# Supplementary Material

Joel Eliason, Michele Peruzzi, Arvind Rao

# 1 Background: Multitype Gibbs Point Process Models and Relationship to SHADE

## 1.1 Multitype Gibbs Point Process Models

Multitype Gibbs point process (MGPP) models provide a general framework for modeling spatial patterns where interactions between points influence their configuration (Baddeley et al., 2015; **?**). Unlike inhomogeneous Poisson processes, Gibbs processes explicitly model dependencies between points via pairwise interaction potentials.

### 1.1.1 Mathematical Formulation

For a spatial point pattern $\mathbf{x} = \{x_1, \ldots, x_n\}$ with type marks $m_i \in \{1, \ldots, K\}$, a multitype Gibbs point process is characterized by its conditional intensity function:

$$\lambda_k(v \mid \mathbf{x}) = \beta_k \exp\left\{ \sum_{x_i \in \mathbf{x}} \log \gamma_{m_i,k}(\|v - x_i\|) \right\}, \tag{1}$$

where $\beta_k > 0$ is the baseline intensity for type $k$ and $\gamma_{j,k}(r)$ is the pairwise interaction function between types $j$ and $k$ at distance $r$. Values $\gamma_{j,k}(r) > 1$ indicate attraction, $\gamma_{j,k}(r) < 1$ indicate repulsion, and $\gamma_{j,k}(r) = 1$ indicate independence.

### 1.1.2 Key Limitation: Symmetry Assumption

A fundamental constraint of standard Gibbs point processes is the symmetry condition:

$$\gamma_{j,k}(r) = \gamma_{k,j}(r) \quad \text{for all types } j, k \text{ and distances } r. \tag{2}$$

This requirement ensures the joint density defines a valid probability measure.

**Biological implication**: MGPPs *cannot* distinguish directional spatial associations. If tumor cells cluster around blood vessels, but vessels are not preferentially located near tumors, a symmetric MGPP estimates only the average of these distinct directional effects—a severe limitation for studying directional biological processes.

In practice, MGPPs also typically assume parametric interaction functions (e.g., Strauss process with $\gamma_{j,k}(r) = \gamma_{j,k}\mathbb{I}(r \leq R)$) with fixed interaction radii. Inference for Gibbs models is challenging due to intractable normalizing constants, though the Berman-Turner logistic regression approximation (**?**) provides a computationally efficient alternative to MCMC over point configurations.

### 1.1.3 Hierarchical Gibbs Models: Asymmetry via Type Ordering

One approach to modeling asymmetric spatial interactions is to impose an ordering on cell types (**?**). In a hierarchical Gibbs model, types are indexed $1, 2, \ldots, K$ with the convention that type $k$ may depend on types $1, \ldots, k-1$ but not on types $k+1, \ldots, K$. The conditional intensity for type $k$ becomes:

$$\lambda_k(v \mid \mathbf{x}_{<k}) = \beta_k \exp\left\{\sum_{j<k} \sum_{x_i \in \mathbf{x}_j} \log \gamma_{j,k}(\|v - x_i\|)\right\}, \tag{3}$$

where $\mathbf{x}_{<k}$ denotes the configuration of all types preceding $k$ in the ordering. This permits directional interactions: type $j$ can influence type $k$ (for $j < k$) without requiring symmetric reciprocal effects. For example, if vasculature is ordered before immune cells, the model can capture how vessel locations structure immune infiltration without requiring vessels to preferentially locate near immune cells.

Hierarchical Gibbs models in the literature typically employ parametric interaction functions with fixed radii and have focused on single-image estimation rather than multilevel data structures.

## 1.2 How SHADE Extends the MGPP Framework

SHADE builds on the conceptual foundation of MGPPs but introduces three key innovations:

### 1.2.1 Asymmetric (Directional) Interactions

Unlike MGPPs, which require $\gamma_{j,k}(r) = \gamma_{k,j}(r)$, SHADE explicitly models asymmetric spatial associations via directional spatial interaction curves:

$$\text{SIC}_{A_k \to B}(s) = \sum_{p=1}^{P} \delta_{A_k}^{(p)} \phi_p(s), \tag{4}$$

where $\text{SIC}_{A_k \to B}(s)$ quantifies how type-$A_k$ source cells at distance $s$ affect the log-intensity of type-$B$

target cells. Critically, $\text{SIC}_{A_k \to B}(s)$ and $\text{SIC}_{B \to A_k}(s)$ are estimated independently, capturing directional processes like immune recruitment by tumors or structural constraints imposed by vasculature.

#### 1.2.2 Flexible Basis Function Expansions

Rather than parametric forms with fixed interaction radii, SHADE uses smooth radial basis functions $\phi_p(s)$ (e.g., Gaussian or B-splines) with data-driven coefficients $\delta_{A_k}^{(p)}$. This yields smooth, adaptable interaction curves avoiding restrictive parametric assumptions while hierarchical priors prevent overfitting.

#### 1.2.3 Multilevel Bayesian Hierarchical Structure

Standard MGPPs and hierarchical Gibbs models focus on single-image estimation. SHADE instead models spatial interactions across nested levels of biological organization:

$$\delta_{A_k}^{(m,p)} \sim \mathcal{N}(\gamma_{A_k}^{(n(m),p)}, \sigma^2_{\text{image},p}) \qquad \text{(image-level)} \tag{5}$$

$$\gamma_{A_k}^{(n,p)} \sim \mathcal{N}(\psi_{A_k}^{(g(n),p)}, \sigma^2_{\text{patient},p}) \qquad \text{(patient-level)} \tag{6}$$

$$\psi_{A_k}^{(g,p)} \sim \mathcal{N}(\mu_p, \sigma^2_{\text{cohort},p}) \qquad \text{(cohort-level)} \tag{7}$$

This hierarchical Bayesian framework, implemented via the Berman-Turner logistic approximation and Hamiltonian Monte Carlo, enables:

- **Borrowing strength** across images within patients and patients within cohorts
- **Full posterior inference** on derived quantities (SICs) at any hierarchical level, with simultaneous credible bands
- **Cohort-level comparisons** of spatial organization patterns between biological groups
- **Quantification of heterogeneity** via variance parameters at each level

This enables interpretable, uncertainty-aware inference on biologically meaningful spatial interaction patterns across multiple scales.

### 1.3 Summary of Key Differences

Table 1 summarizes the key methodological differences between standard MGPP models, hierarchical Gibbs models, and the SHADE framework.

Table 1: Comparison of spatial point process modeling approaches

| Feature | Symmetric MGPP | Hierarchical Gibbs | SHADE |
|---|---|---|---|
| **Interaction symmetry** | Symmetric: $\gamma_{j,k}(r) = \gamma_{k,j}(r)$ | Asymmetric via type ordering | Asymmetric: independent $\text{SIC}_{A\to B}(s)$ |
| **Interaction function** | Parametric with fixed radii | Parametric with fixed radii | Flexible basis expansions |
| **Data structure** | Single image | Single image | Multilevel: image $\subset$ patient $\subset$ cohort |
| **Posterior inference** | Not typically Bayesian | Not typically Bayesian | Full posterior with credible bands, heterogeneity quantification |

Together, these innovations enable SHADE to overcome fundamental limitations of existing point process methods for analyzing spatial organization in multiplexed tissue imaging data, where directional biological processes, smooth distance-dependent interactions, and hierarchical uncertainty quantification are essential.

# 2 Interpreting Example Spatial Interaction Curves

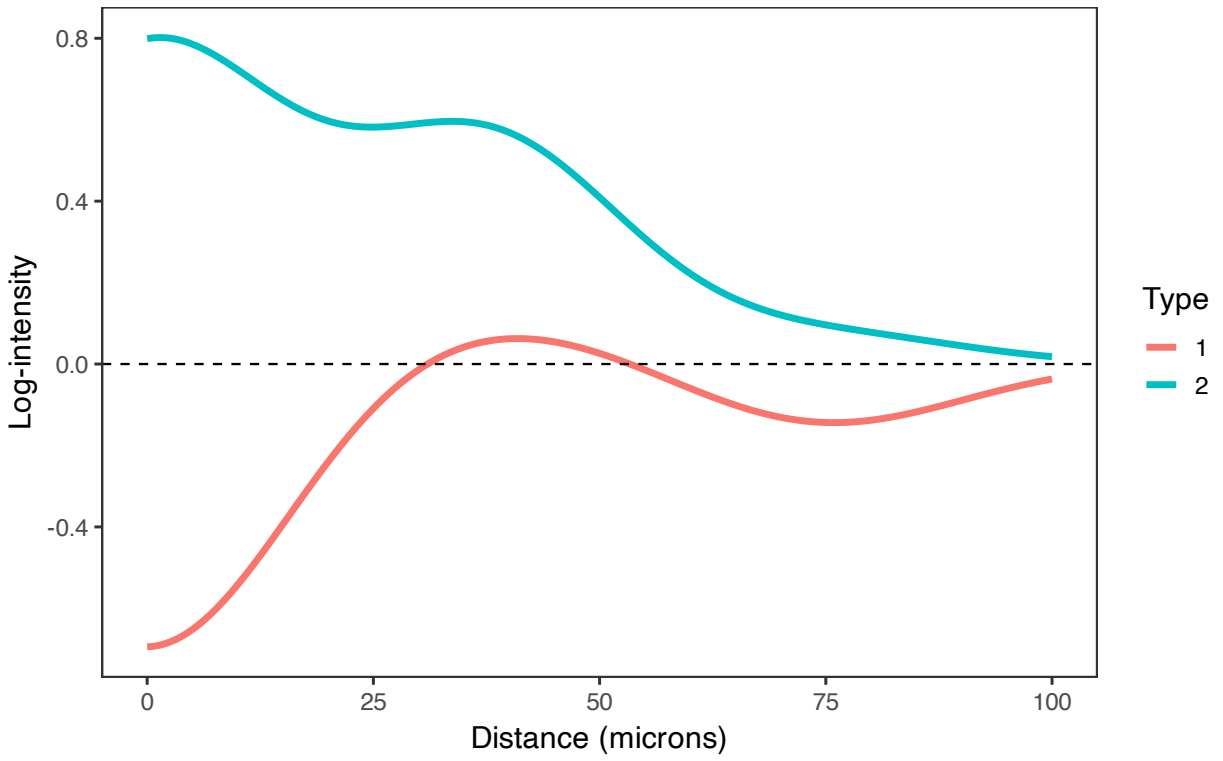
Figure S1 shows two example SICs representative of those estimated from real data. Type 2 cells (cyan) exhibit a strong positive association at short distances, consistent with clustering behavior observed in immune infiltration around tumor cells. In contrast, Type 1 cells (red) show negative association at close range, suggestive of exclusion zones, as might be induced by physical barriers or competitive interactions in the tissue microenvironment.

Figure S2 demonstrates how these spatial interactions manifest in tissue space. Panels 1 and 2 show the spatial predictors $\boldsymbol{q}_{A_1}(v)$ and $\boldsymbol{q}_{A_2}(v)$ generated by the individual SICs from Figure S1—each reflecting the localized contribution to the log-intensity of type $B$ due to each source. Each panel is analogous to a summand from the sum on the right hand side of (2). Panel 3 shows the additive combination of these effects, illustrating how multiple source cell types jointly shape a heterogeneous spatial intensity landscape. The target cells are observed to cluster in high-intensity regions and avoid areas of low expected density.

# 3 Extended Methods and Results for Simulation Studies

## 3.1 Simultaneous Credible Bands for Spatial Interaction Curves

To quantify uncertainty in estimated spatial interaction curves (SICs), we constructed *simultaneous 95% credible bands* using posterior samples from our Bayesian model. While pointwise credible intervals can quantify uncertainty at individual distances, they do not account for multiple comparisons across the distance




Supplementary Figure S1. Estimated SICs for two source cell types. Type 2 shows a strong positive association with the target population at short distances, while Type 1 exhibits negative association at short range. These contrasting patterns highlight the flexibility of the SIC framework for capturing biologically meaningful spatial structure.

domain and tend to overstate confidence when evaluating the entire curve. In contrast, simultaneous credible bands provide a global probabilistic guarantee that the entire SIC lies within the band over its domain with high posterior probability.

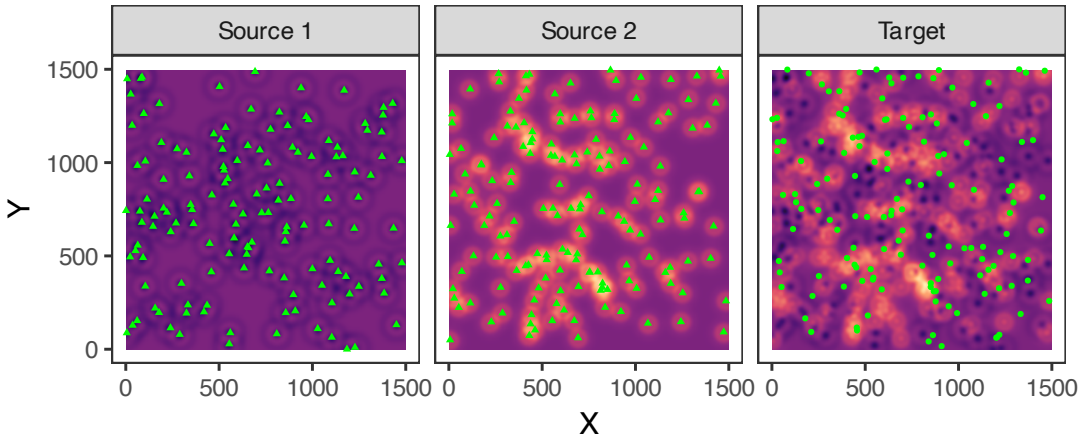
Let $\widehat{\mathrm{SIC}}_{A_k \to B}(s)$ denote the posterior mean and $\hat{\sigma}_{A_k \to B}(s)$ the posterior standard deviation of the SIC at distance $s$, evaluated over a fixed grid of distances $s \in [r_{\min}, r_{\max}]$. For each posterior draw $j$, we computed the standardized residual

$$Z^{(j)}(s) = \frac{\mathrm{SIC}^{(j)}_{A_k \to B}(s) - \widehat{\mathrm{SIC}}_{A_k \to B}(s)}{\hat{\sigma}_{A_k \to B}(s)},$$

and then calculated the *maximum absolute standardized deviation* across the distance domain:

$$T^{(j)} = \sup_{s \in [r_{\min}, r_{\max}]} \left| Z^{(j)}(s) \right|.$$

This statistic captures the largest fluctuation, in units of local posterior uncertainty, that the sampled SIC exhibits relative to the posterior mean. By repeating this computation over posterior draws, we obtain the empirical distribution of the supremum deviation under the posterior. The critical value $c$ for the




Supplementary Figure S2. Spatial interaction fields implied by SICs from two source cell types (Panels 1 and 2) and their additive combination (Panel 3). The combined field governs the spatial distribution of target cells, which are more likely to occur in regions of elevated log-intensity.

simultaneous band is then defined as the $(1-\alpha)$-quantile of the empirical distribution of $T^{(j)}$:

$$c = \text{quantile}_{1-\alpha}\left(\{T^{(1)}, \ldots, T^{(J)}\}\right).$$

The resulting *simultaneous credible band* is:

$$\text{SIC}_{A_k \to B}(s) \in \left[\widehat{\text{SIC}}_{A_k \to B}(s) \pm c \cdot \hat{\sigma}_{A_k \to B}(s)\right], \quad \text{for all } s \in [r_{\min}, r_{\max}].$$

This construction ensures that with posterior probability $1-\alpha$, the entire SIC remains within the band across the full distance domain. It provides a conservative yet interpretable summary of global uncertainty, correcting for the multiple comparisons that arise when interpreting the SIC across many distances.

This approach is conceptually analogous to Scheffé's method for simultaneous inference in classical linear models, but adapted to functional estimation and implemented through posterior simulation. Simultaneous credible bands are particularly useful for assessing whether a spatial interaction is consistently positive, negative, or null over specific distance ranges, and for identifying features such as consistent attraction or repulsion that are unlikely to be due to noise. They also serve as a diagnostic for the informativeness of the data: wider bands reflect higher posterior uncertainty, particularly in regions where data are sparse or spatial patterns are weak.

All figures in the main manuscript showing SICs (Figures 2, 3, **??**, and 9) display simultaneous 95% credible bands computed via this method.

## 3.2 Screening Strategies for Multiple Cell Type Pairs

In exploratory spatial analyses involving many cell type pairs, we often wish to prioritize pairs for further investigation based on the estimated SICs and their uncertainty. While there are many ways to summarize spatial associations for this purpose, we propose three complementary summary measures based on simultaneous credible bands that address different biological questions:

### 3.2.1 Peak Location and Magnitude

To identify where the strongest interaction occurs, we locate the distance at which the SIC reaches its maximum absolute value, along with the magnitude at that location. Formally, for a given source–target pair $(A_k \to B)$, define:

$$s^* = \underset{s}{\operatorname{argmax}} |\widehat{\mathrm{SIC}}_{A_k \to B}(s)|, \quad M = |\widehat{\mathrm{SIC}}_{A_k \to B}(s^*)|,$$

and assess statistical significance by checking whether the simultaneous credible band excludes zero at $s^*$. This identifies the distance of maximal association and its strength, and is particularly useful for identifying pairs with strong localized effects—for example, immune cells that show pronounced clustering around tumor cells at a specific distance range.

### 3.2.2 Persistence Over Biologically Relevant Distance Ranges

To assess whether an interaction is consistently positive or negative over a pre-specified biologically meaningful distance range $I = [s_a, s_b]$, we compute the posterior probability that the SIC maintains a consistent sign throughout the interval. For each posterior draw $j$, we evaluate the SIC at all distances within $I$ and record whether the entire curve is positive (or negative):

$$\Pi_I^{(+)} = \frac{1}{J} \sum_{j=1}^{J} \mathbb{I}\left\{ \min_{s \in I} \mathrm{SIC}^{(j)}_{A_k \to B}(s) > 0 \right\}, \quad \Pi_I^{(-)} = \frac{1}{J} \sum_{j=1}^{J} \mathbb{I}\left\{ \max_{s \in I} \mathrm{SIC}^{(j)}_{A_k \to B}(s) < 0 \right\}.$$

We then define $\Pi_I = \max(\Pi_I^{(+)}, \Pi_I^{(-)})$ as the persistence score. Values $\Pi_I \geq 0.95$ indicate strong evidence for persistent directional association throughout the interval—that is, robust associations that are unlikely to be artifacts of noise or multiple testing. For instance, in tumor immunology, one might define $I = [10, 50]$ $\mu$m to capture local microenvironmental interactions.

#### 3.2.3 Overall Strength

To quantify the cumulative magnitude of a spatial interaction across all distances where it is statistically significant, we integrate the absolute effect size over significant regions:

$$\text{Strength}(A_k \to B) = \int_{I_{\text{sig}}} |\widehat{\text{SIC}}_{A_k \to B}(s)|\, ds,$$

where $I_{\text{sig}}$ is the union of all intervals where the simultaneous credible band excludes zero. This captures cumulative association strength and provides a single summary of the overall importance of a spatial interaction. Cell type pairs with high strength scores exhibit either strong localized effects or more moderate effects that persist across many distances.

#### 3.2.4 Usage in Practice

These three measures can be computed across all source–target pairs and visualized as heatmaps to facilitate systematic comparison. Peak location and magnitude would require either two separate heatmaps (one for $s^*$, one for $M$) or a combined visualization (e.g., color-coding magnitude with symbol size indicating distance). The persistence and overall strength measures each yield a single scalar value per pair and map directly to heatmap intensities. These visualizations allow pairs to be ranked and prioritized for detailed examination.

The choice among these measures depends on the biological question: peak magnitude for identifying strongest interactions, persistence for testing specific distance-based hypotheses, or overall strength for comparing cumulative effects. They are complementary and can be used together to provide multiple perspectives on spatial organization patterns.

### 3.3 Extended Methods for Hyperparameter Studies and Study of Explicit Modeling of Multilevel Structure

#### 3.3.1 Generating Target Points

Target cell locations are drawn from an inhomogeneous Poisson process with intensity:

$$\lambda(v) = \exp\left(\sum_{k=1}^{K}\sum_{p=1}^{P} \boldsymbol{q}_{A_k}^{\top}(v)\boldsymbol{\delta}_{A_k}^{(m)} + \beta_0^{(m)}\right), \tag{8}$$

where $\beta_0^{(m)}$ is a normalization offset to ensure that the expected number of target points matches $N_{\text{points}}$ in image $m$. It is computed as:

$$\beta_0^{(m)} = \log\left(\frac{N_{\text{points}}}{\int_W \exp\left(\sum_{k=1}^{K}\sum_{p=1}^{P} \boldsymbol{q}_{A_k}^{\top}(v)\boldsymbol{\delta}_{A_k}^{(m)}\right) dv}\right). \tag{9}$$

#### 3.3.2 Default Parameter Settings

Unless otherwise specified, the following defaults were used in simulation experiments:

- **Spatial domain:** $S = 1500$.
- **Number of source cell types:** $K = 2$.
- **Basis functions:** $P = 3$ radial basis functions with bandwidth 15 and support up to 75 microns.
- **Points per cell type per image:** $N_{\text{points}} = 150$.
- **Variance parameters:**
  - $\sigma_{\text{cohort},p} = 0.5$
  - $\sigma_{\text{patient},p} = 0.1$
  - $\sigma_{\text{image},p} = 0.1$
- **Number of patients and images:** variable across experiments.

### 3.4 Extended Methods for Comparison of Spatial Pattern Detection Accuracy Across Methods and Conditions

#### 3.4.1 Hierarchical Data Generation

We simulate spatial interaction coefficients $\delta^{(m,p)}$ at three hierarchical levels (for image $m$ and basis function $p \in \{1, 2, 3\}$):

$$\psi^{(g,p)} = \begin{cases} -1.5,\ -1.0,\ -0.5 & \text{if group } g = \text{Responder} \\ \ \ 1.5,\ \ \ 1.0,\ \ \ 0.5 & \text{if group } g = \text{Non-responder} \end{cases}$$

$$\gamma^{(n,p)} \sim \mathcal{N}(\psi^{(g(n),p)},\ \sigma^2_{\text{patient}}), \quad \sigma_{\text{patient}} = 0.1$$

$$\delta^{(m,p)} \sim \mathcal{N}(\gamma^{(n(m),p)},\ \sigma^2_{\text{image}}), \quad \sigma_{\text{image}} = 0.1$$

Here, $g(n)$ denotes the group of patient $n$, and $n(m)$ denotes the patient corresponding to image $m$.

### 3.4.2 Spatial Pattern Generation

Cell patterns were generated using spatially varying intensity functions within $1500 \times 1500$ $\mu$m$^2$ observation windows. T cells and B cells were distributed as independent Poisson processes with densities $\lambda_T$ and $\lambda_B$. Tumor cell locations were generated using a spatially varying intensity surface based on T cell locations:

$$\lambda_{\text{tumor}}(s; n) = \exp\left(\beta_0 + \sum_p \delta^{(n,p)} \phi_p(s)\right)$$

where $\phi_p(s)$ represents radial basis functions with $\sigma = 15$ $\mu$m and centered at $\mu = (0, 40, 80)\mu m$. The baseline intensity $\beta_0$ was calibrated to achieve target tumor cell densities, as in (9).

### 3.4.3 Experimental Design

We employed a factorial design varying the following:

- **Cell Density:** High (150 cells) vs. Low (15 cells) for both T cells and tumor cells.
- **Images per Patient:** 1, 2, or 3 tissue sections.
- **Replication:** 30 independent simulations per condition.

Each simulation included 20 patients per group (40 total), generating datasets with realistic sample sizes for multiplexed imaging studies.

### 3.4.4 Evaluation Metrics

**SHADE Performance:** We assessed SHADE's ability to recover true spatial interaction curves by constructing pointwise 95% credible bands around posterior estimates. Success was defined as correctly identifying the sign of spatial associations (negative for responders, positive for non-responders) at target distances (20, 40, 60 $\mu$m) using pointwise credible intervals.

**G-cross Comparison:** We implemented envelope tests using 39 simulations of point patterns generated according to complete spatial randomness (CSR) to construct 95% confidence envelopes around the null $G(r)$ function between T cells and tumor cells. Detection success required deviations outside of these null envelopes, with clustering detection for non-responders and repulsion detection for responders.

**Statistical Analysis:** Power was calculated as the average over all images of the proportion of correctly classified interactions at target distances. We compared both methods' sensitivity to varying cell densities and sample sizes. Additionally, we evaluated coverage (the proportion of true SIC values falling within the 95% simultaneous credible bands) and type I error rates (the proportion of null interactions incorrectly identified as significant) for all methods.

### 3.5 Hyperparameter Study 1 - The effect of cell count and dummy point ratio on inference quality

We investigated how inference quality is affected by the number of observed cells per cell type ($N_t$) and the ratio of dummy points to real points ($R_d$) used for quadrature. Simulations were run for all combinations of $N_t \in \{20, 80, 150, 300\}$ and $R_d \in \{0.5, 1, 2, 5, 10\}$, with five replicates per setting. Each dataset contained three cell types, with the third serving as the target for spatial interaction estimation. The number of patient groups was fixed at 1, with 40 patients per group and two images per patient. All other simulation parameters were set to defaults.

Average root mean squared error (RMSE) was computed for spatial interaction coefficients at the image ($\delta^{(m,p)}_{t_1 \to t_2}$), patient ($\gamma^{(n,p)}_{t_1 \to t_2}$), and cohort ($\psi^{(g,p)}_{t_1 \to t_2}$) levels. As shown in Figure S3, RMSE for image-level coefficients decreased consistently with increasing $R_d$, suggesting that finer quadrature grids improve inference at the image level. This trend held across all values of $N_t$.

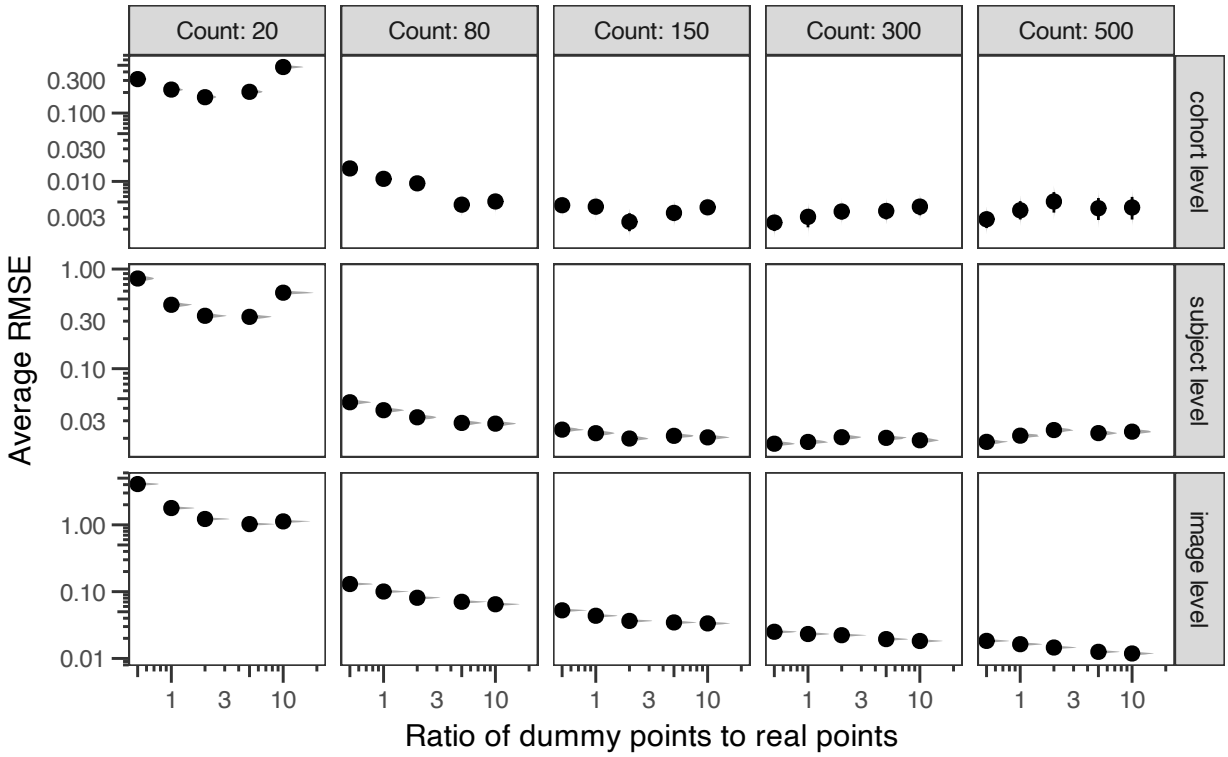
In contrast, inference quality for patient- and cohort-level coefficients slightly worsened when $R_d$ exceeded 2–5, particularly at very low or very high values of $N_t$. This suggests diminishing returns—and potential instability—for higher-level parameter estimation when dummy point counts are excessively large. The magnitude of this effect was modest but visible (note the log scale on the RMSE axis).

To further investigate these trends, we analyzed RMSE stratified by spatial scale (short, medium, long) at each hierarchical level (Figure S4). Improvements in image-level inference with increasing $R_d$ were most pronounced at longer interaction distances. For higher-level parameters, performance remained relatively stable across spatial scales but showed a slight increase in RMSE at short ranges when dummy point ratios were high.

A representative cohort-level SIC estimate is shown in Figure S5, illustrating how estimation accuracy varies with $N_t$ and $R_d$. Inference was notably poorer at low cell counts, with increased bias and wider credible intervals. Optimal performance—reflected in reduced bias and uncertainty—was observed at moderate $N_t$ (80–300) and a wide range of dummy point ratios, consistent with the RMSE patterns observed in Figure S3.

### 3.6 Hyperparameter Study 2 - The effect of number of patients and images per patient on inference quality

Next, we examined the sensitivity of inference quality to the total number of patients as well as the number of images per patient. Here, we set the number of points per type to 150 and the ratio of dummy points to actual points to 2, while keeping the rest of the default parameters the same, as outlined in the Supplement. We then simulated 15 realizations each of every combination of the number of images per patient $N_{\text{images per patient}}$

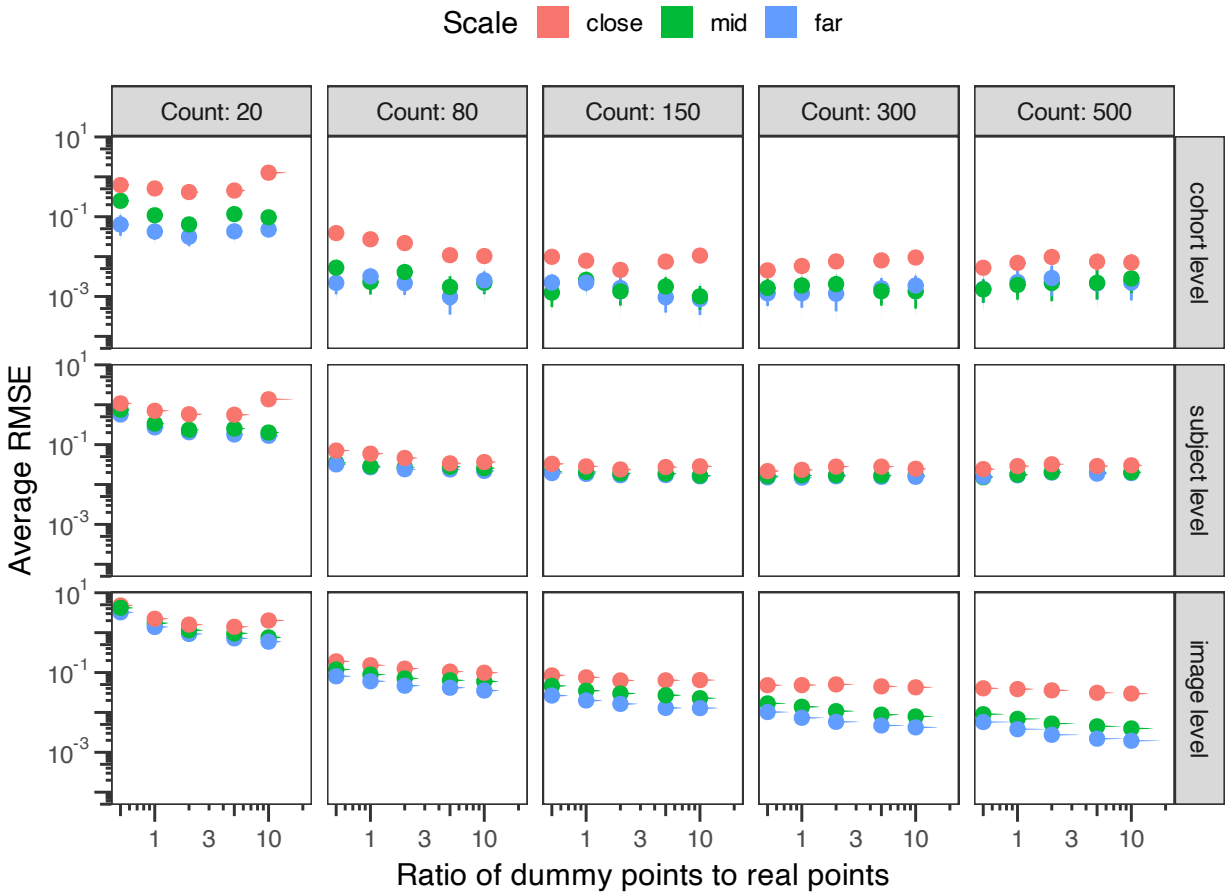



Supplementary Figure S3. Average RMSE for spatial interaction coefficients across hierarchical levels: image-level ($\delta^{(m,p)}_{t_1 \to t_2}$), patient-level ($\gamma^{(n,p)}_{t_1 \to t_2}$), and cohort-level ($\psi^{(g,p)}_{t_1 \to t_2}$), under different numbers of cells per type and dummy-to-real point ratios.

and the total number of patients $N_{\text{patients}}$, where $N_{\text{images per patient}} \in \{1, 2, 4\}$ and $N_{\text{patients}} \in \{10, 20, 40\}$, according to our simulation procedure described in the Supplement.

We found some interesting trends in the average RMSE as the number of patients per patient group increased. For cohorts that only had one image per patient, average RMSE decreased as the number of patients increased (Figure S6). However, for patients that had 2 images, the average RMSE was highest for the highest number of patients, which we found to be a counterintuitive finding. Furthermore, for patients with 4 images, having a cohort with 40 patients was associated with the lowest RMSE, as expected, though the second-lowest was, unexpectedly, having a cohort with 10 patients, rather than 20.

For $\delta^{(m,p)}_{t_1 \to t_2}$ coefficients, longer-range coefficients seemed to have the lowest average RMSE (Figure S7), while close-range coefficients had the worst quality. RMSE, however, was very close to constant across the number of patients, though it did decrease slightly as the number of images increased. RMSE for $\gamma^{(n,p)}_{t_1 \to t_2}$ coefficients decreased as the number of images increased, though, as we noticed earlier, the RMSE in this case has a nonlinear relationship with the number of patients. The RMSE of $\psi^{(g,p)}_{t_1 \to t_2}$ coefficients exhibited the same counterintuitive nonlinear association with number of patients, though in a more pronounced way.

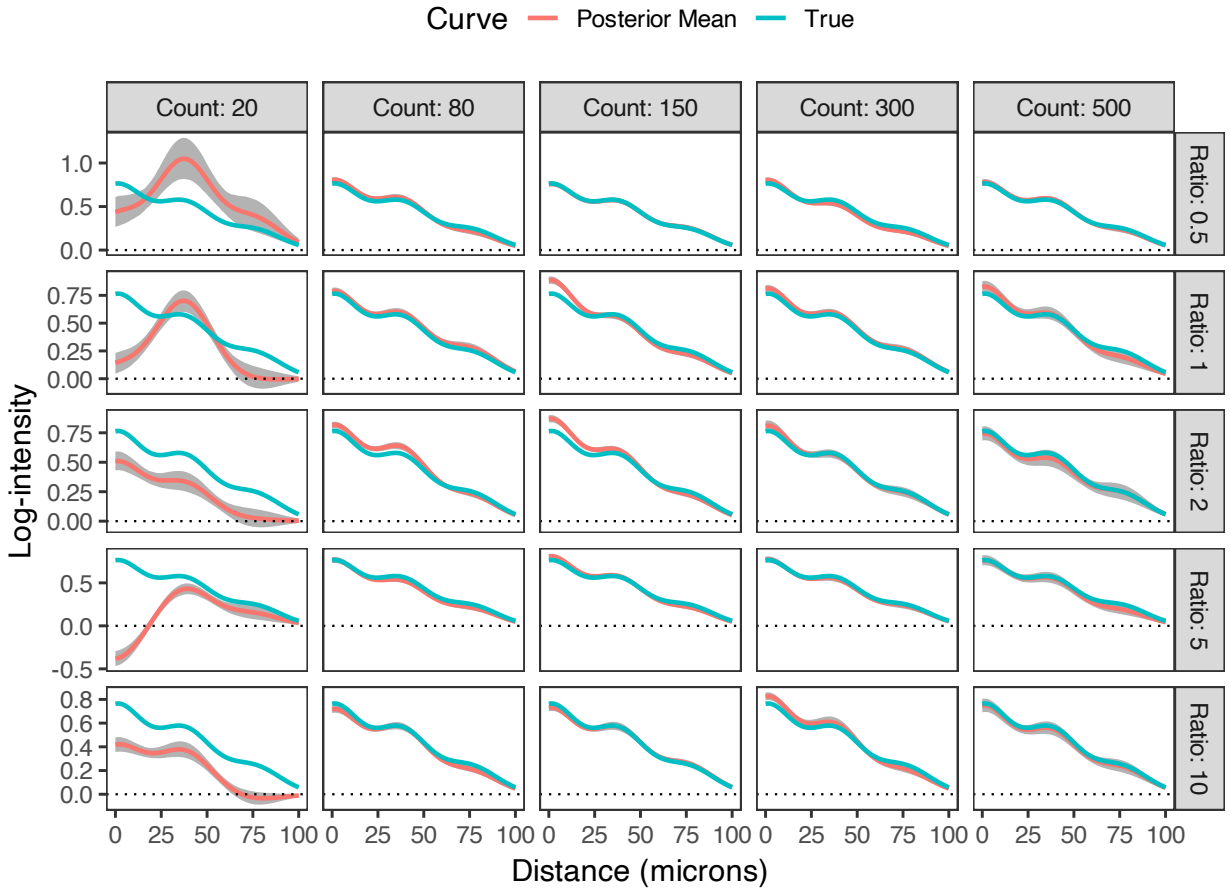



Supplementary Figure S4. RMSE for spatial interaction coefficients across spatial scales, shown separately for each hierarchical level: $\delta^{(m,p)}_{t_1 \to t_2}$, $\gamma^{(n,p)}_{t_1 \to t_2}$, and $\psi^{(g,p)}_{t_1 \to t_2}$. Results are shown for varying cell counts and dummy point ratios.

Finally, we demonstrate an example of an estimated global SIC from this simulation study (Figure S8). Here, we can see that the variance of the SIC estimate decreases both as the number of patients increases and the the number of images increases, which matches our expectations. However, we can also see that there is a slight amount of persistent bias at close range for simulations in which the number of patients equals 40 - this may be indicative of too much shrinkage during estimation and may be the reason for the nonlinear association of RMSE with number of patients.

## 3.7 Robustness to Hard-Core Repulsion

To assess robustness to violations of the Poisson assumption, we simulated spatial point patterns where both source and target cells were generated using a Strauss hard-core process ($r_{\text{hc}} = 10$ microns) rather than a Poisson process. This introduces complete spatial regularity at distances below the hard-core radius,



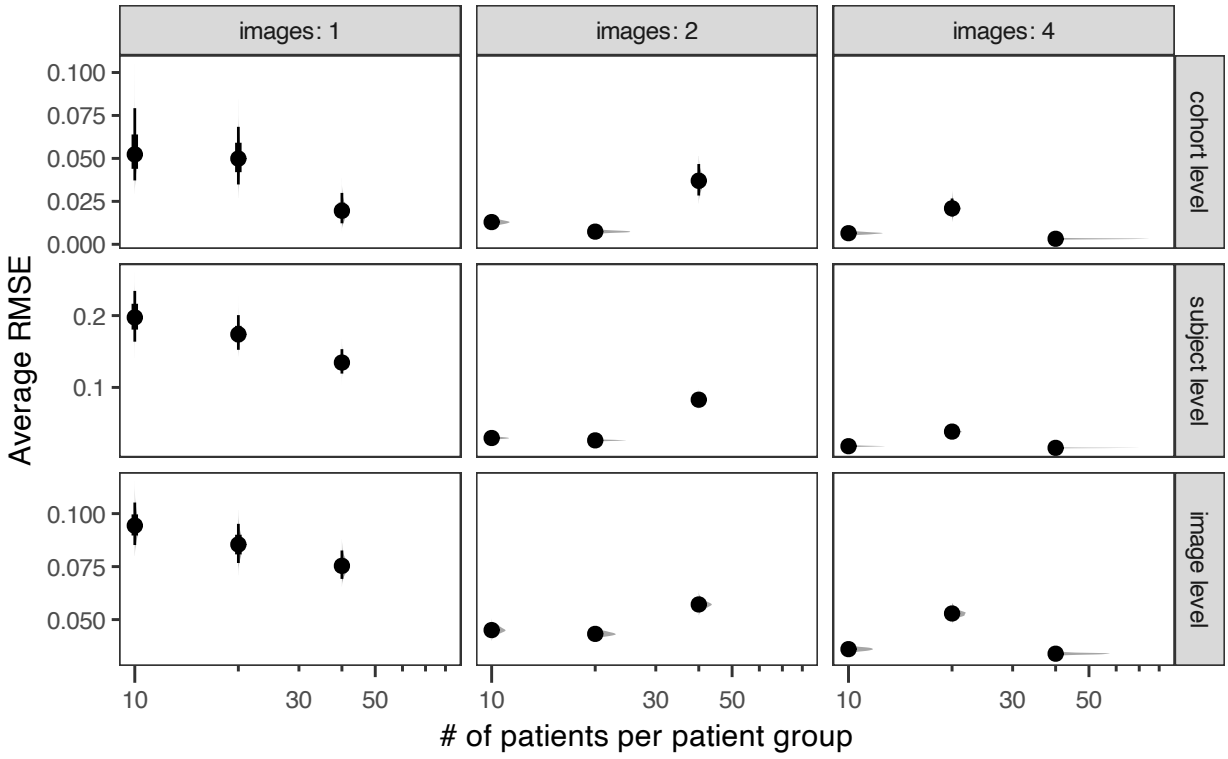

Supplementary Figure S5. Example of estimated cohort-level SIC from dummy point simulations.

mimicking the physical constraint that cell centroids cannot be arbitrarily close.

**[PLACEHOLDER: Full methods and results to be added after simulation is run.]** Preliminary analysis indicates that SHADE successfully recovers spatial interaction effects at distances above MIN_INTERACTION_RADIUS despite the hard-core violation at shorter distances, validating this exclusion approach.

## 3.8 Computational Scaling: Timing Experiments

To assess the practical computational feasibility of SHADE for large-scale spatial datasets, we conducted timing experiments measuring both feature construction time and full model fitting time across simulated datasets ranging from 5,000 to 250,000 cells.




Supplementary Figure S6. Average RMSE for spatial interaction coefficients as a function of dataset size, varying the number of patients and number of images per patient. Results are shown for image-level ($\delta^{(m,p)}_{t_1 \to t_2}$), patient-level ($\gamma^{(n,p)}_{t_1 \to t_2}$), and cohort-level ($\psi^{(g,p)}_{t_1 \to t_2}$) parameters.

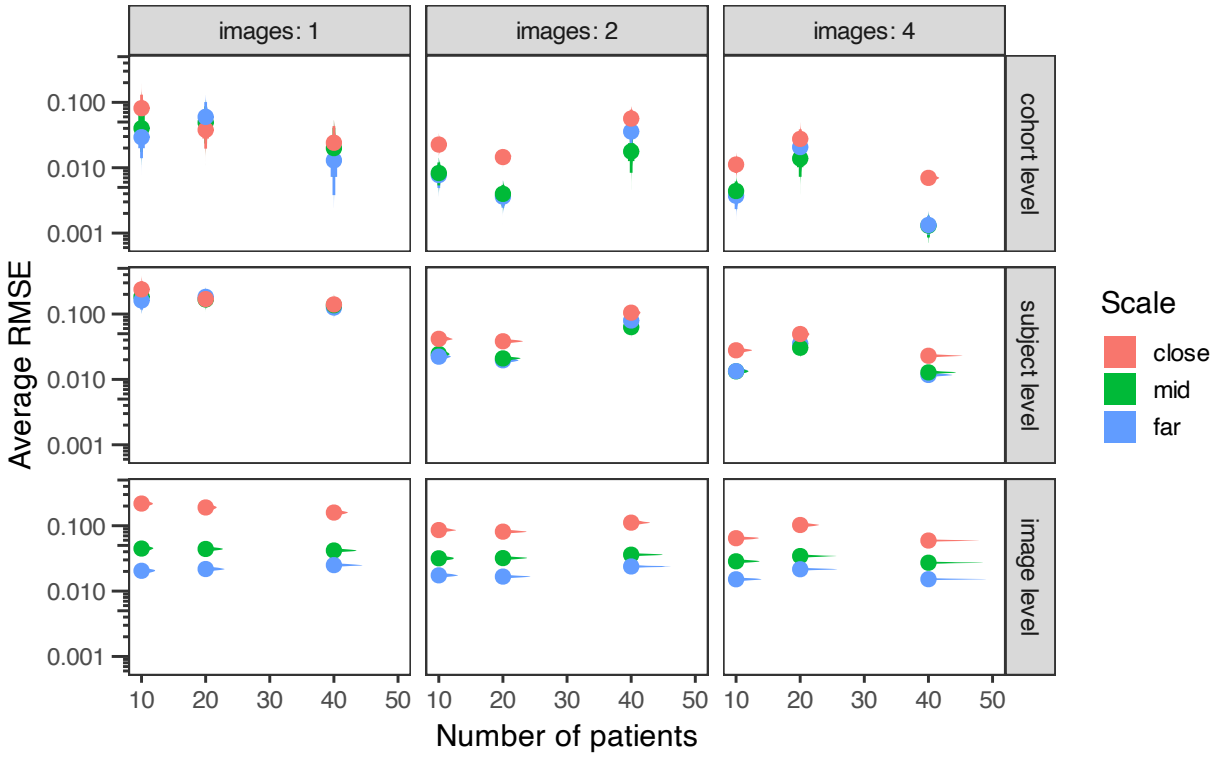
### 3.8.1 Methods

We generated synthetic spatial point patterns with 3 cell types (2 source types, 1 target type) across six cell count conditions: 5,000, 10,000, 25,000, 50,000, 100,000, and 250,000 total cells. To reflect realistic experimental designs, we used a hierarchical structure with 40 patients and 4 images per patient (160 total images), with total cell counts distributed across all images. For each condition, we simulated 20 independent replicates using a fixed spatial window (1500 × 1500 units) per image, so that cell density increased proportionally with cell count. Each pattern was generated with spatial interactions defined by three radial basis functions, matching the setup used in our main simulations.

For each replicate, we measured two key computational steps:

1. **Feature construction time**: The time required to compute pairwise distances between focal and source cells and construct the interaction feature matrix $\boldsymbol{q}_{A_k}(v)$ for all spatial locations (observed cells and dummy points). This step involves distance matrix computation via `spatstat.geom::crossdist` followed by basis function evaluation and summation.

2. **Total model fitting time**: The end-to-end time from loading the prepared data to obtaining posterior




Supplementary Figure S7. Average RMSE across spatial scales for spatial interaction coefficients at the image ($\delta^{(m,p)}_{t_1\to t_2}$), patient ($\gamma^{(n,p)}_{t_1\to t_2}$), and cohort ($\psi^{(g,p)}_{t_1\to t_2}$) levels, across varying numbers of patients and images per patient.

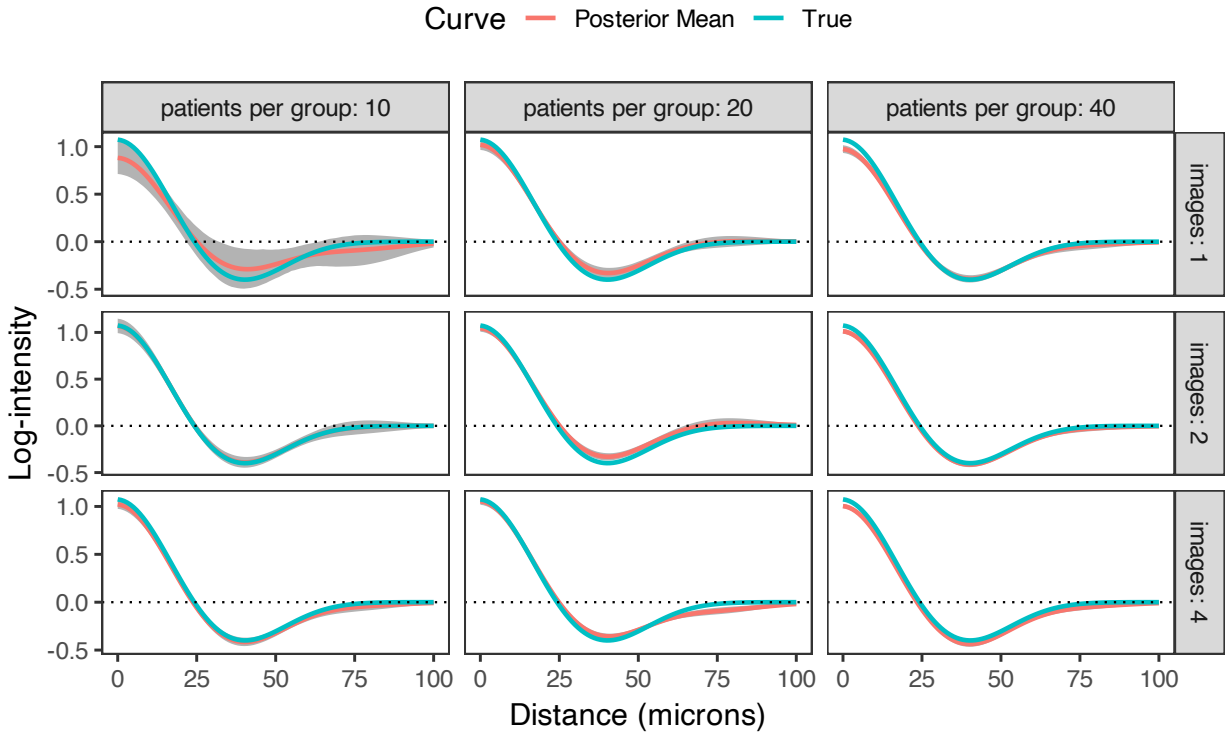
draws from the fitted SHADE model. Models were fit using variational inference with 1,000 posterior draws on a single CPU core (no parallelization).

All timing measurements were performed on a high-performance computing cluster with consistent hardware specifications to ensure comparability across conditions. Timing was recorded using the `tictoc` R package.

### 3.8.2 Results

Figure S9 shows the scaling behavior of feature construction and model fitting times as a function of the total number of cells. Feature construction time exhibited near-quadratic scaling with an empirical exponent of 1.46 (95% CI from log-log regression), consistent with the $\mathcal{O}(n_{\text{focal}} \times n_{\text{source}})$ complexity of computing pairwise distances. Total model fitting time scaled sublinearly with an empirical exponent of 0.85, substantially better than the naive quadratic expectation, due to the efficiency of the variational inference algorithm and reuse of the distance matrix across all basis functions and source types.

These results demonstrate that SHADE remains computationally tractable even for large-scale datasets.




Supplementary Figure S8. Example of an estimated cohort-level SIC from the dataset size simulation study, illustrating how increasing the number of patients and images per patient reduces the variance in SIC estimates.

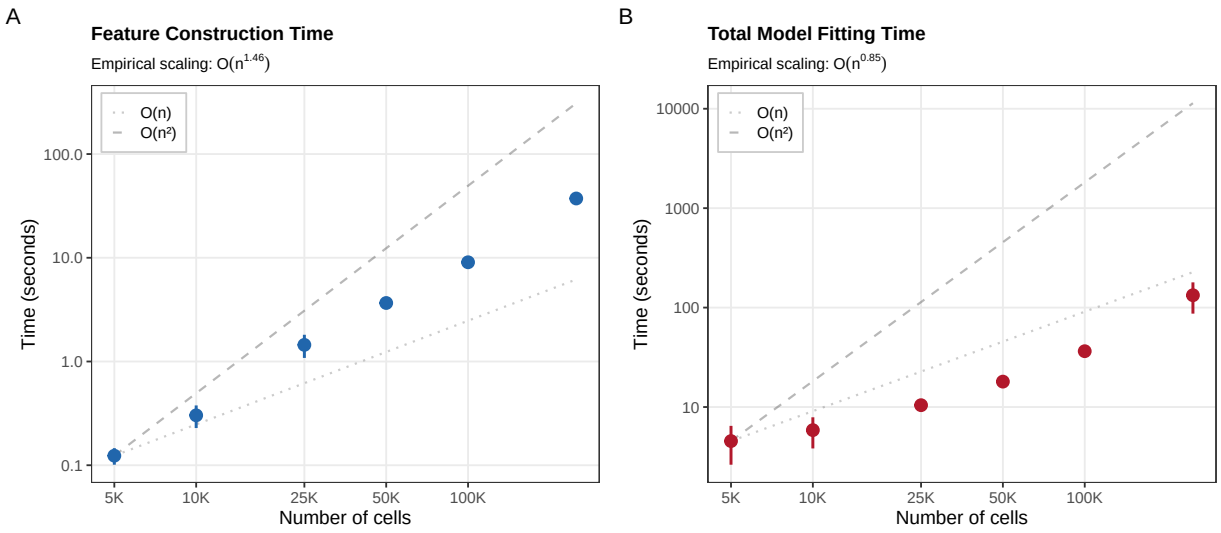
At 100,000 cells—substantially larger than most contemporary spatial profiling datasets—mean total fitting time was 36.4 seconds (±4.1 SD). Even at 250,000 cells, mean fitting time remained 133.1 seconds (±46.2 SD), on the order of minutes rather than hours, making SHADE practical for routine application to large-scale tissue imaging data. Feature construction accounted for approximately 9.1 seconds (100K cells) and 37.3 seconds (250K cells), representing roughly 25–28% of total runtime.

## 3.9 Coverage and Type I Error Performance

In addition to detection power (Figure 5 in the main text), we evaluated the calibration of SHADE's uncertainty quantification and the control of false positive rates across all simulation conditions.

### 3.9.1 Coverage Performance

Figure S10 shows coverage rates for SHADE Hierarchical and SHADE Flat models. Coverage is defined as the proportion of simulated images in which the 95% simultaneous credible band fully contains the true SIC at all distances in the 0–75 $\mu$m range. Results reveal adaptive calibration that depends on source density

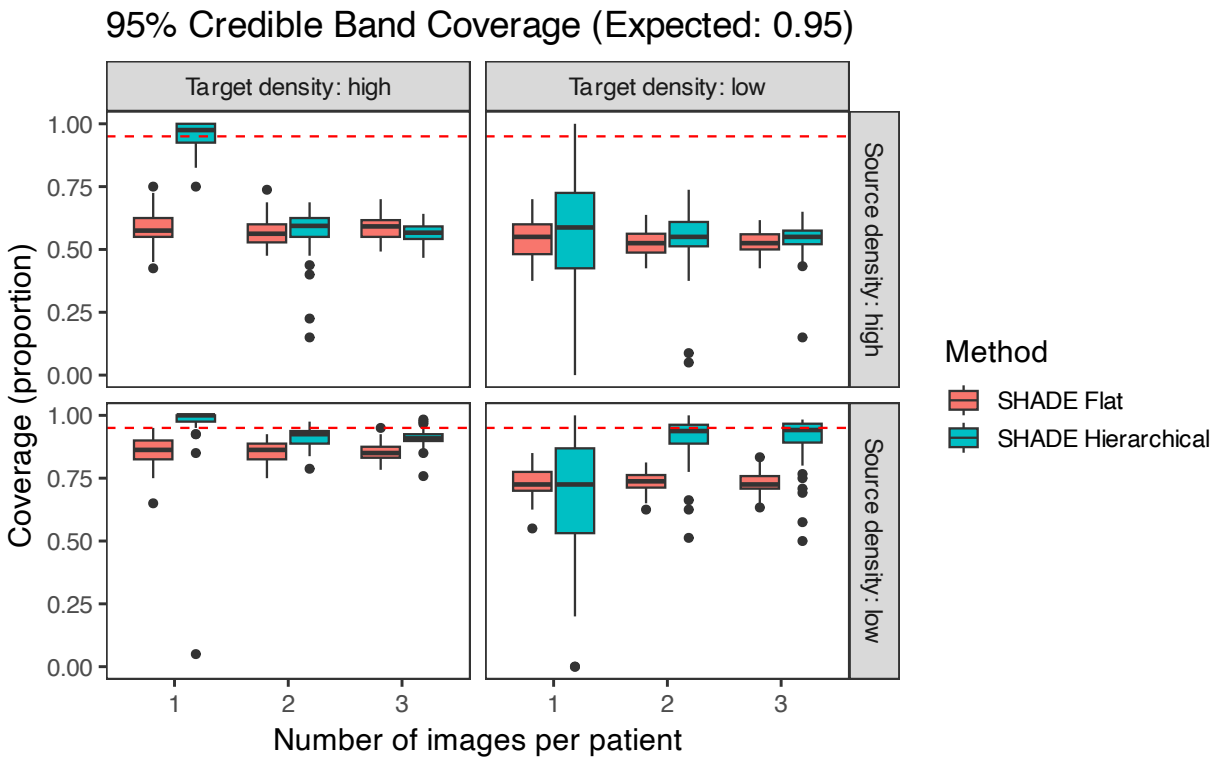



Supplementary Figure S9. Computational scaling of SHADE as a function of cell count. **(A)** Feature construction time shows near-quadratic scaling ($O(n^{1.46})$) with cell count, consistent with the $\mathcal{O}(n_{\text{focal}} \times n_{\text{source}})$ complexity of pairwise distance computation. **(B)** Total model fitting time scales sublinearly ($O(n^{0.85})$) due to parallelization and distance matrix reuse. Points show mean timing across 20 replicates per condition; error bars show $\pm 1$ standard deviation. Dashed and dotted reference lines indicate theoretical $O(n^2)$ and $O(n)$ scaling for comparison.

(conditioning cell type), target density (modeled cell type), and number of images per patient.

**Effect of source and target density.** When source density is low and target density is high, median coverage ranges from 86–100% across different numbers of images, with SHADE achieving perfect 100% coverage when only 1 image is available per patient. This reflects appropriate conservatism when conditioning information is limited. Conversely, when source density is high and target density is low (SHADE's most powerful scenario), coverage drops to 53–59%, as the method trades calibration for sensitivity when abundant conditioning information enables effective hierarchical pooling. When both densities are low, coverage varies dramatically by number of images: with 1 image, coverage is 73%; with 2–3 images, it improves to 94%, demonstrating that hierarchical pooling across multiple images stabilizes uncertainty quantification.

**Effect of number of images.** With 1 image per patient, SHADE Hierarchical exhibits extreme calibration: either overly conservative (100% coverage when source is low and target is high) or poorly calibrated (73% when both are low). With 2–3 images per patient, coverage stabilizes to more consistent values (53–94% depending on density conditions). SHADE Flat maintains consistently poor coverage (53–86%) regardless of number of images, demonstrating that hierarchical pooling is essential for proper calibration.




Supplementary Figure S10. Coverage performance of SHADE Hierarchical and SHADE Flat models across simulation conditions. Boxplots show the proportion of images in which 95% simultaneous credible bands fully contain the true SIC across the entire 0–75 $\mu$m range, stratified by source cell density (rows: conditioning cell type) and target cell density (columns: modeled cell type), with number of images per patient (1, 2, or 3) shown on the x-axis. The red dashed line indicates the nominal 95% level. Median coverage ranges from 53–100%, with adaptive calibration: highest coverage (up to 100%) when source density is low, lowest coverage (53–59%) when source density is high and target density is low, precisely where detection power is highest. Coverage for SHADE Hierarchical improves substantially with multiple images per patient when both densities are low (1 image: 73%; 2–3 images: 94%).

### 3.9.2 Type I Error Control

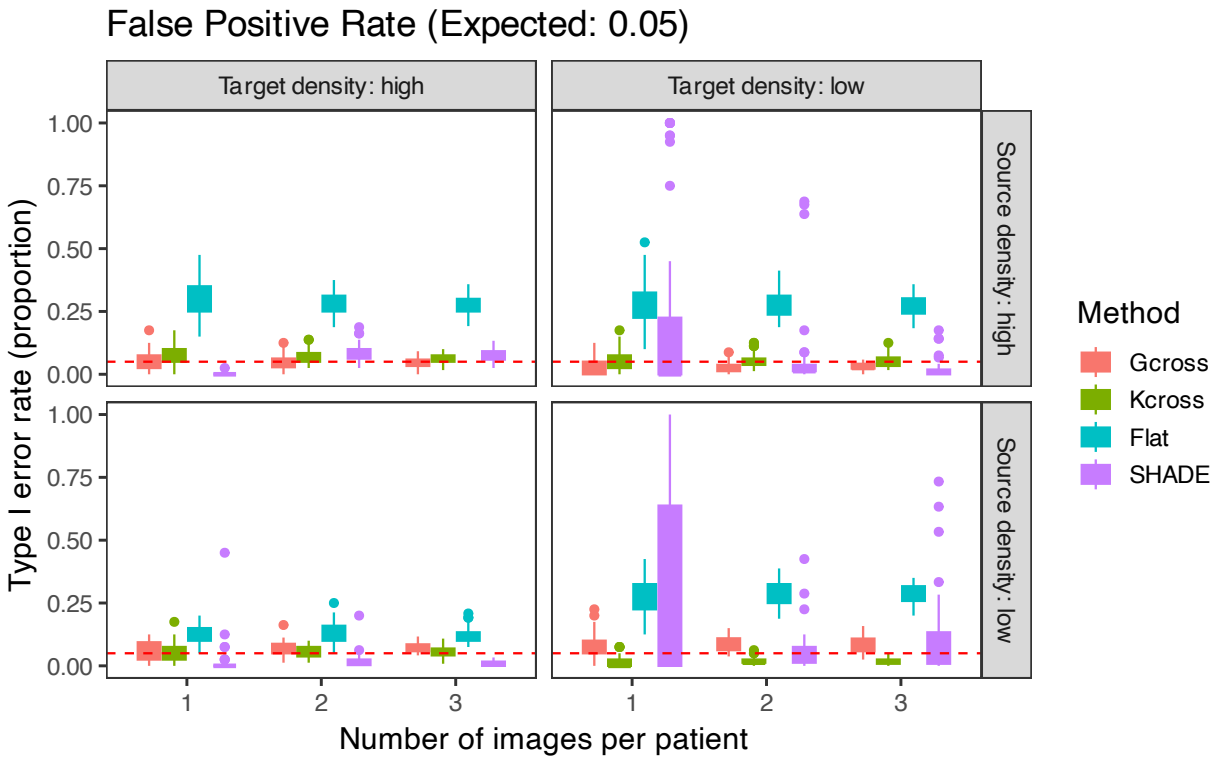
Figure S11 shows type I error rates for all four methods. Type I error is defined as the proportion of null simulations (true SIC = 0 everywhere) in which the method incorrectly detects a non-zero spatial interaction.

**SHADE Hierarchical with 2–3 images per patient** maintains well-controlled median type I error rates (0.8–7.5%) comparable to G-cross (2.5–8.8%) and K-cross (1.3–6.7%). The most conservative control occurs when source density is low and target density is high (median 0.8–1.3%), while the highest rates occur when both densities are high (6.2–7.5%). However, SHADE Hierarchical exhibits greater variability (IQR 0.02–0.12) compared to envelope tests (IQR 0.02–0.05), indicating occasional liberal inference despite good typical-case performance.

**Effect of number of images.** With only 1 image per patient, SHADE Hierarchical shows extreme type I error behavior: perfectly controlled (0%) when source density is low, but with high variability (IQR up to

0.64) when both densities are low. With 2–3 images per patient, type I error rates stabilize substantially, demonstrating that hierarchical pooling requires multiple images for reliable calibration.

**SHADE Flat** shows substantially inflated median type I error rates (27–29%) across all conditions and numbers of images, demonstrating the critical importance of hierarchical pooling for maintaining proper calibration. The flat model's poor performance underscores that estimating SICs independently per image leads to severe miscalibration.




Supplementary Figure S11. Type I error rates across simulation conditions. Boxplots show the proportion of null simulations (true SIC = 0) in which methods incorrectly detect non-zero spatial interactions, for SHADE Hierarchical, SHADE Flat, $G$-cross, and $K$-cross methods. Results are stratified by source cell density (rows: conditioning cell type) and target cell density (columns: modeled cell type), with number of images per patient (1, 2, or 3) shown on the x-axis. The red dashed line indicates the nominal 5% level. SHADE Hierarchical with 2–3 images per patient exhibits well-controlled median type I error rates (0.8–7.5%) comparable to envelope tests (G-cross: 2.5–8.8%; K-cross: 1.3–6.7%), though with greater variability. With only 1 image per patient, SHADE Hierarchical shows extreme behavior (0% or high variability). SHADE Flat shows severe inflation (median 27–29%) regardless of number of images.

### 3.9.3 Interpretation

These calibration results reveal that SHADE Hierarchical exhibits **adaptive calibration** that depends critically on data characteristics: source density (amount of conditioning information), target density, and number of images per patient.

**The source density effect.** Source density determines the quality of conditioning information available for hierarchical pooling. When source density is high, abundant conditioning information enables aggressive but effective information sharing, yielding perfect power (100%) with reduced coverage (53–59%). When source density is low, limited conditioning information leads to appropriate conservatism, with higher coverage (86–100%) but power that depends on having multiple images for pooling.

**The multiple images requirement.** SHADE Hierarchical requires at least 2 images per patient for stable, reliable performance. With only 1 image per patient, the method exhibits extreme and unpredictable behavior across scenarios: overly conservative in some cases (100% coverage, 0% type I error, but poor power when source is low) and poorly calibrated in others (high type I error variability when both densities are low). With 2–3 images per patient, performance stabilizes across all metrics: power improves substantially (from 31% to 100% when source is low and target is high), coverage becomes more consistent (94% when both densities are low), and type I error rates become well-controlled (0.8–7.5%). This stabilization occurs because hierarchical pooling across multiple images provides more reliable estimates of patient-level and population-level effects.

**Comparison with SHADE Flat.** The hierarchical model's critical importance is demonstrated by SHADE Flat's consistent poor performance: severe type I error inflation (27–29%) and inadequate coverage (53–86%) regardless of density conditions or number of images. Estimating SICs independently per image without hierarchical pooling leads to systematic miscalibration that cannot be resolved by having more images per patient.

**Practical recommendations.** For exploratory spatial analyses with at least 2 images per patient and at least one cell type at moderate to high density, SHADE Hierarchical offers superior power with acceptable median type I error control (0.8–7.5%, comparable to envelope tests). The method's adaptive calibration—conservative when evidence is weak, aggressive when hierarchical pooling provides strong signal—is appropriate for hypothesis generation. For confirmatory studies requiring stringent and uniform error control, or when only 1 image per patient is available, envelope tests provide more stable (though conservative) performance. Studies with sparse source cells and only 1 image per patient represent a particularly challenging scenario where envelope tests may be preferable.

## 3.10 Robustness to Spatial Confounding

We tested SHADE's robustness to a realistic model misspecification: discrete spatial compartments (e.g., tumor islands, stromal regions, tissue crypts) that create baseline differences in target cell density independent of source-target interactions. In real tissue, such compartment structure is common but rarely measured or

explicitly modeled. We simulated patterns where both a true source-target interaction existed AND compartment effects influenced target density (multiplicative effects on baseline density with strength 0.8, 1.2, or 1.5), then fit SHADE models that ignored the compartment structure entirely. All simulations used 3 images per patient (the stable regime) and 3 spatial compartments defined by Voronoi tessellation.

Three key metrics assessed robustness: (1) detection power—can SHADE still identify the true interaction?, (2) coverage—do 95% credible bands contain the true source-target interaction curve?, and (3) Type I error—when we simulated patterns with compartment effects but NO source-target interaction, how often did SHADE falsely detect one?

#### 3.10.1 Compartment Confounding Produces Regime-Dependent Bias

When target density is high and source density is high, SHADE achieves excellent detection power (100% median) but severely undercovers (43–52% coverage vs. expected 95%; Figure S13) and exhibits elevated false positive rates that increase with compartment strength (11.7% $\rightarrow$ 17.1% Type I error as compartment effect increases from 0.8 to 1.5; Figure S14). In this data-rich regime, SHADE confidently estimates spatial interactions, but the estimated curves conflate the true source-target relationship with compartment-induced spatial heterogeneity. The credible bands exclude the true interaction curve because SHADE incorrectly attributes compartment effects to the source-target relationship—a classic confounding bias.

#### 3.10.2 Low Target Density Provides Partial Robustness Through Uncertainty

When target density is low, coverage improves substantially (SHADE Hierarchical: 82–93% vs. 43–57% in high target scenarios; Figure S13) and Type I error drops dramatically (1.7–5.8% vs. 11.7–17.1%; Figure S14). Sparse target cell data yields wider credible bands that, while less powerful for detection, better encompass the bias introduced by unmeasured compartment structure. This represents a fortunate side effect of data scarcity: greater uncertainty provides some protection against misattributing compartment effects.

#### 3.10.3 The Low Source / Low Target Regime Shows Conservative Behavior

When both cell types are sparse, SHADE Hierarchical achieves excellent coverage (92–93%) and well-controlled Type I error (1.3–2.9%), but power drops substantially (27–30% median; Figure S12). With limited conditioning information and sparse target data, hierarchical pooling cannot effectively borrow strength across images, leading to conservative inference that maintains calibration at the cost of detection sensitivity. Notably, even envelope tests struggle in this regime (G-cross: 43–45%, K-cross: 23–26% power), suggesting fundamental limitations of spatial point process methods when data are extremely sparse.

#### 3.10.4 SHADE Flat Exhibits Consistently Higher False Positive Rates

Across all regimes, SHADE Flat's Type I error rates (26–40%) exceed those of SHADE Hierarchical (0.8–17.1%; Figure S14), indicating that hierarchical pooling provides some protection against false discoveries even when model assumptions are violated. However, neither version fully controls Type I error in the presence of compartment confounding when target density is high, as both incorrectly detect interactions that arise solely from unmeasured spatial structure. Figure S15 shows the overall trend: Type I error increases with compartment effect strength across all methods, with SHADE Flat consistently exhibiting the highest false positive rates.

#### 3.10.5 Implications for Practice

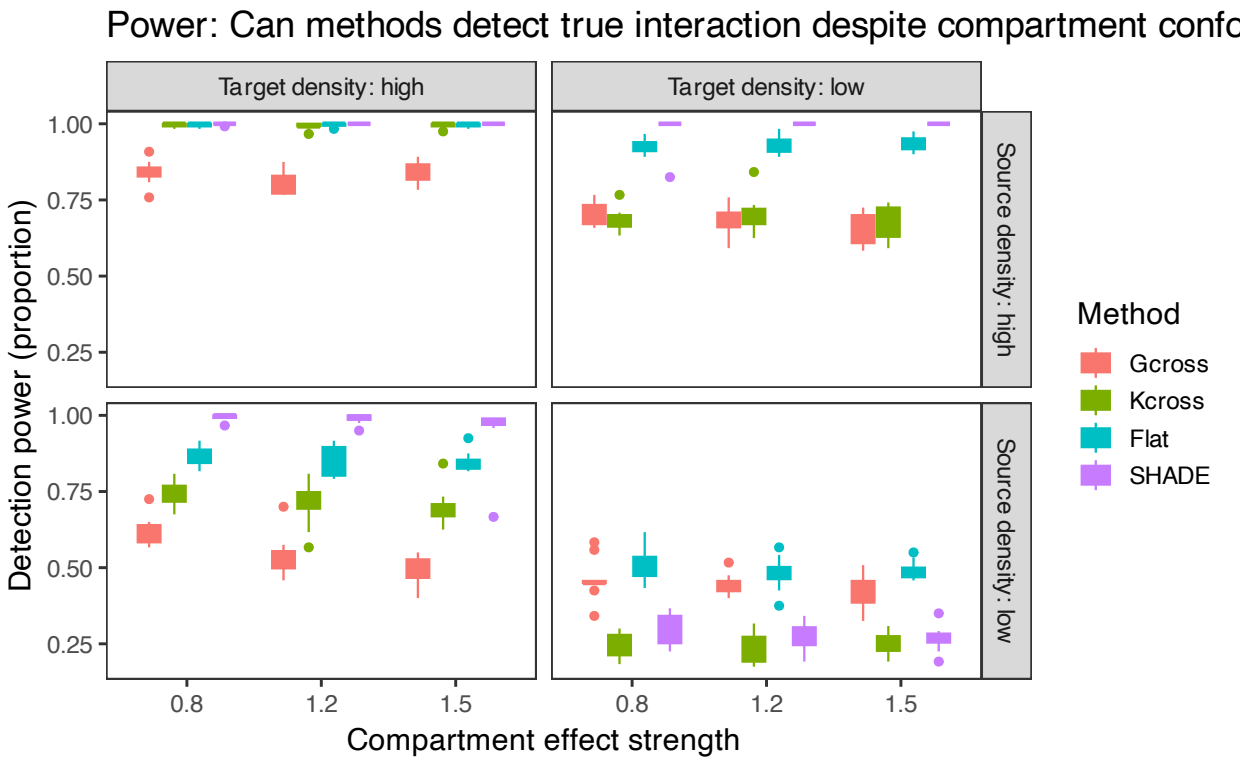
These results demonstrate that unmeasured spatial heterogeneity—compartments, regional density differences, or tissue architecture effects—can produce substantial bias in estimated spatial interactions. When such structure is suspected, analysts should: (1) explicitly model compartments if boundaries can be identified (e.g., tumor/stroma annotations), (2) interpret strong positive findings with caution in high-density scenarios where confounding bias is most severe, or (3) conduct sensitivity analyses by comparing SHADE estimates with and without suspected confounders included. The regime-dependent nature of the bias suggests that data characteristics (target/source densities, number of images) interact with model misspecification to determine whether SHADE maintains calibration or produces biased inference.

# 4 Extended results for colorectal cancer analysis

## 4.1 Detailed description of colorectal cancer dataset and model preparation

The colorectal cancer (CRC) dataset used in this study is a publicly available collection of multiplexed tumor tissue images from 35 patients (Schürch et al., 2020). Each patient contributed four images, each derived from separate biopsies, yielding a total of 140 images. Images were annotated with single-cell resolution across 16 cell types and 56 protein markers, resulting in a multilevel structure: images nested within patients, patients nested within two immune phenotype groups—Crohn's-like reaction (CLR) and diffuse inflammatory infiltration (DII).

The dataset contains approximately 200,000 cells. For analysis, we focused on the eight most abundant cell types. Cell labels were refined to better reflect marker-based characterization: "stroma" cells were reclassified as hybrid epithelial-mesenchymal (E/M) cells based on co-expression of cytokeratin and vimentin (Kuburich et al., 2024), while "smooth muscle" cells were relabeled as cancer-associated fibroblasts (CAFs)

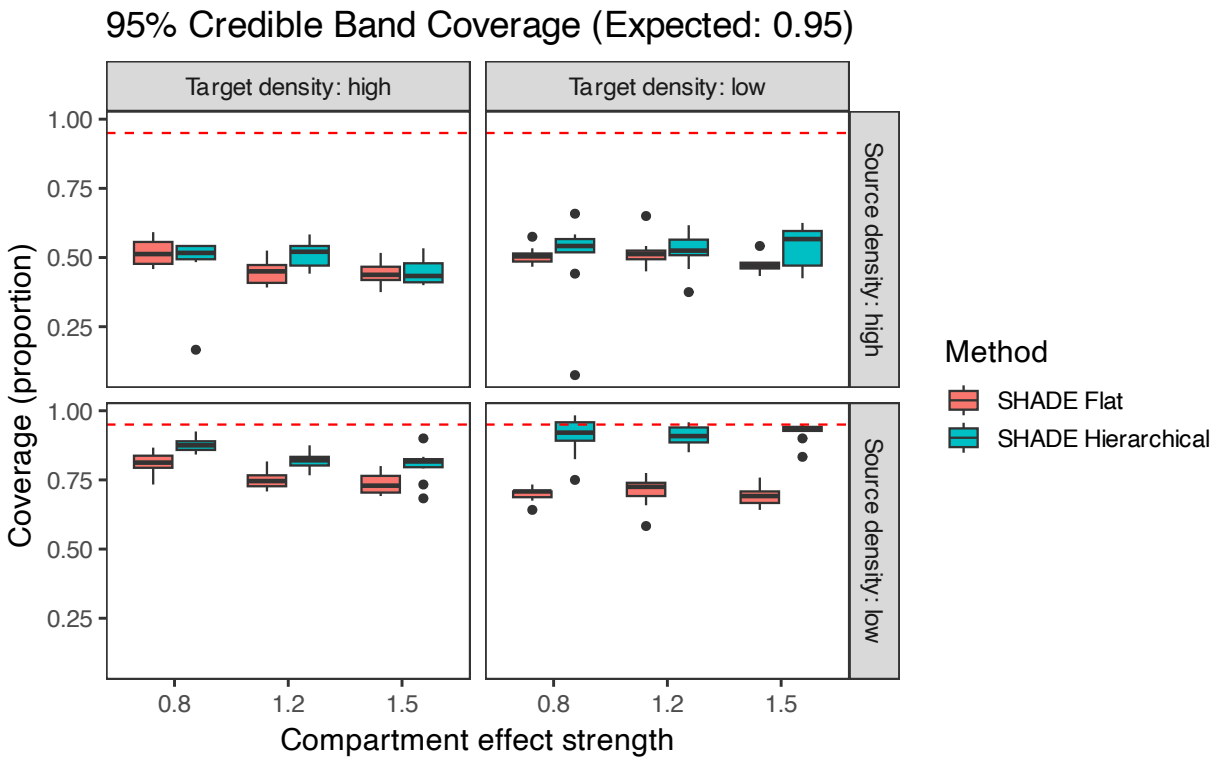



Supplementary Figure S12. Detection power for SHADE and envelope test methods in the presence of compartment confounding. Boxplots show the proportion of datasets in which methods correctly identify non-zero spatial interactions despite the presence of unmeasured compartment structure. Results are stratified by source cell density (rows: conditioning cell type) and target cell density (columns: modeled cell type), with compartment effect strength (0.8, 1.2, 1.5) shown on the x-axis. All simulations used 3 images per patient and 3 spatial compartments. SHADE Hierarchical maintains perfect power (100% median) when either source or target density is high, but power drops to 27–30% when both densities are low.

due to expression of $\alpha$-SMA and vimentin (Cao et al., 2025). Additional cell types included CD163$^+$ macrophages (TAMs), CD8$^+$ T cells, granulocytes, memory CD4$^+$ T cells, tumor cells, and vasculature.

We selected target populations (CD8$^+$ T cells, memory CD4$^+$ T cells, and granulocytes) based on their functional relevance to anti-tumor immunity, and source populations (vasculature, tumor cells, CAFs, TAMs, hybrid E/M cells) based on their roles in tissue architecture and immune modulation. Vasculature structures infiltration pathways, tumor cells and CAFs contribute to immune exclusion, TAMs modulate local inflammation and immune suppression, and hybrid E/M cells may influence spatial dynamics through motility and stromal interactions.

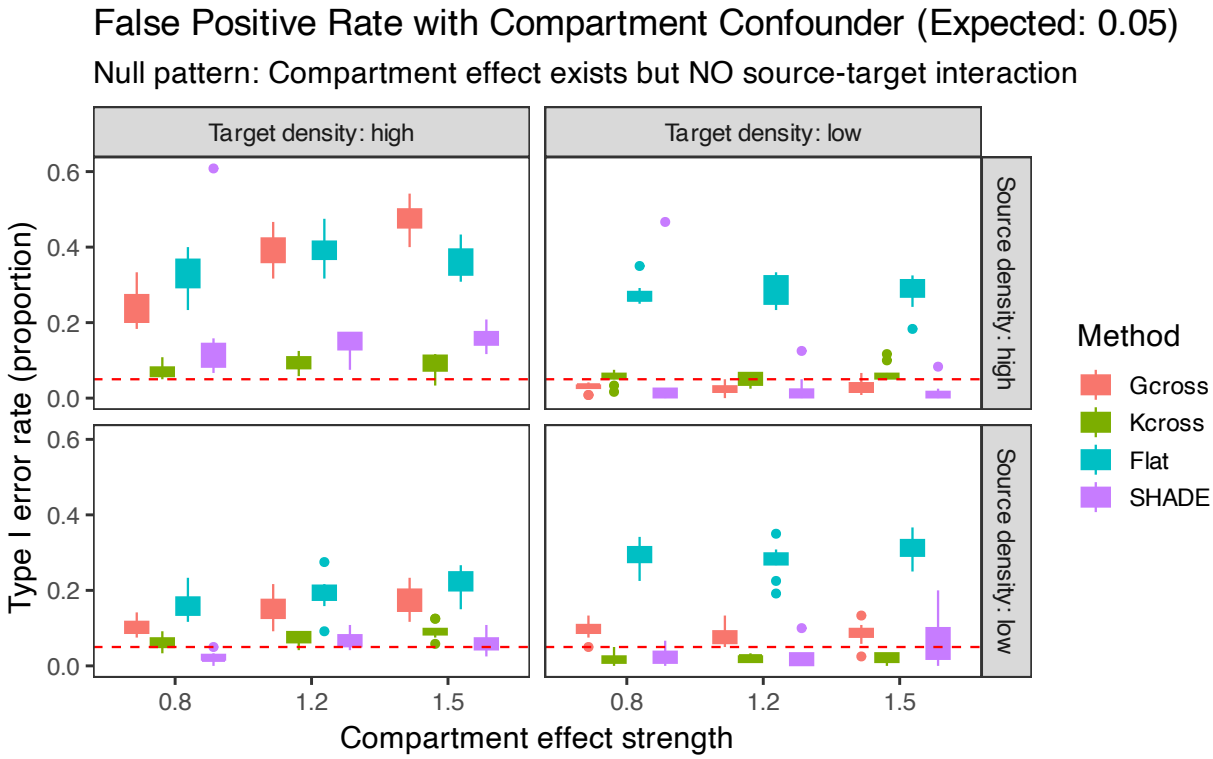
For each target cell type, we constructed a quadrature scheme by generating 1,000 dummy points per image per cell type. To capture distance-dependent spatial interactions, we constructed interaction features $\boldsymbol{q}_{A_k}(v)$ using a set of three radial basis functions $\phi_p$. Data preprocessing included normalization of coordinates and preparation of covariates and interaction features.



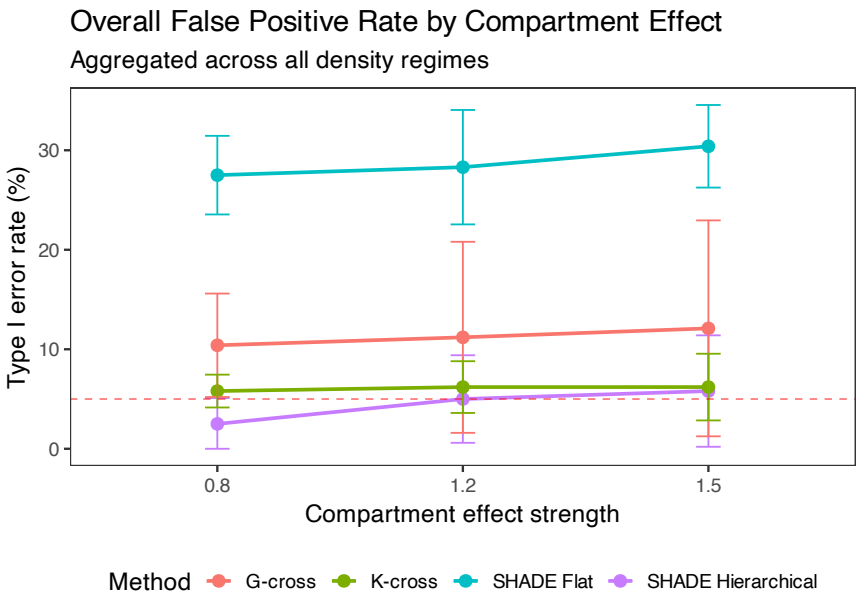

Supplementary Figure S13. Coverage performance in the presence of compartment confounding. Boxplots show the proportion of images in which 95% simultaneous credible bands fully contain the true source-target SIC. Results are stratified by source and target cell densities, with compartment effect strength on the x-axis. The red dashed line indicates the nominal 95% level. When target density is high and source density is high, SHADE severely undercovers (43–52% coverage), incorrectly attributing compartment effects to source-target interactions. When target density is low, coverage improves substantially (82–93%) as wider credible bands encompass confounding bias. When both densities are low, excellent coverage (92–93%) reflects appropriate conservatism.

Model fitting was performed using variational inference with 1,000 posterior draws after fitting. SICs were estimated jointly for all source cell types with respect to each target, providing interpretable, distance-resolved summaries of spatial associations. We set $r_{\min}$ to the dataset's mean cell radius (8–12 $\mu$m depending on cohort) and restricted visualization to $r \geq r_{\min}$.

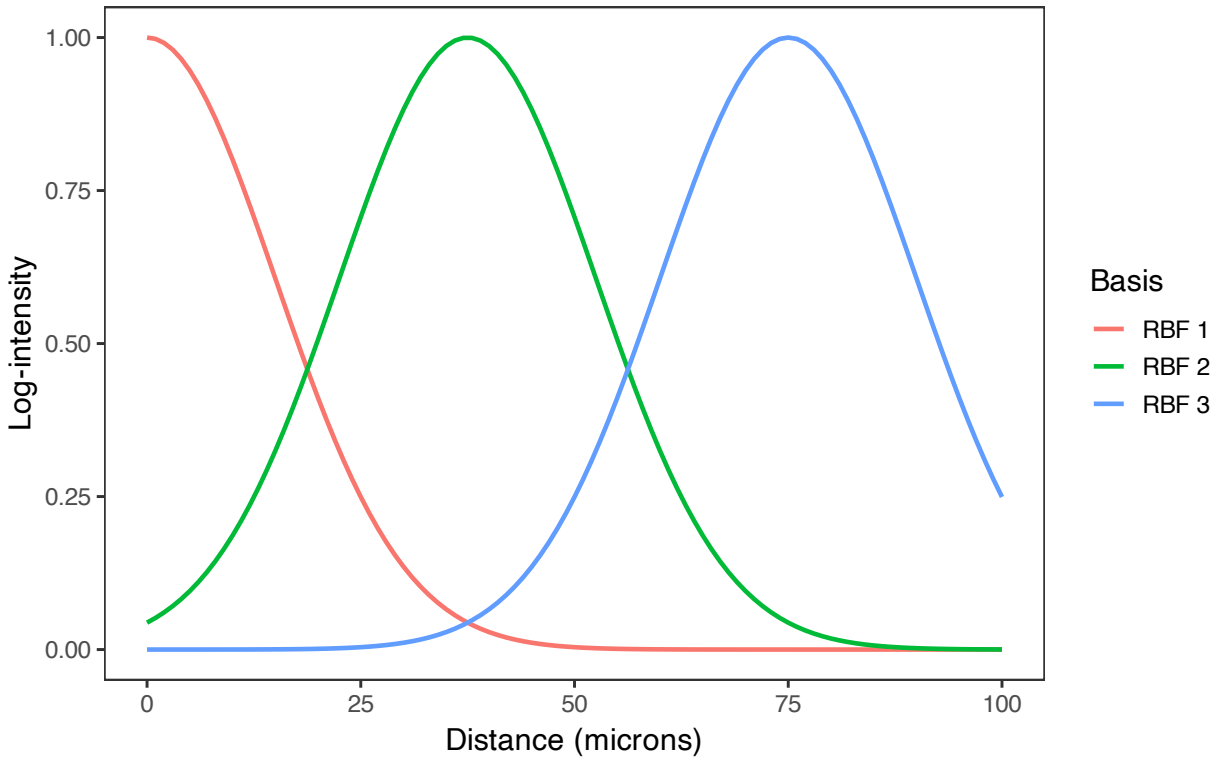
## 4.2 Supplementary Figures



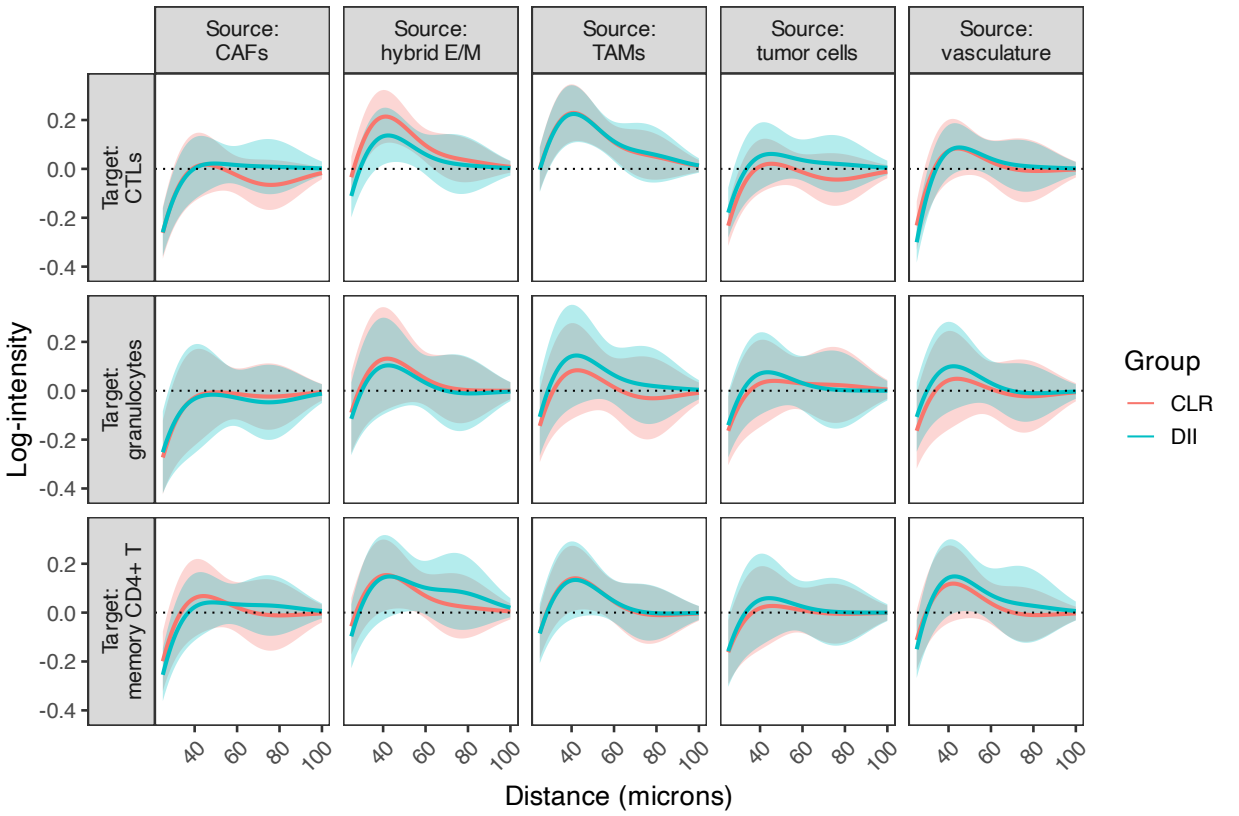

Supplementary Figure S14. Type I error rates when only compartment effects exist (no true source-target interaction). Boxplots show the proportion of null simulations in which methods incorrectly detect non-zero spatial interactions due to unmeasured compartment structure. The red dashed line indicates the nominal 5% level. When both target and source densities are high, SHADE Hierarchical exhibits elevated false positive rates that increase with compartment strength (11.7–17.1%). When target density is low, Type I error is well-controlled (1.7–5.8%). SHADE Flat shows consistently inflated Type I error rates (26–40%) across all conditions, while envelope tests maintain more stable control (G-cross: 2.1–47.1%; K-cross: 1.7–10%).



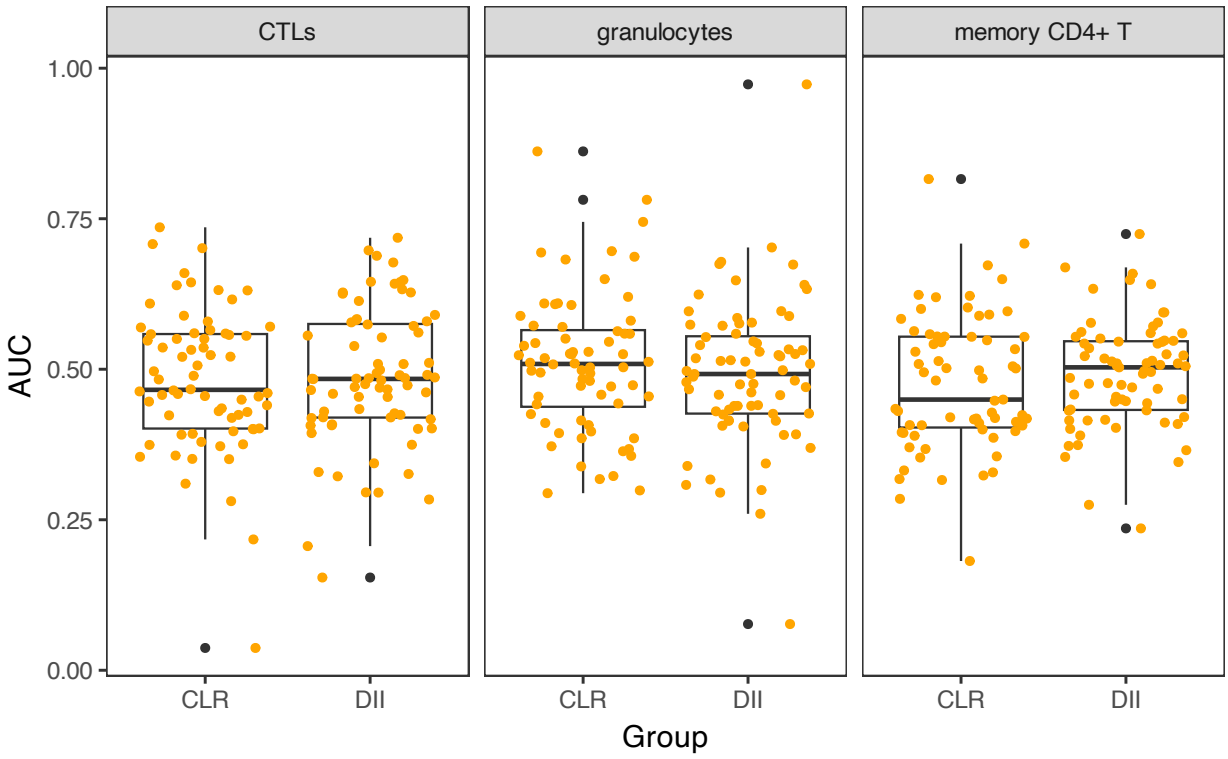

Supplementary Figure S15. Overall Type I error rates by compartment effect strength, aggregated across all density regimes. Line plot shows median Type I error (points) with IQR error bars. All methods exhibit increasing false positive rates as compartment effect strength increases, with SHADE Flat consistently showing the highest rates (26–33%). SHADE Hierarchical maintains the best control (6.2–9.6% median) but still exceeds the nominal 5% level. This demonstrates that unmeasured spatial confounding produces systematic bias that worsens with confounder strength.




Supplementary Figure S16. Radial basis functions $\phi_p(s)$ used to compute distance-based interaction features $\boldsymbol{q}_{A_k}(v)$ in the CRC analysis. These basis functions define the spatial resolution of the SICs.




Supplementary Figure S17. Cohort-level SICs ($\psi^{(g,p)}_{t_1 \to t_2}$) estimated for all source–target cell type pairs in the CRC dataset, stratified by CLR and DII patient groups.




Supplementary Figure S18. Distribution of AUCs by cell type and patient group, evaluating prediction performance of SHADE's conditional intensity model when predicting the spatial organization of each target cell type based on the spatial distribution of all other types.

# 5 Priors for Variance Parameters

We place half-normal priors on the standard deviations of the spatial interaction coefficients at each level of the hierarchy:

$$\sigma_{\text{cohort},p} \sim \text{Half-Normal}(\xi_p, \xi_p), \tag{10}$$

$$\sigma_{\text{patient},p} \sim \text{Half-Normal}(1.5, 10), \tag{11}$$

$$\sigma_{\text{image},p} \sim \text{Half-Normal}(1, 10). \tag{12}$$

where is a user-specified location/scale (in our CRC analysis, we used $\xi_p \in \{5, 3, 1\}$).

# 6 Quantifying Variability in Spatial Interactions

We define SICs hierarchically across cohort, patient, and image levels. For a given source–target cell type pair $(A, B)$, to capture hierarchical variation, we define *patient-level* and *image-level* SICs as:

$$\text{SIC}^{(n)}_{A_k \to B}(s) = \sum_{p=1}^{P} \gamma^{(n,p)}_{A_k} \phi_p(s) \tag{13}$$

$$\text{SIC}^{(m,n(m)}_{A_k \to B}(s) = \sum_{p=1}^{P} \delta^{(m,p)}_{A_k} \phi_p(s) \tag{14}$$

where $\phi_p(s)$ are spline basis functions, and $\gamma^{(n,p)}_{A,B}$ and $\delta^{(m,p)}_{A,B}$ are patient- and image-level coefficients, respectively. To quantify heterogeneity in SICs across patients and images, we compute robust measures of variability at each spatial distance $s$, and summarize them by taking the median across all distances. Specifically:

**Between-patient (within-cohort) variability.** Let $\mathcal{C}$ index cohorts, and let $\mathcal{N}_c$ denote the set of patients belonging to cohort $c \in \mathcal{C}$. For each spatial distance $s$, we compute the cohort-level median SIC as:

$$\overline{\text{SIC}}^{(c)}_{A_k \to B}(s) = \text{median}_{n \in \mathcal{N}_c} \left( \text{SIC}^{(n)}_{A_k \to B}(s) \right)$$

The between-patient (within-cohort) variability for a given cell type pair $(A, B)$ is then defined as:

$$\text{MAD}_{\text{patient}}(A, B) = \text{median}_s \left\{ \text{MAD}_n \left[ \text{SIC}^{(n)}_{A_k \to B}(s) - \overline{\text{SIC}}^{(c(n))}_{A_k \to B}(s) \right] \right\} \tag{15}$$

**Between-image (within-patient) variability.** Similarly, for each patient $n$, we compute the patient-level median SIC:

$$\overline{\mathrm{SIC}}^{(n)}_{A_k \to B}(s) = \mathrm{median}_{m:n(m)=n} \left( \mathrm{SIC}^{(m,n)}_{A_k \to B}(s) \right)$$

The between-image variability is defined as the MAD of image-level SICs from the patient median:

$$\mathrm{MAD}_{\mathrm{image}}(A, B) = \mathrm{median}_s \left\{ \mathrm{MAD}_m \left[ \mathrm{SIC}^{(m,n(m))}_{A_k \to B}(s) - \overline{\mathrm{SIC}}^{(n(m))}_{A_k \to B}(s) \right] \right\} \tag{16}$$

These robust statistics provide interpretable summaries of spatial heterogeneity at each level of the hierarchy while mitigating sensitivity to outliers and small-sample variability. We report these values for each cell type pair and visualize them using heatmaps (Figure 7).