Machine Learning Methods to Understand the Tumour Microenvironment in Pancreatic Cancer using Whole-Slide Multiplexed Imaging

Sophia Mengjia Li
sophiamjia.li@mail.utoronto.ca
Department of Cell & Systems Biology, Department of Computer Science
University of Toronto

Primary Supervisor: Dr. Kieran Campbell
Lunenfeld-Tanenbaum Research Institute, University of Toronto, Toronto, Ontario

Project Mentor: Shanza Ayub
Lunenfeld-Tanenbaum Research Institute, University of Toronto, Toronto, Ontario

A research proposal submitted in the co-sponsoring Department of Cell & Systems Biology and Department of Computer Science in partial fulfillment of the requirements for the award of Honours Bachelor of Science in Bioinformatics and Computational Biology of University of Toronto.

University of Toronto, Toronto, ON
May 2025

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is an aggressive and highly lethal cancer, accounting for over 90% of pancreatic cancer cases (Sarantis et al., 2020). Poor prognosis, with an average 5-year survival rate of less than 10%, is largely attributed to late diagnosis, genetic complexity, and metastatic potential that allows it to spread early and rapidly (Miller et al., 2016). The failure of traditional treatments such as chemotherapy, surgery, and radiation to significantly improve survival outline its place as one of the most lethal diseases (Sarantis et al., 2020).

The tumour microenvironment (TME) of PDAC plays a critical role in disease progression, therapeutic resistance, and patient outcomes (Uzunparmak & Sahin, 2019). It is a dense, complex, and spatially heterogeneous immunosuppressive network of cells and extracellular components characterized by extensive desmoplasia and the dominance of immunosuppressive cells such as Tregs, MDSCs, and TAMs (Sarantis et al., 2020). The disruptive nature of the TME creates significant obstacles for both conventional and novel treatments such as immune-based, vaccine-based, and adoptive T cell therapies such as CAR-T which are rendered with limited clinical benefits (Sarantis et al., 2020). Particularly, its highly heterogeneous nature leads to significant differences in tumour behaviour and treatment response among patients, making the TME an important target for clinical investigation (Sarantis et al., 2020).

Understanding the spatial organization of the TME is crucial towards understanding cell type interactions, tumour progression, and its impact on treatment response (Sarantis et al., 2020). Spatial patterns are closely linked to key tumour-immune interactions which cannot be quantified by genomic, transcriptomic, nor cellular composition data, highlighting the importance of spatial proteomics in its investigation (Janeiro et al., 2024). To analyze spatial organization and distribution, imaging mass cytometry (IMC) is widely adopted in PDAC research for high-throughput biomarker analysis and investigation, enabling the quantification of hundreds of proteins while preserving spatial context (Erreni et al., 2024). Traditional IMC workflows involve selecting regions of interest (ROIs) from whole tissue slides by extracting small tissue cores. These cores are assembled into tissue microarrays (TMAs), a high-throughput platform widely adopted in oncology to facilitate parallel analysis (Voduc et al., 2008). IMC is then performed on multiple cores simultaneously, producing images at single-cell resolution (Erreni et al., 2024). Cell segmentation of these images is a critical initial step in current workflows, serving as the foundation for subsequent downstream analysis (Erreni et al., 2024).

An important limitation of TMA-IMC is its potential to miss intratumoral heterogeneity (Khouja et al., 2010). The tissue cores sample only small, selected regions of the tumour tissue, limiting their ability to capture broader spatial heterogeneity and complexity motifs of the TME (Nagarkar et al., 2016; Ciesielska et al., 2021). More importantly, the ROIs are selected by pathologists to be representative and considered sufficient in capturing the full heterogeneity found in the whole tissue slide (Permuth-Wey et al., 2010). However, the reliability of capturing tumour markers with TMA-IMC depends strongly on the type of cancer and the complexity of the TME (Khouja et al., 2010). While adequate for detecting common features, TMAs can underestimate intra-tumour heterogeneity, especially for rare or uneven distributed features (Lee et al., 2019). This is particularly consequential in cancers with significant spatial variation and highly complex microenvironments such as PDAC, where focal or heterogeneous marker expression may be missed (Ciesielska et al., 2010). This introduces the possibility of sampling bias and can lead to underestimation or misrepresentation of true tumour diversity, with wide-reaching impacts on downstream analysis. While several studies have evaluated the relationship between the number of TMA cores and adequate representation of TME heterogeneity, these approaches are inherently limited by their restricted sampling (Lee et al., 2019). Thus, spatial profiling approaches that examine entire tumour sections is crucial to critically evaluate the true representativeness of TMA-IMC in evaluating TME heterogeneity (Sarantis et al., 2020).

Whole-slide imaging mass cytometry (WS-IMC) now enables comprehensive, multiplex spatial profiling across entire tumour sections, overcoming the limited field of view (FOV) of TMA-based approaches but at the cost of reduced resolution, making single-cell segmentation unfeasible (Motta et al., 2024). This introduces new analytical challenges, as current IMC analysis methods assume single-cell resolution and rely on cell segmentation as a first step, which is not feasible as the spatial resolution of WS-IMC data is insufficient to resolve individual cells and their boundaries (Motta et al., 2024). Consequently, there is a need to create novel computational tools to analyze WS-IMC images and extract meaningful biological insights from them. These novel analysis pipelines will be essential for capturing the full spatial heterogeneity present in tissue sections, providing critical insights in the investigation of highly complex and heterogeneous TMEs, such as those found in PDAC. A key future application will be the comparative assessment of data derived from WS-IMC versus TMA-IMC to assess the adequacy and representativeness of TMAs in assessing tumour heterogeneity. As current benchmarking methods rely on simulated sampling, leveraging WS-IMC as the more spatially resolved standard is critical towards revealing limitations of TMA-IMC that standard evaluation approaches cannot detect (Lee et al., 2019).

Literature Review

Overview of Multiplexed Imaging Technologies

Highly-multiplexed imaging (HMI) technologies have revolutionized cancer research by elucidating spatial tumour heterogeneity with significant consequences for patient outcomes (Bollhagen & Bodenmiller, 2024). These methods allow for the quantification of dozens to hundreds of protein markers within intact tissue sections, preserving spatial context at single-cell resolution (Bollhagen & Bodenmiller, 2024). This spatially resolved data supports the identification of cell phenotypes and biomarkers that would otherwise be obscured in bulk or dissociative analyses such as single-cell RNA sequencing (scRNA-seq) (Williams et al., 2022). As accumulating evidence suggests that cellular localization and spatial relationships strongly influence the functional status of cells in the TME, multiplexed imaging and spatial analysis methods have become widely adopted in both research and clinical settings (Jing et al., 2025).

HMI technologies are largely categorized by method of detection: optical microscope-based approaches such as PhenoCycler (formerly CODEX) and tissue-based cyclic immunofluorescence (t-CyCIF), and laser ablation-based approaches such as multiplexed ion beam imaging (MIBI) and imaging mass cytometry (IMC) (Semba & Ishimoto, 2024). Commercialized by Standard BioTools as the Hyperion product line, IMC utilizes a panel of antibodies tagged with non-naturally occurring heavy metal isotypes to stain tissue sections, where an ultraviolet laser ablates the tissue (Semba & Ishimoto, 2024). The resulting aerosol is ionized and analyzed by time-of-flight (TOF) mass spectrometry, allowing for the simultaneous detection of more than 40 protein targets in a single round of staining and imaging (Bollhagen & Bodenmiller, 2024). Images are acquired at single-cell resolution of approximately 1 µm per pixel, upon which cell segmentation is performed prior to downstream analysis. Prominently, the use of nonbiological isotopes facilitates high multiplexing capacity without spectral overlap and allows for background noise to be disregarded (Bollhagen & Bodenmiller, 2024).

Current IMC Data Analysis Workflow

As IMC produces highly complex images, several steps of analysis must be performed to extract biological meaning from the spatially resolved data (Milosevic, 2023). While a plethora of pipelines and tools exist to analyze this data, standard workflows typically involve steps of data preprocessing, cell segmentation, and downstream analysis (Milosevic, 2023).

To address common artifacts such as background noise, low signal spillover (crosstalk), and back effects, preprocessing is essential to ensure data quality and comparability (Milosevic, 2023). Popular tools such as CATALYST address crosstalk, while MAUI and IMC-Denoise manage other preprocessing tasks (Milosevic, 2023).

Cell segmentation is the most critical step in IMC analysis workflows, as the accuracy of downstream analyses heavily relies on segmentation quality (Milosevic, 2023). Leading methods include DeepCell, which employs deep learning, and workflows combining Ilastik and CellProfiler for pixel-based classification (Milosevic, 2023).

Many downstream tasks exist to investigate the spatially resolved data, most commonly including cell clustering, phenotyping, and dimensionality reduction prior to differential and spatial analysis (Milosevic, 2023). Integrated platforms such as CytoMAP and HistoCAT offer comprehensive functionality for these tasks, including clustering and dimensionality reduction (Milosevic, 2023). For cell phenotyping, specialized tools like Astir provide automated single-cell annotation, reflecting the abundance of available approaches (Milosevic, 2023).

Whole-slide multiplexing imaging

Recent advances in whole-slide multiplexed imaging have enabled comprehensive spatial profiling of tissue sections, but analyzing these data presents unique challenges. Traditional methods, such as ASHLAR, reconstruct whole slides by stitching together smaller FOVs using immunofluorescence as a spatial reference (Liu et al., 2023). However, these approaches are labour intensive, error-prone, and not widely accessible (Liu et al., 2023). Historically, IMC was limited by technical constraints that restricted ROI size (Bollhagen & Bodenmiller, 2024). The introduction of the Hyperion XTi system by Standard BioTools marks a significant advancement, offering the first commercial solution for WS-IMC (Motta et al., 2024). This system's tissue mode (TM) enables imaging of entire tumor slides, albeit at a reduced resolution (5 µm) compared to cell mode (CM) (1 µm). While this trade-off makes single-cell resolution unfeasible, it allows for highly multiplexed spatial profiling across entire tissue sections, surpassing the coverage of tissue microarray (TMA) approaches (Motta et al., 2024).

The shift to lower resolution in WS-IMC introduces new analytical challenges. Current methods of analyzing IMC data assume single-cell resolution and rely on accurate cell segmentation as a first step, something that is not currently possible for reduced resolution whole-slide images. This necessitates the development of novel computational tools that can extract meaningful features from WS-IMC images. Especially with the importance of spatial distribution and tumour heterogeneity in highly complex cancers like PDAC, these tools are essential to gaining new insights into their TMEs. WS-IMC has further potential for the validation of the adequacy, reliability, and representativeness of TMA-IMC, providing whole-slide ground truth for the true heterogeneity of the tumour.

Machine Learning in spatial tissue analysis

Machine learning methods are widely adopted in processing, analyzing, and interpreting high-dimensional, spatially resolved protein data generated by MS-based and imaging-based approaches (Mou et al., 2022). For IMC analysis pipelines, deep learning approaches have become increasingly popular with state-of-the-art performance, beginning to dominate single-cell segmentation, feature extraction, and classification tasks over classical models for pixel-level classification (Mou et al., 2022). Several popular open-source tools leverage deep learning approaches, such as DeepCell, which makes use of a Panoptic backbone, and Astir, which uses a variational autoencoder (Milosevic, 2023). Other tools such as Squidpy integrate several machine-learning-based steps in comprehensive analytical pipelines (Milosevic, 2023).

There is considerable potential to adapt machine learning methods developed for TMA-IMC to WS-IMC (Magness et al., 2024). As existing feature extraction approaches for TMA-IMC depend on single-cell resolution and therefore are not applicable for WS-IMC, novel computer vision techniques must be developed (Erreni et al., 2024). Established tools for downstream tasks such as clustering and spatial analysis can be applied upon these features, enabling more comprehensive tissue architecture mapping and biomarker discovery which is highly valuable for advancing spatial proteomics research (Motta et al., 2024).

Objectives

The primary objective of this project is to create novel computational tools that can extract features from WS-IMC images and relate them to clinical variables in cancer to offer insights into the relationship of tissue context with PDAC. This is described by several objectives:

1. Design machine learning tools capable of extracting meaningful spatial features from lower-resolution WS-IMC images.
2. Relate the extracted features to clinical variables using established TMA-IMC workflows.
3. Compare the features extracted from matched WS-IMC and TMA-IMC images.

Methods

Dataset and Image Preprocessing

Matched WS-IMC and TMA-IMC images from 22 PDAC patients were obtained by Ferris Nowlan and Noor Shakfa on behalf of the Jackson Lab. Immune, epithelial, and CAF panels were aggregated in the whole-slide image panel, where they performed spillover correctional and blank acquisition removal.

Raw IMC data (MCD files) will be exported as multi-channel TIFFs and loaded using the tifffile package (Gohlke, 2025). Pixel-level intensities will be extracted using NumPy (Harris et al., 2020), winsorized with SciPy (Virtanen et al., 2020), then normalized via min-max scaling with scikit-learn (Pedregosa et al., 2011). An AnnData object will be created with normalized intensities for standard preprocessing at the pixel level using ScanPy (Wolf et al., 2018).

Unsupervised CNN Training and Feature Extraction

A model based on a convolutional neural network (CNN) architecture will be trained for the novel application to WS-IMC data, utilizing patch-based self-supervised contrastive learning for feature extraction from highly multiplexed images. Preprocessed whole-slide images will be partitioned into 1000 µm^2 (200 px^2) patches using a sliding window approach, replicating the average core size of the TMA-IMC images. Using PyTorch Lightning, a contrastive learning framework such as SimCLR, MoCo, or BYOL will be implemented to train a standard CNN backbone such as ResNet or EfficientNet (Falcon et al., 2019). Following training, the CNN will be applied to all patches across the dataset to generate patch-level feature embeddings

Downstream Analysis

Patch-level embeddings generated by the trained CNN will serve as input for downstream analyses. Unsupervised clustering of the embeddings will be performed, where the clusters obtained will be used to find spatial patterns. Furthermore, aggregated patch-level features at the whole-slide and patient level will be integrated with clinical metadata, upon which correlation tests and group comparisons will be performed to identify associations between image-derived features and clinical variables.

Comparative Analysis of TMA-IMC and WS-IMC

To systemically investigate and compare the information obtained from TMA-IMC and WS-IMC, several quantitative approaches will be employed for measuring information gain.

1. Calculate centroids of patch-level feature clusters from WS-IMC images and cell-level clusters from TMA-IMC images and compare them using similarity metrics such as Euclidean distance.
2. Compute similarity metrics between marker correlation matrices derived from TMA-IMC and WS-IMC to assess concordance of marker relationships between TMA cores and whole-slide patches.
3. Quantify information gain per sampled WS-IMC patch based on changes in total variation, cluster composition, and average marker expression.

The results of these analyses will be used to construct a curve of diminishing returns, illustrating the relationship between the number of TMA cores and the information captured by WS-IMC as ground-truth.

Expected Results

Anticipated Findings

It is anticipated that the application of novel computational tools WS-IMC data will reveal a more comprehensive view of the spatial heterogeneity and tissue context within PDAC tumours than is possible with TMA-IMC approaches. These tools are expected to extract clinically relevant features from WS-IMC data, providing new insights into how the full complexity of the tumour microenvironment relates to clinical variables. This may uncover spatial patterns and associations that are not detectable using TMA-based workflows, thereby advancing our understanding of tumour biology and its clinical implications.

It is also expected that the project will clarify the extent to which TMA-based approaches are sufficient for capturing tumour heterogeneity. The adequacy of TMA-IMC is likely to depend on the spatial complexity of the tumour microenvironment, with a higher number of TMA cores required to achieve representative sampling in more heterogeneous tumours. By directly comparing TMA-IMC to WS-IMC and using whole-slide data as the ground truth, this project will quantitatively describe the relationship between TMA representativeness and the number of cores sampled. These findings will inform best practices for spatial tissue analysis and guide future study designs in cancer research.

Potential Impact

By achieving these objectives, this project will deliver the first dedicated computational tool for the analysis of WS-IMC data. The methods and workflows developed will establish a foundational framework for interpreting WS-IMC images, enabling comprehensive spatial profiling of the TME beyond the limitations of traditional ROI or TMA approaches. This work will set the stage for future research leveraging WS-IMC in cancer studies and support the integration of whole-slide spatial data into broader cancer research and clinical applications.

Additionally, by directly comparing WS-IMC with TMA-IMC and evaluating the adequacy and representativeness of TMA-selected regions, this project will provide important validation for the TMA-IMC approach. If TMA-IMC is shown to reliably capture the key features of the whole tissue, it will reinforce its continued use in research and clinical settings. Conversely, if significant discrepancies are found, the results will highlight the need for more comprehensive spatial profiling methods. In either case, these findings will inform best practices for spatial tissue analysis and guide future study designs in cancer research.

References

Bollhagen, A., Engler, S., & Bodenmiller, B. (2024). Highly multiplexed tissue imaging in precision oncology and translational cancer research. Cancer Discovery, 14(6), 1082–1099. https://doi.org/10.1158/2159-8290.cd-23-1165

Ciesielska, U., Nowinska, K., Piotrowska, A., Pula, B., Paprocka, M., Krecicki, T., Podhorska-Okolow, M., & Dziegiel, P. (2021). Comparison of TMA technique and routine whole slide analysis in evaluation of proliferative markers expression in laryngeal squamous cell cancer. In Vivo, 35(6), 3264–3272. https://doi.org/10.21873/invivo.12628

Erreni, M., Fumagalli, M. R., Zanini, D., Candiello, E., Tiberi, G., Parente, R., D'Anna, R., Magrini, E., Marchesi, F., Cappello, P., & Doni, A. (2024). Multiplexed imaging mass cytometry analysis in preclinical models of pancreatic cancer. International Journal of Molecular Sciences, 25(3), 1389. https://doi.org/10.3390/ijms25031389

Falcon, W. A., & The PyTorch Lightning team. (2019). PyTorch Lightning. GitHub repository. https://github.com/Lightning-AI/lightning

Gohlke, C. (2025). Tifffile: A Python library to read and write TIFF files. Zenodo. https://doi.org/10.5281/zenodo.15522340

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Janeiro, A. L., Wong, E. M., Jiménez-Sánchez, D., Solorzano, C. O., Lozano, M. D., Teijeira, A., Schalper, K. A., Melero, I., & Andrea, C. E. (2024). Spatially resolved tissue imaging to analyze the tumor immune microenvironment: Beyond cell-type densities. Frontiers in Immunology, 15, Article 11149121. https://doi.org/10.3389/fimmu.2024.11149121

Jing, S.-y., Wang, H.-q., Lin, P., Yuan, J., Tang, Z.-x., & Li, H. (2025). Quantifying and interpreting biologically meaningful spatial signatures within tumor microenvironments. npj Precision Oncology, 9, Article 68. https://doi.org/10.1038/s41698-025-00857-1

Lee, A. T. J., Jones, R. L., Heseltine, K., Thway, K., Shipley, J., Henshaw, R., Huang, P. H., & Jones, R. L. (2019). The adequacy of tissue microarrays in the assessment of inter- and intra-tumoural heterogeneity of infiltrating lymphocyte burden in leiomyosarcoma. Scientific Reports, 9, 9646. https://doi.org/10.1038/s41598-019-50888-5

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., & Angelo, M. (2023). Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering. Nature Communications, 14, Article 4618. https://doi.org/10.1038/s41467-023-40068-5

Magness, A., Colliver, E., Enfield, K. S. S., Lee, C., Shimato, M., Daly, E., Moore, D. A., Sivakumar, M., Valand, K., Levi, D., Hiley, C. T., Hobson, P. S., van Maldegem, F., Reading, J. L., Quezada, S. A., Downward, J., Sahai, E., Swanton, C., & Angelova, M. (2024). Deep cell phenotyping and spatial analysis of multiplexed imaging with TRACERx-PHLEX. Nature Communications, 15, Article 5135. https://doi.org/10.1038/s41467-024-48870-5

Miller, K. D., Siegel, R. L., Lin, C. C., Mariotto, A., Kramer, J. L., Rowland, J., Stein, K., Alteri, R., & Jemal, A. (2016). Cancer treatment and survivorship statistics, 2016. CA: A Cancer Journal for Clinicians, 66(4), 271–289. https://doi.org/10.3322/caac.21349

Milosevic, V. (2023). Different approaches to Imaging Mass Cytometry data analysis. Bioinformatics Advances, 3(1), vbad044. https://doi.org/10.1093/bioadv/vbad044

Motta, V., Raza, Q., Zabinyakov, N., Pfister, T., Parsotam, N., Howell, D., Lim, L., & Loh, C. (2024). Whole slide imaging modes for Imaging Mass Cytometry reveal cellular diversity of the tumor immune microenvironment in mouse glioblastoma. The Journal of Immunology, 212(1 Supplement), 0777_5471. https://doi.org/10.4049/jimmunol.212.Supplement.0777_5471

Mou, M., Pan, Z., Lu, M., Sun, H., Wang, Y., Luo, Y., & Zhu, F. (2022). Application of machine learning in spatial proteomics. Journal of Chemical Information and Modeling, 62(24), 6092–6107. https://doi.org/10.1021/acs.jcim.2c01161

Nagarkar, D. B., Mercan, E., Weaver, D. L., Brunyé, T. T., Carney, P. A., Rendi, M. H., Beck, A. H., Frederick, P. D., Shapiro, L. G., & Elmore, J. G. (2016). Region of interest identification and diagnostic agreement in breast pathology. Modern Pathology, 29(9), 1004–1011. https://doi.org/10.1038/modpathol.2016.85

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Sarantis, P., Koustas, E., Papadimitropoulou, A., Papavassiliou, A. G., & Karamouzis, M. V. (2020). Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. World Journal of Gastrointestinal Oncology, 12(2), 173–181. https://doi.org/10.4251/wjgo.v12.i2.173

Semba, T., & Ishimoto, T. (2024). Spatial analysis by current multiplexed imaging technologies for the molecular characterisation of cancer tissues. British Journal of Cancer, 131(14), 1737–1747. https://doi.org/10.1038/s41416-024-02882-6

Uzunparmak, B., & Sahin, I. H. (2019). Pancreatic cancer microenvironment: a current dilemma. Clinical and Translational Medicine, 8(1), 2. https://doi.org/10.1186/s40169-019-0221-1

Virtanen, P., Gommers, R., Oliphant, T.E. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2

Voduc, D., Kenney, C., & Nielsen, T. O. (2008). Tissue microarrays in clinical oncology. Seminars in Radiation Oncology, 18(2), 89–97. https://doi.org/10.1016/j.semradonc.2007.10.006

Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., & Haque, A. (2022). An introduction to spatial transcriptomics for biomedical research. Genome Medicine, 14(1), 68. https://doi.org/10.1186/s13073-022-01075-1

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 19(1), 15. https://doi.org/10.1186/s13059-017-1382-0