Spatial Transcriptomics Simulation Model

We developed a computational framework for generating large-scale, biologically realistic spatial transcriptomics datasets. The model synthesizes gene expression counts for G genes across a $W \times H$ two-dimensional grid of spatial locations (spots), providing explicit control over spatial expression patterns, sparsity, and count variability.

Spatial Covariance Modeling

For genes with spatially localized expression, we modeled the spatial correlation using the Matérn covariance kernel. To ensure clear parameterization, we define the Matérn function [PMID: 40399936; PMID: 39865128] as a correlation function $\kappa_\nu(r)$, which is a normalized form of the covariance kernel with unit variance ($\sigma^2 = 1$). For two spots separated by Euclidean distance $r = \| \mathbf{s}_i - \mathbf{s}_j \|$, the correlation is:

$$\kappa_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu} r}{\ell} \right)$$

where $\nu > 0$ is the smoothness parameter, $\ell > 0$ is the characteristic length scale, $K_\nu(\cdot)$ is the modified Bessel function of the second kind, and $\Gamma(\cdot)$ is the gamma function. The length scale ℓ controls the range of spatial correlation, while the smoothness parameter ν controls the differentiability of the spatial field.

To ensure numerical stability, we evaluated the kernel using its integral representation, which is well-suited to Gauss-Laguerre quadrature. Using $n = 64$ quadrature nodes $\{x_j, w_j\}$, we approximated the kernel as:

$$\kappa_\nu(r) \approx \frac{1}{2^{\nu-1}\Gamma(\nu)} \sum_{j=1}^{64} w_j \, x_j^{\nu-1} \exp\left(-\frac{\nu r^2}{4\ell^2 x_j} \right)$$

For each (ℓ, ν) pair, kernel values were precomputed radially and truncated when falling below a user-specified cutoff c_{cut}, defining an adaptive spatial support radius.

Generation of Spatially Variable Gene Expression

The latent mean expression $\mu_g(\mathbf{s})$ for gene g at spot $\mathbf{s}$ is modeled as a sum of a baseline expression level and contributions from a set of expression hotspots H_g:

$$\mu_g(\mathbf{s}) = \mu_{\text{baseline},g} + \sum_{m \in H_g} A_m \, \kappa_\nu(\| \mathbf{s} - \mathbf{h}_m \|)$$

where $\mu_{\text{baseline},g}$ is the baseline expression, $\{\mathbf{h}_m\}$ are the hotspot coordinates, and $\{A_m\}$ are the corresponding amplitudes that determine the peak expression intensity.

To model technical artifacts, we introduced zero inflation and stochasticity. First, the latent mean $\mu_g(\mathbf{s})$ was set to zero with a probability p_{drop} to mimic technical dropout. The final observed integer count $X_g(\mathbf{s})$ was then generated by sampling from a Poisson distribution:

$$X_g(\mathbf{s}) \sim \text{Poisson}(\mu_g(\mathbf{s}))$$

The resulting mean $\mu_g(\mathbf{s})$ can be zero either due to dropout or if the spot is sufficiently distant from all hotspots.

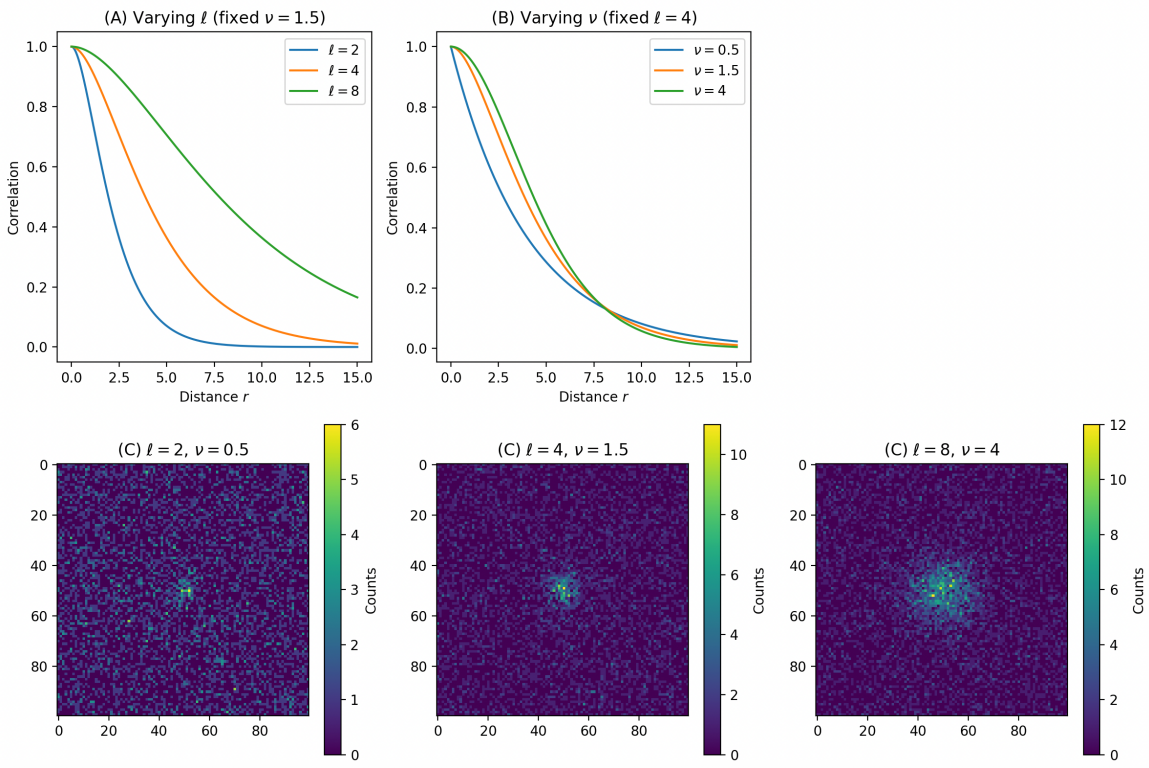

Figure 1 | Simulation of spatial transcriptomics data with tunable Matérn parameters. *(**A**) Effect of varying the correlation length scale ℓ (with $\nu = 1.5$ fixed). Larger ℓ values yield broader spatial correlation. (**B**) Effect of varying the smoothness parameter ν (with $\ell = 4$ fixed). Larger ν values produce smoother spatial fields. (**C**) Simulated 100×100 spatial expression maps for different (ℓ, ν) combinations, showing hotspot locations (bright central pixels) and the decay of expression intensity. Hotspots are generated by convolving their locations with a precomputed Matérn kernel $\kappa_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{\ell}\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}r}{\ell}\right)$, followed by zero inflation and Poisson sampling.*

Simulation Parameterization

Spatial Expression Parameters: The smoothness ν, characteristic length scale ℓ, and kernel cutoff c_{cut} were set to 1.5, 2.5, and 10^{-3}, respectively. The baseline expression

$\mu_{\text{baseline},g}$ and amplitude was 1.0 and 4.0. The dropout rate p_{drop} can be set by users. For power evaluation, one can set one or more hotspot was placed at grid location (30,30). For Type I error evaluation, no hotspots were included.