

Understanding the spatial determinants of the Oxford Classic prognostic signature for high-grade serous ovarian cancer

Received: 27 October 2025

Accepted: 27 May 2026

Cite this article as: Stihl, A., Yau, C. Understanding the spatial determinants of the Oxford Classic prognostic signature for high-grade serous ovarian cancer. *Commun Med* (2026). <https://doi.org/10.1038/s43856-026-01708-1>

Alexandru Stihl & Christopher Yau

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Understanding the spatial determinants of the Oxford Classic prognostic signature for high-grade serous ovarian cancer

Alexandru Stihi¹ and Christopher Yau^{1,2,*}

¹Nuffield Department for Women's and Reproductive Health, University of Oxford, Oxford, UK.

²Health Data Research UK, London, UK.

*Corresponding author: christopher.yau@wrh.ox.ac.uk

May 22, 2026

Abstract

Background: The Oxford Classic (OxC) prognostic signature classifies high-grade serous ovarian cancer (HGSOC) into five transcriptional programs, with epithelial-to-mesenchymal transition (EMT) marking poor prognosis. While successful in bulk transcriptomics, the spatial organisation of these programs within the tumour microenvironment remains unexplored. **Methods:** We developed the Signature-guided Zero-inflated Beta Variational Autoencoder (Sig-ZIB-VAE), a deep learning deconvolution method tailored for spatial transcriptomics data, and applied it to a large-scale HGSOC cohort comprising 94 tumours to quantify spatial cellular organisation. Prognostic significance was assessed using penalised Cox proportional hazards regression integrating clinical, molecular, and spatial features. **Results:** Here we show that EMT cells form dense homotypic clusters broadly depleted from stromal and immune neighbourhoods, yet maintain selective monocyte co-localisation at cluster boundaries. EMT-high tumours display enhanced spatial reorganisation characterised by increased clustering and connectivity, forming locally concentrated mesenchymal-rich domains. Survival analysis confirms EMT-high status as an adverse prognostic factor. **Conclusions:** Critically, spatial metrics of immune cell organisation—particularly monocyte connectivity and clustering—provide substantially stronger prognostic discrimination than EMT proportion alone, demonstrating that tumour microenvironment architecture supersedes cellular composition in determining clinical outcomes in HGSOC.

Plain language summary

Ovarian cancer is often found late and can be hard to treat. The Oxford Classic is a test that looks at patterns of gene activity in cancer cells and can identify a more aggressive type of tumour cell linked to poorer survival. Until now, it was unclear whether how these tumour cells are arranged within the tumour also affects patient outcomes.

In this study, we applied the Oxford Classic to tumour samples while also analysing where different cancer and immune cells are located. We found that tumours with more aggressive cancer cells were linked to worse survival, confirming that the Oxford Classic works in this new setting. Importantly, we also showed that how immune cells are physically arranged within the tumour gives extra information about patient outcomes, beyond simply counting cell types. This suggests that the spatial organisation of cells in ovarian tumours plays an important role in determining how the disease progresses.

Introduction

High-grade serous ovarian cancer (HGSOC) is the most common and aggressive subtype of epithelial ovarian cancers, accounting for the majority of ovarian cancer-related deaths worldwide [1]. Recent evidence indicates that one site of origin for HGSOC is the secretory epithelium of the distal fallopian tube [2]. This insight has motivated application of deep molecular analysis to examine the biology of fallopian tube epithelial (FTE) cells and precursor lesions to better capture the cellular programs underpinning HGSOC [3, 4, 5].

The Oxford Classic (OxC) prognostic signature for HGSOC is based on a 52-gene panel [6] that was derived from single-cell expression profiling of normal FTE cells. OxC associates tumours with one of five transcriptional programs: C3 (differentiated), C4 (KRT17 subtype), C10 (cell cycle), ciliated, and epithelial-to-mesenchymal transition (EMT). Among these, the EMT program has emerged as a marker of particularly poor prognosis in HGSOC, reflecting its role in invasion, metastasis, and immune evasion [7, 8]. OxC has previously been successfully applied to bulk tumour transcriptomes, where deconvolution methods estimate the relative contributions of these FTE-derived programs to each tumour. However, while this approach captures overall abundance of programs such as EMT, it ignores the spatial dimension — potentially overlooking critical aspects of tumour heterogeneity, architecture, and microenvironmental interactions.

Recently, [9] used *in situ* imaging with high-plex RNA detection at single-cell resolution [10] to map the spatial organisation of high-grade serous ovarian cancer and profiled more than 2.5 million cells in 130 tumours from 94 patients using the CosMx system. This included a discovery subset in which single molecule imaging was applied to formalin-fixed and paraffin-embedded (FFPE) tissue microarrays (TMAs) and the expression of 960 genes was measured across 491,792 cells from 94 of the tumours. We sought to utilise this resource to investigate the existence of any spatial determinants of the Oxford Classic signature.

In this work, we translate the OxC prognostic signature to the 960-gene set on the CosMx spatial assay to investigate whether OxC programs, and EMT in particular, exhibit non-random spatial distributions within the tumour microenvironment of HGSOC. We develop a deep learning-based deconvolution method tailored for spatial transcriptomics data and integrate OxC-derived signatures with CosMx profiles to characterise the spatial organisation of EMT and other FTE programs across tumour tissue. Through this approach, we test the hypothesis that the OxC classification, initially designed for bulk transcriptomics, also encodes meaningful spatial information when applied to high-resolution tissue data. We show that EMT cells form spatially clustered, mesenchymal-rich domains that are broadly depleted from stromal and immune neighbourhoods, while maintaining selective colocalisation with monocytes at cluster boundaries. Survival analysis confirms that EMT-high status is associated with poor prognosis and demonstrates that spatial organisation provides stronger prognostic discrimination than EMT proportion alone. These findings indicate that the OxC classification encodes spatially meaningful biological organisation and that tumour microenvironment architecture is a key determinant of clinical outcomes in HGSOC.

Methods

Analytical overview

To investigate the spatial organisation of OxC cell states in ovarian cancer, we developed a modified 20-gene signature panel by identifying the overlap between the original 52-gene OxC panel and the 960-gene CosMx spatial assay. Marker genes for the five FTE malignant subtypes (C3, C4, C10, ciliated, and EMT) were selected using the criteria described in Hu et al. [6], with a relaxed dispersion threshold compared to the original OxC framework, thus maximising gene overlap while maintaining signature fidelity. The resulting gene set was used to construct a signature matrix, capturing the average expression profile of each cell state. The analytical pipeline used in this work proceeded through several stages (see Figure 1A): starting from single-cell spatial transcriptomics data, we applied local

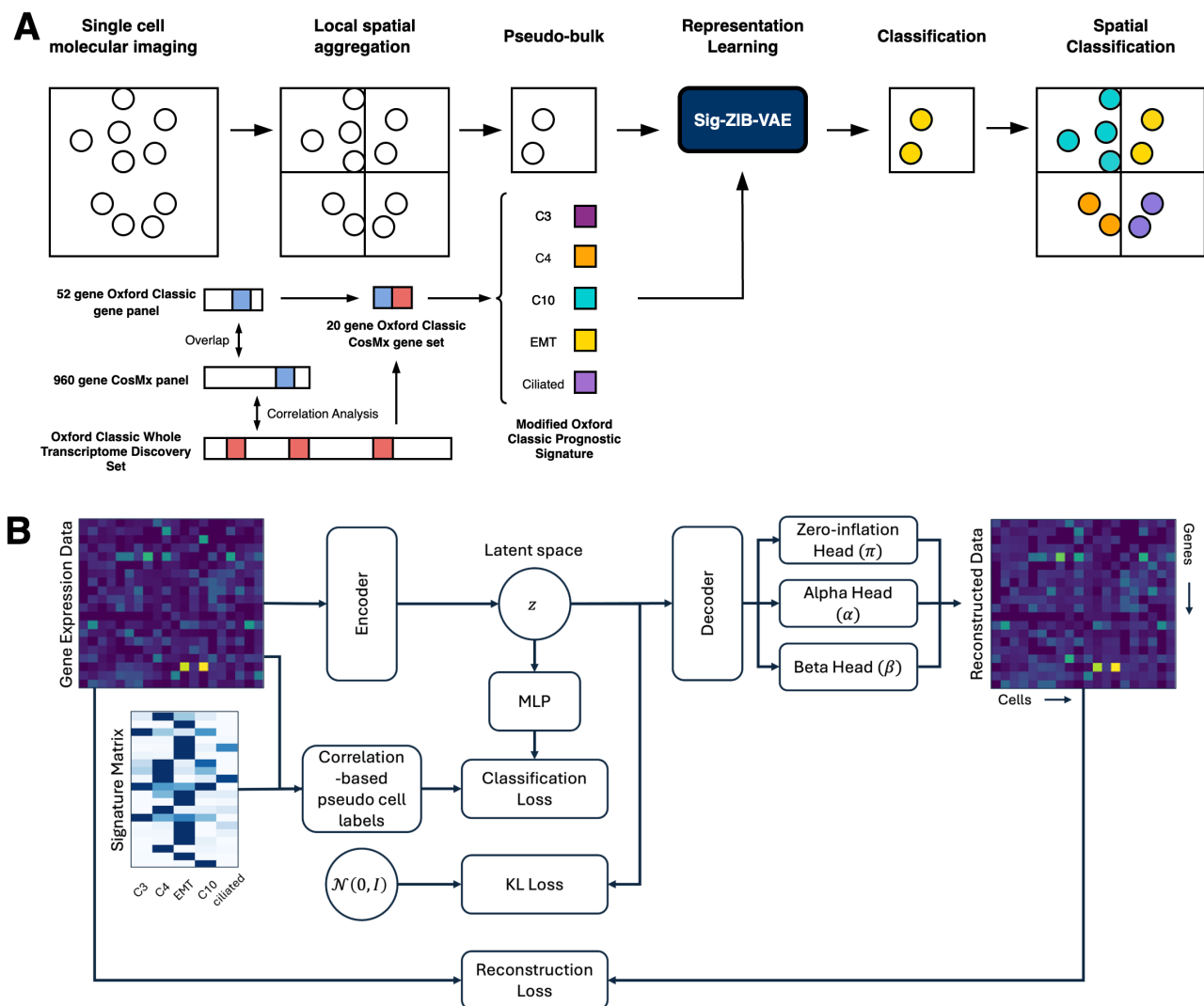


Figure 1: (A) Analysis pipeline. Single-cell CosMx expression profiles are locally aggregated on a fixed grid to generate spatial tiles (pseudo-bulk) and restricted to a modified 20-gene Oxford Classic (OxC) panel obtained by intersecting the original 52-gene OxC set with the 960-gene CosMx discovery panel (Methods). Tile-level profiles are input to a signature-guided zero-inflated Beta variational autoencoder (Sig-ZIB-VAE) to learn low-dimensional representations and assign OxC programme probabilities (C3, C4, C10, EMT, ciliated). Predicted programme labels are mapped back to native cell coordinates to produce spatial program maps, from which cell interaction graphs are constructed and multi-scale spatial metrics are quantified. (B) Schematic overview of the Signature-guided Zero-Inflated Beta Variational Autoencoder (Sig-ZIB-VAE). The model encodes normalized gene expression data \mathbf{X} into a latent representation \mathbf{z} using an encoder network. Latent variables are sampled via the reparameterization trick and decoded through a shared trunk with three output heads estimating the zero-inflation probability (π) and Beta distribution parameters (α , β), which are used for data reconstruction. In parallel, \mathbf{z} is passed through a classifier predicting cell-type probabilities, supervised by pseudo-labels derived from correlations with a reference signature matrix \mathbf{S} . Training minimizes a composite loss combining reconstruction, Kullback–Leibler divergence, and classification terms.

spatial aggregation using a fixed grid to generate pseudo-bulk samples. These aggregated expression profiles were then normalized to the $[0, 1]$ interval and analyzed using a Signature-guided Zero-inflated Beta Variational Autoencoder (Sig-ZIB-VAE) framework.

The Sig-ZIB-VAE architecture (Figure 1B) explicitly models the bimodal, zero-inflated characteristics of the spatially-aggregated data through a multi-task learning framework. As such, the encoder network maps normalized gene expression profiles to a continuous latent space via a variational frame-

work, where latent representations are sampled using the reparametrisation trick. The decoder consists of a shared trunk, followed by three specialized output heads that parameterize a zero-inflated Beta distribution: one head predicts zero-inflation probabilities (π) via *sigmoid* activation, where two additional heads generate the shape parameters (α and β) of the Beta distribution through *Softplus* transformations. To incorporate biological priors, we implemented a multi-layer perceptron (MLP) classifier operating on the latent space that predicts cell-type probabilities based on correlation-derived pseudo-labels generated by comparing input profiles to the reference signature matrix. This model is trained by jointly optimising three objectives: (1) reconstruction fidelity, through zero-inflated Beta likelihood optimisation, (2) latent space regularisation via Kullback–Leibler (KL) divergence with a standard Gaussian prior, and (3) classification accuracy relative to signature-correlated labels. This integrated approach enables the model to learn biologically interpretable latent representations while efficiently processing large-scale spatial datasets and capturing continuous transitions between cell states.

Following deconvolution, spatial neighbourhood graphs were constructed to quantify the tissue architecture, representing cells as nodes and cell-cell interactions as edges. Thus, four complementary spatial metrics were used: neighbourhood enrichment (assessing non-random spatial associations via permutation testing), co-localization quotient (CLQ, measuring directional infiltration patterns), Ripley’s L function (evaluating clustering behaviour across spatial scales via area under the curve), and degree centrality (quantifying cluster connectivity within the tissue network). Finally, to assess the prognostic significance of spatial features, we integrated baseline clinical covariates (age, stage, neoadjuvant chemotherapy, BRCA1/2 status, and tumour mutational burden) with EMT cell proportions and top-performing spatial metrics into penalized Cox proportional hazards models using elastic net regularisation. Missing clinical data were addressed through multiple imputation by chained equations (MICE) with 20 imputed datasets, and model coefficients were pooled to provide robust survival predictions accounting for both model and imputation uncertainty.

Signature Matrix, Dataset and Preprocessing

The signature matrix was acquired following the procedure detailed in [6]. The marker genes for the five FTE subtypes were first selected using the following cut-off points: (1) \log_2 fold change ≥ 1.5 , (2) false discovery rate (FDR) ≤ 0.01 ; (3) dispersion ≥ 0.025 that was measured by variable loadings in the principal component analysis of The Cancer Genome Atlas (TCGA) RNAseq data [11]; (4) the correlation with each other markers was less than 0.9 to avoid redundancy. It should be noted that the dispersion cut-off threshold was lowered here (compared to the 0.2 threshold used in [6]), to improve overlap between the CosMx panel and the OxC signature, yielding 20 shared genes instead of 11. The signature matrix was then computed using the BSEQ-sc method [12] which measures the average expression levels of the marker genes in each cell subtype. As a benchmark, CIBERSORTx [13] was used to perform deconvolution analysis, which used the aforementioned signature matrix and *pseudo-bulk* data generated averaging TPM-normalized expression values across all malignant cells from the same patient per gene. This analysis generated scores of five transcriptomic signatures in each *pseudo-bulk* tumour sample, which can be interpreted as the proportion of the corresponding cell state within the tumour (see Supplementary Section 1).

Spatial transcriptomics data were obtained from the Discovery dataset described in Yeh et al. [9, 14]. This dataset includes expression profiles of 960 genes across 94 tumours. For the present study, only cells that passed the original quality control filters were retained, and low-confidence cells were excluded from downstream analyses. Consequently, the final dataset comprised 491,792 cells, which included 314,191 malignant cells, 28,676 T cells and natural killer (NK) cells, 16,373 B cells, 45,549 monocytes, 606 mast cells, 72,861 fibroblasts/stromal cells and 13,536 endothelial cells. For brevity, detailed descriptions of data acquisition and preprocessing are provided in Yeh et al. [9] and are not repeated here. The overlapping gene set between the CosMx panel and the OxC signature included the following genes: APOA1, DCN, DPP4, ENG, IGFBP5, IGFBP6, IL1B, IL1RN, KRT23, LDLR, MGP, MYH11, PIGR, PTGS1, RAMP1, RGS5, S100A4, SPP1, TPM2, UBE2C.

Next, cell count data underwent a two-step normalization procedure to facilitate meaningful comparisons. First, to account for differential expression magnitudes across genes, we applied gene-specific normalization by calculating scaling factors as the reciprocal of the total counts for each gene across all cells. The expression values were then multiplied by these factors, effectively standardizing the contribution of each gene. Second, to ensure compatibility with the signature matrix, each cell’s gene expression profile was rescaled to the $[0, 1]$ interval. This normalization strategy preserves the relative expression patterns while enabling direct comparison with the reference signatures used in the *pseudo-bulk* deconvolution analysis.

Finally, to analyse spatial gene expression patterns, we implemented a grid-based spatial aggregation approach across malignant cells. This methodology enabled us to capture localized transcriptional signatures while maintaining adequate statistical power. As such, for each FOV ($0.9\text{mm} \times 0.6\text{mm}$), a square grid with fixed dimensions ($45\mu\text{m}$) was centered on the FOV to ensure optimal coverage (see Figure 2). Within each grid cell, malignant cells were identified based on the cell designation established in [9]. To ensure statistical robustness, only grid cells containing at least 3 malignant cells were retained for downstream analysis. For each qualifying grid cell, we aggregated the normalized expression data for the 20 genes of interest and normalized by the total number of malignant cells in that specific grid. The resulting gene expression matrix is also provided in Figure 2. This approach allowed us to generate spatially resolved expression profiles that could be subjected to deconvolution analysis.

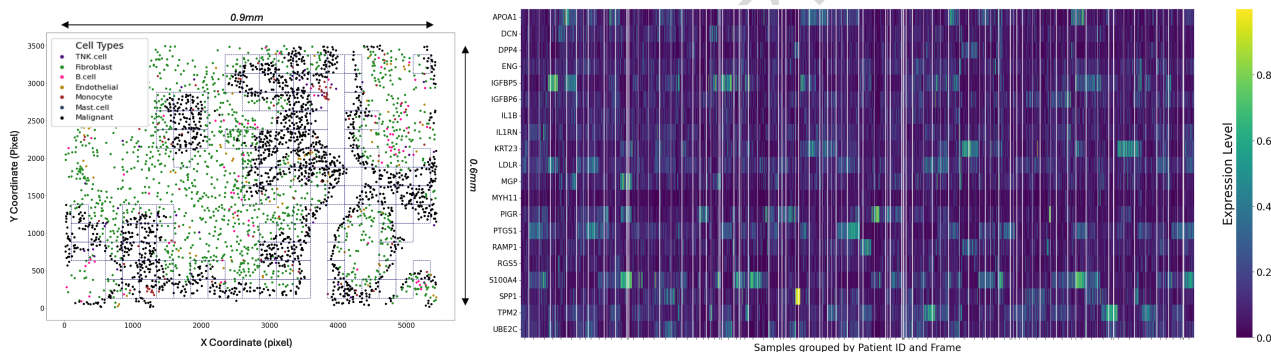


Figure 2: **Left:** Example FOV showing the spatial grid used for aggregating the expression data. Here, only the grid sections where malignant cells were detected are displayed. **Right:** Expression heatmap of selected genes across patient samples. The heatmap displays normalized expression levels of 20 genes (rows) across individual samples (columns) grouped by Patient ID and field of view (FOV). Expression values are normalized by total cell counts in every spatial grid cell. The white vertical lines are strictly used for delimitation purposes. Tick lines are demarcating boundaries between different patients, while thin lines are separating different FOVs from within the same patient.

While the OxC framework successfully leveraged CIBERSORT [15] to deconvolve bulk HGSOc data using single-cell signatures from FTE cells, several limitations inherent to traditional these deconvolution methods necessitate a more sophisticated approach. Statistical deconvolution methods such as CIBERSORT or CIBERSORTx [16] rely on support vector regression with fixed reference matrices. This becomes computationally prohibitive as datasets scale to thousand of samples and cannot adapt to dataset-specific characteristics [17]. Additionally, these methods enforce discrete cell type assignments rather than capturing the continuous spectrum of cell states, limiting their ability to model biological processes like epithelial-mesenchymal transition that are critical in HGSOc [6]. In contrast, the methodology presented in this work addresses these limitation though a deep-learning based architecture that efficiently processes large-scale datasets, while its multi-task learning objective simultaneously optimizes gene expression reconstruction, latent representation learning, and cell-type classification. This integrated approach not only improves computational efficiency but also learns a continuous latent space that may capture gradual transitions between cell states—providing richer biological insights than traditional discrete deconvolution methods while maintaining interpretability

through signature-guided learning.

Under University of Oxford institutional policy, no ethics review was required because this study involved only secondary analysis of previously published, fully de-identified data. The discovery dataset described in Yeh et al. [9] were obtained from archival clinical FFPE tumor tissues, retrospectively procured from archival storage under an Institutional Review Board (IRB)-approved protocol (number 44615) and made publicly available [14].

Sig-ZIB-VAE Description

Following the normalization procedure described previously, the resulting data demonstrated characteristics consistent with a zero-inflated beta distribution, with all values bounded within the $[0, 1]$ interval. As such, inspired by the work of [18, 19], this paper proposes the Signature-guided Zero-inflated Beta Variational Autoencoder (Sig-ZIB-VAE), a scalable method for tumour microenvironment deconvolution analysis, specifically developed for spatially aggregated gene expression data. This modeling procedure explicitly models the bimodal characteristics of the data, i.e., dropout events represented through zero-inflation and continuous values following a Beta distribution for expressed genes. Here, three objectives are jointly optimized in this multi-task framework: gene expression reconstruction, interpretable latent representation-guided through prior biological knowledge encoded in reference signatures, and cell-type classification. The mathematical formulation of the model is described below.

Let $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_K] \in [0, 1]^{P \times K}$ denote a $P \times K$ gene expression signature matrix involving P genes and K cell type signatures. Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in [0, 1]^{P \times N}$ denote a $P \times N$ normalized gene expression matrix involving P genes and N cells.

Then, for each cell n , we model the expression of gene p as a zero-inflated Beta distribution:

$$x_{np} \sim \begin{cases} 0, & \text{with probability } \pi_{np} \\ \text{Beta}(\alpha_{np}, \beta_{np}), & \text{with probability } 1 - \pi_{np} \end{cases} \quad (1)$$

where $\pi_{np} \in [0, 1]$ is the zero-inflation probability, and $\alpha_{np}, \beta_{np} > 0$ are the shape parameters of the Beta distribution. The graphical representation of the proposed model can be seen in Figure 1B. The encoder, denoted by f_{encoder} is a multi-layer neural network with parameters θ_e that maps the gene expression profile $\mathbf{x}_n \in [0, 1]^P$ to a latent representation:

$$\mathbf{h}_n^{(e)} = f_{\text{encoder}}(\mathbf{x}_n; \theta_e) \quad (2)$$

Then, the parameters of the latent distribution for each data point n are computed as linear transformations of the encoder hidden representations $\mathbf{h}_n^{(e)}$:

$$\boldsymbol{\mu}_n = \mathbf{W}_\mu \mathbf{h}_n^{(e)} + \mathbf{b}_\mu \quad (3a)$$

$$\log \boldsymbol{\sigma}_n^2 = \mathbf{W}_\sigma \mathbf{h}_n^{(e)} + \mathbf{b}_\sigma \quad (3b)$$

where $\boldsymbol{\mu}_n \in \mathbb{R}^d$ and $\boldsymbol{\sigma}_n^2 \in \mathbb{R}^d$ denote the mean and variance of the approximate posterior distribution, $\mathbf{W}_\mu, \mathbf{W}_\sigma \in \mathbb{R}^{d \times h}$ are learnable weight matrices, $\mathbf{b}_\mu, \mathbf{b}_\sigma \in \mathbb{R}^d$ are bias terms. Here, h denotes the encoder hidden dimension, while d is the dimensionality of the latent space. The latent variable \mathbf{z}_n is then sampled using the reparameterization trick [20], which enables backpropagation through the stochastic sampling step:

$$\mathbf{z}_n = \boldsymbol{\mu}_n + \boldsymbol{\sigma}_n \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_d) \quad (4)$$

Here \odot denotes element-wise multiplication and $\boldsymbol{\epsilon}$ is a noise vector drawn from a standard d -dimensional Gaussian distribution.

Next, the decoder network maps the latent representation \mathbf{z}_n to a lower-dimensional hidden representation:

$$\mathbf{h}_n^{(d)} = f_{\text{decoder}}(\mathbf{z}_n; \theta_d) \quad (5)$$

where f_{decoder} is a neural network with parameters θ_d . The decoder consists of a shared trunk (common layers) followed by three specialized output heads, each responsible for producing the parameters of the zero-inflated Beta distribution:

$$\pi_n = \sigma(f_{\text{zero}}(\mathbf{h}_n^{(d)}; \theta_\pi)) \quad (6a)$$

$$\alpha_n = \text{Softplus}(f_\alpha(\mathbf{h}_n^{(d)}; \theta_\alpha)) + \epsilon_{\min} \quad (6b)$$

$$\beta_n = \text{Softplus}(f_\beta(\mathbf{h}_n^{(d)}; \theta_\beta)) + \epsilon_{\min} \quad (6c)$$

where $\pi_n \in (0, 1)$ is the probability of observing a structural zero, $\alpha_n, \beta_n > 0$ are the shape parameters of the Beta distribution, $f_{\text{zero}}, f_\alpha, f_\beta$ are output layers (typically linear mappings) with learnable parameters $\theta_\pi, \theta_\alpha, \theta_\beta$. $\sigma(\cdot)$ denotes the sigmoid function, while the Softplus function ensures positivity and ϵ_{\min} is a small positive constant added for numerical stability. This structure allows the decoder to jointly capture the zero-inflation mechanism (π_n) as well as the continuous variability in the data through the Beta distribution parameters (α_n, β_n). Finally, the reconstructed expression values are computed as the expected value of the zero-inflated Beta distribution, given below:

$$\mathbb{E}[x_{np}] = (1 - \pi_{np}) \frac{\alpha_{np}}{\alpha_{np} + \beta_{np}} \quad (7)$$

The Sig-ZIB-VAE model requires three key inputs: (1) the aggregated gene expression profiles from the dataset, (2) a reference signature matrix \mathbf{S} containing cell type-specific molecular signatures, and (3) cell type labels derived from a correlation-based preprocessing step.

The correlation-based procedure is an essential step that transforms continuous expression data into discrete cell type assignments. For each observed gene expression profile, Pearson correlation coefficients are computed against all reference signatures in \mathbf{S} . The resulting similarity scores are then normalized with the Softmax function to obtain probability distributions over cell types. The cell type label is defined as the index corresponding to the maximum probability. This scheme ensures that the assigned labels capture the degree of similarity between observed expression patterns and known reference signatures, thereby providing biologically meaningful supervision signals for training.

To incorporate the information from the reference signatures, we introduce a cell type classifier operating on the latent representation:

$$\mathbf{c}_n = \text{Softmax}(f_{\text{classifier}}(\mathbf{z}_n; \theta_c)), \quad (8)$$

where \mathbf{z}_n is the latent representation vector, $f_{\text{classifier}}$ is a neural network parameterized by θ_c , and $\mathbf{c}_n \in \mathbb{R}^K$ denotes the predicted probability distribution over K candidate cell types. The inclusion of this classifier encourages the latent space to align with biologically relevant cell type distinctions as defined by the reference signatures.

Training of the Sig-ZIB-VAE proceeds by minimizing a composite loss function that balances three key components: (1) reconstruction fidelity, (2) latent space regularisation via the Kullback–Leibler (KL) divergence, and (3) classification accuracy with respect to correlation-derived labels. This is formalized in Equation 9:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{class}} \mathcal{L}_{\text{class}}, \quad (9)$$

where the hyperparameters λ_{KL} and λ_{class} control the relative weighting of the KL and classification terms. The reconstruction component ensures that essential information in the observed expression profiles is preserved. Given the zero-inflated nature of transcriptomics data, we model the likelihood using a mixture of a point mass at zero and a Beta distribution for positive entries. The corresponding negative log-likelihood is decomposed according to Equations 10, where the indicator function $\mathbb{I}(\cdot)$ is used to separate the zero and nonzero cases and λ_{zero} is used to further control the weighting of the zero-loss term:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{zero}} + \mathcal{L}_{\text{Beta}}, \quad (10a)$$

$$\mathcal{L}_{\text{zero}} = -\lambda_{\text{zero}} \sum_{n,p} \left[\mathbb{I}(x_{np} = 0) \log \pi_{np} + \mathbb{I}(x_{np} > 0) \log(1 - \pi_{np}) \right], \quad (10b)$$

$$\mathcal{L}_{\text{Beta}} = - \sum_{n,p} \mathbb{I}(x_{np} > 0) \log p_{\text{Beta}}(x_{np}; \alpha_{np}, \beta_{np}), \quad (10c)$$

The KL divergence term - provided in Equation 11 regularises the approximate posterior $q_{\phi}(\mathbf{z}_n | \mathbf{x}_n)$ to remain close to the standard Gaussian prior, thereby promoting a smooth and disentangled latent space. This encourages global structure in the latent representation, facilitating generalisation and interpretability.

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q_{\phi}(\mathbf{z}_n | \mathbf{x}_n) \| \mathcal{N}(0, I)). \quad (11)$$

Finally, the classification loss component explicitly guides the latent space to align with biologically meaningful cell type distinctions derived from the signature matrix \mathbf{S} . Pseudo-labels are generated by correlating each expression profile \mathbf{x}_n with all reference signatures, followed by softmax normalisation and maximum selection:

$$y_n = \arg \max_k \text{softmax}(\text{Corr}(\mathbf{x}_n, \mathbf{S}))_k. \quad (12)$$

The classification loss then penalises discrepancies between the classifier predictions and these correlation-derived labels as seen in Equation 13:

$$\mathcal{L}_{\text{class}} = - \sum_c \mathbf{1}_{\{y_n=c\}} \log p_{\psi}(y_n = c | \mathbf{z}_n), \quad (13)$$

where $p_{\psi}(y_n | \mathbf{z}_n)$ is the classifier's output distribution over cell types. Classification accuracy, used to evaluate model performance, is computed as the proportion of samples for which the predicted label $\arg \max_k p_{\psi}(y_n = k | \mathbf{z}_n)$ matches the pseudo-label y_n derived from the signature correlation procedure.

By jointly optimising reconstruction fidelity, latent space regularisation, and signature-consistent classification, the composite loss provides a principled framework for analysing zero-inflated expression data. Moreover, the incorporation of correlation-derived labels ensures that the model captures interpretable, biologically grounded cell type distinctions, which are subsequently used in downstream spatial analysis.

Spatial arrangement analysis

The work presented here further extends the OxC classification system to spatial transcriptomics by leveraging spatial neighbourhood graphs to analyse the spatial patterns and interactions between the different cell types across tissues. Here, spatial neighbour relationships were computed based on Euclidean distances, using a fixed radius approach ($30\mu\text{m}$). Each cell was represented as a node in a spatial graph, with edges denoting cell-cell interactions. Edges were established between nodes if their corresponding cells lay within the fixed radius. A total of 98 FOVs, measuring $0.9\text{mm} \times 0.6\text{mm}$ were selected, representing a total number of 58 patients (see supplementary material in [9]). Following the graph computation for all available FOV, a multi-metric approach was employed to quantify the tissue architecture, using four complementary metrics.

Neighbourhood enrichment

Firstly, neighbourhood enrichment analysis was used to evaluate whether pair of cells are spatially adjacent more or less frequently than expected by chance. Cell-cell associations are inferred from spatial connectivity graph, where each node represents a cell and edges encode proximal spatial relationships. For each pair of cell types i and j , the observed number of neighbouring interactions x_{ij} is computed by summing the number of edges connecting nodes of these two types. To assess statistical significance, the cell type labels are randomly permuted across the graph (typically 1,000 times), while preserving the underlying connectivity structure. The mean μ_{ij} and standard deviation σ_{ij} of x_{ij} under the null distribution are estimated from these permutations. A z-score is computed for each cell type pair:

$$Z_{ij} = \frac{x_{ij} - \mu_{ij}}{\sigma_{ij}} \quad (14)$$

This z-score reflects the degree of enrichment (positive values) or depletion (negative values) of interactions between cell types i and j relative to random expectation, indicating whether their spatial co-occurrence is statistically significant [21].

Co-localization tests

Spatial pairwise co-localization patterns were assessed using the co-localization quotient (CLQ) [22]. $CLQ_{a \rightarrow b}$ is a quantitative representation of spatial co-occurrence of ‘target’ and ‘infiltrating’ cell types, denoted as a and b , respectively. The CLQ captures the degree of cell type b co-localizing with cell type a as a ratio of the observed versus the expected number of cell-type b within a radius of $30\mu m$ of cell type a . The CLQ is mathematically defined as:

$$CLQ_{a \rightarrow b} = \frac{C_{b \rightarrow a}/N_a}{N_b/(N-1)} \quad (15)$$

where $C_{b \rightarrow a}$ is the number of cell type b within a $30\mu m$ radius of cells of cell type a . N is the total number of cells in a tissue sample, and N_a and N_b denote number of cells of cell type a and cell type b respectively. CLQ values were calculated for malignant cells infiltrating Fibroblast and Endothelial cells.

Ripley’s L statistics

Ripley’s L function is a second-order spatial statistic used to assess whether a set of spatial points exhibits clustering, regularity (dispersion), or complete spatial randomness at a given spatial scale. It provides a scale-normalized transformation of Ripley’s K function, facilitating interpretation across different radii. The L function is defined as:

$$L(t) = \sqrt{\frac{K(t)}{\pi}} \quad (16)$$

where the K function is given by:

$$K(t) = \frac{1}{\lambda} \sum_{i \neq j} \frac{I(d_{ij} < t)}{n} \quad (17)$$

In Equation 17, d_{ij} denotes the Euclidean distance between points i and j in a dataset of n points and t is the spatial scale (radius) of interest. The term λ represent the average point density, typically estimated as $\lambda = n/A$, where A is the total area encompassing all points. The indicator function $I(\cdot)$ equals 1 when the condition is true (i.e., when two points lie within distance t), and 0 otherwise. To summarize the overall spatial behavior of each cell type across all scales, we computed the area under the Ripley’s L function curve (AUC). This approach captures the cumulative extent of spatial

aggregation or dispersion exhibited by a cell type across distances. Higher AUC values reflect stronger or more persistent spatial clustering over the evaluated distance range.

Degree centrality metrics

The final spatial metric computed in this work is represented by degree centrality [23], which is a graph-theoretic metric, that quantifies the extent to which a cluster is connected to the rest of the network. In the context of spatial organisation, it can be used to assess how integrated or isolated a particular cluster of cells is relative to its surrounding environment.

Let $G = (V, E)$, denote an undirected graph, here V is the set of nodes (or cells in the case of this work) and E is the set of edges representing pairwise correlations or spatial properties. Let $S \subseteq V$ denote a subset of nodes belonging to a specific cluster c_s . Define $N(S)$ as the set of all nodes in the graph that are neighbours of any node in S , including intra-cluster and inter-cluster connections.

Degree centrality of cluster S is then defined as the fraction of nodes outside the cluster that are directly connected to at least one member of S , as described in Equation 18.

$$C_{def}(S) = \frac{|N(S) - S|}{|V| - |S|} \quad (18)$$

Here $|N(S) - S|$ represents the number of non-cluster nodes that are adjacent or functionally integrated within the tissue, suggesting greater interaction with other clusters. Conversely, lower values suggest spatial or functional isolation, though it should be noted that this metric is inherently sensitive to cluster size and cell type abundance, as larger or more prevalent populations have greater opportunity to establish inter-cluster connections.

Boundary connectivity decomposition

To characterise which cell populations contribute to the inter-cluster connectivity captured by degree centrality, the cell-type composition of heterotypic neighbours was directly tallied from the spatial graph. For each FOV, all non-focal cells directly connected to at least one focal cell were identified. The cell-type composition of these boundary neighbours (observed proportion) was compared to the overall non-focal tissue composition (expected proportion) to obtain an observed-to-expected (O/E) ratio for each cell type, where values greater than one indicate over-representation at cluster boundaries and values less than one indicate under-representation. This compositional analysis complements the pairwise CLQ metric by providing a simultaneous breakdown of boundary contacts across all cell types, and is distinct from the permutation-based neighbourhood enrichment z-scores described above, which test whether the *total number* of heterotypic edges between two populations exceeds random expectation across the entire FOV.

Statistics and Reproducibility

Unless otherwise specified, all statistical tests comparing spatial metrics derived from the methods described above were performed using GraphPad Prism (version 10.5.0) [24]. For transparency, biological replicates were independent tumours. When multiple fields of view were available per tumour, spatial metrics were averaged to obtain a single patient-level estimate.

Survival Modeling

The survival analysis incorporated clinical, genomic, and spatial features to assess their prognostic significance in ovarian cancer outcomes. The baseline clinical features included the age at diagnosis, the stage of the disease and whether subjects were under treatment with neoadjuvant chemotherapy. In addition, the baseline model also included BRCA1/2 mutation status, together with the tumour mutational burden (TMB) score. Post diagnosis treatment variables including bevacizumab combination therapy and PARP inhibitor treatment were excluded from the primary baseline model to avoid

treatment selection bias, as these represent post-diagnostic treatment decisions, rather than baseline prognostic attributes. For clarity, the baseline survival features were extracted from [9]. Next, the baseline model was further augmented with spatiomolecular information. This included the overall proportion of EMT cells, as well as the spatial organisation metrics presented in Methods. Survival analysis was performed using the *scikit-survival* package [25], where time-to-event outcomes were specified using overall survival as the primary endpoint, with follow-up time measured in days from diagnosis to death or last follow-up.

Spatial metrics were computed independently for each FOV as described in Methods. For patients with multiple FOVs, metric values were averaged to obtain patient-level summary statistics, yielding a single value per spatial feature per patient. A total of 20 candidate spatial features were evaluated (Supplementary Table 13), comprising: (i) Ripley’s L AUC for 7 cell populations (EMT, monocytes, T/NK cells, B cells, mast cells, fibroblasts, endothelial cells); (ii) degree centrality at 30 μ m for the same 7 cell types; and (iii) pairwise metrics between EMT cells and six stromal/immune populations (neighborhood enrichment z-scores and CLQ values for EMT-monocyte, EMT-T/NK, EMT-fibroblast, EMT-endothelial, EMT-B cell, and EMT-mast cell interactions). Features with > 40% missing values were excluded, and remaining missing values were imputed using median imputation.

A hierarchical modeling approach was employed to systematically evaluate the prognostic contribution of different feature sets. The analysis proceeded in four sequential stages: (1) a baseline clinical model incorporating age, disease stage, neoadjuvant chemotherapy status, BRCA1/2 mutation status, and log-transformed TMB scores; (2) an enhanced model augmenting the baseline features with EMT cell proportion; (3) spatial feature screening, in which individual spatial organisation metrics were added to the extended model to assess their independent prognostic value; and (4) construction of a final comprehensive model that integrated baseline clinical features, EMT cell proportion, and top-performing spatial features. Spatial features were retained in the final model if they met two criteria: (i) they yielded a C-index improvement over the EMT-augmented model, and (ii) they demonstrated selection stability in more than 90% of the imputed datasets, ensuring robust and reproducible associations. To prevent redundancy, pairs of highly correlated spatial features with Spearman’s correlation coefficients ≥ 0.85 were identified and only one feature from each pair was retained for analysis.

Moreover, to validate the prognostic significance of EMT cell proportions, patients were stratified into EMT-high and EMT-low groups using the median EMT score, consistent with the OxC framework. Kaplan–Meier survival curves with log-log confidence intervals were generated, and differences between groups were tested using the log-rank method to determine whether the survival trends previously reported in the OxC study could be reproduced in this spatially characterized dataset.

To evaluate the prognostic contribution of clinical and spatial covariates, penalized Cox proportional hazard models were implemented using elastic net regularisation. Missing data in the clinical variables were first addressed using Multiple Imputation by Chained Equations (MICE) with 20 imputed datasets. Categorical variables were numerically encoded for imputation: age ($\leq 65 = 0$, $> 65 = 1$), disease stage (III = 3, IV = 4), neoadjuvant chemotherapy status (Untreated = 0, Treated = 1), and BRCA mutation status (Benign = 0, Pathogenic = 1). tumour mutational burden was log-transformed ($\log(\text{TMB} + 0.01)$) prior to imputation. The iterative imputation was performed using Bayesian Ridge regression [26], for a maximum of 100 iterations. Finally, imputed categorical values were constrained to valid ranges and rounded to maintain categorical integrity.

The elastic net penalty combines l_1 (LASSO) and l_2 (Ridge) regularisation, by optimizing the penalized partial likelihood:

$$\arg \max_{\beta} \log PL(\beta) - \alpha \left(\underbrace{r \sum_{j=1}^p |\beta_j|}_{\ell_1 \text{ term}} + \frac{1-r}{2} \underbrace{\sum_{j=1}^p \beta_j^2}_{\ell_2 \text{ term}} \right) \quad (19)$$

where $PL(\beta)$ is the partial likelihood function of the Cox model, β_1, \dots, β_p are the coefficients for p features and $\alpha \geq 0$ is the hyperparameter that controls the amount of shrinkage. Here, $r \in [0, 1]$ was used to control the relative ratio between l_1 and l_2 penalties, hence combining the subset selection property of the LASSO penalty with the regularisation strength of the Ridge penalty. In the case of this work, the relative ratio was set to $r = 0.05$, hence imposing a more Ridge-like penalty, while retaining some variable selection capability. The optimal regularisation parameter α was selected through out-of-bag bootstrap cross-validation using 200 bootstrap replicates across 100 candidate α values spanning the range $[0.01 \alpha_{max}, \alpha_{max}]$, where α_{max} is automatically determined by the algorithm. For each imputed dataset, predictor variables were standardized prior to model fitting, with α selected by maximizing out-of-bag Harrell's C-index. Model performance and coefficients were pooled across 20 imputed datasets to obtain robust estimates that account for both model uncertainty and imputation variability.

Results

Sig-ZIB-VAE model performance and validation

To rigorously assess the proposed Sig-ZIB-VAE deconvolution method, data were partitioned into stratified training and validation sets, with 20% of the data held out for validation purposes. To examine model stability and reproducibility, ten independent training runs were performed using different random initializations.

The model demonstrated robust performance with mean absolute validation error of $4.03\% \pm 0.38\%$ and classification accuracy of $96.54\% \pm 0.42\%$ across all runs. The narrow standard deviations indicate consistent convergence independent of initialization, supporting the stability of the multi-objective optimization framework. The optimal model (selected based on lowest validation loss) achieved 3.7% reconstruction error and 96.57% classification accuracy on held-out data. The reconstruction and classification performance of the selected model can be seen in Figure 3B and 3C. We also obtained comparable results to using CIBERSORTx [13] (Supplementary Information, Section 1 and Supplementary Figure 5).

Analysis of the learned distributional parameters revealed successful adaptation to gene-specific expression patterns (as shown in Supplementary Figures 1 and 2). Genes with high sparsity (such as DDP4, MYH11 or RGS5, for example) showed elevated zero-inflation probabilities ($\pi > 0.3$), while their Beta distributions, when expressed, captured the full range of possible values. Conversely, more constitutively expressed genes (such as IGFBP6, TPM2 or UBE2C) exhibited lower zero-inflation probabilities, with Beta parameters that appropriately modeled their expression distributions. Notably, beta parameters across all genes showed considerable variation, reflecting the diversity of expression patterns in the dataset.

Latent space analysis using t-SNE projection revealed distinct cell type segregation (Figure 3D). Four of the five reference signatures were positioned at the periphery of their respective clusters, representing archetypal profiles of pure cell states. The C10 signature occupied a more central location; while this could reflect its biological role as a transient, proliferative state bridging other epithelial programs [6], it is important to note that t-SNE is a stochastic, non-linear dimensionality reduction method whose fine-scale topology should not be over-interpreted. The central positioning may also partly arise from partial overlap among transcriptional programs in the reference signature matrix (Figure 1B), where proliferative gene expression features are shared across C10, C3, and EMT subtypes. To formally assess the stability of the latent space and rule out stochastic visualization artifacts, we performed a complementary seed-sensitivity analysis (Supplementary Figure 3). Briefly, t-SNE projections generated using multiple random initializations exhibited consistent global structure after Procrustes alignment [27, 28], with stable inter-cell type relationships. Moreover, a seed-independent PCA visualization reproduced the same qualitative organisation of cell types.

More robust evidence that the latent space preserves biologically meaningful structure comes from gene-specific expression gradients across the t-SNE projection (Supplementary Figure 4). Genes ex-

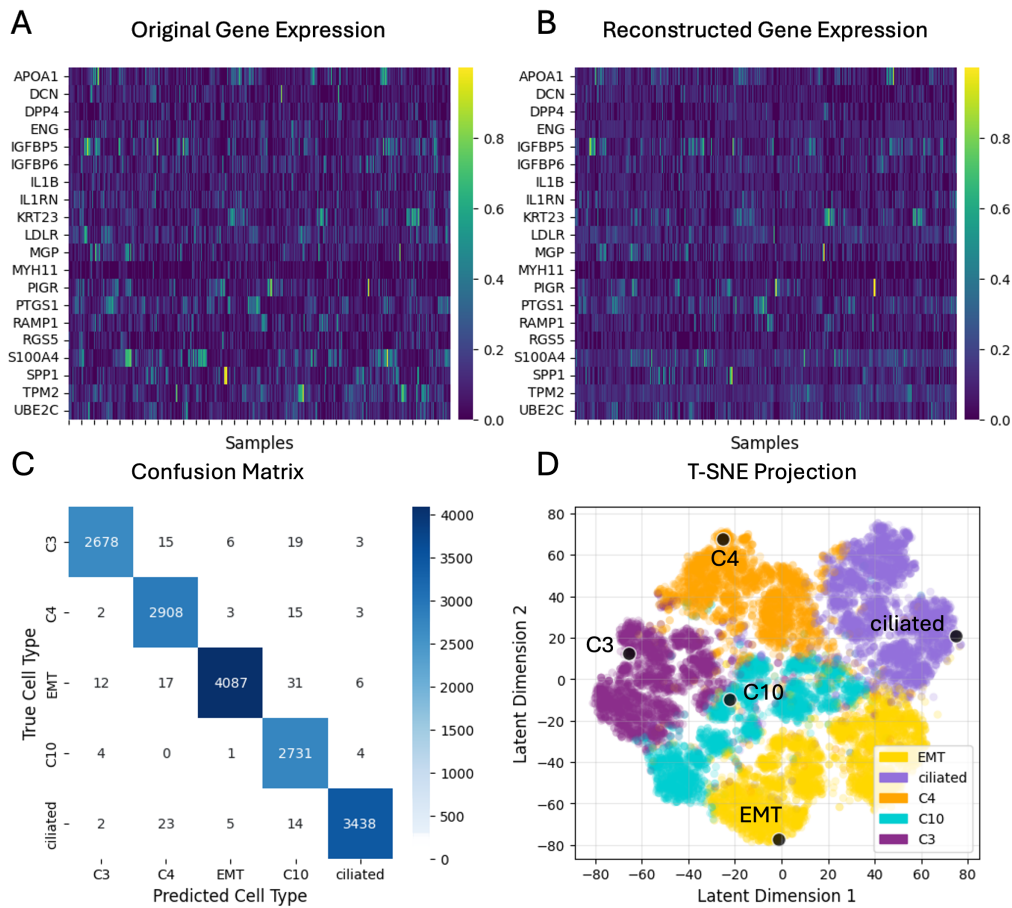


Figure 3: **A:** Original gene expression heatmap; **B:** Reconstructed gene expression demonstrating high fidelity. Expression levels are indicated by the colormaps on the right of the gene panels; **C:** Confusion matrix showing the classification performance across all cell types; **D:** t-SNE projection colored by predicted cell types, revealing distinct clustering. Signature cells are represented by the black dots.

hibit spatially structured patterns that align with expected cell type-specific transcriptional programs: for example, UBE2C (a cell cycle marker) shows concentrated high expression in the C10 region, the high expression of PTGS1 overlaps the C3 cluster boundaries, PIGR localizes to the C4 cluster, TPM2 (a mesenchymal marker) concentrates in the EMT cluster, while KRT23 expression is enriched in the ciliated cell region cluster. These continuous expression gradients provide direct, gene-level validation that the learned latent representations capture and organize cells according to their transcriptional programs, independent of t-SNE's stochastic visualization properties.

Spatial Analysis

Figure 4 shows two example fields of view (FOVs) of FTE tissue architectures, where the five OxC molecular signatures were inferred using the proposed multi-task Sig-ZIB-VAE framework. It is important to note that *Malignant - Other* cells, which are primarily located along the periphery of the FOV, were intentionally excluded from the aggregation process and were not used in any downstream analysis. This exclusion arises from the centering of spatial grids on each FOV, a deliberate design strategy aimed at reducing border effects and maintaining uniform spatial coverage across all samples [29]. Following spatial neighbourhood graph construction, the spatial organisation of EMT cells within HGSOc tissues was examined using neighbourhood enrichment, CLQ values, Ripley's L AUC, and degree centrality metrics, as previously described. Formal definitions of these metrics are provided in the Methods section, with a concise overview presented in Figure 5.

Firstly, neighbourhood enrichment analysis was used to quantify whether pairs of cell types are spa-

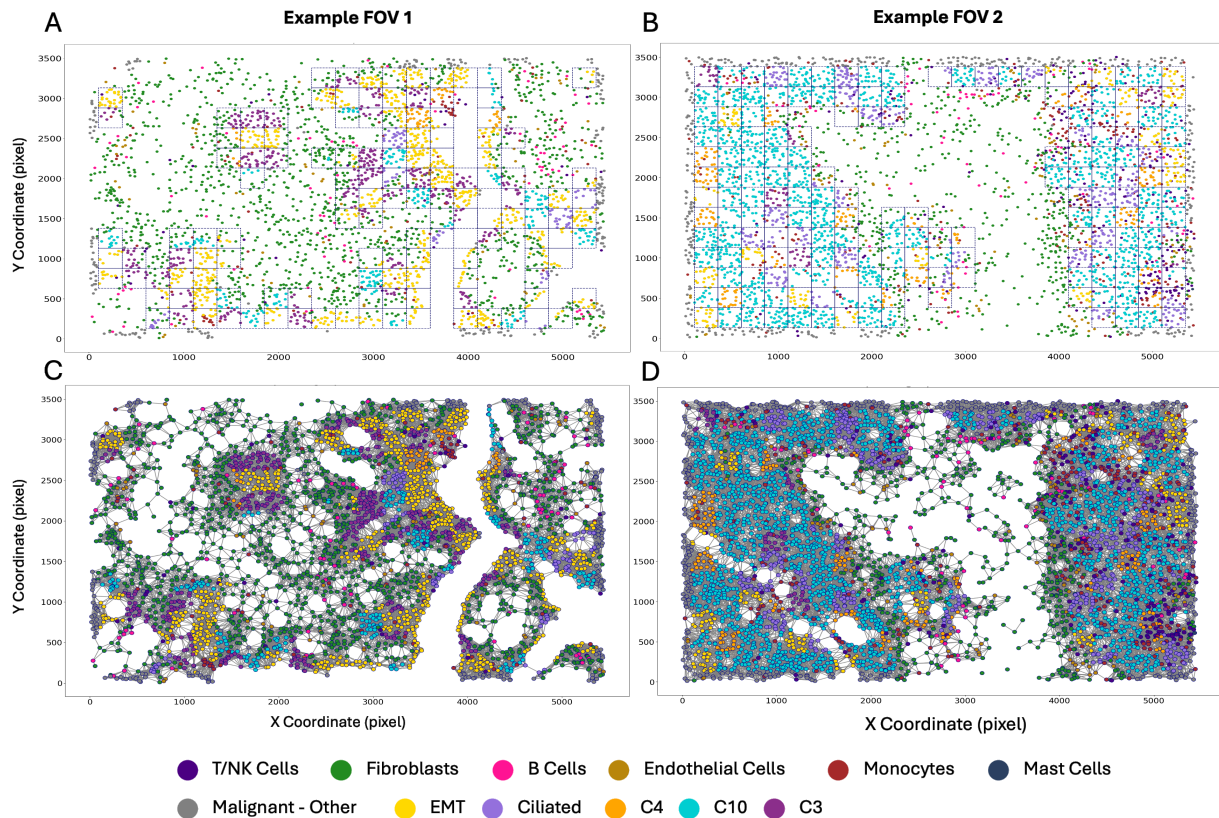


Figure 4: Spatial arrangement of cells across 2 fields of view (FOVs) from unique individuals. (A) and (B) show the grid used for cell aggregations across both examples; (C) and (D) show the corresponding spatial graphs. Cells are represented as nodes, while neighbourhood relations between cells are represented as edges.

tially adjacent more or less frequently than expected by chance, using 1,000 random permutations while preserving the underlying graph connectivity and computing a z-score. Across all comparisons, EMT cells displayed negative enrichment scores, indicating consistent spatial depletion relative to random expectation (Supplementary Table 1). Collectively, these results demonstrate that predicted EMT cells occupy spatially isolated or stromal-depleted regions, showing minimal adjacency to fibroblast, endothelial, or immune compartments. The strongest depletion occurred in relation to fibroblasts (median = -11.23), followed by monocytes (median = -4.62) and T/NK cells (median = -4.22). Endothelial and B cells exhibited moderate depletion (median = -3.58 each), while mast cells were least affected (median = -1.85 ; all $p < 0.0001$). It is worth noting, however, that the latter score likely reflects the overall sparsity of mast cells. These results are illustrated in Figure 5A, where the median enrichment values for all cell classes remained below zero, confirming an overall absence of localized EMT enrichment within mixed cellular niches.

Second, cell-cell co-localisation was quantified using the co-localization quotient (CLQ), which represents the ratio between the observed and expected number of neighbouring cells within a $30\mu m$ radius, thereby reflecting spatial preference. Across all EMT-cell type pairs, CLQ values were greater than one, indicating frequent co-occurrence beyond random expectation (Supplementary Table 2). Among all tested cell types, EMT cells exhibited the highest co-localization with monocytes (median = 17.37 , $p < 0.0001$), followed by T/NK cells (median = 16.45 , $p < 0.0001$). Fibroblasts (median = 9.48 , $p < 0.0001$), endothelial cells (median = 12.61 , $p < 0.0001$), and B cells (median = 5.17 , $p < 0.0001$) also showed significant enrichment relative to random mixing, whereas mast cells did not (median = 13.01 , $p = 0.25$). These findings suggest that EMT-like cells are not randomly dispersed within the tissue but display selective spatial proximity to specific stromal and immune populations—particularly monocytes and T/NK cells—within otherwise depleted microenvironmental niches (Figure 5B).

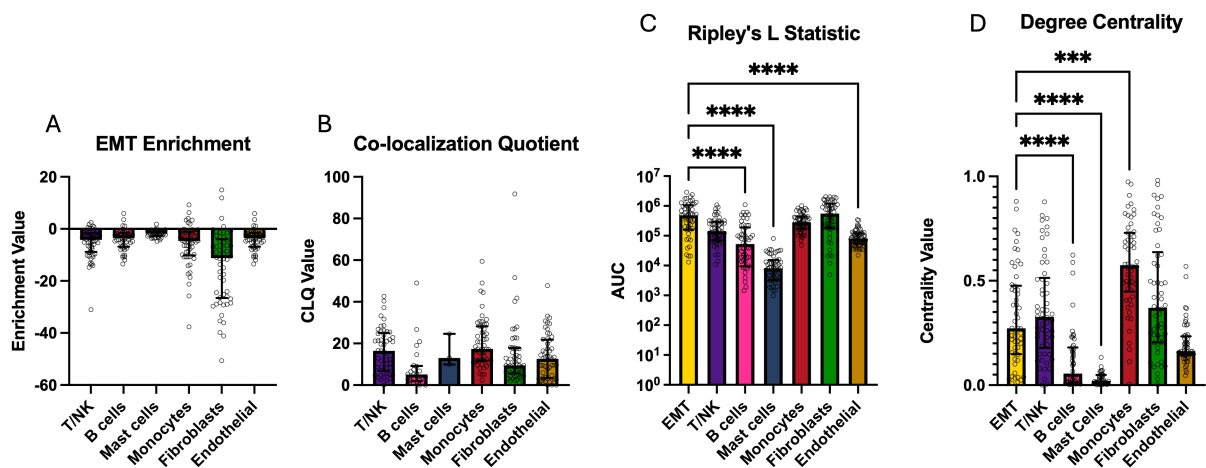


Figure 5: Distribution of spatial organisation metrics across the full cohort. Scatter dot plots show the overall distributions of (A) neighborhood enrichment values ($n_{T/NK} = 54$, $n_B = 48$, $n_{Mast} = 38$, $n_{Monocytes} = 54$, $n_{Fibroblasts} = 54$, $n_{Endothelial} = 48$ biologically independent tumours), (B) co-localization quotients (CLQ) ($n_{T/NK} = 49$, $n_B = 25$, $n_{Mast} = 3$, $n_{Monocytes} = 52$, $n_{Fibroblasts} = 51$, $n_{Endothelial} = 53$ biologically independent and (D) degree centrality ($n_{EMT} = 54$, $n_{T/NK} = 54$, $n_B = 48$, $n_{Mast} = 38$, $n_{Monocytes} = 54$, $n_{Fibroblasts} = 54$, $n_{Endothelial} = 54$ biologically independent tumours) across major immune and stromal cell types. Individual data points represent biologically independent tumours. Bars indicate the median; error bars denote the interquartile range (25th–75th percentiles). Statistical significance was determined using two-sided Kruskal–Wallis tests followed by Dunn’s multiple comparisons test, with a significance threshold value of $\alpha = 0.05$. Significant differences are indicated as: $p < 0.001$ (***), $p < 0.0001$ (****).

Collectively, the CLQ and neighbourhood enrichment analyses reveal a dual pattern of EMT spatial organisation. Negative enrichment scores indicate that EMT cells preferentially cluster with other EMT cells rather than forming heterotypic neighborhoods with stromal or immune populations. However, positive CLQ values - particularly for monocytes (*median* = 17.37) and T/NK cells (*median* = 16.45) - demonstrate that specific cell types are nonetheless enriched within $30\mu m$ of EMT cells. This pattern is consistent with EMT cells forming self-aggregating domains where the majority of neighbours are other EMT cells (explaining the neighbourhood depletion), while maintaining selective spatial proximity to monocytes and fibroblasts specifically at cluster peripheries or within mixed micro-environmental niches. Notably, both metrics operate at the same $30\mu m$ spatial scale, indicating that the observed pattern reflects cellular composition preferences rather than scale-dependent spatial organisation.

Third, spatial clustering within each cell population was assessed using Ripley’s L function, computed across increasing spatial radii. To facilitate comparison across cell types, the area under the curve (AUC) of Ripley’s L was determined for each population, providing a global measure of spatial aggregation, where higher AUC values indicate stronger or more persistent clustering over distance. Across all compartments, predicted EMT cells and fibroblasts exhibited the highest AUC values (Supplementary Table 3). Monocytes showed intermediate clustering, whereas T/NK and B cells displayed comparatively weaker aggregation. Endothelial and mast cells exhibited the lowest AUC values, indicative of near-random or weakly clustered spatial distributions (Figure 5C). Statistical comparison of AUC values across cell types revealed significant differences among several populations (Supplementary Table 5). In particular, EMT cells were significantly more clustered than B cells, mast cells, and endothelial cells ($p < 0.0001$), while showing comparable clustering to fibroblasts and monocytes ($p > 0.99$). These findings might suggest that EMT cells and fibroblasts form localized clusters within the tissue microarchitecture, whereas immune and endothelial compartments display more dispersed distributions.

Finally, degree centrality was computed to assess the relative connectivity of each cellular compartment within the tissue spatial graph. This metric quantifies the proportion of inter-cluster connections

maintained by a given population, thereby reflecting its level of spatial integration within the broader microenvironment. Across all cell types, monocytes and fibroblasts exhibited the highest centrality values (medians = 0.57 and 0.37, respectively), indicating strong inter-cluster connectivity and central positioning within the spatial network (Supplementary Table 4). EMT cells and T/NK cells showed moderate connectivity (medians = 0.27 and 0.33, respectively), suggesting partial integration with surrounding clusters. In contrast, B cells (median = 0.06) and mast cells (median = 0.02) displayed lower centrality scores, consistent with peripheral or spatially isolated distributions, while endothelial cells exhibited intermediate centrality (median = 0.16). Statistical comparison (Supplementary Tables 4 and 6) revealed that EMT cells were significantly more connected than B cells ($p < 0.0001$) and mast cells ($p < 0.0001$), but less connected than monocytes ($p = 0.0004$). No significant differences were observed between EMT-like cells and fibroblasts, endothelial cells, or T/NK cells ($p > 0.49$ for all). These results indicate that EMT cells occupy an intermediate topological position within the tumour spatial network, being more integrated than sparsely distributed immune subsets, yet less connected than stromal populations acting as key structural hubs.

To further characterise which cell populations contribute to EMT inter-cluster connectivity, the cell-type composition of non-EMT neighbours at EMT cluster boundaries was compared to the background tissue composition. The majority of heterotypic connections involved other malignant subtypes (67.4%), while the remaining 32.6% involved stromal and immune populations. Among non-malignant cell types, monocytes were the only population over-represented at EMT boundaries relative to their tissue abundance (observed at 1.27-fold their expected frequency), whereas fibroblasts were under-represented (0.66-fold). The full results are provided in Supplementary Table 11.

To examine the interdependence between spatial metrics, pairwise correlations were computed across all features (Figure 7B). The resulting matrix revealed moderate positive correlations between clustering (Ripley's L AUC), degree centrality, and EMT-stromal co-localization indices, indicating that cell populations exhibiting strong self-aggregation also tend to be spatially integrated within the tissue network. In contrast, neighbourhood enrichment scores showed weaker or inverse correlations with these measures, reflecting their distinct sensitivity to large-scale depletion rather than local clustering. Overall, these analyses might highlight a multifaceted spatial phenotype of EMT cells, characterized by broad depletion from stromal and immune niches, yet persistent small-scale clustering and moderate inter-cluster connectivity, suggestive of selective integration within localized microenvironments.

EMT Stratification analysis

Figure 6 presents the statistical analysis of the spatial distribution differences between EMT-Low and EMT-High subgroups. For clarity, stratification into EMT-Low and EMT-High subgroups was performed using the median EMT proportion value, derived from the present deconvolution method, in line with the methodology described in [6]. Neighbourhood enrichment showed significantly higher depletion of EMT cells from fibroblast-rich and monocyte-rich neighbourhoods in the EMT-High group ($p < 0.0001$ and $p < 0.001$ respectively), with smaller, but significant depletion of endothelial and T/NK cells ($p = 0.0098$ and $p = 0.0119$ respectively). CLQ values did not differ significantly between EMT strata for any cells, indicating preservation of small-radius co-localisation. Not surprisingly, Ripley's L AUC was significantly increased for the EMT-high subgroup ($p = 0.0001$), reflecting the computational dependence of this metric on EMT cell abundance (Supplementary Table 5). In addition, fibroblast populations also exhibited higher Ripley's L AUC values in EMT-High tumours ($p = 0.0173$), consistent with enhanced local aggregation of fibroblast-rich stromal domains. Finally, EMT-High samples exhibited significantly higher degree centrality values ($p = 0.0001$), reflecting a denser network of EMT-derived connections within the tissue architecture, in accordance with the mathematical definition of this metric. Centrality changes in other cellular compartments were not statistically significant, indicating that this effect is specific to the increased representation and spatial integration of EMT-like cells rather than a global reorganisation of the spatial network. Overall, these results suggest that EMT-high tumours are characterised by large-scale depletion of EMT cells from stromal and immune neighbourhoods, preservation of short-range co-localisation, increased intra-population clustering of EMT and fibroblast cells, and higher network connectivity. These findings

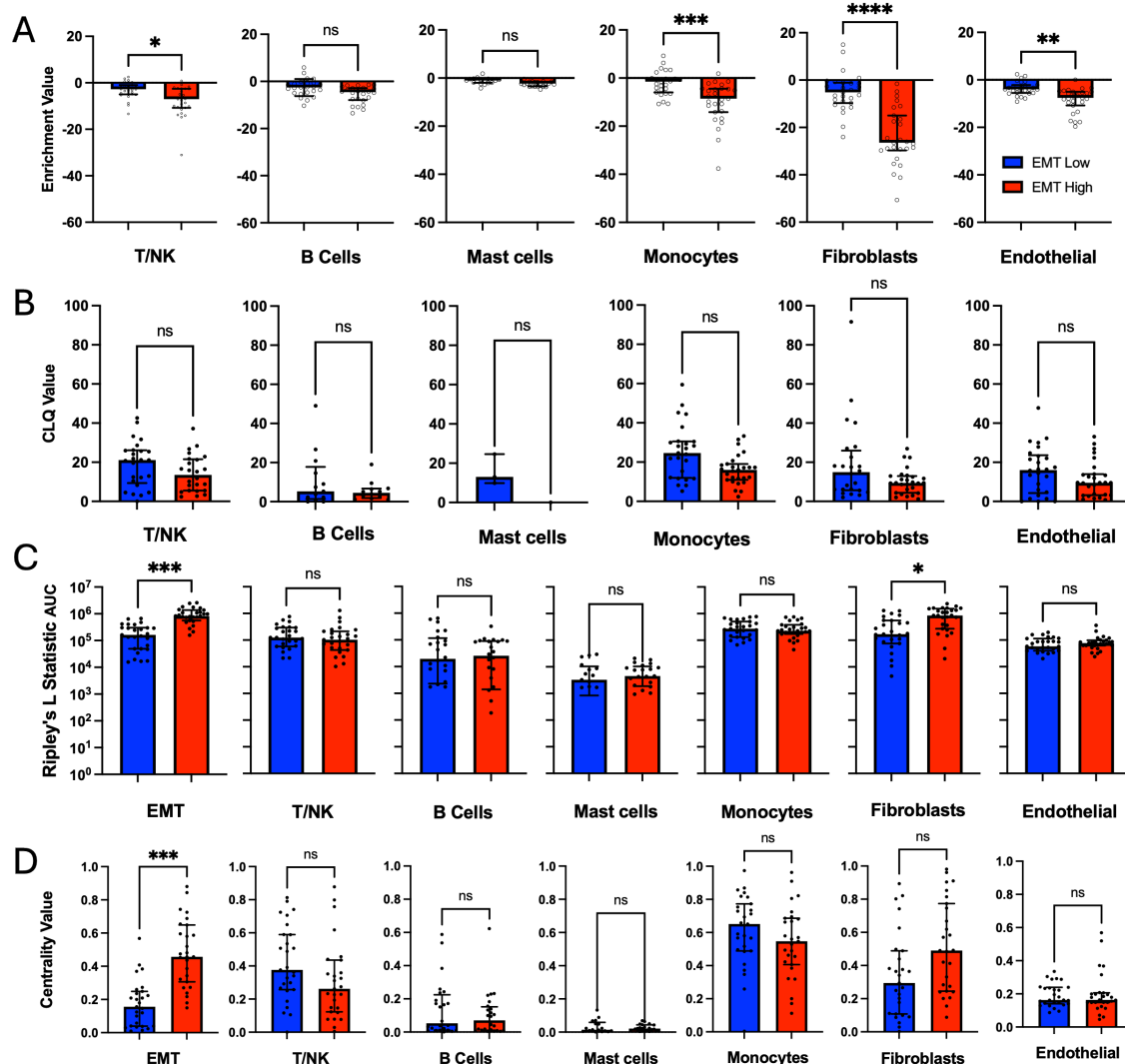


Figure 6: Comparison of spatial organisation metrics between EMT-high and EMT-low tumours, showing differences in neighborhood enrichment (A), cell–cell co-localization (CLQ) (B), Ripley’s L statistic (area under the curve) (C), and degree centrality (D) across major immune and stromal cell types. Individual data points represent biologically independent tumours. Bars indicate the median; error bars denote the interquartile range (25th–75th percentiles). Statistical comparisons were performed between EMT-high (red) and EMT-low (blue) groups using two-sided Kruskal–Wallis tests followed by Dunn’s multiple comparisons test with adjustment for multiple testing ($\alpha = 0.05$). Significant differences are indicated as: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), $p < 0.0001$ (****); *ns* = not significant. The number of biologically independent tumours per group (n_{low}/n_{high}) were as follows: (A) T/NK (27/27), B cells (24/24), Mast cells (17/21), Monocytes (27/27), Fibroblasts (27/27), Endothelial (27/27); (B) T/NK (25/24), B cells (13/12), Mast cells (3/1), Monocytes (25/27), Fibroblasts (24/27), Endothelial (26/27); (C–D) EMT (27/27), T/NK (27/27), B cells (24/24), Mast cells (17/21), Monocytes (27/27), Fibroblasts (27/27), Endothelial (27/27).

may indicate that EMT-high tumours undergo spatial reorganisation, marked by formation of locally clustered, mesenchymal-rich domains rather than homogeneous cellular mixing (full results tables are given in Supplementary Tables 7-10).

Survival Analysis

This section presents the results of the multi-faceted survival analysis. First, Kaplan–Meier survival analysis (Figure 7A) revealed that patients in the EMT-high group had significantly worse overall survival outcomes than those in the EMT-low group (log-rank test: $\chi^2 = 4.11$, $p = 0.043$). These

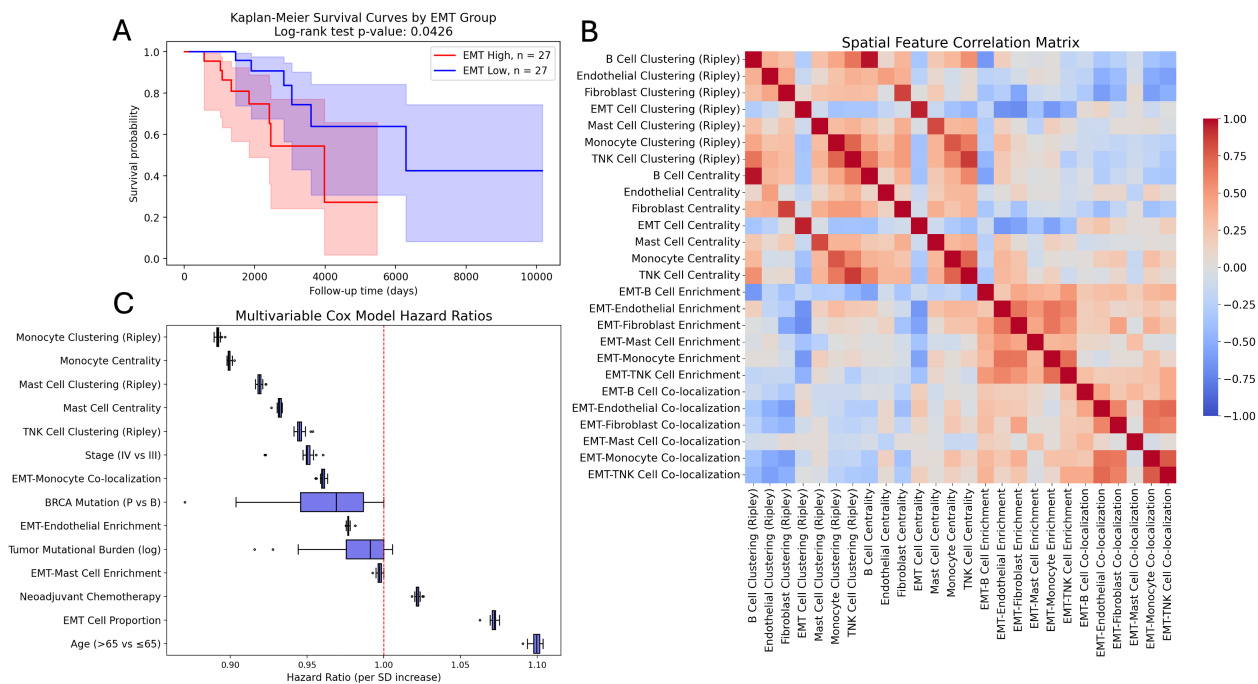


Figure 7: Survival analysis results. **(A)** - Kaplan–Meier curves stratified by EMT proportion (median split; High $n = 27$ vs Low $n = 27$; log-rank $p = 0.0426$). The shaded areas denote the 95% confidence intervals. Notably, this analysis confirms that the modified EMT signature used in this study preserved the prognostic stratification originally reported in the OxC study [6], with EMT-high tumours exhibiting significantly poorer overall survival. **(B)** - Spearman rank correlation matrix candidate spatial tumour microenvironment features. **(C)** Hazard ratios (HRs) from the final penalized Cox model integrating baseline clinical variables, EMT cell proportions, and top-performing spatial features ($n = 54$ biologically independent tumours; boxplots show the distribution of HRs across $m = 20$ multiply-imputed datasets). For boxplots, boxes represent the interquartile range (25th–75th percentiles) with the median indicated by the central line; whiskers denote the 2.5–97.5 percentile range, and outliers are shown as individual points. The vertical red dashed line at unity HR denotes no association with hazard; values to the left indicate decreased risk, whereas values to the right denote increased risk.

results confirm that EMT-high status is associated with poor prognosis, consistent with findings from the OxC study [6], further validating the use of the EMT proportion as an independent prognostic factor in HGSOc. Additionally, to verify that the modified 20-gene signature matrix used in this study retains the same prognostic discriminative ability as the original 52-gene OxC panel, an external validation was performed on *non-spatial* bulk-expression cohorts drawn from the original OxC study [6], combining eight independent datasets, including The Cancer Genome Atlas (TCGA) dataset [11], the Australian Ovarian Cancer Study (AOCS) dataset [30] and six additional microarray datasets from the CuratedOvarianData database [31]. EMT proportions were re-estimated using the 20-gene signature matrix and dichotomised by the within-cohort median. In the pooled meta-cohort ($n = 1657$), EMT-high cases again exhibited significantly inferior overall survival (unstratified log-rank: $\chi^2 = 20.57$, $p = 5.76 \times 10^{-6}$; stratified by study: $\chi^2 = 21.15$, $p = 4.24 \times 10^{-6}$). Full details and the corresponding Kaplan–Meier curves are provided in Supplementary Section 2 and Supplementary Figure 7.

Spatial Metrics Prognostics

As described in the Methods Section, the baseline elastic net Cox model—incorporating age, stage, neoadjuvant chemotherapy status, BRCA1/2 mutation status, and log-transformed TMB - was first tuned for regularisation strength. At the optimal penalty ($\alpha = 0.61$), the mean out-of-bag C-index was 0.506, indicating near-random discrimination and confirming that clinical and genomic covari-

ates alone provide limited prognostic separation in this cohort. Refitting at this penalty across all 20 imputed datasets yielded an apparent Harrell’s C-index of 0.670 ± 0.017 . Incorporating EMT cell proportions raised this to 0.713 ± 0.015 —an absolute gain of $\sim 6.4\%$ —demonstrating that EMT burden contributes independent prognostic information beyond conventional factors, with higher EMT proportions consistently associated with poorer survival (pooled HR = 1.101 ± 0.003).

Following correlation-based filtering (Figure 7B), individual spatial metrics were systematically added to the EMT-augmented model. Monocyte spatial organisation emerged as the dominant prognostic signal: both degree centrality and Ripley’s L -based clustering achieved the highest discriminative performance (C-index: 0.771 ± 0.013 and 0.761 ± 0.012 , respectively), with complete selection stability across imputations and hazard ratios indicating a protective association (HR ≈ 0.84 – 0.85). Mast cell and T/NK cell organisation provided additional, consistent prognostic information at slightly lower discrimination (C-index range: 0.731 – 0.743), while interaction metrics between EMT cells and immune or endothelial populations offered only marginal improvements—providing suggestive but limited evidence for the prognostic relevance of direct tumour–immune spatial relationships. Individual feature contributions are detailed in Supplementary Table 13.

In the final integrated model, spatial metrics dominated the prognostic landscape. The strongest protective effects came from monocyte organisation, with clustering and degree centrality each corresponding to an approximate 10–11% hazard reduction per standard deviation increase (HR = 0.89 and 0.90). Mast cell and T/NK cell spatial features contributed additional, smaller protective effects (HR = 0.92 – 0.95). Among non-spatial covariates, age and EMT proportion remained adverse prognostic factors (HR = 1.10 and 1.07), while stage and BRCA mutation status showed paradoxical protective associations (HR = 0.95 and 0.96), likely reflecting treatment-related confounding.

Discussion

This paper proposed a sequential approach for extending the OxC classification framework into the spatial domain by developing a deep learning–based deconvolution method specifically tailored for spatial transcriptomics data. As such, here, we firstly introduced the Sig-ZIB-VAE model as a deep generative modeling approach for spatially aggregated gene expression data, by combining a zero-inflated Beta likelihood with amortised variational inference, thereby explicitly capturing zero-inflation, while respecting the $[0, 1]$ support for proportion-like inputs.

It should be noted that in contrast to popular single-cell VAE-based methods that assume count distributions (such as NB/ZINB in scVI/scVAE) [18, 32], our Beta parametrisation is tailored to normalised measurements and should not be confused with the disentanglement-oriented β -VAE family [33]. By coupling the decoder with signature-guided supervision in the latent space, Sig-ZIB-VAE leverages prior biological knowledge, analogously to reference-driven deconvolution and annotation frameworks - such as CIBERSORT/MuSiC for bulk data and SPOTlight, Stereoscope, and cell2location for spatial assays - yet retains the benefits of a fully generative model (uncertainty-aware reconstructions, data simulation, and joint optimization of reconstruction and classification) [15, 34, 35, 36, 37].

Compared with regression-based deconvolution approaches, the multi-task objective presented here encourages biologically structured latent space (akin to weakly supervised approaches like CellAssign/scANVI), while scaling efficiently to large datasets via mini-batch stochastic optimisation [18, 38, 39]. Moreover, the generative modeling approach presented here offers significant advantages for future single-cell applications. By explicitly learning the underlying gene expression data, the model can also generate pseudo-realistic samples that preserve both sparsity patterns and continuous expression characteristics. This capability enables principled data augmentation to address class imbalance issues, which are common in clinical datasets and improve model robustness to out-of-distribution samples from different experimental conditions or technologies, another critical limitation identified by [40]. Furthermore, the generative framework facilitates privacy-preserving data sharing and provides a pathway to develop more robust and generalizable predictive models, as single-cell studies increasingly move towards clinical applications with inherently limited samples sizes and high vari-

ability. Finally, grounding supervision in FTE-derived reference signatures aligns the representation with cell states known to stratify HGSOV, including EMT programs linked to adverse outcomes, thereby supporting downstream spatial analyses and prognostic interpretation [7].

Following deconvolution analysis, spatial quantification of EMT cell distributions uncovered insights into tumour microenvironment organisation, revealing a non-uniform distribution of EMT cells in HGSOV tissues. The results show that EMT cells are broadly depleted from stromal and immune neighbourhoods, yet maintain persistent clustering and selective integration within the tissue network - evidenced by higher EMT and fibroblast clustering, co-localisation with selected cell populations (notably monocytes), and increased degree centrality. These findings suggest that EMT-high tumours exhibit distinct spatial organisation characterized by locally clustered, mesenchymal-rich domains rather than homogeneous cellular distribution.

The spatial organisation of the cells surveyed in this study appears to be consistent with recent spatial transcriptomic studies demonstrating that tissue architecture, beyond bulk compositional analysis, correlates with clinical outcomes, where tumour-stromal niches and cellular neighbourhoods may act as functional units of disease progression [41]. Notably, Withnell et al. [42] identified EMT hotspots in breast cancer that colocalize with hypoxic/angiogenic regions characterized by immune cell insulation. Although our study did not directly evaluate hypoxia-associated markers, the observed spatial distribution of EMT cells - forming dense homotypic clusters that are broadly depleted from stromal and immune neighbourhoods yet maintain selective monocyte co-localization at their peripheries - suggests that similar microenvironmental pressures may underlie spatially compartmentalized mesenchymal programs in HGSOV. However, direct validation of this hypothesis will require future study. Specifically in HGSOV, Zhang et al. [43] revealed that changes in immune cell localization and stromal architecture are associated with early relapse, emphasizing the clinical relevance of spatial organisation beyond cellular abundance.

Beyond describing network topology, the observed spatial patterns suggest distinct biological modes of EMT organisation in HGSOV. EMT-like malignant cells were broadly depleted from stromal and immune neighbourhoods at the tissue scale, yet consistently formed locally clustered, mesenchymal-rich domains that retained selective short-range proximity to monocytes. This multi-scale phenotype argues against indiscriminate mixing and instead supports a model in which EMT programs are stabilised within spatially segregated niches. The observed clustering is consistent with prior evidence that EMT programs can retain intercellular adhesion and promote collective tumour cell behaviour rather than isolated dissemination [44, 45], a mode of organisation relevant to ovarian cancer progression where multicellular spheroids contribute to peritoneal spread [46].

The preservation of EMT-monocyte co-localisation at short spatial radii, despite global depletion, suggests selective interactions at niche boundaries rather than within fully immune-infiltrated regions. One possible mechanism, supported by prior studies, is paracrine crosstalk between mesenchymal tumour cells and macrophage-lineage cells, in which EMT-associated programs recruit monocytes while macrophage-derived factors reinforce mesenchymal transcriptional states [47, 48]. In HGSOV, tumour-associated macrophages also contribute to extracellular matrix remodelling through secretion of proteins such as TGFBI and tenascin C [49]. Conversely, the large-scale depletion of EMT cells from T/NK and B cell neighbourhoods aligns with immune exclusion phenotypes linked to EMT, including TGF- β -mediated restriction of cytotoxic lymphocyte infiltration [50], consistent with the original OxC association between EMT and immunosuppression [6].

Notably, monocyte spatial organisation - rather than EMT-monocyte proximity alone - emerged as the strongest protective prognostic feature, indicating that immune architecture plays a dominant role in shaping clinical outcomes. Higher monocyte clustering and network integration may reflect more organised immune configurations, which in other studies have been associated with improved prognosis in HGSOV [51], whereas dispersed monocyte distributions may indicate diffuse immunosuppressive infiltration. Together, these findings reinforce the concept that spatial organisation of tumour-immune interactions, rather than cellular composition alone, is a critical determinant of disease trajectory.

Critically, the survival analysis showed that EMT-high status (i.e., greater EMT proportion) was associated with poorer prognosis—consistent with previous OxC studies [7, 8] - thereby validating EMT proportion as an independent prognostic factor in HGSOC. Moreover, the spatial organisation of immune cell populations - particularly monocytes - also emerged as a dominant determinant of outcome. Among all features, monocyte integration within their local neighborhood (captured by degree centrality) and their spatial clustering (quantified using Ripley’s L AUC) showed the strongest and most consistent prognostic associations, aside from EMT cell proportion and age covariates. Additional protective effects were observed for mast cell spatial metrics and T/NK cell clustering. This hierarchy of prognostic significance may further suggest that the spatial architecture of the tumour-immune interface supersedes cellular composition alone in determining clinical outcomes. Overall, these findings support our initial hypothesis that the OxC classification—originally derived from bulk transcriptomic data—encodes spatially meaningful biological organisation when projected onto high-resolution tissue data, thereby bridging molecular classification with spatial tissue ecology.

Having presented the outcomes of the OxC extension to the spatial domain, it is also important to acknowledge the limitations of the present work and outline potential future directions. While this analysis employed a subset of the OxC gene panel overlapping with the CosMx signatures, a natural next step would be to expand the evaluation to the full OxC repertoire to capture the complete transcriptional heterogeneity originally defined by Hu et al. [6].

Moreover, several aspects of model evaluation warrant careful consideration. Model performance assessment employed held-out samples from the same spatial transcriptomics dataset, which, while suitable for internal benchmarking, may not fully represent the model’s generalization across independent cohorts or experimental conditions. Although the resulting high classification accuracy is encouraging, it remains derived from pseudo-labels defined by correlations with established signatures rather than ground-truth cell annotations. This circular validation approach - common to deconvolution studies - constrains the ability to quantify absolute classification performance.

Additionally, three methodological considerations affect the interpretation of spatial patterns. First, the grid-based spatial aggregation strategy used for deconvolution may contribute to the observed degree of self-clustering, as cells within the same grid square receive identical program assignments by design. While this approach enhances statistical power for signature detection, it may amplify local homotypic clustering patterns relative to true single-cell-resolution spatial organisation, and the extent to which grid-level assignments versus genuine biological clustering drive the observed EMT self-aggregation patterns warrants further investigation using cell-level deconvolution methods. Second, given the limited scale of the analysed fields of view, the findings are best interpreted as capturing local cellular interactions rather than full tissue architecture. At this mesoscopic scale, spatial statistics reflect intra-regional patterns of cell organisation—such as clustering, adjacency, and network connectivity—but cannot resolve macro-architectural gradients across entire tumour sections. Third, while degree centrality quantifies the overall inter-cluster connectivity of a cell population, it does not distinguish which cell types constitute those connections. To address this, we tallied the cell-type composition of non-EMT neighbours at EMT cluster boundaries and compared it to the background tissue composition, revealing selective monocyte over-representation (Supplementary Table 11). The majority of heterotypic connections nonetheless involved other malignant subtypes, and alternative graph-theoretic metrics such as betweenness centrality or modularity-based community detection may provide complementary insights into the relative contributions of malignant cell mixing and immune-stromal boundary interactions [52]. Thus, identifying biologically relevant tissue regions and quantifying intercellular relationships at the appropriate spatial scale remain important open challenges for future spatially resolved studies [42].

Furthermore, independent validation of spatial clustering and prognostic findings in external cohorts represents an important priority which is, however, currently constrained by data availability. This study represents the first attempt to extend the OxC classification framework to spatial transcriptomics, and to our knowledge, no independent spatial transcriptomics datasets currently exist that simultaneously provide: (1) HGSOC samples profiled at single-cell resolution, (2) gene panels over-

lapping with the OxC signature repertoire, (3) matched clinical outcomes data, and (4) comparable tissue sampling strategies (e.g., TMA-based FOVs capturing tumour microenvironmental heterogeneity). While the EMT prognostic signal was successfully validated across eight independent bulk transcriptomics cohorts (Supplementary Section 2), confirming that the modified 20-gene signature retains discriminative power, the spatial organisation patterns and spatial metric prognostics require replication in future spatial profiling studies as such datasets become available.

Data Availability

Spatial transcriptomics data were obtained from the Discovery dataset described in Yeh et al. [9] and available on Zenodo [14] (<https://doi.org/10.1038/s41590-024-01943-5>). The procedure for acquiring the signature matrix used in [6] can be found on the Github repository (<https://github.com/zhiyhu/scFT-paper>). Source data for Figures 3-7 are provided in Supplementary Data 1.

Code Availability

The code used to generate the results presented in this paper is made available at a Github repository (https://github.com/cwcyau/oxc_spatial) and are also archived on Figshare [53] (<https://doi.org/10.6084/m9.figshare.31988754>).

Acknowledgements

We thank Professor Ahmed Ahmed for discussions about this work.

Funding

The authors are supported by an EPSRC Turing AI Acceleration Fellowship (Grant Ref: EP/V023233/1) and through project NIHR173695 funded by the Better Methods, Better Research (BMBR) programme, an MRC and NIHR partnership. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Author Contributions Statement

AS proposed and implemented the method, performed numerical experiments, wrote the initial draft and revised the manuscript. CY supervised the project, revised the initial draft, and wrote the final manuscript.

Competing Interests Statement

The authors declare no competing interests.

References

- [1] Ursula A Matulonis, Anil K Sood, Lesley Fallowfield, Brooke E Howitt, Jalid Sehouli, and Beth Y Karlan. Ovarian cancer. *Nature reviews Disease primers*, 2(1):1–22, 2016.
- [2] Michelle McMullen, Katherine Karakasis, Robert Rottapel, and Amit M Oza. Advances in ovarian cancer, from biology to treatment. *Nature Cancer*, 2(1):6–8, 2021.
- [3] Huy Q Dinh, Xianzhi Lin, Forough Abbasi, Robbin Nameki, Marcela Haro, Claire E Olingy, Heidi Chang, Lourdes Hernandez, Simon A Gayther, Kelly N Wright, et al. Single-cell transcriptomics identifies gene expression networks driving differentiation and tumorigenesis in the human fallopian tube. *Cell reports*, 35(2), 2021.
- [4] Yeh Wang, Peng Huang, Brant G Wang, Tricia Murdock, Leslie Cope, Fang-Chi Hsu, Tian-Li Wang, and Ie-Ming Shih. Spatial transcriptomic analysis of ovarian cancer precursors reveals reactivation of *igfbp2* during pathogenesis. *Cancer research*, 82(24):4528–4541, 2022.
- [5] Tanjina Kader, Jia-Ren Lin, Clemens B Hug, Shannon Coy, Yu-An Chen, Ino de Bruijn, Natalie Shih, Euihye Jung, Roxanne J Pelletier, Mariana Lopez Leon, et al. Multimodal spatial profiling reveals immune suppression and microenvironment remodeling in fallopian tube precursors to high-grade serous ovarian carcinoma. *Cancer Discovery*, 15(6):1180–1202, 2025.
- [6] Zhiyuan Hu et al. The repertoire of serous ovarian cancer non-genetic heterogeneity revealed by single-cell sequencing of normal fallopian tube epithelial cells. *Cancer cell*, 37(2):226–242, 2020.
- [7] Zhiyuan Hu, Paula Cunnea, Zhe Zhong, Haonan Lu, Oloruntoba I Osagie, Leticia Campo, Mara Artibani, Katherine Nixon, Jennifer Ploski, Laura Santana Gonzalez, et al. The oxford classic links epithelial-to-mesenchymal transition to immunosuppression in poor prognosis ovarian cancers. *Clinical Cancer Research*, 27(5):1570–1579, 2021.
- [8] Lena Rai, Antonella Ravaggi, Eliana Bignotti, Robert L Hollis, Dale W Garsed, Ahwan Pandey, Kyriaki Barbara Papalois, Faheemah Patel, Yasmin Kamel, Leticia Campo, et al. Oxford classic-defined emt risk stratification of high-grade serous ovarian cancer for guiding treatment decisions. *Clinical Cancer Research*, pages OF1–OF15, 2025.
- [9] Christine Yiwen Yeh et al. Mapping spatial organization and genetic cell-state regulators to target immune evasion in ovarian cancer. *Nature immunology*, 25(10):1943–1958, 2024.
- [10] Shanshan He, Ruchir Bhatt, Carl Brown, Emily A Brown, Derek L Buhr, Kan Chantranuvatana, Patrick Danaher, Dwayne Dunaway, Ryan G Garrison, Gary Geiss, et al. High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature biotechnology*, 40(12):1794–1806, 2022.
- [11] Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609, 2011.
- [12] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.
- [13] Chloé B Steen, Chih Long Liu, Ash A Alizadeh, and Aaron M Newman. Profiling cell type abundance and expression in bulk tissues with cibersortx. *Stem Cell Transcriptional Networks: Methods and Protocols*, pages 135–157, 2020.
- [14] Christine Yiwen Yeh et al. Mapping spatial organization and genetic cell state regulators to target immune evasion in ovarian cancer, 2024. URL <https://zenodo.org/records/12613839>.
- [15] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

- [16] Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.
- [17] Yanshuo Chen, Yixuan Wang, Yuelong Chen, Yuqi Cheng, Yumeng Wei, Yunxiang Li, Jiuming Wang, Yingying Wei, Ting-Fung Chan, and Yu Li. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nature Communications*, 13(1):6735, 2022.
- [18] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.
- [19] Tian Tian, Martin Renqiang Min, and Zhi Wei. Model-based autoencoders for imputing discrete single-cell rna-seq data. *Methods*, 192:112–119, 2021.
- [20] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.
- [21] Denis Schapiro, Hartland W Jackson, Swetha Raghuraman, Jana R Fischer, Vito RT Zanotelli, Daniel Schulz, Charlotte Giesen, Raúl Catena, Zsuzsanna Varga, and Bernd Bodenmiller. histocat: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods*, 14(9):873–876, 2017.
- [22] Weiruo Zhang, Irene Li, Nathan E Reticker-Flynn, Zinaida Good, Serena Chang, Nikolay Samusik, Saumyaa Saumyaa, Yuanyuan Li, Xin Zhou, Rachel Liang, et al. Identification of cell types in multiplexed in situ images by combining protein expression and spatial information using celesta. *Nature methods*, 19(6):759–769, 2022.
- [23] Stephen P Borgatti, Filip Agneessens, Jeffrey C Johnson, and Martin G Everett. *Analyzing social networks*. SAGE publications Ltd, 2024.
- [24] Mary L Swift. Graphpad prism, data analysis, and scientific graphing. *Journal of chemical information and computer sciences*, 37(2):411–412, 1997.
- [25] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL <http://jmlr.org/papers/v21/20-729.html>.
- [26] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.
- [27] Wojtek Krzanowski. *Principles of multivariate analysis*, volume 23. OUP Oxford, 2000.
- [28] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.
- [29] Erich Kummerfeld, Leland Williams, Yinzhaoh Wang, Samuel T Peters, Elizabeth Schmidt, Mickayla DuFresne-To, David Bernlohr, Paul Robbins, Sayeed Ikramuddin, Oyedele Adeyi, et al. Artifacts in spatial transcriptomics data: their detection, importance, prevalence, and prevention. *Briefings in Bioinformatics*, 26(4):bbaf306, 2025.
- [30] Richard W Tothill, Anna V Tinker, Joshy George, Robert Brown, Stephen B Fox, Stephen Lade, Daryl S Johnson, Melanie K Trivett, Dariush Etemadmoghadam, Bianca Locandro, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical cancer research*, 14(16):5198–5208, 2008.
- [31] Benjamin Frederick Ganzfried, Markus Riester, Benjamin Haibe-Kains, Thomas Risch, Svitlana Tyekucheva, Ina Jazic, Xin Victoria Wang, Mahnaz Ahmadifar, Michael J Birrer, Giovanni Parmigiani, et al. curatedovariandata: clinically annotated data for the ovarian cancer transcriptome. *Database*, 2013, 2013.

- [32] Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020.
- [33] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.
- [34] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):380, 2019.
- [35] Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu, Ivo Gut, and Holger Heyn. Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic acids research*, 49(9):e50–e50, 2021.
- [36] Alma Andersson, Joseph Bergensträhle, Michaela Asp, Ludvig Bergensträhle, Aleksandra Jurek, José Fernández Navarro, and Joakim Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications biology*, 3(1):565, 2020.
- [37] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology*, 40(5):661–671, 2022.
- [38] Allen W Zhang, Ciara O’Flanagan, Elizabeth A Chavez, Jamie LP Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, et al. Probabilistic cell-type assignment of single-cell rna-seq for tumor microenvironment profiling. *Nature methods*, 16(10):1007–1015, 2019.
- [39] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1):e9620, 2021.
- [40] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.
- [41] Elaine Stur, Sara Corvigno, Mingchu Xu, Ken Chen, Yukun Tan, Sanghoon Lee, Jinsong Liu, Emily Ricco, Daniel Kraushaar, Patricia Castro, et al. Spatially resolved transcriptomics of high-grade serous ovarian carcinoma. *Science*, 25(3), 2022.
- [42] Eloise Withnell and Maria Secrier. Spottedpy quantifies relationships between spatial transcriptomic hotspots and uncovers environmental cues of epithelial-mesenchymal plasticity in breast cancer. *Genome Biology*, 25(1):289, 2024.
- [43] Yuanfu Zhang, Shu Sun, Yue Qi, Yifan Dai, Yangyang Hao, Mengyu Xin, Rongji Xu, Hongyan Chen, Xiaoting Wu, Qian Liu, et al. Characterization of tumour microenvironment reprogramming reveals invasion in epithelial ovarian carcinoma. *Journal of Ovarian Research*, 16(1):200, 2023.
- [44] Kevin J Cheung and Andrew J Ewald. A collective route to metastasis: Seeding by tumor cell clusters. *Science*, 352(6282):167–169, 2016.
- [45] Nicole M Aiello and Yibin Kang. Context-dependent emt programs in cancer metastasis. *Journal of Experimental Medicine*, 216(5):1016–1026, 2019.
- [46] Marcin P Iwanicki, Rachel A Davidowitz, Mei Rosa Ng, Achim Besser, Taru Muraanen, Melissa Merritt, Gaudenz Danuser, Tan Ince, and Joan S Brugge. Ovarian cancer spheroids use myosin-generated force to clear the mesothelium. *Cancer discovery*, 1(2):144–157, 2011.
- [47] Shicheng Su, Qiang Liu, Jingqi Chen, Jianing Chen, Fei Chen, Chonghua He, Di Huang, Wei

- Wu, Ling Lin, Wei Huang, et al. A positive feedback loop between mesenchymal-like cancer cells and macrophages is essential to breast cancer metastasis. *Cancer cell*, 25(5):605–620, 2014.
- [48] Bin-Zhi Qian and Jeffrey W Pollard. Macrophage diversity enhances tumor progression and metastasis. *Cell*, 141(1):39–51, 2010.
- [49] Anna Mary Steitz, Alina Steffes, Florian Finkernagel, Annika Unger, Leah Sommerfeld, Julia M Jansen, Uwe Wagner, Johannes Graumann, Rolf Müller, and Silke Reinartz. Tumor-associated macrophages promote ovarian cancer cell migration by secreting transforming growth factor beta induced (tgfb β) and tenascin c. *Cell death & disease*, 11(4):249, 2020.
- [50] Sanjeev Mariathasan, Shannon J Turley, Dorothee Nickles, Alessandra Castiglioni, Kobe Yuen, Yulei Wang, Edward E Kadel Iii, Hartmut Koeppen, Jillian L Astarita, Rafael Cubas, et al. Tgf β attenuates tumour response to pd-11 blockade by contributing to exclusion of t cells. *Nature*, 554(7693):544–548, 2018.
- [51] Lin Zhang, Jose R Conejo-Garcia, Dionyssios Katsaros, Phyllis A Gimotty, Marco Massobrio, Giorgia Regnani, Antonis Makrigiannakis, Heidi Gray, Katia Schlienger, Michael N Liebman, et al. Intratumoral t cells, recurrence, and survival in epithelial ovarian cancer. *New England journal of medicine*, 348(3):203–213, 2003.
- [52] Qi Gao, Shasha Yuan, Shengjun Li, and Juan Wang. Integrating modularity maximization and contrastive learning for identifying spatial domain from spatial transcriptomics. *Expert Systems with Applications*, page 129426, 2025.
- [53] Alexandru Stihi and Christopher Yau. Understanding the spatial determinants of the oxford classic prognostic signature for high-grade serous ovarian cancer., 2026. URL <https://doi.org/10.6084/m9.figshare.31988754>.