrix $\mathbf{X}$ where rows represent focal cells and columns represent neighbour type abundances.

4.2.1.2 *Spillover correction*

Lateral spillover or transcript diffusion occurs when mRNA from neighboring cells contaminates the focal cell's measurements, leading to technical artifacts that can confound spatial analyses. To account for this, we constructed spillover covariates S_{ik} that capture contamination exclusively from other cell types. Specifically, we derived $\mathbf{S}$ from the neighborhood abundance matrix $\mathbf{X}$

by setting $S_{ik} = X_{ik}$ if cell i is not of type k, and $S_{ik} = 0$ otherwise. This ensures that a cell type cannot spill over into itself, as homotypic contamination is indistinguishable from true expression and thus not modeled as spillover.

4.2.1.3 *Edge correction*

Cells near image boundaries have truncated neighbourhoods. We applied an isotropic area-fraction correction using the ratio of the observable neighbourhood area to the complete circular area.

4.2.2 *Hierarchical mixed models for gene expression*

We developed hierarchical generalised linear mixed models (GLMMs) to decompose gene expression into interpretable components. Expression counts y_{ig} for cell i and gene g were modelled using a negative binomial distribution to accommodate the overdispersion characteristic of single-cell count data.

4.2.2.1 *Negative binomial parameterisation*

The negative binomial distribution can be derived as a Poisson-Gamma mixture, where the Poisson rate parameter follows a Gamma distribution, yielding a marginal distribution with variance $\mathrm{Var}(Y) = \mu + \mu^2/\phi$, where μ is the mean and ϕ is the dispersion parameter. As $\phi \to \infty$, the distribution converges to Poisson; smaller values of ϕ indicate greater overdispersion. We estimated gene-specific dispersion parameters ϕ_g jointly with the mean model parameters, allowing each gene to have its own overdispersion level. This parameterisation explicitly models the quadratic mean-variance relationship inherent to count data, providing more appropriate inference than Gaussian assumptions which implicitly assume constant variance (Nakagawa et al., 2017).

4.2.2.2 *Single sample model*

For individual tissue samples, expression counts follow a negative binomial distribution $y_{ig} \sim \text{NB}(\mu_{ig}, \phi_g)$ with log-mean expression modelled as:

$$\log \mu_{ig} = \beta_{0g} + \sum_{k=1}^{K} \beta_{kg} \mathbf{S}_{ik} + b^{(0)}_{c_i,g} + \sum_{k=1}^{K} b^{(k)}_{c_i,g} \mathbf{X}_{ik} \tag{35}$$

where β_{0g} represents the global baseline expression for gene g, β_{kg} quantifies spillover contamination, $b^{(0)}_{c_i,g}$ captures cell type-specific gene expression baselines with $b^{(0)}_{c,g} \sim \mathcal{N}(0, \sigma^2_{0g})$, and $b^{(k)}_{c_i,g}$ captures how proximity to cell type k affects expression in cells of type c_i with $b^{(k)}_{c,g} \sim \mathcal{N}(0, \sigma^2_{kg})$. The Gaussian random effects operate on the log-link scale, inducing multiplicative effects on the expected counts while the negative binomial likelihood captures residual overdispersion not explained by the random effects structure.

4.2.2.3 *Multi sample model with condition effects*

For identifying spatially associated changes in cell state across multiple images associated with some condition, we extended the model to incorporate a negative binomial distribution $y_{ig} \sim \text{NB}(\mu_{ig}, \phi_g)$ with log-mean expression:

$$\log \mu_{ig} = \beta_{0g} + \sum_{k} \beta_{kg} \mathbf{S}_{ik} + b^{(0)}_{c_i,g} + b^{(R)}_{c_i,g} R_i + \sum_{k} b^{(k)}_{c_i,g} \mathbf{X}_{ik} + \sum_{k} b^{(R\times k)}_{c_i,g} (R_i \mathbf{X}_{ik}) + \nu_{m(i),g} \tag{36}$$

where $R_i \in \{0, 1\}$ indicates treatment condition, $b^{(R)}_{c_i,g}$ captures cell type-specific treatment effects with $b^{(R)}_{c,g} \sim \mathcal{N}(0, \sigma^2_{Rg})$, $b^{(R\times k)}_{c_i,g}$ represents treatment-dependent proximity effects with $b^{(R\times k)}_{c,g} \sim \mathcal{N}(0, \sigma^2_{R\times k,g})$, and $\nu_{m(i),g}$ accounts for image-level variation with $\nu_{m,g} \sim \mathcal{N}(0, \tau^2_g)$. The baseline and proximity random effects $b^{(0)}_{c,g} \sim \mathcal{N}(0, \sigma^2_{0g})$ and $b^{(k)}_{c,g} \sim \mathcal{N}(0, \sigma^2_{kg})$ are as defined in the single-sample model.

4.2.2.4 *Model fitting*

All GLMMs were fitted using restricted maximum likelihood (REML) estimation via the `glmmTMB` package in R (Brooks et al., 2017). REML was chosen over

maximum likelihood (ML) because it provides unbiased estimates of variance components by accounting for the loss of degrees of freedom from fixed effect estimation, which is particularly important when the number of random effect levels is moderate relative to sample size (Nakagawa et al., 2017). Convergence was assessed via gradient checks, and models failing to converge were flagged for manual inspection.

4.2.3 *Variance decomposition framework*

4.2.3.1 *Link-scale decomposition*

We developed a variance decomposition framework operating on the linear predictor (log) scale to quantify the relative importance of each model component, following the approach outlined by Nakagawa et al. (2017) for GLMMs. Decomposing variance on the link scale rather than the response scale is appropriate for non-Gaussian models because the linear predictor is where the fixed and random effects are additive. For fixed effects, we computed:

$$SS_{\text{fixed}}^{(k)} = \sum_{i \in \mathcal{F}} (\mathbf{X}_{ik} \hat{\beta}_k)^2 \tag{37}$$

For random effects:

$$SS_{\text{random}}^{(mlt)} = b_{mlt}^2 \cdot \sum_{i \in \mathcal{F}} z_{ilt}^2 \tag{38}$$

where m indexes grouping factors (cell type or image ID), l indexes levels within groups, t indexes terms, and $\mathcal{F}$ denotes the optional focal cell type subset. The design matrix element z_{ilt} equals $\mathbf{X}_{ik}$ for proximity effects, R_i for treatment effects, and $R_i \cdot \mathbf{X}_{ik}$ for treatment-proximity interactions.

4.2.3.2 *Biological component blocks*

Model components were organised into biologically interpretable blocks that differ between single-sample and multi-sample analyses.

For single-sample analyses, variance is partitioned into four blocks. The Celltype block contains random intercepts capturing baseline expression differences between cell types. The Spatial state block contains random slopes quantifying how proximity to each neighbour type modulates expression within each focal cell type. The Spillover block captures technical contamination from spillover. All unexplained variation and minor terms are aggregated into the Residuals block.

For multi-sample analyses comparing treatment conditions, the variance decomposition extends to six blocks. The Celltype and Spillover blocks remain as described above. However, proximity effects are now split into baseline and treatment-specific components: the Spatial state block captures proximity-induced expression changes in the control condition, while the Responder spatial state block captures how these proximity effects differ under treatment. Additionally, the Responder status block quantifies the direct effect of treatment on each cell type independent of spatial context. This decomposition allows separation of treatment effects that are cell-autonomous versus those that depend on cellular neighbourhoods.

For analyses focusing on specific cellular interactions, only terms corresponding to specified neighbour types are retained in the Cell state and Responder spatial state blocks, with all other terms aggregated into Residuals. This filtering allows precise quantification of specific cell-cell interaction strengths while maintaining the interpretability of the variance decomposition.

4.2.3.3 *Multi-gene aggregation*

To obtain transcriptome-wide multivariate estimates, variance components were aggregated across genes. Let $SS_j^{(m)}$ denote the sum of squares for gene j and model component m, where component m indexes all fixed and random ef-

fect terms (e.g., baseline, proximity effects, treatment effects, interactions, and image-level effects). These were aggregated using residual variance weighting:

$$SS_{total}^{(m)} = \sum_{j=1}^{p} w_j \cdot SS_j^{(m)} \tag{39}$$

where weights $w_j = 1/\max(\hat{\sigma}^2_{resid,j}, \epsilon)$ with $\hat{\sigma}^2_{resid,j} = SS_{resid,j}/n_j$ give greater influence to genes with better model fits. The proportion of total variance explained by each block was then computed as:

$$pct_m = \frac{SS_{total}^{(m)}}{\sum_k SS_{total}^{(k)}} \tag{40}$$

This metric quantifies the relative contribution of each biological process to overall transcriptional variation, providing interpretable effect sizes that are comparable across different experimental conditions and cell type combinations.

4.2.4 *Gene prioritisation via MCSD*

To identify individual genes driving specific variance components, we developed the Multi-Component Spectral Decomposition (MCSD) scoring framework. For each gene j and observation i within focal cell type f, we extracted contributions from the relevant variance block:

$$C_{ij} = \sum_{t \in block} b_{jt} \cdot z_{it} \tag{41}$$

After optional centring by subtracting the mean μ_j, we performed singular value decomposition on the resulting matrix $\mathbf{Z}$ and computed gene-specific scores:

$$MCSD_j = \sum_{k=1}^{r} \mathbf{V}_{jk}^2 \cdot w_k \tag{42}$$

where $\mathbf{V}$ contains the right singular vectors and $w_k = \sigma_k^2 / \sum_{l=1}^{r} \sigma_l^2$ represents the variance proportion explained by component k. Genes were pre-filtered to retain those with prevalence above 0.10 and variance above 1.0 within the focal scope, ensuring robust decomposition.

4.2.5 *Pairwise cell-cell interaction analysis*

To systematically characterise cell-cell interactions, we analysed each focal-neighbour pair (f, h) where $f \neq h$. After subsetting data to focal cell type f and retaining only neighbour type h terms in the variance decomposition, we computed the proportion of variance attributable to the specific cellular interaction. This produces a complete interaction matrix quantifying the transcriptional impact of each pairwise spatial relationship.

4.2.6 *Evaluation data*

4.2.6.1 *Xenium breast cancer*

We first applied our framework to a high-resolution *in situ* transcriptomic dataset introduced in (Liu et al., 2025a), which profiled the human breast tumour microenvironment using Xenium Spatial Molecular Imaging (SMI). The original study combined serial tissue sections and histology-anchored integration to map immune, stromal, and epithelial compartments across invasive and non-invasive regions. For our analysis, we selected one image containing ~ 103,000 segmented cells with spatial coordinates and curated gene expression profiles. Cells were annotated into 18 discrete types, which we consolidated into 9 cell-types (e.g., $CD4^+$ and $CD8^+$ merged into a single T-cell group), and spatial analyses were performed using raw transcript counts and cell type annotations.

4.2.6.2 *CosMx melanoma*

We analysed a CosMx Spatial Molecular Imaging (SMI) dataset introduced in the SIMVI study (Dong et al., 2025b), which profiled human melanoma across 25 patient sections collected after immune checkpoint inhibitor treatment. For this work, we focused on 14 patients with well-defined clinical outcomes (stable disease, SD: $n = 7$; progressive disease, PD: $n = 7$). The dataset comprised $\sim$ 31,000 cells annotated with expression of over 1000 genes, spatial coordinates, and initial labels for 17 discrete types. Cell types with less than 100 cells were excluded and closely related populations merged, yielding six cell types (Tumour, Macrophage, T cell, B cell, Endothelial, Fibroblast) for downstream modelling. Pixel coordinates were converted to micrometres by multiplying (x, y) values by 0.12028, and spatial analyses were performed using raw transcript counts and cell type annotations.

4.3 RESULTS

To identify cell-type-specific spatial interactions and their transcriptional consequences, we developed PACE, a hierarchical mixed-modelling framework that decomposes gene expression variance into interpretable components (Figure 4.1). PACE quantifies whether cells alter their gene expression based on spatial neighbors by partitioning expression into three sources: baseline cell type identity, spatial state (proximity-driven expression changes), and spillover contamination. The framework constructs neighborhood abundance matrices within defined radii and uses these as covariates in a hierarchical generalised linear mixed models for gene-by-gene variance decomposition. PACE provides pairwise analysis showing how proximity to specific neighbor types affects each focal cell type's spatial state, comprehensive variance decomposition quantifying the relative contribution of each component, and gene-level analysis through Multi-Component Spectral Decomposition (MCSD) to identify which genes

drive spatial relationships. This multi-resolution approach reveals both which cell type pairs exhibit the strongest spatial interactions and which genes mediate those proximity-driven changes.

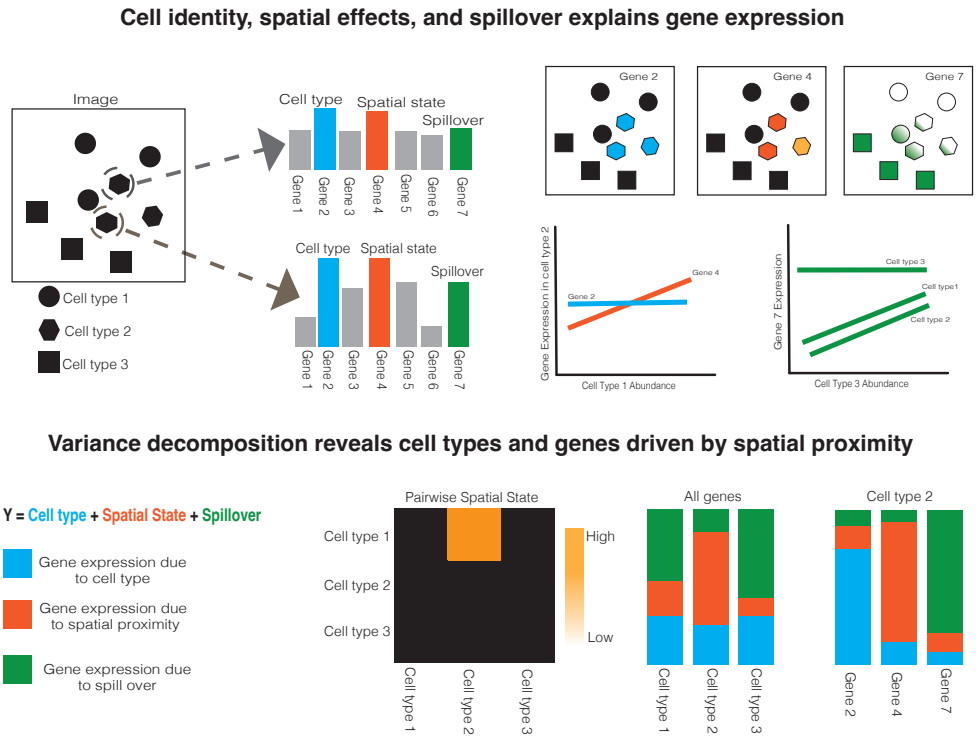

Figure 4.1: Overview of the PACE framework for decomposing spatial gene expression variance. The conceptual schematic illustrates how PACE partitions gene expression (Y) into three components: cell type identity, spatial state, and spillover with accompanying scatter plots demonstrating relationships between cell type abundance and gene expression patterns. Variance decomposition visualizations include: bar charts showing the relative contribution of each component across different cell types; a pairwise spatial state heatmap revealing which focal-neighbor cell type combinations drive the strongest proximity-dependent expression changes; and gene-level variance decomposition identifying specific genes whose expression is most influenced by spatial context. By separating these contributions, PACE quantitatively attributes variance across genes and cell types, identifying those most sensitive to spatial context versus technical artifacts and revealing the magnitude of spatial effects between specific cell type pairs.

4.3.1 *Myoepithelial cells exhibit strongest spatial response to tumor proximity in breast cancer*

Myoepithelial and dendritic cells showed the highest sensitivity to their spatial microenvironment among all breast cancer cell types analyzed. We applied PACE to a human breast cancer Xenium image comprising approximately 103,000 segmented cells across nine cell types. Variance decomposition revealed that cell type identity explained 10-20% of expression variance across most cell types, with myoepithelial and dendritic cells showing proximity-driven contributions of 1-2% of total variance (Figure 4.2A). While this spatial contribution appears modest, it represents a meaningful biological signal given the high dimensionality of the data.

The myoepithelial-tumor cell interaction dominated all other pairwise spatial relationships in the breast cancer microenvironment. Pairwise analysis examining how each focal cell type's expression changes when near each possible neighbor type revealed that the myoepithelial-tumor pairing exhibited the strongest spatial signal (proportion = 0.0125), substantially higher than other interactions such as dendritic-T cell (0.0068) or B cell-T cell (0.0044) pairings (Figure 4.2B). This analysis isolates the spatial state component for each focal-neighbor combination separately, enabling precise quantification of specific cell-cell interactions.

Canonical myoepithelial keratins show proximity-dependent downregulation while luminal marker appears unexpectedly. Using Multi-Component Spectral Decomposition (MCSD) to score genes based on their contribution to spatial state variance, we identified unexpected KRT7 upregulation (MCSD = 0.170) as the top-ranked gene mediating the myoepithelial-tumor spatial interaction (Figure 4.2C). This finding is biologically surprising since KRT7 is a luminal epithelial marker not typically expressed in myoepithelial cells, suggesting either cell misidentification, aberrant expression, or technical artifacts. Given this un-

expected result, we focused our analysis on the canonical myoepithelial markers KRT5 (MCSD = 0.158) and KRT14 (0.155), which ranked second and third respectively. Gene-specific variance decomposition showed KRT5 and KRT14 with 3-5% spatial contribution while KRT7 showed 9% (Figure 4.2D), though the latter's biological relevance in myoepithelial cells remains questionable. The presence of KRT7 signal may indicate luminal cell contamination in regions of disrupted tissue architecture or potential phenotypic plasticity at tumor interfaces requiring further validation.

4.3.2 *Canonical myoepithelial keratins show coordinated downregulation at tumor interfaces*

Myoepithelial cells progressively lose basal keratin expression with increasing tumor proximity. Visualization of KRT5 and KRT14 expression in myoepithelial cells across bins of increasing tumor neighbor count revealed consistent downregulation of both canonical markers (Figure 4.2E, F). Both KRT5 and KRT14 expression showed progressive decrease from low ([0,2]) to high ([8,24]) tumor density regions. This coordinated downregulation of basal keratins indicates loss of myoepithelial differentiation markers at tumor interfaces, consistent with the documented breakdown of myoepithelial barrier function during invasion.

Basal keratin suppression localizes precisely to tumor-myoepithelial interfaces. Spatial maps confirmed that KRT5 and KRT14 downregulation occurs specifically at sites of direct tumor-myoepithelial contact (Figure 4.3A, B). Both markers showed reduced expression along tumor-facing edges of disrupted ducts (orange insets) while maintaining higher expression in myoepithelial cells within intact ducts (blue insets) distant from tumor clusters. The spatial restriction of these changes to tumor interfaces, with preservation of normal expression in non-adjacent regions, demonstrates localized loss of myoepithelial identity

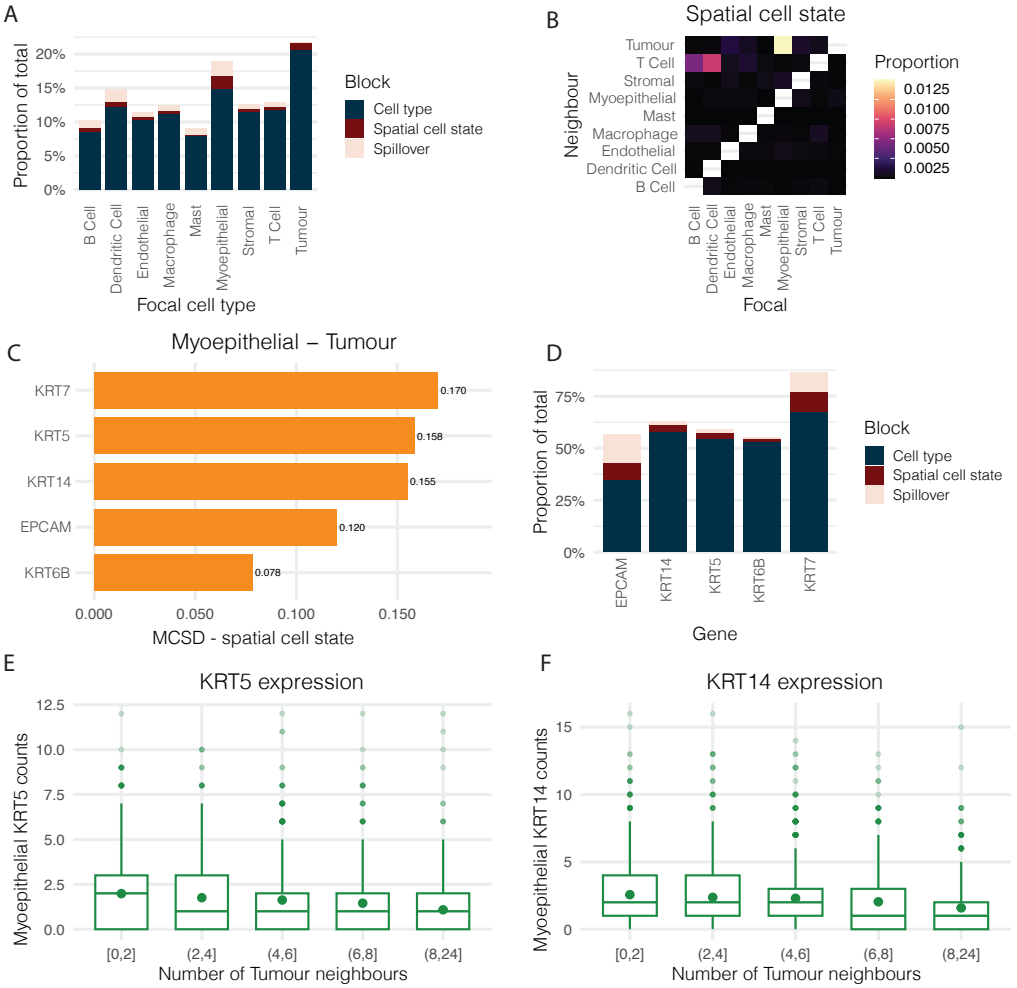

Figure 4.2: Variance decomposition and spatial modelling of breast cancer microenvironment. (A) Variance decomposition across cell types shows that myoepithelial and dendritic cells exhibit the highest spatial signal while tumor cells show minimal spillover, consistent with their dense clustering. (B) Pairwise spatial state heatmap showing the proportion of variance explained by proximity to each neighbor cell type for each focal cell type. The tumor-myoepithelial interaction exhibits the strongest spatial signal among all cell type pairs. (C) Multi-Component Spectral Decomposition (MCSD) ranking of genes whose expression in myoepithelial cells changes with tumor proximity, identifying KRT7, KRT5, and KRT14 as the top drivers of spatial transcriptional changes. (D) Gene-level variance decomposition showing the relative contribution of cell type identity, spatial state, and spillover to the expression variance of each gene. (E-F) Expression of KRT5 and KRT14 in myoepithelial cells stratified by tumor neighbor density. Cells were binned based on the number of tumor cells within a 25 micrometre radius: [0,2] represents myoepithelial cells with 0-2 tumor neighbors (minimal contact), (2,4] with 3-4 neighbors, (4,6] with 5-6 neighbors, (6,8] with 7-8 neighbors, and (8,24] with 9 or more tumor neighbors (high tumor density).

markers rather than global dedifferentiation. This pattern aligns with previous reports of progressive myoepithelial marker loss (p63, calponin, α-SMA) during the transition from DCIS to invasive carcinoma, though the unexpected KRT7 signal warrants careful interpretation and validation of cell type assignments in these transitional zones.

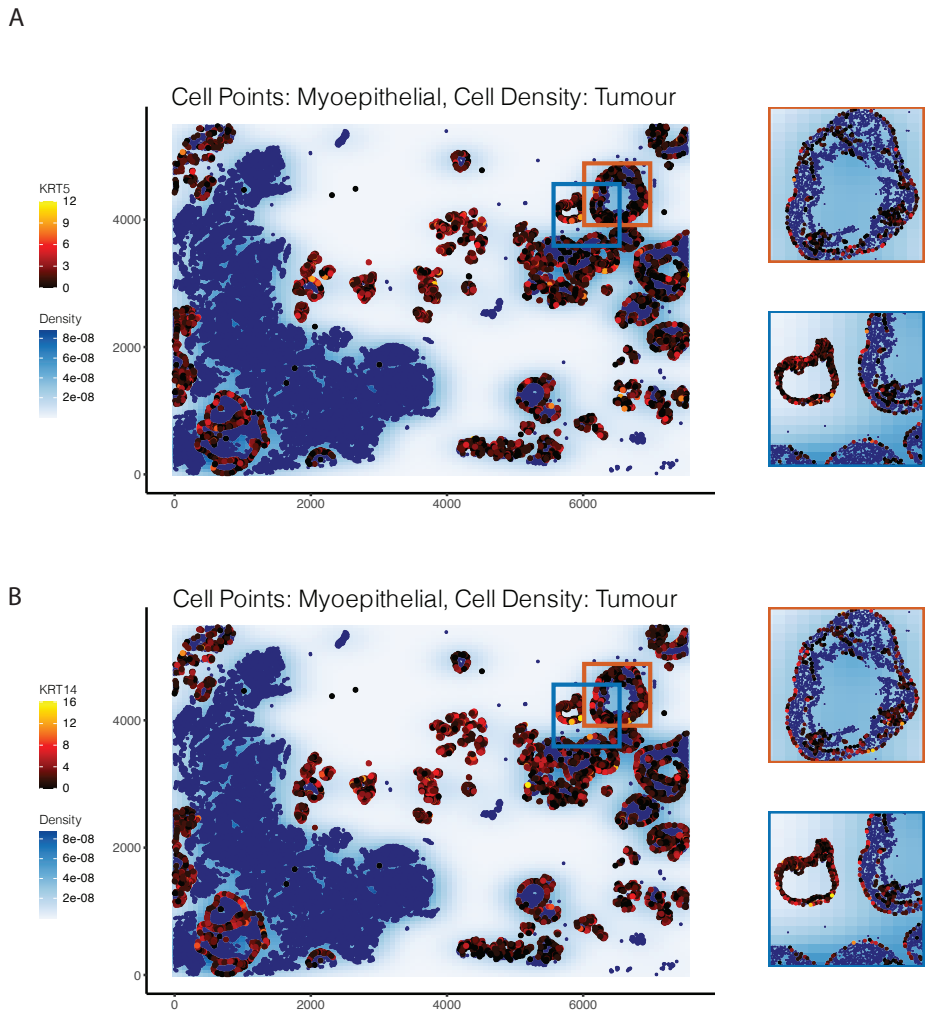

Figure 4.3: Spatial localisation of keratin expression in myoepithelial cells. Scatter plots show myoepithelial cells (points) colored by gene expression overlaid on tumor cell density (background shading). The orange insets show ducts with high tumour density while the blue insets show ducts with low tumour density. (A) KRT5 expression in myoepithelial cells decreases at tumor-duct interfaces. (B) KRT14 expression in myoepithelial cells decreases at the same interfaces while remaining higher in ducts distant from tumor cells. These spatial patterns suggest proximity-dependent changes in keratin expression consistent with loss of basal characteristics.

4.3.3 *Fibroblasts and macrophages show strongest spatial variance in melanoma*

Stromal and immune populations exhibited the highest spatial dependency among melanoma cell types. To demonstrate multi-sample application of PACE, we analyzed the CosMx spatial transcriptomics dataset of melanoma patients with progressive disease (PD) and stable disease (SD). Variance decomposition revealed that fibroblasts and macrophages showed spatial variance contributions of 1-3% total variance, the highest among all cell types (Figure 4.4A). We quantified this by partitioning expression variance into spatial cell state (baseline proximity effects) and responder spatial state (treatment-specific proximity effects) components. In contrast, T cells and B cells showed spatial contributions below 1%, potentially reflecting their lower abundance and fewer spatially responsive genes.

Technical spillover varies inversely with cell density patterns. The lymphocyte populations exhibited the highest spillover contributions at approximately 3-4%, while tumor cells displayed the lowest spillover (Figure 4.4A). This pattern reflects the sparse distribution of lymphocytes in tissue, making them more susceptible to contamination from neighboring cells. Conversely, the minimal tumor spillover is consistent with their dense clustering and spatial homogeneity. Since homotypic interactions are excluded from spillover estimation, densely packed tumor cells show minimal contamination artifacts.

4.3.4 *Fibroblast activation patterns distinguish treatment outcomes*

Cancer-associated fibroblast markers show opposite spatial responses in progressive versus stable disease. Fibroblast-tumor interactions emerged as the most prominent pairwise relationship, displaying the highest responder spatial state proportion across all focal-neighbor cell type combinations (Figure 4.4B). Multi-Component Spectral Decomposition (MCSD) analysis identified canoni-

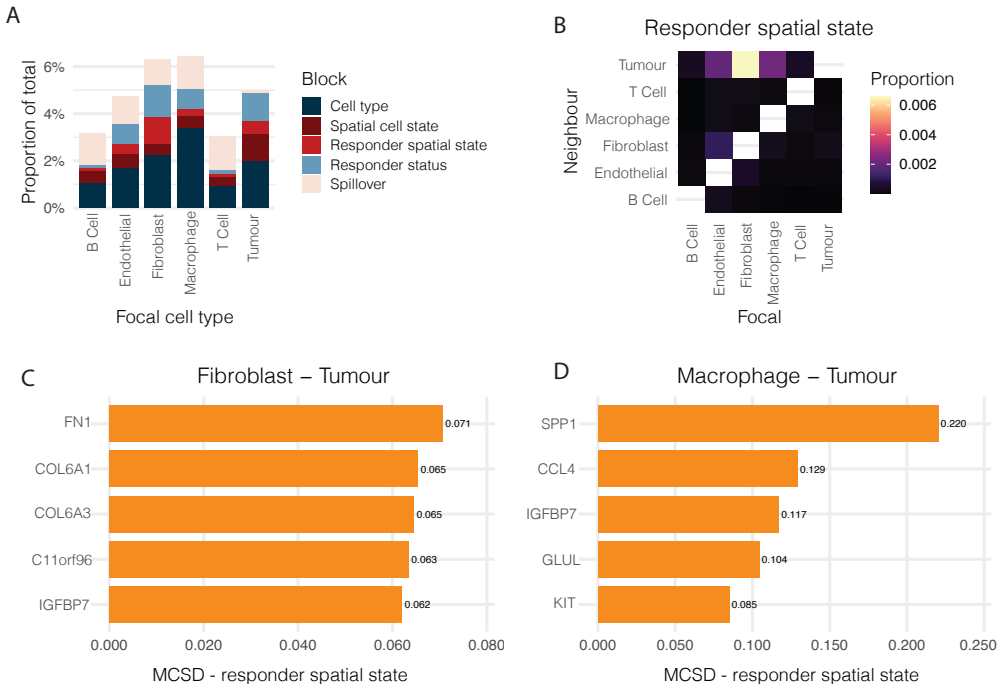

Figure 4.4: Variance decomposition and spatial interaction profiles in melanoma comparing progressive disease (PD) and stable disease (SD). (A) Variance decomposition across cell types, with stacked bars showing the additive contribution of each component (cell type, spatial state, responder spatial state, spillover, and residual) to total expression variance. Fibroblasts and macrophages exhibit the highest spatial signal while tumor cells show minimal spillover, consistent with their dense clustering. (B) Pairwise responder spatial state heatmap showing the proportion of variance attributable to treatment-specific proximity effects for each focal-neighbor cell type pair. The fibroblast-tumor and macrophage-tumor interactions exhibit the strongest differential spatial signals between PD and SD. (C) Multi-Component Spectral Decomposition (MCSD) ranking of genes whose expression in fibroblasts changes with tumor proximity in a treatment-dependent manner, identifying FN1, COL6A1, and COL6A3 as top drivers of spatial transcriptional changes. (D) MCSD ranking for macrophage-tumor interactions identifies SPP1, CCL4, and IGFBP7 as genes whose spatial response to tumor proximity differs between PD and SD.

cal CAF genes as the top spatially responsive drivers, including FN1 (MCSD = 0.071), COL6A1 (0.065), and COL6A3 (0.065) (Figure 4.4C). These genes encode extracellular matrix proteins that are established markers of activated CAFs.

FN1 expression diverges between disease states based on tumor proximity. In progressive disease samples, FN1 expression increased with tumor neighbor density, whereas in stable disease samples, expression decreased slightly (Figure 4.5A). Spatial visualization in a representative PD sample confirmed elevated FN1 expression specifically in tumor-dense regions (Figure 4.5B). Gene-specific variance decomposition showed that FN1, COL6A1, and COL6A3 had substantial contributions from both spatial cell state and responder spatial state variance blocks, indicating tight linkage to local tumor proximity (Figure 4.5C). These findings suggest therapeutic resistance associates with CAF activation in regions of high tumor density.

4.3.5 *SPP1 expression in macrophages distinguishes disease progression*

Tumor-associated macrophage markers show elevated expression in progressive disease regardless of spatial location. Macrophage-tumor interactions ranked second in strength based on responder spatial state proportions (Figure 4.4B). Multi-Component Spectral Decomposition (MCSD) analysis identified SPP1 (MCSD = 0.220) as the top spatial driver, followed by immunomodulatory genes including CCL4, IGFBP7, GLUL, and KIT (Figure 4.4D). SPP1, also known as osteopontin, is a hallmark of tumor-associated macrophages implicated in immunosuppression and therapy resistance.

SPP1 maintains elevation across spatial contexts in progressive disease with a distinctive non-monotonic pattern. Macrophages in PD samples showed consistently elevated SPP1 expression compared to SD across all tumor proximity zones, with highest expression at low tumor density (0-20 neighbors), decreased expression at intermediate density (21-30 neighbors), and partial recovery at

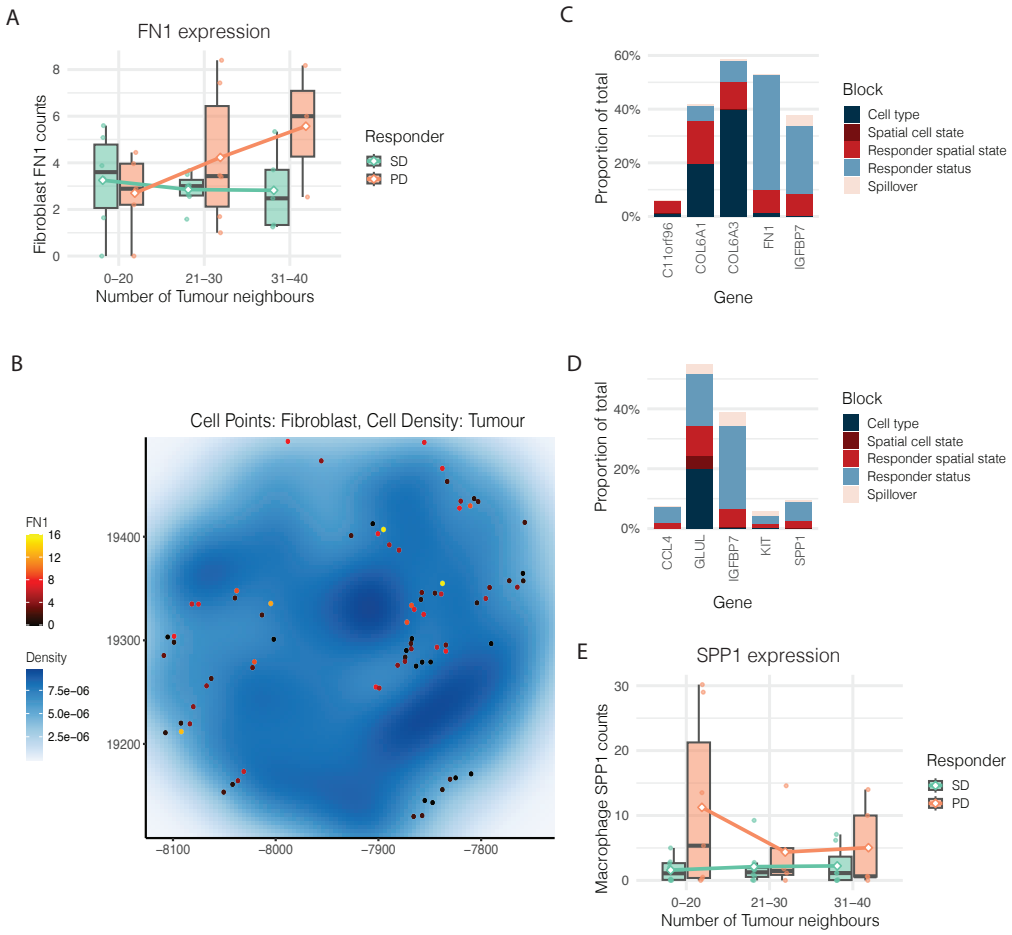

Figure 4.5: Disease-specific spatial expression programs in melanoma. (A) FN1 expression in fibroblasts stratified by tumor neighbor count shows an increasing trend in PD and decreasing trend in SD. (B) Spatial map of FN1 expression in a representative PD sample reveals strong upregulation specifically at tumor-dense interfaces. (C) Gene-level variance decomposition of top fibroblast drivers (FN1, COL6A1, COL6A3) shows notable spatial or responder spatial state contributions. (D) Variance decomposition of top macrophage drivers (SPP1, CCL4, IGFBP7) shows dominant spatial components, particularly for SPP1. (E) SPP1 expression in macrophages stratified by tumor neighbor density shows elevated expression throughout PD tissues with a non-monotonic spatial pattern, while SD maintains consistently low expression.

high density (31-40 neighbors) (Figure 4.5E). SD samples maintained uniformly low SPP1 expression regardless of spatial context. This sustained elevation of SPP1 in PD tissues, despite spatial variation, provides a quantitative marker distinguishing progressive from stable disease that aligns with prior reports linking SPP1-expressing macrophages to checkpoint blockade failure.

4.4 DISCUSSION

We developed PACE to quantify the extent to which cells alter their gene expression programs based on the identity of their spatial neighbors. While existing methods can identify spatially variable genes or aggregate neighborhood effects, they cannot determine which specific cell type pairs drive spatial signals, how much variance each interaction explains, or separate biological proximity effects from technical contamination inherent to dense tissues. PACE addresses these limitations through hierarchical mixed models that partition expression variance into cell type identity, pairwise spatial interactions, and technical spillover, providing percentage breakdowns that enable researchers to prioritize biologically meaningful signals.

Applied to breast cancer and melanoma datasets, PACE revealed proximity-driven phenotypic transitions invisible to bulk analysis: myoepithelial cells at tumor borders showed progressive downregulation of canonical basal markers KRT5 and KRT14, quantifying loss of myoepithelial identity, while melanoma-associated fibroblasts exhibited opposite FN1 responses to tumor proximity in progressive versus stable disease. These findings demonstrate how cell-type-resolved variance decomposition enables quantification of cell type-specific spatial interactions in complex tissues.

The ability to quantify cell-type-specific spatial interactions represents an advance over existing methods. While population-level approaches, such as SVCA (Arnol et al., 2019), partition variance into spatial and non-spatial components,

they cannot distinguish whether observed effects arise from proximity to immune cells or stromal cells. This resolution may be particularly valuable for therapeutic development, as disrupting specific cellular interactions could offer more targeted strategies than broadly modulating spatial effects. PACE's hierarchical framework simultaneously captures cell-autonomous and proximity-dependent effects, providing quantifiable effect sizes (e.g., 5% of KRT5 variance attributable to tumour proximity) rather than abstract embeddings from latent variable methods. This quantitative approach prioritises interactions for experimental validation.

The unexpected detection of KRT7 in cells annotated as myoepithelial raises important considerations about cell identification in disrupted tissue architecture. KRT7 is an established luminal epithelial marker absent from normal myoepithelial cells, which instead express basal keratins including KRT5, KRT14, and KRT17. The KRT7 signal in our myoepithelial population could reflect several possibilities including misidentification of luminal cells as myoepithelial due to segmentation challenges at disrupted tumor interfaces, technical spillover from adjacent KRT7-high tumor cells despite our correction methods, or genuine aberrant expression representing phenotypic plasticity. While the latter would be biologically intriguing, suggesting partial luminal transdifferentiation, the more parsimonious explanation involves technical artifacts or annotation errors. The coordinated downregulation of canonical myoepithelial markers KRT5 and KRT14 provides more reliable evidence of proximity-driven changes, demonstrating progressive loss of basal identity at tumor interfaces consistent with myoepithelial barrier dysfunction during invasion.

The coordinated downregulation of KRT14 and KRT5 in tumour-adjacent myoepithelial cells represents a phenotypic change that warrants further investigation. KRT14 and KRT5 typically mark myoepithelial identity (Nguyen et al., 2018), and their loss has been associated with compromised barrier function (Boecker and Buerger, 2002). This loss of basal markers appears distinct from

classical epithelial-mesenchymal transition (Sarrio et al., 2008). The spatial localisation of these changes at tumour interfaces, with preservation of normal phenotype in distant ducts, indicates localised reprogramming. Recent evidence shows basal-like breast cancers can originate from luminal progenitors acquiring basal characteristics (Van der Veen et al., 2024). Our observations directly demonstrate compromised myoepithelial barrier function. The progressive decrease in KRT5 and KRT14 across tumour density gradients provides quantitative measures of myoepithelial dedifferentiation.

In melanoma, differential expression of SPP1 in macrophages and FN1 in fibroblasts distinguished progressive from stable disease. While SIMVI analysis of this dataset identified spatial programs in melanoma (Dong et al., 2025a), it aggregated all spatial effects into a single latent representation without distinguishing which cell type pairs drove the signal or quantifying their relative contributions. PACE revealed that macrophage-tumor and fibroblast-tumor interactions specifically accounted for the strongest spatial variance, with SPP1 and FN1 as the key mediating genes.

SPP1 has been reported to suppress CD8+ T cell proliferation and IFN-γ production through CD44 engagement (Klement et al., 2018; Moorman et al., 2020). Our analysis revealed that SPP1 expression in macrophages remained elevated throughout PD tissues compared to SD across all tumor proximity zones. The spatial pattern within PD showed highest expression at low tumor density with a dip at intermediate density. While this non-monotonic distribution requires further investigation, the consistent elevation in PD versus SD aligns with reports linking SPP1-expressing macrophages to poor prognosis and checkpoint blockade failure (Lyu et al., 2025).

The activation of FN1+ CAFs provides complementary resistance mechanisms through PI3K/AKT signalling and ECM barrier formation seen in pancreatic cancer (Wang et al., 2024). The divergent FN1 responses to tumor proximity, increasing in progressive but decreasing in stable disease, suggest fundamental

differences in CAF programming between treatment outcomes. The concurrent identification of COL6A1 and COL6A3 aligns with reports linking these extracellular matrix components to poor prognosis (Chen et al., 2021; Zhang et al., 2023a). Together, these spatial programs indicate that resistance involves coordinated remodelling of both immune and stromal compartments, with specific spatial distributions that PACE can quantify at the gene and cell type level.

Several technical decisions influence the interpretation of our findings. The 25 μm neighbourhood radius captures immediate cell contacts and short-range paracrine signalling while avoiding dilution by distant cells. However, different biological processes operate across varying scales. Juxtacrine signalling requires direct contact (<5 μm) while cytokine gradients can extend beyond 100 μm. Additionally, our 2D analysis may miss relevant interactions from cells positioned above or below the focal plane, as tissues are inherently three-dimensional structures where cells separated in the z-axis could be in direct contact or within signalling range. The uniform kernel treats all neighbours equally, which may not capture distance-dependent signal decay. Edge correction assumes uniform density, which can potentially introduce bias at tissue interfaces where cell packing changes.

Computational cost scales primarily with cell count. Gene-wise model fitting was parallelised across genes and completed within approximately one hour for both the single-sample breast cancer analysis (103,000 cells) and multi-sample melanoma analysis (31,000 cells) on a standard laptop (Apple M4, 24 GB RAM). These requirements are tractable for current spatial transcriptomics datasets, though larger atlas-scale studies may benefit from GPU acceleration or approximate inference methods such as variational Bayes. The framework requires discrete cell type annotations, which may not fully capture continuous phenotypic variation observed in some tissues. Future extensions incorporating probabilistic assignments or continuous state variables could address this limitation. The PACE framework is currently implemented as analysis scripts in R;

development of a documented software package with standardised interfaces would facilitate broader adoption and reproducibility. Such a package could include functions for neighbourhood construction, model fitting, variance decomposition, and visualization, enabling researchers to apply these methods to their own spatial transcriptomics datasets.

PACE provides a framework for quantitative analysis of spatial transcriptomics data by decomposing expression variance into interpretable components at cell type resolution. The method identifies coordinated multi-gene programs that respond to specific neighboring cell types, moving beyond single-gene analyses to capture systemic transcriptional changes driven by spatial proximity. By explicitly modelling pairwise cell type interactions while correcting for technical spillover contamination, PACE can distinguish genuine proximity-driven expression changes from technical artifacts (Bai et al., 2021b; Damond et al., 2019a). The observations regarding myoepithelial transitions and immunosuppressive programs would be challenging to identify through existing approaches that either aggregate spatial effects across all neighbors or analyze genes individually (Arnol et al., 2019; Tanevski et al., 2022). Notably, PACE can identify complex spatial patterns including non-monotonic relationships, as demonstrated by SPP1+ macrophages showing highest expression at low tumor density with intermediate suppression in progressive disease. As spatial technologies continue advancing in resolution and coverage (Moses and Pachter, 2022; Palla et al., 2022), variance decomposition frameworks that provide cell type-specific, multi-gene insights will become increasingly valuable for understanding how tissue organization influences cellular behaviour in health and disease.

5

CONCLUSION

This thesis aimed to develop analytical frameworks that characterise disease-associated cellular phenotypic modulation, specifically how individual cells alter their molecular profiles, functional states, and spatial behaviours in response to pathological conditions, through interpretable and robust computational models. The increasing adoption of high-dimensional single-cell technologies, including mass cytometry and spatially resolved imaging assays, has brought forth unprecedented opportunities for biological discovery. However, these opportunities are accompanied by an equally significant set of computational and statistical challenges. As data becomes more complex, noisy, and diverse across cohorts and technologies, the demand for methods that balance scalability, precision, and interpretability continues to grow. The methodology proposed in this thesis responds to that demand by contributing tools that explicitly account for technical noise, facilitate generalisation across cohorts, and produce outputs that can be directly mapped to existing biological knowledge. Each contribution presented in this thesis reflects an effort to bridge the gap between complex data structures and biological insight, ensuring that the models we use can both predict and explain.

Chapter 2 demonstrated that the choice of similarity metric influences cell type discovery in imaging cytometry, with correlation-based metrics consistently outperforming distance-based metrics across 15 benchmarked datasets. The FuseSOM framework leverages this observation through multiview integration, combining complementary similarity measures to achieve more robust clustering than any single metric alone. Such findings suggest reconsidering

the widespread use of Euclidean distance in tools like FlowSOM (Van Gassen et al., 2015a). The results indicate that some existing analyses may have overlooked biologically meaningful populations due to the suboptimal selection of metrics. The evaluation across IMC, MIBI-TOF, CODEX, and seqFISH platforms provides principles for cell type identification that apply across different technologies.

Chapter 3 presents dioscRi, a transferable deep learning framework for predicting clinical outcomes from multi-parameter cytometry data. The critical challenge in cytometry-based biomarker discovery is that models trained on one cohort often fail when applied to new datasets due to batch effects and technical variation. DioscRi addresses this through two innovations. The first is a transferable normalisation scheme that learns invariant representations without requiring explicit batch labels, with the second being an overlapping group LASSO that preserves biological interpretability by structuring features according to cell type hierarchies. Applied to coronary artery disease, dioscRi achieved superior performance compared to CellCNN (Arvaniti and Claassen, 2017b) and DeepCNN (Hu et al., 2020b), demonstrating that carefully designed architectures can balance predictive accuracy with mechanistic transparency. The framework's ability to incorporate both cell type proportions and average marker expression provides complementary views of immune variation, revealing associations that would be missed if considered separately.

Chapter 4 develops PACE, a framework designed to quantify whether and how cells alter their gene expression programs when positioned near specific neighbouring cell types. The central biological question PACE addresses is whether spatial proximity drives transcriptional changes, for instance, whether myoepithelial cells adopt different expression profiles when adjacent to tumour cells versus when isolated in normal ducts. PACE achieves this through variance decomposition that separates gene expression into components attributable to intrinsic cell type identity, proximity-induced changes (spatial state), and tech-

nical spillover contamination. When myoepithelial cells were positioned near tumour cells in breast cancer, PACE revealed progressive loss of basal identity, characterised by coordinated downregulation of the canonical myoepithelial keratins KRT5 and KRT14, demonstrating proximity-driven dedifferentiation at tumour interfaces. In melanoma, PACE identified that fibroblasts increase FN1 expression near tumour cells specifically in progressive disease, while decreasing it in stable disease. Additionally, SPP1 expression in macrophages remained elevated throughout progressive disease tissues compared to stable disease, with a non-monotonic spatial pattern showing highest expression at low tumor density. These proximity-dependent expression changes demonstrate that spatial context functions as a regulatory mechanism in disease, providing quantitative measures of cellular interactions that enable targeted therapeutic strategies.

Three connecting themes emerge across these contributions. First, the explicit modelling of technical artefacts is important for biological discovery. Whether batch effects in mass cytometry or lateral spillover in spatial platforms, each method demonstrates that computational frameworks benefit from accounting for platform-specific noise characteristics. The principle likely extends beyond the specific technologies studied here. As new measurement platforms emerge, understanding and correcting their unique artefact profiles will remain important (Luecken and Theis, 2019a; Heumos et al., 2023).

Second, the interplay between model complexity and interpretability requires careful consideration throughout the analytical process. While deep learning offers representational capacity, the black-box nature of many architectures limits their utility in clinical and research settings where mechanistic understanding is helpful (Murdoch et al., 2019). The methods developed here demonstrate that interpretability need not be sacrificed for performance. DioscRi's hierarchical structure maps directly to cell type relationships while maintaining competitive prediction accuracy, allowing researchers to trace which specific cell pop-

ulations and markers drive clinical associations. PACE's variance components precisely quantify the contribution of cell type identity versus proximity to specific neighbours to a gene's expression, providing percentage breakdowns that can warrant further investigation. The emphasis on interpretable design aligns with calls for explainable AI in biomedicine (Holzinger et al., 2019; Rudin, 2019), particularly as these technologies progress toward clinical deployment, where regulatory approval and physician adoption benefit from understanding model decisions.

The third theme concerns how biological systems exhibit inherent hierarchical organisation that computational methods should respect. Cell types develop through branching lineages where T cells differentiate into CD4+ and CD8+ subsets, which further specialise into memory and effector populations. Gene expression programs operate in coordinated modules where transcription factors regulate multiple downstream targets simultaneously. In tissues, cells influence each other through direct contact and paracrine signalling in ways that depend on both cell types involved. Methods that treat cells as independent data points or features as unstructured lists miss these fundamental relationships. The group LASSO in dioscRi leverages cell type hierarchies to enhance feature selection, while PACE explicitly models the differences between fibroblast-tumour interactions and macrophage-tumour interactions. These approaches suggest that incorporating biological structure into computational frameworks improves both performance and interpretability (Wolf et al., 2019; Setty et al., 2019). Future methods might benefit from prioritising biological plausibility in their design rather than treating single-cell data as generic high-dimensional datasets.

The biological discoveries enabled by these methods illustrate their potential application to disease understanding and therapeutic development. The progressive loss of myoepithelial identity identified by PACE, characterised by proximity-driven downregulation of KRT5 and KRT14 at tumour interfaces, pro-

vides quantitative evidence for myoepithelial barrier dysfunction during breast cancer invasion. The spatial localisation of these changes, with preservation of normal keratin expression in distant ducts, demonstrates that dedifferentiation occurs specifically where tumour cells contact epithelial barriers. In melanoma, the identification of $SPP1^{+}$ TAMs and $FN1^{+}$ CAFs aligns with recent reports that these populations contribute to immunosuppressive niches (Bill et al., 2023; Zhang et al., 2023b). The spatial persistence of SPP1 expression across tumour densities in progressive disease, contrasted with low persistence in stable disease, suggests fundamental differences in cellular programming between treatment outcomes. Together, these observations demonstrate how spatial profiling technologies, when coupled with appropriate analytical frameworks, can quantify cellular interactions that inform therapeutic strategies targeting the tumour microenvironment.

Several directions emerge for extending this work. Multi-modal integration combining transcriptomic, proteomic, and metabolomic data at single-cell resolution will require methods that harmonise features across molecular scales while preserving modality-specific information (Stuart et al., 2019; Argelaguet et al., 2021). Current approaches that force modalities into shared latent spaces may lose unique biological signals (Hao et al., 2021b; Cao et al., 2022a), suggesting that extensions of dioscRi's hierarchical framework could maintain modality-specific features while learning shared representations. Three-dimensional tissue analysis presents additional opportunities, as biological processes including immune migration and vascular networks are inherently three-dimensional (Schott et al., 2024; Xu et al., 2025; Zhao et al., 2022). Extending PACE's variance decomposition to 3D neighbourhoods would require addressing computational scalability and registration across serial sections, with spillover correction becoming even more critical as reconstruction introduces additional error sources.

Temporal dynamics represent another frontier for spatial methods. Tissues undergo continuous remodelling through cell division, death, migration, and differentiation that static snapshots cannot fully capture. Methods that infer temporal trajectories from spatial data or integrate time-series measurements could reveal how cellular interactions evolve during development or disease progression (Schiebinger et al., 2019; Forrow and Schiebinger, 2021). Combining PACE's variance decomposition with trajectory inference might identify when specific spatial programs activate during tumor progression or which cell-cell interactions drive therapeutic resistance. These extensions share a common requirement for methods that balance biological complexity with computational tractability while maintaining the interpretability essential for hypothesis generation and clinical translation.

As technologies continue to evolve toward higher resolution, broader coverage, and multi-modal integration, the principles explored here may help guide the creation of computational frameworks that transform raw measurements into biological understanding. Realising this potential requires methods that can handle the scale, complexity, and noise inherent in single-cell measurements while producing interpretable insights. The journey from data to insight involves both technical and biological considerations. Each computational choice, from similarity metrics to normalisation strategies to model architectures, influences what biological patterns can be discovered. This thesis presents three methods addressing specific challenges in the analytical workflow, from cell type identification through clinical prediction to spatial interaction modelling, demonstrating that the explicit handling of technical artefacts and incorporation of biological structure can improve both performance and interpretability. By developing methods that account for the complexity of biological systems while remaining computationally tractable and interpretable, we work toward a more quantitative understanding of cellular behaviour that could contribute to human health. The frameworks presented in this thesis represent contributions toward that goal, providing tools for the community to extract insights from the

data generated by modern single-cell technologies. The continued development of such methods will be important as we work to translate single-cell biology into clinical applications.

A

APPENDIX FOR CHAPTER 2

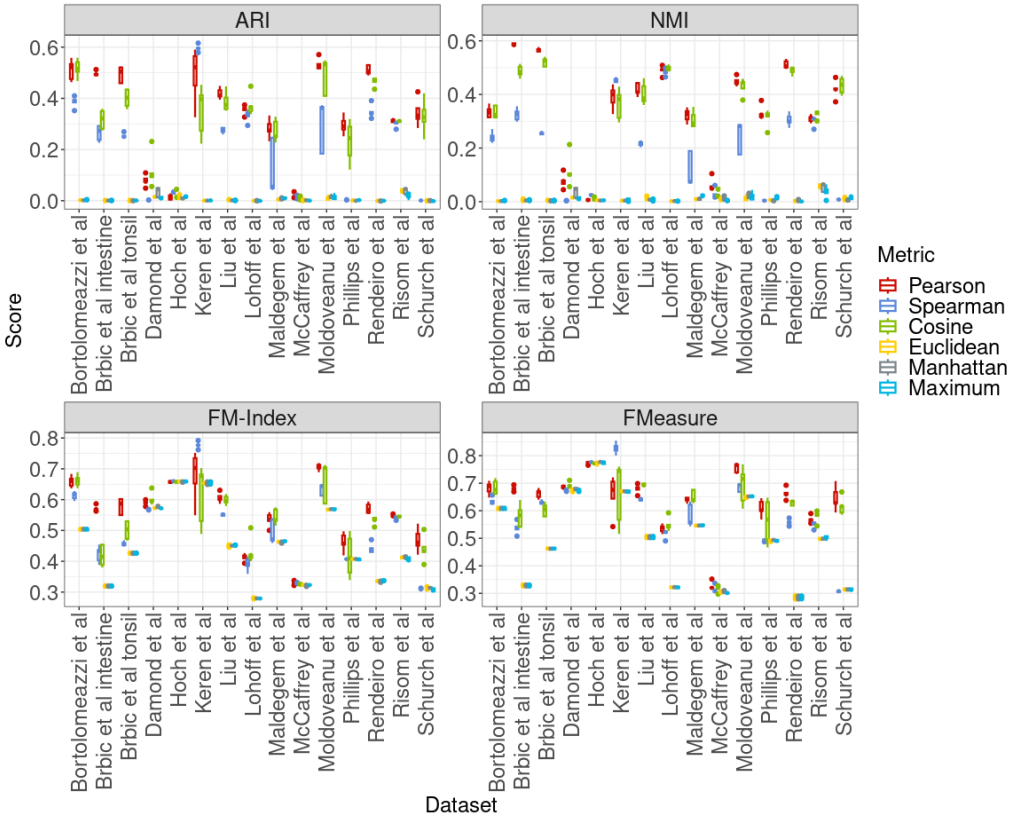

Figure S1: Boxplots of benchmarking similarity metrics on agglomerative clustering of 15 multiplexed imaging datasets. Each dataset was subsetted to 20K cells five times, and the distribution of clustering scores was plotted. Correlation-based distances consistently have a higher average score compared to Euclidean-based distances.

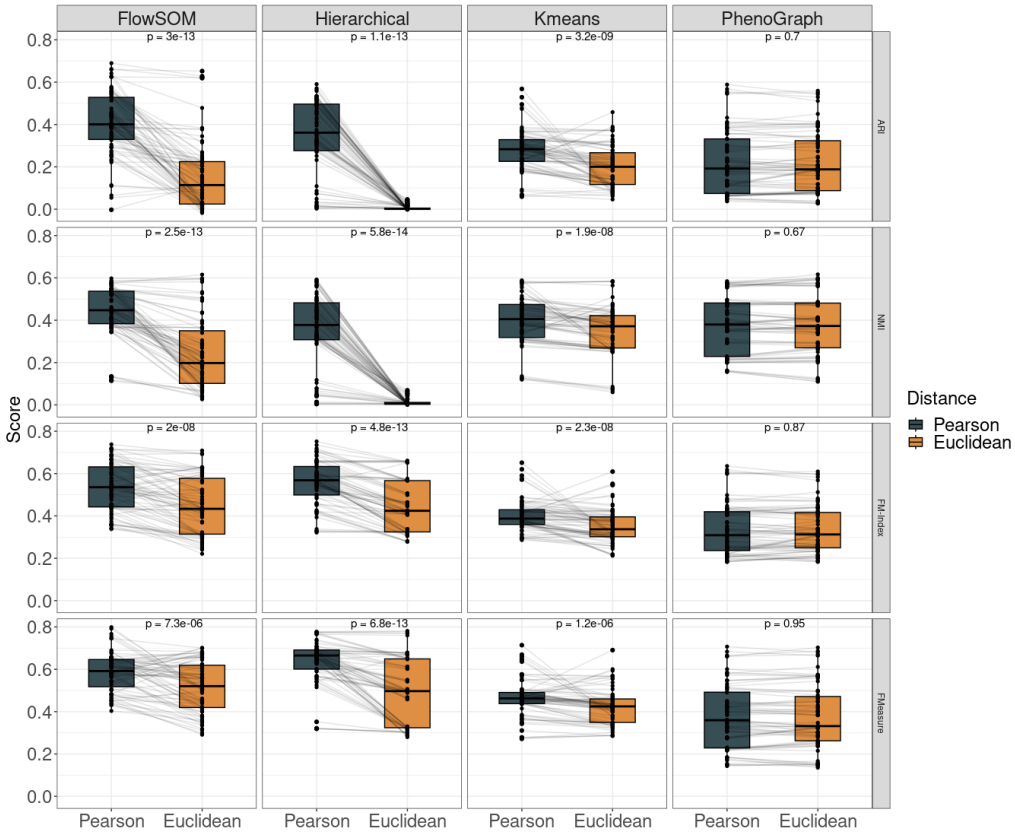

Figure S2: Paired boxplots of clustering performance of four clustering methods using Pearson correlation and Euclidean across four evaluation metrics (ARI, NMI, FM-Index, and FMeasure). Actual range of p-values has been shown. For FlowSOM, Hierarchical clustering, and Kmeans, there is a statistically significant difference in performance between Pearson and Euclidean using the Wilcoxon rank-sum test. This is not evident for PhenoGraph

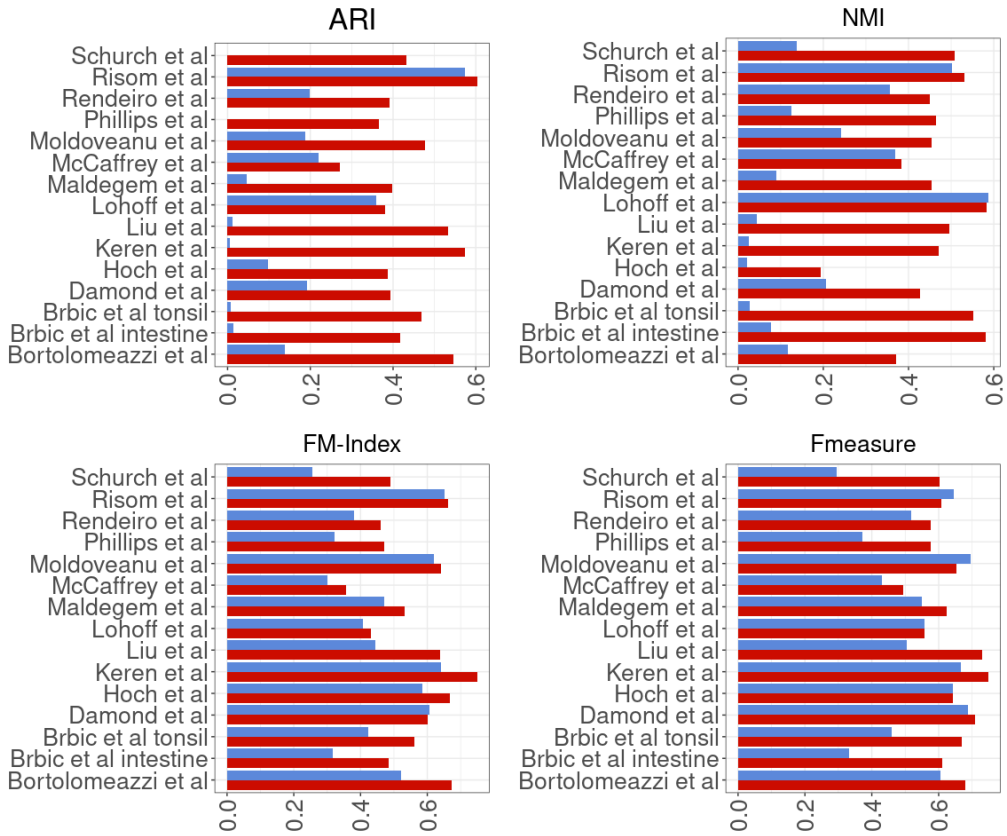

Figure S3: Bar plots of average clustering performance for FuseSOM and FlowSOM showed for all 15 datasets across four performance metrics (ARI, NMI, FM-Index, and FMeasure). For a majority of the datasets, FuseSOM outperforms FlowSOM.

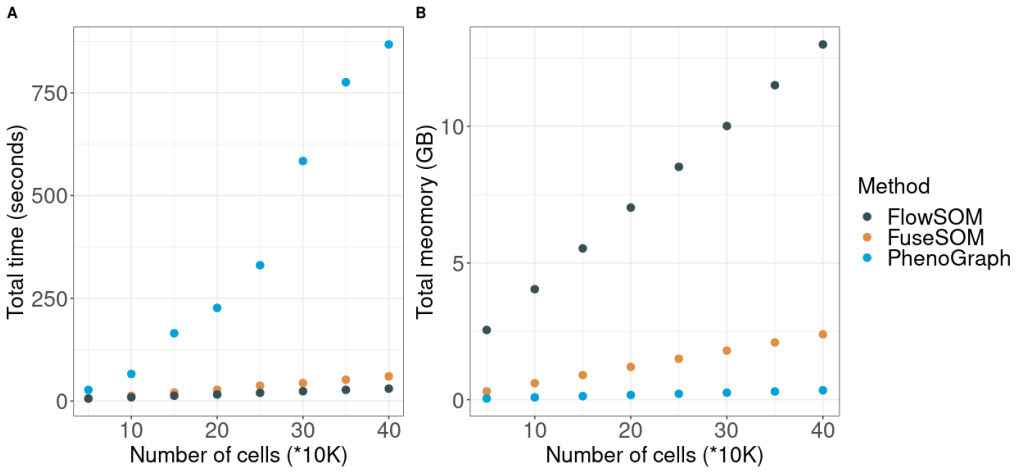

Figure S4: (A) Scatter plot of total running time for FuseSOM, FlowSOM, and PhenoGraph across an increasing number of cells. (B) Scatter plot of total memory usage for FuseSOM, FlowSOM, and PhenoGraph across an increasing number of cells.

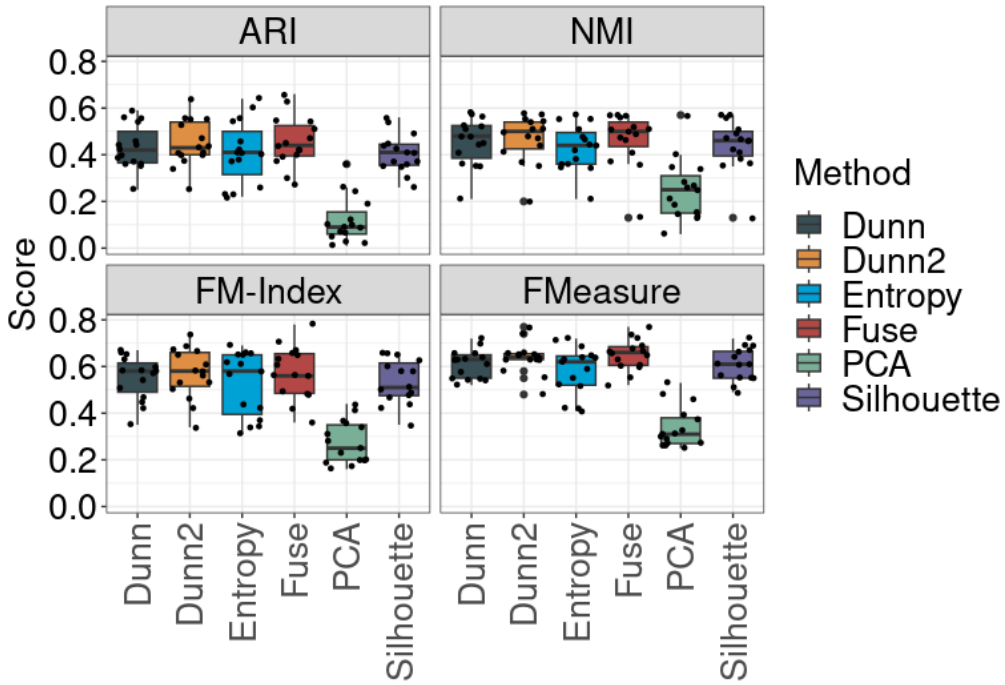

Figure S5: Boxplots of clustering performance of different multiview weighting schemes. There appears to be no significant improvement in performance using other weighting schemes.

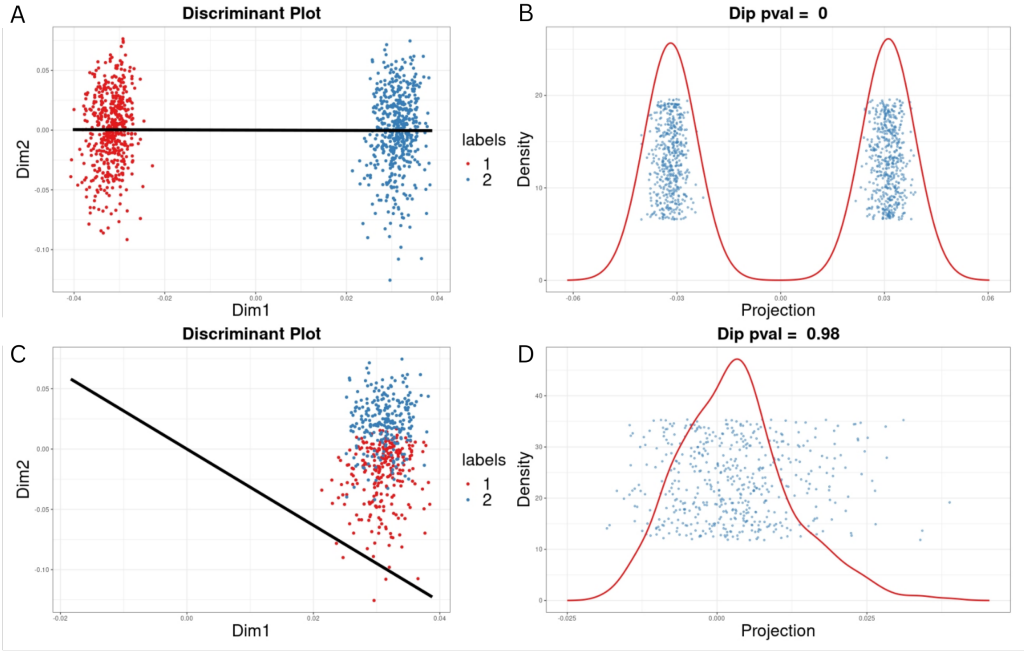

Figure S6: An example of discriminant coordinates cluster number estimation. (A) Shows two classes well separated by a discriminant line in black. (B) Shows the Dip test applied to the projection of the two classes onto this discriminant line in (A). Note that a significant p-value ($p < 0.05$) is obtained for this case. (C) Shows two classes not well separated by a discriminant line in black. (D) Shows the Dip test applied to the projection of the two classes onto this discriminant line in (C). Note that a non-significant p-value ($p > 0.05$) is obtained for this case.

B

APPENDIX FOR CHAPTER 3

B.1 SUPPLEMENTARY TABLES

Name	Layers	Clusters	Cells	MMD-λ	Batch	Input	α	λ
Wagner et al – Breast Cancer	26-22-19	11	10,000	0.1	32	30	0.8	0.050
CMV - Study SDY519	19-15-12	11	10,000	0.1	32	23	0.4	0.057
Bioheart-CT Discovery	23-19-16	11	10,000	0.01	32	27	1.0	0.052
Matthew et al – COVID-19	14-12-11	11	10,000	0.1	32	16	0.4	0.053

Table S1: Hyper-parameters used for each dataset when fitting dioscRi MMD-VAE normalization and Overlap-Group Lasso

Name	Technology	Description	Samples	Outcome
Wagner et al – Breast Cancer	CyTOF	Predicting tumor in breast cancer samples	168	Tumor/non-tumor
CMV - Study SDY519	CyTOF	Predicting positive vs negative CMV titer results in influenza samples	472	Positive/negative
Bioheart-CT Discovery	Mass cytometry	Discovery study for BioHEART-CT	111	Gensini_bin = 1/0
Bioheart-CT Validation	Mass cytometry	Validation study for BioHEART-CT	58	Gensini_bin = 1/0
Matthew et al – COVID-19	Flow cytometry	Profile of CD8+ Non-Naive T Cells to distinguish recovered from COVID-19 vs. healthy	214	COVID-19 recovered/healthy

Table S2: Benchmark datasets. Five published datasets were used to benchmark and compare the performance of dioscRi to other deep-learning models. "Name" refers to each dataset throughout the manuscript

B.2 SUPPLEMENTARY FIGURES

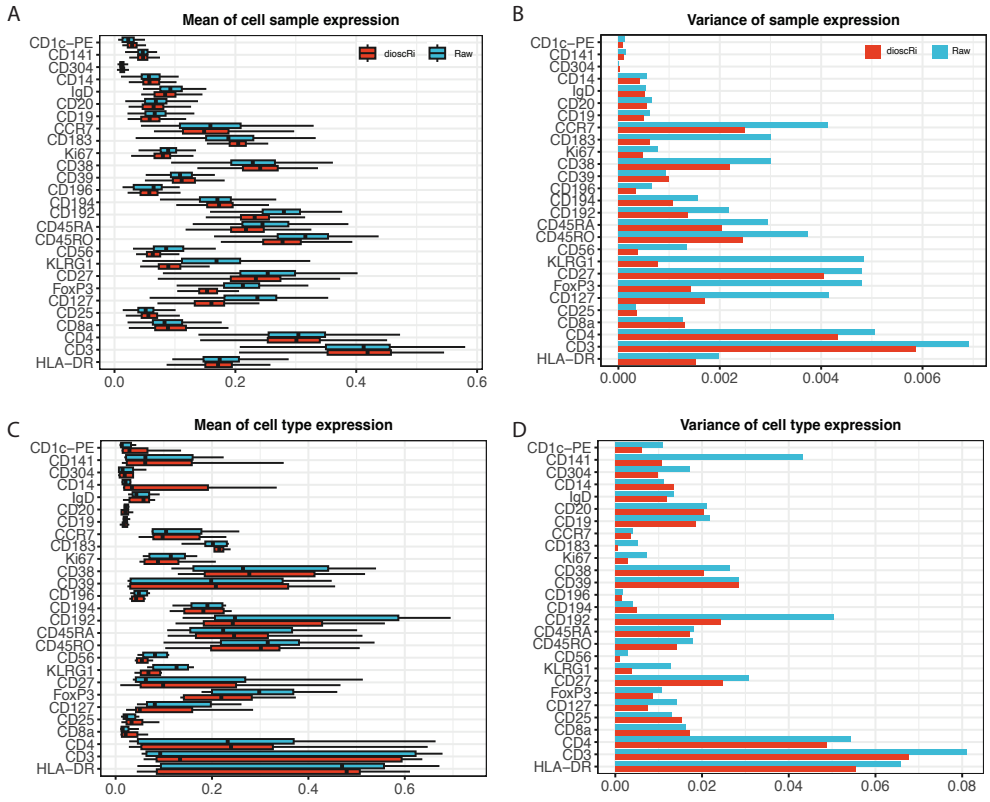

Figure S1: Distribution plots demonstrating the effectiveness of dioscRi normalisation for the discovery cohort using the original manual gates. (A) Each bar shows distribution of the mean expression of the indicated marker within the 111 samples in the discovery cohort. (B) Variance of the means shown in (A). (C) Each bar shows distribution of the mean expression of the indicated marker within the 11 manually gated populations in the discovery cohort. (D) Variance of the means shown in (C)..

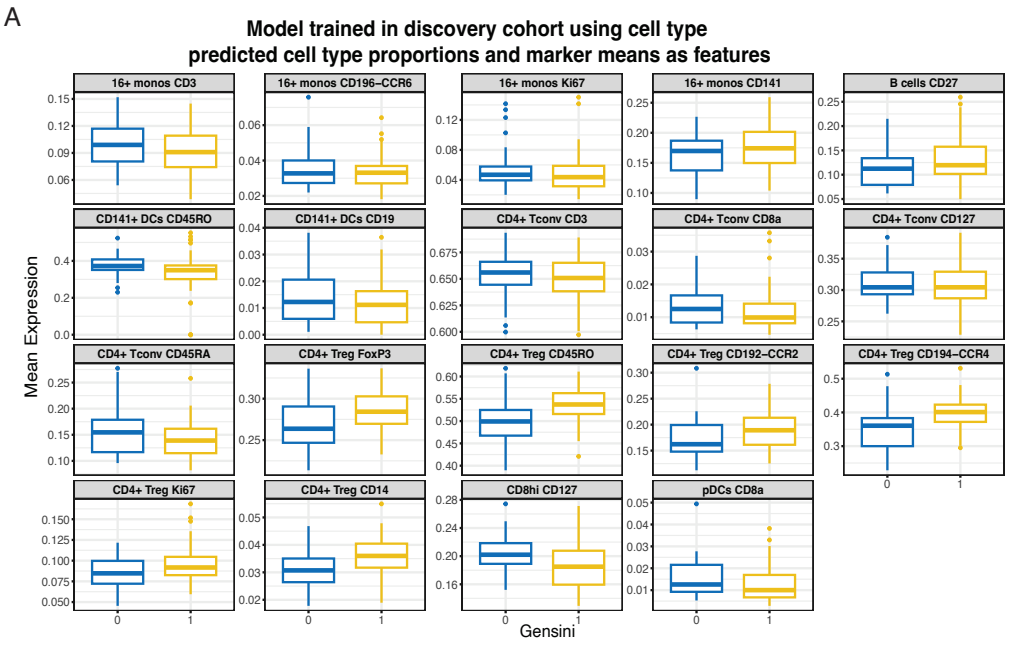

Figure S2: (A) Sample level boxplots of non-zero coefficients of the marker means from the overlapping group lasso model illustrated in Figure 3B-C, split by Gensini (0 = CAD-, 1 = CAD+).

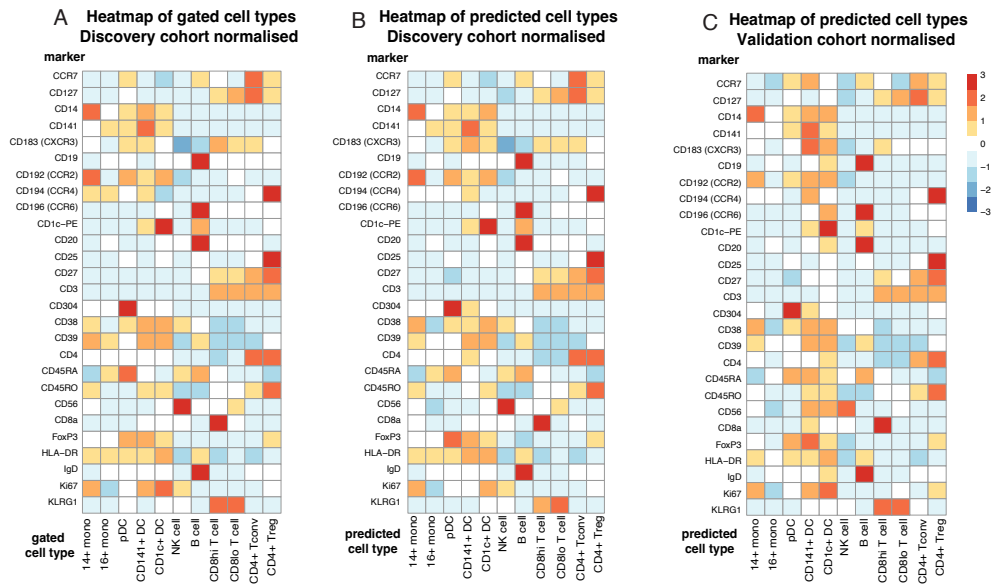

Figure S3: Heatmaps of mean marker expression for normalized data showing 11 cell types from the discovery and validation cohorts. Color scale indicates z-scored expression values: red represents high expression, blue represents low expression, and white indicates intermediate expression. (A) Original manually gated populations, discovery cohort. (B) Predicted cell types, discovery cohort. (C) Predicted cell types, validation cohort.

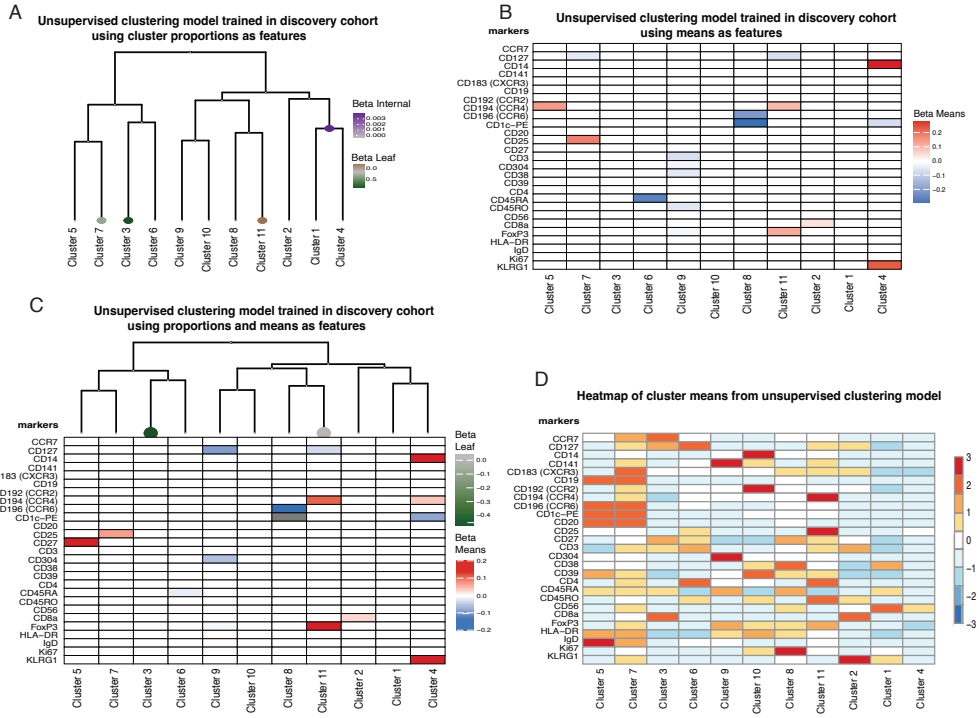

Figure S4: Overlapping group LASSO identifies cell-type associations with CAD status in unsupervised cell clusters. In (A,C), each coefficient corresponds to a cluster proportion: negative values (green) denote higher abundance in CAD– samples, while positive values (purple) denote higher abundance in CAD+ samples. In (B,C), marker level coefficients are shown: negative values (blue) associate with CAD– samples, and positive values (red) associate with CAD+ samples. (D) Heat map of marker expression in clusters.

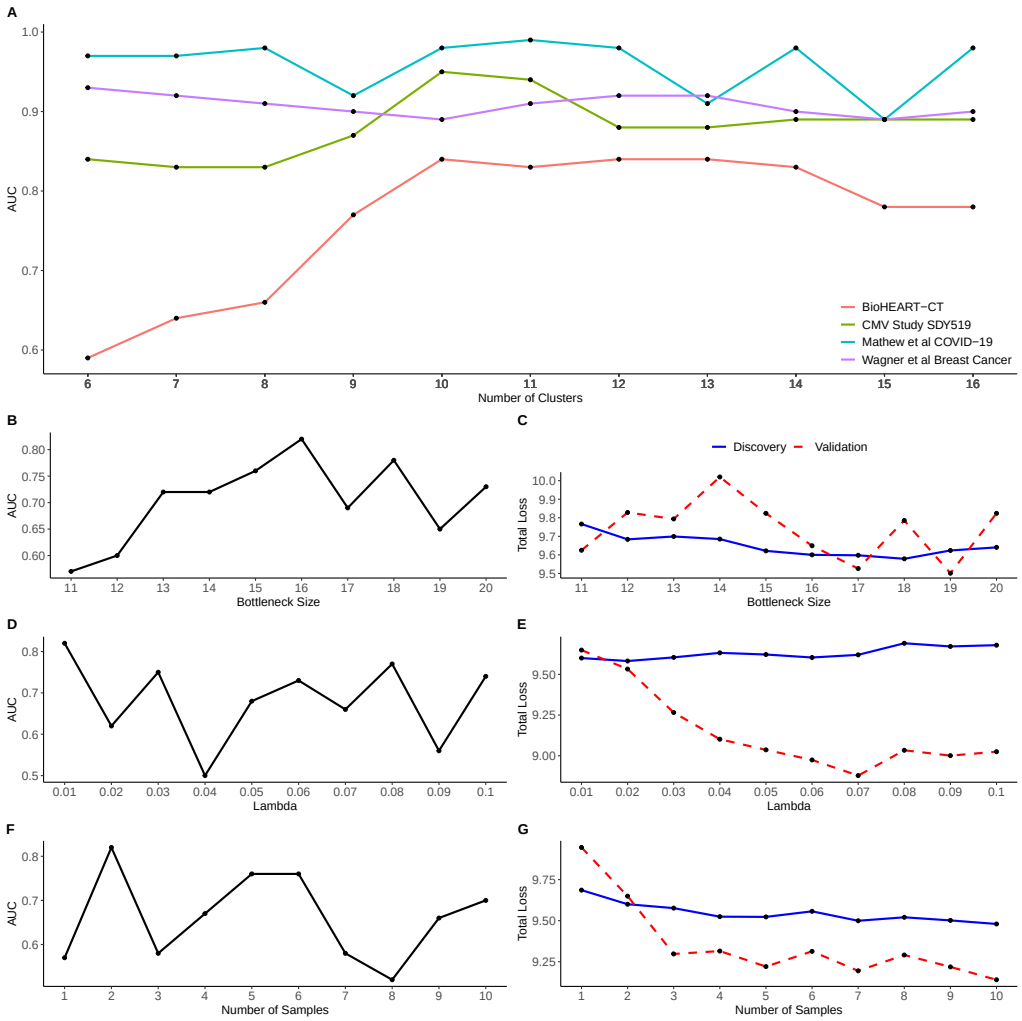

Figure S5: Hyperparameter sweeps identify key parameters affecting dioscRi performance. A) AUC values for CAD status prediction against the number of clusters used during modeling. B) AUC values for CAD status prediction against the bottleneck size for the MMD-VAE normalization. C) Total loss for the discovery and validation studies against the bottleneck size for the MMD-VAE normalization. D) AUC values for CAD status prediction against Lambda for the MMD-VAE normalization. E) Total loss for the discovery and validation studies against Lambda for the MMD-VAE normalization. F) AUC values for CAD status prediction against the number of reference samples for the MMD-VAE normalization. G) Total loss for the discovery and validation studies against the number of reference samples for the MMD-VAE normalization

BIBLIOGRAPHY

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome Biology*, 20(1):194.

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., Scheuermann, R. H., FlowCAP Consortium, and DREAM Consortium (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, 10(3):228–238.

Alpert, A., Pickman, Y., Leipold, M., Rosenberg-Hasson, Y., Ji, X., Gaujoux, R., Rabani, H., Starosvetsky, E., Kveler, K., Schaffert, S., Furman, D., Caspi, O., Rosenschein, U., Khatri, P., Dekker, C. L., Maecker, H. T., Davis, M. M., and Shen-Orr, S. S. (2019). A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nature Medicine*, 25(3):487–495.

Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., and Powell, J. E. (2019). scpred: accurate supervised method for cell-type classification from single-cell rna-seq data. *Genome Biology*, 20(1):264.

Ameijeiras-Alonso, J., Crujeiras, R. M., and Rodríguez-Casal, A. (2019). Mode testing, critical bandwidth and excess mass. *TEST*, 28(3):900–919.

Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., et al. (2014). Multiplexed ion beam imaging of human breast tumors. *Nature Medicine*, 20(4):436–442.

Argelaguet, R., Cuomo, A. S., Stegle, O., and Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nature Biotech-*

nology, 39(10):1202–1215.

Arnol, D., Schapiro, D., Bodenmiller, B., et al. (2019). Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell Reports*, 29(1):202–211.

Arora, R., Cao, C., Kumar, M., Sinha, S., Chanda, A., McNeil, R., Omstead, A., Cao, K., Byron, S. A., Deshpande, A., et al. (2023a). Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nature Communications*, 14(1):5029.

Arora, R. et al. (2023b). Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nature Communications*, 14(1):5029.

Arvaniti, E. and Claassen, M. (2017a). Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature Communications*, 8(1):1–10.

Arvaniti, E. and Claassen, M. (2017b). Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature Communications*, 8(1):1–10.

Atkuri, K. R., Stevens, J. C., and Neubert, H. (2014). Mass cytometry: A highly multiplexed single-cell technology for advancing drug development. *Drug Metabolism and Disposition*, 43(2):227–233.

Baharlou, H., Canete, N. P., Cunningham, A. L., Harman, A. N., and Patrick, E. (2019). Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies. *Frontiers in Immunology*, 10:2657.

Bai, Y., Zhu, B., Rovira-Clave, X., Chen, H., Markovic, M., Chan, C. N., Su, T.-H., McIlwain, D. R., Estes, J. D., Keren, L., and et al. (2021a). Adjacent cell marker lateral spillover compensation and reinforcement for multiplexed images. *Frontiers in Immunology*, 12.

Bai, Y., Zhu, B., Rovira-Clave, X., et al. (2021b). Technical note: Evaluation and mitigation of signal spillover in imaging mass cytometry. *Journal of Immunological Methods*, 499:113163.

Bailur, J. K., McCachren, S. S., Pendleton, K., Vasquez, J. C., Lim, H. S., Duffy, A., Doxie, D. B., Kaushal, A., Foster, C., DeRyckere, D., Castellino, S., Kemp, M. L., Qiu, P., Dhodapkar, M. V., and Dhodapkar, K. M. (2020). Risk-associated alterations in marrow t cells in pediatric leukemia. *JCI Insight*, 5(16).

Bannon, D., Moen, E., Schwartz, M., Borba, E., Kudo, T., Greenwald, N., Vijayakumar, V., Chang, B., Pao, E., Osterman, E., et al. (2021). Deepcell kiosk: scaling deep learning-enabled cellular image analysis with kubernetes. *Nature Methods*, 18(1):43–45.

Batth, I. S., Meng, Q., Wang, Q., Torres, K. E., Burks, J., Wang, J., Gorlick, R., and Li, S. (2020). Rare osteosarcoma cell subpopulation protein array and profiling using imaging mass cytometry and bioinformatics analysis. *BMC Cancer*, 20(1).

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E.-a. D., Krutzik, P. O., Finck, R., et al. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696.

Bhattacharya, S., Dunn, P., Thomas, C. G., Smith, B., Schaefer, H., Chen, J., Hu, Z., Zalocusky, K. A., Shankar, R. D., Shen-Orr, S. S., Thomson, E., Wiser, J., and Butte, A. J. (2018). Immport, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific Data*, 5(1).

Bill, R., Wirapati, P., Messemaker, M., Roh, W., Zitti, B., Duval, F., Kiss, M., Park, J. C., Saal, T. M., Hoelzl, J., et al. (2023). Cxcl9: Spp1 macrophage polarity

identifies a network of cellular programs that control human cancers. *Science*, 381(6657):515–524.

Black, S., Phillips, D., Hickey, J. W., Kennedy-Darling, J., Venkataraaman, V. G., Samusik, N., Goltsev, Y., Schürch, C. M., and Nolan, G. P. (2021). CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nature Protocols*, 16(8):3802–3835.

Boecker, W. and Buerger, H. (2002). Evidence of progenitor cells of glandular and myoepithelial cell lineages in the human adult female breast epithelium. *Cell Proliferation*, 35:465–481.

Bonner, W. A., Hulett, H. R., Sweet, R. G., and Herzenberg, L. A. (1972). Fluorescence activated cell sorting. *Review of Scientific Instruments*, 43(3):404–409.

Bortolomeazzi, M., Keddar, M. R., Montorsi, L., Acha-Sagredo, A., Benedetti, L., Temelkovski, D., Choi, S., Petrov, N., Todd, K., Wai, P., Kohl, J., Denner, T., Nye, E., Goldstone, R., Ward, S., Wilson, G. A., Al Bakir, M., Swanton, C., John, S., Miles, J., Larijani, B., Kunene, V., Fontana, E., Arkenau, H.-T., Parker, P. J., Rodriguez-Justo, M., Shiu, K.-K., Spencer, J., and Ciccarelli, F. D. (2021). Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts. *Gastroenterology*, 161(4):1179–1193.

Brbić, M., Cao, K., Hickey, J. W., Tan, Y., Snyder, M. P., Nolan, G. P., and Leskovec, J. (2021). Annotation of Spatially Resolved Single-cell Data with STELLAR. preprint, Bioinformatics.

Brestoff, J. R. and Frater, J. L. (2022). Contemporary challenges in clinical flow cytometry: Small samples, big data, little time. *The Journal of Applied Laboratory Medicine*, 7(4):931–944.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmtmb

balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

Burkhardt, D. B., Stanley III, J. S., Tong, A., et al. (2022). Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, 40(4):619–629.

Butler, A., Hoffman, P., Smibert, P., et al. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., et al. (2022a). Unsupervised removal of systematic background noise from droplet-based single-cell experiments using cellbender. *Nature Methods*, 19(3):323–325.

Cao, Y., Geddes, T. A., Yang, J. Y. H., and Yang, P. (2020). Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, 2(9):500–508.

Cao, Y., Lin, Y., Patrick, E., Yang, P., and Yang, J. Y. H. (2022b). scfeatures: multi-view representations of single-cell and spatial data for disease outcome prediction. *Bioinformatics*, 38(20):4745–4753.

Chan, A., Jiang, W., Blyth, E., Yang, J., and Patrick, E. (2021). treekor: identifying cellular-to-phenotype associations by elucidating hierarchical relationships in high-dimensional cytometry data. *Genome Biology*, 22(1):324.

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090.

Chen, P. et al. (2021). H3k27 acetylation activated-col6a1 promotes osteosarcoma lung metastasis by repressing stat1 and activating pulmonary cancer-associated fibroblasts. *Theranostics*, 11(3):1473–1492.

Chen, W.-T. et al. (2020). Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182(4):976–991.

Cheng, L., Karkhanis, P., Gokbag, B., Liu, Y., and Li, L. (2022). Dgcytof: Deep learning with graphic cluster visualization to predict cell types of single cell mass cytometry data. *PLOS Computational Biology*, 18(4):e1008885.

Chevrier, S., Crowell, H. L., Zanotelli, V. R., et al. (2018). Compensation of signal spillover in suspension and imaging mass cytometry. *Cell Systems*, 6(5):612–620.

Damond, N., Engler, S., Zanotelli, V. R., et al. (2019a). A map of human type 1 diabetes progression by imaging mass cytometry. *Cell Metabolism*, 29(3):755–768.

Damond, N., Engler, S., Zanotelli, V. R., Schapiro, D., Wasserfall, C. H., Kusmartseva, I., Nick, H. S., Thorel, F., Herrera, P. L., Atkinson, M. A., and Bodenmiller, B. (2019b). A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. *Cell Metabolism*, 29(3):755–768.e5.

Dincer, A. B., Janizek, J. D., and Lee, S.-I. (2020). Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*, 36(Supplement_2):i573–i582.

Domínguez Conde, C. et al. (2022a). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197.

Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., Howlett, S. K., Suchanek, O., Polanski, K., King, H. W., Mamanova, L., Huang, N., Szabo, P. A., Richardson, L., Bolt, L., Fasouli, E. S., Mahbubani, K. T., Prete, M., Tuck, L., Richoz, N., Tuong, Z. K., Campos, L., Mousa, H. S., Needham, E. J., Pritchard, S., Li, T., Elmentaite, R., Park, J., Rahmani, E., Chen, D., Menon, D. K., Bayraktar, O. A., James, L. K., Meyer, K. B., Yosef, N., Clatworthy, M. R., Sims, P. A., Farber, D. L., Saeb-Parsy, K., Jones, J. L., and Teich-

mann, S. A. (2022b). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197.

Dong, X. et al. (2025a). Simvi disentangles intrinsic and spatial-induced cellular states in spatial omics data. *Nature Communications*, 16:58089.

Dong, X. et al. (2025b). Simvi disentangles intrinsic and spatial-induced cellular states in spatial omics data. *Nature Communications*, 16:58089.

Elmarakeby, H. A., Hwang, J., Arafeh, R., et al. (2021). Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352.

Eng, C. H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G., and Cai, L. (2019a). Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568:235–239.

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

Ergen, C. et al. (2025). Resolving cellular cross-contamination in spatial transcriptomics. *Nature Methods*, 22:89–99.

Etemad, K. and Chellappa, R. (1997). Discriminant analysis for recognition of human face images. *Journal of the Optical Society of America A*, 14(8):1724.

Fan, Y., Andrusivova, Z., Wu, Y., Chai, C., Larsson, L., He, M., Zhao, Z., and Lundberg, E. (2023a). Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science*, 380(6648):eadf0704.

Fan, Y. et al. (2023b). Expansion sequencing: Spatially precise in situ transcriptomics in intact tissues. *Science*, 380(6648):eadf0704.

Finak, G., Langweiler, M., Jaimes, M., et al. (2016). Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium. *Scientific Reports*, 6(1):20686.

Fischer, D. S., Schaar, A. C., and Theis, F. J. (2023). Modeling intercellular communication in tissues using spatial graphs of cells. *Nature Biotechnology*, 41(3):332–336.

Forrow, A. and Schiebinger, G. (2021). Lineageot is a unified framework for lineage tracing and trajectory inference. *Nature Communications*, 12(1):1–10.

Fowlkes, E. B. and Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569.

Freytag, S., Tian, L., Lönnstedt, I., et al. (2018). Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data. *F1000Research*, 7:1297.

Fujita, A., Takahashi, D. Y., and Patriota, A. G. (2014). A non-parametric method to estimate the number of clusters. *Computational Statistics & Data Analysis*, 73:27–39.

Gagnon, D. (2025). Achieving reliable cross-laboratory flow cytometry instrument standardization. *Genetic Engineering & Biotechnology News*.

Gayoso, A., Lopez, R., Xing, G., et al. (2022). A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166.

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

Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., Varga, Z., Wild, P. J., Günther, D., and Bodenmiller, B. (2014b). Highly multiplexed imaging of

tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417–422.

Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., and et al. (2014c). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., et al. (2018). Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981.e15.

Greenwald, N. F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., Fullaway, C. C., McIntosh, B. J., Leow, K. X., Schwartz, M. S., et al. (2022). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology*, 40(4):555–565.

Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296.

Han, L., Qiu, P., Zeng, Z., Jorgensen, J. L., Mak, D. H., Burks, J. K., Schober, W., McQueen, T. J., Cortes, J., Tanner, S. D., Roboz, G. J., Kantarjian, H. M., Kornblau, S. M., Guzman, M. L., Andreeff, M., and Konopleva, M. (2015). Single-cell mass cytometry reveals intracellular survival/proliferative signaling in flt3-itd-mutated aml stem/progenitor cells. *Cytometry Part A*, 87(4):346–356.

Hao, Y., Hao, S., Andersen-Nissen, E., et al. (2021a). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021b). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.

Hartigan, J. A. and Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics*, 13(1).

He, S. et al. (2022). High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature Biotechnology*, 40(12):1794–1806.

Herzenberg, L. A., Tung, J., Moore, W. A., Herzenberg, L. A., and Parks, D. R. (2006). Interpreting flow cytometry data: a guide for the perplexed. *Nature Immunology*, 7(7):681–685.

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., et al. (2023). Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572.

Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature Biotechnology*, 37(6):685–691.

Hoch, T., Schulz, D., Eling, N., Gómez, J. M., Levesque, M. P., and Bodenmiller, B. (2022). Multiplexed imaging mass cytometry of the chemokine milieus in melanoma characterizes features of the response to immunotherapy. *Science Immunology*, 7(70):eabk1692.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312.

Hripcsak, G. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Hsu, C.-H., Chen, J., Lee, K.-J., and Donahue, L. R. (2025). Therapy-induced ecm remodeling creates a transient immune barrier in residual melanoma. *Advanced Science*.

Hu, Z., Bhattacharya, S., and Butte, A. J. (2022). Application of machine learning for cytometry data. *Frontiers in Immunology*, 12.

Hu, Z., Tang, A., Singh, J., Bhattacharya, S., and Butte, A. J. (2020a). A robust and interpretable end-to-end deep learning model for cytometry data. *Proceedings of the National Academy of Sciences*, 117(35):21373–21380.

Hu, Z., Tang, A., Singh, J., Bhattacharya, S., and Butte, A. J. (2020b). A robust and interpretable end-to-end deep learning model for cytometry data. *Proceedings of the National Academy of Sciences*, 117(35):21373–21380.

Hunter, M. V., Moncada, R., Weiss, J. M., Yanai, I., and White, R. M. (2021). Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nature Communications*, 12(1):6278.

Iyengar, S. et al. (2024). Spatiomark: Quantifying spatial effects on gene expression. *Bioinformatics*, 40(2):btae024.

Jackson, H. W. et al. (2020). The single-cell pathology landscape of breast cancer. *Nature*, 578(7796):615–620.

Janesick, A. et al. (2023). High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353.

Kalinowski, T., Allaire, J., and Chollet, F. (2024). *keras3: R Interface to 'Keras'*. R package version 1.1.0.

Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., and Decker, S. (2020). Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, 22(1):393–415.

Keren, L., Bosse, M., Thompson, S., Risom, T., Vijayaragavan, K., McCaffrey, E., Marquez, D., Angoshtari, R., Greenwald, N. F., Fienberg, H., Wang, J., Kambham, N., Kirkwood, D., Nolan, G., Montine, T. J., Galli, S. J., West, R., Bendall, S. C., and Angelo, M. (2019). MIBI-TOF: A multiplexed imaging

platform relates cellular phenotypes and tissue structure. *Science Advances*, 5(10):eaax5851.

Kim, T., Chen, I., Lin, Y., Wang, C. Y., Yang, J. Y. H., and Yang, P. (2019a). Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in Bioinformatics*, 20(6):2316–2326.

Kim, T., Chen, I. R., Lin, Y., Wang, A. Y.-Y., Yang, J. Y. H., and Yang, P. (2019b). Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in Bioinformatics*, 20(6):2316–2326.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes.

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201.

Klement, J. D. et al. (2018). An osteopontin/cd44 immune checkpoint controls cd8+ t cell activation and tumor immune evasion. *Journal of Clinical Investigation*, 128(12):5549–5560.

Kopf, A., Fortuin, V., Somnath, V. R., and Claassen, M. (2021). Mixture-of-experts variational autoencoder for clustering and generating from similarity-based representations on single cell data. *PLOS Computational Biology*, 17(6):e1009086.

Korsunsky, I., Millard, N., Fan, J., et al. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296.

Kott, K. A., Chan, A. S., Vernon, S. T., Hansen, T., Kim, T., de Dreu, M., Gunasegaran, B., Murphy, A. J., Patrick, E., Psaltis, P. J., Grieve, S. M., Yang, J. Y.,

Fazekas de St Groth, B., McGuire, H. M., and Figtree, G. A. (2023). Mass cytometry analysis reveals altered immune profiles in patients with coronary artery disease. *Clinical & Translational Immunology*, 12(11).

Kott, K. A., Vernon, S. T., Hansen, T., Yu, C., Bubb, K. J., Coffey, S., Sullivan, D., Yang, J., O'Sullivan, J., Chow, C., Patel, S., Chong, J., Celermajer, D. S., Kritharides, L., Grieve, S. M., and Figtree, G. A. (2019). Biobanking for discovery of novel cardiovascular biomarkers using imaging-quantified disease burden: protocol for the longitudinal, prospective, bioheart-ct cohort study. *BMJ Open*, 9(9):e028649.

Kronstad, L. M., Seiler, C., Vergara, R., Holmes, S. P., and Blish, C. A. (2018). Differential induction of ifn- and modulation of cd112 and cd54 expression govern the magnitude of nk cell ifn- response to influenza a viruses. *The Journal of Immunology*, 201(7):2117–2131.

Kvålseth, T. (2017). On Normalized Mutual Information: Measure Derivations and Properties. *Entropy*, 19(11):631.

Lee, B. H., Kelly, G., Bradford, S., Davila, M., Guo, X. V., Amir, E.-A. D., Thrash, E. M., Solga, M. D., Lannigan, J., Sellers, B., Candia, J., Tsang, J., Montgomery, R. R., Tamaki, S. J., Sigdel, T. K., Sarwal, M. M., Lanier, L. L., Tian, Y., Kim, C., Hinz, D., Peters, B., Sette, A., and Rahman, A. H. (2019). A modified injector and sample acquisition protocol can improve data quality and reduce inter-instrument variability of the helios mass cytometer. *Cytometry Part A*, 95(9).

Lee, E., Chern, K., Nissen, M., Wang, X., Huang, C., Gandhi, A. K., Bouchard-Côté, A., Weng, A. P., and Roth, A. (2023). Spatialsort: A bayesian model for clustering and cell population annotation of spatial proteomics data. *Bioinformatics*, 39(Supplement$_1$) : i131–i139.

Leipold, M. D., Obermoser, G., Fenwick, C., et al. (2018). Comparison of cytof assays across sites: Results of a six-center pilot study. *Journal of Immunological Methods*, 453:37–43.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., ad D. Amir, E., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., Finck, R., Gedman, A. L., Radtke, I., Downing, J. R., Pe'er, D., and Nolan, G. P. (2015a). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., Finck, R., Gedman, A. L., Radtke, I., Downing, J. R., Pe'er, D., and Nolan, G. P. (2015b). Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197.

Lewis, S. M., Asselin-Labat, M.-L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V. C., Merino, D., Rogers, K. L., and Naik, S. H. (2021). Spatial omics and multiplexed imaging to explore cancer biology. *Nature Methods*, 18(9):997–1012.

Li, H., Shaham, U., Stanton, K. P., Yao, Y., Montgomery, R. R., and Kluger, Y. (2017). Gating mass cytometry data by deep learning. *Bioinformatics*, 33(21):3423–3430.

Liu, C. C., Bosse, M., Kong, A., Kagel, A., Kinders, R., Hewitt, S. M., Varma, S., van de Rijn, M., Nowak, S. H., Bendall, S. C., and Angelo, M. (2022). Reproducible, high-dimensional imaging in archival human tissue by multiplexed ion beam imaging by time-of-flight (MIBI-TOF). *Laboratory Investigation*, 102(7):762–770.

Liu, C. C., Greenwald, N. F., Kong, A., McCaffrey, E. F., Leow, K. X., Mrdjen, D., Cannon, B. J., Rumberger, J. L., Varra, S. R., and Angelo, M. (2023). Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering. *Nature Communications*, 14(1).

Liu, P., Pan, Y., Chang, H.-C., Wang, W., Fang, Y., Xue, X., Zou, J., Toothaker, J. M., Olaloye, O., Santiago, E. G., McCourt, B., Mitsialis, V., Presicce, P., Kallapur, S. G., Snapper, S. B., Liu, J.-J., Tseng, G. C., Konnikova, L., and Liu, S. (2024). Comprehensive evaluation and practical guideline of gating methods for high-

dimensional cytometry data: manual gating, unsupervised clustering, and auto-gating. *Briefings in Bioinformatics*, 26(1).

Liu, Y. et al. (2025a). High-parameter spatial multi-omics through histology-anchored integration. *bioRxiv*. Preprint.

Liu, Y. et al. (2025b). Spatiotemporal dynamics of wound healing. *Nature Cell Biology*, 27:123–135.

Lohoff, T., Ghazanfar, S., Missarova, A., Koulena, N., Pierson, N., Griffiths, J. A., Bardot, E. S., Eng, C.-H. L., Tyser, R. C. V., Argelaguet, R., Guibentif, C., Srinivas, S., Briscoe, J., Simons, B. D., Hadjantonakis, A.-K., Göttgens, B., Reik, W., Nichols, J., Cai, L., and Marioni, J. C. (2022). Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nature Biotechnology*, 40(1):74–85.

Lopez, R., Regier, J., Cole, M. B., et al. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058.

Luecken, M. D. and Theis, F. J. (2019a). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746.

Luecken, M. D. and Theis, F. J. (2019b). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746.

Lyu, A. et al. (2025). Evolution of myeloid-mediated immunotherapy resistance in prostate cancer. *Nature*, 637:1207–1217.

Ma, X. et al. (2023). Pan-cancer spatial transcriptomics reveals cancer-associated fibroblast plasticity. *Cancer Cell*, 41:1234–1248.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Maecker, H. T., McCoy, J. P., and Nussenblatt, R. (2012). Standardizing immunophenotyping for the human immunology project. *Nature Reviews Immunology*, 12(3):191–200.

Maecker, H. T. and Trotter, J. (2006). Flow cytometry controls, instrument setup, and the determination of positivity. *Cytometry Part A*, 69A(10):1037–1042.

Mathew, D., Giles, J. R., Baxter, A. E., Oldridge, D. A., Greenplate, A. R., Wu, J. E., Alanio, C., Kuri-Cervantes, L., Pampena, M. B., D'Andrea, K., Manne, S., Chen, Z., Huang, Y. J., Reilly, J. P., Weisman, A. R., Ittner, C. A. G., Kuthuru, O., Dougherty, J., Nzingha, K., Han, N., Kim, J., Pattekar, A., Goodwin, E. C., Anderson, E. M., Weirick, M. E., Gouma, S., Arevalo, C. P., Bolton, M. J., Chen, F., Lacey, S. F., Ramage, H., Cherry, S., Hensley, S. E., Apostolidis, S. A., Huang, A. C., Vella, L. A., Betts, M. R., Meyer, N. J., Wherry, E. J., Alam, Z., Addison, M. M., Byrne, K. T., Chandra, A., Descamps, H. C., Kaminskiy, Y., Hamilton, J. T., Noll, J. H., Omran, D. K., Perkey, E., Prager, E. M., Pueschl, D., Shah, J. B., Shilan, J. S., and Vanderbeck, A. N. (2020). Deep immune profiling of covid-19 patients reveals distinct immunotypes with therapeutic implications. *Science*, 369(6508).

McCaffrey, E. F., Donato, M., Keren, L., Chen, Z., Delmastro, A., Fitzpatrick, M. B., Gupta, S., Greenwald, N. F., Baranski, A., Graf, W., Kumar, R., Bosse, M., Fullaway, C. C., Ramdial, P. K., Forgó, E., Jojic, V., Van Valen, D., Mehra, S., Khader, S. A., Bendall, S. C., van de Rijn, M., Kalman, D., Kaushal, D., Hunter, R. L., Banaei, N., Steyn, A. J. C., Khatri, P., and Angelo, M. (2022). The immunoregulatory landscape of human tuberculosis granulomas. *Nature Immunology*, 23(2):318–329.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278.

Melssen, W., Wehrens, R., and Buydens, L. (2006). Supervised Kohonen networks for classification problems. *Chemometrics and Intelligent Laboratory Sys-*

tems, 83(2):99–113.

Miljkovic, D. (2017). Brief review of self-organizing maps. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1061–1066, Opatija, Croatia. IEEE.

Miron, M., Kumar, B. V., Meng, W., Granot, T., Carpenter, D. J., Senda, T., Chen, D., Rosenfeld, A. M., Zhang, B., Lerner, H., Friedman, A. L., Hershberg, U., Shen, Y., Rahman, A., Luning Prak, E. T., and Farber, D. L. (2018). Human lymph nodes maintain tcf-1hi memory t cells with high functional potential and clonal diversity throughout life. *The Journal of Immunology*, 201(7):2132–2140.

Moffitt, J. R., Lundberg, E., and Heyn, H. (2022). The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics*, 23(12):741–759.

Mokari, M., Mohammadzade, H., and Ghojogh, B. (2018). Recognizing Involuntary Actions from 3D Skeleton Data Using Body States. *Scientia Iranica*, 0(0):0–0.

Moldoveanu, D., Ramsay, L., Lajoie, M., Anderson-Trocme, L., Lingrand, M., Berry, D., Perus, L. J., Wei, Y., Moraes, C., Alkallas, R., Rajkumar, S., Zuo, D., Dankner, M., Xu, E. H., Bertos, N. R., Najafabadi, H. S., Gravel, S., Costantino, S., Richer, M. J., Lund, A. W., Del Rincon, S. V., Spatz, A., Miller, W. H., Jamal, R., Lapointe, R., Mes-Masson, A.-M., Turcotte, S., Petrecca, K., Dumitra, S., Meguerditchian, A.-N., Richardson, K., Tremblay, F., Wang, B., Chergui, M., Guiot, M.-C., Watters, K., Stagg, J., Quail, D. F., Mihalcioiu, C., Meterissian, S., and Watson, I. R. (2022). Spatially mapping the immune landscape of melanoma using imaging mass cytometry. *Science Immunology*, 7(70):eabi5072.

Moorman, H. R. et al. (2020). Osteopontin: A key regulator of tumor progression and immunomodulation. *Cancers*, 12(11):3379.

Moses, L. and Pachter, L. (2022). Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Nakagawa, S., Johnson, P. C. D., and Schielzeth, H. (2017). The coefficient of determination r2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134):20170213.

Nguyen, Q. H. et al. (2018). Profiling human breast epithelial cells using single cell rna sequencing identifies cell diversity. *Nature Communications*, 9(1):2028.

Nielsen, F. (2016). Hierarchical Clustering. In *Introduction to HPC with MPI for Data Science*, pages 195–211. Springer International Publishing, Cham.

Nowicka, M., Krieg, C., Weber, L. M., Hartmann, F. J., Guglietta, S., Becher, B., Levesque, M. P., and Robinson, M. D. (2019). Minimizing batch effects in mass cytometry data. *Frontiers in Immunology*, 10:2367.

NPJ Precision Oncology Editorial Team (2025). Quantifying and correcting slide artifacts in spatial transcriptomics. *NPJ Precision Oncology*, 9:12.

Ogishi, M., Yang, R., Gruber, C., Zhang, P., Pelham, S. J., Spaan, A. N., Rosain, J., Chbihi, M., Han, J. E., Rao, V. K., Kainulainen, L., Bustamante, J., Boisson, B., Bogunovic, D., Boisson-Dupuis, S., and Casanova, J.-L. (2021). Multibatch cytometry data integration for optimal immunophenotyping. *The Journal of Immunology*, 206(1):206–213.

Oliveira, G. et al. (2025). High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nature Genetics*.

Pachitariu, M. and Stringer, C. (2022). Cellpose 2.0: how to train your own model. *Nature Methods*, 19(12):1634–1641.

Palla, G., Fischer, D. S., Regev, A., and Theis, F. J. (2022). Spatial components of molecular tissue biology. *Nature Biotechnology*, 40(3):308–318.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLOS Genetics*, 2(12):1–20.

Pedersen, C. B., Dam, S. H., Barnkob, M. B., Leipold, M. D., Purroy, N., Rassenti, L. Z., Kipps, T. J., Nguyen, J., Lederer, J. A., Gohil, S. H., Wu, C. J., and Olsen, L. R. (2022). cycombine allows for robust integration of single-cell cytometry datasets within and across technologies. *Nature Communications*, 13(1).

Pentimalli, T. et al. (2025). Combining 3d imaging and spatial transcriptomics. *Nature Methods*, 22:45–56.

Perfetto, S. P., Chattopadhyay, P. K., and Roederer, M. (2004). Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology*, 4(8):648–655.

Phillips, D., Matusiak, M., Gutierrez, B. R., Bhate, S. S., Barlow, G. L., Jiang, S., Demeter, J., Smythe, K. S., Pierce, R. H., Fling, S. P., Ramchurren, N., Cheever, M. A., Goltsev, Y., West, R. B., Khodadoust, M. S., Kim, Y. H., Schürch, C. M., and Nolan, G. P. (2021). Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma. *Nature Communications*, 12(1):6726.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The human cell atlas. *eLife*, 6:e27041.

Rendeiro, A. F., Ravichandran, H., Bram, Y., Chandar, V., Kim, J., Meydan, C., Park, J., Foox, J., Hether, T., Warren, S., Kim, Y., Reeves, J., Salvatore, S., Mason, C. E., Swanson, E. C., Borczuk, A. C., Elemento, O., and Schwartz, R. E. (2021). The spatial landscape of lung pathology during COVID-19 progression. *Nature*, 593(7860):564–569.

Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.2.9.

Risom, T., Glass, D. R., Averbukh, I., Liu, C. C., Baranski, A., Kagel, A., McCaffrey, E. F., Greenwald, N. F., Rivero-Gutiérrez, B., Strand, S. H., Varma, S., Kong, A., Keren, L., Srivastava, S., Zhu, C., Khair, Z., Veis, D. J., Deschryver, K., Vennam, S., Maley, C., Hwang, E. S., Marks, J. R., Bendall, S. C., Colditz, G. A., West, R. B., and Angelo, M. (2022). Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell*, 185(2):299–310.e18.

Robertson, H., Kim, H. J., Li, J., Robertson, N., Robertson, P., Jimenez-Vera, E., Ameen, F., Tran, A., Trinh, K., O'Connell, P. J., Yang, J. Y. H., Rogers, N. M., and Patrick, E. (2024). Decoding the hallmarks of allograft dysfunction with a comprehensive pan-organ transcriptomic atlas. *Nature Medicine*, 30(12):3748–3757.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., et al. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467.

Rossi, F. (2012). *yasomi: Yet Another Self Organising Map Implementation*. R package version 0.3/r39.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., and Teichmann, S. A. (2017). The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Saiselet, M. et al. (2020). Transcriptional output, cell types densities and normalization in spatial transcriptomics. *Journal of Molecular Cell Biology*, 12(11):906–908.

Samadani, A.-A., Kubica, E., Gorbet, R., and Kulić, D. (2013). Perception and Generation of Affective Hand Movements. *International Journal of Social Robotics*, 5(1):35–51.

Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L., and Nolan, G. P. (2016). Automated mapping of phenotype space with single-cell data. *Nature Methods*, 13(6):493–496.

Sarrio, D. et al. (2008). Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Research*, 68(4):989–997.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.

Schott, M., León-Periñán, D., Splendiani, E., Strenger, L., Licha, J. R., Pentimalli, T. M., Schallenberg, S., Alles, J., Samut Tagliaferro, S., Boltengagen, A., Ehrig, S., Abbiati, S., Dommerich, S., Pagani, M., Ferretti, E., Macino, G., Karaiskos, N., and Rajewsky, N. (2024). Open-st: High-resolution spatial transcriptomics in 3d. *Cell*, 187(15):3953–3972.e26.

Schuyler, R. P., Jackson, C., Garcia-Perez, J. E., et al. (2019). Minimizing batch effects in mass cytometry data. *Frontiers in Immunology*, 10:2367.

Schürch, C. M., Bhate, S. S., Barlow, G. L., Phillips, D. J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D. R., Kinoshita, S., Samusik, N., Goltsev, Y., and Nolan, G. P. (2020). Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell*, 182(5):1341–1359.e19.

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with palantir. *Nature Biotechnology*, 37(4):451–460.

Silverman, B. W. (1981). Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):97–99.

Simpson, G. L. and Oksanen, J. (2021). *analogue: Analogue and weighted averaging methods for palaeoecology*. R package version 0.17-6.

Smith, M. (2023). *ReductionWrappers: Wrapper exposing several Python dimensional reduction tools*. R package version 2.5.4.

Spitzer, M. H. and Nolan, G. P. (2016). Mass cytometry: single cells, many features. *Cell*, 165(4):780–791.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J. P., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods*, 9(3):386–396.

Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.

Sugar, C. A. and James, G. M. (2003). Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *Journal of the American Statistical Association*, 98(463):750–763.

Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, 17(2):193–200.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell rna-seq in the past decade. *Nature Protocols*, 13(4):599–604.

Takahashi, K., Kochin, B. F., Shankar, G., Ching, K. L., Mustapha, I., Cobos-Uribe, C., Röltgen, K., Piechocka-Trocha, A., Hill, B., Lefteri, D. A., et al. (2021). Multi-batch cytometry data integration for optimal immunophenotyping. *The Journal of Immunology*, 206(1):206–213.

Tanevski, J. et al. (2022). Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome Biology*, 23(1):97.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., et al. (2009). mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382.

Thrun, M. C. and Stier, Q. (2021). Fundamental clustering algorithms suite. *SoftwareX*, 13:100642.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Tieleman, T. and Hinton, G. (2012). *Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude*.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2017). Wasserstein auto-encoders.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):1–12.

Tracy, C. A. and Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159(1):151–174.

Tran, H. T. N., Ang, K. S., Chevrier, M., et al. (2020a). A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome Biology*, 21(1):1–32.

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., and Chen, J. (2020b). A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome Biology*, 21(1):1–32.

Trussart, M., Teh, C. E., Tan, T., et al. (2020). Cycombine: a robust method for combining multiple cytometry datasets. *Bioinformatics*, 36(10):3104–3111.

Van der Veen, E. E. et al. (2024). Oncogene activated human breast luminal progenitors contribute basally located myoepithelial cells. *Breast Cancer Research*, 26:187.

Van Gassen, S., Callebaut, B., Van Helden, M. J., et al. (2015a). Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645.

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., and Saeys, Y. (2015b). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645.

van Maldegem, F., Valand, K., Cole, M., Patel, H., Angelova, M., Rana, S., Colliver, E., Enfield, K., Bah, N., Kelly, G., Tsang, V. S. K., Mugarza, E., Moore, C., Hobson, P., Levi, D., Molina-Arcas, M., Swanton, C., and Downward, J. (2021). Characterisation of tumour microenvironment remodelling following oncogene inhibition in preclinical studies with imaging mass cytometry. *Nature Communications*, 12(1):5906.

Verschoor, C. P., Lelic, A., Bramson, J. L., and Bowdish, D. M. E. (2015). An introduction to automated flow cytometry gating tools and their implementation. *Frontiers in Immunology*, 6.

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergén, L., Navarro, J. F., et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16(10):987–990.

Wagner, J., Rapsomaniki, M. A., Chevrier, S., Anzeneder, T., Langwieder, C., Dykgers, A., Rees, M., Ramaswamy, A., Muenst, S., Soysal, S. D., Jacobs, A., Windhager, J., Silina, K., van den Broek, M., Dedes, K. J., Rodríguez Martínez, M., Weber, W. P., and Bodenmiller, B. (2019). A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell*, 177(5):1330–1345.e18.

Wang, I. H. et al. (2025). Spatial transcriptomics reveals human cortical layer and area specification. *Nature*.

Wang, K. Y. X., Pupo, G. M., Tembe, V., Patrick, E., Strbenac, D., Schramm, S.-J., Thompson, J. F., Scolyer, R. A., Yang, J. Y. H., Mann, G. J., and Yang, P. (2022a). Cross-platform omics prediction procedure: a statistical machine learning framework for wider implementation of precision medicine. *npj Digital Medicine*, 5(1):85.

Wang, S. et al. (2022b). Spatial architecture of tumor-infiltrating lymphocytes predicts clinical outcome. *Nature Medicine*, 28:1456–1467.

Wang, X. et al. (2024). Fn1 from cancer-associated fibroblasts orchestrates pancreatic cancer metastasis via integrin-pi3k/akt signaling. *Frontiers in Oncology*, 15:1595523.

Watson, E. R., Mora, A., Taherian Fard, A., and Mar, J. C. (2022). How does the structure of data impact cell–cell similarity? Evaluating how structural properties influence the performance of proximity metrics in single cell RNA-seq data. *Briefings in Bioinformatics*, 23(6):bbac387.

Weber, L. M. and Robinson, M. D. (2016). Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, 89(12):1084–1096.

Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., and Haque, A. (2022). An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1):68.

Willie, E., Rao, S., Figtree, G., Yang, J., de St Groth, B. F., McGuire, H., and Patrick, E. (2025). dioscri enables transferable prediction of clinical outcomes in multi-parameter cytometry data.

Willie, E., Yang, P., and Patrick, E. (2023). The impact of similarity metrics on cell-type clustering in highly multiplexed in situ imaging cytometry data. *Bioinformatics Advances*, 3(1).

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. (2019). Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):1–9.

Wu, H. et al. (2021). Spatiotemporal immune zonation of the human kidney. *Science*, 365(6460):1461–1466.

Xu, X., Su, J., Zhu, R., Li, K., Zhao, X., Fan, J., and Mao, F. (2025). From morphology to single-cell molecules: high-resolution 3d histology in biomedicine. *Molecular Cancer*, 24(1).

Yao, Z. et al. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241.

Yuan, M. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Zahedi, S. et al. (2024). Technical artifacts in spatial omics data. *Nature Biotechnology*, 42:234–245.

Zeng, Y. and Breheny, P. (2016). Overlapping group logistic regression with applications to genetic pathway selection. *Cancer Informatics*, 15(1):179–187.

Zhang, J. et al. (2023a). Single-cell analysis reveals the col11a1+ fibroblasts are cancer-specific fibroblasts that promote tumor progression. *Frontiers in Pharmacology*, 14:1121586.

Zhang, T., Warden, A. R., Li, Y., and Ding, X. (2020). Progress and applications of mass cytometry in sketching immune landscapes. *Clinical and Translational Medicine*, 10(6).

Zhang, Y., Chen, H., Mo, H., Hu, X., Gao, R., Zhao, Y., Liu, B., Niu, L., Sun, X., Yu, X., et al. (2023b). Single-cell analyses reveal key immune cell subsets associated with response to pd-l1 blockade in triple-negative breast cancer. *Cancer Cell*, 41(12):2166–2181.

Zhao, S., Song, J., and Ermon, S. (2017). Infovae: Information maximizing variational autoencoders.

Zhao, T., Chiang, Z. D., Morriss, J. W., LaFave, L. M., Murray, E. M., Del Priore, I., Meli, K., Lareau, C. A., Nadaf, N. M., Li, J., et al. (2022). Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature*, 601(7891):85–91.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049.

Zihni, C., Mills, C., Matter, K., and Balda, M. S. (2016). Tight junctions: from simple barriers to multifunctional molecular gates. *Nature Reviews Molecular Cell Biology*, 17(9):564–580.