SimSpace: a comprehensive in-silico spatial omics data simulation framework

Tianxiao Zhao[1,2], Katherine Zhang[1,2], Michelle Hollenberg[1,2], Wen Zhou[3], and David Fenyö*[1,2]

[1]Institute for Systems Genetics, NYU Grossman School of Medicine, New York, NY, USA
[2]Department of Biochemistry and Molecular Pharmacology, NYU Grossman School of Medicine, New York, NY, USA
[3]Department of Biostatistics, NYU School of Global Public Health, New York, NY, USA

Abstract

Tissue function is tightly linked to the spatial organization, and our knowledge of the relationship between tissue structure and function has increased due to the rapid technological development of spatial transcriptomics approaches. A bottle-neck is, however, the analysis of the complex and high-dimensional spatial transcriptomics data. Many computational tools have been proposed for spatial analysis but evaluating them remains challenging due to the scarcity of datasets with known ground truth. To address this, we developed *SimSpace*, a flexible simulation framework that can generate synthetic spatial cell maps with categorical cell type labels and biologically meaningful organization. Cell type spatial patterns are simulated using a Markov Random Field model, enabling the control of spatial autocorrelation and interaction between cell types. This approach captures a broad range of tissue architectures, from well-separated niches to spatially mixed environments. We evaluate the simulated data using spatial statistics metrics (eg. entropy, Moran's I) and benchmark the results against real spatial transcriptomics and proteomics datasets. Our findings show that the simulations closely mirror key spatial properties observed in biological tissues. Additionally, we demonstrate the utility of our framework as a testbed for benchmarking spatial gene detection and cell type deconvolution methods. By providing reproducible, ground-truth-controlled datasets, *SimSpace* supports the development, validation, and comparison of computational methods in spatial omics.

Introduction

Spatial omics technologies are rapidly advancing and have revealed that tissue function is intimately tied to the spatial organization of cells. Technologies like spatial transcriptomics and other spatial omics techniques have enabled the profiling of molecular features in their native spatial context, providing unprecedented insights into tissue architecture and cellular interactions. Understanding these spatial patterns is crucial for many biology questions, including tissue development, embryogenesis, neuroscience, and immunological dynamics [1–4].

Many computational methods and tools have been developed to analyze spatial omics data [5, 6]. Hundreds of methods, with many focusing on spatial omics data analysis, have been proposed for tasks like deconvolution [7–11], spatial variable gene detection [12–17], spatial imputation [18,19], and spatial domain recognition [20–22]. Evaluating these tools, however, remains challenging due to the limited availability of datasets with known ground truth.

In-silico simulations, in this case, offer a promising solution by generating realistic synthetic spatial omics datasets for benchmarking method performance. Currently, many methodology studies are using their simulation settings either too simplistic [23–26] or biased by assumptions that are too specific [14,27,28], resulting in a biased benchmarking result. Several emerging simulation tools have been designed for this task, and these can be classified into two major types, the reference-based and reference-free (or de novo) approaches.

*Corresponding author: David Fenyö, email:David@fenyolab.org

The reference-based approaches, including simulation tools like scDesign [29] and SRTsim [30], always require a spatial omics reference dataset to generate corresponding omics-level and spatial-level simulations, which results in a simulated dataset that has almost identical spatial coordinates and patterns as the reference data. This limits the ability of reference-based tools to generate diverse spatial patterns and cell type distributions, as the simulated datasets are constrained by the reference data.

In contrast, reference-free approaches such as scCube [31] and scMultiSim [32] rely on simplified assumptions for generating spatial features. For example, scCube employs a basic Gaussian distribution to model cell type spatial distributions in its reference-free mode, which does not adequately capture the intricate spatial organization and heterogeneity observed in real tissues. This approach also lacks the ability to model spatial heterogeneity and cell-cell interactions across distinct tissue regions or niches, which are essential for faithfully representing tissue architecture [33]. Similarly, scMultiSim utilizes a straightforward time series model to generate spatial patterns by simulating one cell at each time point, which may not sufficiently reflect the complex spatial relationships among cells. Overall, these reference-free methods are limited in their ability to recapitulate the biological complexity and cellular spatial proximity inherent in real tissue environments, which is crucial for accurately simulating spatial omics data. Additionally, there is no good benchmark for these reference-free simulation tools, making it difficult to systematically evaluate their performance and biological realism.

To address these challenges, we present SimSpace, a comprehensive simulation framework for generating synthetic spatial cell maps with biologically realistic organization. SimSpace adopts a hierarchical modeling strategy: it first simulates spatial cell distributions using a Markov Random Field approach, enabling precise control over niche composition, spatial proximity, and cell-cell interactions. It then generates the corresponding omics profiles, including transcriptomic and proteomic data, to support downstream analyses. The framework supports both reference-free and reference-based simulation modes, allowing users to create diverse tissue architectures ranging from well-defined niches to spatially intermixed environments across resolutions (**Fig. 1A**). Moreover, SimSpace also supports three-dimensional simulations, capturing the full complexity of tissue structure.

We systematically evaluate SimSpace using multiple spatial statistics and benchmark its outputs against real datasets, including a Xenium [34] human breast tumor dataset (spatial transcriptomics) and a CODEX [35] (also known as PhenoCycler-Fusion) human endometrial tumor dataset (spatial proteomics). These evaluations demonstrate SimSpace's utility for benchmarking computational methods such as cell type deconvolution and spatially variable gene detection. SimSpace is implemented as an open-source Python package with flexible parameterization and modular design, making it accessible for a wide range of spatial omics research applications. Together, these features position SimSpace as a valuable resource for generating realistic, ground-truth-controlled spatial datasets to support rigorous benchmarking and method development in spatial omics.

Results

Overview of SimSpace

SimSpace is a comprehensive framework designed to simulate spatial omics data through a hierarchical modeling strategy (**Supp Fig. 1**). For the spatial component, SimSpace first generates spatial niches within a defined spatial grid, capturing spatial autocorrelation and niche-level interactions. Subsequently, within each niche, single-cell phenotypes are simulated to model cell–cell interaction patterns. Additional simulation parameters, such as cell densities and spatial perturbations, are incorporated to generate the final spatial coordinates of individual cells (**Fig. 1B**). For the omics component, SimSpace simulates cellular omics profiles based on spatial coordinates and cell type assignments, utilizing either built-in probabilistic models or external simulation tools. The resulting output is a spatial omics dataset comprising spatial coordinates, cell type labels, and corresponding omics profiles. This dataset can be used for a variety of downstream applications, including the validation and benchmarking of computational methods such as cell type deconvolution, spatial variable gene detection, and spatial clustering.

SimSpace's simulation framework is built upon Markov Random Field (MRF) models, which are widely used probabilistic graphical models for capturing spatial dependencies among random variables arranged on a lattice or grid [36–38]. In the context of spatial omics data, MRFs enable the modeling of local interactions, as each variable (e.g., cell phenotypes) is conditionally dependent only on its immediate neighbors rather than on distant elements (**Fig. 2A**), thereby reflecting the inherent spatial organization of biological tissues.

This approach is particularly well-suited for simulating spatial omics data, as it allows for the generation of spatial patterns that accurately capture the spatial proximity and cell-cell interactions. Unlike existing simulation tools, SimSpace offers a flexible and extensible framework capable of generating both de novo and reference-based spatial omics datasets with novel but biologically realistic spatial architectures. In the absence of reference data, SimSpace can simulate spatial patterns using user-defined MRF parameters and other key factors such as cell densities and spatial autocorrelation. Alternatively, when a reference spatial omics dataset is available, SimSpace can infer model parameters to generate simulated spatial patterns that are different but closely recapitulate the spatial features and biological relevance of the reference. This versatility enables SimSpace to produce a broad spectrum of spatial omics datasets with varying resolutions, spatial organizations, and complexities to support diverse research applications.

Reference-free simulation of spatial omics data

SimSpace enables reference-free simulation of spatial omics data, allowing users to generate synthetic datasets with customizable spatial patterns and cell type distributions independent of any reference dataset. This mode is particularly advantageous when reference data are unavailable or when there is a need to model specific spatial architectures not observed in existing datasets. The reference-free simulation workflow begins with the specification of key hyperparameters, including the number of spatial niches, cell types, the neighborhood structure, and the spatial grid configuration. Using these parameters, SimSpace generates spatial cell maps exhibiting diverse and biologically relevant spatial organizations (**Fig. 2B**). Subsequently, omics profiles for individual cells are simulated in a reference-free manner (**Fig. 2C**), utilizing either the built-in Gamma-Poisson probabilistic model or external single-cell simulation tools such as splatter [39]. To incorporate spatial context into the simulated profiles, SimSpace offers an optional adjustment step based on a spatial ligand-receptor interaction model, which modifies the expression levels according to cell positions and phenotypes. Such interactions are known to play critical roles in spatial transcriptomics [40, 41]. This step enhances the biological realism of the reference-free simulated data by coupling molecular variation to spatial context. Collectively, the reference-free simulation mode provides a flexible platform for generating spatial omics datasets with user-defined spatial and phenotypic characteristics, facilitating method development and benchmarking in the absence of empirical references.

To evaluate the ability of SimSpace to generate spatial patterns that resemble those observed in real spatial omics data, we created eight simulation datasets in reference-free mode using randomly generated parameters. These simulated datasets were compared to a real spatial omics dataset, specifically the Xenium human breast tumor dataset (**Supp Fig. 2**), by analyzing the cumulative distributions of Moran's Index (MI) across cell types. MI is a widely used metric for quantifying spatial autocorrelation and characterizing the degree of spatial clustering among cell phenotypes within a sample [40, 42, 43]. For each cell type in both the simulated and real datasets, we computed MI values and compared their cumulative distributions. The results demonstrate that SimSpace is capable of producing simulated datasets with MI distributions closely matching those of the real dataset (**Fig. 2D**), as indicated by a non-significant Kolmogorov-Smirnov test (p-value $= 0.0604 > 0.05$) between the SimSpace and Xenium datasets. For comparison, we also created eight simulated datasets using scCube [31], a reference-free spatial simulation tool, and assessed their MI distributions relative to the Xenium dataset. In this case, the MI distributions of scCube simulations differed significantly from those of the real data (p-value $= 5.95 \times 10^{-12}$), suggesting that scCube may not fully capture the spatial characteristics of biological tissues. These results indicate that SimSpace can generate spatial patterns that more accurately recapitulate the spatial organization observed in real spatial omics data compared to existing reference-free simulation tools.

Three-dimensional spatial omics data simulation

SimSpace extends its simulation capabilities to three-dimensional spatial omics data, enabling the modeling of tissue architectures with increased biological realism. In this mode, the simulation framework generalizes the MRF model to three spatial dimensions, allowing users to specify a 3D spatial grid, including the number of layers and the neighborhood structure in all three axes. Both reference-free and reference-based simulation modes are supported, with spatial patterns, cell type distributions, and omics profiles generated in a manner analogous to the two-dimensional case. The resulting 3D spatial omics datasets capture complex spatial relationships and heterogeneity inherent to volumetric tissues (**Fig. 2E,F**). This functionality facilitates the benchmarking and development of computational methods designed for three-dimensional spatial omics

data analysis.

Reference-based simulation of spatial omics data

SimSpace further enables reference-based simulation of spatial omics data, allowing users to generate synthetic spatial datasets that closely recapitulate the spatial organization of a given reference sample through an optimization-driven approach (**Fig. 3A**). In this mode, SimSpace first simulates spatial patterns by minimizing the contrastive loss between key spatial statistics, specifically Moran's Index and neighborhood entropy, computed for the simulated and reference datasets. The optimization employs a genetic algorithm to iteratively refines the parameters of the MRF model, including both niche-level and cell-type interaction matrices, as well as additional relevant parameters, to achieve the best match to the reference. Preservation of these spatial statistics ensures that the simulated dataset retains the essential biological features of the reference. At the local scale, neighborhood entropy quantifies the diversity of cell types within the immediate vicinity of each cell, while at the global scale, Moran's Index measures spatial autocorrelation of cell type distributions. By first computing these metrics from the reference dataset and then optimizing the simulation parameters accordingly, SimSpace produces spatial patterns that faithfully reflect the spatial characteristics of the original sample.

When omics data are available in the reference dataset, SimSpace can generate corresponding omics profiles by leveraging the simulated spatial patterns through various built-in or external simulation tools that enables the multimodal single-cell data simulation, such as scDesign3 [29]. The resulting simulated omics profiles are then integrated with the generated spatial coordinates and cell type annotations to produce the final spatial omics dataset.

To evaluate the performance of SimSpace in reference-based simulation, we selected a 1mm $\times$ 1mm tile from the Xenium human breast tumor dataset [34] as the reference and generated a corresponding simulated dataset using SimSpace (**Fig. 3B**). For comparison, we also generated a simulated dataset using scCube, which produces omics profiles in a reference-based manner but generates spatial patterns using its reference-free approach. We hypothesized that an effective simulation should accurately recapitulate key biological features of the reference dataset, such as cell type annotations, and exhibit no significant differences from the reference in downstream analyses.

To assess this, we applied several spatial clustering methods, including BANKSY [21], SpatialPCA [22], and SEDR [44], to the reference and both simulated datasets. The clustering results demonstrated that SimSpace-generated data were closely aligned with the reference, showing minimal batch effects, whereas scCube simulations exhibited a clear batch effect (**Fig. 3C, Supp Fig. 3**). To further investigate whether these differences were attributable to the omics data or the spatial patterns, we replaced the simulated omics profiles in both SimSpace and scCube datasets with the corresponding Xenium reference data, matching cell types at random. Spatial clustering performed on these modified datasets yielded results consistent with the original analysis, indicating that the observed batch effect in scCube is primarily due to differences in spatial patterns rather than omics profiles (**Supp Fig. 4**). Together, these results demonstrate that SimSpace can faithfully reproduce the spatial organization of the reference dataset, generating spatial omics simulations that closely resemble the original data—a level of fidelity not achieved by other reference-based simulation tools.

Benchmarking computational methods using SimSpace simulated datasets

In contrast to existing simulation tools, SimSpace enables the generation of an unlimited number of spatial omics datasets with customizable spatial patterns and properties. This flexibility positions SimSpace as a powerful in silico benchmarking platform for spatial omics computational methods. By producing a diverse array of simulated datasets with known ground truth, SimSpace facilitates rigorous evaluation and comparison of computational approaches under controlled and reproducible conditions. To illustrate its utility, we apply SimSpace for benchmarking two key spatial omics analysis tasks: cell type deconvolution and spatial variable gene detection.

Cell type deconvolution

Cell type deconvolution is a fundamental task in spatial transcriptomics analysis, aiming to estimate the relative abundances of distinct cell types within each spatial location, particularly in datasets with multicellular resolution [45,46]. In this study, we systematically evaluated five widely used cell type deconvolution methods: cell2location [7], RCTD (in both 'full' and 'multi' modes) [8], CARD [9], Seurat [10], and spa-

tialDWLS [11]. These methods represent state-of-the-art approaches for deconvolving cell type composition in spatial omics data and have been extensively benchmarked in recent comparative studies [47, 48].

To systematically benchmark these methods, we employed SimSpace to generate a series of reference-based simulated datasets derived from randomly selected tiles of the Xenium human breast tumor dataset. To capture a broad range of biological and spatial diversity, we varied both the number of cell types and the spatial distribution patterns for each reference tile, resulting in 16 distinct simulated datasets (**Supp Fig. 5**). Each simulated dataset was further processed by applying four different convolution kernel sizes, thereby modeling a range of spatial resolutions. The deconvolution methods were then applied to these convolved datasets, and their performance was assessed by comparing the predicted cell type proportions to the ground truth proportions from the original SimSpace simulations (**Fig. 4A**). Evaluation metrics included Pearson correlation coefficient (PCC), root mean square error (RMSE), and Jaccard Index (JI), were calculated for each convolved spot to quantitatively assess the accuracy of cell type proportion estimates.

The benchmarking results indicate that most deconvolution methods achieve high Pearson correlation coefficients (PCC) with the ground truth cell type proportions, with RCTD and CARD consistently demonstrating the highest accuracy (**Fig. 4B, C, Supp Fig. 6**). Comparable trends are observed for the root mean square error (RMSE) metric. Notably, while most methods exhibit improved performance with increasing convolution kernel sizes, Seurat and spatialDWLS perform optimally at smaller kernel sizes, suggesting that these methods may be more sensitive to spatial resolution and potentially less suitable for datasets with larger spot sizes, such as Visium. To further assess the ability of each method to detect the presence of specific cell types, we evaluated performance using the Jaccard Index (JI). Consistent with the PCC and RMSE results, Seurat and spatialDWLS achieve the highest JI at smaller kernel sizes, whereas other methods show increased JI with larger kernel sizes. Additionally, we investigated the impact of cell type complexity by varying the number of cell types in the simulated datasets from 8 to 14. The results reveal that all methods tend to perform better with fewer cell types (**Fig. 4D**), which is expected given the increased complexity and heterogeneity associated with a greater number of cell types, making accurate deconvolution more challenging.

In summary, our benchmarking of cell type deconvolution methods using SimSpace-simulated datasets demonstrates that most approaches are capable of accurately recovering cell type proportions across a range of spatial resolutions and complexities. Among the evaluated methods, RCTD and CARD consistently exhibited the highest accuracy across varying kernel sizes. Additionally, we observed that increasing the number of cell types in the dataset led to a reduction in deconvolution accuracy for all methods, reflecting the greater challenge posed by higher cellular complexity. These results underscore the value of SimSpace for generating controlled, ground-truth datasets that enable systematic and rigorous evaluation of spatial deconvolution algorithms under diverse experimental conditions.

Spatial variable gene detection

Spatial variable gene (SVG) detection is another common task in spatial transcriptomics data analysis, where the goal is to identify genes that exhibit non-random, informative spatial patterns in the sample [49, 50]. In this study, we benchmarked six different spatial variable gene detection methods, including Hotspot [12], SpatialDE [13], scBSP [14, 15], SPARK [16], Giotto [17], and Moran's I implemented in the Seurat package [51].

To systematically benchmark these spatial variable gene detection methods, we utilized the Xenium human breast tumor dataset as a reference to generate SimSpace datasets under both reference-based and reference-free simulation modes (**Fig. 5A**). For the reference-based scenario, we generated a simulated dataset that closely recapitulates the spatial patterns and gene expression profiles of the Xenium reference. Each SVG detection method was then applied to both the simulated and reference datasets, enabling direct comparison of their outputs and assessment of method reproducibility. In parallel, we generated eight additional simulated datasets using reference-free spatial patterns while maintaining reference-based gene expression profiles. SVG detection methods were applied to these datasets to evaluate the robustness of each approach and to investigate the impact of varying spatial patterns on SVG detection performance.

Different SVG detection methods employ distinct statistical models and underlying assumptions, resulting in p-values with varying scales and distributions. To enable rigorous and unbiased comparison across methods, we implemented a rank-based normalization procedure: for each method, p-values were ranked from lowest to highest and then mapped onto a common scale [52]. This approach facilitates direct comparison of results across both methods and datasets. Using the reference-based simulated dataset, we assessed the concordance

of normalized p-values among the different methods (**Fig. 5B**). While there was general agreement in the identification of significant genes, notable discrepancies in gene rankings were observed, particularly for SpatialDE and scBSP, which exhibited rankings distinct from those produced by the other methods. These differences underscore the impact of methodological choices on SVG detection outcomes.

To further evaluate reproducibility, we compared the rank-normalized p-values obtained from the reference-based simulated dataset to those from the original Xenium reference dataset. With the exception of SpatialDE and Moran's Index, most methods demonstrated strong concordance between simulated and reference data, as reflected by Pearson correlation coefficients (PCC) ranging from 0.6 to 0.9 (**Fig. 5C, Supp Fig. 7**). Consistent results were observed using Kendall's tau, a rank-based correlation metric, which also indicated high agreement for the majority of methods. Notably, scBSP achieved the highest correlation with the reference dataset, suggesting near-perfect reproducibility. To further assess binary classification performance, we identified genes as 'significant' SVGs based on a p-value threshold of 0.05 and quantified the overlap between simulated and reference datasets. All evaluated methods achieved high precision and recall, with F1 scores between 0.8 and 0.9 (**Fig. 5C**). Collectively, these findings demonstrate that most spatial variable gene detection methods yield highly reproducible results between SimSpace-generated simulated data and real datasets, underscoring the utility of SimSpace as a robust simulation platform for benchmarking spatial analysis tools.

For the reference-free simulated datasets, we applied the same rank-based normalization procedure to the p-values generated by each method and evaluated the consistency of SVG detection across datasets (**Fig. 5D**). To quantitatively assess concordance, we computed Kendall's W, which measures agreement among multiple raters, as well as the average Spearman correlation. Among the methods evaluated, scBSP exhibited the highest level of consistency across reference-free simulations, indicating strong robustness to variations in spatial patterns. Hotspot also demonstrated substantial concordance, while the remaining methods showed comparatively lower levels of agreement. These results suggest that scBSP and Hotspot are more robust to changes in spatial organization, whereas other methods may be more sensitive to the underlying spatial structure of the data.

In summary, our benchmarking of spatial variable gene (SVG) detection methods using SimSpace-simulated datasets demonstrates that most approaches yield highly reproducible and robust results when compared to real spatial omics data. Notably, methods such as scBSP and Hotspot exhibited the highest levels of consistency and robustness to variations in spatial patterns, whereas others, including SpatialDE and Moran's I, were more sensitive to changes in spatial structure. The application of rank-based normalization to p-values facilitated equitable comparison across methods and datasets, revealing that the majority of evaluated approaches can reliably identify spatially informative genes under both reference-based and reference-free simulation scenarios. Together, these findings highlight the utility of SimSpace for generating controlled, ground-truth spatial omics datasets, thereby enabling systematic evaluation and comparison of SVG detection algorithms and supporting the advancement of spatial analysis methodologies.

SimSpace for spatial proteomics data simulation

Besides spatial transcriptomics, SimSpace can also be applied to simulate spatial proteomics datasets, which is important for understanding the spatial organization and tissue environment [53]. By incorporating protein expression data and spatial information, SimSpace can generate realistic spatial proteomics datasets that reflect the complex interactions between different cell types in their microenvironment. To demonstrate this, we applied SimSpace to a selected tile of a CODEX (CO-Detection by indEXing) human endometrial tumor dataset, which contains spatial proteomics data with 31 protein markers and 206 cells with 9 phenotypes. CODEX is a highly multiplexed imaging technology that enables simultaneous detection of dozens of protein markers at single-cell resolution within intact tissue sections, where data are collected by iterative cycles of antibody staining, imaging, and fluorophore removal, allowing for comprehensive spatial profiling of the tumor microenvironment [54]. We generated a reference-based simulated dataset based on the reference tile from the CODEX dataset, and then applied SimSpace to simulate the spatial patterns of the tumor samples and the corresponding protein marker expression profiles in the tumor microenvironment (**Fig. 6A**).

By leveraging the MRF model, SimSpace encodes the spatial organization of the tissue into the potential parameters, which can be interpreted as spatial proximity between different cell types (**Fig. 6B**). A positive potential indicates a strong attraction between two cell types, while a negative potential indicates spatial exclusion. For example, the parameter matrix inferred from the CODEX reference dataset reveals strong

attraction between cytotoxic T cells and CD4+ T cells ($\theta = 0.92$), as well as between endothelial and stromal cells ($\theta = 1.40$), suggesting that these cell types tend to co-localize in the microenvironment. These interaction patterns are faithfully recapitulated in the SimSpace-generated tissue, where the simulated organization of selected cell types closely mirrors the empirical CODEX data (**Fig. 6C**). For instance, the simulated clustering of T cells and the alignment of endothelial and stromal cells are consistent with those observed in the reference sample and other studies [55, 56], underscoring SimSpace's capacity to generate biologically realistic spatial architectures based on inferred interaction parameters. Conversely, tumor cells exhibit marked spatial exclusion from most immune cell types ($\theta < 0$) except for neutrophils and cytotoxic T cells, indicating a distinct spatial organization that is consistent with the known features of endometrial tumor microenvironments [57, 58]. Importantly, this illustrates how SimSpace can be used as an exploratory tool for better understanding spatial biology by identifying and interpreting spatial interaction patterns present in reference datasets.

Discussion

SimSpace fills a critical gap in spatial omics research by enabling the systematic generation of synthetic datasets with known ground truth, controllable spatial organization, and flexible omics profiles. Our results demonstrate that SimSpace's MRF-based modeling approach can recapitulate a wide spectrum of tissue architectures, from highly structured niches to spatially intermixed environments, and accurately reproduce key spatial statistics like Moran's Index and neighborhood entropy. This fidelity is evident in both reference-free and reference-based simulation modes, where SimSpace outperforms existing tools like scCube in matching the spatial properties of real datasets, as shown by statistical comparisons with Xenium spatial transcriptomics data.

A major strength of SimSpace is its hierarchical simulation framework, which decouples spatial niche generation, cellular spatial pattern modeling, and omics profile simulation. This modularity allows users to independently control spatial autocorrelation, cellular spatial proximity, cell density, and marker gene expression, or to fit these parameters to a reference dataset. The ability to simulate three-dimensional spatial omics data further extends SimSpace's applicability to emerging 3D spatial technologies, supporting the benchmarking of methods that require volumetric tissue context. Through comprehensive benchmarking, we demonstrated that SimSpace-generated datasets provide a rigorous and reproducible platform for evaluating computational methods. In cell type deconvolution, SimSpace enabled the systematic assessment of multiple algorithms across varying spatial resolutions and cell type complexities, revealing that methods such as RCTD and CARD are robust to changes in spatial scale, while others are more sensitive to kernel size or cell type diversity. For spatial variable gene (SVG) detection, SimSpace allowed us to compare the reproducibility and robustness of six widely used methods under both reference-based and reference-free scenarios. Our results highlight that while most SVG methods are consistent across simulated and real data, some (e.g., scBSP and Hotspot) are more robust to changes in spatial pattern, whereas others are more sensitive to spatial structure.

Importantly, SimSpace's application to spatial proteomics data, as demonstrated with the CODEX human endometrial tumor dataset, illustrates its versatility beyond transcriptomics. By inferring and encoding cell-cell interaction matrices from reference data, SimSpace not only simulates realistic spatial architectures but also provides interpretable parameters that can reveal underlying biological organization, such as immune cell exclusion or stromal-endothelial co-localization in the tumor microenvironment.

Despite these advances, several limitations should be noted. First, the current implementation of SimSpace primarily focuses on spatial omics data simulation, and while it can generate corresponding omics profiles, the accuracy of these profiles depends on the underlying probabilistic models or external simulation tools used. Future work will aim to enhance the omics simulation models, potentially incorporating additional biological knowledge such as gene regulatory networks or signaling pathways to improve the fidelity of simulated profiles. Second, while SimSpace supports both reference-free and reference-based simulation modes, the latter requires a reference dataset to guide the spatial pattern generation. This may limit its applicability in scenarios where such references are not available or where users wish to explore novel spatial patterns not present in existing data. In such cases, the reference-free mode can still be used, but the spatial patterns may not fully reflect the complexity of target real biological tissues and may require careful parameter tuning to achieve realistic results. Finally, while SimSpace provides a powerful framework for simulating and representing spatial omics data, it is important to recognize that no simulation tool can fully capture

the complexity of biological systems. As such, SimSpace should be used as a complementary tool alongside empirical data to validate and benchmark computational methods, rather than as a replacement for real experimental datasets.

In summary, SimSpace provides a flexible, extensible, and biologically informed platform for spatial omics data simulation. By enabling the generation of reproducible, ground-truth datasets with tunable spatial and molecular features, SimSpace empowers the development, benchmarking, and comparison of computational methods, and supports deeper exploration of spatial organization in complex tissues.

Method

SimSpace framework

The SimSpace model simulates spatial omics data in a hierarchical manner. In general, we donate the joint distribution of a cell's omics and spatial information as:

$$f(\sigma, c, x, y) = f(\sigma|c, x, y)f(c|x, y)f(x, y) \quad (1)$$

where σ is the cellular omics profile, c is the cell phenotype, and (x, y) is the cell coordinates. $f(\cdot)$ stands for the joint distribution of the variables. The first term $f(\sigma|c, x, y)$ represents the omics profile given the cell type and spatial coordinates, which can be modeled using reference-based single-cell omics simulation tools or reference-free probabilistic models. The second term $f(c|x, y)$ represents the cell type distribution given the spatial coordinates, which can be modeled using a Markov Random Field model to capture spatial autocorrelation and interactions between cell types. The third term $f(x, y)$ represents the spatial distribution of cells, which can be defined based on a user-specified spatial pattern and the cell density parameters.

In this framework, each component of the joint distribution can be flexibly specified to reflect different biological assumptions or experimental designs. To perform the simulation, in a reverse order, we first generate the spatial coordinates (x, y) of the cells, then simulate the cell types c based on the spatial coordinates, and finally generate the omics profiles σ based on the cell types and spatial coordinates. Specifically, the simulation process can be summarized as follows:

$$\begin{aligned}(\hat{x}, \hat{y}) &\sim f(x, y, \boldsymbol{\gamma})\\ \hat{c} &\sim f(c|\hat{x}, \hat{y}, \boldsymbol{\theta})\\ \hat{\sigma} &\sim f(\sigma|\hat{c}, \hat{x}, \hat{y}, \boldsymbol{\beta})\end{aligned}$$

where $\hat{x}$ and $\hat{y}$ are the simulated spatial coordinates, $\hat{c}$ is the simulated cell type, and $\hat{\sigma}$ is the simulated omics profile. The parameters $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are the model parameters for the spatial distribution, cell type distribution and omics profile distribution, respectively.

Markov Random Field model

To simulate the spatial distribution of cell phenotypes as well as other features like niches, SimSpace uses Markov Random Field models for the spatial modeling. A Markov Random Field is a probabilistic graphical model that represents the dependencies between random variables arranged on a lattice or grid, which has been widely used in image analysis and computer vision where spatial information is crucial [37, 59].

We model the spatial configuration of cell phenotypes as a Markov Random Field (MRF) defined over a two-dimensional grid, where each grid location corresponds to a cell position, and the cell phenotype at each position is treated as a random variable. Specifically, for each cell position $i \in \mathcal{L}$ in the given space $\mathcal{L}$, we define a random variable c_i that represents the phenotype of the cell at that position.

$$P(c_i = k) = P(\text{cell i is of type } k) \geq 0$$
$$\sum_{k=1}^{n} P(c_i = k) = 1$$

where k is the cell type and n is the total number of cell types. A configuration $\mathbf{c} = \{c_i\}_{i\in\mathcal{L}}$ is a complete assignment of cell types to all positions in $\mathcal{L}$. Then, for each cell position i, we define a neighborhood structure $\mathcal{N}_i$ that contains itself and the neighboring cells of cell i. The neighborhood structure functions as the clique in the undirected graphical model and can be defined by a distance threshold, where all cells within the threshold distance are considered neighbors. Studies have shown that such neighborhood definitions are effective in capturing local spatial relationships especially in spatial analysis [60, 61].

Given the definition, the MRF model holds the local Markov property, which states that a variable is conditionally independent of all other variables given its neighbors [62]:

$$P(c_i|c_{\mathcal{N}_i}, c_{\mathcal{L}\setminus\mathcal{N}_i}) = P(c_i|c_{\mathcal{N}_i}) \quad (2)$$

where $c_{\mathcal{N}_i}$ is the set of phenotype in the neighborhood of location i, and $c_{\mathcal{L}\setminus\mathcal{N}_i}$ is the set of phenotype outside the neighborhood of location i. This means that the phenotype of a cell only depends on its neighborhood structure including itself, which is a reasonable assumption in spatial omics data where local interactions are often more relevant than distant ones.

This conditional independence allows the positive joint distribution over phenotype configurations to be factorized over local cliques. In our setting, it results in the following factorization:

$$P(\mathbf{c}) = \frac{1}{Z}\prod_{i\in\mathcal{L}} \zeta(c_{\mathcal{N}_i}) \quad (3)$$

where $\zeta(c_{\mathcal{N}_i})$ is a local potential function defined over the labels in neighborhood $\mathcal{N}_i$ and Z is the partition function:

$$Z = \sum_{\mathbf{c}} \prod_{i\in\mathcal{L}} \zeta(c_{\mathcal{N}_i})$$

Given the positive condition and Markov property, the MRF can factorize and the potential function can

be expressed in an exponential form [36]:

$$\zeta(c_{\mathcal{N}_i}) = \exp(\phi \sum_{j \in \mathcal{N}_i} \theta_{c_i c_j}) \tag{4}$$

$$\begin{aligned} P(\mathbf{c}) &= \frac{1}{Z} \prod_{i \in \mathcal{L}} \zeta(c_{\mathcal{N}_i}) \\ &= \frac{1}{Z} \prod_{i \in \mathcal{L}} \exp(\phi \sum_{j \in \mathcal{N}_i} \theta_{c_i c_j}) \\ &= \frac{1}{Z} \exp(\phi \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{N}_i} \theta_{c_i c_j}) \end{aligned}$$

where ϕ here is a global smoothness parameter that controls the degree of interaction between neighboring phenotypes. The potential parameters $\theta_{c_i c_j}$ encodes the pairwise affinity between phenotypes c_i and c_j, reflecting biological preferences for spatial proximity.

Such $\theta_{c_i c_j}$ can be learned from the data or specified based on prior knowledge if available in a reference-based setting. Otherwise, it can be initialized randomly to simulate a prior distribution. The MRF model can then be used to simulate the spatial distribution of niches and cell types. Specifically, we use a Gibbs sampler to sample the cell types based on the potential function (4). The Gibbs sampling iteratively samples each cell type based on the conditional distribution given its neighborhood, which is defined by the MRF model. The sampling process can be summarized as follows:

1. Randomly assign initial phenotypes to all cell positions.
2. For n_{iter} iterations:
   (a) For each cell i:
      i. For each phenotype k, compute $\zeta_k = \exp\left(\phi \sum_{j \in \mathcal{N}_i} \theta_{k,c_j}\right)$.
      ii. Normalize ζ_k to obtain the conditional probability:
      $$P(c_i = k \mid c_{\mathcal{N}_i}) = \frac{\zeta_k}{\sum_{k'} \zeta_{k'}}$$
      iii. Sample a new phenotype c_i for cell i from the categorical distribution defined by $P(c_i = k \mid c_{\mathcal{N}_i})$.
3. Return the final phenotype assignments.

This iterative procedure ensures that the simulated spatial configuration reflects the local interactions encoded by the MRF potential parameters.

Reference-free simulation

In a reference-free scenario, the number of spatial niches m and the number of cell phenotypes k will be firstly defined, together with the neighborhood structure $\mathcal{N}$ and the spatial gird $\mathcal{L}$. The grid has a default setting of a 100×100 square grid, while it can also be customized to other shapes as long as each pixel represents a potential cell position. The neighborhood structure $\mathcal{N}$, customized by the maximal distance of interaction pairs, functions during the MRF process, where a larger neighborhood stands for a boarder cell-cell interaction range.

Niche simulation. To simulate niches across space, we applied a niche-level MRF model with the given niche potential parameters Θ_{niche} as the transition matrix. Specifically, given the number of spatial niches m, a symmetric matrix $\Theta_{m \times m} = \{\theta_{i,j} | i, j \in \{1, \ldots, m\}\}$ will be generated from a random generator. Then, a Gibbs sampler is performed to generate the niche distributions from the niche-level MRF model.

Cell type simulation. Within each niche, the phenotype interaction scores will be generated to simulate the different cell-cell interaction patterns in different niches. For each niche, a symmetric matrix $\Theta_{n \times n} = \{\theta_{i,j} | i, j \in \{1, \ldots, n\}\}$ will then be generated as well from the random generator. Another Gibbs sampler will then be performed to simulate cells' spatial distributions.

Cell density and coordinates simulation. Next, a cell density sampler will be applied to simulate the different cell densities for each phenotype. The k density parameters $\{d_1, d_2, \ldots, d_k\}$ for each phenotype play as the expected fraction of cells kept in the final simulation result. Finally, perturbations on the cells' coordinates are added to turn the gird-like into continuous spatial coordinates. The perturbation can be defined by a Gaussian distribution with a user-defined standard deviation σ_{perturb}, which will be added to the original coordinates of each cell.

Omics data simulation. Once the spatial patterns are simulated, the omics metadata will be simulated. Multiple omics simulation tools can be used to generate the omics data, like splatter [39] and other reference-free omics simulation model.

Meanwhile, SimSpace supports a mixed probabilistic model for omics data de-novo simulation: first, SimSpace defines G synthetic features (genes, proteins, etc.) with two classes, universal and marker features. A universal features u_g has its expected expression levels λ_{u_g} defined by a Gamms distribution with a global shape and scale parameter (α_u, β_u). Then for cell i, the actual expression levels $\sigma_{u_g,i}$ will

be sampled from a Poisson distribution:

$$\lambda_{u_g} \sim \text{Gamma}(\alpha_u, \beta_u), \quad \sigma_{u_g,i} \sim \text{Poisson}(\lambda_{u_g})$$

For a marker features m_g and its corresponding cell type c_g, the expected expression levels in the corresponding cell type λ_{m_g,c_g} are defined by a Gamma distribution with a shape and scale parameter (α_m, β_m). The actual expression levels $\sigma_{m_g,i}$ will then be sampled from a Poisson distribution:

$$\begin{aligned} \lambda_{m_g} &\sim \text{Gamma}(\alpha_u, \beta_u) \\ \lambda'_{m_g} &\sim \text{Gamma}(\alpha_m, \beta_m) \\ \sigma_{m_g,i} &\sim \begin{cases} \text{Poisson}(\lambda'_{m_g}), & \forall i : c_i = c_g \\ \text{Poisson}(\lambda_{m_g}), & \text{otherwise} \end{cases} \end{aligned}$$

The parameters $(\alpha_u, \beta_u, \alpha_m, \beta_m)$ can be defined by the user to control the expression levels of universal and marker features. The final omics data will then be generated by combining the universal and marker features for each cell.

To incorporate spatial information into the omics data, SimSpace can also simulate spatial interaction effects between simulated genes. Specifically, a given amount of genes g_{lr} will be selected as ligand-receptor pairs, and their expression levels will be enhanced by a spatial interaction term. The interaction term is defined as:

$$\begin{aligned} \sigma'_{g_l,i} &= \sigma_{g_l,i} \cdot \left(1 + \frac{\mu \cdot \bar{\sigma}_{g_r,\mathcal{N}_i}}{\bar{\sigma}_{g_r}}\right) \\ \sigma'_{g_r,i} &= \sigma_{g_r,i} \cdot \left(1 + \frac{\mu \cdot \bar{\sigma}_{g_r,\mathcal{N}_i}}{\bar{\sigma}_{g_l}}\right) \end{aligned}$$

where $\sigma_{g_l,i}$ is the expression level of ligand gene g_l in cell i, $\bar{\sigma}_{g_l,\mathcal{N}_i}$ is the average expression level of ligand gene g_l in the neighborhood of cell i, and $\bar{\sigma}_{g_l}$ is the average expression level of ligand gene g_l across all cells, and so for the receptor gene g_r. The parameter μ controls the strength of the global spatial interaction effect, which can be defined by the user. The final omics data will then be generated by combining the universal, marker features and the spatial interaction effects for each cell.

Spatial statistics calculation

To evaluate the spatial patterns of the simulated data, we calculate several spatial statistics, including Moran's Index and neighborhood entropy. These statistics are calculated based on the spatial coordinates and cell types of the provided dataset.

Moran's Index. Moran's Index is a measure of spatial autocorrelation, which quantifies the degree of similarity between neighboring cells. It is defined as:

$$I_{c_k} = \frac{n}{W} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(c_{i,k} - \bar{c_k})(c_{j,k} - \bar{c_k})}{\sum_{i=1}^{n}(c_{i,k} - \bar{c_k})^2}$$

where n is the number of cells, c_k is the cell type k, $c_{i,k}$ is the binary indicator of cell type k for cell i, w_{ij} is the spatial weight between cells i and j, and W is the sum of all w_{ij}. The spatial weight is defined by a KNN graph, where $w_{ij} = 1$ if cell i and cell j are neighbors, and $w_{ij} = 0$ otherwise.

Neighborhood Entropy. Neighborhood entropy is a measure of the diversity of cell types in the neighborhood of a cell. It is defined as:

$$H_i = -\sum_{k=1}^{K} p_{i,k} \log(p_{i,k})$$

where $p_{i,k}$ is the proportion of cell type k in the neighborhood of cell i, and K is the total number of cell types. The neighborhood is defined by a KNN graph, where the neighborhood of cell i includes its nearest neighbors. Densities at different neighborhood entropy threshold are calculated to represent the overall neighborhood entropy distribution, which can be marked as:

$$H_{q_j} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(q_{j-1} \leq H_i \leq q_j)$$

where H_i is the neighborhood entropy of cell i, q_j is the j-th threshold of the neighborhood entropy distribution. The neighborhood entropy distribution can be used to evaluate the diversity of cell types in the spatial patterns at sample scale.

Reference-based simulation

In a reference-based framework, spatial patterns are fitted by minimizing the contrastive loss of Moran's Index and neighborhood entropies between the original reference and fitted simulation. The optimization process uses the genetic algorithm, which iteratively updates the parameters of the niche and cell-type MRF model as well as cell densities to minimize the loss function. The optimization process can be summarized as follows:

$$\begin{aligned} \hat{\theta} &\leftarrow \text{argmin}_\theta \mathcal{L}(\theta) \\ \mathcal{L}(\theta) &= \mathcal{L}_{\text{MI}}(\hat{c}, c_{\text{ref}}) + \mathcal{L}_{\text{NE}}(\hat{c}, c_{\text{ref}}) \\ \mathcal{L}_{\text{MI}}(\hat{c}, c_{\text{ref}}) &= \sum_{k=1}^{K} |I_{c_k}(\hat{c}) - I_{c_k}(c_{\text{ref}})| \\ \mathcal{L}_{\text{NE}}(\hat{c}, c_{\text{ref}}) &= \sum_{j=1}^{Q} |H_{q_j}(\hat{c}) - H_{q_j}(c_{\text{ref}})| \end{aligned}$$

where $\hat{c}$ is the simulated cell type distribution, c_{ref} is the reference cell type distribution, and θ is the set of parameters to be optimized. The optimization process will then iteratively update the parameters until convergence or until a maximum number of iterations is reached.

Spatial pattern simulation. Once the parameter $\hat{\theta}$ is optimized, the spatial patterns can be simulated like reference-free settings using the MRF model with the optimized parameters. The niche-level MRF model will be used to simulate the spatial distribution of niches, and the cell-level MRF model will be used to simulate the spatial distribution of cell types within each niche. The cell densities will also be simulated based on the optimized parameters.

Omics data simulation. With spatial data generated, the omics profile will be simulated through the specified reference-based single-cell simulation tool. The simulated omics data will then be combined with the spatial coordinates and cell types to generate the final spatial omics data.

Three-dimensional data simulation

To simulate three-dimensional spatial omics data, we extend the two-dimensional MRF model to a three-dimensional MRF model. The neighborhood structure $\mathcal{N}$ is defined in three dimensions, and the corresponding potential functions ζ are defined in three dimensions as well. Specifically, the potential function is still in the same form as equation 4, where now the neighborhood $\mathcal{N}_i$ now includes all adjacent voxels in three-dimensional space. The simulation process, including Gibbs sampling and parameterization, remains unchanged except for the extension to 3D coordinates (x, y, z). This allows SimSpace to generate realistic three-dimensional spatial omics datasets with controllable spatial structure and cell-cell interactions.

Spatial clustering and visualization

SpatialPCA and BANKSY are used to perform the spatial clustering of simulated and real spatial omics data. We first concatenate the spatial coordinates and the omics data of SimSpace, scCube, and Xenium reference data into a single dataset, and then perform clustering methods on the concatenated matrix. For spatialPCA, we selected top 100 highly variable genes to perform PCA while keeping other parameters as default. The first two principal components are then used for the plotting. For BANKSY, we tested different λ setting, 0.2, 0.5, and 0.8, and visualized the results using UMAP plots.

Additionally, to isolate the spatial effects from the omics data, we replaced the simulated omics data of SimSpace and scCube with the Xenium reference data. Specifically, for each cell in SimSpace and scCube, we replaced the simulated omics data with the corresponding Xenium reference data from a randomly selected cell that has the same cell type. The replaced omics data is then used for the spatial clustering and visualization.

Cell type deconvolution methods

We benchmarked five different cell type deconvolution methods using SimSpace's simulated datasetes, including cell2location, RCTD, CARD, Seurat, and spatialDWLS. Each method was applied following their provided tutorials and default parameters: for cell2location, we trained the model with 3000 epoches for the reference data, with a batch size of 256. Next, in the spatial mapping, we kept the default parameters *N_cells_per_location* as 30 and *detection_alpha* as 20. The deconvolution training took 5000 epoches; for RCTD, we kept every spot and gene by setting *CELL_MIN_INSTANCE* as 1. Also, both 'multi' and 'full' mode of RCTD were performed seperately; for CARD, we set the *minCountGene* and *minCountSpot* as 1; for Seurat, *SCTransform* was used to normalize the reference and spot data, and the first 30 principal components were used for the anchor transfering; for spatialDWLS, we computed the cell-type mean expression matrix from the Xenium reference data, which was then used to deconvolute the dataset with *n_cell* as 20.

The performance of each method is evaluated by comparing the predicted cell type proportions with the ground truth cell type proportions in the simulated data. The evaluation metrics include Pearson correlation coefficient (PCC), rooted mean square error (RMSE), and Jaccard index which is defined as:

$$\begin{aligned}
\text{PCC}(A, B) &= \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \bar{B})^2}} \\
\text{RMSE}(A, B) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - B_i)^2} \\
\text{Jaccard}(A, B) &= \frac{|A \cap B|}{|A \cup B|} \\
&= \frac{\sum_{i=1}^{n}\mathbb{I}(A_i)\mathbb{I}(B_i)}{\sum_{i=1}^{n}\left[\mathbb{I}(A_i) + \mathbb{I}(B_i) - \mathbb{I}(A_i)\mathbb{I}(B_i)\right]}
\end{aligned}$$

where A and B denote the ground truth and predicted cell type proportions, respectively, and n is the number of cell types. $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition is met (in this case, the cell type is present and $A_i > 0$) and 0 otherwise.

Spatial variable gene detection methods

We benchmarked six different spatial variable gene detection methods using SimSpace's simulated datasetes, including Hotspot, SpatialDE, scBSP, SPARK, Giotto, and Moran's I implemented in the Seurat package. While following the recommended pipelines for each method, we used the default parameters as suggested in their respective documentation. Specifically, for SPARK we set the *percentage* as 0 to keep all the genes while screening for SVGs.

To evaluate the performance of each method, we compared the detected SVGs with the ground truth SVGs in the simulated data. The evaluation metrics include Pearson correlation coefficient (PCC), Kendall's tau, and F1 score. The PCC is calculated between the rank-normalized p-values of the detected SVGs and the ground truth SVGs.

Kendall's tau is used to measure the concordance of the rankings of the detected SVGs and the ground truth SVGs, which is defined as:

$$\tau = \frac{(C - D)}{\frac{1}{2}n(n-1)}$$

where C is the number of concordant pairs, D is the number of discordant pairs, and n is the number of genes.

F1 score is used to evaluate the binary classification performance of the SVG detection methods, where a gene is considered an SVG if its p-value is below a threshold of 0.05. The F1 score is defined as:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision is the proportion of true positive SVGs among the detected SVGs, and Recall is the proportion of true positive SVGs among the ground truth SVGs.

Kendall's coefficient of concordance (W) is a non-parametric statistic used to assess the degree of agreement among multiple raters or judges when ranking a set of items. Kendall's W was employed to quantify the consistency of rankings assigned to genes across different reference-free simulation datasets. The coefficient is calculated by first summing the ranks assigned to each gene by all raters, then applying the formula:

$$W = \frac{12S}{m^2(n^3 - n)}$$

where m is the number of raters (in this case, the number of simulation datasets), n is the number of genes, and S is the sum of squared differences between the ranks assigned to each gene and the average rank across all raters:

$$S = \sum_{j=1}^{n} \left(R_j - \frac{m(n+1)}{2} \right)^2$$

where R_j is the sum of ranks assigned to gene j by all raters. The value of S is computed by taking the sum of the squared differences between each gene's rank and the average rank across all genes, multiplied by the number of raters. The coefficient W is then normalized to fall within the range of 0 to 1, where 0 indicates no agreement among raters and 1 indicates perfect agreement.

Data collection and preprocessing

Xenium human breast tumor dataset. The dataset was downloaded from 10X Genomics website: https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast. The cell type annotation can be found from its original publication.

CODEX human endometrial tumor dataset. The CODEX human endometrial tumor dataset used in this study is part of an ongoing project and will be made publicly available upon publication. The data contains 31 protein markers and 11.1 million cells with 13 cell types, which are annotated based on the original study.

Code availability

The SimSpace package is available at https://github.com/TianxiaoNYU/simspace as a Python package, where users can find the source code, documentation, and examples for using the package. The code for the analysis presented in this study will be made publicly available upon publication.

References

[1] Z. K. Tuong, K. W. Loudon, B. Berry, N. Richoz, J. Jones, X. Tan, Q. Nguyen, A. George, S. Hori, S. Field, A. G. Lynch, K. Kania, P. Coupland, A. Babbage, R. Grenfell, T. Barrett, A. Y. Warren, V. Gnanapragasam, C. Massie, and M. R. Clatworthy. Resolving the immune landscape of human prostate at a single-cell level in health and cancer. *Cell Rep*, 37(12):110132, 2021.

[2] L. A. Walsh and D. F. Quail. Decoding the tumor microenvironment with spatial technologies. *Nat Immunol*, 24(12):1982–1993, 2023.

[3] Z. Yao, C. T. J. van Velthoven, M. Kunst, M. Zhang, D. McMillen, C. Lee, W. Jung, J. Goldy, A. Abdelhak, M. Aitken, K. Baker, P. Baker, E. Barkan, D. Bertagnolli, A. Bhandiwad, C. Bielstein, P. Bishwakarma, J. Campos, D. Carey, T. Casper, A. B. Chakka, R. Chakrabarty, S. Chavan, M. Chen, M. Clark, J. Close, K. Crichton, S. Daniel, P. DiValentin, T. Dolbeare, L. Ellingwood, E. Fiabane, T. Fliss, J. Gee, J. Gerstenberger, A. Glandon, J. Gloe, J. Gould, J. Gray, N. Guilford, J. Guzman, D. Hirschstein, W. Ho, M. Hooper, M. Huang, M. Hupp, K. Jin, M. Kroll, K. Lathia, A. Leon, S. Li, B. Long, Z. Madigan, J. Malloy, J. Malone, Z. Maltzer, N. Martin, R. McCue, R. McGinty, N. Mei, J. Melchor, E. Meyerdierks, T. Mollenkopf, S. Moonsman, T. N. Nguyen, S. Otto, T. Pham, C. Rimorin, A. Ruiz, R. Sanchez, L. Sawyer, N. Shapovalova, N. Shepard, C. Slaughterbeck, J. Sulc, M. Tieu, A. Torkelson, H. Tung, N. Valera Cuevas, S. Vance, K. Wadhwani, K. Ward, B. Levi, C. Farrell, R. Young, B. Staats, M. M. Wang, C. L. Thompson, S. Mufti, C. M. Pagan, L. Kruse, N. Dee, S. M. Sunkin, L. Esposito, M. J. Hawrylycz, J. Waters, L. Ng, K. Smith, B. Tasic, X. Zhuang, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature*, 624(7991):317–332, 2023.

[4] A. Lomakin, J. Svedlund, C. Strell, M. Gataric, A. Shmatko, G. Rukhovich, J. S. Park, Y. S. Ju, S. Dentro, V. Kleshchevnikov, V. Vaskivskyi, T. Li, O. A. Bayraktar, S. Pinder, A. L. Richardson, S. Santagata, P. J. Campbell, H. Russnes, M. Gerstung, M. Nilsson, and L. R. Yates. Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature*, 611(7936):594–602, 2022.

[5] L. Moses and L. Pachter. Museum of spatial transcriptomics. *Nat Methods*, 19(5):534–546, 2022.

[6] L. Liu, A. Chen, Y. Li, J. Mulder, H. Heyn, and X. Xu. Spatiotemporal omics for biology and medicine. *Cell*, 187(17):4488–4519, 2024.

[7] V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, R. Elmentaite, A. Lomakin, V. Kedlian, A. Gayoso, M. S. Jain, J. S. Park, L. Ramona, E. Tuck, A. Arutyunyan, R. Vento-Tormo, M. Gerstung, L. James, O. Stegle, and O. A. Bayraktar. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol*, 40(5):661–671, 2022.

[8] D. M. Cable, E. Murray, L. S. Zou, A. Goeva, E. Z. Macosko, F. Chen, and R. A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol*, 40(4):517–526, 2022.

[9] Y. Ma and X. Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat Biotechnol*, 40(9):1349–1359, 2022.

[10] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, 3rd Mauck, W. M., Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902 e21, 2019.

[11] R. Dong and G. C. Yuan. Spatialdwls: accurate deconvolution of spatial transcriptomic data. *Genome Biol*, 22(1):145, 2021.

[12] D. DeTomaso and N. Yosef. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Syst*, 12(5):446–456 e9, 2021.

[13] V. Svensson, S. A. Teichmann, and O. Stegle. Spatialde: identification of spatially variable genes. *Nat Methods*, 15(5):343–346, 2018.

[14] J. Wang, J. Li, S. T. Kramer, L. Su, Y. Chang, C. Xu, M. T. Eadon, K. Kiryluk, Q. Ma, and D. Xu. Dimension-agnostic and granularity-based spatially variable gene identification using bsp. *Nat Commun*, 14(1):7367, 2023.

[15] Jinpu Li, Mauminah Raina, Yiqing Wang, Chunhui Xu, Li Su, Qi Guo, Ricardo Melo Ferreira, Michael T Eadon, Qin Ma, Juexin Wang, and Dong Xu. scbsp: A fast and accurate tool for identifying spatially variable features from high-resolution spatial omics data. *bioRxiv*, 2025.

[16] S. Sun, J. Zhu, and X. Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat Methods*, 17(2):193–200, 2020.

[17] R. Dries, Q. Zhu, R. Dong, C. L. Eng, H. Li, K. Liu, Y. Fu, T. Zhao, A. Sarkar, F. Bao, R. E. George, N. Pierson, L. Cai, and G. C. Yuan. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol*, 22(1):78, 2021.

[18] N. Moriel, E. Senel, N. Friedman, N. Rajewsky, N. Karaiskos, and M. Nitzan. Novosparc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nat Protoc*, 16(9):4177–4200, 2021.

[19] Manik Kuchroo, Danielle F. Miyagishima, Holly R. Steach, Abhinav Godavarthi, Yutaka Takeo, Phan Q. Duy, Tanyeri Barak, E. Zeynep Erson-Omay, Scott Youlten, Ketu Mishra-Gorur, Jennifer Moliterno, Declan McGuone, Murat Günel, and Smita Krishnaswamy. sparc recovers human glioma spatial signaling networks with graph filtering. *bioRxiv*, page 2022.08.24.505139, 2022.

[20] J. Hu, X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara, and M. Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods*, 18(11):1342–1351, 2021.

[21] V. Singhal, N. Chou, J. Lee, Y. Yue, J. Liu, W. K. Chock, L. Lin, Y. C. Chang, E. M. L. Teo, J. Aow, H. K. Lee, K. H. Chen, and S. Prabhakar. Banksy unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nat Genet*, 56(3):431–441, 2024.

[22] L. Shang and X. Zhou. Spatially aware dimension reduction for spatial transcriptomics. *Nat Commun*, 13(1):7203, 2022.

[23] T. Vu, S. Seal, T. Ghosh, M. Ahmadian, J. Wrobel, and D. Ghosh. Funspace: A functional and spatial analytic approach to cell imaging data using entropy measures. *PLoS Comput Biol*, 19(9):e1011490, 2023.

[24] A. Sampath Kumar, L. Tian, A. Bolondi, A. A. Hernandez, R. Stickels, H. Kretzmer, E. Murray, L. Wittler, M. Walther, G. Barakat, L. Haut, Y. Elkabetz, E. Z. Macosko, L. Guignard, F. Chen, and A. Meissner. Spatiotemporal transcriptomic maps of whole mouse embryos at the onset of organogenesis. *Nat Genet*, 55(7):1176–1185, 2023.

[25] G. Yan, S. H. Hua, and J. J. Li. Categorization of 34 computational methods to detect spatially variable genes from spatially resolved transcriptomics data. *Nat Commun*, 16(1):1141, 2025.

[26] W. Wang, S. Zheng, S. C. Shin, J. C. Chavez-Fuentes, and G. C. Yuan. Ontrac characterizes spatially continuous variations of tissue microenvironment through niche trajectory analysis. *Genome Biol*, 26(1):117, 2025.

[27] F. W. Townes and B. E. Engelhardt. Nonnegative spatial factorization applied to spatial genomics. *Nat Methods*, 20(2):229–238, 2023.

[28] L. Shang, P. Wu, and X. Zhou. Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics. *Nat Commun*, 16(1):1059, 2025.

[29] D. Song, Q. Wang, G. Yan, T. Liu, T. Sun, and J. J. Li. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nat Biotechnol*, 42(2):247–252, 2024.

[30] J. Zhu, L. Shang, and X. Zhou. Srtsim: spatial pattern preserving simulations for spatially resolved transcriptomics. *Genome Biol*, 24(1):39, 2023.

[31] J. Qian, H. Bao, X. Shao, Y. Fang, J. Liao, Z. Chen, C. Li, W. Guo, Y. Hu, A. Li, Y. Yao, X. Fan, and Y. Cheng. Simulating multiple variability in spatially resolved transcriptomics with sccube. *Nat Commun*, 15(1):5021, 2024.

[32] H. Li, Z. Zhang, M. Squires, X. Chen, and X. Zhang. scmultisim: simulation of single-cell multi-omics and spatial data guided by gene regulatory networks and cell-cell interactions. *Nat Methods*, 2025.

[33] D. Haviv, J. Remsik, M. Gatie, C. Snopkowski, M. Takizawa, N. Pereira, J. Bashkin, S. Jovanovich, T. Nawy, R. Chaligne, A. Boire, A. K. Hadjantonakis, and D. Pe'er. The covariance environment defines cellular niches for spatial inference. *Nat Biotechnol*, 43(2):269–280, 2025.

[34] A. Janesick, R. Shelansky, A. D. Gottscho, F. Wagner, S. R. Williams, M. Rouault, G. Beliakoff, C. A. Morrison, M. F. Oliveira, J. T. Sicherman, A. Kohlway, J. Abousoud, T. Y. Drennon, S. H. Mohabbat, Teams x Development, and S. E. B. Taylor. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun*, 14(1):8353, 2023.

[35] G. Gut, M. D. Herrmann, and L. Pelkmans. Multiplexed protein maps link subcellular organization to cellular states. *Science*, 361(6401), 2018.

[36] P Clifford and JM Hammersley. Markov fields on finite graphs and lattices, 1971.

[37] G. R. Cross and A. K. Jain. Markov random field texture models. *IEEE Trans Pattern Anal Mach Intell*, 5(1):25–39, 1983.

[38] Victor Freguglia and Nancy Lopes Garcia. Inference tools for markov random fields on lattices: The r package mrf2d. *Journal of Statistical Software*, 101(8):1 – 36, 2022.

[39] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome Biol*, 18(1):174, 2017.

[40] Z. Li, T. Wang, P. Liu, and Y. Huang. Spatialdm for rapid identification of spatially co-expressed ligand-receptor and revealing cell-cell communication patterns. *Nat Commun*, 14(1):3995, 2023.

[41] Z. Cang, Y. Zhao, A. A. Almet, A. Stabell, R. Ramos, M. V. Plikus, S. X. Atwood, and Q. Nie. Screening cell-cell communication in spatial transcriptomics via collective optimal transport. *Nat Methods*, 20(2):218–228, 2023.

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

[43] B. F. Miller, D. Bambah-Mukku, C. Dulac, X. Zhuang, and J. Fan. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. *Genome Res*, 31(10):1843–1855, 2021.

[44] H. Xu, H. Fu, Y. Long, K. S. Ang, R. Sethi, K. Chong, M. Li, R. Uddamvathanak, H. K. Lee, J. Ling, A. Chen, L. Shao, L. Liu, and J. Chen. Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Med*, 16(1):12, 2024.

[45] P. L. Stahl, F. Salmen, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, A. Borg, F. Ponten, P. I. Costea, P. Sahlen, J. Mulder, O. Bergmann, J. Lundeberg, and J. Frisen. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[46] S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[47] L. Yan and X. Sun. Benchmarking and integration of methods for deconvoluting spatial transcriptomic data. *Bioinformatics*, 39(1), 2023.

[48] H. Li, J. Zhou, Z. Li, S. Chen, X. Liao, B. Zhang, R. Zhang, Y. Wang, S. Sun, and X. Gao. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nat Commun*, 14(1):1548, 2023.

[49] G. Yan, S. H. Hua, and J. J. Li. Categorization of 34 computational methods to detect spatially variable genes from spatially resolved transcriptomics data. *Nat Commun*, 16(1):1141, 2025.

[50] C. Chen, H. J. Kim, and P. Yang. Evaluating spatially variable gene detection methods for spatial transcriptomics data. *Genome Biol*, 25(1):18, 2024.

[51] Y. Hao, S. Hao, E. Andersen-Nissen, 3rd Mauck, W. M., S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, and R. Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587 e29, 2021.

[52] D. Philtron, Y. Lyu, Q. Li, and D. Ghosh. Maximum rank reproducibility: A nonparametric approach to assessing reproducibility in replicate experiments. *J Am Stat Assoc*, 113(523):1028–1039, 2018.

[53] E. Lundberg and G. H. H. Borner. Spatial proteomics: a powerful discovery tool for cell biology. *Nat Rev Mol Cell Biol*, 20(5):285–302, 2019.

[54] Y. Goltsev, N. Samusik, J. Kennedy-Darling, S. Bhate, M. Hale, G. Vazquez, S. Black, and G. P. Nolan. Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981 e15, 2018.

[55] Y. Liu, A. Sinjab, J. Min, G. Han, F. Paradiso, Y. Zhang, R. Wang, G. Pei, Y. Dai, Y. Liu, K. S. Cho, E. Dai, A. Basi, J. K. Burks, K. I. Rajapakshe, Y. Chu, J. Jiang, D. Zhang, X. Yan, P. A. Guerrero, A. Serrano, M. Li, T. H. Hwang, A. Futreal, J. A. Ajani, L. M. Solis Soto, A. A. Jazaeri, H. Kadara, A. Maitra, and L. Wang. Conserved spatial subtypes and cellular neighborhoods of cancer-associated fibroblasts revealed by single-cell spatial multi-omics. *Cancer Cell*, 43(5):905–924 e6, 2025.

[56] C. C. Hughes. Endothelial-stromal interactions in angiogenesis. *Curr Opin Hematol*, 15(3):204–9, 2008.

[57] M. V. Patel, Z. Shen, M. Rodriguez-Garcia, E. J. Usherwood, L. J. Tafe, and C. R. Wira. Endometrial cancer suppresses cd8+ t cell-mediated cytotoxicity in postmenopausal women. *Front Immunol*, 12:657326, 2021.

[58] H. Guan, Q. Xiong, J. Xiong, Y. Liu, and W. Zhang. Cd8+ t cell activation in endometrial cancer: prognostic implications and potential for personalized therapy. *Front Immunol*, 16:1542669, 2025.

[59] S. Krishnamachari and R. Chellappa. Multiresolution gauss-markov random field models for texture segmentation. *IEEE Trans Image Process*, 6(2):251–67, 1997.

[60] Q. Zhu, S. Shah, R. Dries, L. Cai, and G. C. Yuan. Identification of spatially associated subpopulations by combining scrnaseq and sequential fluorescence in situ hybridization data. *Nat Biotechnol*, 2018.

[61] Z. Wu, A. E. Trevino, E. Wu, K. Swanson, H. J. Kim, H. B. D'Angio, R. Preska, G. W. Charville, P. D. Dalerba, A. M. Egloff, R. Uppaluri, U. Duvvuri, A. T. Mayer, and J. Zou. Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nat Biomed Eng*, 6(12):1435–1448, 2022.

[62] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

Figures

Figure 1

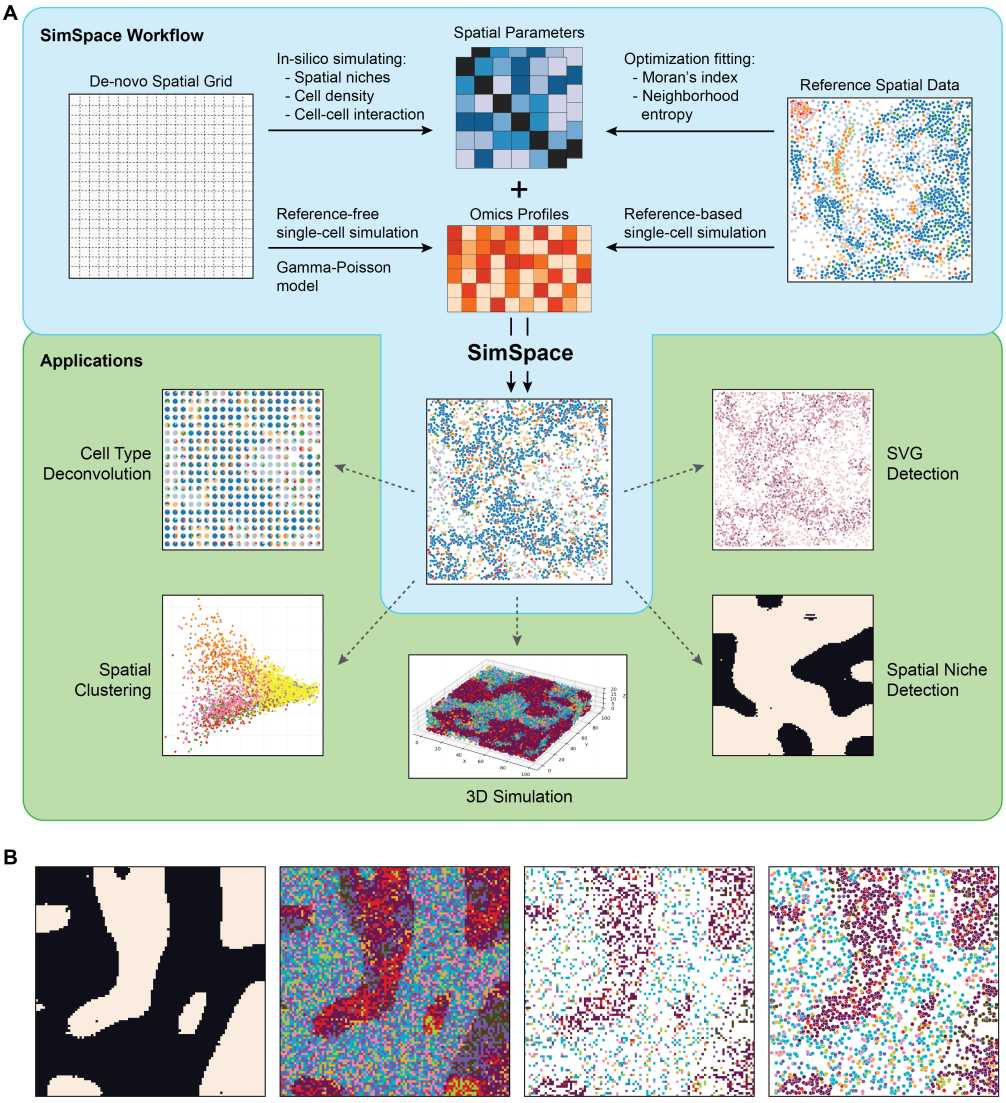

Figure 1: Overview of the SimSpace framework and its downstream applications. (A) The SimSpace framework consists of two main components: the spatial pattern simulator and the omics profile simulator. SimSpace supports both reference-free and reference-based simulation modes. A de novo spatial grid is first initialized, upon which spatial parameters (e.g., spatial niches, cell density, and cell-cell interaction) are simulated. These parameters are optionally fit to reference spatial datasets by optimizing spatial autocorrelation metrics such as Moran's index and neighborhood entropy. Simulated omics profiles are generated from the single-cell simulation tools and integrated with spatial patterns to produce a spatial omics dataset. The resulting datasets enable various downstream applications, including spatial clustering, cell type deconvolution, spatially variable gene detection, spatial niche identification, and 3D spatial simulations. (B) A SimSpace simulation example workflow. First, the niches are simulated using a niche-level Markov Random Field model, which generates spatial niches with different interaction patterns. Then, the cell types are simulated within each niche using a cell-level simulation model, which captures the interactions between neighboring cell types. Next, cell densities are simulated, and a perturbation is applied to the cell coordinates to generate a continuous spatial distribution eventually.

Figure 2

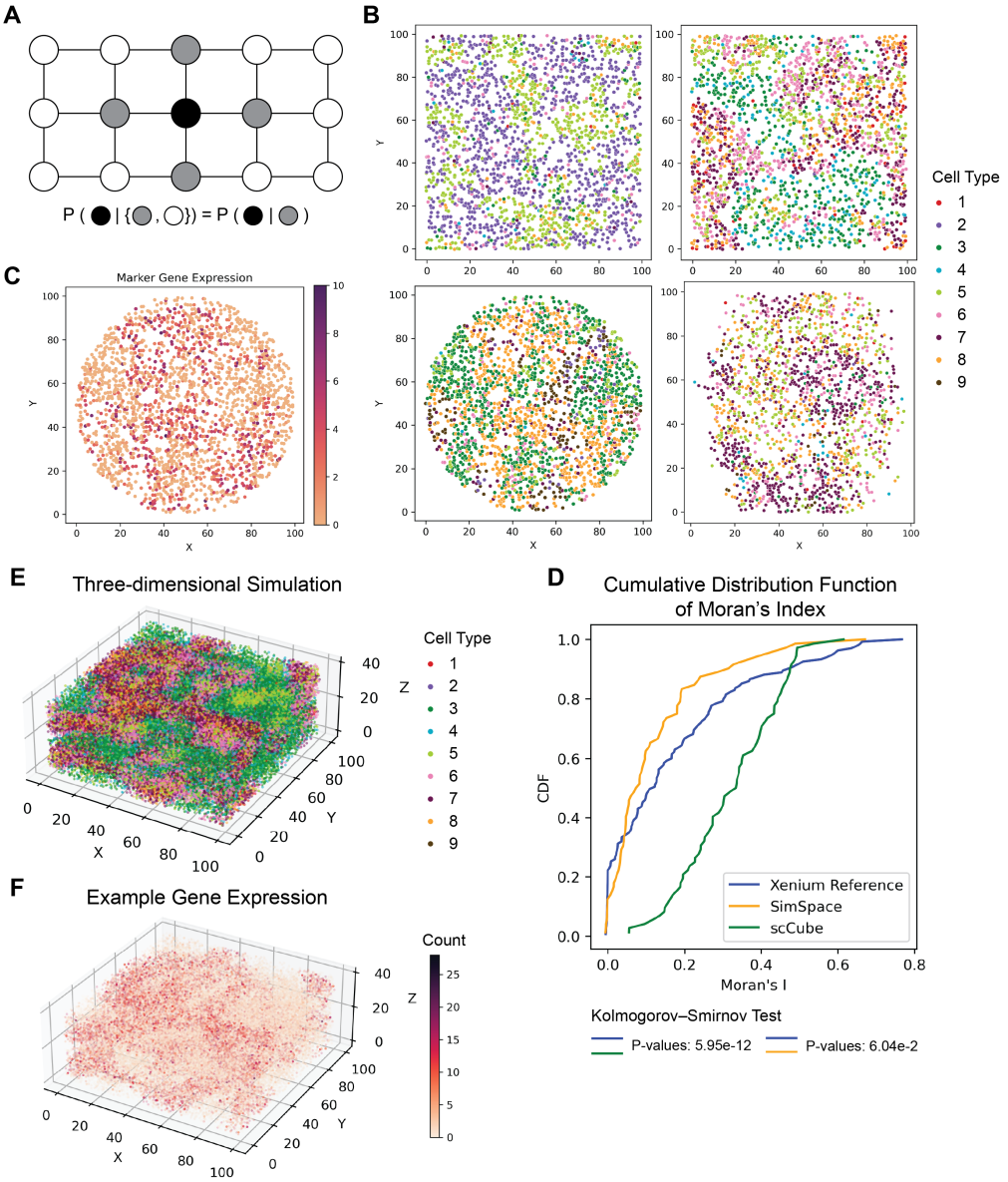

Figure 2: SimSpace enables the simulation of biologically realistic spatial patterns and omics profiles in a reference-free setting. (A) Schematic illustrating the use of a Markov Random Field (MRF) model in SimSpace to generate spatial patterns by enforcing local neighborhood dependencies, where the probability of a cell type assignment is conditioned on its neighbors. (B) Representative examples of simulated two-dimensional spatial maps generated under varying spatial parameters, demonstrating the framework's ability to produce diverse and tunable cell type arrangements; distinct cell types are indicated by different colors. (C) Simulated gene expression values for a representative marker gene display spatial gradients and local variability consistent with biological observations. (D) Comparative analysis of spatial patterns in real and simulated data using the cumulative distribution function (CDF) of Moran's I scores; SimSpace simulations more closely recapitulate the spatial autocorrelation observed in the Xenium reference dataset compared to scCube, as indicated by the Kolmogorov-Smirnov test. (E) Example of a three-dimensional spatial simulation generated by SimSpace, showing spatial niches across Z-slices annotated by cell type. (F) Simulated gene expression in three-dimensional space reveals spatial heterogeneity that mirrors the complexity of real tissue architecture.

Figure 3

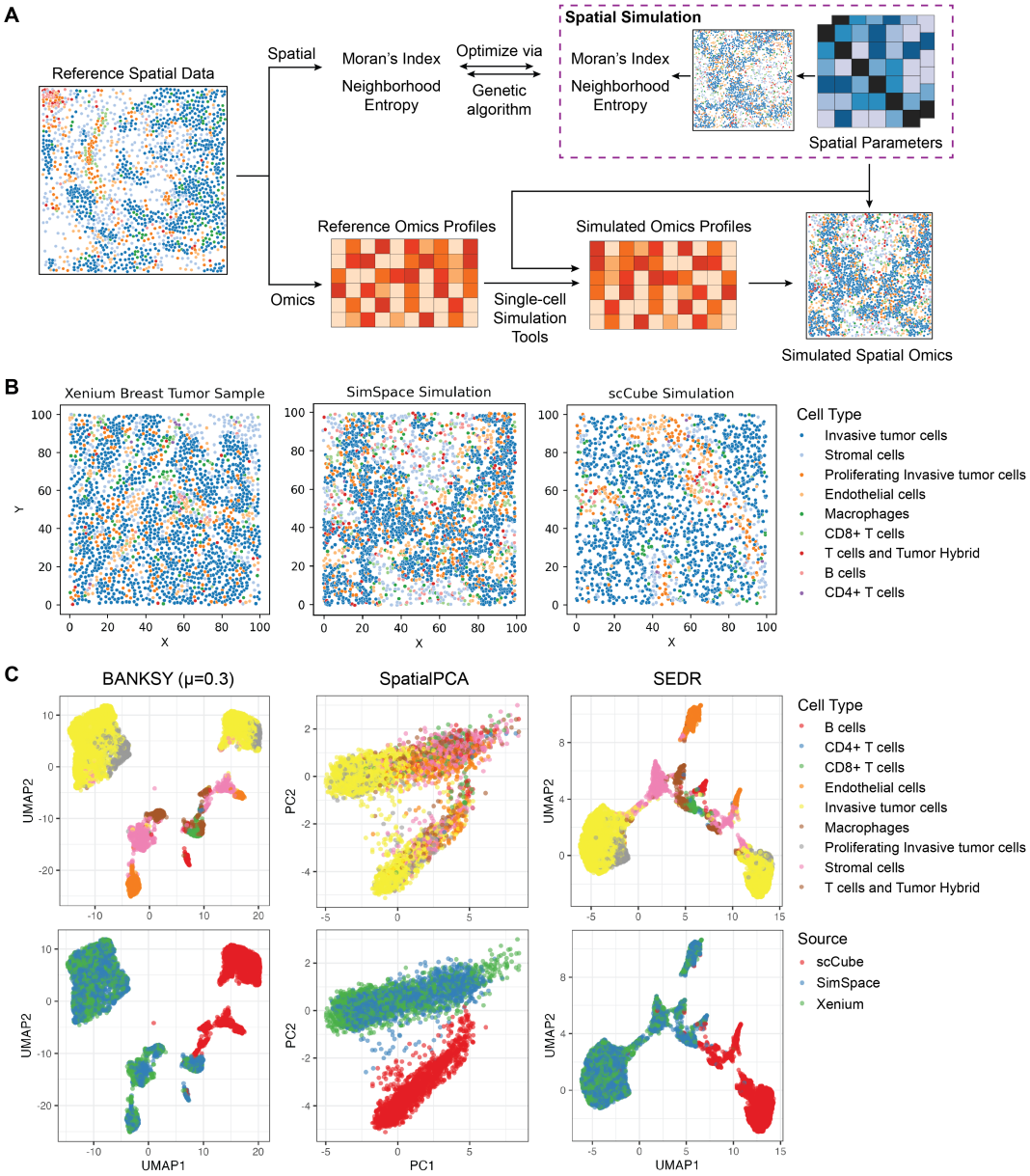

Figure 3: SimSpace enables reference-based spatial simulation and generates realistic, unbiased spatial omics data. (A) Reference-based simulation workflow. Given a real spatial transcriptomics dataset, SimSpace estimates spatial parameters (Moran's index, neighborhood entropy) from the reference spatial coordinates and fits them via a genetic algorithm. These fitted spatial features are then combined with omics profiles—simulated from the reference—to generate realistic spatial omics datasets. (B) Cell type maps from real and simulated data. Reference tile from the Xenium breast tumor sample (left) is compared with simulations from SimSpace (middle) and scCube (right). SimSpace recapitulates key spatial organization features of the reference while introducing variability, unlike scCube which shows spatial patterns inconsistent with the reference. (C) Evaluation using spatial clustering methods. Three tools (BANKSY, SpatialPCA, and SEDR) are applied to data from SimSpace, Xenium, and scCube. The top row shows clustering by cell type; the bottom row indicates data source. SimSpace simulations yield spatial embeddings that resemble those of the Xenium reference, yet do not cluster exclusively by source, suggesting that SimSpace simulations are both realistic and free of bias to the reference dataset.

Figure 4

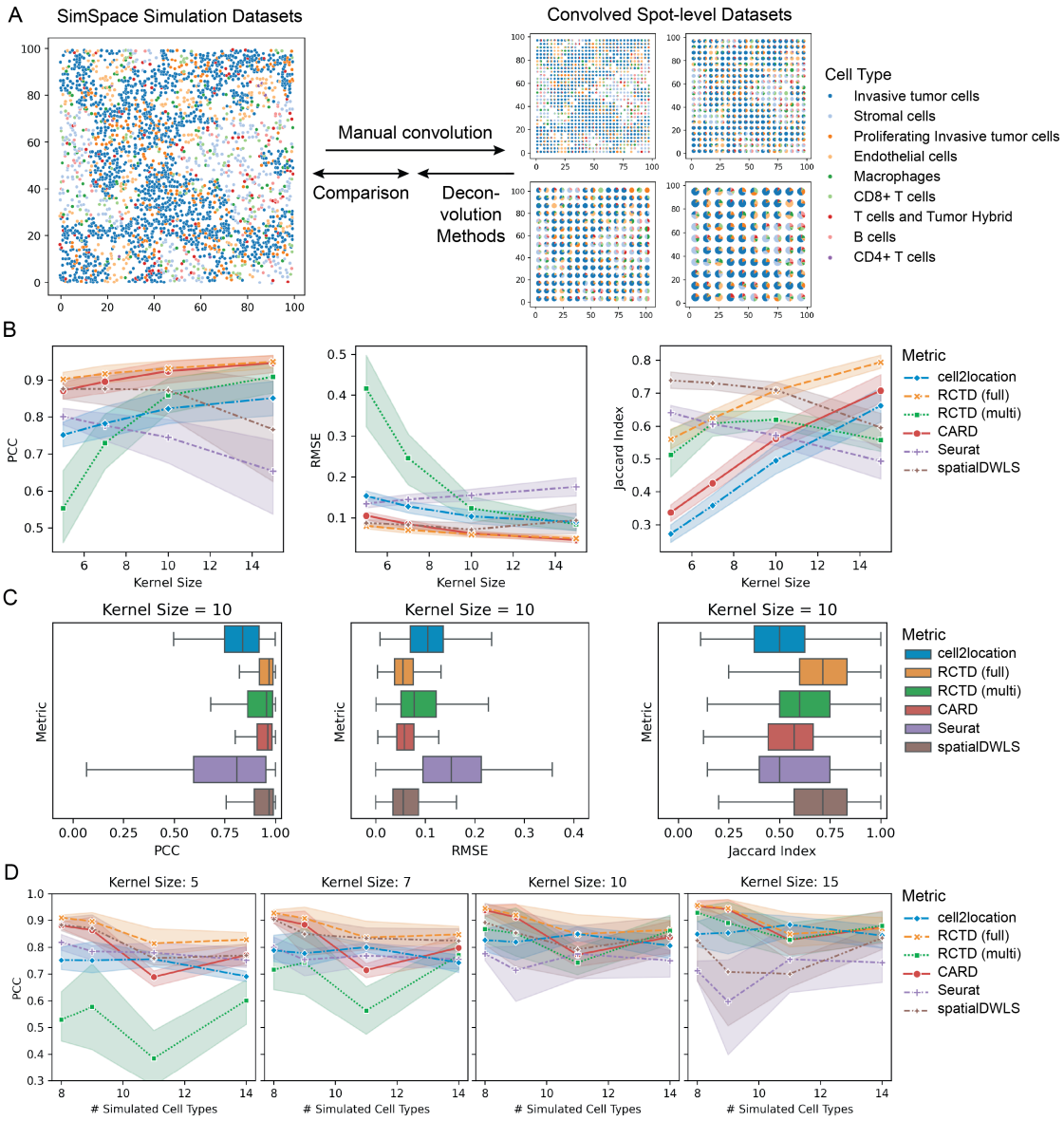

Figure 4: Benchmarking spatial cell type deconvolution methods using SimSpace simulations. (A) Benchmarking framework. Simulated single-cell resolution data are aggregated into spot-level measurements via manual convolution with varying kernel sizes to mimic real spatial transcriptomics. These convolved datasets serve as ground truth for evaluating deconvolution methods. (B) Performance across deconvolution methods evaluated using three metrics—Pearson correlation coefficient (PCC), root mean squared error (RMSE), and Jaccard index—across kernel sizes (spot resolutions). SimSpace simulations reveal performance trends across methods, including strong performance from RCTD and CARD. (C) Comparison at fixed kernel size (10). Boxplots summarize PCC, RMSE, and Jaccard index across all methods. CARD and cell2location achieve high accuracy and consistency across metrics. (D) Robustness across varying cell type complexity. Performance of each method is shown as a function of the number of simulated cell types under four different kernel sizes. RCTD full mode, CARD, and cell2location demonstrate robustness across increasing complexity, while performance for other methods varies more widely.

Figure 5

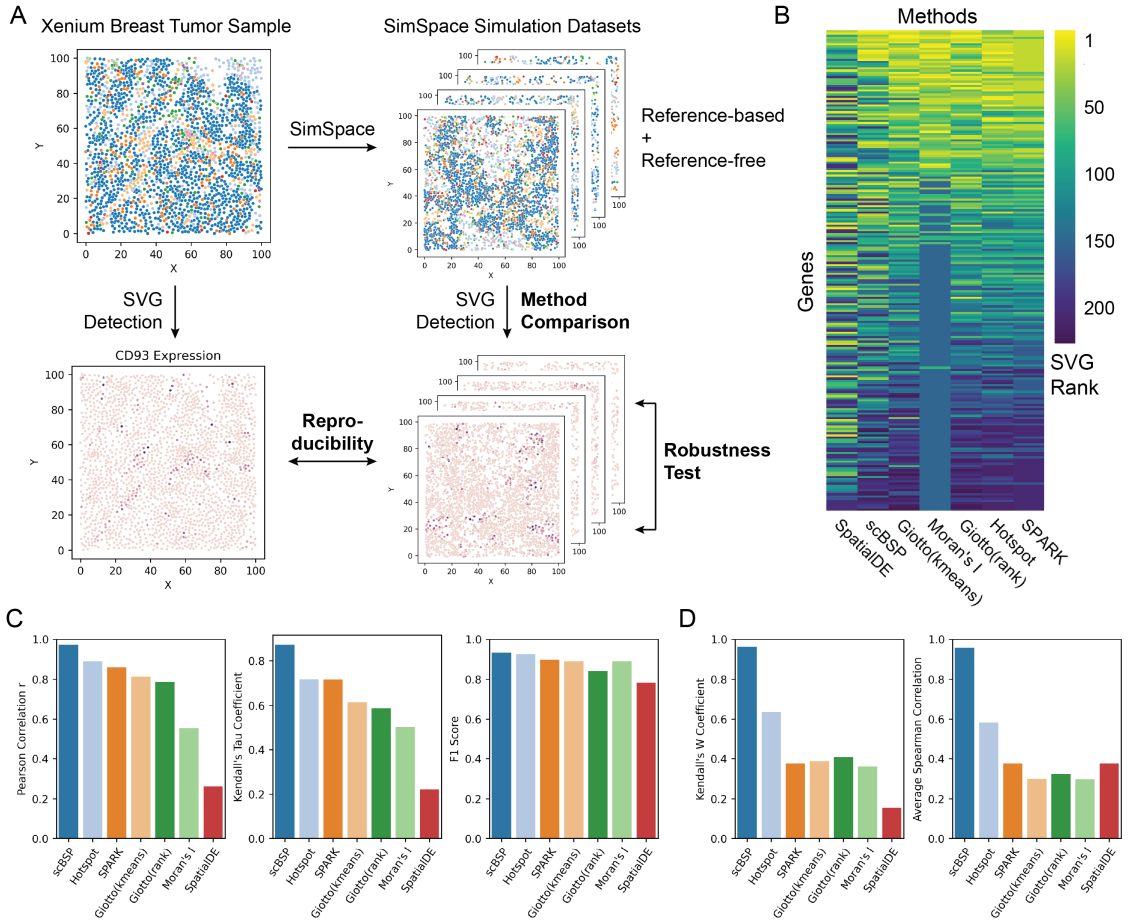

Figure 5: Benchmarking spatially variable gene (SVG) detection using SimSpace simulations. (A) Benchmarking framework. A Xenium human breast tumor sample is used to generate SimSpace datasets under both reference-based and reference-free simulation modes. SVG detection is performed on the original Xenium dataset and on simulated datasets to compare methods and assess reproducibility and robustness. (B) Comparison of SVG rankings across methods. Rank-normalized p-values from a reference-based SimSpace simulation are visualized across seven methods: SPARK, Hotspot, Giotto (rank), Moran's I, Giotto (k-means), scBSP, and SpatialDE. Although top-ranked genes show general overlap, several methods like scBSP and SpatialDE shows discrepancies in gene prioritization. (C) Reproducibility between simulated and reference data. Across methods, Pearson correlation (left), Kendall's Tau (middle), and F1 score (right) are computed between the Xenium reference and reference-based SimSpace simulation. Most methods demonstrate high reproducibility, with scBSP achieving the strongest concordance across all metrics. (D) Robustness across reference-free simulations. Consistency of SVG rankings across eight reference-free SimSpace datasets is quantified using Kendall's W (left) and average Spearman correlation (right). scBSP and Hotspot show the highest robustness to variations in spatial structure, while other methods display more variability.

Figure 6

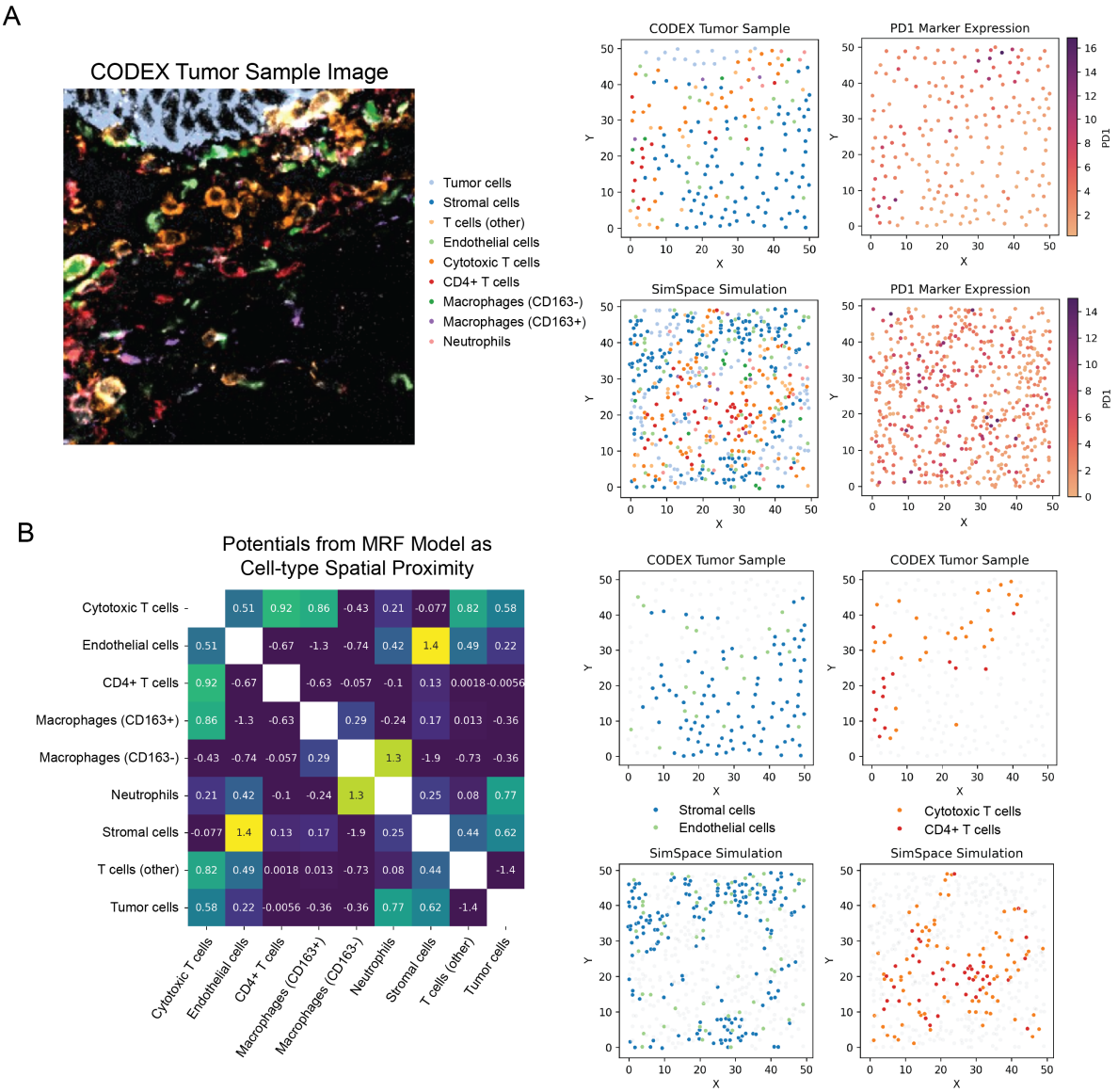

Figure 6: SimSpace models and simulates spatial proteomics data to reveal microenvironmental cell-cell interactions. (A) Reference-based simulation of spatial proteomics. A representative tile from a CODEX endometrial tumor sample (left) is used to generate a SimSpace simulation incorporating both spatial cell distributions and protein marker expression. PD1 expression is shown for both real and simulated datasets, demonstrating the preservation of spatial expression patterns across cell types. (B) Quantifying spatial cell-cell interactions. Using SimSpace's underlying MRF model, an interaction matrix is inferred to represent pairwise spatial co-localization strengths between cell types. Strong spatial attraction is observed between cytotoxic and CD4+ T cells, and between endothelial and stromal cells, while tumor cells show spatial exclusion from most immune populations. Spatial distributions of selected cell types (endothelial and stromal cells, cytotoxic and CD4+ T cells) are shown in both the CODEX reference and SimSpace simulation, demonstrating realistic tissue architecture reconstruction. These results demonstrate SimSpace's ability not only to simulate spatial proteomics data, but also to serve as a tool for interpreting spatial interaction patterns in complex tissues.

Supplementary Information

Supplementary Figure 1

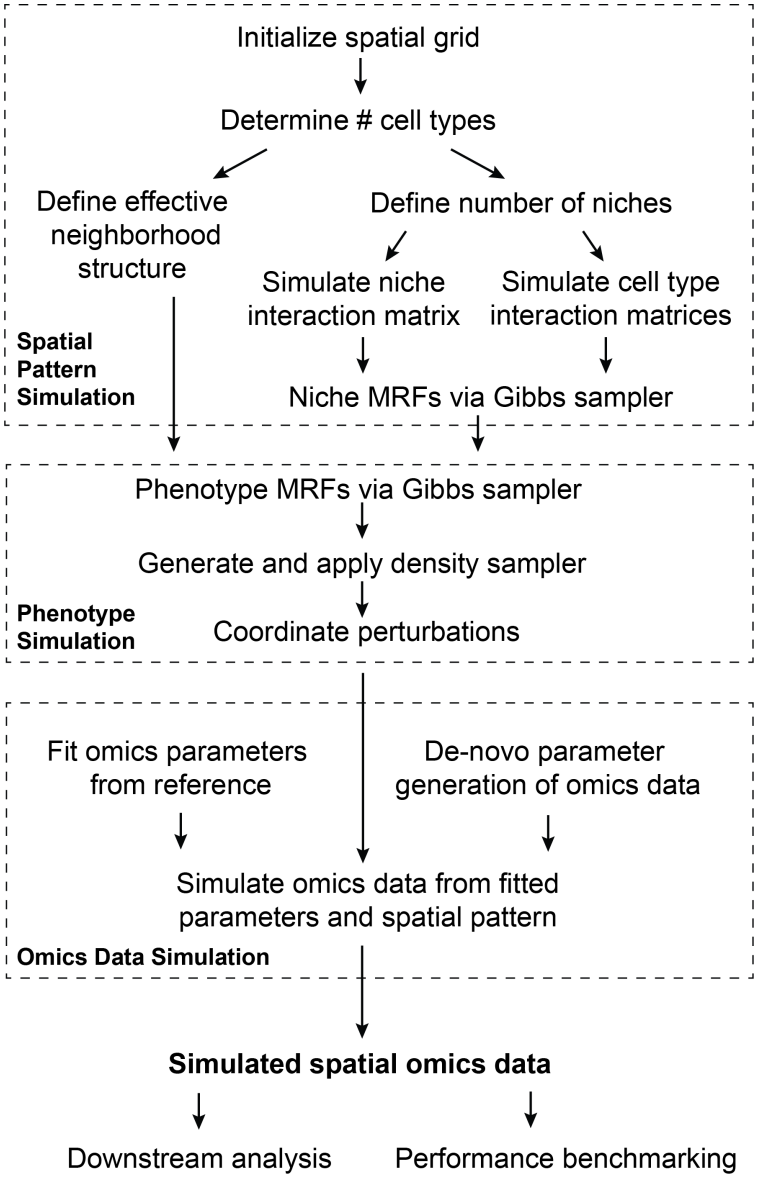

Figure S1: Overview of the SimSpace generative framework. Flowchart summarizing the three main stages of SimSpace simulation: (1) Spatial Pattern Simulation: A spatial grid is initialized, followed by user-defined settings or fitted parameters from reference data for the number of cell types, neighborhood structure, and niche/cell-type interactions. Markov Random Fields (MRFs) are simulated via Gibbs sampling to generate structured spatial distributions. (2) Phenotype Simulation: Phenotype-level MRFs are constructed, followed by density sampling and coordinate perturbation to emulate realistic cellular density and variability. (3) Omics Data Simulation: Omics parameters are either fitted from a reference dataset or generated de novo. Simulated spatial patterns are paired with expression profiles to produce final spatial omics datasets.

Supplementary Figure 2

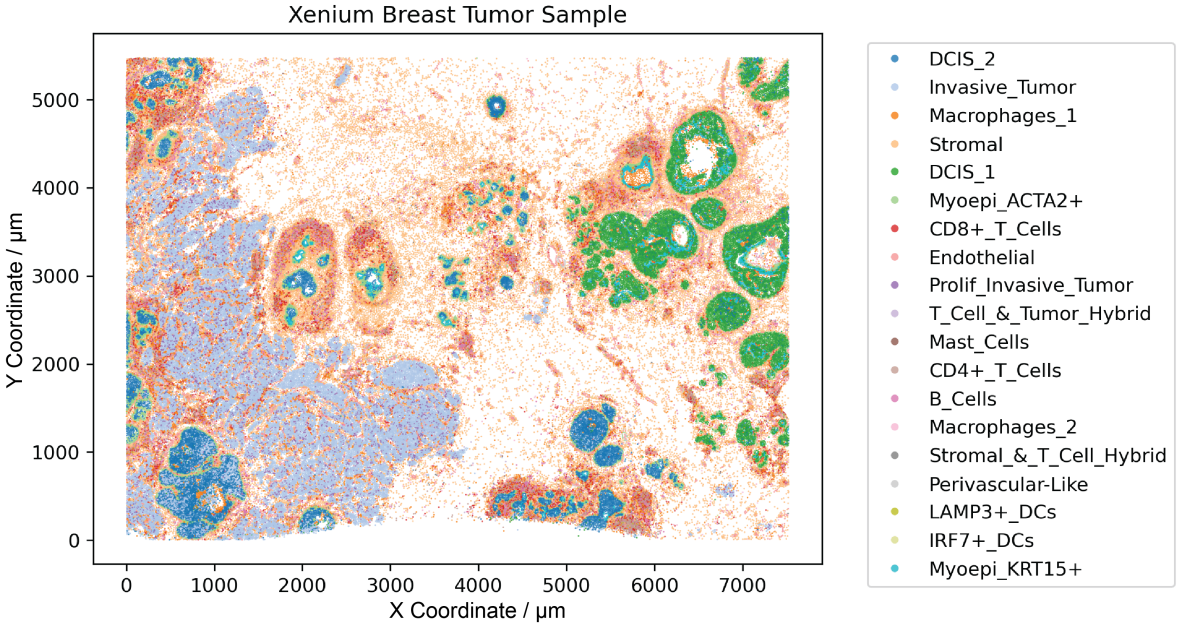

Figure S2: Cell-type annotation of the Xenium breast tumor sample. Visualization of the full-resolution Xenium breast tumor sample colored by cell type labels, highlighting the complex spatial organization of tumor, stromal, and immune populations. Cell annotations are based on the original Xenium dataset publication.

Supplementary Figure 3

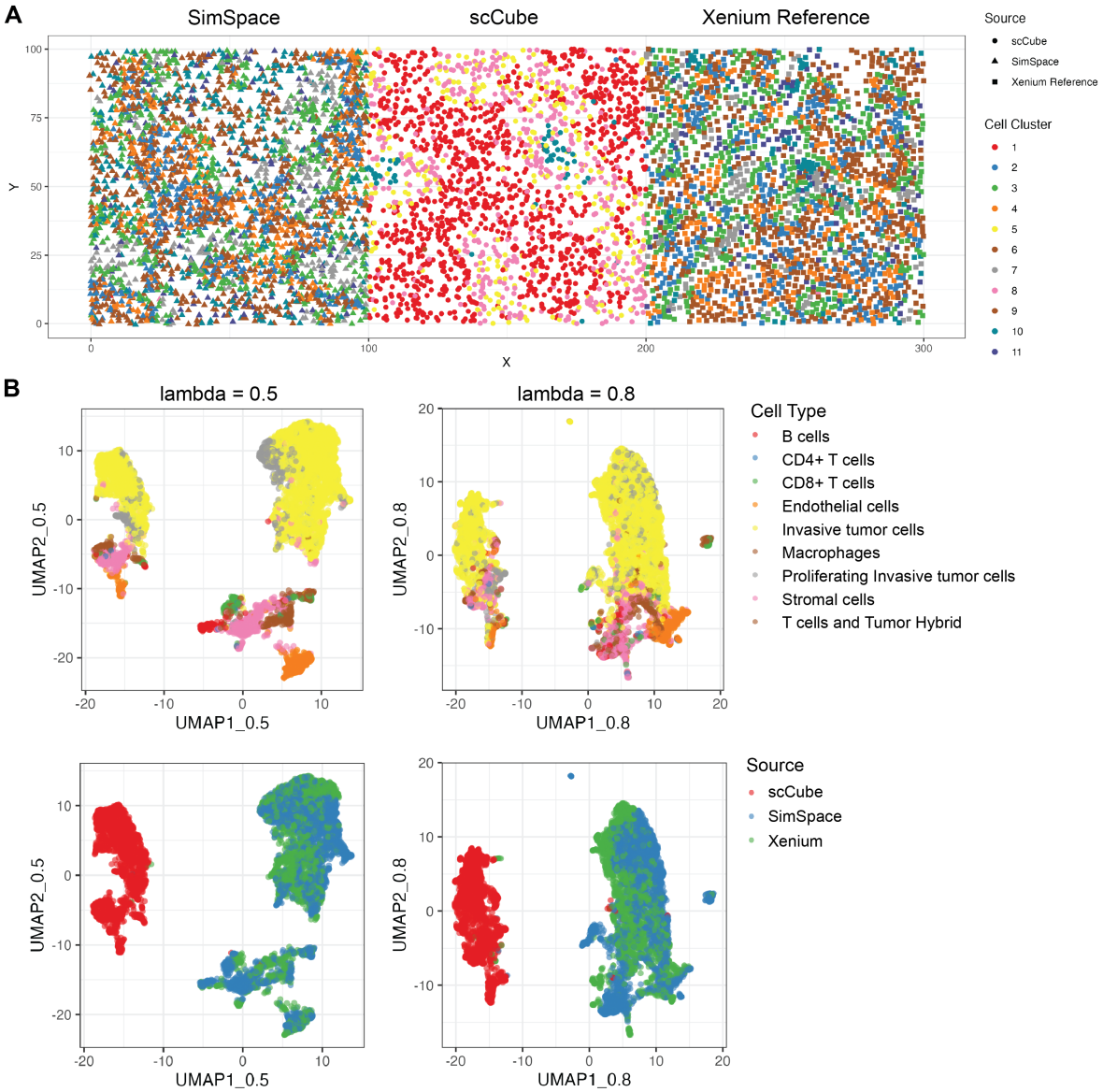

Figure S3: Comparison of spatial and latent structure across simulated and real datasets. (A) Spatial visualization of cell clusters. Cell cluster assignments are shown for SimSpace, scCube, and Xenium reference datasets. SimSpace recapitulates diverse spatial patterns and heterogeneity similar to the Xenium reference, while scCube shows limited spatial variability and over-clustered regions. (B) BANKSY UMAP embeddings colored by cell type (top) and data source (bottom). Embeddings were generated using two a mixing parameter ($\lambda = 0.5$ and $\lambda = 0.8$) that controls the importance of cells' own expression and neighborhood expression effects. SimSpace and Xenium cells intermingle in latent space, indicating high similarity in expression and structure. In contrast, scCube cells form separate clusters, suggesting batch-specific artifacts or limited biological realism. These results support the fidelity of SimSpace simulations in capturing both spatial and transcriptomic structure.

Supplementary Figure 4

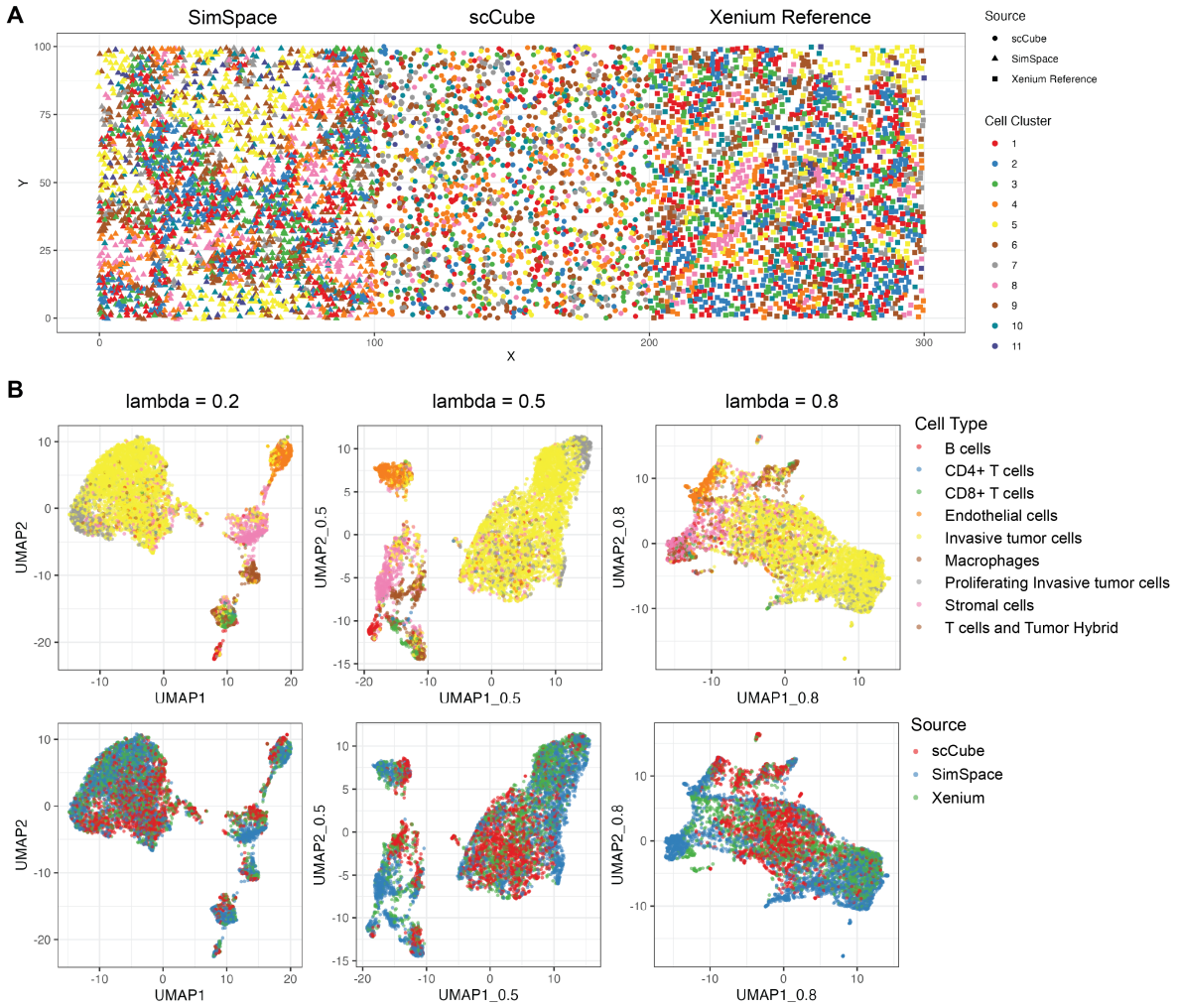

Figure S4: Testing the impact of spatial patterns on batch effects in spatial clustering. (A) Spatial distributions of cell clusters across SimSpace, scCube, and Xenium datasets. In this analysis, the original simulated omics profiles were replaced with expression profiles from the Xenium reference dataset, randomly matched by cell type. (B) BANKSY UMAP embeddings with shared expression profiles across datasets. Embeddings were computed under different mixing parameter ($\lambda = 0.3, 0.5, 0.8$). Top: colored by cell type; Bottom: colored by dataset source. Despite identical omics inputs, scCube cells remain separated from SimSpace and Xenium across all λ values, confirming that the observed batch effect arises primarily from differences in spatial patterns rather than gene expression. In contrast, SimSpace retains strong alignment with the Xenium reference, highlighting its ability to generate biologically realistic spatial architectures.

Supplementary Figure 5

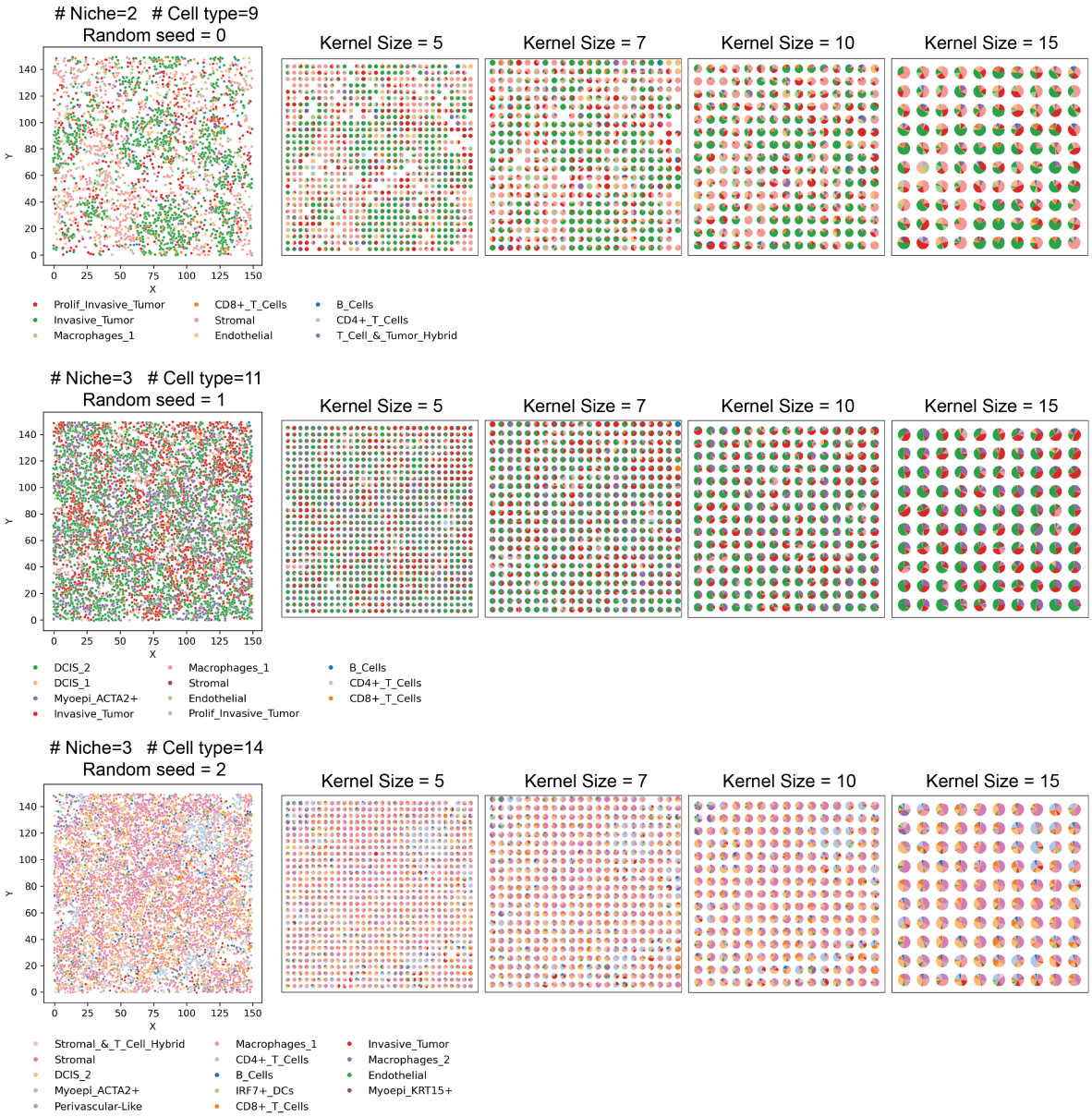

Figure S5: Example SimSpace-generated datasets and their convolved spot-level representations. Three representative simulations derived from Xenium breast tumor tiles are shown, each with different numbers of cell types, spatial niches, and random seeds. For each simulation (rows), the leftmost panel shows the single-cell spatial layout, while the subsequent panels display spot-level data generated by applying manual convolution with kernel sizes of 5, 7, 10, and 15. These examples illustrate how spatial resolution and cellular composition vary across simulated datasets.

Supplementary Figure 6

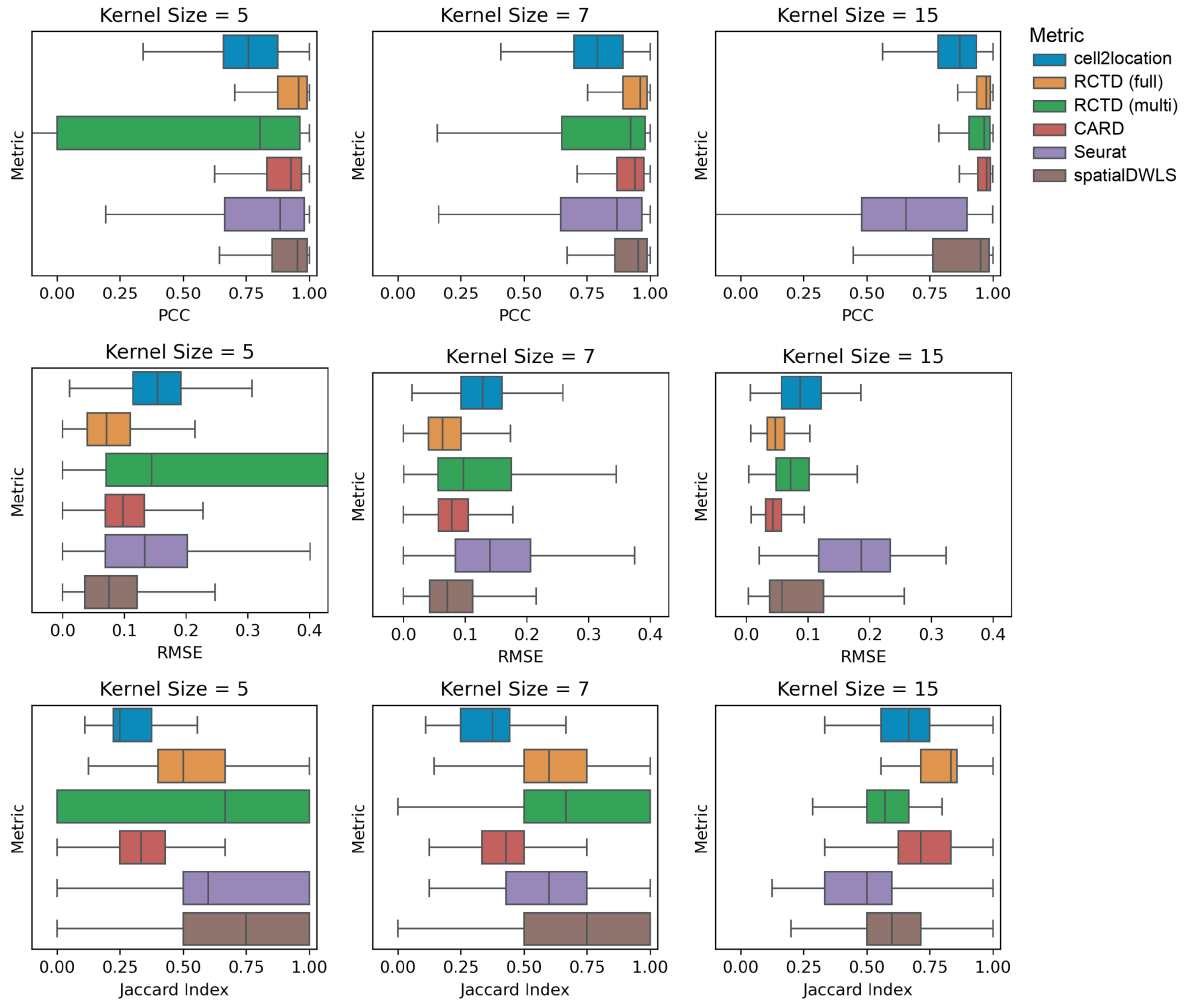

Figure S6: Performance of cell type deconvolution methods across different kernel sizes. Boxplots show benchmarking results for six deconvolution methods (cell2location, RCTD [full and multi-mode], CARD, Seurat, and spatialDWLS) evaluated on SimSpace-generated datasets. Performance is assessed using Pearson correlation coefficient (PCC, top row), root mean squared error (RMSE, middle row), and Jaccard index (bottom row) across three kernel sizes: 5, 7, and 15. Each method's distribution reflects variation across multiple simulated datasets.

Supplementary Figure 7

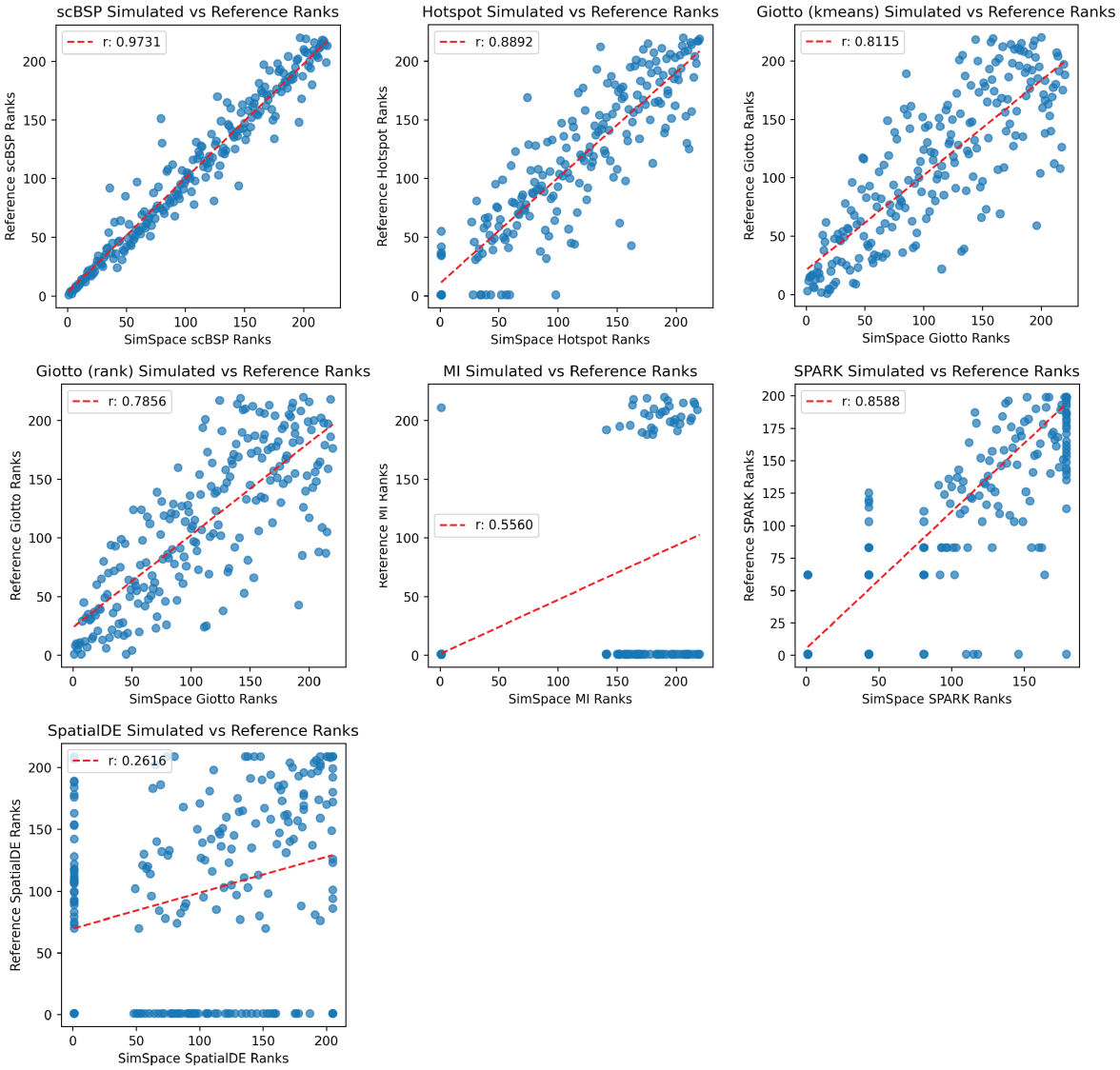

Figure S7: Comparison of SVG rankings between SimSpace simulations and the Xenium reference dataset. Scatter plots compare gene ranks derived from SimSpace reference-based simulations versus the original Xenium reference across seven SVG detection methods: scBSP, Hotspot, Giotto (kmeans), Giotto (rank), Moran's I, SPARK, and SpatialDE. Each point represents one gene; dotted red lines indicate linear regression fit. Pearson correlation coefficients (r) are shown for each method. scBSP and Hotspot achieve the highest concordance with the reference, while SpatialDE and Moran's I exhibit lower agreement, suggesting greater sensitivity to spatial variation.