Does spatial organisation of the tumour microenvironment add stable prognostic information beyond cell composition? A patient-grouped, nested-validation study of breast-cancer imaging mass cytometry

July 2026

Abstract

Background. Highly multiplexed tissue imaging has made it possible to quantify not only *which* cells populate a breast tumour but *how* tumour, stromal and immune cells are spatially arranged. A recurring claim in spatial biology is that this micro-architecture carries prognostic information beyond simple cell abundance. We tested this claim under a deliberately strict, leakage-controlled evaluation, framing the analysis as a decision about whether a spatial-feature panel should advance to external validation rather than as an attempt to build a deployable clinical predictor.

Methods. Using Version 2 of the publicly archived Basel imaging mass cytometry (IMC) cohort (Zenodo `10.5281/zenodo.4607374`; 285 patients, 376 tissue-microarray cores, 844,498 phenotyped single cells), we first validated single-cell phenotypes, coordinates and tumour/stroma masks on a stratified subset of seven cores spanning all four clinical subtypes and three histological grades. We then computed a locked panel of edge-corrected marked spatial statistics—cross-type Ripley's L deviations at $r = 20/50/100$ px and neighbourhood-enrichment z-scores calibrated by 1000 within-image label permutations—aggregated cores to patients by cell-count-weighted means, and compared three strictly nested Elastic-Net Cox proportional-hazards models (A: clinical; B: +composition; C: +spatial) under patient-grouped 5×5 nested cross-validation. Out-of-fold (OOF) discrimination was summarised by Harrell's C-index with 1000-resample bootstrap 95% confidence intervals (CIs).

Results. Adding cell-type composition to the clinical baseline produced a stable improvement in OOF discrimination (Model B C-index = 0.764 [0.712–0.813] versus Model A 0.725 [0.668–0.779]; $\Delta\text{C-index}_{B-A} = +0.039$, 95% CI [+0.006, +0.073], excluding zero). In contrast, the full 33-feature spatial panel *degraded* discrimination relative to composition ($\Delta\text{C-index}_{C-B} = -0.029$ [−0.078, +0.011]) and provided no gain over the clinical baseline ($\Delta\text{C-index}_{C-A} = +0.010$ [−0.046, +0.060]). Model C carried 33 features for only 80 events (feature-to-event ratio 0.41), and the multi-scale Ripley's L statistics were strongly collinear.

Conclusions. On a cohort of this scale, high-dimensional spatial descriptors did not add *stable* prognostic value and, if anything, harmed out-of-fold generalisation, whereas cell-type composition did. We therefore return a **SIMPLIFY** decision: advance the clinical+composition model—and, optionally, a highly reduced spatial subset—to external validation, rather than the full spatial panel. The study illustrates how patient-grouped nested validation guards against a common optimism in spatial-prognostics research.

Keywords: imaging mass cytometry; breast cancer; tumour microenvironment; spatial statistics; Ripley's L; neighbourhood enrichment; survival analysis; nested cross-validation; prognostic modelling.

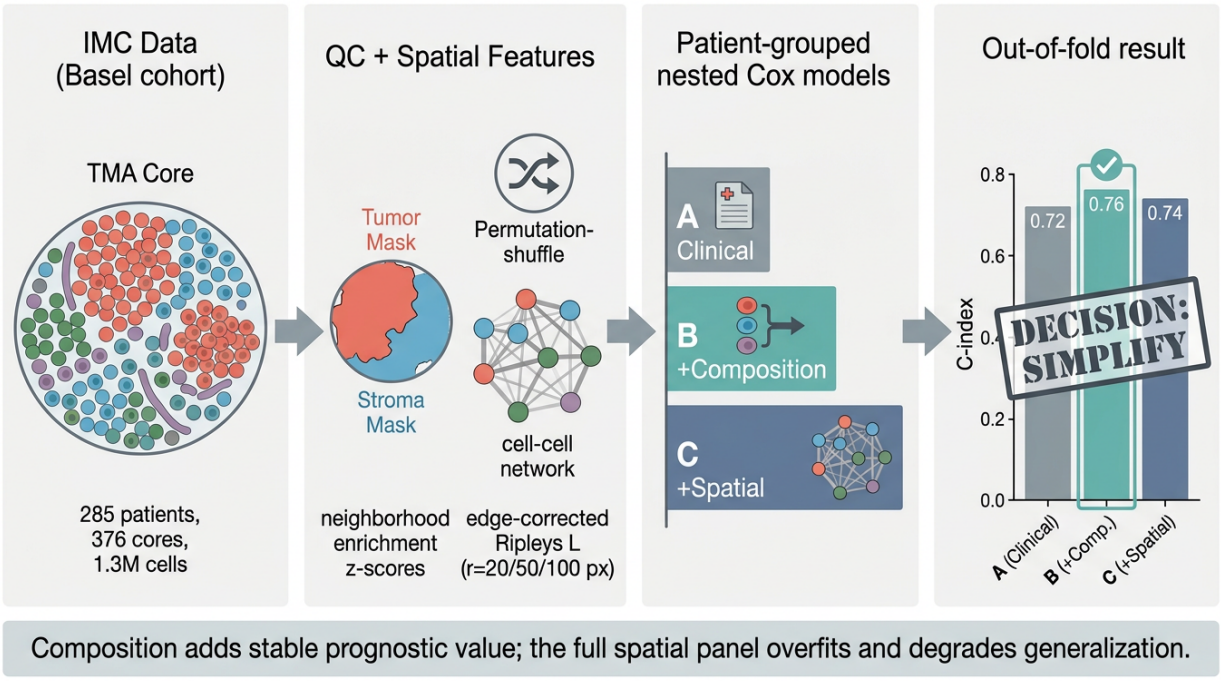


Figure 1: Graphical abstract. From single-cell IMC data of the Basel breast-cancer cohort, phenotypes, masks and coordinates are validated on a stratified subset; permutation-calibrated neighbourhood enrichment and edge-corrected Ripley's L features are computed and aggregated to patients; three strictly nested Elastic-Net Cox models are compared under patient-grouped nested cross-validation. Out-of-fold discrimination is highest for the clinical+composition model (Model B); the full spatial panel (Model C) does not improve—and slightly degrades—generalisation, motivating a SIMPLIFY decision for external validation.

1 Introduction

Breast cancer is the most commonly diagnosed malignancy worldwide, with an estimated 2.3 million new cases and 685,000 deaths in 2020, making accurate prognostication a persistent clinical priority [1]. Contemporary risk stratification rests on a compact set of pathology variables— tumour size, histological grade, nodal status and receptor status [2, 3]—and on the molecular intrinsic subtypes that reorganised our understanding of the disease two decades ago [4, 5]. These variables are robust, inexpensive and reproducible, but they summarise the tumour largely as a bulk entity and are agnostic to the cellular society in which malignant cells reside.

A substantial body of work now establishes that the composition of the immune infiltrate is prognostic. Tumour-infiltrating lymphocytes (TILs) predict outcome and treatment response, particularly in triple-negative and HER2-positive disease, and standardised scoring guidelines have been issued [6, 7]. In colorectal cancer the "Immunoscore"—built from the type, density and location of cytotoxic and memory T cells—outperforms classical TNM staging, and comparable spatially-resolved immune metrics have been proposed across tumours [8, 9]. These observations have crystallised into a broader hypothesis: that the *spatial organisation* of the tumour microenvironment (TME)—how tumour, stromal and immune populations are geometrically arranged relative to one another—encodes clinically relevant information that bulk composition alone misses [10].

Highly multiplexed tissue imaging has made this hypothesis directly testable. Imaging mass cytometry (IMC) couples laser ablation with mass-spectrometric detection of metal-tagged antibodies

to quantify dozens of proteins at subcellular resolution while preserving tissue architecture [11]. Related platforms—multiplexed ion beam imaging (MIBI) and co-detection by indexing (CODEX)—have similarly enabled single-cell spatial maps of tumour tissue [12, 13]. Landmark studies using these technologies have argued that spatial structure is prognostic: Keren and colleagues distinguished "compartmentalised" from "mixed" immune architectures in triple-negative breast cancer that stratified survival [14]; Schürch and colleagues showed that coordinated cellular neighbourhoods at the colorectal invasive front predicted outcome beyond cell counts [15]; and quantitative multiplex immunohistochemistry linked myeloid-inflamed spatial complexity to poor prognosis [16]. In breast cancer specifically, the single-cell pathology landscape defined by IMC [17], its phenogenomic extension [18], and the recurrent TME "structures" associated with genomics and outcome [19] have all been interpreted as evidence that spatial architecture is an independent prognostic axis.

Yet the strength of this conclusion depends critically on how it is evaluated. Spatial-feature engineering is expansive: for four broad cell classes there are ten unordered neighbourhood pairs, each measurable at multiple length scales and with multiple summary statistics, so a spatial panel readily reaches dozens of correlated covariates. When these are entered into a survival model on a cohort with a limited number of events, several well-characterised statistical hazards converge. Low event-per-variable ratios inflate coefficient variance and overfitting risk [20, 21]; strong collinearity among multi-scale spatial descriptors destabilises coefficient estimates; and, most insidiously, reusing the same resampling procedure for feature/hyperparameter selection and for performance reporting produces optimistically biased estimates [22]. Repeated tissue cores from the same patient add a further trap: if replicate images are treated as independent, or if scaling and encoding are fit on the full dataset, information leaks across the train/test boundary. Reporting standards for prognostic models emphasise exactly these points [23–25].

In this study we ask a deliberately narrow and decision-oriented question: using the publicly archived Basel IMC cohort, does the spatial organisation of tumour, stromal and immune cells add *stable* prognostic information beyond cell composition and standard pathology variables, and—if so—should a spatial panel advance to external validation? We do not attempt to build a deployable clinical predictor. Instead we (i) validate phenotypes, masks and coordinates on a stratified image subset; (ii) engineer a locked panel of edge-corrected marked spatial statistics and permutation-calibrated neighbourhood enrichment; (iii) aggregate replicate cores to the patient level; and (iv) compare composition-only with composition-plus-spatial survival models under patient-grouped nested cross-validation, quantifying incremental out-of-fold performance with uncertainty. The output is a triaged *advance / simplify / stop* decision. As we show, the answer on this cohort is a rigorous negative: composition adds stable value, but the full spatial panel does not, and the discipline of nested validation is what makes that distinction visible.

2 Methods

Figure 2 summarises the end-to-end analysis. All computation used a fixed random seed (42); scaling and encoding were fit strictly within training partitions; and every intermediate panel was locked and versioned before modelling.

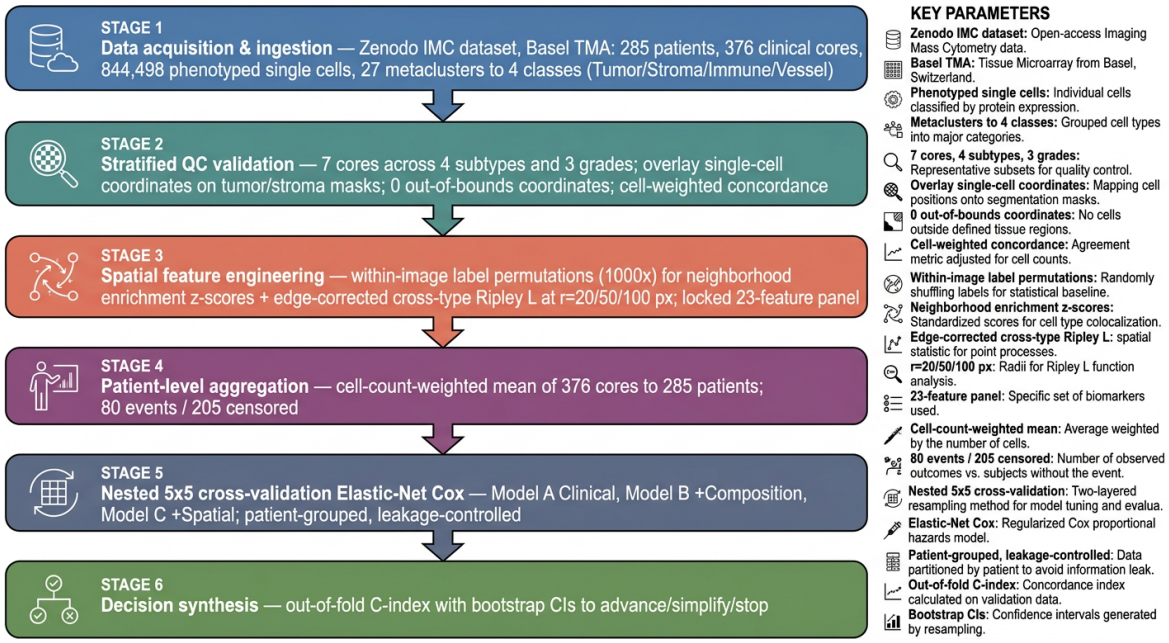


Figure 2: Analysis pipeline. Six locked stages from data acquisition to decision synthesis, with the principal parameters of each stage. Feature engineering, patient aggregation and the modelling comparison were frozen before out-of-fold performance was inspected.

2.1 Dataset and cohort definition

We used the imaging mass cytometry breast-cancer resource of Jackson, Fischer et al. [17], archived on Zenodo (record `4607374`; DOI `10.5281/zenodo.4607374`; Version 2). The resource comprises two tissue microarrays (TMAs)—a Basel and a Zurich cohort— profiled by IMC with a panel of metal-tagged antibodies, together with per-cell coordinates, per-cell metacluster phenotypes, tumour/stroma segmentation masks, and clinical metadata. Because the record totals approximately 48 GB and is dominated by raw image stacks and histoCAT sessions that are not required for cell-level survival analysis, essential comma-separated tables were streamed selectively from the largest archive via HTTP range requests, while the small label, location and mask archives were downloaded in full. Single-cell expression is stored in long format (one row per cell × marker); phenotype and location tables are joined on a unique cell identifier, cores on a core identifier, and images to patients on a patient identifier (PID). The per-cell coordinates and phenotypes derive from the resource's published single-cell segmentation and clustering pipeline; segmentation is itself a recognised source of variability in multiplexed imaging, and contemporary deep-learning segmenters now approach human-level performance [26].

Profiling confirmed 855,668 located Basel single cells across 381 imaged cores, of which 844,498 (98.7%) carry one of 27 annotated metaclusters that we mapped to four broad classes—Tumour, Stroma, Immune and Vessel—with no unmapped labels. Critically, overall-survival follow-up is populated only for the Basel cohort; the Zurich TMA lacks `OSmonth` and status entirely and therefore cannot serve as a survival cohort. All survival modelling was consequently restricted to Basel, and control cores present in the imaging tables but absent from patient metadata (e.g. liver and slide-control cores) were excluded. After restriction to the 376 Basel clinical cores present in the patient metadata and aggregation to unique patients (below), the analysis cohort comprised **285 patients with 80 death events and 205 censored observations** (28.1% event rate). Table 1

summarises the cohort.

Table 1: Analysis cohort (Basel IMC, Version 2). Composition percentages are cell-weighted over phenotyped cells; the four broad classes are the units of the spatial analysis.

Characteristic	**Value**
Patients (survival cohort)	285
Death events / censored	80 / 205
Event rate	28.1%
Clinical cores analysed	376
Cores per patient (1 / 2 / 3 / 4)	201 / 78 / 5 / 1
Phenotyped single cells (Basel)	844,498
Annotated metaclusters → broad classes	27 → 4
Broad classes	Tumour, Stroma, Immune, Vessel
Overall-survival follow-up (median)	75 months (range 0–233)
Clinical variables	age, tumour size, grade, ER, PR, HER2, subtype

2.2 Stratified quality control and validation

Before large-scale feature engineering, we verified that single-cell coordinates and phenotypes were consistent with the physical tumour/stroma masks. Each core's mask was located by transforming the full-stack filename (`_full.tiff`→`_full_mask_AllTumorFilled.tiff`), robustly mapping 367 of 376 Basel clinical cores (nine missing masks were recorded). Masks are binary `uint8` arrays in which value 1 denotes the filled tumour region and 0 the stroma/non-tumour region at approximately 1 μm/px; each cell was assigned the mask value at $\lfloor$Location_Center_Y$\rfloor$, $\lfloor$Location_Center_X$\rfloor$.

From eligible tumour cores (known clinical type; disease status "tumour"; grade $\in \{1, 2, 3\}$; ≥ 500 cells; mask tumour-fraction in $[0.05, 0.98]$), a deterministic greedy-coverage sampler selected seven cores spanning all four clinical subtypes (HR+HER2−, HR+HER2+, HR−HER2+, TripleNeg) and all three grades. For each selected core we produced a 300-dpi overlay of the semi-transparent tumour/stroma mask with single cells coloured by broad class, and computed cell-weighted concordance between phenotype class and mask region, together with coordinate sanity checks (out-of-bounds, negative and duplicate coordinates).

2.3 Permutation-based spatial feature engineering

We engineered a locked spatial panel for all 376 Basel clinical cores using coordinates and broad-class labels. Two families of statistics were computed to separate genuine spatial organisation from mere cell abundance.

Neighbourhood enrichment with within-image permutations. For each core we built a contact graph connecting cells within a 20-px ($\approx$20 μm) radius using a `cKDTree`. For each of the ten unordered class pairs (A, B) we counted observed contacts and compared them to a null distribution generated by **1000 within-image label permutations**, in which cell coordinates were held fixed and class labels were shuffled. The permutation loop was fully vectorised (batched pair enumeration with a code-offset `bincount` trick), allowing the whole cohort to run in under three minutes. The feature is the standardised enrichment

$$Z_{AB} = \frac{\mathrm{Obs}_{AB} - \mu_{AB}^{\mathrm{perm}}}{\sigma_{AB}^{\mathrm{perm}}}, \tag{1}$$

with $Z_{AB} = 0$ when $\sigma_{AB}^{\text{perm}} = 0$. Positive values indicate spatial attraction (co-localisation) and negative values avoidance (segregation). By construction, these permutation z-scores condition on the observed composition of each image and therefore isolate arrangement from abundance—the design principle underlying neighbourhood-permutation tests in tools such as histoCAT and Squidpy [27, 28].

Edge-corrected marked spatial statistics. We complemented enrichment with cross-type Ripley's statistics, the canonical second-order summaries of marked point processes [29, 30]. For the tumour–immune, tumour–stroma and immune–stroma pairs we estimated the border/reduced-sample edge-corrected cross-type K function, converted it to the variance-stabilised L function, symmetrised over class ordering, and reported the deviation from complete spatial randomness,

$$\Delta L_{AB}(r) = L_{AB}(r) - r, \tag{2}$$

at radii $r \in \{20, 50, 100\}$ px. Values above zero indicate co-localisation and below zero segregation at scale r. Edge correction follows established point-process practice for bounded windows [31, 32].

The locked panel (`spatial_feature_panel.csv`) contains 376 rows and 23 spatial features—4 composition fractions, 10 neighbourhood-enrichment z-scores and 9 cross-type Ripley's L deviations—with zero missing or infinite values and composition fractions summing to one. A machine-readable feature dictionary and all parameters were recorded alongside the panel.

2.4 Patient-level aggregation

Because survival is recorded per patient while imaging is per core, and 84 patients contributed replicate cores, the 376-core panel was aggregated to 285 patients. The primary aggregate is the **cell-count-weighted mean** of each core's features (weight = number of cells), which represents the pooled cellular microenvironment; for composition fractions this weighted mean equals the fraction obtained by pooling all of a patient's cells. An unweighted mean was stored as a secondary robustness variant. Composition fractions summed to one at the patient level (maximum deviation 4.4×10^{-16}). Clinical covariates (age, tumour size, grade, ER/PR/HER2 status, clinical type) were verified identical across a patient's replicate cores. Survival time was taken from `OSmonth`; the event indicator mapped death from any cause to 1 and alive/alive-with-metastases to 0. Survival columns contained no missing values; a small number of categorical covariates were mode-imputed (ERStatus ×1, PRStatus ×6, clinical type ×6), leaving zero missing values across all analysis columns.

2.5 Survival modelling and nested validation

We fit Elastic-Net penalised Cox proportional-hazards models [33–35] using `lifelines` [36]. Three strictly nested feature blocks were compared, each a superset of the previous:

Model A — Clinical (10 features): age, tumour size, grade, and one-hot-encoded ER, PR, HER2 status and clinical subtype.

Model B — +Composition (14 features): adds the Tumour, Stroma, Immune and Vessel fractions.

Model C — +Spatial (33 features): adds the 10 neighbourhood-enrichment z-scores and 9 cross-type Ripley's L deviations ($r = 20/50/100$ px).

Discrimination was estimated by **patient-grouped 5×5 nested cross-validation**. The outer 5-fold split (stratified on event × clinical type, with an event-only fallback) partitioned patients into held-out test folds; within each outer training partition an inner 5-fold loop grid-searched the penalty strength (`penalizer`$\in \{0.01, 0.05, 0.1, 0.5, 1.0\}$) and mixing parameter (`l1_ratio`$\in \{0.1, 0.5, 0.9\}$), selecting the combination with the best mean inner-fold Harrell C-index. `StandardScaler` and one-hot encoding were fit on the training partition of each fold only. Because aggregation is to unique patients, each patient appears in exactly one test fold, so there is no cross-fold leakage of replicate cores.

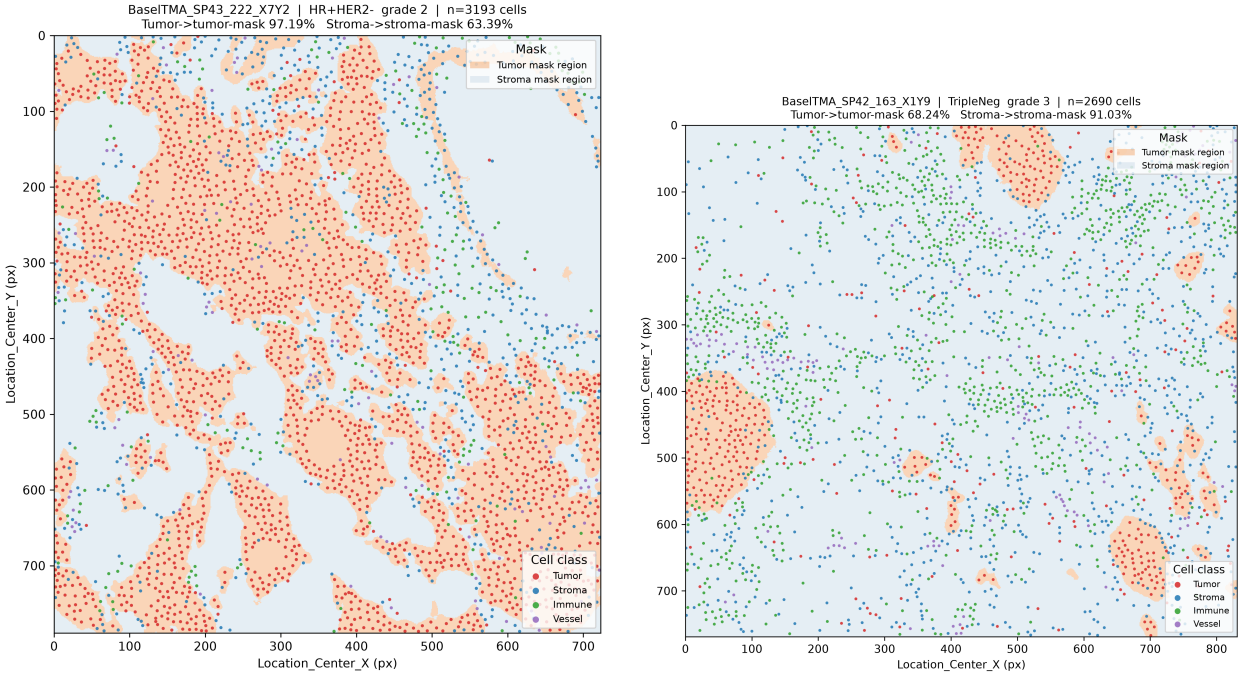
2.6 Performance evaluation and decision framework

Out-of-fold (OOF) risk scores were pooled across all 285 patients, and discrimination was summarised by Harrell's C-index [23]; time-dependent ROC methods provide a complementary view of predictive accuracy [37]. Uncertainty on the pooled C-index and on the incremental C-index between nested models (ΔC-index$_{B-A}$, ΔC-index$_{C-A}$, ΔC-index$_{C-B}$) was quantified by 1000-resample bootstrap 95% CIs. An incremental gain was deemed *stable* only if its bootstrap CI excluded zero. These metrics fed a pre-specified triage: **ADVANCE** the full spatial panel if it stably improved OOF discrimination over composition; **SIMPLIFY** to the best lower-dimensional model if composition—but not the full spatial panel—added stable value; or **STOP** if no block improved on the clinical baseline. The deliverable is this decision for external validation, not a locked clinical predictor, in keeping with prognostic-model reporting guidance [24, 25].

3 Results

3.1 Phenotypes, masks and coordinates are internally consistent

Phenotype coverage was high: 98.69% of located Basel cells carried a broad-class label, with no unmapped classes. Across the seven stratified QC cores, cell-mask concordance was biologically sensible given that the "AllTumorFilled" mask is a coarse *regional* rather than per-cell segmentation: cell-weighted, 80.3% of tumour cells fell inside the tumour region, while 75.5% of stromal, 85.6% of immune and 87.4% of vessel cells fell inside the stroma/non-tumour region. Coordinate integrity was perfect—zero out-of-bounds, negative or duplicate coordinates across all selected cores. Figure 3 shows a representative overlay in which tumour cells clearly aggregate within tumour-mask blobs and immune/stromal cells populate the intervening stroma, confirming that phenotype labels, physical masks and spatial coordinates are mutually registered and safe for downstream spatial statistics.



(a) HR+HER2−, grade 2 (3193 cells) **(b)** TripleNeg, grade 3 (2690 cells)

Figure 3: Quality-control overlays. Single-cell coordinates (Tumour red, Stroma blue, Immune green, Vessel purple) overlaid on the semi-transparent tumour (orange) and stroma (grey-blue) mask regions for two of the seven stratified QC cores. Tumour cells cluster within tumour-mask regions while immune and stromal cells occupy the stroma, confirming registration of phenotypes, masks and coordinates.

3.2 The spatial panel recovers expected tumour-microenvironment biology

The locked panel reproduced well-known TME organisation (Figure 4). Mean neighbourhood-enrichment z-scores showed strong self-aggregation of tumour cells into nests (Tumour–Tumour $z = +28.1$) and clustering of immune and vascular cells (Immune–Immune +13.6; Vessel–Vessel +11.4). Cross-class contacts were predominantly avoided: tumour–immune ($z = -13.6$, reaching -100 in strongly immune-excluded cores) and tumour–stroma ($z = -19.9$) enrichment were markedly negative, quantifying immune exclusion and tumour/stroma compartmentalisation. Edge-corrected cross-type Ripley's L deviations were consistent: immune–stroma $\Delta L(r) > 0$ at all scales (immune cells residing in stroma) whereas tumour–immune deviations were predominantly negative (segregation). Composition fractions varied widely across cores (median tumour and stroma fractions near 0.4, immune near 0.1), confirming that both abundance and arrangement are heterogeneous across the cohort and that the two are captured by distinct feature families.

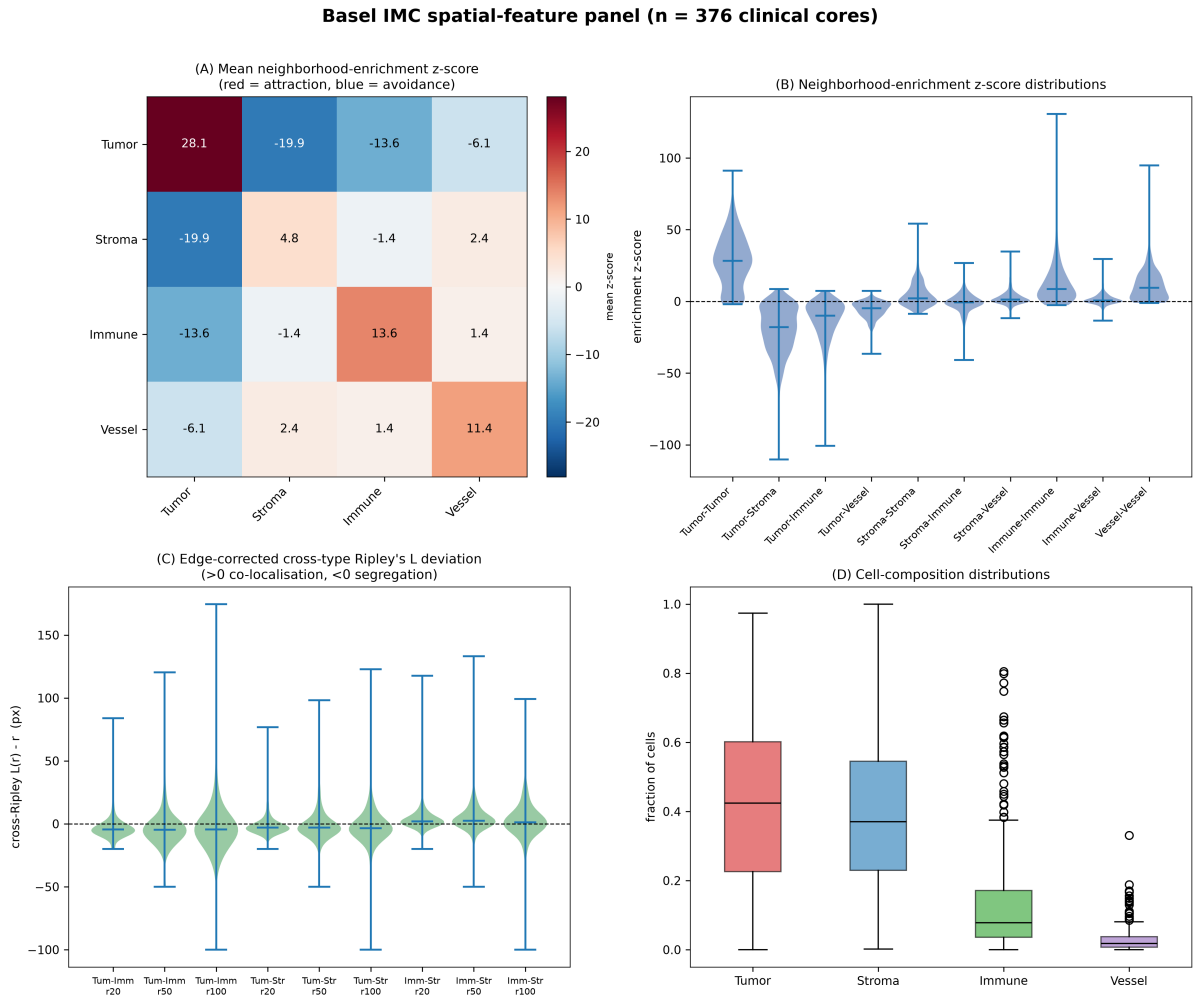


Figure 4: Locked spatial-feature panel (376 cores). (A) Mean neighbourhood-enrichment z-score matrix (red attraction, blue avoidance): strong tumour self-aggregation and tumour–immune/–stroma exclusion. (B) Distributions of the ten enrichment z-scores. (C) Edge-corrected cross-type Ripley's L deviations $\Delta L(r) = L(r) - r$ at $r = 20/50/100$ px. (D) Cell-composition fractions by broad class.

At the patient level (Figure 5), several features tracked with clinical subtype—for example, triple-negative tumours tended toward higher immune fractions—but distributions by survival status (deceased versus censored) overlapped substantially for the individual spatial features, foreshadowing the modelling result that spatial descriptors carry limited *independent* prognostic signal.

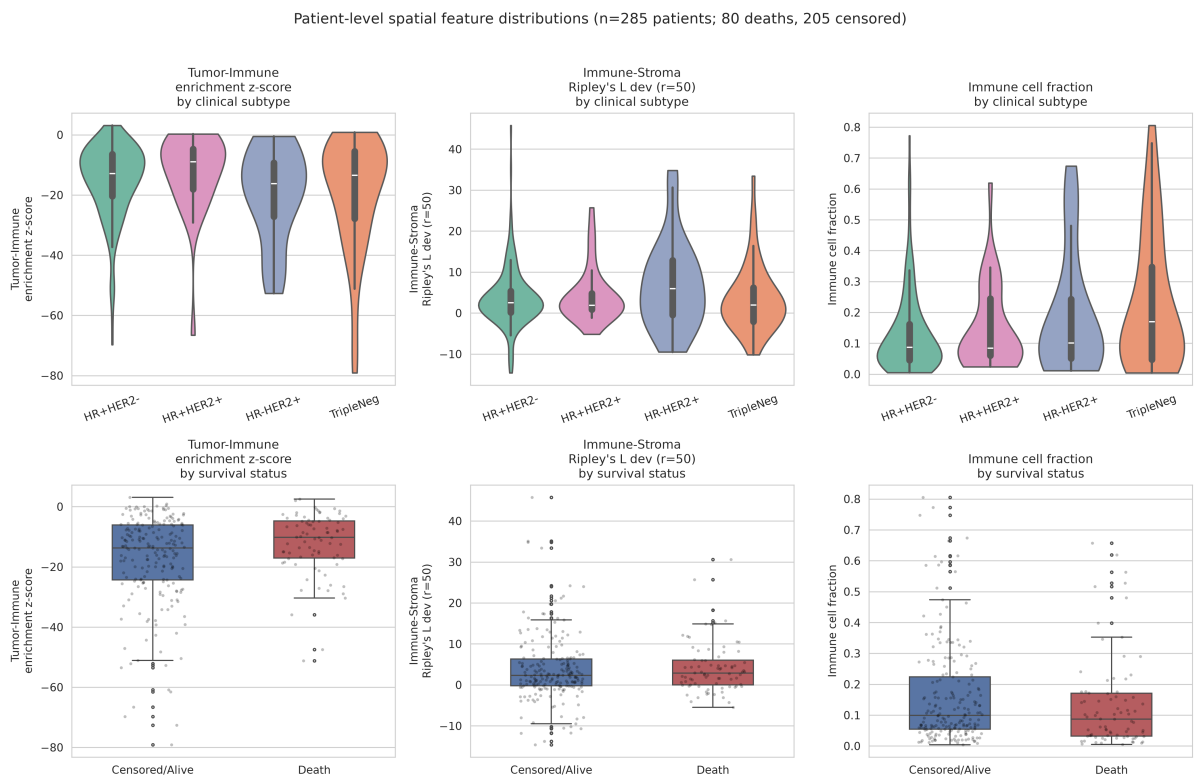


Figure 5: Patient-level feature distributions (285 patients). Selected features by clinical subtype (top) and by survival status (bottom). Subtype structure is visible, but survival-status distributions overlap for individual spatial features, consistent with the limited incremental value observed in modelling.

3.3 Composition adds stable prognostic value; the full spatial panel does not

Under patient-grouped nested cross-validation, out-of-fold discrimination followed a clear, non-monotone pattern (Figure 6A; Table 2). The clinical baseline (Model A) achieved a pooled OOF C-index of **0.725** (95% CI 0.668–0.779). Adding the four composition fractions (Model B) raised discrimination to **0.764** (0.712–0.813), the best of the three models. Adding the 19 spatial features (Model C) *lowered* discrimination to **0.735** (0.672–0.793).

The incremental analysis makes the decision explicit. The composition gain over clinical was stable and statistically supported—$\Delta\text{C-index}_{B-A} = +0.039$ (95% CI $[+0.006, +0.073]$), with a bootstrap CI excluding zero. In contrast, the full spatial panel offered no gain over the clinical baseline ($\Delta\text{C-index}_{C-A} = +0.010$, $[-0.046, +0.060]$) and, relative to composition, *degraded* discrimination ($\Delta\text{C-index}_{C-B} = -0.029$, $[-0.078, +0.011]$). The per-fold pattern was consistent rather than driven by a single split: Model B equalled or exceeded Model C in four of five outer folds, and the inner grid search selected substantially stronger penalisation for Model C (typically retaining 23–26 non-zero coefficients out of 33 with small penalties, but collapsing to as few as 3 when the inner loop favoured heavy ℓ_1 shrinkage), a hallmark of an overparameterised feature block fighting collinearity.

Table 2: Out-of-fold discrimination and incremental C-index under patient-grouped 5×5 nested cross-validation (Basel; 285 patients, 80 events). Bootstrap: 1000 resamples. An incremental gain is *stable* when its 95% CI excludes zero.

Model	Features	OOF C-index	95% CI
A: Clinical	10	0.725	[0.668, 0.779]
B: +Composition	14	**0.764**	**[0.712, 0.813]**
C: +Spatial	33	0.735	[0.672, 0.793]
Incremental C-index (mean [95% CI])			
Δ(B−A) composition gain		+0.039	[+0.006, +0.073] **(excludes 0)**
Δ(C−A) full spatial vs clinical		+0.010	[−0.046, +0.060]
Δ(C−B) marginal spatial gain		−0.029	[−0.078, +0.011]

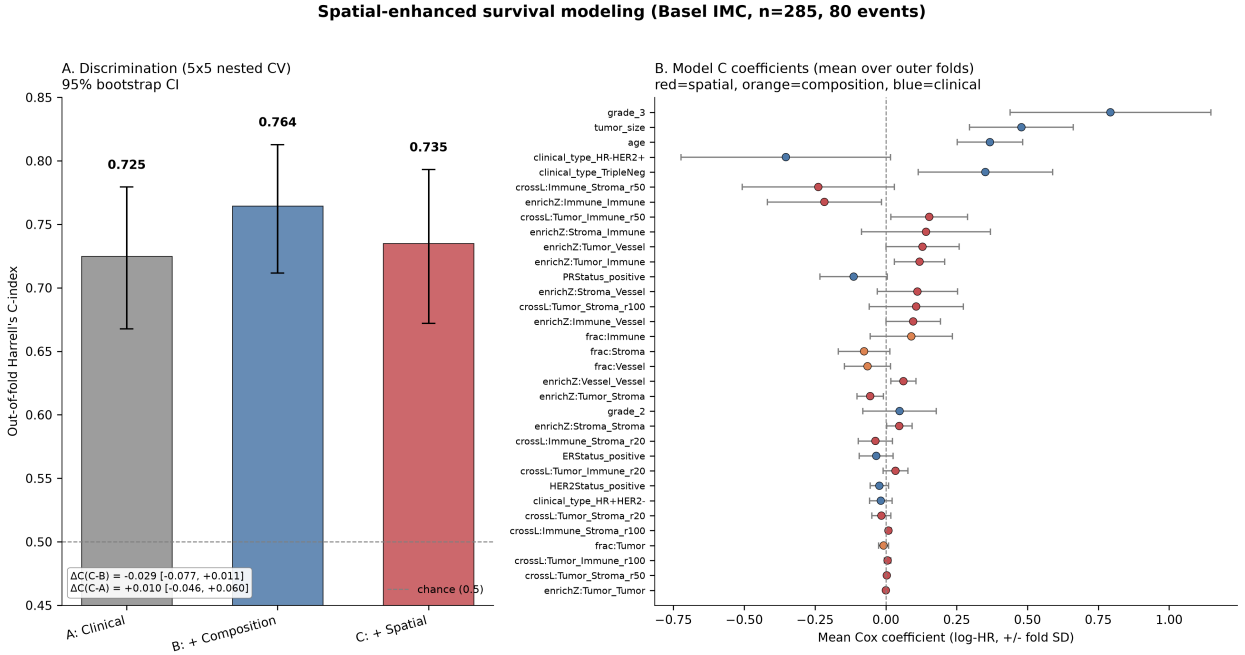


Figure 6: Nested-validation performance and Model C coefficients. (A) Pooled out-of-fold Harrell C-index with 1000-resample bootstrap 95% CIs for the three nested models; Model B (+composition) is highest, and the incremental panels show a stable composition gain (ΔB−A CI excludes 0) but a non-significant, negative marginal spatial effect (ΔC−B). (B) Mean Model C Cox coefficients (log hazard ratio ± fold SD) coloured by block (blue clinical, orange composition, red spatial): grade 3, tumour size and age dominate, while individual spatial coefficients are small and unstable across folds.

3.4 Prognostic weight concentrates in classical pathology variables

Averaged Cox coefficients across the outer folds (Figure 6B; Table 3) explain why. In every model, prognostic weight concentrated in the classical pathology variables: histological grade 3 carried the largest hazard (mean HR ≈ 1.8–2.2), followed by tumour size and age, each selected in all or most folds. In Model B the composition fractions entered with small but consistently signed coefficients (e.g. a protective tendency for higher vessel fraction), and—because Model B is compact—age, tumour size, grade 3 and the triple-negative indicator were selected in *every* fold (selection frequency 1.0), a stability the larger Model C could not match. In Model C the spatial coefficients were

individually small, frequently changed sign across folds, and were selected inconsistently; the largest spatial effects (e.g. immune–stroma ΔL at $r = 50$ px, tumour–immune enrichment) were precisely the collinear, scale-redundant terms most prone to variance inflation. The net effect is that spatial features consumed model capacity without contributing stable, generalisable signal.

Table 3: Representative mean Cox coefficients (log hazard ratio) and selection frequency across the five outer folds. Positive coefficients indicate higher risk. Spatial terms in Model C are small and unstably selected.

Model	Feature	Mean coef.	Sel. freq.
A	grade 3	+0.705	0.8
A	tumour size	+0.359	0.8
A	age	+0.341	0.8
B	tumour size	+0.369	1.0
B	age	+0.304	1.0
B	grade 3	+0.593	1.0
B	clinical type TripleNeg	+0.157	1.0
B	frac Vessel	−0.046	0.4
C	grade 3	+0.793	1.0
C	tumour size	+0.479	1.0
C	age	+0.367	1.0
C	enrichZ Immune–Immune	−0.217	0.8
C	crossL Immune–Stroma r=50	−0.238	0.8
C	enrichZ Tumour–Immune	+0.119	0.8

3.5 Decision: SIMPLIFY

Applying the pre-specified triage, the evidence maps onto a **SIMPLIFY** verdict. Composition adds stable value over the clinical baseline (ΔC-index$_{B-A}$ CI excludes zero), so the analysis does not STOP. But the full spatial panel neither improves on the clinical baseline nor on composition—and its point estimate is lower than composition—so the analysis does not ADVANCE the full panel. The recommended external-validation candidate is therefore **Model B (clinical + composition)**, the best OOF performer, optionally accompanied by a *highly reduced* spatial subset (a single biologically motivated scale such as $r = 50$ px, or a low-dimensional PCA-compressed spatial block) to test whether any compact spatial signal survives out-of-fold validation. The full 33-feature Model C should not advance as-is.

4 Discussion

Using a publicly archived breast-cancer IMC cohort and a deliberately strict evaluation, we asked whether the spatial organisation of tumour, stromal and immune cells adds *stable* prognostic information beyond cell composition and standard pathology. The answer, on this cohort, is a clear and instructive negative. Cell-type composition improved out-of-fold discrimination over the clinical baseline by a margin whose bootstrap CI excluded zero, but a high-dimensional panel of permutation-calibrated neighbourhood enrichment and edge-corrected Ripley's L features did not—and in fact slightly degraded generalisation relative to composition alone. We interpret this not as evidence that tissue architecture is biologically irrelevant, but as evidence that its *stable, incremental, patient-level prognostic value is small* on a cohort of this size, and is easily overwhelmed by the statistical cost of estimating it.

4.1 Why the spatial panel degraded generalisation

Three mutually reinforcing mechanisms explain the result. First, the feature-to-event ratio is unfavourable: Model C estimates 33 coefficients from only 80 events (ratio 0.41), far below the density at which penalised survival models estimate coefficients reliably [20, 21]. Second, the spatial descriptors are strongly collinear—cross-type Ripley's L at $r = 20$, 50 and 100 px are by construction nested cumulative statistics, and neighbourhood enrichment and L deviations measure overlapping notions of co-localisation—so even Elastic-Net shrinkage cannot cleanly allocate weight, producing the sign flips and unstable selection we observed across folds. Third, and most importantly, our evaluation refused to reward this instability: patient-grouped nested cross-validation fits scaling, encoding and hyperparameters strictly within training folds and reports performance only on held-out patients, so any apparent gain from over-flexible spatial features cannot be laundered through the same data used to select them [22, 23]. Much of the enthusiasm for spatial prognostics in the literature rests on in-sample associations, univariate comparisons, or discovery cohorts in which features and thresholds were chosen; our design shows how that enthusiasm can evaporate under out-of-fold scrutiny.

4.2 Reconciliation with prior spatial-prognostics studies

Our finding is compatible with, rather than contradictory to, landmark studies that emphasise spatial architecture [14, 15, 19]. Those studies typically operate at a different granularity—defining discrete, biologically interpretable neighbourhood or "structure" archetypes, often over finer cell-type taxonomies and, in several cases, in specific subtypes such as triple-negative disease—rather than entering dozens of continuous pairwise spatial statistics into a single cohort-wide survival model. It is entirely plausible that a small number of carefully chosen, biologically grounded spatial motifs carry prognostic value that our broad four-class continuous panel dilutes. Composition itself is partly a spatial phenomenon: the immune fraction that proved valuable here is the aggregate of the TIL signal that decades of pathology have shown to be prognostic [6–8]. Our result is thus best read as a caution about *how* spatial information is operationalised and validated, not as a claim that geometry never matters.

4.3 Strengths

The study's principal strength is methodological discipline. Phenotypes, masks and coordinates were validated before any modelling; the spatial panel was locked and documented before performance was inspected; neighbourhood enrichment was calibrated against within-image label permutations so that arrangement is separated from abundance; spatial statistics were edge-corrected following established point-process practice [29–31]; replicate cores were aggregated to unique patients so that no patient straddled the train/test boundary; and discrimination was reported out-of-fold with bootstrap uncertainty on both absolute and incremental performance. Framing the deliverable as an *advance/simplify/stop* decision rather than a putative clinical predictor aligns the analysis with prognostic-model reporting standards [24, 25] and makes the negative result actionable.

4.4 Limitations

Several limitations temper the conclusions. First, the analysis is single-cohort: the Basel TMA is the only arm with overall-survival follow-up, and the Zurich TMA therefore cannot externally validate the models—indeed, providing that external comparison is the entire purpose of the decision we

return. Second, our spatial vocabulary is intentionally coarse (four broad classes, three cross-type pairs, three radii); a finer phenotype taxonomy, discrete neighbourhood archetypes, or graph-based descriptors might capture prognostic geometry that our continuous summaries miss, and testing a compact version of such features is exactly the "simplify" path we recommend. Third, the regional "AllTumorFilled" mask is not a per-cell segmentation, so the QC concordances are expected to be partial; this does not affect the coordinate-based spatial statistics but does limit mask-based inferences. Fourth, categorical covariates were mode-imputed on the full dataset rather than within folds, a minor residual leakage that a fully hardened pipeline would close. Finally, Harrell's C-index measures ranking, not calibration, and can be insensitive to clinically meaningful changes; a fuller evaluation would add time-dependent AUC and calibration assessment [37].

4.5 Implications

For investigators building spatial-prognostic models on cohorts of a few hundred patients and fewer than a hundred events, our results argue for parsimony: establish that composition adds value before layering on high-dimensional geometry; pre-register a locked feature panel; validate at the patient level with nested resampling; and report incremental performance with uncertainty. For the breast-cancer IMC resource specifically, the immediately testable hypotheses are that (i) the clinical+composition model transports to an external cohort, and (ii) a single, biologically motivated spatial scale or a PCA-compressed spatial block—rather than the full panel—adds compact, generalisable signal. These are the experiments a "simplify" decision is meant to license.

5 Conclusion

In a patient-grouped, leakage-controlled analysis of the Basel breast-cancer IMC cohort, cell-type composition added stable prognostic value beyond standard pathology (out-of-fold C-index $0.725 \rightarrow 0.764$; ΔC-index$= +0.039$, 95% CI $[+0.006, +0.073]$), whereas a 33-feature panel of edge-corrected marked spatial statistics and permutation-calibrated neighbourhood enrichment did not, and marginally degraded out-of-fold generalisation ($\Delta\text{C-index}_{C-B} = -0.029$, $[-0.078, +0.011]$). The combination of an unfavourable feature-to-event ratio and strong collinearity among multi-scale spatial descriptors overwhelmed any incremental signal, and the nested-validation design was precisely what made this visible. We therefore return a **SIMPLIFY** decision for external validation: advance the clinical+composition model—and, at most, a highly reduced spatial subset— rather than the full spatial panel. Beyond this specific cohort, the study offers a template for honest spatial-prognostics research, in which locked features, patient-level aggregation, nested resampling and explicit uncertainty convert an appealing hypothesis into a disciplined, reproducible decision.

Data and code availability

The imaging mass cytometry dataset is publicly available on Zenodo (record `4607374`, DOI `10.5281/zenodo.4607374`, Version 2). All derived panels (locked spatial-feature panel, patient-level features, out-of-fold predictions, model coefficients and the decision-synthesis summary) and the analysis scripts (QC validation, spatial feature engineering, patient aggregation, nested Cox modelling and decision synthesis) were generated with a fixed random seed (42) and are organised for reproduction.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. doi: 10.3322/caac.21660.

[2] C. W. Elston and I. O. Ellis. Pathological prognostic factors in breast cancer. I. the value of histological grade in breast cancer: Experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991. doi: 10.1111/j.1365-2559.1991.tb00229.x.

[3] Armando E. Giuliano, James L. Connolly, Stephen B. Edge, Elizabeth A. Mittendorf, Hope S. Rugo, Lawrence J. Solin, Donald L. Weaver, David J. Winchester, and Gabriel N. Hortobagyi. Breast cancer—major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA: A Cancer Journal for Clinicians*, 67(4):290–303, 2017. doi: 10.3322/caac.21393.

[4] Charles M. Perou, Therese Sørlie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Christian A. Rees, Jonathan R. Pollack, Douglas T. Ross, Hilde Johnsen, Lars A. Akslen, Øystein Fluge, Alexander Pergamenschikov, Cheryl Williams, Shirley X. Zhu, Per E. Lønning, Anne-Lise Børresen-Dale, Patrick O. Brown, and David Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000. doi: 10.1038/35021093.

[5] Therese Sørlie, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Thor Thorsen, Hanne Quist, John C. Matese, Patrick O. Brown, David Botstein, Per Eystein Lønning, and Anne-Lise Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19): 10869–10874, 2001. doi: 10.1073/pnas.191240098.

[6] R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, S. Wienert, G. Van den Eynden, F. L. Baehner, F. Penault-Llorca, E. A. Perez, E. A. Thompson, W. F. Symmans, A. L. Richardson, J. Brock, C. Criscitiello, H. Bailey, M. Ignatiadis, G. Floris, J. Sparano, Z. Kos, T. Nielsen, D. L. Rimm, K. H. Allison, J. S. Reis-Filho, S. Loibl, C. Sotiriou, G. Viale, S. Badve, S. Adams, K. Willard-Gallo, and S. Loi. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: Recommendations by an International TILs Working Group 2014. *Annals of Oncology*, 26(2):259–271, 2015. doi: 10.1093/annonc/mdu450.

[7] Carsten Denkert, Gunter von Minckwitz, Silvia Darb-Esfahani, Bianca Lederer, Barbara I. Heppner, Karsten E. Weber, Jan Budczies, Jens Huober, Frederick Klauschen, Jenny Furlanetto, Wolfgang D. Schmitt, Jens-Uwe Blohmer, Thomas Karn, Berit M. Pfitzner, Sherko Kümmel, Knut Engels, Andreas Schneeweiss, Arndt Hartmann, Aurelia Noske, Peter A. Fasching, Christian Jackisch, Marion van Mackelenbergh, Peter Sinn, Christian Schem, Claus Hanusch, Michael Untch, and Sibylle Loibl. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: A pooled analysis of 3771 patients treated with neoadjuvant therapy. *The Lancet Oncology*, 19(1):40–50, 2018. doi: 10.1016/S1470-2045(17)30904-X.

[8] Jérôme Galon, Anne Costes, Fatima Sanchez-Cabo, Amos Kirilovsky, Bernhard Mlecnik, Christine Lagorce-Pagès, Marie Tosolini, Matthieu Camus, Anne Berger, Philippe Wind, Franck Zinzindohoué, Patrick Bruneval, Paul-Henri Cugnenc, Zlatko Trajanoski, Wolf-Herman Fridman,

and Franck Pagès. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795):1960–1964, 2006. doi: 10.1126/science.1129139.

[9] Gabriela Bindea, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, Maximilian Waldner, Anna C. Obenauf, Helen Angell, Tessa Fredriksen, Lucie Lafontaine, Anne Berger, Patrick Bruneval, Wolf Herman Fridman, Christoph Becker, Franck Pagès, Michael R. Speicher, Zlatko Trajanoski, and Jérôme Galon. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, 39(4):782–795, 2013. doi: 10.1016/j.immuni.2013.10.003.

[10] Mikhail Binnewies, Edward W. Roberts, Kelly Kersten, Vincent Chan, Douglas F. Fearon, Miriam Merad, Lisa M. Coussens, Dmitry I. Gabrilovich, Suzanne Ostrand-Rosenberg, Catherine C. Hedrick, Robert H. Vonderheide, Mikael J. Pittet, Rakesh K. Jain, Weiping Zou, T. Kevin Howcroft, Elisa C. Woodhouse, Robert A. Weinberg, and Matthew F. Krummel. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nature Medicine*, 24(5): 541–550, 2018. doi: 10.1038/s41591-018-0014-x.

[11] Charlotte Giesen, Hao A. O. Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J. Schüffler, Daniel Grolimund, Joachim M. Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J. Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11 (4):417–422, 2014. doi: 10.1038/nmeth.2869.

[12] Michael Angelo, Sean C. Bendall, Rachel Finck, Matthew B. Hale, Chuck Hitzman, Alexander D. Borowsky, Richard M. Levenson, John B. Lowe, Scot D. Liu, Shuchun Zhao, Yasodha Natkunam, and Garry P. Nolan. Multiplexed ion beam imaging of human breast tumors. *Nature Medicine*, 20(4):436–442, 2014. doi: 10.1038/nm.3488.

[13] Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*, 174(4):968–981, 2018. doi: 10.1016/j.cell.2018.07.010.

[14] Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, Allison Kurian, David Van Valen, Robert West, Sean C. Bendall, and Michael Angelo. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*, 174(6):1373–1387, 2018. doi: 10.1016/j.cell.2018.08.039.

[15] Christian M. Schürch, Salil S. Bhate, Graham L. Barlow, Darci J. Phillips, Luca Noti, Inti Zlobec, Pauline Chu, Sarah Black, Janos Demeter, David R. McIlwain, Shigemi Kinoshita, Nikolay Samusik, Yury Goltsev, and Garry P. Nolan. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359, 2020. doi: 10.1016/j.cell.2020.07.005.

[16] Takahiro Tsujikawa, Sushil Kumar, Rohan N. Borkar, Vahid Azimi, Guillaume Thibault, Young Hwan Chang, Ariel Balter, Rie Kawashima, Gina Choe, David Sauer, Edward El Rassi, Daniel R. Clayburgh, Molly F. Kulesz-Martin, Eric R. Lutz, Lei Zheng, Elizabeth M. Jaffee, Patrick Leyshock, Adam A. Margolin, Motomi Mori, Joe W. Gray, Paul W. Flint, and Lisa M. Coussens. Quantitative multiplex immunohistochemistry reveals myeloid-inflamed tumor-immune complexity associated with poor prognosis. *Cell Reports*, 19(1):203–217, 2017. doi: 10.1016/j.celrep.2017.03.037.

[17] Hartland W. Jackson, Jana R. Fischer, Vito R. T. Zanotelli, H. Raza Ali, Robert Mechera, Savas D. Soysal, Holger Moch, Simone Muenst, Zsuzsanna Varga, Walter P. Weber, and Bernd Bodenmiller. The single-cell pathology landscape of breast cancer. *Nature*, 578(7796):615–620, 2020. doi: 10.1038/s41586-019-1876-x.

[18] H. Raza Ali, Hartland W. Jackson, Vito R. T. Zanotelli, Esther Danenberg, Jana R. Fischer, Helen Bardwell, Elena Provenzano, Oscar M. Rueda, Suet-Feung Chin, Samuel Aparicio, Carlos Caldas, and Bernd Bodenmiller. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nature Cancer*, 1(2):163–175, 2020. doi: 10.1038/s43018-020-0026-6.

[19] Esther Danenberg, Helen Bardwell, Vito R. T. Zanotelli, Elena Provenzano, Suet-Feung Chin, Oscar M. Rueda, Andrew Green, Emad Rakha, Samuel Aparicio, Ian O. Ellis, Bernd Bodenmiller, Carlos Caldas, and H. Raza Ali. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature Genetics*, 54(5):660–669, 2022. doi: 10.1038/s41588-022-01041-y.

[20] Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, 1996. doi: 10.1016/S0895-4356(96)00236-3.

[21] Richard D. Riley, Kym I. E. Snell, Joie Ensor, Danielle L. Burke, Frank E. Harrell, Karel G. M. Moons, and Gary S. Collins. Minimum sample size for developing a multivariable prediction model: Part II—binary and time-to-event outcomes. *Statistics in Medicine*, 38(7):1276–1296, 2019. doi: 10.1002/sim.7992.

[22] Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010. URL `https://www.jmlr.org/papers/v11/cawley10a.html`.

[23] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[24] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G. M. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*, 350:g7594, 2015. doi: 10.1136/bmj.g7594.

[25] Gary S. Collins, Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B. Reitsma, Maarten van Smeden, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

[26] Noah F. Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J. McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, Cole Pavelchek, Sunny Cui, Isabella Camplisson, Omer Bar-Tal, Jaiveer Singh, Mara Fong, Gautam Chaudhry, Zion Abraham, Jackson Moseley, Shiri Warshawsky, Erin Soon, Shirley Greenbaum, Tyler Risom, Travis Hollmann, Sean C. Bendall, Leeat Keren, William Graf,

Michael Angelo, and David Van Valen. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology*, 40(4):555–565, 2022. doi: 10.1038/s41587-021-01094-0.

[27] Denis Schapiro, Hartland W. Jackson, Swetha Raghuraman, Jana R. Fischer, Vito R. T. 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. doi: 10.1038/nmeth.4391.

[28] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. Squidpy: A scalable framework for spatial omics analysis. *Nature Methods*, 19(2):171–178, 2022. doi: 10.1038/s41592-021-01358-2.

[29] B. D. Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–212, 1977. doi: 10.1111/j.2517-6161.1977.tb01615.x.

[30] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R.* Chapman and Hall/CRC, Boca Raton, FL, 2015. doi: 10.1201/b19708.

[31] Peter Haase. Spatial pattern analysis in ecology based on Ripley's K-function: Introduction and methods of edge correction. *Journal of Vegetation Science*, 6(4):575–582, 1995. doi: 10.2307/3236356.

[32] Adrian Baddeley and Rolf Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005. doi: 10.18637/jss.v012.i06.

[33] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x.

[34] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

[35] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39 (5):1–13, 2011. doi: 10.18637/jss.v039.i05.

[36] Cameron Davidson-Pilon. lifelines: Survival analysis in Python. *Journal of Open Source Software*, 4(40):1317, 2019. doi: 10.21105/joss.01317.

[37] Patrick J. Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005. doi: 10.1111/j.0006-341X.2005.00318.x.