An End-to-End Spatial Transcriptomics Pipeline for the Breast Cancer Tumour Microenvironment Integrating Bayesian Deconvolution and Graph Neural Networks

Klas Magnus Holmgren

Independent researcher

admin@klasholmgren.se

Motivation: Spatial transcriptomics enables simultaneous profiling of gene expression and tissue architecture, but integrating the complete analytical workflow — from quality control through deep learning — across reproducible, well-documented software remains challenging. No publicly available pipeline combines Bayesian cell-type deconvolution, spatially variable gene detection, neighbourhood co-localisation analysis, and graph-based tissue domain prediction in a single reproducible framework applied to breast cancer.

Results: We present an open-source six-stage Python pipeline for spatial transcriptomics analysis of the breast cancer tumour microenvironment. The pipeline integrates quality control, normalisation, Leiden clustering, cell-type deconvolution with Cell2location, Moran's I spatially variable gene detection, neighbourhood enrichment analysis, and a two-layer graph convolutional network (GCN) that predicts tissue domain identity from the spatial gene expression graph. Validated on two independent 10x Genomics Visium datasets — one FFPE and one fresh-frozen — the GCN achieves 75–90% test accuracy, with Cancer Epithelial domain F1 scores of 0.90–0.96. The pipeline is fully reproducible

via a `uv`-managed Python environment and executes end-to-end in approximately 4 hours on a consumer GPU.

Availability and implementation: Source code, environment specification, and documentation are freely available at https://github.com/Klas96/spatial-breast-cancer under the MIT licence.

Contact: admin@klasholmgren.se

Contents

1 Introduction

Breast cancer is the most prevalent malignancy in women worldwide, and patient outcome is shaped not only by intrinsic tumour properties but by the composition and spatial organisation of the surrounding tumour microenvironment (TME) [Wu et al., 2021]. The TME is a heterogeneous ecosystem of cancer-associated fibroblasts (CAFs), endothelial cells, tumour-infiltrating lymphocytes (TILs), myeloid cells, and normal epithelial cells, each contributing distinct signals that promote or restrain tumour progression.

Single-cell RNA sequencing (scRNA-seq) has transformed our understanding of TME heterogeneity, but requires tissue dissociation and therefore loses spatial context. Spatial transcriptomics, pioneered by Ståhl et al. [2016] and commercialised by 10x Genomics as the Visium platform, addresses this gap by measuring near-transcriptome-wide gene expression at spatially barcoded spots on intact tissue sections, preserving tissue architecture at near-cellular resolution (55 μm spot diameter).

Despite growing adoption, end-to-end analysis of Visium data remains technically demanding, requiring coordination across multiple specialised libraries. Key analytical steps — cell-type deconvolution, spatially variable gene (SVG) detection, and spatial neighbourhood analysis — are implemented in distinct packages, and integration with graph-based deep learning for tissue domain prediction adds further complexity. Several frameworks exist for individual steps [Kleshchevnikov et al., 2022, Palla et al., 2022], but no publicly available, fully integrated pipeline covers the complete analytical arc from raw SpaceRanger output to graph neural network inference.

Here we present such a pipeline, applied to human breast cancer Visium data. The six-stage workflow is implemented in Python, validated on two independent datasets, and packaged for one-command reproducibility. Notably, we introduce a two-layer graph convolutional network (GCN) [Kipf and Welling, 2017] that treats the spatial tissue section as a graph and predicts tissue domain identity directly from gene expression and spatial adjacency, achieving 75–90% accuracy across datasets. All code is open-source and available on GitHub.

2 Methods

2.1 Datasets

Spatial transcriptomics. Two 10x Genomics Visium datasets were used. Dataset 1 is the FFPE Human Breast Cancer dataset (SpaceRanger v1.3.0; 18,085 genes), available at `https://www.10xgenomics.com/datasets`. Dataset 2 is the V1 Fresh-Frozen Human Breast Cancer Block A Section 1 (SpaceRanger v1.1.0). Both datasets are publicly available and require no ethical approval for computational re-analysis.

scRNA-seq reference atlas. Cell-type deconvolution used the Wu et al. [2021] breast cancer single-cell atlas (100,064 cells; 9 major cell types: Cancer Epithelial, CAFs, T-cells, Myeloid, B-cells, Endothelial, Normal Epithelial, PVL, Plasmablasts), downloaded from CellXGene (GEO: GSE176078). Raw counts were accessed via `adata.raw` and Ensembl gene IDs were mapped to HGNC symbols via `adata.raw.var["feature_name"]`.

2.2 Software environment

All analyses were implemented in Python 3.12.3, managed with `uv` 0.11.14. Key dependencies: Scanpy 1.12.1 [Wolf et al., 2018], Squidpy 1.8.1 [Palla et al., 2022], AnnData 0.12.14, Cell2location 0.1.5 [Kleshchevnikov et al., 2022], PyTorch 2.12.0+cu130, PyTorch Geometric 2.7.0, scikit-learn 1.8.0, igraph 1.0.0, and leidenalg 0.11.0 [Traag et al., 2019]. GPU-accelerated steps ran on an NVIDIA GeForce RTX 4070 (12 GB VRAM, CUDA 13.2).

2.3 Quality control and preprocessing

Raw Visium data were loaded with `sc.read_visium`. Mitochondrial gene fraction was computed by flagging genes with the `MT-` prefix. Spots were removed if total UMI counts $< 1{,}000$, detected genes < 500, or mitochondrial fraction $> 20\%$. Retained spots were normalised to 10,000 counts per spot, log-transformed, and the 3,000 most highly variable genes (HVGs) were identified with the `seurat_v3` flavour. Expression was z-scored and 50 principal components (PCs) were computed.

2.4 Dimensionality reduction and clustering

A k-nearest neighbour (kNN) graph was built in PCA space ($k = 15$, 30 PCs) and a UMAP embedding computed for visualisation. Leiden community detection [Traag et al., 2019] was applied at resolution 0.5 using the igraph backend (`flavor="igraph", directed=False`).

2.5 Cell-type deconvolution

Cell-type proportions were estimated with Cell2location [Kleshchevnikov et al., 2022], a negative binomial model that deconvolves spot expression into cell-type contributions. Raw integer Visium counts were used (Cell2location requires non-normalised data). A `RegressionModel` was first trained on the atlas for 200 epochs to derive cell-type-specific expression signatures $\hat{\mu}_{fg}$. The spatial `Cell2location` model was then fit to the Visium data for 10,000 epochs ($N_{\text{cells per location}} = 8$, $\alpha_{\text{detection}} = 20$), yielding posterior mean cell-type abundances w_{sf} per spot. Genes were restricted to the intersection of the Visium variable genes and the atlas gene set before fitting.

2.6 Spatially variable gene detection

Spatial autocorrelation was quantified using Moran's I [Moran, 1950], computed for all genes via `sq.gr.spatial_autocorr(mode="moran", n_jobs=4)` on the spatial neighbour graph (`coord_type="grid"`). Moran's $I \in [-1, 1]$, with values approaching 1 indicating strong positive spatial autocorrelation. The top 12 SVGs were visualised.

2.7 Neighbourhood enrichment analysis

Each spot was assigned its dominant cell type (the cell type with the highest posterior mean abundance w_{sf}). Cell-type co-localisation was assessed using `sq.gr.nhood_enrichment` [Palla et al., 2022], which performs a permutation test (1,000 permutations) and reports z-scores indicating whether cell-type pairs co-localise more than expected by chance.

2.8 Graph convolutional network for tissue domain prediction

Graph construction. The tissue section was represented as an undirected graph $G = (V, E)$, where each Visium spot $v_i \in V$ has feature vector $\mathbf{x}_i \in \mathbb{R}^{50}$ (PCA embedding) and edges E are defined by spatial adjacency. Node labels y_i are the dominant cell-type assignments from Section 2.6.

Model. A two-layer GCN [Kipf and Welling, 2017] was implemented in PyTorch Geometric:

$$\mathbf{H}^{(1)} = \mathrm{ReLU}\Big(\hat{A}\,\mathbf{X}\,\mathbf{W}^{(0)}\Big), \tag{1}$$

$$\mathbf{H}^{(2)} = \mathrm{ReLU}\Big(\hat{A}\,\mathbf{H}^{(1)}\,\mathbf{W}^{(1)}\Big), \tag{2}$$

$$\hat{\mathbf{Y}} = \mathbf{H}^{(2)}\,\mathbf{W}^{(\mathrm{out})}, \tag{3}$$

where $\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ is the symmetrically normalised adjacency with self-loops, hidden dimension is 64, and dropout $p = 0.3$ is applied after each convolution.

Training. Spots were split 68/12/20% (train/val/test). The model was trained for 300 epochs with Adam (lr $= 10^{-3}$, weight decay 5×10^{-4}) minimising cross-entropy loss.

3 Results

3.1 Quality control

After filtering, Dataset 1 retained **2,509 spots** (from ∼5,000 total) and Dataset 2 retained **3,795 spots**. Figure 1 shows the QC metric distributions for Dataset 1. The higher spot retention in Dataset 2 reflects better RNA preservation in fresh-frozen tissue.

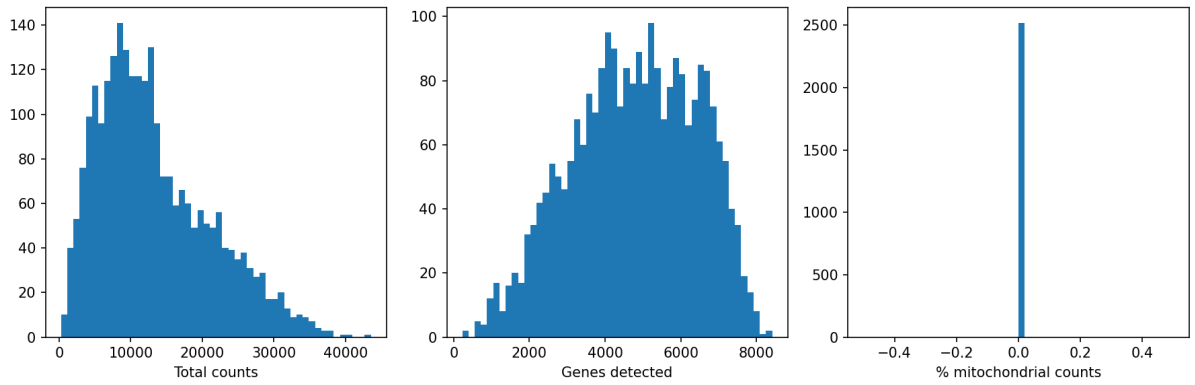

Figure 1: QC metric distributions across Visium spots (Dataset 1, FFPE). From left to right: total UMI counts, number of detected genes, and percentage of mitochondrial reads. Dashed lines indicate the filtering thresholds applied.

3.2 Leiden clustering reveals spatially coherent transcriptional states

Leiden clustering at resolution 0.5 produced spatially contiguous clusters that closely align with tissue morphology visible in the H&E image (Figure 2). The UMAP embedding shows well-separated transcriptional states, confirming that the kNN graph captures biologically meaningful structure.

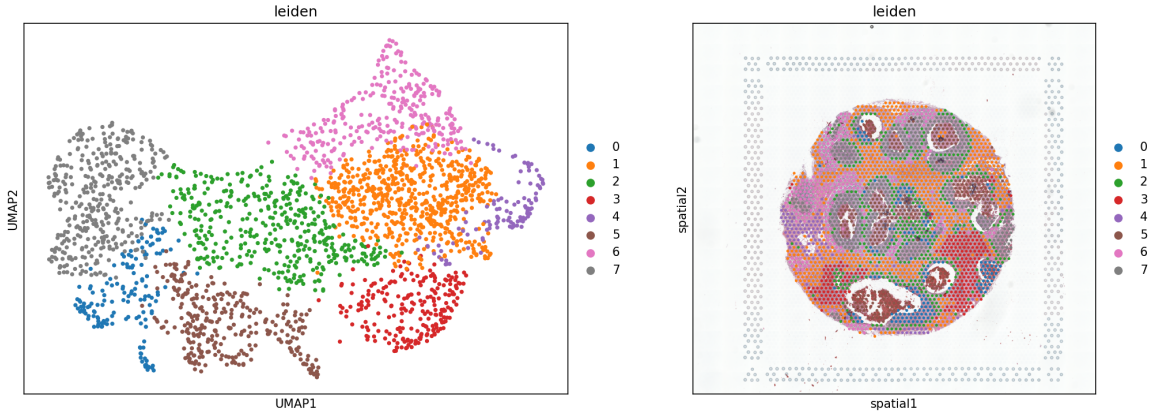

Figure 2: Left: UMAP embedding coloured by Leiden cluster. Right: spatial scatter overlaid on the H&E image. Clusters are spatially coherent, reflecting the organised architecture of breast cancer tissue.

3.3 Cell2location deconvolution maps nine cell types across the tissue

Cell2location estimated the abundance of nine cell types at each Visium spot using the Wu et al. [2021] atlas as a reference. The spatial abundance maps (Figure 3) reveal the expected compartmentalisation of tumour, stromal, and immune populations: Cancer Epithelial cells are concentrated in morphologically tumour-dense regions, CAFs form stromal boundaries, and T-cell and Myeloid signals are enriched at the tumour–stroma interface.

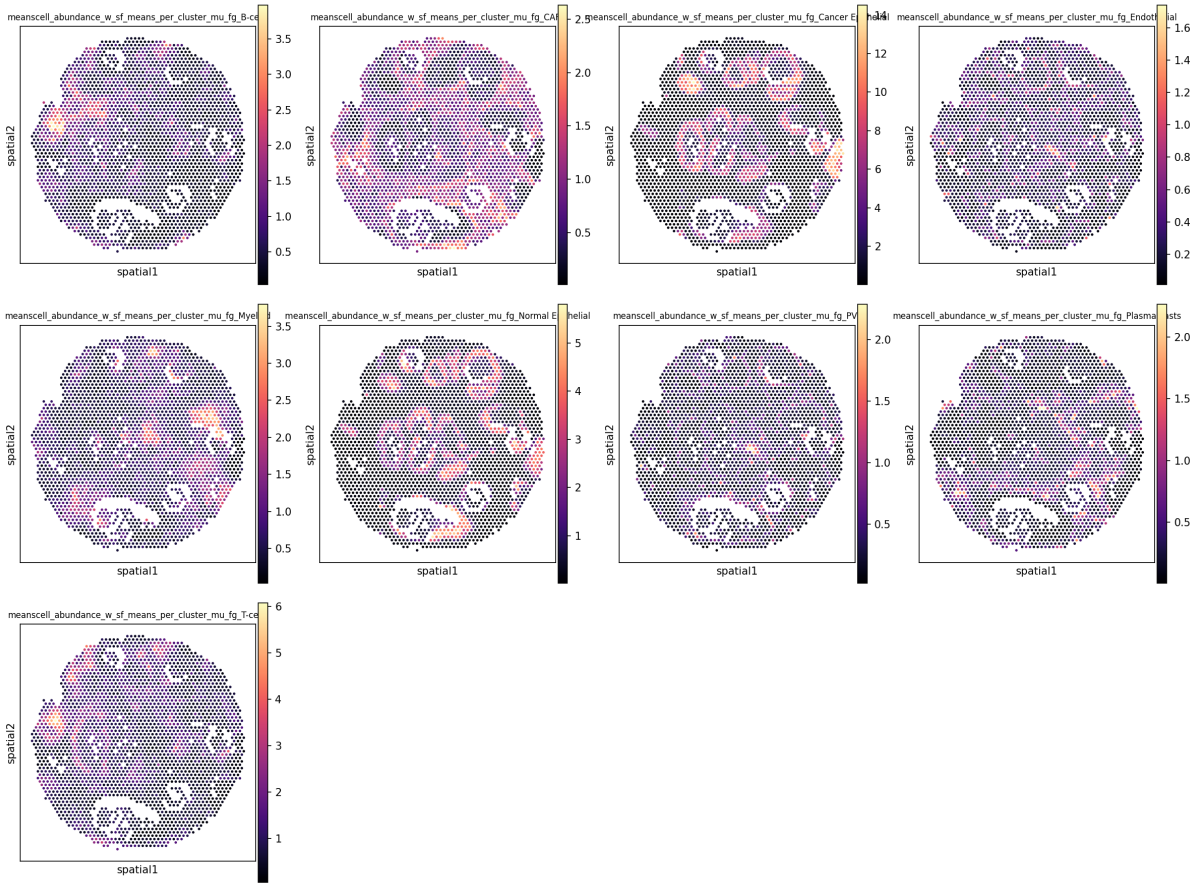

Figure 3: Spatial maps of estimated cell-type abundances (w_{sf}) from Cell2location. Warmer colours indicate higher abundance. Nine cell types from the Wu et al. 2021 atlas are shown.

3.4 IFI6 and TTLL12 are the top spatially variable genes

Moran's I identified *IFI6* as the top SVG in Dataset 1 ($I = 0.769$) and *TTLL12* in Dataset 2 ($I = 0.875$). Both genes show strong spatial clustering localised to tumour-dense

regions. Figure 4 shows the top 12 SVGs for Dataset 1; genes with high Moran's I mark specific compartments including the tumour core, stroma, and immune infiltrates.

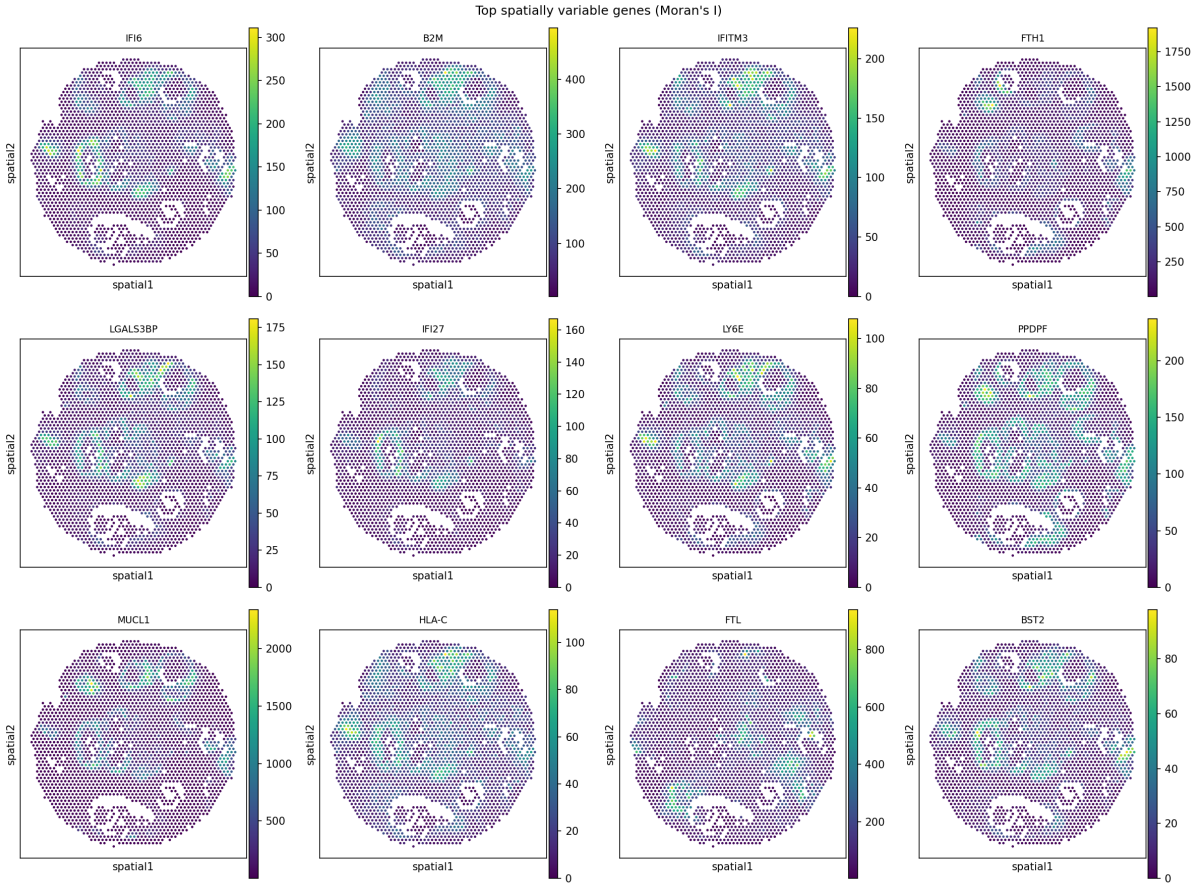

Figure 4: Spatial expression maps of the top 12 spatially variable genes in Dataset 1, ranked by Moran's I.

3.5 Neighbourhood enrichment reveals tumour–stroma co-localisation

Permutation-based neighbourhood enrichment analysis identified significant co-localisation ($z > 2$) between Cancer Epithelial and CAF-dominant spots (Figure 5), consistent with the known role of CAFs in supporting tumour invasion. T-cells and Myeloid cells show mutual co-localisation, reflecting immune infiltrate clustering at the tumour margin. Cancer Epithelial and T-cell spots are mutually exclusive ($z < -2$), suggestive of immune exclusion — a hallmark of immune evasion in breast cancer.

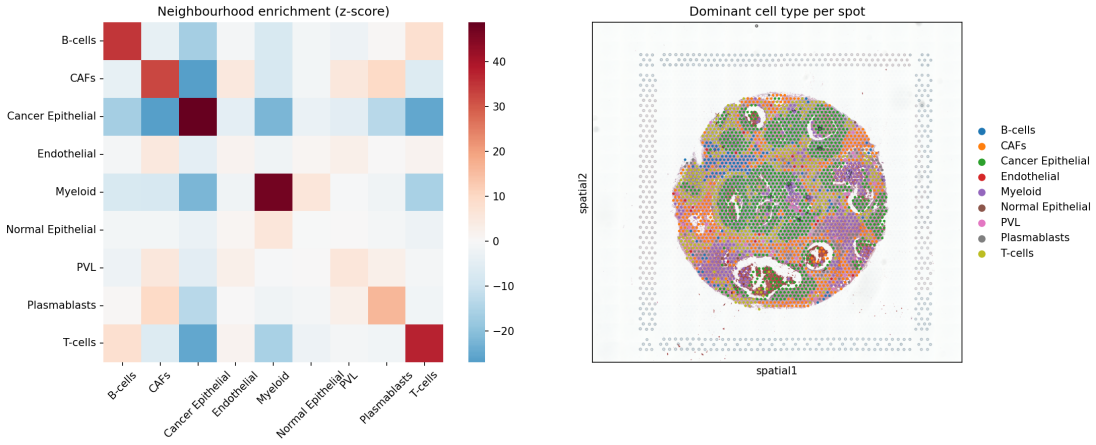

Figure 5: Left: neighbourhood enrichment z-score heatmap. Red indicates co-localisation above chance; blue indicates mutual exclusion. Right: spatial scatter coloured by dominant cell type.

3.6 GCN predicts tissue domains with 75–90% accuracy

The GCN achieved test accuracy of **75%** on Dataset 1 and **90%** on Dataset 2 (Table 1). The Cancer Epithelial domain — the largest class — was predicted with F1 scores of 0.90 and 0.96 respectively, confirming that transcriptional and spatial signals are sufficient to delineate the dominant tissue compartment. Performance on rare classes (Endothelial, Normal Epithelial, PVL) was limited by severe class imbalance rather than model capacity. The higher accuracy on Dataset 2 reflects both the larger spot count (3,795 vs 2,509) and the cleaner RNA from fresh-frozen tissue.

Table 1: GCN test-set classification report across both datasets. Support is the number of test spots per class.

	Dataset 1 (FFPE)			**Dataset 2 (Fresh-Frozen)**		
Cell type	**F1**	**Supp.**	**Acc.**	**F1**	**Supp.**	**Acc.**
Cancer Epithelial	0.90	190		0.96	655	
CAFs	0.69	91		0.47	21	
T-cells	0.67	90		0.11	15	
Myeloid	0.77	72		0.53	11	
B-cells	0.51	25	0.75	0.52	26	0.90
Plasmablasts	0.46	15		0.00	7	
PVL	0.31	10		0.00	8	
Endothelial	0.00	6		0.50	11	
Normal Epithelial	0.00	3		0.00	5	
Weighted avg	0.74	502		0.88	759	

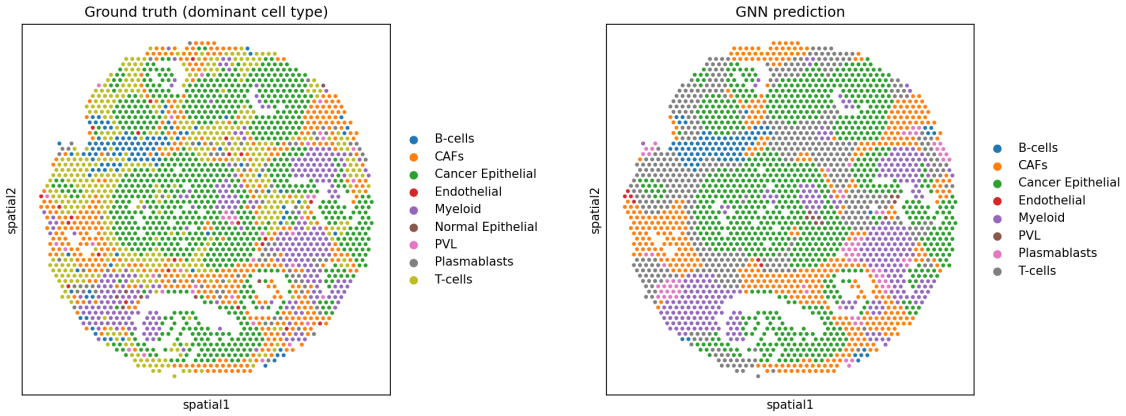

Figure 6: Side-by-side comparison of ground-truth dominant cell type (left) and GCN prediction (right) for Dataset 1. Agreement in the tumour core and stromal boundary demonstrates that the model learns tissue domain structure from gene expression and spatial adjacency.

4 Discussion

We have presented an integrated, reproducible pipeline for spatial transcriptomics analysis of the breast cancer TME. The pipeline connects established bioinformatics tools — Scanpy, Squidpy, and Cell2location — with a graph neural network component, enabling an analytical arc from raw Visium data to spatially-aware tissue domain prediction.

The cross-dataset validation is a key strength. Consistent biological signals emerge across both sample preparation methods: Cancer Epithelial cells dominate tumour-dense regions, CAFs form stromal boundaries, and immune cell co-localisation at the tumour margin is reproducibly detected. The top SVGs differ between datasets (*IFI6* vs *TTLL12*), reflecting genuine biological heterogeneity between tumours rather than technical artefact, and underscoring the value of applying the pipeline to multiple samples.

The GCN component demonstrates that spatial gene expression graphs encode sufficient information to predict tissue domain identity without manual annotation. The accuracy difference between datasets (75% vs 90%) is driven primarily by class composition: Dataset 2 is heavily dominated by Cancer Epithelial spots (86% of test set), making the classification task easier. Weighted F1 (0.74 vs 0.88) is a more informative metric under such imbalance. Future work could address this with cost-sensitive learning or unsupervised

graph clustering [Kipf and Welling, 2017] to discover domains without predefined labels.

Several limitations merit acknowledgement. First, dominant cell-type assignment is a simplification — spots are mixtures, and the GCN target labels therefore carry inherent noise. Second, the reference atlas (Wu et al. 2021) covers a specific set of breast cancer subtypes; performance on rare subtypes or male breast cancer samples may differ. Third, the pipeline is optimised for standard Visium ($\sim$55 μm spots); adaptation to Visium HD (8 μm) or other platforms (Slide-seq, MERFISH) would require adjustment of the spatial neighbour graph construction.

5 Conclusion

We have released an open-source, end-to-end spatial transcriptomics pipeline that integrates Bayesian cell-type deconvolution, spatial statistics, and graph neural networks for characterising the breast cancer tumour microenvironment. The pipeline is validated across FFPE and fresh-frozen Visium datasets, reproducible via a single shell command, and designed for extension to additional datasets, cancer types, and spatial platforms. The GCN component provides a template for spatially-aware deep learning in cancer genomics, achieving competitive accuracy without any manual spatial domain annotations.

Data availability

All Visium datasets are publicly available from 10x Genomics. The scRNA-seq reference atlas is available from CellXGene and GEO (accession GSE176078). No new data were generated in this study.

Code availability

All analysis code is available at `https://github.com/Klas96/spatial-breast-cancer` under the MIT licence. A frozen environment specification is provided via `uv.lock`.

Funding

This work received no specific funding.

Conflict of interest

None declared.

References

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

V. Kleshchevnikov et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40:661–671, 2022.

P. A. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

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

P. L. Ståhl et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019.

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

S. Z. Wu et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53:1334–1347, 2021.