ST-Assign: a probabilistic model for joint cell type identification in spatial transcriptomics and single-cell RNA sequencing data

Agnieszka Geras[1,2] **and Ewa Szczurek**[2,✉]

[1]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
[2]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland

Abstract
We present ST-Assign, a novel computational tool for joint cell-type annotation in single-cell RNA sequencing and cell-type mixture decomposition in spatial transcriptomics data. The model integrates the two data sources to enhance cell-type identification. It accounts for shared variables such as gene expression profiles and leverages prior knowledge about marker genes. We formulate the model as a generative graph-based structure and employ Markov chain Monte Carlo for the inference. We demonstrate the model's utility on mouse brain data, achieving simultaneous cell-type annotation and decomposition of cell-type mixtures. ST-Assign provides valuable insights into cell populations within organisms, improving our understanding of cellular heterogeneity.

probabilistic model | MCMC sampling | spatial transcriptomics data | cell types | single-cell RNA-seq data
Correspondence: ***szczurek@mimuw.edu.pl***

Background

Today, scientists are able to investigate individual cells to gain deeper insights into the diverse cell populations that comprise organisms. In particular, single-cell RNA-seq data (scRNA-seq) is commonly used to examine the extent of transcriptional heterogeneity between cells and identify novel sub-populations within them (1). Cell type annotation is a vital step in the scRNA-seq data analysis pipeline. It is typically conducted through manual data inspection and expert decision-making, which can, unfortunately, be laborious and susceptible to potential biases (2). As a solution, computational tools have been proposed to facilitate automatic, probabilistic cell-type identification, such as CellAssign (3) and scSorter (4).
While scRNA data lack information on cells' position within the tissue, spatial transcriptomics (ST) technology (5) enables its retention. Nonetheless, the mini-bulk nature of ST poses the challenge of decomposing hidden cell-type mixtures within each spatial location (spot). To overcome this, we have developed Celloscope (6), specifically designed to infer the prevalence of each cell type from ST data.
When scRNA-seq and ST data originate from the same tissue, an identical set of cell types can be assigned to individual cells in scRNA-seq or deconvoluted in each ST spot. Importantly, variables involved in the data-generating process for these two separate datasets overlap. For example, gene expression profiles characteristic of each cell type rule the observations in both scRNA-seq and ST data. Therefore, a model that would integrate scRNA-seq and ST data based on such shared variables has the potential to enhance the statistical power of cell-type identification.
This calls for a novel approach for joint single-cell type identification and cell-type mixture decomposition.

ST-Assign jointly annotates cell types to single cells and decomposes cell-type mixtures. Here we propose a novel hierarchical Bayesian probabilistic model, called ST-Assign, integrating data from the two sources:

1. single-cell RNA-seq data,
2. spatial transcriptomics data;

and a marker-gene-driven computational method to jointly perform the two tasks that until now were carried out independently:

1. cell type annotation in single-cell RNA-seq data,
2. cell-type mixture decomposition in spatial transcriptomics spots.

ST-Assign aims to combine signals from two data sources obtained via different experimental techniques with distinct yet complementary strengths and weaknesses. ST-Assign preserves single-cell resolution and, at the same time, benefits from spatial information. Moreover, datasets generated on different platforms are expected to borrow statistical strength from one another and, consequently, be less affected by the experimental noise.
We assume the list of marker genes and their specific cell types are given *a priori*, but gene expression profiles of those cell types are unknown beforehand. As a result, the expression level for each marker gene in each cell type is estimated as a part of

the model inference. The postulated model is generative and takes into account the dependency of the observed gene expression on hidden variables such as the cell type assignment and cell type mixtures' composition. Moreover, ST-Assign is specifically designed to address the effects of differences between the two experimental techniques by introducing dedicated variables. Conditional probability distributions expressing the dependence structure were used to develop the model's inference procedure via Markov chain Monte Carlo (the Metropolis-within-Gibbs algorithm). On top of that, we provide a biological showcase of ST-Assign usage concerning mouse brain data.
The model takes as input an unannotated scRNA-seq dataset, an ST dataset taken from the same tissue type, prior knowledge about marker genes encoded as a binary matrix, and, optionally, preliminary estimates for the number of all cells within each ST spot. The model output constitutes cell-type annotations in single-cell RNA-seq and the count of each cell type in each ST spot (cell-type decomposition).

Formulation of the ST-Assign model

ST-Assign represents all variables of interest as random variables (Table 1). We use a graph-based representation to express the conditional dependence structure between them (Figure 1).
Since ST-Assign integrates data from the two sources, naturally, the model variables can be divided into three components: (i) variables modeling ST data, (ii) variables modeling scRNA-seq data, and (iii) variables shared between the two components mentioned above. Naturally, the component modeling spatial transcriptomics data is analogous to Celloscope's structure; however, we introduced substantial changes concerning how gene expression given by mixtures of cells is modeled. In the new ST-Assign model, we explicitly account for the number of cells of every type in each ST spot. To this end, we employ the multinomial distribution to model the cell type composition of a spot.

Observed variables		
C_{gs}	Negative binomial	gene expression for gene g in spot s
C'_{gc}	Negative binomial	gene expression for gene g in cell c
B_{gt}	-	indicator whether gene g is a marker for cell type t
Hidden variables		
θ_{st}	Gamma	unnormalised abundance of cell type t in spot s
λ_0	-	base gene expression level, shared across all genes and cell types
Λ_{gt}	-	over-expression for a marker gene g in its specific cell type t
N_s	Poisson	number of cells in spot s
ρ_g	-	gene-dependent cross-platform factor for gene g
ρ_0	-	shared cross-platform factor for non-markers
N_s	Poisson	number of cells in spot s
$M_{s:}$	Multinomial	cell type composition in spot s
p_g	-	success probability parameter for gene g
Z_{st}	Bernoulli	indicator variable stating presence of cell type t in spot s
π_{st}	Beta	prior for the probability of presence of cell type t in a spot s
T_c	Categorical	cell type of cell c
D	Dirichlet	prior distribution on T_c
Deterministic variables		
h_{st}	Dirichlet	fraction of cell type t in a spot s

Table 1. Observed, hidden, and deterministic variables in ST-Assign, their corresponding probability distributions, and their interpretations.

Hyperparameters	
α	adjustment for the average number of cell types present in ST spot
a	rate parameter for cell types present in ST spot
b	scale parameter for cell types present in ST spot
a_0	rate parameter for cell types absent in ST spot
b_0	scale parameter for cell types present in ST spot
l_s	estimate for the total number of cells in ST spot
d	concentration parameters for cell types in scRNA-seq data

Table 2. Description of hyperparameters of ST-Assign variables.

Let $s \in \{s_1, s_2, \ldots, s_M\}$ index spots, $c \in \{c_1, \ldots, c_P\}$ cells and $g \in \{g_1, g_2, \ldots g_K\}$ genes. We assume t_N cell types are present in the considered tissue, indexed by $t \in \{t_1, t_2, \ldots, t_N\}$. Again, matrix B encodes prior knowledge about marker genes, and B_{gt} takes the value 1 if gene g is a marker for cell type t and 0 otherwise.

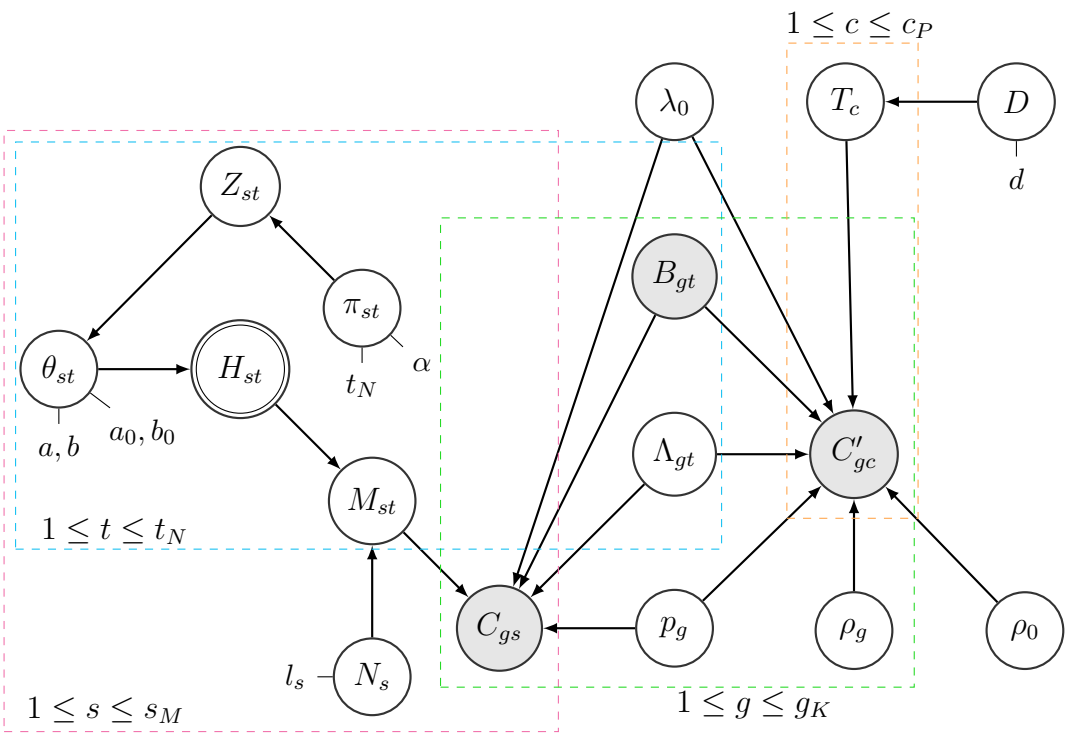

Fig. 1. The graphical representation of ST-Assign. Double-circled nodes correspond to deterministic variables, gray to observed, and the remaining nodes correspond to hidden (inferred) variables. Arrows represent probabilistic dependencies. ST-Assign variables are described in Table 1 and its hyperparameters in Table 2. Dotted magenta, blue, green and orange rectangles (plates) represent variables that repeat in the model.

Similarly, as in the case of Celloscope, Λ_{gt} denotes the over-expression of gene g when it is a marker for type t, and λ_0 represents the base expression level of any gene shared across all the genes. We preserve the central assumption about the increased expression of the marker genes in their specific cell types compared to other cell types. Therefore, we assume the expected expression of a single cell c of type t equals:

$$\Lambda_{gt} B_{gt} + \lambda_0.$$

Consequently, if gene g is a marker gene for cell type t, its expression in cell c equals $\lambda_0 + \Lambda_{gt}$. Otherwise, $B_{gt} = 0$, and it is equal to λ_0. The variable Λ_{gt} is shared between the two components modeling ST and scRNA-seq data.

ST-Assign serves two purposes: delineating cell-type composition of spots in ST data, represented by the hidden variable M_{st}, and cell-type annotation in scRNA-seq data, defined by the hidden variable T_c. Therefore, M_{st} and T_c are the two central variables in the model. M is a matrix with s_M rows and t_N columns. The value of an element M_{st} is the number of cells in spot s that are of cell type t. One row of the matrix M, denoted as $M_{s:} = [M_{st_1}, \ldots, M_{st_N}]$ represents the hidden composition of spot s. The number of all cells of all types within spot s is denoted N_s. Clearly, a given row's entries sum up to N_s:

$$N_s = \sum_{t \in \text{Types}} M_{st}.$$

We impose:

$$N_s \sim \text{Pois}(l_s),$$

where l_s denotes the preliminary estimates of the number of all cells for each spot. When it comes to modeling scRNA-seq data, T_c is a categorical variable indicating the cell type of cell c:

$$T_c \mid D \sim \text{Categorical}(D)$$

on which we impose Dirichlet prior

$$D \sim \text{Dirichlet}(d)$$

with the concentration parameter d. D and d are vectors of sizes equal to the number of cell types t_N:

$$D = [D_{t_1},\, D_{t_2},\, \ldots, D_{t_N}],$$

$$d = [d_{t_1},\, d_{t_2},\, \ldots, d_{t_N}].$$

For gene g, its expression in spot s, denoted C_{gs}, is measured as the count of reads that map to gene g from spot s. Likewise, the expression of gene g in cell c, denoted as C_{gc}, is measured as the count of reads that map to gene g from cell c. The gene expression matrices C and C' are modeled as observed variables. A row $C_{g:} = [C_{gs_1}, \ldots, C_{gs_M}]$ represents the expression of gene g across spots and a column $C_{s:} = [C_{g_1 s}, \ldots, C_{g_K s}]$ represents the gene expression in spot s across marker genes. Analogously, a row $C'_{g:} = [C'_{gc_1}, \ldots, C'_{gc_P}]$ represents the expression of gene g across cells and a column $C'_{:c} = [C'_{g_1 c}, \ldots, C'_{g_K c}]$ represents expression of cell c across marker genes.

Let us denote C^t_{gs} as the number of counts coming from a cell of type t. We assume that the random variable C_{gs} depends on the hidden cell-type composition in spot s, modeled as $[M_{st_1}, \ldots, M_{st_N}]$. Therefore, C_{gs} can be expressed as a sum of independent random variables:

$$C_{gs} \overset{d}{=} M_{st_1} C^{t_1}_{gs} + M_{st_2} C^{t_2}_{gs} + \cdots + M_{st_n} C^{t_n}_{gs},$$

where $\overset{d}{=}$ denotes equality in distribution. In the subsequent, we shall use the following two remarks:

Remark 1: $X \sim NB(r,p)$ is equivalent to $X \sim NB\left(\frac{1-p}{p}\mu, p\right)$, where $\mu = \frac{p}{1-p} r$.

Remark 2: Let X, Y be independent random variables satisfying $X \sim NB(r_1, p)$, $Y \sim NB(r_2, p)$. Then:

$$X + Y \sim NB(r_1 + r_2, p).$$

We assume:

$$C_{gs} \mid M_{s:}, B_{g:}, \Lambda_{g:}, \lambda_0, p_g \sim \mathrm{NB}\left(\frac{1-p_g}{p_g} \sum_t M_{st}(\lambda_0 + B_{gt}\Lambda_{gt}),\, p_g\right),$$

where p_g (success probability for gene g) accounts for the over-dispersion in gene expression data and depends only on the gene in question. Similarly to the variable Λ_{gt}, the variable p_g, in the graph's structure, is a child of both C_{gs} and C'_{gc}, and as a consequence, is shared between the two model's components, accounting for ST and scRNA-seq data.

We use a simple feature allocation model to represent the presence of cell types in spots. We set

$$\pi_{st} \sim Beta\left(\frac{\alpha}{t_N}, 1\right). \tag{1}$$

Let Z_{st} indicate whether type t is present in spot s (takes value 1, if t is present in s and value 0 otherwise). We set:

$$Z_{st} \sim Bernoulli(1, \pi_{st}). \tag{2}$$

Let θ_{st} denote the unnormalised abundance of type t in spot s. In the case when a given type t is present in spot s ($Z_{st} = 1$) we expect θ_{st} to take values more distant from 0 as compared to a situation when $Z = 0$. Therefore, we define a conditional distribution for θ_{st} given the value of Z_{st} in the following manner:

$$\begin{aligned} \theta_{st} \mid Z_{st} = 1 &\sim Gamma(a, b) \\ \theta_{st} \mid Z_{st} = 0 &\sim Gamma(a_0, b_0) \end{aligned} \tag{3}$$

where a, b, a_0, b_0 are chosen so that values sampled from $Gamma(a,b)$ are significantly larger than small values sampled from $Gamma(a_0, b_0)$. Let us denote $\Theta_{s:} = [\theta_{st_1}, \theta_{s_{t_2}}, \ldots, \theta_{st_N}]$ as one row of matrix Θ describing unnormalised abundance of all cell types in spot s.

To model M_{st}, we use the multinomial distribution with the number of trials given as the number of all cells in spot s (N_s) and the event probability parameter $H_{s:}$

$$M_{st} \mid N_s, H_s \sim \mathrm{Multinomial}(N_s, H_{s:}),$$

where $H_{s:}$ is deterministic and computed based on the unnormalized abundance of cell types within spot s:

$$H_{s:} = \frac{\theta_{s:}}{\sum_{t=1}^{N} \theta_{st}}.$$

The three parameters: θ_{st}, Z_{st} and π_{st} are defined in precisely the same manner as in Celloscope (6).

To model gene expression counts in scRNA-seq data, we use the negative binomial distribution. We assume that:

$$C'_{gc} \mid T_c = t,\, B_{gt},\, \Lambda_{gt},\, \lambda_0,\, p_g, \rho_g, \rho_0 \sim NB\left(\frac{1-p_g}{p_g}\left(\rho_0\lambda_0 + \rho_{gt}\Lambda_{gt}B_{gt}\right),\, p_g\right), \tag{4}$$

where ρ_g denotes cross-platform factor accounting for differences in sequencing depth between the two techniques. Here, we distinguish between gene-dependent factors, different for each marker gene (ρ_g) and a shared factor for non-markers (ρ_0).

The model's parameters inference

For inferring the hidden variables, we use the Metropolis-within-Gibbs sampler (Methods). Sampling steps for ST-Assign are provided in Algorithm 1. In the Algorithm 1, variables correspond to the above-described random variables in the model, and additionally, variable J corresponds to the number of iterations.

Algorithm 1 Metropolis-Hastings within Gibbs Sampler of ST-Assign parameters

Require: B, C, C' J
procedure MH-WITHIN-GIBBS(B, C, C', J)
  for $j \leftarrow 1$ to J **do**
    Sample $\pi^{(j)} \sim p(\pi \mid \text{MB}(\pi^{(j-1)}))$
    Sample $Z^{(j)} \sim p(Z \mid \text{MB}(Z^{(j-1)})$
    $\Theta^{(j)} \leftarrow$ MH-SINGLE-STEP($\Theta^{(j-1)}$)
    $[M^{(j)}, N^{(j)}] \leftarrow$ MH-SINGLE-STEP($M, N \mid M^{(j-1)}, N^{(j-1)}$)
    $\Lambda^{(j)} \leftarrow$ MH-SINGLE-STEP($\Lambda^{(j-1)}$)
    $\lambda_0^{(j)} \leftarrow$ MH-SINGLE-STEP($\lambda_0^{(j-1)}$)
    $p^{(j)} \leftarrow$ MH-SINGLE-STEP($p^{(j-1)}$)
    $\rho_g \leftarrow$ MH-SINGLE-STEP($\rho_g^{(j-1)}$)
    $\rho_0 \leftarrow$ MH-SINGLE-STEP($\rho_0^{(j-1)}$)
    Sample $T_c^{(j)} \sim p(T_c \mid \text{MB}(T_c^{(j-1)}))$
    Sample $D^{(j)} \sim p(D \mid \text{MB}(D^{(j-1)}))$
  end for
  return $\pi, Z, \Theta, \Lambda, \lambda_0, p, N, M, T, D, \rho, \rho_0$
end procedure

procedure MH-SINGLE-STEP($x^{(j-1)}$)
  Draw proposal sample $x^* \sim q(. \mid x^{(j-1)})$
  Evaluate acceptance probability r
  Generate $u \sim U[0,1]$
  if $u \leq r$ **then**
    Accept proposed sample: $x^{(j)} \leftarrow x^*$
  else
    Reject proposed sample and replicate previous sample: $x^{(j)} \leftarrow x^{(j-1)}$
  end if
  return $x^{(j)}$
end procedure

Below, we provide the model's equations for full conditional distributions and proposal dstributions used in the Algorithm's implementation.
In the case of the parameters D, T, π, and Z, we are in possession of the explicit formula to sample from. In the case of the variables: Λ, λ_0, ρ, ρ_0 p, N, M, Θ only functions $g(\cdot)$ proportional to the full conditional distributions are known (Methods). These functions are used to compute the acceptance ratios (Methods) while performing accept-reject Metropolis-Hastings steps. To avoid numerical errors, all variables are sampled in the log-space.

Sampling T_c**.**
For a selected cell c:

$$p(T_c \mid C'_{:c}, B, D_c, \Lambda, \lambda_0, p, \rho) \propto p(T_c \mid D) \prod_{g \in \text{Genes}} p(C'_{gc} \mid \Lambda, \lambda_0, p_g, B, T_c, \rho_g, \rho_0) =$$

$$\mathrm{Cat}(T_c \mid D) \prod_{g \in \text{Genes}} \mathrm{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_g \Lambda_{gT_c} B_{gT_c} + \rho_0 \lambda_0), p_g\right) =$$

$$D_{t_1}^{[T_c=t_1]} D_{t_2}^{[T_c=t_2]} \dots D_{t_N}^{[T_c=t_N]} \prod_{g \in \text{Genes}} \mathrm{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_g \Lambda_{gT_c} B_{gT_c} + \rho_0 \lambda_0), p_g\right).$$

Technically, to sample from the Categorical distribution, while avoiding leaving the log-space, we use a technique called the Gumbel-max trick (7, 8), which involves generating random numbers from the Gumbel distribution and performing a softmax operation on the log-probabilities.

Sampling D.
We leverage the conjugate prior property of the Dirichlet distribution with respect to the categorical distribution and we employ the Categorical-Dirichlet model.

$$\begin{aligned} p(D \mid T_:, d) \propto \mathrm{Dirichlet}(D \mid d) \prod_{c \in \text{Cells}} \mathrm{Categorical}(T_c \mid D) \; = \\ \mathrm{Dirichlet}(D_{t_1}, D_{t_2}, \dots, D_{t_N} \mid d_{t_1}, d_{t_2}, \dots, d_{t_N}) \prod_{c \in \text{Cells}} \mathrm{Categorical}(T_c \mid D_{t_1}, D_{t_2}, \dots, D_{t_N}) \propto \\ \prod_{j=1}^{t_N} D_{t_j}^{d_{t_j}-1} \prod_{c \in \text{Cells}} \prod_{j=1}^{t_N} D_{t_j}^{\mathbb{1}\{T_c=t_j\}} = \prod_{j=1}^{t_N} D_{t_j}^{d_{t_j}-1+\sum_{c \in \text{Cells}} \mathbb{1}\{T_c=t_j\}} \end{aligned} \tag{5}$$

Therefore,

$$p(D \mid T_:, d) = \mathrm{Dirichlet}(d'_{t_1}, d'_{t_2}, \dots, d'_{t_N}),$$

where:

$$d'_{t_j} = d_{t_j} + \sum_{c \in \text{Cells}} \mathbb{1}\{T_c = t_j\}.$$

Full conditional distribution for ρ_g.
For a gene g:

$$p(\rho_g \mid C'_{g:}, \Lambda, \lambda_0, p_g, B, T, \rho_0) \propto \prod_{c \in \text{Cells}} \mathrm{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_g \Lambda_{gT_c} B_{gT_c} + \rho_0 \lambda_0), p_g\right).$$

Full conditional distribution for ρ_0.

$$p(\rho_0 \mid C', \Lambda, \lambda_0, p_g, B, T, \rho_g) \propto \prod_{\substack{g \in \text{Genes} \\ c \in \text{Cells}}} \mathrm{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g}(\rho_g \Lambda_{gT_c} B_{gT_c} + \rho_0 \lambda_0), p_g\right).$$

Full conditional distribution for [N_s, M_s].
Recall that the relationship between the variables N_s and M_n is expressed by the following formula:

$$N_s = \sum_{t \in \text{Types}} M_{st}, \tag{6}$$

therefore these variables are updated together in one block. For a selected spot s:

$$p(N_s, M_{s:}, \mid C_{:s}, B, \Theta_{s:}, \Lambda, p, l_s) \propto p(N_s \mid l_s) p(M_{s:} \mid N_s, \Theta_{s:}) \prod_{g \in \text{Genes}} p(C_{gs} \mid \Lambda_{g:}, \lambda_0, B_{g:}, M_{s:}, p)$$

$$\mathrm{Pois}(N_s \mid l_s)\, \mathrm{Multinom}\,(M_{s:} \mid N_s, H_{s:}) \prod_{g \in \text{Genes}} \mathrm{NB}\left(C_{gs} \mid \frac{1-p_g}{p_g} \sum_t M_{st}(\lambda_0 + B_{gt}\Lambda_{gt}), p_g\right). \tag{7}$$

Sampling π_{st}.

$$P\left(\pi_{st} \mid Z_{st}\right) \propto \mathrm{Beta}\left(\pi_{st} \mid \frac{\alpha}{t_N}, 1\right) \mathrm{Bernoulli}\left(Z_{st} \mid 1, \pi_{st}\right) = \mathrm{Beta}\left(\pi_{st} \mid \frac{\alpha}{t_N} + Z_{st}, 2 - Z_{st}\right) \tag{8}$$

Sampling Z_{st}.

As Z_{st} is a discrete, binary random variable, it suffices to consider its two possible values, 0 and 1.

$$\begin{gathered}
P(Z_{st} \mid \theta_{st}, \pi_{st}) \propto P(Z_{st} \mid \pi_{st}) P(\theta_{st} \mid Z_{st}) = \text{Bernoulli}(Z_{st} \mid 1, \pi_{st}) \Gamma(\theta_{st} \mid a, b)^{Z_{st}} \Gamma(\theta_{st} \mid a_0, b_0)^{1-Z_{st}} \\
P(Z_{st} = 1 \mid \pi_{st}, \theta_{st}) \propto \pi_{st} \theta_{st}^{a-1} e^{-b\theta_{st}} = A \\
P(Z_{st} = 0 \mid \pi_{st}, \theta_{st}) \propto (1-\pi_{st}) \theta_{st}^{a_0-1} e^{-b_0\theta_{st}} = B \\
P(Z_{st} = 1 \mid \pi_{st}, \theta_{st}) = \frac{A}{A+B} \quad P(Z_{st} = 0 \mid \pi_{st}, \theta_{st}) = 1 - \frac{A}{A+B} = \frac{B}{A+B}
\end{gathered} \tag{9}$$

Full conditional distribution for $\Theta_{s:}$. For a selected spot s:

$$p(\Theta_{s:} \mid N_s, M_{s:}, Z_{s:}) \propto \text{Multinom}\,(M_{s:} \mid N_s, H_{s:}) \prod_{t \in \text{Types: } Z_{st} \neq 0} \Gamma(\theta_{st} \mid a, b) \prod_{t \in \text{Types: } Z_{st} = 0} \Gamma(\theta_{st} \mid a_0, b_0).$$

Full conditional distribution for Λ_{gt}.

$$\begin{gathered}
p\left(\Lambda_{gt} \mid C_{gt}, C'_{gt}, B_{g:}, \Lambda_{g:}, \lambda_0, p_g, M, \rho, \rho_0, T\right) \propto \\
\prod_{s \in \text{Spots}} p(C_{gs} \mid M_{s:}, \Lambda_{g:}, \lambda_0, p_g, B_{g:}) \prod_{c \in \text{Cells: } T_c = t} p(C'_{gc} \mid \Lambda_{gt}, \lambda_0, p_g, B_{gt}, \rho_g, \rho_0) = \\
\prod_{s \in \text{Spots}} \text{NB}\left(C_{gs} \mid \frac{1-p_g}{p_g} \sum_{t'} M_{st'} (\Lambda_{gt'} B_{gt'} + \rho_0 \lambda_0), p_g\right) \\
\prod_{c \in \text{Cells: } T_c = t} \text{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g} (\rho_g \Lambda_{gt} B_{gt} + \rho_0 \lambda_0), p_g\right).
\end{gathered} \tag{10}$$

Full conditional distribution for λ_0.

$$\begin{gathered}
p\left(\lambda_0 \mid C, C', B, \Lambda, p, M, T, \rho, \rho_0\right) \propto \\
\prod_{g \in \text{Genes}} \left[\prod_{s \in \text{Spots}} p(C_{gs} \mid M_{s:}, \Lambda_{g:}, \lambda_0, p_g, B_{g:}) \prod_{c \in \text{Cells}} p(C'_{gc} \mid \Lambda_{gT_c}, \lambda_0, p_g, B_{gT_c}, \rho_g, \rho_0) \right] = \\
\prod_{g \in \text{Genes}} \left[\prod_{s \in \text{Spots}} \text{NB}\left(C_{gs} \mid \frac{1-p_g}{p_g} \sum_{t} M_{st} (\Lambda_{gt} B_{gt} + \lambda_0), p_g\right) \right. \\
\left. \prod_{c \in \text{Cells}} \text{NB}\left(C'_{gc} \mid \frac{1-p_g}{p_g} (\rho_g \Lambda_{gT_c} B_{gT_c} + \rho_0 \lambda_0), p_g\right) \right].
\end{gathered} \tag{11}$$

Full conditional distribution for p_g. For a gene g:

$$\begin{gathered}
p\left(p_g \mid C_{g:}, C'_{g:}, B_{g:}, \Lambda_{g:}, \lambda_0, M, T\right) \propto \\
\prod_{s \in \text{Spots}} p(C_{gs} \mid M_{s:}, \Lambda_{g:}, \lambda_0, p_g, B_{g:}) \prod_{c \in \text{Cells}} p(C'_{gc} \mid \Lambda_{gt}, \lambda_0, p_g, B_{gt}, T_c, \rho_g, \rho_0).
\end{gathered} \tag{12}$$

The proposal distributions are analogous to proposal distributions used in Celloscope (6) accept for N_s and $M_{s:}$ variables, for which new values have to be proposed together due to the relationship between them (Equations 6 and 7).

Proposal distribution and Hastings ratio for [N_s, M_s]. We propose a three-step procedure for proposing new values of $M_{s:}$ and N_s. First, to sample a new value N_s^* for the variable N_s, we choose the ceiling of the truncated normal distribution as the proposal distribution (similarly as for Celloscope (6)):

$$N_s^* \mid N_s^{\text{prev}} \sim \lceil T \rceil,\ T \sim \text{TNorm}(N_s^{\text{prev}}, \sigma_N),$$

where σ_N controls step-size and N_s^{prev} is the value from the previous iteration. Let $f_{\lceil \text{TNorm} \rceil}(\cdot \mid N_s^{\text{prev}}, \sigma_N)$ denote the density

of $\lceil T \rceil$. Second, we sample a random vector $\epsilon \sim N(0, \sigma_M \mathbb{I})$, where σ_M controls the step-size. Let us denote

$$\mu_s(M_{s:}, \epsilon) = \frac{M_{s:} + \epsilon_t}{\sum\limits_{t \in \text{Types}} (M_{st} + \epsilon_t)}.$$

Third, we sample new values $M^*_{s:}$ of the variable $M_{s:}$ from the multinomial distribution with the number of trials parameter equal to N^*_s and the event probability parameter computed as a normalized cell-type composition from the previous iteration perturbed with ϵ:

$$M^*_{s:} \sim \text{Multinomial}(N^*_s, \mu_s(M^{\text{prev}}_{s:}, \epsilon)).$$

To sum up, the procedure to acquire new values for $[N_s, M_{s:}]$ is as follows:

1. Draw a proposal sample $N^*_s \sim f_{\lceil \text{TNorm} \rceil}(\cdot \mid N^{\text{prev}}_s, \sigma_N)$.

2. Draw $\epsilon \sim \text{N}(0, \sigma_M \mathbb{I})$.

3. Draw $M^*_{s:} \sim \text{Multinomial}(N^*_s, \mu_s(M^{\text{prev}}_{s:}, \epsilon))$.

Finally, the Hastings is computed:

$$\frac{f_{\lceil \text{TNorm} \rceil}(N^{\text{prev}}_s \mid N^*_s, \sigma_N)}{f_{\lceil \text{TNorm} \rceil}(N^*_s \mid N^{\text{prev}}_s, \sigma_N)} \frac{f_{\text{Multinom}}\left(M^{\text{prev}}_{s:} \mid N^{\text{prev}}_s, \mu_s(M^*, \epsilon^{\text{prev}})\right)}{f_{\text{Multinom}}\left(M^*_s \mid N^*_s, \mu_s(M^{\text{prev}}_s, \epsilon^*)\right)},$$

where f_{Multinom} denotes the probability mass function of the multinomial distribution, ϵ^* and ϵ^{prev} the values of ϵ, which perturbed the event probability parameter in the current and previous iteration, respectively.

Simulation study

To demonstrate the accurate performance of ST-Assign, we conducted a study using simulated data. We explored six different data simulation scenarios that focused on various aspects of the model. These scenarios differed in terms of the average gene expression level (Λ parameter), ranging from decreased Λ through moderate Λ to increased Λ. By default, the larger the average simulated Λ, the larger we set the cross-platform effect ρ. For each Λ *mode*, we also simulated datasets assuming no cross-platform effects ($\rho = 1$). For each simulation setting, we generated 10 replicates of simulated data and evaluated the model

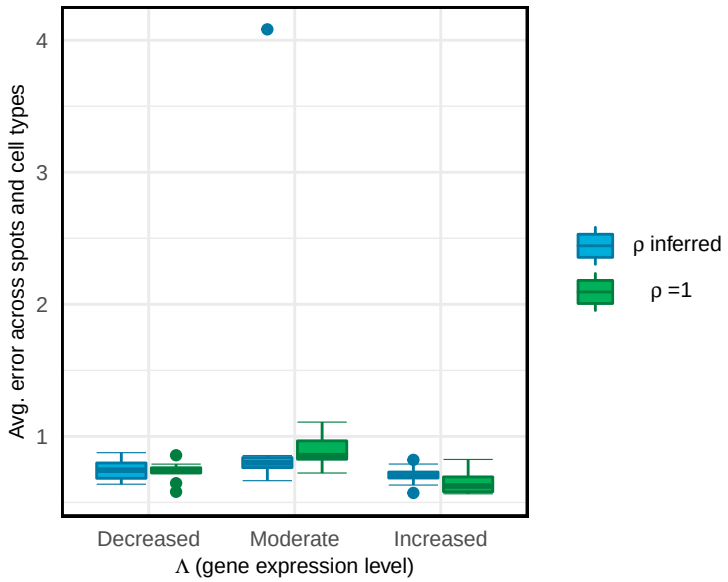

Fig. 2. Excellent performance of ST-Assign on simulated data. A Box-plots represent distributions of average absolute error (y-axis; computed using Equation 13) for different settings of ρ parameter (colors) and data simulation scenarios (x-axis).

performance of the inference of hidden variables. Specifically, to assess the accuracy, we computed the average absolute error

across all spots for each replicate k using the following equation:

$$\epsilon_k = \frac{\sum_{i=1}^{s_M} \sum_{j=1}^{t_N} |M_{ij}^{(k)} - \hat{M}_{ij}^{(k)}|}{t_N s_M}. \tag{13}$$

In this equation, $M_{ij}^{(k)}$ represents the true value of the number of cells of type j in spot i for the k-th simulated dataset, and $\hat{M}_{ij}^{(k)}$ is the number of cells of type j in spot i for the k-th dataset estimated by ST-Assign.
We observed excellent performance of ST-Assign in marker-gene-based cell type deconvolution (Figure 2), with an average error of approximately 0.5 cells across all scenarios. The error rarely exceeded one cell. However, we observed one outlier in the case of moderate Λ. In this particular setup, the error was significantly higher, reaching around 4 cells. Despite this one outlier, overall, the error was very low. The excellent performance of the model was confirmed in another task of assigning cell types to single cells. In this task, ST-Assign demonstrated 100% accuracy for all cells in all the simulated dataset scenarios we considered.

Applying ST-Assign on mouse brain data

To demonstrate the usefulness of ST-Assign model, we conducted a case study on mouse brain data, integrating data from two independent sources:

(i) spatial transcriptomics data from the anterior mouse brain (sagittal section) from 10x Genomics (9),

(ii) single-cell RNA-seq data from an adolescent mouse brain (10).

We applied ST-Assign for simultaneous cell-type annotation in dataset (ii) and cell-type mixtures decomposition in dataset (i). Specifically, the model was used to obtain the number of cells of each type in ST spots and assign cell types to individual cells in scRNA-seq data.

Analysed data. We utilized ST data obtained from sagittal sections of the anterior mouse brain provided by 10x Genomics. The primary objective of this case study was to investigate the feasibility of simultaneously decomposing ST cell-type mixtures and annotating single cells according to marker genes without a reference in the form of an already analyzed scRNA-seq dataset with annotations. To achieve this goal, we examined a smaller yet representative subset of spots whose cell type compositions are already known, thanks to the analysis performed with Celloscope.
We considered 893 spots. Across them, a minimum of 2, on average 26.2, and a maximum of 71 genes had non-zero expression. The minimal total expression per spot was 2, maximal was 9150, while an average was 319.6. We employed a scRNA-seq dataset that covered all brain regions of an adolescent mouse brain (10). We obtained the gene expression data from `http://mousebrain.org/adolescent/`, which also featured information on cell types and their specific marker genes. In this case study, we restricted our analysis to the major cell types previously identified by Celloscope in the chosen subset of spots (Table 3). We also randomly chose subsets of cells to analyze (Table 3). Across cells, a minimum of 8, on average 89, and a maximum of 122 genes had non-zero expression. The minimal total expression per cell was 13, maximal was 13478, while an average total expression was 979.6.

Cell type	Cluster name in (10)	No. of cells in (10)	No. of cells we took
Inhibitory (GABAergic)	MSN1, MSN2, MSN3, MSN4, MSN5, MSN6	5691	1000
Excitatory (glutamatergic) neurons	TEGLU1, TEGLU2, TEGLU3, TEGLU9, TEGLU10, TEGLU11, TEGLU20	5997	1000
Astrocytes	ACTE1, ACTE2	6947	1000
Microglia	MGL1, MGL2	4090	1000
Oligodendrocytes	MOL1	8715	1000
Chorid plexus epithelial cells	CHOR	458	450

Table 3. In (10), the measured single cells were clustered according to gene expression. These clusters, interpreted as cell types or sub-types, were annotated based on deferentially expressed genes. Cell types considered in this study, their respective clusters in (10), the number of cells in the respective clusters in (10), and the number of cells randomly chosen to be given as input to ST-Assign.

The scRNA-seq dataset analyzed in this study exhibited a significantly higher number of zero entries per cell on average compared to spots in ST data, with 110.8 and 48 zero entries on average per cell and spot, respectively. The main reason

behind the observed differences in the number of zero entries should be attributed to the fact that ST data consists of *mini-bulk* measurements and as such, may exhibit less technical noise compared to scRNA-seq data. Additionally, this difference may arise from various technical and biological factors, including differences in sequencing depth and biological variability.
In this case study, 137 genes were selected and modeled as markers. The marker gene sets used previously while applying Celloscope (6) were extended by incorporating genes indicated as markers in (10) for the considered cell types.

Results of cell type identification. The results of the case study on mouse brain data demonstrate the potential of ST-Assign to integrate heterogeneous data sources and facilitate a comprehensive analysis of complex biological systems. The obtained cell type decomposition (Figure 3) is in agreement with previous results acquired with Celloscope (6). This indicates that even with the additional difficult task of assigning types to single cells, ST-Assign performs similarly well in the task of cell type deconvolution in ST data as Celloscope.

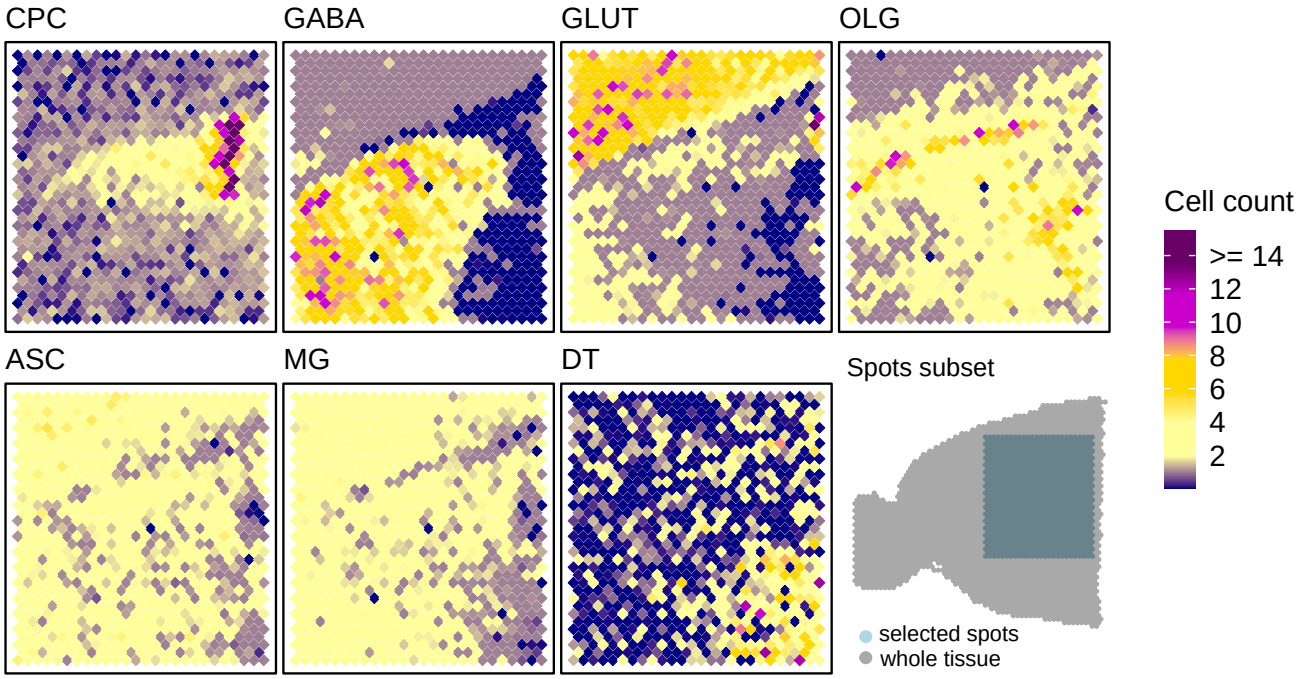

Fig. 3. Spatial composition of cell types in the anterior part of the mouse brain (sagittal section) as inferred by ST-Assign. Heatmaps show the abundance of selected cell types, with dark blue indicating absence, yellow moderate occurrence, and magenta dominance. The cell types are as follows: CPC – Choroid plexus epithelial cells, GABA - GABAergic neurons, GLUT – Glutamatergic neurons, OLG – Oligodendrocytes, ASC – Astrocytes, MG – Microglia, DT – dummy type. The bottom-right inset shows the subset of spots analyzed.

As part of the ST-Assign inference, we performed cell type annotation in the accompanying single-cell dataset. The obtained results are consistent with the clustering performed in (10): the majority of cells were annotated by ST-Assign to their corresponding cell types, as defined by the clustering in (10) (Figure 4A). However, a portion of cells, especially among the Glutamatergic neurons (7% of cells), were annotated as the dummy type. Additionally, two cells that were assigned to Choroid plexus epithelial cells and Microglia respectively in (10), were annotated as Astrocytes by ST-Assign. This discrepancy could provide an interesting avenue for further investigation. In particular, it is possible that some of the cells were not correctly classified to types by the algorithm used by (10).
Furthermore, we conducted an independent clustering of the single-cell dataset under consideration (Table 3) using the Seurat package (11) (Figure 4B) and colored the cells according to the ST-Assign results. In general, the t-SNE clusters were similar to those obtained previously (10). Intriguing, a cluster dominated by GABAergic neurons (colored in beige) exhibited a visibly distinct subcluster annotated as dummy type according to ST-Assign (colored in magenta). Importantly, a group of similar cells assigned to the dummy type could come from an additional, currently not considered cell subtype.

Identification of an additional cell type based on examining results obtained for dummy type. The heatmap representing the dummy type in Figure 3 showed that cells clustered in the bottom right quadrant, again suggesting the possibility of another cell type being present in this region that was previously overlooked. To investigate further, we added an additional 1000 cells annotated as Peptigeneric neurons in (10) to the single-cell dataset. Both single-cell and ST dataset were expanded with 10 genes that were indicated as markers for Peptigeneric neurons in (10). This addition did not alter the cell-type composition obtained previously (Figure 5). However, it did reveal the presence of Peptigeneric neurons in the expected bottom right quadrant, i.e., the same region where solely the dummy type was localized previously.

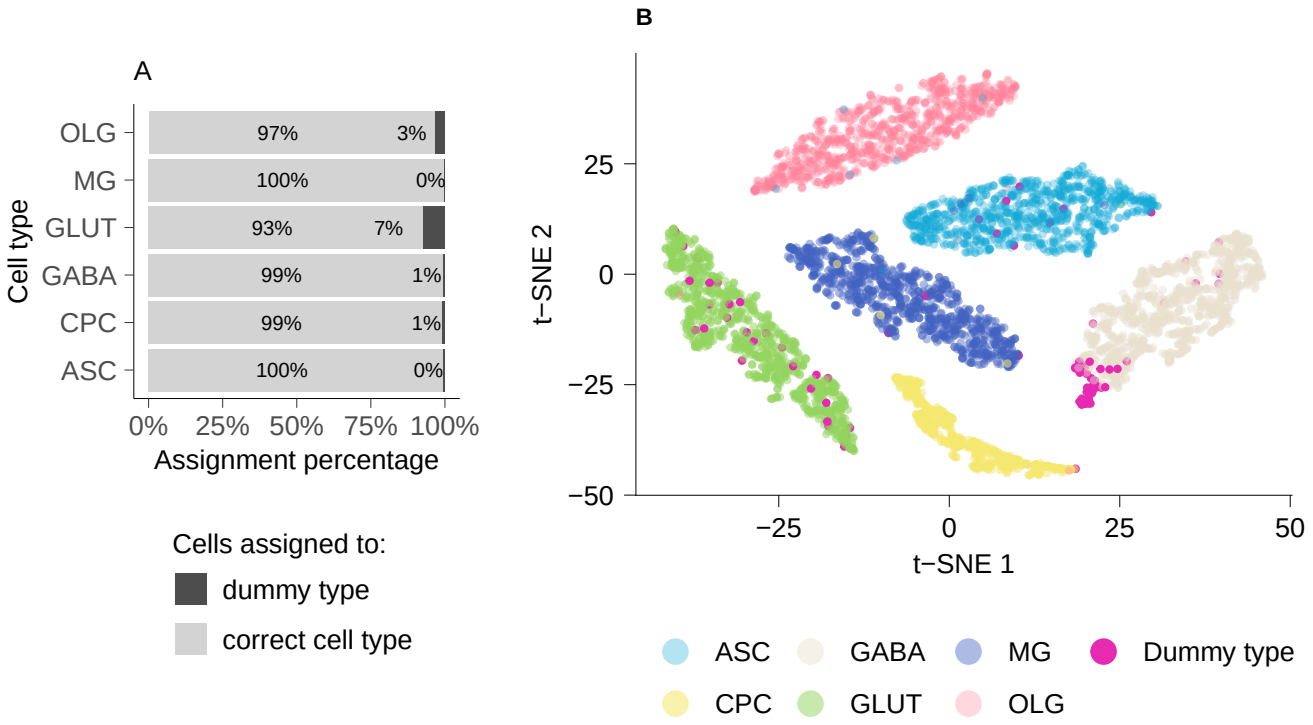

Fig. 4. **A.** Barplots demonstrate the level of agreement between ST-Assign's cell type annotation and the cell types defined by clusters in (10). Light gray indicates cells assigned to the cell type in correspondence with clusters in (10). **B.** An independent clustering results using t-SNE visualization, with colors indicating the ST-Assign cell type annotation for each cell.

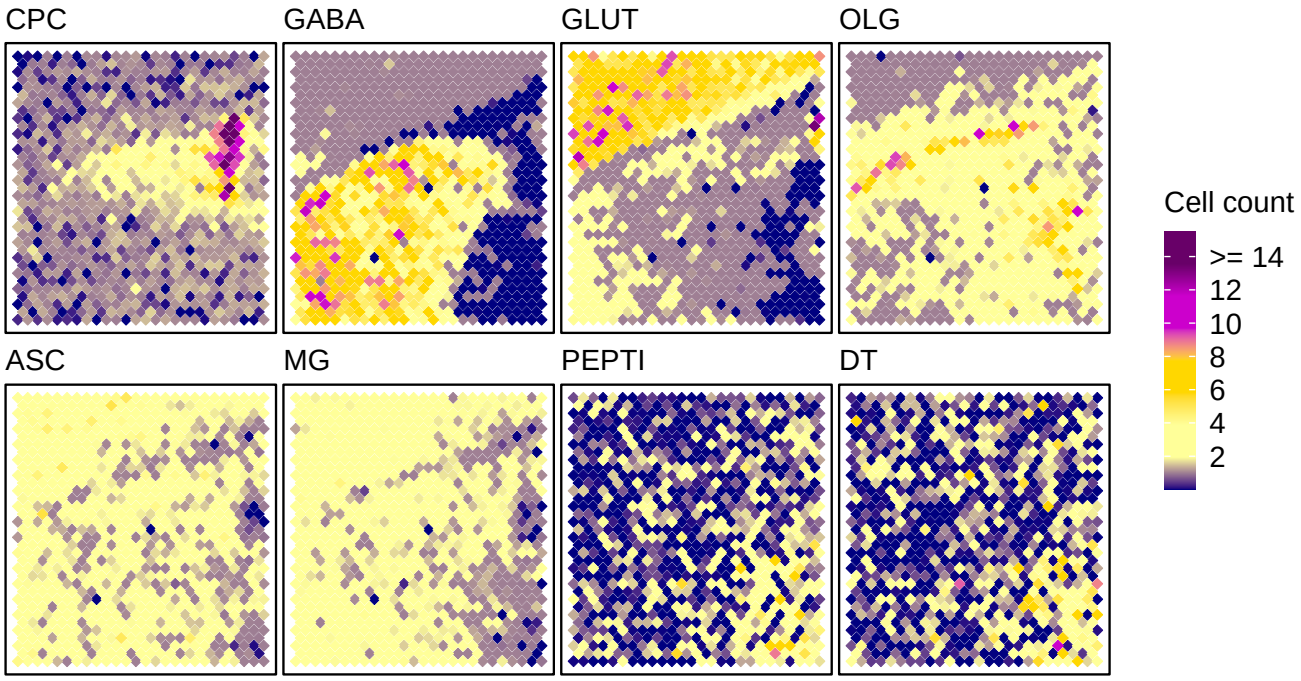

Fig. 5. **Spatial composition of cell types in the anterior part of the mouse brain (sagittal section) as inferred by ST-Assign including an additional cell type compared to Figure 3.** Heatmaps indicate the abundance of selected cell types, with dark blue representing absence, yellow representing moderate occurrence, and magenta indicating dominance. The cell types displayed are the same as in Figure 3, with the addition of Peptigeneric neurons (PEPTI).

Discussion

We presented a novel probabilistic ST-Assign model for concurrent cell-type annotation in scRNA-seq data and cell-type mixture decomposition in ST data. To infer model parameters, we used the Metropolis-within-Gibbs algorithm. We demonstrated

the usefulness of ST-Assign on mouse brain data.
Nevertheless, integrating data from two sources may pose challenges due to technical discrepancies. Cross-platform effects, such as unequal effectiveness with which genes are captured or only partial overlap between sequenced cell types and genes, must be addressed. To this end, we introduced dedicated random variables. Specifically, we added cross-platform factor ρ that alleviates unequal sequencing depth, with which genes are captured.
Once again, as in the case of Celloscope, we introduced the so-called *dummy type*, which accounts for cells beyond the scope of considered cell types. Importantly, we leveraged it to expand the considered set of cell types. Moreover, by design, ST-Assign considers only a limited subset of genes (expected to be highly expressed); therefore, it is easy to check that those considered genes are expressed in the two considered data sets.
Compared to Celloscope, the new model is supplied with additional information on marker gene expression measured in single cells. This new information is expected to boost the performance of inferring gene expression profiles for each cell type, which was challenging in the case of Celloscope due to the presence of cell-type mixtures in spots.
Future research includes applying ST-Assign to larger datasets and investigating discrepancies in cell type annotations. However, the results obtained in the mouse brain case study already indicate ST-Assign as a promising tool for integrating ST and scRNA-seq data for analyzing cell type composition in tissues.

Methods

For inferring the hidden variables we use the Metropolis-within-Gibbs sampler, a Markov chain Monte Carlo (MCMC) algorithm which is a combination of the Gibbs Sampler and the Metropolis-Hastings algorithm (Algorithm 1). Suppose the graphical model contains variables $x_1, \ldots, x_n$. Let $MB(x_i)$ denote the Markov Blanket of x_i, i.e., the set containing its parents, children, and co-parents. Then:

$$p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = p\left(x_i \mid \mathrm{MB}(x_i)\right), \tag{14}$$

i.e., the conditional distribution of x_i given the values of all other variables equals the conditional distribution given the values of the variables from its Markov blanket.
The iterative sampling procedure is as follows: firstly, starting values for $x_1^{(0)}, \ldots, x_n^{(0)}$ are randomly initialised, then, in a given iteration j ($j = 1, 2, \ldots, J$, where J denotes the number of iterations), we take every variable x_i, $i = 1, ..., n$ one by one, in some arbitrary ordering and for each given variable x_i, its value is sampled given the values of the variables from the Markov blanket $\mathrm{MB}\left(x_i^{(j-1)}\right)$ from the previous iteration. As a result each x_i is updated iteratively, up until convergence. There are two options for updating the value of x_i in the j-th iteration:

1. If $p\left(x_i \mid \mathrm{MB}\left(x_i^{(j-1)}\right)\right)$ can be expressed in a closed form, a new value x_i^* is sampled directly from $p\Big(x_i \mid \mathrm{MB}\left(x_i^{(j-1)}\right)\Big)$ and $x_i^{(j)}$ is set to x_i^*.

2. In case we only known a function g proportional to $p\left(x_i \mid \mathrm{MB}(x_i)\right)$

$$p\left(x_i \mid MB\left(x_i^{(j-1)}\right)\right) \propto f(x_i),$$

   we perform a single Metropolis-Hastings accept-reject step (MH-single-step procedure in Algorithm 1). A candidate value x_i^* is sampled from a predefined proposal distribution $q\left(\cdot \mid x_i^{(j-1)}\right)$, and then either accepted with probability given by

$$r = \min\left(1, \frac{g\left(x^*\right) q\left(x^{(j-1)} \mid x^*\right)}{g\left(x^{(j-1)}\right) q\left(x^* \mid x^{(j-1)}\right)}\right) \tag{15}$$

   and $x_i^{(j)} \leftarrow x_i^*$, or the previous value is held: $x_i^{(j)} \leftarrow x_i^{(j-1)}$. After updating x_i, we immediately use the new value for sampling other variables.

Funding. This project has received funding from the Polish National Science Centre PRELUDIUM grant no 2021/41/N/ST6/03619. E.S. acknowledges the support from the Polish National Science Centre SONATA BIS grant No. 2020/38/E/NZ2/00305.

Ethics approval and consent to participate. Not applicable.

Competing interests. Projects in Szczurek lab are co-funded by Merck Healthcare. Otherwise, the authors declare that they have no competing interests.

Bibliography

1. Efthymia Papalexi and Rahul Satija. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35–45, 2018.
2. Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.
3. Allen W. Zhang, Ciara O'Flanagan, Elizabeth A. Chavez, Jamie L. P. Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, Daniel Lai, Anja Mottok, Clementine Sarkozy, Lauren Chong, Tomohiro Aoki, Xuehai Wang, Andrew P. Weng, Jessica N. McAlpine, Samuel Aparicio, Christian Steidl, Kieran R. Campbell, and Sohrab P. Shah. Probabilistic cell-type assignment of single-cell rna-seq for tumor microenvironment profiling. *Nature Methods*, 16(10):1007–1015, Oct 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0529-1.
4. Hongyu Guo and Jun Li. scSorter: assigning cells to known cell types according to marker genes. *Genome Biology*, 22(1):69, Feb 2021. doi: 10.1186/s13059-021-02281-7.
5. Patrik Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, Jose Fernandez Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Costea, Pelin Akan Sahlén, Jan Mulder, Olaf Bergmann, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353:78–82, 07 2016.
6. Agnieszka Geras, Shadi Darvish Shafighi, Kacper Domżał, Igor Filipiuk, Łukasz Rączkowski, Paulina Szymczak, Hosein Toosi, Leszek Kaczmarek, Łukasz Koperski, Jens Lagergren, Dominika Nowis, and Ewa Szczurek. Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics data. *Genome Biology*, 24(1):120, May 2023. doi: 10.1186/s13059-023-02951-8.
7. Ryan Adams. The gumbel-max trick for discrete distributions.
8. Tim. Answer to: How do i sample from a discrete (categorical) distribution in log space?
9. Satija Lab. *stxBrain.SeuratData: 10X Genomics Visium Mouse Brain Dataset*, 2019. R package version 0.1.1.
10. Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Häring, Emelie Braun, Lars E. Borm, Gioele La Manno, Simone Codeluppi, Alessandro Furlan, Kawai Lee, Nathan Skene, Kenneth D. Harris, Jens Hjerling-Leffler, Ernest Arenas, Patrik Ernfors, Ulrika Marklund, and Sten Linnarsson. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014.e22, Aug 2018. doi: 10.1016/j.cell.2018.06.021.
11. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 2021. doi: 10.1016/j.cell.2021.04.048.