Spatially aligned random partition models on spatially resolved transcriptomics data

Yunshan Duan[1], Shuai Guo[2], Hao Yan[2], Wenyi Wang[2], and Peter Mueller[1, *]

[1]Department of Statistics and Data Science, University of Texas at Austin

[2]The University of Texas MD Anderson Cancer Center

[*]Corresponding author: Peter Mueller, pmueller@math.utexas.edu

Abstract

We propose spatially aligned random partition (SARP) models for clustering multiple types of experimental units, incorporating dependence in a subvector of the cluster-specific parameters, e.g., a subvector of spatial information, as in the motivating application. The approach is developed for inference about co-localization of immune, stromal, and tumor cell sub-populations. The aim is to understand the recruitment of immune and stromal cell subtypes by tumor cells, formalized as spatial dependence of the corresponding homogeneous cell subpopulations. This is achieved by constructing Bayesian nonparametric random partition models for the different types of cells, with a hierarchically structured prior introducing the desired dependence. Specifically, we use Pitman-Yor priors and add dependence in the base measure for spatial features, while leaving the base measure corresponding to gene expression features a priori independent across different types of cells. Details of the model construction are designed to lead to a convenient MCMC algorithm for posterior inference. Simulation studies show favorable performance in identifying co-localization between types of cells. We apply the proposed approach with colorectal cancer (CRC) data and discover subtypes of immune and stromal cells that are spatially aligned with specific tumor regions.

1 Introduction

Current literature on Bayesian inference for random partitions includes extensive work on joint models for multiple random partitions allowing for dependence. However, there is no principled approach for inference about specific clusters in a partition of one type of experimental units relating to clusters in a random partition of another type of experimental units. We require such inference for a motivating application to spatial transcriptomics data, to answer questions about specific subtypes of tumor cells recruiting specific immune cell subtypes. To address

this question, we develop a Bayesian nonparametric (BNP) inference approach for spatially aligned random partitions across multiple sets of experimental units. That is, the proposed inference implements model-based clustering on separate sets of units, introducing dependence in a subvector of the cluster specific parameters corresponding to spatial location (or any other shared feature sub-space – for simplicity, without loss of generality, we will throughout assume spatial dependence). In the motivating application, the experimental units are immune, stromal, and tumor cells from tumor tissue, with the inference question being how subtypes of these cells co-localize with one another. We integrate spatial transcriptomics (ST) and single-cell RNA sequencing (scRNA-seq; SC) data to obtain both gene expression profiles and spatial coordinates for individual cells. Motivated by the observation that immune and stromal cells often cluster around specific tumor regions, we aim to characterize the spatial co-localization patterns among subpopulations of these three cell types. In our framework, the spatial alignment of cell subtypes is formalized as dependence in random partitions of the spatial subvector, while gene expression remains independent. This investigation aims to advance our understanding of the tumor microenvironment and the immunological interactions that influence tumor progression.

We implement the desired inference under a Bayesian paradigm. Bayesian inference on random partitions is often implemented by way of setting up a mixture model $F(y) = \sum_h^\infty w_h f(y \mid \theta_h)$. For statistical inference it is convenient to write $F(y)$ as a mixture of the kernel $f(y \mid \theta)$ with respect to a discrete mixing measure $G(\theta) = \sum_h w_h \delta_{\theta_h}$. Here δ_θ is a point mass at θ. Under a Bayesian approach the model construction is completed with a prior probability model for the random mixing measure G. Prior models for random probability measures are known as non-parametric Bayesian (BNP) models [1]. The most widely used BNP prior, specifically for model-based clustering is the Dirichlet process (DP) prior [1, Chapter 4]. However, the law of the random partition induced by the DP, known as the Chinese restaurant process (CRP), is controlled by only a single parameter [2]. Several alternative models generalize DP mixture models to allow for more flexibility in clustering, using, for example, the Pitman-Yor process (PYP) [3], and the normalized inverse Gaussian as special cases [4].

Cancer tissues consist of diverse populations of epithelial, stromal, and immune cells, each encompassing multiple distinct cell types and characteristics. Gaining insights into the specific roles of these cellular components and their interactions is essential for furthering our understanding of cancer biology. To better understand the interaction of different cell types, we set up a random partition of tumor, immune, and stromal cells, allowing for correlation of different subtypes. Specifically, we represent cell populations composed of homogeneous cell subtypes by way of mixtures of kernels with respect to mixing measures $G_j,\ j = 1, 2, 3$, for tumor, stromal and immune cells, respectively. In this context, we represent the recruitment of immune and stromal cell subtypes by tumor cells as spatial dependence. In the mixture

model this becomes spatial dependence of the locations of (some) immune and stromal cell subpopulations with the locations of tumor cell subpopulations. The desired dependence is formalized as a dependent prior $p(G_1, G_2, G_3)$ across the mixtures. Many dependent priors for a family $\mathcal{G} = \{G_1, \ldots, G_J\}$ of random probability measures have been studied in the BNP literature over the past 25 years. Some of the earliest constructions introduce the desired dependence by way of shared hyperparameters [5, 6]. Another early construction is based on additive decomposition models using a mixture of common versus idiosyncratic parts based on DPs [7]. Variations to ensure DP marginal distribution are proposed in Kolossiatis et al. [8], and extensions to completely random measures (CRM) [9] are introduced in Lijoi et al. [10]. Other types of widely used dependent models include the hierarchical DP (HDP) [11], the nested DP (NDP) [12] and their variations. Another widely used construction for dependent random probability measures is introduced in the seminal paper by MacEachern [13] as the dependent DP (DDP) model for problems when the random probability measures are indexed by covariates as $\mathcal{G} = \{G_x;\ x \in X\}$. A discussion on the properties of the various dependent priors constructions has been included in [14].

However, the motivating application in the upcoming discussion requires more structure than what is naturally accommodated in these approaches. For start, there is an asymmetry in the desired inference. Also dependence is restricted to a spatial subvector of linked subpopulations of the immune, stromal, and tumor cells, but gene expression profiles should remain a priori independent. None of the earlier mentioned constructions favors such structure as we need it here for a meaningful biological interpretation. This motivates us to propose spatially aligned random partition models that introduce the co-localization of clusters with the desired structure. The proposed framework approaches the problem as an inference problem with dependent partitions, building on the rich literature of dependent random partition models that we briefly reviewed before. However, other approaches are possible. One alternative and very natural approach is to build on an equally rich literature of point processes. This direction is explored, for example, in Osher et al. [15].

We propose spatially aligned random partition (SARP) models to identify clusters of multiple cell types within the same spatial space. The desired structure for this application includes that (1) aligned clusters should exhibit similar but not identical spatial locations; (2) the dependence structure should allow for one type of cells (e.g., tumor cells) to serve as reference with clusters of other types of cells (e.g., immune cells and stromal cells) to be located around reference cell clusters; and (3) alignment should allow multiple clusters of the other types to be spatially aligned with one subtype of the reference type; (4) the alignment is introduced only for a subvector of the features – in our case, the spatial coordinates. The desired structure is achieved by adding dependence in the base measure of Pitman-Yor priors. The construction allows for a natural and efficient posterior simulation algorithm. We apply the proposed SARP

model for inference with colorectal cancer (CRC) data.

The rest of the paper is organized as follows. Section 2 describes the motivating application of studying cell interaction in the tumor micro-environment and the data source. We describe the proposed pipeline and data pre-processing. In Section 3.1 we introduce the proposed SARP model, first for the case of two sets of experimental units. Section 3.2 discusses marginal and asymptotic properties of the model. Section 3.3 extends the model to more than two sets of units, and Section 3.4 introduces the posterior inference algorithm. In Section 4, we summarize simulation studies to validate the proposed inference approach. We report on a sensitivity analysis to assess the impact of prior choices. In Section 5, we apply the proposed method for inference with colorectal cancer (CRC) data. The results find immune and stromal cell subtypes spatially aligned with tumor subtypes.

2 Case Study

CRC is the second most common cause of cancer death in the United States with approximately 153,020 new cases in 2023 [16]. It is well-established that immune cells within the tumor microenvironment (TME) profoundly influence tumor development, progression, and treatment responses in colorectal cancer (CRC) patients. Histological analyses have observed that immune and stromal cells often position themselves around specific tumor regions, suggesting intricate interactions between these cell types. While such spatial relationships have been noted, comprehensive characterization of the precise spatial co-localization of immune, stromal, and tumor cell subpopulations in CRC using spatial transcriptomics (ST) data remains limited. Notably, recent studies have begun to address this gap: A study integrating spatial and single-cell transcriptomics revealed that CRC cells at the invasive front co-localize with macrophages, highlighting spatial intercellular communication between tumor cells and the immune microenvironment [17]. Another study demonstrated that the spatial organization and immune status of the tumor-stroma boundary are distinctive features of different CRC subtypes, influencing disease progression and therapeutic responses [18]. Building upon these insights, our objective is to utilize ST technology to systematically identify and characterize immune and stromal cell subpopulations that are spatially aligned with tumor cells. This approach aims to provide a more detailed understanding of cellular interactions within the CRC TME, potentially uncovering novel therapeutic targets or prognostic indicators.

We use data from two advanced RNA sequencing techniques, including single-cell RNA sequencing (SC) and spatial transcriptomics (ST) data. SC data is now widely used for analysis [19]. More recently, ST data has gained attention for enabling spatial inference of the TME by leveraging additional spatial location information [20, 21]. To assess the spatial interactions of specific cell types, we integrate information from both modalities during preprocessing to obtain

spatially inferred single cell data for further analysis. The SC data is a publicly available CRC dataset from Lee et al. [22]. The ST sample is public domain data from the 10X website (`https://www.10xgenomics.com/datasets/human-colorectal-cancer-11-mm-capture-area-ffpe-2-standard`).

Figure 1 shows the workflow for the implemented pipeline. SC data (y^g_{SC}) provides gene expression profiles for each individual cell in the tumor tissue, while ST data includes both gene expression and spatial information ($y^g_{\text{ST}}, y^s_{\text{ST}}$) for each spot on a tissue grid, representing the cumulative information of multiple cells within that spot. First, standard pre-processing and data cleaning is carried out for both, SC and ST data, using programs implemented in the *Seurat* package. For SC data, only genes that were detected in a minimum of three cells were retained. We excluded low-quality cells by setting thresholds for the number of expressed genes and detected Unique Molecular Identifier (UMI) counts: cells were retained if they expressed between 500 and 6,500 genes and had UMI counts ranging from 500 to 2,000,000. For ST data, genes with all 0s are removed. Additionally, we removed mitochondrial, microRNA, long non-coding, and predicted gene models to focus on high-confidence, protein-coding transcripts to avoid confounding signals from less informative or tissue-specific genes. Next, we use CellTrek [23] to spatially map the SC data leveraging the spatial annotation of ST data. The output from this step are single cells with gene expression and inferred spatial location on the tissue slide, i.e., spatially enriched SC data. Principle Component Analysis (PCA) is used for dimension reduction and Uniform Manifold Approximation and Projection (UMAP) transformations are implemented for visualization (but not for analysis). Based on known markers for immune, stromal, and tumor cells, we split the spatially enriched SC data into three subsets by cell type. Figure 2 shows the spatially enriched SC data with cell type annotation in the CRC datasets. In the two subfigures, single cells are colored by the annotated cell types and plotted with respect to spatial locations and gene expression UMAP scores, respectively. Using these spatially enriched data, the inference goal is then the discovery of spatially aligned random partitions, aligned across tumor, immune and stromal cells.

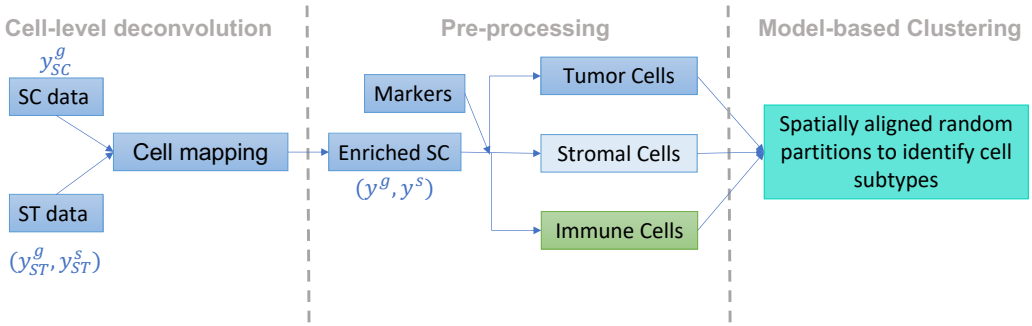


Figure 1: Workflow of the proposed method.

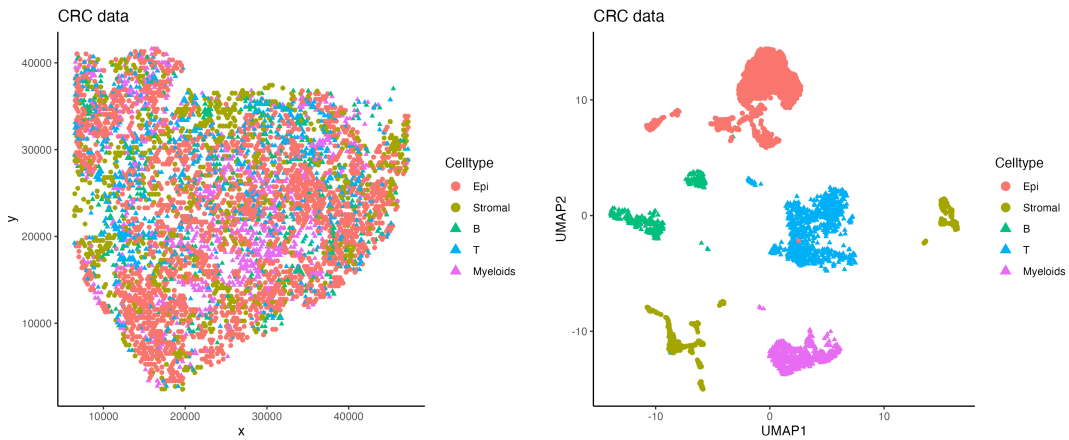


Figure 2: Spatially enriched CRC single cell data after preprocessing, plotted w.r.t. spatial coordinates (left panel) and UMAP scores of gene expression (right panel). B, T and Myeloids refer to immune subpopulations, Epi refers to tumor (epithelial) cells, and Stromal refers to stromal cells.

3 A spatially aligned random partition model (SARP)

3.1 Statistical model

Through cell mapping and preprocessing of ST and SC data, we obtain tumor, stromal and immune single cell data with imputed spatial locations. We first consider inference for two sets of cells, immune and non-immune cells (the latter including tumor and stromal cells), and defer extension to more than two types to later (Section 3.3). The main analysis goal remains inference on spatially aligned partitions for the two types of experimental units, immune and non-immune cells. For each of the two types, we use clustering to discover subtypes representing homogeneous cell subpopulations. The desired spatial alignment of some of these subtypes will reveal how immune cells interact with subtypes of non-immune cells.

We introduce the construction of the spatially aligned random partition models. Let y_{ji} denote the feature vector including gene expression and spatial coordinates for non-immune cells ($j = 1;\ i = 1, \ldots, n_1$) and immune cells ($j = 2;\ i = 1, \ldots, n_2$), respectively. We assume that the features y_{ji} are generated from mixtures $\int f(\cdot \mid \theta)\, dG_j(\theta)$ of a kernel $f(\cdot)$ with respect to discrete mixing measures G_j. We write the mixtures equivalently as hierarchical models with cell-specific latent variables θ_{ji}:

$$y_{ji} \mid \theta_{ji} \sim f(y_{ji} \mid \theta_{ji}), \tag{1}$$

$$\theta_{ji} \mid G_j \sim G_j. \tag{2}$$

The discrete nature of G_j induces ties among the θ_{ji} which create a random partition of the experimental units.

First, consider the special case that y_{ji} only contains the spatial coordinates whose distributions should be aligned across subtypes of the two different experimental units. The dependence across types $j = 1, 2$, can be introduced by a prior on (G_1, G_2). The discrete random probability measure G_1 is assumed to be generated from a Pitman-Yor process (PYP) with non-atomic base measure $G_1^\star$. Conditional on parameters $\theta_{1i} \sim G_1,\ i = 1, \ldots, n_1$, the mixing measure G_2 is assumed to be generated from another PYP,

$$\begin{aligned} G_1 &\sim \mathrm{PYP}(a_1, \sigma_1, G_1^\star) \\ G_2 \mid \theta_{1,1:n_1} &\sim \mathrm{PYP}(a_2 + b_2, \sigma_2, G_2^{\star s}). \end{aligned} \tag{3}$$

The construction of $G_2^{\star s}$ is where we introduce the desired dependence by allowing the atoms of $G_2^{\star s}$ to emerge at locations around θ_{1i}, $i = 1, \ldots, n_1$, or at new locations. This is achieved by assuming a corresponding structure in the base measure $G_2^{\star s}$ (introducing the superindex ^{s} in anticipation of later model extension beyond spatial features). Specifically, the base measure $G_2^{\star s}$ is a mixture of a new base measure $\widetilde{G}_2^{\star s}$ and kernels centered at $\theta_{11}, \ldots, \theta_{1n_1}$, with relative weights a_2 and b_2,

$$G_2^{\star s} \propto a_2 \widetilde{G}_2^{\star s} + b_2 \tilde{f}_1, \quad \widetilde{f}_1(\cdot) \propto \sum_{t=1}^{n_1} K(\cdot \mid \theta_{1t}). \tag{4}$$

The PYP is a generalization of the already mentioned DP where two parameters in the PYP allow for more flexibility in the stick-breaking process. In the application to clustering cells, we set $\sigma_1, \sigma_2 < 0$ to imply a priori a moderate number of clusters for the cell subtypes, favoring larger clusters with reasonable biological interpretations. The proposed model uses a_1, a_2, b_2 and different base measures $G_j^\star$ to allow control and flexibility of the dependence between conditions. Additionally, we use the same parametric family for $K(\cdot)$ as the sampling distribution $f(\cdot)$ in (1). This simplifies comptutation by allowing to treat the atoms in $\theta_{21}, \ldots, \theta_{2n_2}$ that are generated from $K(\cdot \mid \theta_{1i})$ as if they were additional observations, and leads to convenient posterior simulation schemes (details in Section 3.4).

Next we extend the model to the full observed feature vectors $y_{ji} = (y_{ji}^s, y_{ji}^j),\ j = 1, 2$, including spatial coordinates y_{ji}^s and gene expression y_{ji}^j. For example, for the non-immune cells ($j = 1$), we use y_{1i}^s to denote the spatial coordinates of the i-th non-immune cell, and use y_{1i}^1 to denote the gene expression subvector (using distinct superindices ^{j} in anticipation of the upcoming augmentation of the corresponding parameter vector θ_{ji}). Dependence across the two types is restricted to the spatial subvector by the following construction. We assume independent kernels on gene expression features and spatial locations, denoted as f_g and f_s, respectively,

$$y_{ji} \mid \theta_{ji} \sim f_s(y_{ji}^s; \theta_{ji}^s) f_g(y_{ji}^j; \theta_{ji}^j). \tag{5}$$

For both cell populations we introduce an augmented parameter vector θ_{ji} as a triple $\theta_{ji} = (\theta^s_{ji}, \theta^1_{ji}, \theta^2_{ji})$, with θ^s_{ji} indexing the sampling model for the spatial variable, and θ^j_{ji} indexing the model for gene expression for cell type j. Introducing θ^2_{1i} in the mixing measure for (y_{1i}, s_{1i}) and vice versa is done to allow use of the aligned random partition model (2) and (3), which requires G_j to be defined on the same space. We assume independent priors on $(\theta^s_{ji}, \theta^1_{ji}, \theta^2_{ji})$ in base measures $G^\star_j$. The prior distribution for the spatial subvector θ^s_{2i} in $G^\star_2$ induces the desired spatial dependence structure and remains $G^{\star s}_2$ as in (4). The sampling model (5) replaces (1). Together with (2) and (3) this completes the structure of the proposed spatially aligned random partition models (SARP). See below for specific parameric assumptions.

Specifically, in our applicaton, we assume independent Gaussian distributions for spatial coordinates and gene expressions in the sampling model, such that,

$$y_{ji} \mid \theta_{ji} \sim N(y^s_{ji}; \mu^s_{ji}, \Sigma^s_{ji}) N(y^j_{ji}; \mu^j_{ji}, \Sigma^j_{ji}), \quad j = 1, 2, \tag{6}$$

where $\theta_{ji} = (\mu_{ji}, \Sigma_{ji})$, and $\mu_{ji} = (\mu^s_{ji}, \mu^1_{ji}, \mu^2_{ji})$, $\Sigma_{ji} = \text{diag}(\Sigma^s_{ji}, \Sigma^1_{ji}, \Sigma^2_{ji})$. We assume independence for spatial and gene related features in the base measures $G^\star_1$ and $G^\star_2$,

$$G^\star_j(\theta^s, \theta^1, \theta^2) = G^{\star s}_j(\theta^s) \cdot G^{\star 1}_j(\theta^1) \cdot G^{\star 2}_j(\theta^2), \quad j = 1, 2. \tag{7}$$

We re-parametrize the covariance matrix in the Gaussian likelihood for spatial coordinates as $\Sigma^s_{ji} = \Gamma^s_{ji} R^s_{ji} \Gamma^s_{ji}$, where $\Gamma^s_{ji} = \text{diag}(\sigma^s_{ji1}, \ldots, \sigma^s_{jiD})$ is a diagonal matrix of the standard deviations and R^s_{ji} is the correlation matrix. In this way we have more control over the preference on spatial profiles of the cell subpopulations. The priors on spatial feature parameters $\theta^s_{ji} = (\mu^s_{ji}, \Gamma^s_{ji}, R^s_{ji})$ are given by $G^{\star s}_1 = \tilde{G}^{\star s}_1$ and $G^{\star s}_2$ in Equation (4) with

$$\tilde{G}^{\star s}_j(\theta^s_{ji}) = N(\mu^s_{ji}; \mu^s_0, \Gamma^s_{ji}/\kappa_0) \cdot \prod_d \text{Inv-Gamma}(\sigma^s_{jid}; \alpha_{0d}, \beta_{0d}) \cdot \text{Uniform}(R^s_{ji}), \quad j = 1, 2. \tag{8}$$

In the case of bivariate spatial coordinates the uniform on R^s_{ji} reduces to a uniform prior on the scalar correlation coefficient. The kernel $K(\cdot \mid \theta_{1t})$ in Equation (4) takes the same form as $\tilde{G}^{\star s}_j$ in Equation (8), replacing μ^s_0 by μ_{1t}. For gene related features, normal inverse-Wishart (NIW) priors are assumed in the base measures

$$G^{\star z}_j = \text{NIW}(\mu^z_0, \kappa_0, \Phi^z_0, \nu_0), \tag{9}$$

for $z = 1, 2$ and $j = 1, 2$. The inverse gamma prior on σ^s in (8) ensures that implied clusters remain spatially localized. Without a separation of prior specification on θ^s and θ^j, $j = 1, 2$, cluster formation with informative prior on σ^s could be entirely swamped by gene expression features, especially when high-dimensional gene expression features are used.

In summary, the SARP model introduces an aligned random partition model for non-immune and immune cells. The alignment is formalized as dependence in the spatial coordinate

subvector. Note the inherent asymmetry in the model, reflecting the asymmetric, causal nature of the underlying biology with tumor cells recruiting immune cell subpopulations. Finally, the very different nature of the two sets of experimental units is also reflected in the choice of kernels. Immune cells have different gene profiles from non-immune cells even though they are spatially aligned. There is no notion of the dependence extending to the gene expression part of the data vectors.

3.2 Properties

We first discuss the joint distribution of the spatial features y^s_{ji} across both types of cell under the SARP model (focusing on the spatial subvector only we drop the ^{s} superindex for the moment). Choosing $K(\cdot \mid \theta) = \delta_\theta(\cdot)$ in (4) the distribution $\widetilde{f}_1$ reduces to the empirical distribution of $\theta_{1,1:n_1}$. Under this assumption and the special case of the DP prior with discount parameters $\sigma_1 = \sigma_2 = 0$, and if a_1, a_2 and b_2 are chosen accordingly, then $(\theta_{ji};\ j = 1, 2$ and $i = 1, \ldots, n_j)$ arise as an i.i.d. sample from a single mixing measure G with a DP prior. The result clarifies the model structure by characterizing it as clustering all cells jointly. Using $K(\cdot \mid \theta) \neq \delta_\theta$ allows for some spatial dis-location of immune cells relative to non-immune cells.

Proposition 1 *If $\sigma_1 = \sigma_2 = 0$, $a_2 = a_1$, $b_2 = n_1$, $K(\cdot \mid \theta) = \delta_\theta(\cdot)$, and $G^\star_1 = \widetilde{G}^\star_2$, the model in equations* (2), (3) *and* (4) *is equivalent to*

$$\begin{aligned} \theta_{ji} \mid G \sim G, \quad j = 1, 2,\ i = 1, \ldots, n_j, \\ G \sim DP(a_1 G^\star_1) \end{aligned} \tag{10}$$

Proof: Note that the following two models are equivalent.

$$\begin{aligned} x_i \mid G \sim G, \quad i = 1, \ldots, n_1 \\ G \sim DP(aG^\star) \\ G \mid x_{1:n} \sim DP(aG^\star + n_1 \hat{f}_{1:n}) \\ x_i \mid G \sim G, \quad i = n_1 + 1, \ldots, n_1 + n_2 \end{aligned} \quad \leftrightarrow \quad \begin{aligned} x_i \mid G \sim G, \quad i = 1, \ldots, n_1 + n_2 \\ G \sim DP(aG^\star) \end{aligned} \tag{11}$$

where $\hat{f}_{1:n}$ is the empirical distribution of $x_{1:n}$. When $a_2 = a_1 = a$, $b_2 = n_1$, and $G^\star_1 = \widetilde{G}^\star_2 = G^\star$, then model (2), (3) *and* (4) *reduces to* (10) *with $x_i = \theta_{1,i}, i = 1, \ldots, n_1$ and $x_{n_1+i} = \theta_{2,i}$, $i = 1, \ldots, n_2$.*

The SARP model introduces latent $\theta_{j,i}$ generated from discrete random measures with PYP priors as specified in (3). Instead of using the PYP prior, one could consider any suitable Bayesian nonparametric prior, incorporating the same dependence structure on the base measures. For example, finite mixtures have proven useful in many applications. Using a mixture of finite mixtures (MFM) we can characterize posterior asymptotics. Under the MFM

prior, the posterior distribution $\Pi(G_1 \mid y_{11}, \ldots, y_{1n_1})$ and the conditional posterior distribution $\Pi(G_2 \mid y_{21}, \ldots, y_{2n_2}, \theta_{1,1:n_1})$ converge with increasing n_j to the true mixing measure G_j^0 $(j = 1, 2)$, with posterior contraction rates as follows.

Proposition 2 *Under model* (6) *and MFM prior, for* $j = 1, 2$*, we have*

$$\Pi\left(G_1 \in \bar{\mathcal{G}}(\Theta) : W_1(G_1, G_1^0) \lesssim (\log n_1/n_1)^{1/2} \mid y_{11}, \ldots, y_{1n_1}\right) \to 1,$$
$$\Pi\left(G_2 \in \bar{\mathcal{G}}(\Theta) : W_1(G_2, G_2^0) \lesssim (\log n_2/n_2)^{1/2} \mid y_{21}, \ldots, y_{2n_2}, \theta_{1,1:n_1}\right) \to 1,$$

in probability under G_j^0*. Here,* $\bar{\mathcal{G}}(\Theta)$ *denotes the space of all discrete measures, and* W_1 *denotes* L_1 *Wasserstein distance. The relation* $\lesssim$ *denotes inequality up to a constant multiple where the value of the constant is independent of* n_j*.*

Proof: The claim follows from Theorem 3.1 in Guha et al. [24]. For $j = 1$, the SARP model for $y_{11}, \ldots, y_{1n_1}$ could be written as $y_{1i} \mid \theta_{1i} \sim f(y_{1i} \mid \theta_{1i})$, $\theta_{1i} \sim \text{MFM}(\alpha, G_1^\star, p_K)$, which is a standard MFM model. For $j = 2$, given $\theta_{1,1:n_1}$ the SARP model for $y_{21}, \ldots, y_{2n_2}$ could be written as $y_{2i} \mid \theta_{2i} \sim f(y_{2i} \mid \theta_{2i})$, $\theta_{2i} \mid \theta_{1,1:n_1} \sim \text{MFM}(\alpha, G_2^\star, p_K)$, which is also a standard MFM model. Therefore, for models on both types of units, we can establish the posterior contraction rate following the scheme of proof in [24].

The asymptotic properties of the model with PYP priors are not straightforward to establish. However, our experimental results (not presented here) indicate that models employing both types of prior choices exhibit comparable performance in practice.

3.3 Extension to multiple cell types

In the previous discussions, we split the dataset into immune versus non-immune cells, and investigated the subtypes of immune cells that are recruited by certain non-immune subtypes. A natural extension of the model is then to more than two types of cells. In the application to the CRC data, we are interested in three types of cells, including tumor, stromal, and immune cells, and investigate how subpopulations of the latter two cell types are co-located with tumor cell sub-types. We therefore generalize the SARP model to spatially aligned clustering of multiple (≥ 2) types of experimental units where one of the types of units is considered as reference for the spatial co-localization.

Let $j = 1, \ldots, J$ index J types of experimental units. Without loss of generality, assume $j = 1$ is the reference type of units, such that some clusters of type j $(j > 1)$ units should locate around clusters of type 1 units. In our case, $j = 1, 2, 3$, denotes tumor, stromal, and immune cells, respectively. We follow the same construction as in Section 3.1. The sampling model and prior on θ remain as in (1) and (2), and we specify a joint prior on

the mixing measures $p(G_1, \ldots, G_J)$ by assuming independence of G_j $(j > 1)$ given G_1, i.e., $p(G_1, \ldots, G_J) = p(G_1) \prod_{j=1}^{J} p(G_j \mid G_1)$, with

$$\begin{aligned} G_1 &\sim \mathrm{PYP}(a_1, \sigma_1, G_1^\star) \\ G_j &\sim \mathrm{PYP}(a_j + b_j, \sigma_j, G_j^\star), \;\; G_j^\star \propto a_j \widetilde{G}_j^\star + b_j \tilde{f}_1, \;\; j = 2, \ldots, J \end{aligned} \tag{12}$$

with $\widetilde{f}_1$ as in (4), and the base measures $G_1^\star$ and $\widetilde{G}_j^\star$ $(j > 1)$ are specified like $G_1^\star$ and $\widetilde{G}_2^\star$ before, and hyperparameters remain unchanged. Under $J = 2$ the model reduces to the earlier construction.

3.4 Posterior Inference

Posterior inference is implemented as MCMC simulation. We introduce latent cluster membership indicators s_{ji} with $s_{ji} = k$ if $\theta_{ji} = \theta_{jk}^\star$, where $\theta_{jk}^\star,\ k = 1, \ldots, K_j$ are the unique elements of $\theta_{j,1:n_j}$. We can rewrite (1) as

$$y_{ji} \mid \theta^\star, s_{ji} = k \sim f(y_{ji} \mid \theta_{jk}^\star).$$

Under the PYP the implied prior for the indicators s_{ji} can be characterized as $p(s_{ji} = k) = w_{jk}$ with w_{jk} generated by a stick breaking construction, $w_{j1} = V_{j1}, \;\; w_{jk} = \prod_{\ell=1}^{k-1}(1 - V_{j\ell}) \cdot V_{jk}$, and

$$V_{jk} \sim \mathrm{Beta}(1 - \sigma_j, M_j + k\sigma_j),$$

where $M_1 = a_1, M_j = a_j + b_j,\ j = 2, \ldots, J$. The prior models for the atoms $\theta_{jk}^\star$ are

$$\begin{aligned} \theta_{1k}^\star &\sim G_1^\star, \\ \theta_{jk}^\star \mid \theta_{1,1:n_1} &\sim \frac{a_j}{a_j + b_j} \widetilde{G}_j^\star + \frac{b_j}{a_j + b_j} \tilde{f}_1, \;\; j = 2, \ldots, J, \end{aligned} \tag{13}$$

where $\widetilde{f}_1(\cdot) \propto \sum_{t=1}^{n_1} \widetilde{K}(\cdot \mid \theta_{1t})$. Here we rewrite the model with $\widetilde{K}$ on the space of $\theta = (\theta^s, \theta^1, \theta^2)$ for a convenient notation in the MCMC algorithm. The distribution is defined as $\widetilde{K}(\theta \mid \theta_{1t}) = K(\theta^s \mid \theta_{1t}^s) \cdot \tilde{G}_j^{\star 1}(\theta^1) \tilde{G}_j^{\star 2}(\theta^2)$ and the model remains unchanged. We introduce additional indicators $c_{jk} \in \{0, 1, \ldots, n_1\},\ k = 1, \ldots, K_j$ to replace (13) by

$$\theta_{jk}^\star \mid \theta_{1,1:n_1}, c_{jk} = t \sim \begin{cases} \widetilde{G}_j^\star & \text{if } t = 0 \\ \widetilde{K}(\cdot \mid \theta_{1t}) & \text{if } t = 1, \ldots, n_1, \end{cases} \tag{14}$$

with

$$p(c_{jk} = t) \propto \begin{cases} a_j & \text{for } t = 0 \\ b_j / n_1 & \text{for } t = 1, \ldots, n_1 \end{cases} \tag{15}$$

For a better mixing Markov chain we use a blocked Gibbs sampler based on a finite truncation of the PYP random probability measure, truncating at $K = K_{max}$. Using the indicators c_{jk} and s_{ij} the full conditional posterior probabilities can be derived in closed form. This motivates the following Gibbs sampling transition probabilities to define MCMC posterior simulation:

1. Update $\theta^\star$ by generating from the complete conditional posteriors. Let $L_{jk} = \{\ell : c_{j\ell} = i, s_{1i} = k\}$. Then

$$p(\theta^\star_{1k} \mid \dots) \propto \widetilde{G}^\star_1(\theta^\star_{1k}) \prod_{i:s_{1i}=k} f(y_{1i} \mid \theta^\star_{1k}) \prod_{j=2}^{J} \prod_{\ell \in L_{jk}} \widetilde{K}(\theta^\star_{j\ell} \mid \theta^\star_{1k}),$$
$$p(\theta^\star_{jk} \mid c_{jk} = 0, \dots) \propto \widetilde{G}^\star_j(\theta^\star_{jk}) \prod_{i:s_{ji}=k} f(y_{ji} \mid \theta^\star_{jk}), \quad j = 2, \dots, J,$$
$$p(\theta^\star_{jk} \mid c_{jk} = t, \dots) \propto \widetilde{K}(\theta^\star_{jk} \mid \theta_{1t}) \prod_{i:s_{ji}=k} f(y_{ji} \mid \theta^\star_{jk}), \quad t = 1, \dots, n_1, \; j = 2, \dots, J.$$

2. Update s_{1i} for $i = 1, \dots, n_1$,

$$p(s_{1i} = k \mid \dots) \propto p_{1k} f(y_{1i} \mid \theta^\star_{1k}) \prod_{j=2}^{J} \prod_{\ell : c_{j\ell} = i} \widetilde{K}(\theta^\star_{j\ell} \mid \theta^\star_{1k}), \quad k = 1, \dots, K_{max}$$

3. Update s_{ji} for $i = 1, \dots, n_j$, $j = 2, \dots, J$,

$$p(s_{ji} = k \mid \dots) \propto p_{jk} f(y_{ji} \mid \theta^\star_{jk}), \quad k = 1, \dots, K_{max}$$

4. Update $V_{j,k}$ for $j = 1, \dots, J$, $k = 1, \dots, K_{max} - 1$. Let $p_{j1} = V_{j1}$, $p_{jk} = \prod_{\ell=1}^{k-1}(1 - V_{j\ell}) \cdot V_{jk}$, $k = 2, \dots, K_{max} - 1$. Sample

$$V_{jk} \sim \text{Beta}(1 - \sigma_j + m_{jk}, M_j + k\sigma_j + \sum_{\ell=k+1}^{K_{max}} m_{j\ell}),$$

   where m_{jk} is the cardinality of $\{i : s_{ji} = k\}$.

5. Update c_{jk} for all currently imputed clusters, $j = 2, \dots, J$,

$$p(c_{jk} = 0 \mid \dots) \propto \frac{a_j}{a_j + b_j} \widetilde{G}^\star_j(\theta^\star_{jk}),$$
$$p(c_{jk} = t \mid \dots) \propto \frac{b_j}{a_j + b_j} \frac{1}{n_1} \widetilde{K}(\theta^\star_{jk} \mid \theta_{1t}), \quad t = 1, \dots, n_1.$$

Based on the conjugate prior assumptions all complete conditional posterior distributions are available in closed form and can be easily sampled from.

4 Simulation Study

We carry out simulation studies to assess inference under the proposed model. Experiments include analyses with simulated data, with semi-synthetic data and a small sensitivity study. Simulated data allows to verify the ability to discover spatial alignment under realistic conditions, assessing inference by visualization. Simulation with semi-synthetic data based on real single-cell data ensures evaluation under realistic conditions. The sensitivity study explores the impact of alternative prior choices for the weights in the SARP mixture model. In all simulation studies we focus on two types of cells, such as immune versus non-immune cells.

Hyper-parameters: Throughout the simulation studies, we implement the SARP model with parametric families as specified in Section 3.1. For the hyper-parameters in PYP priors, we set $a_1 = 2,\ a_j = 1,\ b_j = 1,\ j = 2, \ldots, J$, discounting parameters $\sigma_j = 2/K_{max}$ where $K_{max} = 10$. The hyper-parameters in the base measures are set as follows: For the parameters on spatial features in Equation (8), μ_0^s is the sample mean, $\kappa_0 = 1$, $\alpha_{0d} = 10^5$, $\beta_{0d} =$ the sample variance $/100\alpha_{0d}$. For the parameters on gene expression in Equation (9), μ_0^z is the sample mean under group z, $\Phi_0^z/100$ is the sample covariance matrix under group z, $\kappa_0 = 1$, $\nu_0 = 2+$ dimension of gene expressions.

Comparison: We compare results under SARP with inference under a Gaussian mixture model (GMM) as a benchmark. The number of clusters for the GMM is determined based on an ELBOW plot. Since the GMM does not include explicit parameters to indicate co-localization we have to add a suitable posterior summary. Given the discovered clusters under the GMM, for the purpose of summarizing simulation results we define the probability of an immune cluster being co-localized with a non-immune cluster (or spatially independent) as proportional to the likelihood of the immune cells in the cluster under the cluster-specific kernel of the non-immune clusters (or the distribution used in $\widetilde{G}_j^\star$). Such cluster-specific co-localization probabilities naturally imply cell-cell co-localization probabilities to be used in the following discussion of reports.

Reporting and summarizing inference on the random partition (s_{ji}) and on the co-localization of clusters (c_{2j}) requires some more clarifications.

Random partition: We report a posterior summary of the random partition by minimizing the variation of information (VI) loss as proposed in Dahl et al. [25], and implemented in the *salso* R package. The reported point estimate $\widehat{s}$ of the random partition characterizes subpopulations of immune and non-immune cells.

Co-location: We define co-localization summaries at the cluster and at the cell-level under the simulation truth and under the inference, using notation P for cluster-level summaries, and Q for cell-level summaries (and PQ for a mixed cluster-cell summary – see below), and a superindex ^{o} to mark the simulation truth. For each immune cell cluster ($j = 2, k$) and

non-immune *cell* t, the posterior probabilities $PQ_{kt} \equiv p(c_{jk} = t \mid \widehat{s}, data)$ summarize inference on co-localization, including characterization of uncertainties. Note that following the structure of the inference model PQ_{kt} is cell-specific for non-immune cell t. Alternatively, a cluster-specific posterior probability of co-localization for immune and non-immune clusters k and ℓ, respectively, can be reported as $P_{k\ell} \equiv \sum_{t \in C_\ell} PQ_{kt}$, with $C_\ell = \{t : s_{1t} = \ell\}$. A limitation of these summaries is the conditioning on cluster labels. Instead, we report a summary that avoids label-switching by focusing on cell-specific co-localization. Let Q denote an $(n_2 \times n_1)$ cell by cell co-localization matrix with $Q_{ij} = p(c_{2s_{2i}} = j \mid data)$, the probability that immune cell i appears in a cluster that is linked with non-immune cell j. The probability of immune cell i being co-localized with any of the non-immune clusters is represented by the row sums of matrix Q, i.e. $q_i = \sum_j Q_{ij}$. We will define a true binary co-localization status P^o between the immune cluster and the non-immune clusters under the ground truth (naturally P^o is a binary matrix), which induces a true cell-cell co-localization matrix Q^o (details when we introduce the simulation scenarios). Also $q_i^o = \sum_\ell Q^o_{s_i,\ell}$ is an indicator that cell i is co-localized with a non-immune cluster under the ground truth. We report the accuracy of estimating q_i^o as ACC $= 1 - \sum_i abs(q_i - q_i^o)/(n_2)$. For a practically more relevant summary of cell-cell co-localization, we report true negative rate (TNR, specificity) and true positive rate (TPR, sensitivity) as a function of a threshold λ on Q_{ij} as follows. Let TN $= \{(i,j) : Q^o_{ij} < \lambda\}$ and $N_0 = |\text{TN}|$, and similarly TP $= \{(i,j) : Q^o_{ij} \geq \lambda\}$ and $N_1 = |\text{TP}|$. Let then TPR $= \sum_{\text{TP}} I(Q_{ij} \geq \lambda)/N_1$, and TNR $= \sum_{\text{TN}} I(Q_{ij} < \lambda)/N_0$. Accordingly, the true classification rate is TCR $= \{\sum_{\text{TP}} I(Q_{ij} \geq \lambda) + \sum_{\text{TN}} I(Q_{ij} < \lambda)\}/(N_1 + N_0)$. In the simulation results we chose λ to achieve a set TNR, and report then compare the corresponding TPR and TCR.

4.1 Simulated data

We simulate data assuming bivariate gene expression features y^g_{ji} and spatial coordinates y^s_{ji}, mainly for the ease of visualization. The first row of Figure 3 shows the simulated cells plotting spatial coordinates and gene expression, respectively. Considering three true non-immune cell clusters, $K_1^o = 3$, and four immune cell clusters, $K_2^o = 4$, we first specify a ground truth for the co-localization between the immune and non-immune clusters as indicated in the first row of Table 1. The first three immune clusters are co-localized with non-immune cluster 2,2,1, respectively, and the fourth cluster is spatially independent. This defines the true cluster-cluster indicators P^o. From P^o, we induce a cell-cell co-localization truth Q^o by spreading any entry $P^o_{s_i,\ell} = 1$ for immune cell i uniformly across all non-immune cells j in cluster ℓ (to be used for the reported comparisons). We then specify a Gaussian mixture model with sampling distributions as assumed in (6) with location parameters of the co-located clusters being close to each other spatially, and then generate the bivariate gene expression features as shown in

Figure 3.

Implementing inference under the SARP and GMM models for this simulated data, we find an estimate for the random partition. The implied clusters are shown in Figures 3c-f. Table 1 summarizes estimated co-localization of the identified cluster and the accuracy of co-localization under SARP and GMM, respectively. In summary, under this easy scenario with spatially well-separated clusters under the simulation truth, both methods report accurate inference.

Co-located non-immune cluster	Immune clusters				Rates of classifying cells			
	1	2	3	4	ACC	TPR	TNR	TCR
Truth	2	2	1					
SARP	x	x	x		0.965	1.000	1.000	1.000
GMM	x	x	x		0.963	0.999	1.000	0.999

Table 1: Co-localization between clusters in simulated bivariate data. The symbol "x" denotes the immune cluster is co-localized with one of the non-immune clusters. Blank entries indicate absence of spatial correlation with any non-immune cluster. The specific indices of the linked clusters can not be compared, due to label switching. Instead the last four columns report the accuracy of estimating the cell-by-cell co-location matrix Q. See the text for a definition of the rates.

4.2 Semi-simulated data based on CRC

We generate semi-synthetic data based on the spatially enriched single cell data shown in Section 2 as a reference. In short, we start with K-means estimates of cluster arrangements for immune and non-immune cells. To create the simulation truth we replace the immune cell cluster centroids $\theta^{\star}_{2k}$ by simulating cluster specific parameters $\theta^{\star}_{2k}$ under (14) and (15): we (i) use the estimated $\theta^{\star}_{1\ell}$; (ii) simulate co-localization indicators $PQ^{o}_{k\ell}$; and (iii) generate $\theta^{\star}_{2k}$ under (15). In more detail, using PCA we get $y^{g}_{ji} \in \Re^{p}$ for $p = 20$ gene expression features, and use the following procedure to generate the semi-simulated data. Using the non-immune cells from the reference dataset we implement K-means to generate a ground truth for non-immune cell clusters, including corresponding cluster centers $\theta_{1,i}$. The K-means algorithm is then also applied to the immune cells. We only save the number of clusters, K_2, cluster memberships $s_{2,i}$, and cell-specific residuals $\epsilon_{2,i} \equiv y^{s}_{2,i} - \theta_{2,i}$, but generate (new) simulated immune cell cluster centers $\theta^{\star}_{2k}$ by first randomly generating true co-localization status $PQ^{o}_{k\ell}$, and then, given $PQ^{o}_{k\ell}$ we generate new immune cluster centers $\theta^{\star}_{2,k}$ using (14) and (15). Note that the simulation truth on $PQ^{o}_{k\ell}$ implies a simulation truth Q^{o}_{ij} for cell-cell co-localization. Finally, spatial coordinates for immune cells in the semi-synthetic data set are then generated

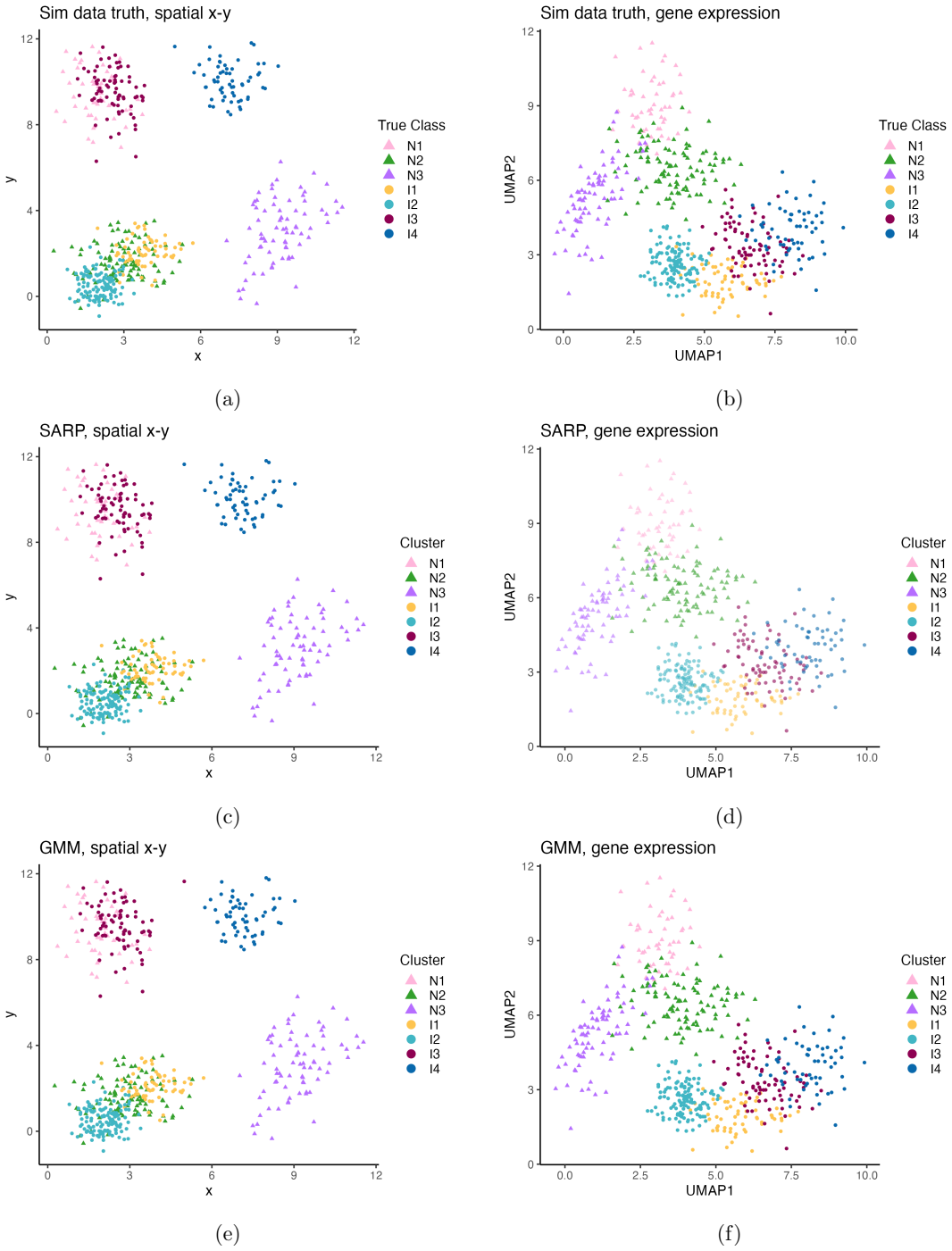


Figure 3: Simulation truth and estimated partitions using SARP and GMM on 2D data. In the first column the cells are plotted w.r.t. spatial coordinates and in the second column w.r.t. gene expressions.

as $y^s_{2,i} = \theta_{2,k} + \epsilon_{2,i}$ using the earlier saved residuals and cluster centers.

We evaluate the accuracy of identifying co-localization comparing SARP and GMM. Figure A1 in the Appendix shows the clustering result and Table 2 summarizes the estimated co-localization between the clusters. Under this semi-simulated scenario that closely mimics the actual data SARP still identifies cell subpopulations and the spatial alignment, even with the relatively high dimensional gene expression and less clearly separated bivariate spatial clusters.

Co-located non-immune cluster	Immune clusters										Rates of classifying cells			
	1	2	3	4	5	6	7	8	9	10	ACC	TPR	TNR	TCR
Truth	8	5	5	4	9	5	1	–	–	–				
SARP	x	x	x	x	x		x	x	x		0.84	0.32	0.96	0.93
GMM		x				–	–	–	–	–	0.55	0.14	0.96	0.92

Table 2: Co-localization across clusters and accuracy in simulated semi-synthetic data. Same as Table 1, for the semi-simulated data.

5 CRC data analysis

We fit the proposed SARP model with the CRC data introduced in Section 2, aiming to discover immune and non-immune cell subpopulations, and to understand their spatial alignment. The CRC data (after preprocessing) is shown in Figure 2. In this application the experimental units (indexed by i) are the single cells, and their gene expression and inferred spatial locations are the data. For gene expressions y^g_{ji} we use the top 20 PCA dimensions. For the SARP, we use $K_{max} = 15$, and the setting of all other hyper-parameters as described in the simulation studies.

Posterior inference on the partitions reports the following homogeneous cell sub-populations. A point estimate of the partition, i.e., cluster assignment, is obtained as described in the simulation studies. Clusters with ignorable cardinality (cardinality < 10) are combined with the closest larger clusters.

Clustering into homogeneous cell sub-populations: The allocation of cells to these clusters can be seen in Figure 4a. We identify 15 non-immune cell clusters and 15 immune cell clusters. We annotate the identified clusters by considering the cell type annotation (using known marker gene as implemented in the annotation under Seurat clustering [26]) of all cells within each cluster, and annotate the identified clusters with the main cell types found in each cluster (clusters are found to be almost pure subsets of the annotated cell types, making the cluster annotation biologically meaningful). The annotated clusters include 11 Epi, 4 Stromal, 6 Myeloid, 7 T-, and 2 B-cell subtypes. The spatial alignment is shown in Figure 4b and we

include more discussions on the interpretation with the extended model below.

Extension to three types: Next we extend the scope of the analysis by considering three types of cells, distinguishing tumor ($j = 1$), immune ($j = 2$), and stromal ($j = 3$) cells. We implement the extended version of the model with $J = 3$ types to investigate co-localization of immune and stromal cell sub-types with tumor subtypes. The setting of the hyperparameters and posterior summaries remains unchanged as in the simulation studies. Inferred cell clusters are shown in Figure 5a. Under the extension to $J = 3$ types we identify a finer resolution of cell sub-populations. Inference under the extended model with $J = 3$ allows more detailed conclusions on how different subtypes of cells interact with tumor cells. The estimated clusters are shown in Figure 5a, and the clusters are annotated using the same approach described in the previous paragraph. We identify 15 epithelial clusters, 14 stromal clusters and 15 immune cell clusters.

Spatial alignment: Given the cell subtypes, the estimated spatial alignment is shown in Figure 5b. The figure shows the subtypes of immune and stromal cells recruited by different tumor subtypes, including cells that are co-located with epithelial cell subtypes Epi5, 7, 8, 10, respectively. Clusters are plotted w.r.t. the spatial coordinates over the histology image. For example, on the bottom of the figure, we show one myeloid and one stromal cell subtype, are spatially aligned with the epithelial cell subtype 10.

Uncertainty quantification: Besides point estimates of inferred spatial alignment, uncertainty in the reported co-localization can be reported as follows. For example, immune cluster T4 has a probability of 96% of being spatially aligned with Epi13 (Figure B2) and 4% of being not spatially aligned. That is, although sub-population T4 is (*a posteriori*) more likely spatially co-localized with Epi13, there is a remaining 4% probability of being spatially independent.

For a more in-depth biological interpretation of the observed co-localization patterns, we performed differential expression (DE) analysis among the spatially defined epithelial subsets and identified distinct markers genes per group (Figure 6). Epithelial cells were categorized based on their spatial enrichment for specific TME components: Epi-Inner (surrounded predominantly by other epithelial cells), Epi-Mixed-I (in proximity to mixed immune populations), Epi-Myeloid, Epi-T, Epi-B, and Epi-Stromal. We focused on confident associations supported by robust marker gene expression and literature validation. For example, Epi-Stromal cells expressed high levels of *CXCL1/2/3*, a chemokine axis previously shown to mediate epithelial-stromal crosstalk and promote a pro-tumorigenic microenvironment [27]. Epi-Myeloid cells were enriched for *AGR2* gene, which encodes a protein implicated in macrophage-epithelial communication that promotes an immunosuppressive niche [28]. Meanwhile, Epi-T cells upregulated *CEACAM1* and *CEACAM6*, two cell adhesion molecules known to inhibit antitumor T cell activity and contribute to immune evasion [29]. These results recapitulate well-established epithelial- stromal/immune interactions, supporting the biological relevance of our spatial clas-

sification. Furthermore, our framework enables deeper exploration of epithelial phenotypic diversity shaped by local microenvironments, offering new opportunities to define ecosystem-specific vulnerabilities that have the potential to inform therapeutic targeting.

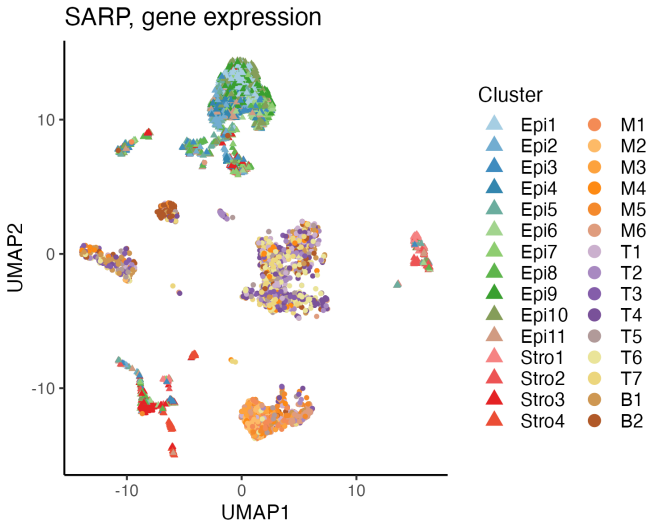
6 Discussion

We introduced a spatially aligned random partition model (SARP). SARP addresses a gap in existing methods by introducing inference on regressing clusters within a partition of one type of experimental unit to clusters formed by a random partition of another type of experimental unit. Inference under the SARP model identifies clusters on multiple sets of experimental units allowing alignment on a subvector of spatial features, but independence with respect to other features. While we focus on spatial alignment, the model and inference remain equally valid for any other partition of multivariate outcomes. The construction is asymmetric, assuming that one type of experimental units is a natural reference for a generative model of co-location.

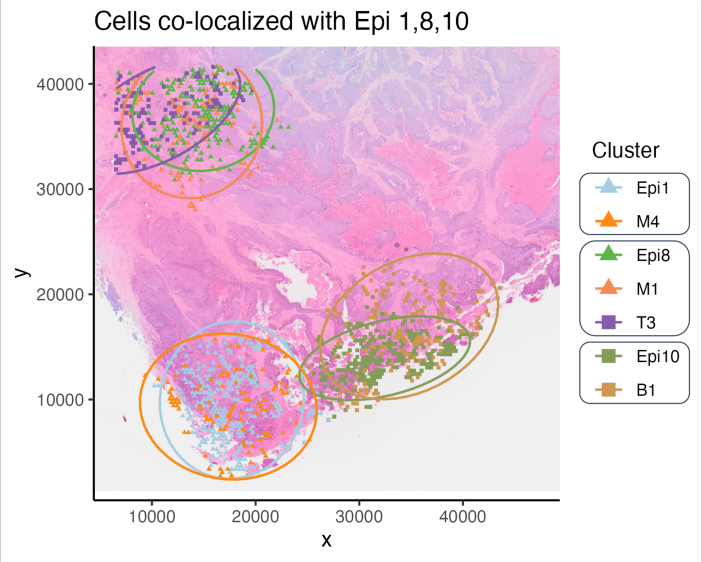
We suggested the use of PYP priors with negative discount parameter to induce an upper bound on the number of identified clusters. The choice was motivated by the application, based on the resolution of cell subtypes that we are interested in (and that is biologically meaningful to interpret). This use of an upper bound on the number of clusters is not required for the model, and could be dropped if desired, simply by using a non-negative discount parameter.

More complicated structure of the spatial co-localization could be considered if needed. For example, with $J > 2$ sets of experimental units, instead of taking one type of unit as a reference one could allow for co-localization between all possible pairs of clusters of different sets of unit. Formally this could be introduced by using, for example, a CAR type of model across all types of experimental units.

Finally, several limitations remain. One limitation is the assumed a priori independence of cluster locations. For easier interpretation it could be desireable to instead consider a repulsive prior to favor distinct and well separated clusters [30–32]. Another important limitation in the context of the application to the CRC data is the multi-step nature of the bioinformatics pipeline to impute the spatial annotation of the single cell data based on available spatial transcriptomics data. The pipeline ignores the uncertainty in this mapping. Finally, the model casts the problem as one of pairwise co-location. Alternative approaches could construct networks of cell sub-types.

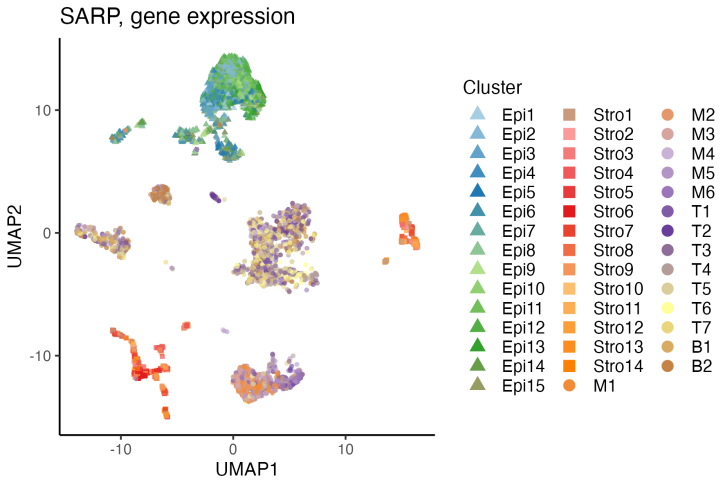


(a)

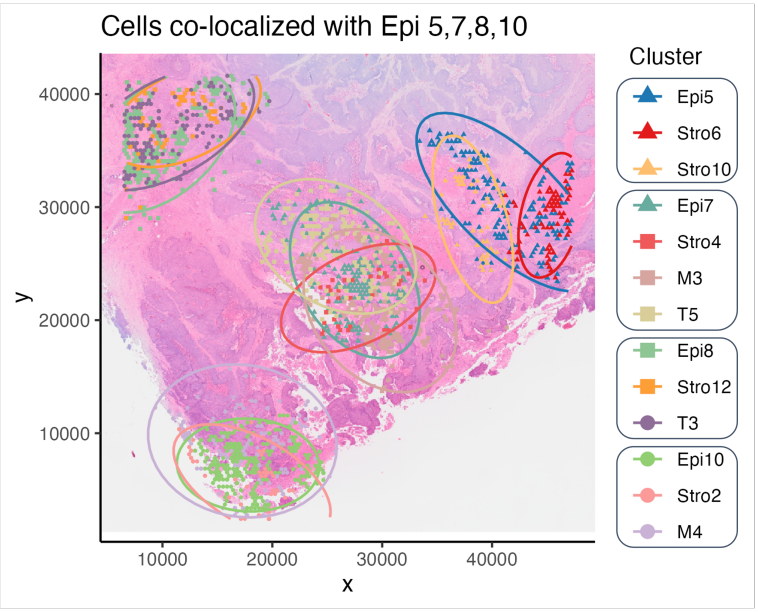


(b)

Figure 4: Inference results for the CRC data under the SARP model with two cell types. (a) Estimated clusters plotted w.r.t. UMAP scores of gene expression, colored by estimated clusters. The clusters are indexed by the annotated celltypes. (b) Estimated spatial alignment across the discovered clusters under the 2-type SARP model. The figure show immune cell clusters co-localized with Epi1, Epi8, and Epi10, respectively. All cells are plotted with respect to spatial coordinates. The contours are 95% probability contours of the inferred distribution of the subpopulations on spatial coordinates. For reference, the background shows the histology image. See the appendix for spatial alignment for the remaining tumor cell clusters.



(a)



(b)

Figure 5: Inference results for the CRC data under the extended SARP model. Same as Figure 4 for the extended SARP model.

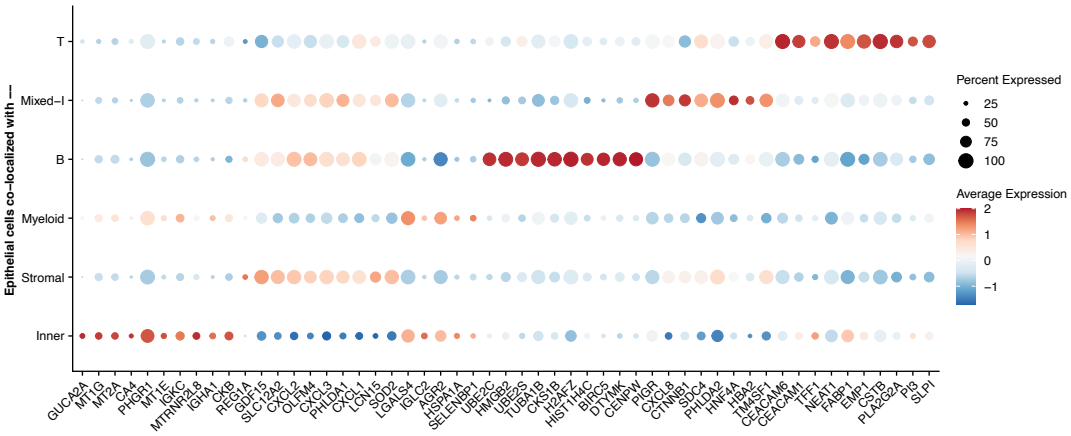


Figure 6: Differential gene expression analysis for epithelial cells by different inferred micro-environment. Epithelial cells are classified into subtypes based on their co-localization with surrounding cells: those recruiting immune cells, including "Myeloid", "T", "B", and "Mixed-I" (mixed immune types); those surrounded by "Stromal" cells; and "Inner" subtypes located within epithelial clusters.

References

[1] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2017.

[2] Pierpaolo De Blasi, Stefano Favaro, Antonio Lijoi, Ramsés H Mena, Igor Prünster, and Matteo Ruggiero. Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229, 2013.

[3] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[4] Alexander Gnedin and Jim Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138:5674–5685, 2006.

[5] D.M. Cifarelli and E. Regazzini. Problemi statistici non parametrici in condizioni di scambialbilità parziale e impiego di medie associative. Technical report, Quaderni dell'Istituto di Matematica Finanziaria, Torino, 1978.

[6] Pietro Muliere and Sonia Petrone. A Bayesian predictive approach to sequential search for

an optimal dose: Parametric and nonparametric models. *Journal of the Italian Statistical Society*, 2:349–364, 1993.

[7] Peter Müller, Fernando Quintana, and Gary Rosner. A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):735–749, 2004.

[8] Michalis Kolossiatis, Jim E Griffin, and Mark FJ Steel. On Bayesian nonparametric modelling of two correlated distributions. *Statistics and Computing*, 23:1–15, 2013.

[9] John FC Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1993. ISBN 0-19-853693-3. Oxford Science Publications.

[10] Antonio Lijoi, Bernardo Nipoti, and Igor Prünster. Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20:1260–1291, 8 2014.

[11] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[12] Abel Rodriguez, David B Dunson, and Alan E Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

[13] S.N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA, 1999. American Statistical Association.

[14] Yunshan Duan, Shuai Guo, Wenyi Wang, and Peter Müller. Immune profiling among colorectal cancer subtypes using dependent mixture models. *Journal of the American Statistical Association*, pages 1–14, 2025.

[15] Nathaniel Osher, Jian Kang, Santhoshi Krishnan, Arvind Rao, and Veerabhadran Baladandayuthapani. SPARTIN: a Bayesian method for the quantification and characterization of cell type interactions in spatial pathology data. *Frontiers in Genetics*, 14:1175603, 2023.

[16] Rebecca L Siegel, Nikita Sandeep Wagle, Andrea Cercek, Robert A Smith, and Ahmedin Jemal. Colorectal cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(3): 233–254, 2023.

[17] Yuki Ozato, Yasuhiro Kojima, Yuta Kobayashi, Yuuichi Hisamatsu, Takeo Toshima, Yusuke Yonemura, Takaaki Masuda, Kouichi Kagawa, Yasuhiro Goto, Mitsuaki Utou, et al. Spatial and single-cell transcriptomics decipher the cellular environment containing

HLA-G+ cancer cells and SPP1+ macrophages in colorectal cancer. *Cell Reports*, 42(1), 2023.

[18] Yu Feng, Wenjuan Ma, Yupeng Zang, Yanying Guo, Young Li, Yixuan Zhang, Xuan Dong, Yi Liu, Xiaojuan Zhan, Zhizhong Pan, et al. Spatially organized tumor-stroma boundary determines the efficacy of immunotherapy in colorectal cancer patients. *Nature Communications*, 15(1):10259, 2024.

[19] Geng Chen, Baitang Ning, and Tieliu Shi. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics*, 10:317, 2019.

[20] Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

[21] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19 (5):534–546, 2022.

[22] Hae-Ock Lee, Yourae Hong, Hakki Emre Etlioglu, Yong Beom Cho, Valentina Pomella, Ben Van den Bosch, Jasper Vanhecke, Sara Verbandt, Hyekyung Hong, Jae-Woong Min, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nature Genetics*, 52(6):594–603, 2020.

[23] Runmin Wei, Siyuan He, Shanshan Bai, Emi Sei, Min Hu, Alastair Thompson, Ken Chen, Savitri Krishnamurthy, and Nicholas E Navin. Spatial charting of single-cell transcriptomes in tissues. *Nature Biotechnology*, pages 1–10, 2022.

[24] Aritra Guha, Nhat Ho, and XuanLong Nguyen. On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4):2159–2188, 2021.

[25] David B Dahl, Devin J Johnson, and Peter Müller. Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics*, 31(4):1189–1201, 2022.

[26] Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, and Rahul Satija. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 2023.

[27] Chenjia Zhou, Ying Gao, Peilun Ding, Tao Wu, and Guang Ji. The role of CXCL family members in different diseases. *Cell Death Discovery*, 9(1):212, 2023.

[28] Mengqi Jia, Xiaojing Chen, Wenxue Guo, Dapeng Ma, Peng Wang, Huanmin Niu, Changhong Liu, Xianjuan Lin, QiQi Lu, Jing Wang, et al. AGR2-mediated cell-cell

communication controls the antiviral immune response by promoting the thiol oxidation of TRAF3. *Redox Biology*, 82:103581, 2025.

[29] Jessica Pinkert, Hans-Henning Boehm, Mark Trautwein, Wolf-Dietrich Doecke, Florian Wessel, Yingzi Ge, Eva Maria Gutierrez, Rafael Carretero, Christoph Freiberg, Uwe Gritzan, et al. T cell-mediated elimination of cancer cells by blocking CEACAM6–CEACAM1 interaction. *Oncoimmunology*, 11(1):2008110, 2022.

[30] Yanxun Xu, Peter Müller, and Donatello Telesca. Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, 72(3):955–964, 2016.

[31] José J Quinlan, Fernando A Quintana, and Garritt L Page. On a class of repulsive mixture models. *Test*, 30:445–461, 2021.

[32] Mario Beraha, Raffaele Argiento, Federico Camerlenghi, and Alessandra Guglielmi. Bayesian mixtures models with repulsive and attractive atoms. *arXiv preprint arXiv:2302.09034*, 2023.