Bayesian Statistical Modeling of Spatially Resolved Transcriptomics Data

Xi Jiang
xij@smu.edu

Follow this and additional works at: https://scholar.smu.edu/hum_sci_statisticalscience_etds

Part of the Applied Statistics Commons, Biostatistics Commons, and the Statistical Models Commons

Recommended Citation
Jiang, Xi, "Bayesian Statistical Modeling of Spatially Resolved Transcriptomics Data" (2023). Statistical Science Theses and Dissertations. 44.
https://scholar.smu.edu/hum_sci_statisticalscience_etds/44

This Dissertation is brought to you for free and open access by the Statistical Science at SMU Scholar. It has been accepted for inclusion in Statistical Science Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
BAYESIAN STATISTICAL MODELING OF
SPATIALLY RESOLVED TRANSCRIPTOMICS DATA
Approved by:

Dr. Guanghua Xiao
Professor in Peter O’Donnell Jr. School of Public Health, UTSW

Dr. Qiwei Li
Assistant Professor in Department of Mathematical Sciences, UTD

Dr. Daniel F. Heitjan
Professor in Department of Statistical Science, SMU & Peter O’Donnell Jr. School of Public Health, UTSW

Dr. Jing Cao
Professor in Department of Statistical Science, SMU

Dr. Chul Moon
Assistant Professor in Department of Statistical Science, SMU

Dr. Lin Xu
Assistant Professor in Peter O’Donnell Jr. School of Public Health, UTSW
BAYESIAN STATISTICAL MODELING OF SPATIALLY RESOLVED TRANSCRIPTOMICS DATA

A Dissertation Presented to the Graduate Faculty of the Dedman College Southern Methodist University

in Partial Fulfillment of the Requirements for the degree of Doctor of Philosophy with a Major in Biostatistics by

Xi Jiang

B.S., Mathematics and Applied Mathematics, University of Chinese Academy of Sciences, China

December 16, 2023
Acknowledgments

I would like to express my deepest gratitude to the individuals and institutions that have played a pivotal role in the completion of my Ph.D. journey. First and foremost, I extend my sincere appreciation to advisors, Dr. Guanghua Xiao and Dr. Qiwei Li. Their knowledge, expertise, unwavering support, and patience have been invaluable throughout my research. Their guidance has not only shaped my research but has also been pivotal in fostering my growth in my future career.

I am grateful to Dr. Lin Xu for providing me with various research ideas and shared his thoughts for many projects. I would also like to thank Dr. Xiaowei Zhan for his guidance through the STAr project. I am also thankful for all collaborators and colleagues at the Quantitative Biomedical Research Center at UT Southwestern. The exchange of ideas and collective efforts have enriched the quality and depth of my research. I am also grateful to the members of my thesis committee, Dr. Daniel Heitjan, Dr. Jing Cao, and Dr. Chul Moon, for their constructive feedback and thoughtful suggestions, which has contributed significantly to the refinement of this work.

My deep and sincere gratitude to my family for their encouragement, love, and understanding through my doctoral studies. I am especially grateful to my boyfriend, Dr. Hao Tian, for being my constant source of support and love. My sincere thanks also go to all my friends and fellow graduate students who provided a supportive community through this academic journey. Their companion has been a source of encouragement and inspiration.

Thank you all for being part of this incredible journey and having left an indelible mark on my Ph.D. experience.
Spatially resolved transcriptomics (SRT) quantifies expression levels at different spatial locations, providing a new and powerful tool to investigate novel biological insights. As experimental technologies enhance both in capacity and efficiency, there arises a growing demand for the development of analytical methodologies.

One question in SRT data analysis is to identify genes whose expressions exhibit spatially correlated patterns, called spatially variable (SV) genes. Most current methods to identify SV genes are built upon the geostatistical model with Gaussian process, which could limit the models’ ability to identify complex spatial patterns. In order to overcome this challenge and capture more types of spatial patterns, in Chapter 2, we introduce a Bayesian approach to identify SV genes via a modified Ising model. The key idea is to use the energy interaction parameter of the Ising model to characterize spatial expression patterns. We use auxiliary variable Markov chain Monte Carlo algorithms to sample from the posterior distribution with an intractable normalizing constant in the model. Simulation studies using both simulated and synthetic data showed that the energy-based modeling approach led to higher accuracy in detecting SV genes than those kernel-based methods. When applied to two real spatial transcriptomics datasets, the proposed method discovered novel spatial patterns that shed light on the biological mechanisms.
Spatial domain identification is another direction in SRT analysis, which enables the transcriptomic characterization of tissue structures and further contributes to the evaluation of heterogeneity across different tissue locations. Current spatial domain analysis of SRT data primarily relies on molecular information and fails to fully exploit the morphological features present in histology images, leading to compromised accuracy and interpretability. To overcome these limitations, in Chapter 3, we develop a multi-stage statistical method called iIMPACT. It includes a finite mixture model to identify and define spatial domains based on AI-reconstructed histology images and spatial context of gene expression measurements, and a negative binomial regression model to detect domain-specific spatially variable genes. Through multiple case studies, we demonstrated iIMPACT outperformed existing methods, confirmed by ground truth biological knowledge. These findings underscore the accuracy and interpretability of iIMPACT as a new clustering approach, providing valuable insights into the cellular spatial organization and landscape of functional genes within SRT data.

Most next-generation sequencing-based SRT techniques are limited to measuring gene expression in a confined array of spots, capturing only a fraction of the spatial domain. Typically, these spots encompass gene expression from a few to hundreds of cells, underscoring a critical need for more detailed, single-cell resolution SRT data to enhance our understanding of biological functions within the tissue context. Addressing this challenge, in Chapter 4, we introduce BayesDeep, a novel Bayesian hierarchical model that leverages cellular morphological data from histology images, commonly paired with SRT data, to reconstruct SRT data at the single-cell resolution. BayesDeep effectively model count data from SRT studies via a negative binomial regression model. This model incorporates explanatory variables such as cell types and nuclei-shape information for each cell extracted from the paired histology image. A feature selection scheme is integrated to examine the association between the morphological and molecular profiles, thereby improving the model robustness. We applied BayesDeep to two real SRT datasets, successfully demonstrating its capability to reconstruct
SRT data at the single-cell resolution. This advancement not only yields new biological insights but also significantly enhances various downstream analyses, such as pseudotime and cell-cell communication.
# TABLE OF CONTENTS

LIST OF FIGURES........................................................................................................... xiii

LIST OF TABLES .................................................................................................................. xx

CHAPTER

1 Introduction ...................................................................................................................... 1

1.1. Spatially Resolved Transcriptomics Technology .................................................. 1

1.2. Applications on Spatially Resolved Transcriptomics Data .................................. 2

1.2.1. Spatially variable gene detection ...................................................................... 2

1.2.2. Spatial domain identification .......................................................................... 4

1.2.3. Gene expression resolution enhancement ....................................................... 5

1.2.4. Others .............................................................................................................. 6

1.3. Advantages of Bayesian Modeling in Analyzing Spatially Resolved Transcriptomics Data ................................................................. 7

1.4. Overview of Projects ............................................................................................ 10

2 BOOST-MI: A Bayesian Modified Ising Model for Spatially Variable Gene Identification ........................................................................................................ 12

2.1. Background ............................................................................................................ 12

2.2. Model .................................................................................................................... 14

2.2.1. Data preparation .............................................................................................. 14

2.2.2. A brief review of the Ising model ................................................................... 17

2.2.3. A modified Ising model to identify spatially variable genes ......................... 18

2.3. Model Fitting ....................................................................................................... 20

2.3.1. MCMC algorithms ......................................................................................... 20
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.2. Posterior inference</td>
<td>21</td>
</tr>
<tr>
<td>2.4. Simulation Study</td>
<td>22</td>
</tr>
<tr>
<td>2.4.1. Data generative model</td>
<td>22</td>
</tr>
<tr>
<td>2.4.2. Prior and algorithm settings</td>
<td>23</td>
</tr>
<tr>
<td>2.4.3. Evaluation metrics</td>
<td>25</td>
</tr>
<tr>
<td>2.4.4. Model performance</td>
<td>26</td>
</tr>
<tr>
<td>2.5. Real Data Analysis</td>
<td>27</td>
</tr>
<tr>
<td>2.5.1. Case study on mouse olfactory bulb spatial transcriptomics dataset</td>
<td>27</td>
</tr>
<tr>
<td>2.5.2. Case study on human breast cancer spatial transcriptomics dataset</td>
<td>32</td>
</tr>
<tr>
<td>2.6. Discussion</td>
<td>35</td>
</tr>
<tr>
<td>3. iIMPACT: A Bayesian Integrative Approach for Spatial Domain Identification and Beyond</td>
<td>37</td>
</tr>
<tr>
<td>3.1. Background</td>
<td>37</td>
</tr>
<tr>
<td>3.2. Model</td>
<td>40</td>
</tr>
<tr>
<td>3.2.1. Overview</td>
<td>40</td>
</tr>
<tr>
<td>3.2.2. Data preparation</td>
<td>42</td>
</tr>
<tr>
<td>3.2.3. A Bayesian normal-multinomial mixture model to identify histology-based spatial domains</td>
<td>45</td>
</tr>
<tr>
<td>3.2.4. A generalized linear regression model to detect domain-specific spatially variable genes</td>
<td>49</td>
</tr>
<tr>
<td>3.3. Model Fitting</td>
<td>51</td>
</tr>
<tr>
<td>3.3.1. MCMC algorithms</td>
<td>51</td>
</tr>
<tr>
<td>3.3.2. Posterior inference</td>
<td>52</td>
</tr>
<tr>
<td>3.4. Simulation Study</td>
<td>53</td>
</tr>
<tr>
<td>3.4.1. Data generative model</td>
<td>53</td>
</tr>
<tr>
<td>3.4.2. iIMPACT and competing method settings</td>
<td>54</td>
</tr>
</tbody>
</table>
3.4.3. Model performance .................................................. 55

3.5. Real Data Analysis ..................................................... 55

3.5.1. Case study on human breast cancer 10x Visium dataset .... 55
3.5.2. Case study on human prostate cancer 10x Visium dataset ... 59
3.5.3. Case study on human ovarian cancer 10x Visium dataset .... 62
3.5.4. Case study on mouse visual cortex STARmap dataset ........ 64

3.6. Discussion ................................................................. 65

4 BayesDeep: A Bayesian Hierarchical Model for Gene Expression Reconstruction at Single-cell Resolution ............................................. 68

4.1. Background ............................................................... 68

4.2. Model ....................................................................... 70

4.2.1. Overview ............................................................... 72
4.2.2. Data preparation ...................................................... 73
4.2.3. A Bayesian negative binomial regression model with grouped observations ........................................... 75

4.3. Model Fitting ............................................................... 79

4.3.1. MCMC algorithms .................................................... 80
4.3.2. Posterior inference ................................................... 80
4.3.3. Predictive inference .................................................. 81

4.4. Simulation Study .......................................................... 82

4.4.1. Model validation on simulated data ............................... 82
4.4.2. Model validation on real spatially resolved transcriptomics data with spot resolution ....................................... 83
4.4.3. Model validation on real spatially resolved transcriptomics data with single-cell resolution .............................. 85

4.5. Real Data Analysis ........................................................ 86
4.5.1. Case study on human breast cancer 10x Visium dataset .......... 86
4.5.2. Case study on human prostate cancer 10x Visium dataset .......... 92
4.6. Discussion .................................................................................. 97

5 Conclusion and Future Directions .................................................. 99

APPENDIX

A Appendix of Chapter 2 ................................................................. 102
A.1. Full Details of Data Normalization ............................................. 102
A.2. Full Details of Data Dichotomization ........................................... 104
A.3. Full Details of MCMC Algorithms ............................................. 106
A.4. Sensitivity Analysis .................................................................... 108
A.5. Supporting Figures and Tables ................................................... 109

B Appendix of Chapter 3 ................................................................. 116
B.1. Full Details of MCMC Algorithms ............................................. 116
B.2. Sensitivity Analysis .................................................................... 120
B.3. Supporting Figures and Tables ................................................... 123

C Appendix of Chapter 4 ................................................................. 132
C.1. Full Details of MCMC Algorithms ............................................. 132
C.2. Simulated Data Generative Model ............................................. 134
C.3. Supporting Figures and Tables ................................................... 135

BIBLIOGRAPHY ............................................................................... 141
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The schematic diagram of the proposed BOOST-MI model</td>
<td>15</td>
</tr>
<tr>
<td>2.2</td>
<td>Common types of two-dimensional lattices. Red spots are the neighbors of the blue spot, while black spots are not.</td>
<td>18</td>
</tr>
<tr>
<td>2.3</td>
<td>Simulation study: The six spatial patterns used to generate the simulated data. (a) and (b) Two artificial patterns with the Gaussian and linear kernel; (c) An artificial pattern generated from the Gaussian process; (d) An artificial pattern with complete attraction pattern, i.e., $\theta \to \infty$; (e) and (f) Two real patterns constructed from the mouse olfactory bulb (MOB) and human breast cancer (BC) study. The boxplots of AUCs achieved by BOOST-MI, Ising, BOOST-GP, BinSpect, SPARK, and SpatialDE under different scenarios in terms of spatial pattern and zero-inflation setting.</td>
<td>24</td>
</tr>
<tr>
<td>2.4</td>
<td>Real data analysis on the mouse olfactory bulb (MOB) dataset: (a) The Venn diagram of SV genes identified by BOOST-MI, SPARK, and BinSpect-rank; (b) Distinct spatial expression patterns summarized on the basis of 734, 772, and 1,415 SV genes identified by BOOST-MI, SPARK, and BinSpect-rank; (c) Distinct spatial expression patterns summarized on the basis of 307 SV genes identified by BOOST-MI only; (d) Distinct spatial expression patterns summarized on the basis of 345 SV genes identified by SPARK only; (e) The associated hematoxylin and eosin (H&amp;E)-stained tissue slides of the analyzed MOB dataset; (f) The top 15 gene ontology (GO) terms in the GO enrichment analysis for those 734 SV genes identified by BOOST-MI, with the red dashed line indicating a significance level of 0.05; (g) The top six genes with the attraction pattern in terms of BF values.</td>
<td>29</td>
</tr>
<tr>
<td>2.5</td>
<td>Real data analysis: The boxplots of Moran’s I’s of SV genes detected by all three methods (i.e., BOOST-MI, SPARK, and BinSpect), each pair of methods, and each method alone for the (a) mouse olfactory bulb and (b) human breast cancer datasets.</td>
<td>31</td>
</tr>
</tbody>
</table>
2.6 Real data analysis on the human breast cancer (BC) dataset: (a) The Venn diagram of SV genes identified by BOOST-MI, SPARK, and BinSpect-rank; (b) Distinct spatial expression patterns summarized on the basis of 302, 293, and 3,278 SV genes identified by BOOST-MI, SPARK, and BinSpect-rank; (c) Distinct spatial expression patterns summarized on the basis of 174 SV genes identified by BOOST-MI only; (d) Distinct spatial expression patterns summarized on the basis of 165 SV genes identified by SPARK only; (e) The associated hematoxylin and eosin (H&E)-stained tissue slides of the analyzed BC dataset; (f) The top 15 gene ontology (GO) terms in the GO enrichment analysis for those 302 SV genes identified by BOOST-MI, with the red dashed line indicating a significance level of 0.05.

3.1 Workflow of iIMPACT: A. iIMPACT starts by combining and processing image profile from AI-reconstructed histology images, and geospatial and molecular profiles from SRT data (circled by dashed lines) to conduct the histology-based spatial domain identification. B. A Bayesian normal-multinomial mixture model with the Markov random field (circled by solid lines) is fitted for histology-based spatial domain identification. Based on the spatial domain identification results, biologically important cellular spatial organization can be characterized, including the domain-specific relative abundance of cell types and interactive zones (circled by dotted lines). C. Domain-specific SVGs are identified by a negative binomial (NB) regression model.

3.2 Simulation study: The boxplots of ARIs achieved by iIMPACT, SpaGCN, BayesSpace, stLearn, and Louvain under different scenarios in terms of sparsity and variability settings.

3.3 Real data analysis on human breast cancer dataset: A. H&E-stained image of the tissue section with ground truth labels from pathologists. B. Spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, with the number of clusters to be five. C. Estimates (posterior means and credible intervals) of domain-specific relative abundance of cell types for the seven cell types observed in the AI-reconstructed histology image. D. Interactive zones (black asterisk spots) defined by iIMPACT. E. Identified interactive zones (black asterisk spots) and other boundary areas of tumor and its adjacent Domain 3, and boxplots of gene expression richness for spots in the interactive zone and other boundaries. F. Gene enrichment analysis between SVGs detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and known breast cancer genes from the COSMIC database. G. Spatial expression patterns of two example SVGs, \textit{COX6C} and \textit{ELF3}, that were only detected by iIMPACT.
3.4 Real data analysis on human prostate cancer dataset: A. H&E-stained image of the tissue section. B. Spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to be five. C. Estimates (posterior means and credible intervals) of domain-specific relative abundance of cell types for the six cell types observed in the AI-reconstructed histology image. D. Interactive zones (black asterisk spots) defined by iIMPACT. E. Identified interactive zones (black asterisk spots) and other boundary areas of Domain 2 and Domain 3, and boxplots of gene expression richness for spots in the interactive zone and other boundaries. F. Gene enrichment analysis between SVGs detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and the known prostate cancer genes from the COSMIC database. G. Spatial expression patterns of two example SVGs, \( \text{EIF3E} \) and \( \text{TBL1XR1} \), that were only detected by iIMPACT.

3.5 Real data analysis on human ovarian cancer dataset: A. H&E-stained image of the tissue section. B. Spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to be five. C. Estimates (posterior means and credible intervals) of domain-specific relative abundance of cell types for the six cell types observed in the AI-reconstructed histology image. D. Interactive zones (black asterisk spots) defined by iIMPACT. E. Identified interactive zones (black asterisk spots) and other boundary areas of tumor and its adjacent Domain 5, and boxplots of gene expression richness for spots in interactive zone and other boundaries. F. Gene enrichment analysis between SVGs detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and the known ovarian cancer genes from the COSMIC database. G. Spatial expression patterns of two example SVGs, \( \text{BCL6} \) and \( \text{CHD4} \), that were only detected by iIMPACT.

3.6 Real data analysis on mouse visual cortex STARmap data: A. Layer structure of the tissue section from the original study. Spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to seven (the number of layers). Manually added square lattice grid when fitting iIMPACT is displayed with dashed lines. B. Interactive zones (black asterisk spots) defined by iIMPACT. C. Gene enrichment analysis between SVGs detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and genes functionally relevant to visual cortex for five layers. D. Spatial expression patterns of two example SVGs, \( \text{Deptor} \) and \( \text{Vamp3} \), that were only detected by iIMPACT.
4.1 Flowchart of the proposed BayesDeep: A. BayesDeep integrates the spot-resolution molecular profile $Y$ from NGS-based SRT data, the single-cell-resolution image profile $X$ from the paired AI-reconstructed histology image, and the spot-cell geospatial profile $G$ to recover gene expression at the single-cell resolution $\Theta$. B. The hierarchical formulation of the BayesDeep model, based on a Bayesian regularized negative binomial regression model with grouped observations. C. BayesDeep estimates the association between the single-cell-resolution molecular and image profiles $B$ and predicts the single-cell-resolution molecular profile $\Theta$. D. Several downstream analyses can be enhanced based on the availability of the single-cell-resolution molecular profile $\Theta$, including identifying distinct cell populations, elucidating the process of tumorigenesis via pseudotime analysis, and exploring the mechanisms of cell-cell communication.

4.2 Overview of model validation, including the validation settings and evaluation results in terms of the Pearson correlation coefficients $\rho$ between the actual and predicted gene expression for BayesDeep, TESLA, and Gaussian Process (GP), respectively. The validation is stratified into three distinct data: A. Simulated data at the spot resolution; B. Human breast cancer 10x Visium data at the spot resolution; C. Mouse visual cortex STARmap data at the single-cell resolution.

4.3 Real data analysis on the human breast cancer 10x Visium data: A. Manual annotation and nuclei identification results by applying HD-staining to the paired histology image; and the two selected tumor and immune regions for an illustrative comparison between BayesDeep and TESLA. B. The actual expression from ST data at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected tumor region for genes $AZGP1$, $MUCL1$, and $TOP2A$. Violin plots display gene expression on the selected tumor region across different cell types. C. The actual expression from ST data at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected immune region for genes $CORO1A$, $IGHG2$, and $TRBC1$. Violin plots display gene expression on the selected immune region across different cell types.

4.4 Real data analysis (continued) and downstream analysis on the human breast cancer 10x Visium data: A. Heatmap of the covariate coefficient matrix $B$ estimated by BayesDeep, indicating the association between gene expression and morphological features extracted from the paired histology image. B. Cell population analysis on the BayesDeep- and TESLA-generated single-cell-resolution gene expression. C. Pseudotime analysis on tumor cells extracted from the cell population analysis. D and E. Cell-cell interactions inferred from the BayesDeep- and TESLA-generated single-cell-resolution gene expression, and their overlap.
4.5 Real data analysis on the human prostate cancer 10x Visium data: A. Manual annotation and nuclei identification results by applying HD-staining to the paired histology image; and the two selected tumor and immune regions for an illustrative comparison between BayesDeep and TESLA. B. The actual expression at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected tumor region for genes ADGRF1, SPON2, and TMEFF2. Violin plots display gene expression on the selected tumor region across different cell types. C. The actual expression at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected immune region for genes CD24, CD47, and CXCR4. Violin plots display gene expression on the selected immune region across different cell types.

4.6 Real data analysis (continued) and downstream analysis on the human prostate cancer 10x Visium data: A. Heatmap of the covariate coefficient matrix $B$ estimated by BayesDeep, indicating the association between gene expression and morphological features extracted from the paired histology image. B. Cell population analysis on the BayesDeep- and TESLA-generated single-cell-resolution gene expression. C. Pseudotime analysis on tumor cells extracted from the cell population analysis. D and E. Cell-cell interactions inferred from the BayesDeep- and TESLA-generated single-cell-resolution gene expression, and their overlap.

A.1 Simulation study: The boxplots of AUCs achieved by different normalization methods in BOOST-MI.

A.2 Simulation study: The boxplots of AUCs achieved by different dichotomization methods in BOOST-MI.

A.3 Simulation study: The heatmaps of averaged AUCs and MCCs achieved by different BOOST-MI hyperparameters $\tau_\omega$ and $\tau_\theta$.

A.4 The scatter plot of the Moran’s I based on the dichotomized data against that based on the normalized data over all genes in the mouse olfactory bulb (MOB) and human breast cancer (BC) datasets.

A.5 The histogram of low-expression spot proportions over all genes in the mouse olfactory bulb (MOB) and human breast cancer (BC) datasets.

A.6 Real data analysis: The hierarchical clustering dendrograms of SV genes detected by BOOST-MI, SPARK, and BinSpect from the (a) mouse olfactory bulb and (b) human breast cancer datasets. Each color represents a unique spatial pattern defined in Figure 2.4 and 2.6.
B.1 Sensitivity analysis: ARIs achieved by iIMPACT when setting the image profile weight $w$ to be zero, and proportion of variance explained under different number of leading principal components in PCA. ......................... 121

B.2 Sensitivity analysis: ARIs achieved by iIMPACT clustering under different choices of the image profile weight $w$ on A. human breast cancer data; and B. mouse visual cortex STARmap data. ......................... 122

B.3 Demonstration of geometric representations of spatial distribution of spots, and definition of spot expanded area in the 10x Visium and classical ST technologies. ................................................................. 123

B.4 Nuclei identification results from the HD-Staining model for human breast cancer data: A. The number of nuclei identified for seven different nuclei classes; B. Histogram of number of cells in each spot expanded area; C. Spatial distribution of spot-level cell type abundance for seven nuclei classes. 124

B.5 Nuclei identification results from the HD-Staining model for human prostate cancer data: A. The number of nuclei identified for six different nuclei classes; B. Histogram of number of cells in each spot expanded area; C. Spatial distribution of spot-level cell type abundance for six nuclei classes. 125

B.6 Histology image of the tissue section with manually annotated labels from pathologists, and histology-based spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to be 3, for human prostate cancer data. ............................... 126

B.7 Nuclei identification results from the HD-Staining model for human ovarian cancer data: A. The number of nuclei identified for six different nuclei classes; B. Histogram of number of cells in each spot expanded area; C. Spatial distribution of spot-level cell type abundance for six nuclei classes. 127

B.8 Histology image of the tissue section with manually annotated labels from pathologists, and histology-based spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to be 2, for human ovarian cancer data. ............................... 128
C.1 Geometric representations of a barcoded spot (in red) and its expanded region (in green) on the spatial transcriptomics (ST) and the improved 10x Visium platforms. For ST, the spot diameter is 55µm with a center-to-center distance of 100µm between two adjacent spots. For 10x Visium, these measures are 100µm and 200µm, respectively. The percentage of area covered by ST barcoded spots can be approximated by the ratio of the red circle area to the green square, calculated as \( \frac{\pi \times (100\mu m)^2}{200\mu m^2} = 19.63\% \). Similarly, the percentage of area covered by 10x Visium barcoded spots can be approximated by the ratio of the red circle to the green hexagon, calculated as \( \frac{\pi \times (55\mu m)^2}{3 \sqrt[3]{2} \times (\frac{2}{\sqrt{3}} \times 100\mu m)^2} = 27.43\% \).

C.2 Overview of model validation for the simulated data: Evaluation results in terms of the root mean square error (RMSE) between the actual and estimated covariate coefficient \( \beta_{ij} \)'s for BayesDeep with and without regularization, respectively.

C.3 Overview of model validation for the human prostate cancer 10x Visium data, including the validation settings and evaluation results in terms of the Pearson correlation coefficients \( \rho \) between the actual and predicted gene expression counts (\( y_{ij} \)'s vs. \( \hat{y}_{ij} \)'s) for BayesDeep, TESLA, and Gaussian Process (GP), respectively.

C.4 Downstream analysis on the human breast cancer 10x Visium data: A. BayesDeep-generated single-cell-resolution expression of the breast cancer stem cell markers \( CD44 \) depicted on the UMAP. B. BayesDeep-generated single-cell-resolution expression of the breast cancer stem cell markers \( CD24 \) depicted on the UMAP. C. Pseudotime analysis on the UMAP.

C.5 Downstream analysis on the human prostate cancer 10x Visium data: A. BayesDeep-generated single-cell-resolution expression of the prostate cancer stem cell markers \( ITGA6 \) depicted on the UMAP. B. BayesDeep-generated single-cell-resolution expression of the prostate cancer stem cell markers \( ALCAM \) depicted on the UMAP. C. Pseudotime analysis on the UMAP.
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1 List of size factors used for normalizing sequence count data.</td>
<td>102</td>
</tr>
<tr>
<td>A.2 List of variance-stabilizing transformations (VST) used for normalizing sequence count data.</td>
<td>103</td>
</tr>
<tr>
<td>A.3 The key notations of the proposed BOOST-MI model.</td>
<td>111</td>
</tr>
<tr>
<td>A.4 Simulation study: Implementation details of BOOST-MI and the competing methods.</td>
<td>112</td>
</tr>
<tr>
<td>A.5 Simulation study: The averaged MCCs (standard deviations) achieved by BOOST-MI, Ising, BOOST-GP, BinSpect, SPARK, and SpatialDE under different scenarios in terms of spatial pattern and zero-inflation setting.</td>
<td>113</td>
</tr>
<tr>
<td>A.6 Real data analysis on the mouse olfactory bulb (MOB) dataset: List of 60 SV genes that had an attraction pattern with a positive interaction parameter in the Ising model.</td>
<td>114</td>
</tr>
<tr>
<td>B.1 Summary of the four real datasets analyzed in the paper.</td>
<td>129</td>
</tr>
<tr>
<td>B.2 Running time (in minutes) of iIMPACT, BayesSpace and SpaGCN on the four real datasets analyzed in the paper.</td>
<td>129</td>
</tr>
<tr>
<td>B.3 The key notations of the proposed iIMPACT.</td>
<td>130</td>
</tr>
<tr>
<td>B.4 Posterior means of parameters (domain-specific relative abundances of cell types $\hat{\Omega}$ and means of low-dimensional representation of gene expression $\hat{M}$) for human breast cancer data, which are applied in generating the simulated data.</td>
<td>131</td>
</tr>
<tr>
<td>C.1 A list of key notations of the proposed BayesDeep.</td>
<td>139</td>
</tr>
<tr>
<td>C.2 A list of explanatory variables utilized by BayesDeep in real data analysis, including cell type classification and nuclei shape features extracted from the paired histology image by HD-staining.</td>
<td>140</td>
</tr>
</tbody>
</table>
I dedicate this to my family, advisors, friends, and people who have supported me throughout my study journey.
Chapter 1
Introduction

1.1. Spatially Resolved Transcriptomics Technology

Spatially resolved transcriptomics (SRT) has recently revolutionized genomics research. SRT enables high-throughput quantitative assessment of the location and abundance of gene activity within a tissue, which can elucidate molecular mechanisms at an unprecedented level of spatial detail. Traditional molecular profiling technologies (e.g., single-cell or single-nuclei RNA sequencing) dissociate tissues, which loses valuable spatial and morphological information of gene expression [1–3], and offers limited gene-level assessment via tissue slides [4]. In contrast, SRT technologies empower the comprehensive characterization of molecular abundance within individual cells while preserving spatial information. Advances in SRT technologies have helped to inform how location-specific information infers gene function and phenotype [5, 6], which have facilitated biological discoveries in fields such as epidemiology and physiology [4, 7–11].

SRT data are generated from multiple technologies, which can be categorized into two types: imaging-based and next generation sequencing (NGS)-based. Imaging-based SRT techniques includes: 1) Sequential fluorescence in situ hybridization (FISH) (seqFISH and seqFISH+) [12], an imaging-based technology that enables the identification of many RNA transcripts sequential rounds of hybridization; 2) Multiplexed error-robust FISH (MER-FISH) [13], another FISH technology with fewer fluorescent channels and more hybridization rounds compared to seqFISH; 3) STARmap (spatially-resolved transcript amplicon read-out mapping) [14], a 3D intact-tissue RNA sequencing approach at single-cell resolution; and etc. NGS-based SRT techniques contains: 1) Spatial transcriptomics (ST) [15] and 4)
high-definition spatial transcriptomics (HDST) [16], two recently developed SRT approaches that obtain high-quality RNA-sequencing data with positional information through arrayed probes with unique positional barcodes. 10x Genomics, a biotechnology company, has developed a new molecular profiling solution called the Visium Spatial Gene Expression, which provides a growing number of publicly available spatial transcriptomics data and a platform for users to perform spatial transcriptomics; 6) Slide-seqV2 [17], an improved version of Slide-seq, which has a higher RNA capture efficiency; and 7) An integration of a microarray-based SRT method and a single-cell RNA sequencing method (denoted as Other) [18], a recently developed method to characterize the cell populations and their spatial organization in a tissue.

1.2. Applications on Spatially Resolved Transcriptomics Data

1.2.1. Spatially variable gene detection

Spatially variable (SV) gene is defined as genes that exhibit spatial variation across the tissue [19]. SV gene identification is a critical aspect of downstream analysis because it provides a comprehensive exploration of SRT datasets. Here, we summarize seven statistical methods tailored for identifying SV genes. These methods can be classified into three categories according to how they define the spatial correlation of gene expressions when modeling SRT data: 1) kernel-based, and 2) other spatial models.

Gene expression levels can be modeled via the Gaussian process where spatial dependency is quantified using various kernel functions. SpatialDE [19] models the normalized expressions with a Gaussian process to identify SV genes. It applies the squared exponential, periodic, and linear kernels to capture spatial dependency. Alternatively, SPARK [20] directly models the gene expression counts via a generalized linear model and employs spatial kernels with different length-scale characteristics to accommodate various spatial patterns.
It has higher power compared to SpatialDE. Moreover, both approaches have low computational cost. BOOST-GP [21] is a Bayesian hierarchical model that uses the Gaussian process (GP) to identify SV genes. Unlike SPARK, it models gene expression counts with a ZINB distribution, so it achieves better performance on highly sparse SRT data. However, the computational cost of BOOST-GP is much higher since it is under the Bayesian framework. These three methods require kernel functions that must be carefully selected, which now brings us to the task of quantifying spatial patterns that are not or cannot be specified by kernel functions.

There also exist other approaches, such as nonparametric or non-model based methods, proposed for SV gene detection. Trendsceek [22] is a nonparametric method without distribution specification, which models data via a marked point process. SV genes are identified by mark-segregation hypothesis testing that are carried out with permutation tests. Trendsceek is computationally intensive and has noticeably unsatisfactory performance compared to other model-based methods [20]. BinSpect [23] is a non-model-based approach grounded on the statistical enrichment of spatial network neighbors by binarizing gene expression levels. The framework of BinSpect is straightforward and it has the highest computational efficiency compared to the other five methods described above. However, the authors of BOOST-MI mention this is an aggressive method that may suffer some misclassification bias. SpaGCN [24] is a spatial domain and SV gene identification approach based on a graph convolutional network. Unlike other SV gene detection methods, it first integrates information from the gene expression profile, location, and pathology image to identify spatial domains. Then, SV genes are defined based on domain-guided differential expression analysis.
1.2.2. Spatial domain identification

Profiling localized gene expression patterns is intricately linked to delineating spatially connected regions or clusters within a tissue based on expression data [25]. The exploration of SRT data necessitates a crucial step of spatial domain identification, achieving through spatial clustering. Here, we summarize the published statistical and machine learning methods for spatial domain identification. These methods can be classified into three categories according to what information they includes: 1) applying molecular profiles in SRT data, 2) applying both molecular and geospatial profiles in SRT data, and 3) applying SRT data and the corresponding histology image.

Traditional clustering methods tailored for single-cell RNA sequencing (scRNA-seq) data [26, 27] utilize the expression across all genes as features of each spot, which do not employ the spatial relationship among spots. Markov random fields (MRF) are employed in certain instances to integrate geospatial profiles into spatial clustering. Hidden Markov random fields (HMRF) is utilized in a published method [28] to identify spatial domains. They construct a neighborhood graph representing spatial relationships among capture locations, with cell states depending on the labels of their immediate neighbors. The HMRF model ensures coherence in both physical and gene expression space, decomposing the graph into multiple components, each representing a spatially variable domain. BayesSpace [29] employs a Bayesian approach with a prior having a function of assigning higher weights to be same cluster for physically closer spatial locations. SC-MEB [30] employs an empirical Bayes approach, assuming Gaussian gene expression at each capture location and encouraging spatial smoothness by penalizing the assignment of neighboring locations to different clusters. stLearn [31] and SpaGCN [24] also utilize features extracted from the histology image. stLearn smoothes expression values at each capture location using information aggregated from neighboring locations, weighted by morphological similarity. Subsequently, standard clustering algorithms such as k-means are applied. SpaGCN is a graph convolu-
tional network (GCN)-based method by integrating the RGB values from histology images and identifying spatial clusters through clustering the output of the graph convolutional layer.

1.2.3. Gene expression resolution enhancement

NGS-based SRT data lack single-cell resolution. The gene expression measured at each spot is from a composition of multiple cells, potentially across different cell types [32]. Enhancing gene expression resolution in SRT data is a critical direction to unravel the intricacies of cellular functions within tissues. Thus, there exists a necessity for spatial gene expression methods that can overcome the inherent lower resolution of the technology. Most of statistical methods to enhance gene expression resolution leverage information from nearby locations. Additionally, high-resolution details on cell morphology can be extracted from the histology images of H&E-stained tissue sections, providing an opportunity to further refine gene expression resolution.

Various methodologies and algorithms have been developed to improve the resolution of gene expression information, advancing our understanding of spatially defined molecular landscapes. BayesSpace [29] is a Bayesian statistical model by leveraging the spatial neighborhood structure to resolve expression at a subspot level. Machine learning and deep learning play pivotal roles in dealing with the histology image and integrating diverse data modalities to explore the spatial heterogeneity of gene expression patterns. XFuse [33] employs a deep generative model to predict super-resolution gene expression at each pixel, leveraging information from histology images. However, it has extremely slow processing time. TESLA [34] is a machine learning framework integrating SRT data and histology image information to generate super-resolution gene expression images. ST-Net [35], a deep learning algorithm, combines SRT data and histology images to capture high-resolution gene expression heterogeneity. ST-Net employs convolutional neural networks and was trained
on a breast cancer SRT dataset. HisToGene [36] is an attention-based model designed for the prediction of super-resolution gene expression from histology images. It applies a modified Vision Transformer model to incorporate the spatial dependency of measured spots through appropriate positional embedding, and consider the spot dependency together with histological features for gene expression prediction. These approaches refine spot-level gene expression counts and some also enable the inference of pixel level molecular profiles, offering invaluable insights for the exploration of the cellular environment.

1.2.4. Others

SRT data analysis encompasses many other diverse directions aimed at unraveling the complex relationships between gene expression patterns and spatial information within tissues.

Due to the limitation on resolution for most NGS-based SRT techniques, another crucial question in the analysis of SRT data is the assignment of cell types for spots. Cell type deconvolution is fundamental to the understanding of tissue structure and mechanism of cell-environment interactions [37]. Besides, cell type information can help to search for spatial gene expression patterns, which provides more biological insights for gene function exploration [38]. Recently, many cell-type deconvolution approaches for SRT data have been developed. They combine the gene expression information from SRT data with some cell-type reference from other genomics data, which introduces more accurate and convincing results. Most of the published methods are leveraging annotated single-cell or single-nucleus RNA sequencing (scRNA-seq or snRNA-seq) data as the labeled reference. Various cell-type estimation approaches have been exploited. Some relies on least square regression, including RCTD [39], spatialDWLS [40], etc, which can measure the mixture of cell types. Other methods are based on machine learning techniques, like v-support vector regression [41]. There also exist some Bayesian methods, such as cell2location [42].
Once cell types are identified through cell-type deconvolution, a natural progression involves investigating how different cell types interact and how their interactions are influenced by spatial proximity [32, 43]. The binding of ligands to receptors facilitates signal transmission, altering the molecular behavior of the recipient cell. The exploration of ligand-receptor pairs, a methodology often employed in single-cell RNA sequencing (scRNA-seq) studies [44], provides insights into intercellular communications. SRT data is able to offer richer information for the study of cell–cell communications in tissues characterized by spatial structures. Several methodologies have recently emerged, leveraging gene expression data, to delve into the intricate landscape of cell–cell communications. CellChat [45] and CellPhoneDB [44] are methods proposed to study cell-cell interactions on scRNA-seq data. SVCA [46] utilizes Gaussian processes to model the variation of gene expression, decomposes the variation into components of different effects and estimates the cell-cell interaction component. SpaOTsc [47] explores cell–cell communications by examining the interaction dynamics among ligand-receptor pairs and their downstream genes. MISTy [48] is a multiview model under machine learning framework, which is capable of learning interaction effects from both neighboring cells and distant cells.

In recent years, it becomes more and more popular to leverage both expression profiles from scRNA-seq data and spatial patterns from SRT data, thus we can transfer knowledge between the two types of data, which benefits the analysis of both data types [49]. Integration of scRNA-seq and SRT data has been applied on cell type deconvolution, spatial clustering, spatial location reconstruction, etc. Some methods are developed to utilize SRT data to help reconstruct spatial locations in the scRNA-seq data, such as GLISS [50], DEEPsc [51], etc.

These diverse directions collectively contribute to a holistic understanding of SRT data, bridging the gap between gene expression and the spatial context within complex biological tissues.
1.3. Advantages of Bayesian Modeling in Analyzing Spatially Resolved Transcriptomics Data

Bayesian modeling offers a powerful and flexible framework for data analysis, distinguishing itself from traditional frequentist methods by incorporating prior knowledge and updating beliefs based on observed data, allowing for a more nuanced and probabilistic representation of spatial relationships [52, 53]. Firstly, Bayesian modeling allows for integrating prior knowledge or beliefs about the parameters for data modeling. This is particularly valuable when dealing with limited data or when there is existing information. Secondly, Bayesian modeling accommodates the incorporation of complex hierarchical structures and random effects. This flexibility is advantageous in data analysis under various context. Additionally, Bayesian modeling, by representing parameters as probability distributions, inherently captures and quantifies uncertainty [54]. This is crucial in data modeling, where uncertainty may arise from various sources such as measurement error or variability in data generating processes. Last but not least, Bayesian models are well-suited for making predictions in unsampled locations. By considering the correlation structure and uncertainty, Bayesian models provide more reliable predictions and credible intervals, which are essential for data interpolation.

Spatial modeling for geospatial data aims at understanding and analyzing complex spatial relationships, which makes it particularly well-suited for applying Bayesian framework [55, 56]. Firstly, the complex spatial dependencies that can be effectively captured through hierarchical structures. For example, spatial modeling can be viewed as a hierarchical specification, with Gaussian processes introduced appropriately at different levels of the specification [55]. For inference on flexible Bayesian model structure, suitable Gibbs sampling or other Markov chain Monte Carlo (MCMC) algorithms can be utilized for model fitting [57]. Moreover, Markov Random Field (MRF), a probabilistic graphical model, is commonly to be used for Bayesian spatial modeling as priors or the underlying hierarchical model specification, due to its characteristics to represent complex dependencies and relationships among
a set of random variables [58]. It assumes local dependency, i.e., the probability distribution of a random variable is conditionally dependent on its neighboring variables, given the values of all other variables in the system. These characteristics of Bayesian modeling make it a suitable choice for modeling the spatial structure for SRT data.

In addition, with the recent rapid technical advances in genome sequencing, Bayesian statistical methods have been widely applied in the analysis of high-throughput bioinformatics data [59]. Firstly, there are essential characteristics of the data generated through genome sequencing technologies like scRNA-seq or SRT: 1) Sparsity is becoming increasingly central to practical data analysis and inference with high-dimensional data. This can be addressed by modeling zero-inflation or using appropriate prior structures for more accurate and stable estimates in Bayesian methods [60, 61]. 2) The expression measurements obtained in SRT platforms are often in the form of counts. Bayesian modeling is particularly advantageous when directly dealing with count data, where Poisson or negative binomial distributions can be applied to account for the specific characteristics of the data [20, 21]. Secondly, Bayesian models can incorporate hierarchical structures, allowing for the modeling of dependencies between different levels of biological information and thus capturing the complexity of biological systems. Moreover, Bayesian methods inherently provide a form of regularization in prior settings, leading to automatic feature selection, which is beneficial for analyzing high-dimensional genomics data [62], such as identifying relevant genes in SRT data [63].

In conclusion, Bayesian modeling offers a powerful and flexible framework for dealing with SRT data. Its adaptability, ability to incorporate prior information, and robustness to unique SRT data characteristics make it an invaluable tool for statistical inference in a wide range of applications for SRT data analysis.
1.4. Overview of Projects

Chapters 2, 3, and 4 describe the modeling approaches for SRT data analysis, focusing on spatially variable gene identification, spatial domain identification, and gene expression resolution enhancement, respectively.

In Chapter 2, we introduce a Bayesian approach to identify SV genes via a modified Ising model. The key idea is to use the energy interaction parameter of the Ising model to characterize spatial expression patterns. We use auxiliary variable Markov chain Monte Carlo algorithms to sample from the posterior distribution with an intractable normalizing constant in the model. Simulation studies using both simulated and synthetic data showed that the energy-based modeling approach led to higher accuracy in detecting SV genes than those kernel-based methods. When applied to two real spatial transcriptomics datasets, the proposed method discovered novel spatial patterns that shed light on the biological mechanisms. Section 2.1 introduce the background, including the development of SRT technology, the statistical methods developed for SV gene detection and their limitation, and the motivation of proposing BOOST-MI. Section 2.2 introduces the two main components of the data preparation for BOOST-MI: sequence count data normalization and relative expression level dichotomization. In Sections 2.2 and 2.3, we introduce the proposed modeling framework and describe the Markov chain Monte Carlo (MCMC) algorithm and the resulting posterior inference. In Section 2.4, we compare BOOST-MI with existing approaches on simulated data. Two real data analyses are shown in Section 2.5. Section 2.6 concludes the article and discusses future research directions.

In Chapter 3, we develop a multi-stage statistical method called iIMPACT. It includes a finite mixture model to identify and define histology-based spatial domains based on AI-reconstructed histology images and spatial context of gene expression measurements, and a negative binomial regression model to detect domain-specific spatially variable genes. Through multiple case studies, we demonstrate iIMPACT outperformed existing methods,
confirmed by ground truth biological knowledge. Section 3.1 briefly surveys the existing work in spatial domain identification and their limitations. Section 3.2 introduces the integrative method for spatial domain and SV gene analysis. Section 3.3 supplies a brief discussion of the MCMC algorithm and the resulting posterior inference. In Section 3.4, we evaluate model performance on simulated data. Four real data analyses are shown in Section 3.5. Our conclusions and discussion are presented in Section 3.6.

In Chapter 4, we propose BayesDeep, a fully Bayesian hierarchical model that infers gene expression for all observed cells by leveraging cellular morphological information extracted from the associated histology image. BayesDeep models the spot-level expression counts via a negative binomial regression model, incorporating cell-typing and shape features extracted from the image using a recently developed deep-learning model. We further employ feature selection by using spike-and-slab priors to improve robustness. By applying BayesDeep to several real SRT datasets, we found that the reconstructed single-cell-resolution gene expressions provide new biological insights and enhance downstream analysis. Section 4.1 discusses the motivation and commonly used approaches in gene expression resolution enhancement. Section 4.2 and 4.3 introduces the model framework and model fitting. The results for the model validation and two real SRT data analysis are presented in Section 4.4 and 4.5, respectively. Section 4.6 discusses the potential future directions.
2.1. Background

Cellular and molecular spatial organizations play essential roles in biological functions. In recent years, spatial molecular profiling (SMP) techniques have made significant breakthroughs, which enable transcriptome measurement in high spatial resolution [4]. Gene expression profiling approaches are no longer limited to tissue-dissociation, which led to the loss of spatial context of the measured gene expression [8]. Sequencing-based SMP platforms, such as spatial transcriptomics (ST) [15] and the improved Visium platform, use spatial barcodes to capture RNA molecules and then synthesize and sequence their complementary DNA molecules. Through this technology, the expression levels of thousands of genes can be measured across hundreds of spatial locations, namely, spots. Those spots are usually arrayed on a two-dimensional grid. Particularly, ST and Visium spots are arranged on the square and triangular lattices, respectively.

SMP techniques enable researchers to study the gene expressions together with their spatial and morphological contexts, which provide new opportunities to advance our understanding of both cellular and molecular spatial organizations [64], and their relationships with diseases [12]. Many new questions can be explored with the emerging SMP techniques. One of the most immediate ones is to identify genes whose expressions exhibit spatially correlated patterns, referred to as spatially variable (SV) genes. The study of spatial patterns in gene expression could reveal significant insights into many aspects, such as embryo development [26], tumor progression [65], and the clinical impact of intra-tumor heterogeneity [66].
Several methods have been developed in recent years to address the above fundamental question in ST studies. Trendsceek [22] is based on marked point processes, and it is computationally intensive [23] and has noticeably unsatisfying performance [20]. BinSpect [23] is an easy and fast computational method based on statistical enrichment of spatial network neighbors after binarizing gene expression levels. Most model-based analyses, such as SPARK [20], SpatialDE [19], and BOOST-GP [21], are built upon the geostatistical model with a Gaussian process (GP), where a kernel must be selected with caution. Among these methods, SpatialDE transforms the measured counts at different locations into normalized data before analysis, SPARK uses a Poisson distribution to model the count data directly, and BOOST-GP accounts for the zero-inflation and mean-variance relationship existing in raw counts under a Bayesian framework. However, the GP-based models rely on ad hoc kernels that limit the models’ ability to identify complex spatial patterns. Furthermore, none of the existing approaches take advantage of the additional spatial structure of the ST data; that is, the gene expression levels are measured on a lattice grid.

To enable the model to identify complex spatial patterns, we developed a novel approach named Bayesian modeling of spatial transcriptomics data via a modified Ising model (BOOST-MI) to identify SV genes in ST studies. It makes computation efficient by taking advantage of the fact that the ST experiments measure gene expression on a lattice. Ahead of fitting BOOST-MI, we need to normalize sequence count data to relative gene expression levels and dichotomize relative expression levels to a binary spatial pattern. Then, BOOST-MI characterizes the binary spatial pattern via inferring the Ising model interaction parameter under a Bayesian framework. The double Metropolis-Hastings (DMH) algorithm [67] is used to sample from the posterior distribution with an intractable normalizing constant in the Ising model. Compared with other existing approaches, BOOST-MI tests the interaction energy parameter in a modified Ising model, which is able to characterize a broader type of spatial patterns than kernel-based modeling approaches. In addition, the proposed method enriches the inference via incorporating priors if necessary and naturally quantifying uncer-
tainties under a Bayesian framework. We demonstrated the advantages of BOOST-MI in a comprehensive simulation study using both simulated data with various spatial patterns and zero-inflation settings and synthetic data from real ST experiments. The proposed model showed outstanding performance compared to the alternatives. Finally, when applied to two real ST datasets, the proposed method discovered novel spatial patterns that shed light on the biological mechanisms.

2.2. Model

This section introduces two data preprocessing steps (see the dashed-line block in Figure 2.1) before fitting BOOST-MI to identify SV genes. Then we review the Ising model and introduce a modified Ising model with external fields to identify SV genes. The schematic diagram is shown in Figure 2.1.

We first summarize the ST data notations. Let an $n \times p$ matrix $Y$ denote the gene expression count table (i.e., the molecular profile). Each entry $y_{ij} \in \mathbb{N}, i = 1, \ldots, n, j = 1, \ldots, p$ is the read count for gene $j$ observed at spot $i$. We use an $n \times 2$ matrix $T$ to represent the geospatial profile, where each row $t_i = (t_{i1}, t_{i2})$ indicates the spot location on the two-dimensional space. All non-boundary spots have the same neighborhood structure. Supposing the gene expression levels are measured on a $L$-by-$W$ square lattice grid, the coordinate of each spot $i$ can be written as $(t_{i1} = l, t_{i2} = w), l = 1, \ldots, L, w = 1, \ldots, W$. If it is not a boundary spot, then its four neighboring spots are at locations $(l \pm 1, w)$ and $(l, w \pm 1)$. Table A.3 in the Appendix A summarizes all notations introduced in the paper.

2.2.1. Data preparation

Normalization is critical to sequence count data analysis. To counteract various artifacts and bias due to biological and technical reasons, we convert each read count to its relative gene expression level, denoted by $\bar{y}_{ij} = y_{ij}/s_i$, where $s_i$ is the size factor of sample $i$, capturing
all nuisance effects. The most straightforward way is to set \( s_i \propto Y_i = \sum_{j=1}^{p} y_{ij} \), i.e., the total number of counts across all genes in each sample (known as sequencing depth or library size), combined with some constraints, such as \( \prod_{i=1}^{n} s_i = 1 \). Note that SPARK suggests this normalization, namely total sum scaling (TSS). We could also consider other estimations on \( s_i \)'s, which have been introduced for mitigating the influence of extremely low and high counts when analyzing bulk RNA-seq data, such as upper-quartiles (Q75) [68], relative log expression (RLE) [69], and weighted trimmed mean by M-values (TMM) [70]. In addition to the above normalization method based on size factor estimation, we could use the method based on variance-stabilizing transformation (VST) (detailed in Section A.1 of the Appendix A). The following sensitivity analysis (see Figure A.1 in the Appendix A) indicates that 1) the methods based on size factor estimation noticeably outperformed the VST-based methods; and 2) the performance based on TSS was not significantly different from Q75, RLE, and TMM. Thus, we suggest using TSS as the default setting for the sake of simplicity.

We then denoise the relative expression levels by partitioning all spots into two groups. Dichotomization is a common practice to denoise the sequence count data. For the analysis of metagenomics data [71, 72], the authors dichotomize the normalized taxonomic abundances to reduce the influence of sequencing noise. ST data also suffer from many biological and technique errors (e.g., dropouts) [73]. Thus, following BinSpect [23], we discretize the normalized expression levels for each gene into two groups. Consequently, this step out-
puts the suitable data type required in the subsequent analysis and makes our approach more robust in the face of over-dispersion and zero-inflation, which are the two important characteristics of ST data. For each gene, we introduce a binary vector $\xi_j = (\xi_{1j}, \ldots, \xi_{nj})$ to represent the dichotomization result based on its relative expression levels $(\tilde{y}_{1j}, \ldots, \tilde{y}_{nj})$, with $\xi_{ij} = 1$ indicating gene $j$ is highly expressed at spot $i$ and $\xi_{ij} = 0$ otherwise. Since there is no consensus on the dichotomization of spots based on either absolute or relative expression level, we propose to estimate $\xi_j$ via fitting a two-component Gaussian mixture model (GMM), $\tilde{y}_{ij} | \xi_{ij}, \mu_j, \sigma_j^2 \sim (1 - \xi_{ij})N(\mu_{j0}, \sigma_{j0}^2) + \xi_{ij}N(\mu_{j1}, \sigma_{j1}^2)$, subjecting to $\mu_{j0} < \mu_{j1}$. Here $\mu_j = (\mu_{j0}, \mu_{j1})$ and $\sigma_j^2 = (\sigma_{j0}^2, \sigma_{j1}^2)$ are the group means and variances that need to be estimated. In addition to the model-based clustering, we also consider $k$-means, which is implicitly based on pairwise distances between relative expression levels. Section A.2 of the Appendix A provides details about the GMM and $k$-means implementations ($k = 2$). The following sensitivity analysis (see Figure A.2 in the Appendix A) indicates that BOOST-MI performed equally well between the two choices.

One concern about the dichotomization step is that it may lead to information loss. To validate that this loss is neglectable, we compared the Moran’s I [74], a measure of spatial autocorrelation, between the normalized and dichotomized ST data. In particular, for each gene $j$,

$$\text{Moran’s I} = \frac{\sum_{i=1}^{n} \sum_{i' = 1}^{n} W_{ii'} (y_{ij} - \bar{y}_j)(y_{i'j} - \bar{y}_j)}{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2}, \quad (2.1)$$

where $\bar{y}_j = \sum_{i=1}^{n} y_{ij} / n$ is the average expression level of gene $j$ over all the $n$ spots. Here, $W = [w_{ii'}]_{n \times n}$ is a binary spatial weight matrix, with $w_{ii'} = 0$, $w_{ii'} = 1$ if spot $i$ and $i'$ are neighbors (i.e., their Manhattan distance $\leq 1$), and $w_{ii'} = 0$ otherwise. Moran’s I ranges from $-1$ to 1. Higher absolute values denote stronger spatial correlation. We computed the Moran’s I based on normalized data and dichotomized data (replace $y_{ij}$ with $\tilde{y}_{ij}$ and $\xi_{ij}$ in Equation (2.1)) for all genes in the two real ST datasets [15] analyzed in this paper. The
graphical representation of the pairwise Moran’s I between the normalized and dichotomized ST data is shown in Figure A.4 in the Appendix A, indicating that the dichotomization does not cause noticeable information loss. Furthermore, we conducted a one-tailed paired Wilcoxon signed-rank test, showing no evidence that Moran’s I at the normalized level is significantly higher than that at the binary level ($p$-value = 1.000 for both datasets).

2.2.2. A brief review of the Ising model

The Ising model, first introduced by [75] and used in statistical mechanics, is a model of interacting binary states on a crystalline lattice [76]. Let $G = (V,E)$ denote a graph with a finite set of vertices $V$ and a set of edges $E$. In statistics, the Ising model is considered an undirected graph such that each vertex is geometrically regular assigned on a lattice and each edge is of the same length [77]. There are three typical types of two-dimensional lattices: square, triangular, and honeycomb (see Figure 2.2). The former two are a special case of Bravais lattices, which can be defined as an infinite array of discrete points generated by a set of discrete translation operations [78]. ST and Visium (an improved ST platform) spots are arranged on square and triangular lattices, where each non-boundary spot has four and six neighbors, respectively. Every vertex will be assigned a binary state (also known as a spin). If there are three or more states, the model is known as the Potts model. Because the spots are assigned with different spins and react with their neighbors’ spins, there exists a measurement of overall energy, named Hamiltonian,

$$H(\xi_j|\theta_j) = -\theta_j \sum_{i \sim i'} I(\xi_{ij} \neq \xi_{i'j}), \quad (2.2)$$

where $i \sim i'$ denotes the collection of all neighboring spot pairs, $\theta_j$ denotes the interaction energy between highly and lowly expressed spots for gene $j$, and $I(\cdot)$ denotes the indicator function. According to the Hammersley–Clifford theorem [58], we can write the probability of observing a particular configuration of $\xi_j$ as $\Pr(\xi_j|\theta_j) = \exp(-H(\xi_j|\theta_j))/\sum_{\xi \in \Xi} \exp(-H(\xi|\theta_j))$,
where $\uplus$ denotes the set of all configurations of spins. An exact evaluation of the normalizing constant (i.e., the denominator of the above equation) requires us to sum over the entire space, consisting of $2^n$ configurations. Thus, $\text{Pr}(\xi_j|\theta_j)$ is intractable even for a small lattice size.

2.2.3. A modified Ising model to identify spatially variable genes

Although Ising models have a wide range of applications in many areas, they are often examined assuming that the underlying abundance of different spins is equivalent. In the context of statistical mechanics, we can view the model as a system of interacting particles in the absence of an external field. However, we found that low-expression spots are usually much higher than high-expression spots due to excessive zeros presented in the ST data. For instance, the averaged proportions of low-expression spots in the two processed real ST datasets [15] analyzed in the paper are 78.6\% and 86.8\%, respectively (see Figure A.5 in the Appendix A). Ignoring this feature could decrease the identification accuracy due to the inclusion of more false positives (see Figure 2.3).

To remedy the oversimplified Ising model, we add an external force in the Hamiltonian,

$$H(\xi_j|\omega_j, \theta_j) = -\theta_j \sum_{i \sim j'} I(\xi_{ij} \neq \xi_{ij'}) - \left[ \omega_{j0} \sum_{i=1}^{n} I(\xi_{ij} = 0) + \omega_{j1} \sum_{i=1}^{n} I(\xi_{ij} = 1) \right], \quad (2.3)$$
where $\omega_j = (\omega_{j0}, \omega_{j1})$ and $\theta_j$ represent the first and second-order intensities. The first term is proportional to the number of neighboring spot pairs with different spins, while the remaining part (in the square bracket) can be viewed as the weighted average of the numbers of spots with different spins. As the energy function in Equation (2.3) is still locally defined, we can write the joint probability on $\xi_j$, up to its normalizing constant, as

$$\Pr(\xi_j | \omega_j, \theta_j) \propto \exp(-H(\xi_j | \omega_j, \theta_j))$$

$$= \exp \left( \theta_j \sum_{i \sim i'} I(\xi_{ij} \neq \xi_{i'j}) + \omega_{j0} \sum_{i=1}^n I(\xi_{ij} = 0) + \omega_{j1} \sum_{i=1}^n I(\xi_{ij} = 1) \right),$$

(2.4)

which serves as the full data likelihood of the proposed BOOST-MI model. To interpret $\omega_j$ and $\theta_j$, we give the conditional probability of observing a high-expression level of gene $j$ at spot $i$,

$$\Pr (\xi_{ij} = 1 | \cdot) \propto \exp \left( \omega_{j1} + \theta_j \sum_{i' \in \text{Nei}(i)} I(\xi_{i'j} = 0) \right),$$

(2.5)

where Nei($i$) denotes the set of all neighbors to spot $i$. Equation (2.5) is essentially a logistic regression, and hence the parameters $\omega_j$ and $\theta_j$ can be interpreted in terms of conditional odds ratios in general. According to Equation (2.5), if $\theta_j = 0$, then $\Pr(\xi_{ij} = 1 | \cdot) = \exp(\omega_{j1})/\exp(\omega_{j0} + \exp(\omega_{j1}))$, implying that each dichotomized expression level $\xi_{ij}$ is independently and identically sampled from a Bernoulli distribution. Thus, no spatial pattern should exhibit, and gene $j$ is a non-SV gene. The underlying abundance of low and high-expression levels in $\xi_j$ are characterized by $\exp(\omega_{j0})/\exp(\omega_{j0} + \exp(\omega_{j1}))$ and $\exp(\omega_{j1})/\exp(\omega_{j0} + \exp(\omega_{j1}))$, respectively. Fixing $\omega_j$, Equation (2.5) reveals that the smaller the $\theta_j$ (i.e., $\theta_j \to -\infty$), the more likely the dichotomized expression level at any spot is concordant with the majority of its neighboring spots’ levels, resulting in a repulsion pattern (i.e., the clustering of spots with the same dichotomized expression level). In contrast, when $\theta_j$ takes a large positive value, we expect an attraction pattern; that is, the exhibition
of clustering among spots with different dichotomized expression levels. Thus, the spatial correlation between the low and high-expression levels of gene $j$ can be quantified by $\theta_j$. Figure 2.1 shows the three typical patterns in terms of $\theta_j$ conditional on $\omega_{j0} = \omega_{j1}$.

In conclusion, BOOST-MI uses $\theta_j$ to characterize the binary spatial pattern defined by the dichotomized gene expression levels $\xi_j$ on the lattice grid $T$. It is noteworthy that the kernel-based methods, such as SPARK, SpatialDE, and BOOST-GP, are only able to identify SV genes with a small subset of repulsion patterns defined by the selected kernel, while BOOST-MI accounts for all attraction and repulsion patterns.

To complete the model specification, we impose $\theta_j \sim \mathcal{N}(0, \tau^2_\theta)$. As for the first-order intensity $\omega_j$, we notice that an identifiability problem arises from Equation (2.5). For example, adding a nonzero constant $c$ into $\omega_{jk}, k = 0, 1$, does not change the conditional probability $\Pr(\xi_{ij} = k | \cdot)$. Thus, we force $\omega_{j1} = 1$ and set a normal prior on $\omega_{j0} \sim \mathcal{N}(1, \tau^2_\omega)$.

2.3. Model Fitting

In this section, we describe the MCMC algorithm for model fitting and the posterior inference. Our inferential strategy allows for simultaneously estimating the first-order intensity $\omega_j$, which reveals the underlying abundance of the low and high-expression levels of gene $j$, and the second-order intensity $\theta_j$ (also known as the interaction parameter), which captures the spatial correlation between the low and high-expression levels. We give the overview of our MCMC algorithms and the resulting posterior inference as below. Note that each gene is examined independently by BOOST-MI.

2.3.1. MCMC algorithms

The full data likelihood function is given in Equation (2.4), which involves an intractable normalizing constant $C(\omega_j, \theta_j) = \sum_{\xi'} \exp(-H(\xi' | \omega_j, \theta_j))$. Taking the two real ST datasets
analyzed in the paper as examples, it needs to sum over $2^{250} \approx 1.8 \times 10^{75}$ and $2^{260} \approx 1.9 \times 10^{78}$ elements, respectively. This makes the Metropolis–Hastings (MH) algorithm infeasible in practice. To overcome this challenge, we employ the double MH (DMH) algorithm [67] to estimate both $\omega_j$ and $\theta_j$ for each gene. The DMH is an asymptotic algorithm, which has been shown to produce accurate results by various spatial models [79,80]. Unlike other auxiliary variable MCMC algorithms [81,82] that also aim to have the normalizing constant ratio canceled, it is more efficient because it does not require drawing the auxiliary variables from a perfect sampler, which is usually computationally expensive. The full details of our MCMC algorithm are given in Section A.3 of the Appendix A.

2.3.2. Posterior inference

Our primary interest lies in the identification of SV genes via making inferences on the interaction parameter $\theta_j$. We obtain the posterior inference by post-processing of the MCMC samples after burn-in. To validate if gene $j$ exhibits a repulsion pattern, we set the null and alternative hypotheses as $\mathcal{M}_0 : \theta_j \geq 0$ and $\mathcal{M}_1 : \theta_j < 0$; while testing an attraction pattern, we set $\mathcal{M}_0 : \theta_j \leq 0$ and $\mathcal{M}_1 : \theta_j > 0$. We could select the model via calculating the Bayes factor (BF) in favor of $\mathcal{M}_1$ over $\mathcal{M}_0$, which is defined as the ratio of posterior odds to prior odds,

$$BF_j = \frac{\Pr(\mathcal{M}_1|\xi_j)}{\Pr(\mathcal{M}_0|\xi_j)} \approx \begin{cases} \frac{\sum_{u=1}^{U} I(\theta_j^{(u)} < 0)/U}{\sum_{u=1}^{U} I(\theta_j^{(u)} \geq 0)/U} & \text{for repulsion} \\ \frac{\sum_{u=1}^{U} I(\theta_j^{(u)} > 0)/U}{\sum_{u=1}^{U} I(\theta_j^{(u)} \leq 0)/U} & \text{for attraction} \end{cases}$$

(2.6)

where the prior odds cancel out as we choose a normal prior on $\theta_j$ centered at zero, and the posterior odds can be approximated using the MCMC samples $\{\theta_j^{(1)}, \ldots, \theta_j^{(U)}\}$. Here $U$ denotes the total number of MCMC iterations after burn-in. The larger the $BF_j$, the more
likely gene $j$ is an SV gene, integrating over the uncertainty in all model parameters. We suggest choosing the BF threshold based on the scale for interpretation [83].

2.4. Simulation Study

We performed a series of simulation studies to evaluate the performance of BOOST-MI and compared it with that of four existing methods: SpatialDE, SPARK, BOOST-GP, and BinSpect. Because of the poor performance of Trendsceek reported in most literature [20, 23], we did not include it here. In addition, we fitted the classical Ising model (with the Hamiltonian defined in Equation (2.2)) under the same Bayesian framework.

2.4.1. Data generative model

We generated simulated data from four artificial spatial patterns (see Figure 2.3(a)-(d)) and two real spatial patterns (see Figure 2.3(e) and (f)). The two real patterns were constructed from the mouse olfactory bulb (MOB) and human breast cancer (BC) datasets [15] analyzed in the next two subsections. The first three artificial patterns (named spot, linear, and GP) were on a $16 \times 16$ square lattice ($n = 256$ spots), while the remaining one named MOB I was on $n = 260$ spots. The MOB II and BC patterns were on $n = 260$ and 250 spots, respectively. We set $p = 100$, among which 15 were SV genes. We followed the data generative schemes [20–22] to simulate the gene expression count table $Y$, which was more consistent with those of kernel-based models, such as BOOST-GP and SPARK, while substantially different from BOOST-MI. For each gene $j$, the log relative expression level at spot $i$ was generated via

$$
\log \tilde{y}_{ij} = \begin{cases} 
\beta_0 + e_i + \epsilon_{ij} & \text{if gene } j \text{ is an SV gene} \\
\beta_0 + \epsilon_{ij} & \text{otherwise}
\end{cases},
$$
where $\beta_0$ denotes the baseline relative expression level and $\epsilon_{ij}$ denotes the non-spatial errors following $N(0, \sigma^2_\epsilon)$. We set $\beta_0 = 2$ and $\sigma_\epsilon = 0.3$. For a non-SV gene, the relative expression levels were from a log-normal (LN) distribution with mean and variance being 2 and 0.3$^2$. Consequently, no spatial correlation should be observed. For an SV gene with the spot pattern, the values of $e_i$’s of the four center spots at $(8,8)$, $(8,9)$, $(9,8)$, and $(9,9)$ were set to $\log 6$, while all others were linearly decreased to zero within a radius of five spots. For an SV gene with the linear pattern, the value of $e_i$ of the most bottom-left spot at $(1,1)$ was set to $\log 6$, while all others were linearly decreased to zero along the diagonal line. For an SV gene with the GP pattern, we firstly randomly generated $e_i$’s from a GP using the squared exponential (SE) kernel (see Figure 2.3(c)), where the length-scale parameter $l$ was set to 1.327 and then rescaled all $e_i$’s so that the middle 80% ranged from 0 to $\log 6$. For an SV gene with the remaining patterns, each spot was dichotomized into low and high-expression levels with $e_i = 0$ and $\log 3$, respectively. To mimic the excess zeros and over-dispersion in the real ST datasets, we sampled each gene expression count $y_{ij}$ from a zero-inflated negative binomial (ZINB) model, $y_{ij} \sim \pi_i (y_{ij} = 0) + (1 - \pi_i) NB(s_i \tilde{y}_{ij}, \phi_j)$, where the size factor $s_i \sim LN(0, 0.2^2)$ and the dispersion parameter $\phi_j$ was from an exponential distribution with mean 10. For the choice of the false zero proportion $\pi_i$, we randomly selected 10%, 30%, or 50% counts and forced their values to zero. Combined with the six patterns and three zero-inflation settings, there were 15 different scenarios. For each scenario, we repeated the above steps to generate 10 replicates.

2.4.2. Prior and algorithm settings

We chose to normalize the raw counts $Y$ using TSS and dichotomize the relative expression levels for each gene using GMM as the default setting. As for BOOST-MI, the prior specifications are $\omega_0 \sim N(1, \tau_\omega^2)$ and $\theta \sim N(0, \tau_\theta^2)$. We set $\tau_\omega = 2.5$ and $\tau_\theta = 1/3$. The former indicated that the underlying proportion of the low-expression spots was ex-
Figure 2.3: Simulation study: The six spatial patterns used to generate the simulated data. (a) and (b) Two artificial patterns with the Gaussian and linear kernel; (c) An artificial pattern generated from the Gaussian process; (d) An artificial pattern with complete attraction pattern, i.e., \( \theta \to \infty \); (e) and (f) Two real patterns constructed from the mouse olfactory bulb (MOB) and human breast cancer (BC) study. The boxplots of AUCs achieved by BOOST-MI, Ising, BOOST-GP, BinSpect, SPARK, and SpatialDE under different scenarios in terms of spatial pattern and zero-inflation setting.
pected between 1% and 99% with a probability of 95%, while the latter ensured about 99% of $\theta_j$’s value ranged from $-1$ to $1$ a priori. A follow-up sensitivity analysis indicated that BOOST-MI was incredibly insensitive to the choice of these two hyperparameters (see Section A.4 and Figure A.3 in the Appendix A). As for the MCMC algorithm, we ran four independent MCMC chains for each gene with 10,000 iterations, discarding the first half as burn-in. We started each chain from a model by randomly drawing all parameters from their prior distributions. Results we report below were obtained by pooling together the MCMC outputs from the four chains. For competing methods, we summarize their implementation details in Table A.4 in the Appendix A and provide the related scripts on our Github page (https://github.com/Xijiang1997/BOOST-MI). Both BOOST-MI and BOOST-GP identify SV genes based on BFs, while BinSpect, SPARK and SpatialDE output $p$-values to guide the selection.

2.4.3. Evaluation metrics

First, we used the area under the curve (AUC) of the receiver operating characteristic (ROC) to evaluate the performance of all methods [84]. The ROC curve was created by plotting the true positive rate against the false positive rate across various thresholds used to select SV genes based on BFs or $p$-values. Second, we classified each gene as an SV or non-SV gene by pinpointing a specified threshold. Specifically, we set the BF cutoff to 150, corresponding to a decisive strength of evidence. To control the type-I error rate, we adjusted $p$-values from SPARK, SpatialDE, and BinSpect using the Benjamini-Hochberg procedure [85] and chose a significance level of $0.05$ as the cutoff. We chose the Matthews correction coefficient (MCC) [86] as the secondary performance metric, because SV genes are usually only a small subset of all genes, making other binary classification metrics not suitable. AUC yields a value between 0 and 1, and MCC value ranges from $-1$ to $1$. For both of them, the larger the value, the more accurate the identification.
2.4.4. Model performance

According to Figure 2.3, which displays the boxplots of AUCs by different methods over ten replicated datasets under each scenario, we concluded as follows. First, BOOST-MI achieved the highest performance in terms of median AUC under 14 out of 18 scenarios, while BOOST-GP had only a marginal advantage over BOOST-MI under the scenario with a medium or high proportion of false zeros and the artificial spot, linear or GP pattern. Second, under the low zero-inflation setting, SPARK had a similar performance with BOOST-MI when the SV genes were generated from linear and BC patterns. However, it suffered from reduced power under the medium and high zero-inflation settings. This clearly suggested that realistic modeling in BOOST-MI contributed to its advantage over all other methods. Third, BinSpect was very sensitive to the choice of clustering methods. For example, the one based on top percentage rank performed significantly better than the one based on \( k \)-means under almost all scenarios, excluding the four scenarios with a high proportion of false zeros and the first four spatial patterns. In contrast, BOOST-MI is considerably robust to different normalization and dichotomization methods (see Figure A.1 and A.2 in the Appendix A). Forth, our BOOST-MI with the modified Hamiltonian consistently outperformed the Ising model with the classical Hamiltonian. Last but not least, SPARK, SpatialDE and BOOST-GP had absolutely no power to detect SV genes with the MOB I pattern, which was defined by a large positive interaction parameter in the modified Ising model. They might miss some important discoveries in real data analysis. Meanwhile, BinSpect had no satisfactory performance to detect such a pattern either, indicating our model-based analysis BOOST-MI could sharpen inferences. Besides, we reported the summaries of all methods’ performance in terms of MCCs in Table A.5 in the Appendix A, respectively. Those results led to similar conclusions.

Regarding the efficiency, the average execution time per gene was 0.004, 0.290, and 0.067 seconds for BinSpect, SPARK, and SpatialDE. BOOST-MI spent 4.5 seconds per gene on
average due to the computationally intensive DMH algorithms. In contrast, BOOST-GP had around two times higher computational cost than BOOST-MI, which spent 11.3 seconds per gene on average. All experiments were implemented on a high-performance computing server with two Intel Xeon CPUs (45 MB cache and 2.10 GHz) and 250 GB memory.

2.5. Real Data Analysis

2.5.1. Case study on mouse olfactory bulb spatial transcriptomics dataset

The first dataset that we applied BOOST-MI to analyze is a publicly available ST dataset in a mouse olfactory bulb (MOB) study [15]. It is accessible on the Spatial Research Lab (http://www.spatialresearch.org). There are 12 replicates in this study. Following the previous studies [19–21], we used the MOB replicate 11, which contains 16,218 genes measured on 262 spots. We applied the quality control steps described below before applying BOOST-MI. First, we excluded spots with fewer than ten total counts across all genes. Then, those genes with more than 80% zero read counts across all spots were dropped. After these two steps, the MOB data had $n = 260$ spots and $p = 9,769$ genes. For other methods, we applied the pre-processing procedures suggested in their papers. We used the same prior specification, algorithm setting, and significance criteria used in the simulation study. We ran four independent MCMC chains and used the potential scale reduction factor (PSRF) [87] to diagnose MCMC convergence. PSRF is a statistic comparing the estimated between-chains and within-chain variances for a model parameter. Its value should be close to one if multiple chains have converged to the target posterior distribution. The PSRFs for all $\omega_j$'s and $\theta_j$'s were below 1.1, clearly suggesting that the MCMC algorithms converged. Then, for each dataset, we pooled together the outputs from the four chains and selected SV genes based on their BFs. We only compared BOOST-MI with SPARK and BinSpect-rank due to the poor performance of SpatialDE and heavy computational burden of BOOST-GP. The
average execution time per gene was 0.004, 0.294, and 4.697 seconds by BinSpect, SPARK, and BOOST-MI, respectively.

BOOST-MI identified 734 SV genes, which was approximately the same number of SV genes detected by SPARK (772 SV genes) and around half of the number of SV genes detected by BinSpect (1,415 SV genes). Figure 2.4(a) is a Venn diagram showing the overlap of detected SV genes by all methods. More than half of the SV genes identified by BOOST-MI (388 out of 734) were also reported by the other two, while there were 221 unique SV genes to BOOST-MI. Most of the SV genes detected by SPARK (618 out of 772) were also included in the result of BOOST-MI or BinSpect, which indicates that SPARK is a relatively conservative method for SV gene identification. BinSpect was shown to be the most aggressive method that reported the most SV genes, more than half of which (746 out of 1,415) were not founded by either alternative.

To further explore the spatial patterns we had found, we performed the agglomerative hierarchical clustering on the SV genes identified by each of the three methods. For better visualization, we followed the data preprocessing [19, 20] to normalize the raw read counts to relative expression levels using log-VST. Next, based on the distance matrix computed from relative expression levels of all pairs of SV genes, a hierarchical clustering dendrogram was constructed. We then determined the number of clusters by cutting the hierarchical clustering dendrogram at a height corresponding to a clear separation. The hierarchical clustering dendrogram is shown in Figure A.6(a) in the Appendix A. Last, we summarized the expression patterns via the averaged relative expression levels within each cluster. As a result, the SV genes detected by BOOST-MI and BinSpect were clustered into five groups, while there were four groups for SPARK, shown in Figure 2.4(b). Consistent with the results in a previous study [20], three major spatial patterns were shown in the first three columns in Figure 2.4(b). Note that the fourth pattern of SPARK (with 80 genes) and BinSpect (with 140 genes) could be merged into their first pattern, respectively. Also, the third and
Figure 2.4: Real data analysis on the mouse olfactory bulb (MOB) dataset: (a) The Venn diagram of SV genes identified by BOOST-MI, SPARK, and BinSpect-rank; (b) Distinct spatial expression patterns summarized on the basis of 734, 772, and 1,415 SV genes identified by BOOST-MI, SPARK, and BinSpect-rank; (c) Distinct spatial expression patterns summarized on the basis of 307 SV genes identified by BOOST-MI only; (d) Distinct spatial expression patterns summarized on the basis of 345 SV genes identified by SPARK only; (e) The associated hematoxylin and eosin (H&E)-stained tissue slides of the analyzed MOB dataset; (f) The top 15 gene ontology (GO) terms in the GO enrichment analysis for those 734 SV genes identified by BOOST-MI, with the red dashed line indicating a significance level of 0.05; (g) The top six genes with the attraction pattern in terms of BF values.
fourth patterns of BOOST-MI had a high similarity. It is noteworthy that a unique pattern (the last pattern of BOOST-MI) could only be detected by BOOST-MI with 194 SV genes, indicating our method had a higher power. To further compare BOOST-MI with SPARK, we repeated the above procedure on those SV genes detected by BOOST-MI or SPARK only. The results are shown in Figure 2.4(c) and (d), respectively. The 307 genes identified only by BOOST-MI could be categorized into two groups, including the unique pattern with 178 genes. SV genes only detected by SPARK showed a strong periodic pattern, suggesting that SPARK might be more sensitive to the smooth periodic spatial pattern.

To have an additional diagnosis of the spatial signals of those detected SV genes, we employed Moran’s I spatial autocorrelation test [21, 74, 88]. First, we split the 1,827 SV genes detected by any of the three methods into seven mutually exclusive groups shown in Figure 2.4(a), i.e., the common SV gene set, three SV gene sets detected by each pair of methods, and three SV gene sets detected by each method alone. We then computed the Moran’s I for each gene. Figure 2.5(a) summarizes the Moran’s I’s of all genes by the seven groups. According to the boxplot, we conclude the following: 1) SV genes identified by all three methods have the strongest spatial correlation on average, implying that those genes are more likely to be true SV genes, and 2) SV genes detected by BOOST-MI alone, in general, have higher spatial correlations than those SV genes reported by SPARK or BinSpect alone, suggesting that BOOST-MI has a stronger ability to capture spatial patterns compared to other two approaches.

Next, using a Python wrapper GSEAPY [89, 90], we performed gene ontology (GO) enrichment analysis of the 734 SV genes identified by BOOST-MI to explore their relevant biological functions. A total of 3,929 mouse GO terms in three components (biological processes, cellular components, and molecular functions) had at least one gene overlap. Controlling the false discovery rate (FDR) at 5%, we found 155 GO terms, the top 15 of which (with the smallest \( p \)-values) are shown in Figure 2.4(f). As with SPARK, many enriched
gene sets were related to synaptic signaling and the nervous system, both of which are significantly associated with the synaptic organization and olfactory bulb development [91]. Examples include chemical synaptic transmission (GO:0007268; adjusted $p$-value $4.69 \times 10^{-7}$) and nervous system development (GO:0007399; adjusted $p$-value $1.18 \times 10^{-3}$).

Lastly, BOOST-MI identified 60 genes that had an attraction pattern with a positive interaction parameter in the modified Ising model. Table A.6 in the Appendix A lists all of them. To analyze the potential biological functions of these SV genes, we performed functional enrichment analysis. There were 603 mouse GO terms and 43 Kyoto Encyclopedia of Genes and Genomes (KEGG) terms with at least one gene overlapping with those SV genes. We also found some statistically significant GO and KEGG terms, which had adjusted $p$-values less than a significance level of 0.05. For example, holo TFIIH complex (GO:0005675) and nucleotide excision repair KEGG term were significantly enriched (adjusted $p$-value 0.040 and 0.013, respectively). In all, these discoveries highlighted the advantage of BOOST-MI.
2.5.2. Case study on human breast cancer spatial transcriptomics dataset

The second dataset is an available ST dataset in a human breast cancer (BC) study [15], which is also accessible on the Spatial Research Lab (http://www.spatialresearch.org). There are four layers available and we used the BC layer 2, which contains 14,789 genes measured on 251 spots. We applied the same quality control steps described in Section 2.5.1. The BC data had \( n = 250 \) spots and \( p = 2,280 \) genes. We applied the same pre-processing procedures for other methods and the same prior specification, algorithm setting, and significance criteria as mentioned. To check convergence, we ran four chains and the PSRFs for all \( \omega_j \)'s and \( \theta_j \)'s were below 1.1. We also compared our result with BinSpect and SPARK. The average execution time per gene was 0.004, 0.272, and 4.369 seconds by BinSpect, SPARK, and BOOST-MI, respectively.

There were 302 SV genes identified by BOOST-MI, which was slightly larger than the number of SV genes detected by SPARK (293 SV genes) and around one-tenth of the number of SV genes detected by BinSpect (3,278 SV genes). A Venn diagram, as shown in Figure 2.6(a), indicates that over 40% of the SV genes detected by BOOST-MI (124 out of 302) were in common with the result obtained by both SPARK and BinSpect. There were only 27 SV genes identified only by BOOST-MI. Once again, it demonstrated that SPARK is a conservative method and BinSpect is an aggressive one, since nearly 90% of the SV genes detected by SPARK (260 out of 293) were also included in the results of BOOST-MI or BinSpect, while nearly 90% of the SV genes reported by BinSpect (2,874 out of 3,278) were found by neither BOOST-MI nor SPARK.

We reported more detailed analysis results, following the same procedure when studying the MOB dataset. Each row in Figure 2.6(b) shows the distinct expression patterns detected by BOOST-MI, SPARK, and BinSpect, respectively. There were five, four, and five groups obtained by performing the agglomerative hierarchical clustering. The hierarchical clustering dendrogram is shown in Figure A.6(b) in the Appendix A. It is noteworthy that BOOST-
Figure 2.6: Real data analysis on the human breast cancer (BC) dataset: (a) The Venn diagram of SV genes identified by BOOST-MI, SPARK, and BinSpect-rank; (b) Distinct spatial expression patterns summarized on the basis of 302, 293, and 3,278 SV genes identified by BOOST-MI, SPARK, and BinSpect-rank; (c) Distinct spatial expression patterns summarized on the basis of 174 SV genes identified by BOOST-MI only; (d) Distinct spatial expression patterns summarized on the basis of 165 SV genes identified by SPARK only; (e) The associated hematoxylin and eosin (H&E)-stained tissue slides of the analyzed BC dataset; (f) The top 15 gene ontology (GO) terms in the GO enrichment analysis for those 302 SV genes identified by BOOST-MI, with the red dashed line indicating a significance level of 0.05.
MI detected all four patterns discovered by SPARK. However, BOOST-MI detected 102 SV genes with the first pattern, while SPARK reported only 52 SV genes with a similar pattern. BOOST-MI identified fewer SV genes than SPARK for the second and third patterns, while the number of SV genes was almost the same for the fourth pattern. The last pattern of BOOST-MI was unique, and it was approximately the reversed pattern of the first one. Additionally, we repeated the same procedure on the SV genes only detected by BOOST-MI and SPARK, which is shown in Figure 2.6(c) and (d). In this BC study, BOOST-MI reported no SV genes with an attraction pattern.

Following the same procedure described in Section 2.5.1, we computed the Moran’s I for all 3,341 genes in the seven groups (shown in Figure 2.4(b)) and summarized the results in Figure 2.5(b). They clearly suggest again that: 1) SV genes identified by all three methods have the strongest spatial correlation on average, and 2) SV genes detected by BOOST-MI alone tend to have higher spatial correlations than those SV genes reported by the alternatives.

Finally, we performed GO enrichment analysis of the 302 SV genes identified by BOOST-MI. A total of 2,477 human GO terms had at least one gene overlap with those identified SV genes. At an FDR of 5%, 116 GO terms were found. Figure 2.6(f) shows the top 15 GO terms with the smallest adjusted $p$-values. SPARK discovered many enriched gene sets, which were related to extracellular matrix organization and immune responses [20]. Although these GO terms were not shown in Figure 2.6(f), BOOST-MI did detect the same related terms with a significant result (e.g., the adjusted $p$-value for extracellular matrix organization (GO:0030198) was $1.53 \times 10^{-14}$). Furthermore, more virus-related GO terms were found to be significant in our analysis. There is strong evidence that many types of the virus may have a causal relationship with human breast cancers [92]. For example, virus life cycle GO term (GO:0019058) was significantly enriched in the reported SV genes by BOOST-MI,
while not statistically significant for SV genes identified by SPARK (the adjusted $p$-value was only 0.829).

2.6. Discussion

In this paper, we develop a multi-stage method to identify SV genes from ST data. Instead of characterizing an SV gene via a pre-specified kernel by most existing methods, we define a spatial pattern via the Hamiltonian energy in the Ising model with external fields. BOOST-MI is flexible as it provides users multiple choices of data preprocessing procedures and model priors. Meanwhile, BOOST-MI is insensitive to different settings and feasible for balancing the sensitivity and specificity of SV gene detection. Besides, it is considerably robust to various biological and technique noise and biases. In the simulation study, BOOST-MI had a noticeable advantage over alternative methods, especially when there were a great number of zeros in the data. This is very encouraging since BOOST-MI does not directly model excess zeros. BOOST-MI led to more discoveries in real data analysis, such as novel spatial patterns that had never been reported and novel SV genes that kernel-based methods are unable to detect.

Several extensions of our model are worth investigating. First, the proposed model could be extended to model $k$ discrete gene expression levels via a Potts model or its modified version to characterize finer spatial patterns. The number of components $k$ can even be estimated [93]. Second, based on the multi-stage BOOST-MI, a fully Bayesian hierarchical model could be developed to directly model the ST count data to account for the uncertainties associated with the dichotomization. Besides, the model could enable a joint selection of SV genes by incorporating knowledge on relationships among genes (e.g., gene-gene network), following the idea proposed by a recent work [62]. As the model complexity increases, advanced Bayesian computation methods such as variational Bayes [94] are thus needed to reduce the computational burden and make the joint model feasible in real-world applications. Finally, since our approach requires the gene expression measured on a lattice grid, it
is necessary to generalize our model to detect SV genes from SMP data produced by other platforms based on the single-molecule fluorescence in situ hybridization (FISH), such as sequential FISH (seqFISH) [95] and multiplexed error-robust FISH (MERFISH) [13], through the hidden Ising or Potts models [80]. Those SMP techniques measure the expression levels of hundreds of genes on thousands of cells, which are irregularly scattered in a planar space. These future directions could potentially further improve the performance of BOOST-MI.
3.1. Background

Spatially resolved transcriptomics (SRT), a new generation of RNA-sequencing analysis techniques, provides biological information at the cellular level while preserving the organization of the tissue and cellular microenvironment [3, 4, 96, 97]. One category of SRT methods builds upon next-generation sequencing (NGS)-based SRT techniques, including spatial transcriptomics (ST) [15], 10x Visium (an improved ST platform), Slide-seq [98], Slide-seqV2 [17], and high-definition spatial transcriptomics (HDST) [16]. These techniques capture RNA molecules via spatially arrayed barcoded probes. The barcoded areas, namely spots, cover a group of cells and are usually arrayed on a two-dimensional grid. Another category of SRT platforms is based on imaging techniques, such as seqFISH [95], MERFISH [13], and STARmap [14]. They measure the expression level for hundreds to thousands of genes at the single-cell resolution with detailed spatial organization information. With these advancements, SRT techniques have been widely applied to facilitate discoveries of novel insights in biomedical studies.

A central challenge for SRT data analysis is to define clinically or biologically meaningful spatial domains by partitioning regions with similar molecular and/or histological characteristics, because the spatial domain identification serves as the foundation for several important downstream analyses, including but not limited to the domain-based differential expression analysis, trajectory analysis, and functional pathway analysis [25, 99]. However, current state-of-the-art methods typically focus on achieving this goal solely by analyzing SRT molecular profiles, such as gene expressions, while neglecting the valuable morphological
or biological information present in the associated histology images. For example, the Seurat package, the most prevalent single-cell RNA sequencing data analysis pipeline [26, 27], utilizes only the high-throughput gene expression of each spot for clustering analysis but does not leverage any information from the associated histology images in this analysis. On the other hand, several recently developed methods, such as stLearn [31], hidden-Markov random field [28], BayesSpace [29], and SpaGCN [24], integrate spatial information or various features extracted from the histology image into the clustering analysis of SRT data. However, those image features, such as RGB color values, do not explicitly reveal detailed morphological information (e.g., cell locations and types) and therefore fail to provide biologically relevant insights.

Different from molecular information, histology images characterize cellular structures and tissue microenvironments, which have been proven valuable in clinical diagnosis and prognosis [100, 101]. Computer vision algorithms have enabled us to automatically segment cell nuclei from digital histology images at a large scale [102]. Recent developments in deep convolutional neural networks (e.g., H-DenseUNet [103], Micro-Net [104], Hover-Net [105], and HD-Staining model [100]) have further integrated the automatic identification, classification, and feature extraction of each observed nucleus in a histology image. In practice, a histology-based spatial domain (e.g., tissue) is defined as a group of cells with similar morphological and molecular context as a unit. Thus, we hypothesize that integrating spot-level molecular profiles and cellular-level image profiles from AI-reconstructed histology images could enhance the spatial domain identification in terms of both accuracy and interpretability.

Another challenge for SRT data analysis is to identify spatially variable genes (SVGs), which represent genes with spatially correlated expression patterns [20, 21]. Recently developed methods, such as SpatialDE [19], SPARK [20], BOOST-GP [21], and BOOST-MI [106], characterize the global spatial dependency of a gene in the whole domain while ignoring the
spatial pattern heterogeneity due to cellular organization, which could be fully observed in AI-reconstructed histology images. SpaGCN [24] proposed domain-guided differential expression analysis to detect SVGs without a rigorous statistical framework. Therefore, there is an urgent need to develop reliable statistical method to detect domain-specific SVGs.

This paper proposes a two-stage statistical approach by integrating Image and Molecular Profiles to Analyze and Cluster spatial Transcriptomics data, or iIMPACT for short. The first stage is implementing a Bayesian finite mixture model to allocate all spots into mutually exclusive clusters, namely histology-based spatial domains. We decompose each mixture component into two sub-components to integrate image and molecular profiles. In particular, a multinomial sub-component is employed to model cell type abundance available in histology images. Following BayesSpace [29], we use a normal sub-component to model the low-dimensional representation of normalized gene expression from the matching SRT molecular profile. The Bayesian model also adopts a Markov random field prior (MRF) to encourage neighboring spots to be clustered in the same histology-based spatial domain. The spots’ neighborhood structure can be straightforwardly defined from the NGS-based SRT geospatial profile, as spots are usually arrayed on square or triangular lattices. Through the resulting posterior inference, we obtain histology-based spatial domains and their interactive zones, while characterizing each identified histology-based spatial domains by inferring its underlying domain-specific relative abundance of cell types. The second stage is implementing a negative binomial (NB) regression model to search for domain-specific SVGs, which are differentially expressed between a given histology-based spatial domain identified in the first stage and all others. This approach directly models the numbers of read counts (used as a proxy for gene expression) in the SRT molecular profile to achieve minimum information loss. iIMPACT could also be extended to analyze imaging-based SRT data via some special handling. Compared with existing state-of-the-art methods, iIMPACT is able to fully leverage information from the nuclei segmentation procedure on the histology images for clustering analysis and has strong biological interpretability. Applying iIMPACT
on four datasets from different SRT platforms (summarized in Table B.1 in Appendix B), we confirmed that iIMPACT performed better on both spatial domain identification and domain-specific SVG detection than state-of-the-art methods. We further demonstrated that iIMPACT could capture biological features at both the spatial domain level and gene level. Therefore, by integrating image and molecular information, iIMPACT facilitates the discovery of new biological insights from SRT datasets.

3.2. Model

In this section, we first define the molecular and geospatial profiles from NGS-based SRT data (e.g., spatial transcriptomics and the improved 10x Visium platform) and the image profile from the matching AI-reconstructed histology image. Then we discuss how to construct the corresponding profiles from imaging-based SRT (e.g., STARmap) data. After that, we detail the statistical models used in the two stages of iIMPACT. Table B.3 in the Appendix B summarizes all key notations introduced in this section.

3.2.1. Overview

iIMPACT is a two-stage statistical method to analyze SRT data, with its workflow shown in Figure 3.1. It includes two stages – histology-based spatial domain identification by a Bayesian normal-multinomial mixture model and domain-specific SVG detection by a negative binomial regression model.

To achieve the above goals, iIMPACT utilizes the morphological context of histology images and the spatial context of gene expression measurements, referring to the image and molecular profiles in Figure 3.1A and throughout the paper. In particular, the molecular profile refers to the low-dimensional representation of normalized gene expression values at the spot level (denoted by $Y$), which is obtained by a pre-specified dimension reduction technique, such as principal component analysis (PCA). The accompanying SRT geospatial
Figure 3.1: Workflow of iIMPACT: A. iIMPACT starts by combining and processing image profile from AI-reconstructed histology images, and geospatial and molecular profiles from SRT data (circled by dashed lines) to conduct the histology-based spatial domain identification. B. A Bayesian normal-multinomial mixture model with the Markov random field (circled by solid lines) is fitted for histology-based spatial domain identification. Based on the spatial domain identification results, biologically important cellular spatial organization can be characterized, including the domain-specific relative abundance of cell types and interactive zones (circled by dotted lines). C. Domain-specific SVGs are identified by a negative binomial (NB) regression model.
profile that records all spots’ locations is processed as an adjacent matrix (denoted by $G$) representing the spots’ neighborhood structure. iIMPACT requires the locations and types of all cell nuclei in the matching histology image. Combining with the geospatial profile, we can generate the image profile (denoted by $V$), which indicates the spot-level cell type abundance, i.e., the numbers of different cell types within a spot and its expanded area.

In the first stage, we employ a Bayesian normal-multinomial mixture model with the MRF prior [58, 107] to identify the histology-based spatial domains (Figure 3.1B) and interactive zones, corresponding to those spots with less confidence to be allocated to any histology-based spatial domains. Through model parameter estimation, iIMPACT can infer the underlying relative abundance of cell types at each histology-based spatial domain to provide a reference to distinguish their histological types. In the second stage, an NB regression model is fitted for each gene and each histology-based spatial domain of interest, where domain-specific SVGs can be defined (Figure 3.1C).

3.2.2. Data preparation

**Molecular profile $Y$:** In general, the spot-level molecular profile of NGS-based SRT data can be represented by an $N \times P$ count table $C$, where each entry $c_{ij} \in \mathbb{N}$, $i = 1, \ldots, N$, $j = 1, \ldots, P$ is the read count for gene $j$ measured at spot $i$. To account for nuisance effects across spots, including sequencing depth, amplification and dilution efficiency, and reverse transcription efficiency, we normalize each read count $c_{ij}$ to its relative level $\tilde{c}_{ij} = c_{ij}/s_i$, where $s_i$ is the total sum of counts across all genes at spot $i$, $s_i = \sum_{j=1}^{P} c_{ij}$, although other normalization methods are acceptable. Then, the relative gene expression $\tilde{c}_{ij}$ are further log transformed to approximately conform to normality. Following the preprocessing steps in BayesSpace [29], we select the top 2,000 most highly variable genes in terms of their relative expression and perform principal component analysis (PCA), or other dimension reduction techniques (e.g., t-SNE [108] or UMAP [109]), to obtain the low-
dimensional representation of normalized gene expression denoted by an $N \times P'$ matrix $Y$, where each entry $y_{ij} \in \mathbb{R}$, $i = 1, \ldots, N$, $j = 1, \ldots, P'$ is the value of the $j$–th top principal component (PC) at spot $i$. We choose to model the PCs in $Y$ rather than the raw count table $C$ to avoid the use of complex finite mixture models with feature selection based on cumbersome multivariate distributions. Here, we recommend modeling the top three PCs ($P' = 3$) for simplicity. A sensitivity analysis on the human breast cancer data (see Section B.2 in Appendix B) shows that larger $P'$ only provided marginal improvements in clustering performance.

**Image profile $V$:** To integrate the image profile into iIMPACT, we applied a nuclei segmentation and identification algorithm, the histology-based digital (HD)-Staining model [100], to extract cellular features from images. The HD-Staining model is a trained deep-learning model implemented by the mask regional convolutional neural network (Mask R-CNN) architecture [110] for the tumor morphological microenvironment to segment the nuclei of different types of cells, such as immune, tumor and stromal cells. The model was first trained using histology images from lung adenocarcinoma patients in the National Lung Screening Trial study, which has nuclei of six different cell types manually labeled by pathologists. Although the model was originally trained by lung cancer data, it has been improved and verified to be widely adapted to histology image datasets with other cancer types, such as breast cancer, head and neck cancer, ovarian cancer, prostate cancer, and other carcinomas. The HD-Staining model takes a batch of high-resolution histology image patches of a tissue section as input and simultaneously segments and classifies cell nuclei on this image patch. It provides the locations and types for all identified nuclei in the whole histology image. To match the molecular information measured at spots, which only take less than half area (e.g., the area of all spots in 10x Visium platform is about 38% of the whole domain area), we count cells with different types within each spot and its expanded area (see Figure B.3 in Appendix B) so that all the cellular information can be utilized. The result is summarized into an $N \times Q$ count matrix $V$, namely cell abundance table, where
each entry $v_{iq} \in \mathbb{N}$, $i = 1, \ldots, N$, $q = 1, \ldots, Q$ is the number of cells with type $q$ observed at spot $i$ and its expanded area. iIMPACT leverages the single-cell level histology information from the image profile to enhance spatial domain identification.

**Geospatial profile $G$:** Spots are the round area of barcoded mRNA capture probes where gene expression is measured. The spatial distribution of spots is arrayed on a square or triangular lattice. We denote the SRT geospatial profile by an $N \times 2$ matrix $T$, where each row $t_i = (t_{i1}, t_{i2})$ gives the $x$ and $y$ coordinates of the spot $i$ on a two-dimensional Cartesian plane. ST and 10x Visium spots are arranged on square and triangular lattice grids, respectively. Thus, defining a neighborhood structure provides an alternative way to represent the geospatial profile $G$. In particular, $G$ is an $N \times N$ binary adjacent matrix, where each entry $g_{ii'} = 1$ if spot $i$ and $i'$ are neighbors (i.e., the Euclidean distance $\sqrt{(t_{i1} - t_{i'1})^2 + (t_{i2} - t_{i'2})^2}$ between spot $i$ and $i'$ is less than a threshold) and $g_{ii'} = 0$ otherwise. Note that each diagonal entry $g_{ii'}$ is equal to zero. There are four and six neighbors for each non-boundary spot from the ST and 10x Visium platforms, respectively. With this neighborhood structure $G$ as our geospatial profile, the spatial information can be easily integrated into Bayesian cluster analysis via an appropriate prior setting.

**Special handling to imaging-based SRT data:** Imaging-based SRT techniques usually have a higher spatial resolution than NGS-based SRT techniques, which is capable of profiling mRNA at the single-cell level. Data from some imaging-based platforms might provide the spatial distribution and types of cells on the tissue section in the original study. To fit iIMPACT to imaging-based SRT data such as STARmap [14], we manually add a square lattice grid with appropriate size to the whole domain and consider each square unit as a spot (see Figure 3.6A). Note that those 'spots' fill the whole domain; thus, there is no gap between two adjacent spots. For STARmap data in Section 3.5, the grid size was chosen to be $750 \times 750$ pixels, resulting in $N = 170$ spots. Each non-boundary spot has four neighboring spots. We define $G$ with each entry $g_{ii'} = 1$ if spot $i$ and $i'$ are neighbors. To
construct the molecular profile $Y$, we first normalize, transform, and reduce the dimension of the gene expression counts at the single-cell level, and then average the resulting values across all cells within each spot. To obtain the 'image' profile $V$, we directly count the cells with different types in each spot.

3.2.3. A Bayesian normal-multinomial mixture model to identify histology-based spatial domains

The first stage of iIMPACT is to use a Bayesian finite mixture model to partition the whole domain into $K$ mutually exclusive histology-based spatial domains. In general, a finite mixture model [111,112] generates random variables from a weighted sum of $K$ independent distributions that belong to the same parametric family,

$$f(x_i|z_i, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \left[ \pi(z_i = k) f_k(x_i|z_i = k, \theta_k) \right],$$

where $z = (z_1, \ldots, z_N)^T$ denotes the latent variables specifying the identity of the mixture component $f_k$, characterized by $\theta_k$, to each observation $x_i$. In the context of this paper, $x_i = \{y_i \in \mathbb{R}^P, v_i \in \mathbb{N}^Q\}$ represents the observed molecular and image profiling data, and $z_i = k$ indicates that spot $i$ belongs to histology-based spatial domain $k$. Since there are two modalities $Y$ and $V$, we decompose the mixture component $f_k$ into two sub-components described below. In addition, we incorporate the information from the geospatial profile $G$ into the prior placed over the auxiliary variable $z$, encouraging the neighboring spots to be in the same histology-based spatial domain. The number of histology-based spatial domains $K$ is decided by prior biological knowledge when available or otherwise by the elbow of the Bayesian information criterion (BIC) plot [113].
Modeling the molecular profile $Y$: We use a multivariate normal (MN) sub-component for modeling the low-dimensional gene expression $y_i$ at spot $i$:

$$y_i | z_i = k, \mu_k, \Sigma_k \sim MN(\mu_k, \Sigma_k),$$

where $\mu_k = (\mu_{k1}, \ldots, \mu_{kP'})^T$, $\mu_{kp} \in \mathbb{R}$ is the domain-specific mean vector and $\Sigma_k$ is the $P' \times P'$ domain-specific variance-covariance matrix, requiring positive definiteness. For computational efficiency, we specify a normal prior for $\mu_k$ conditional on $\Sigma_k$, and an inverse-Wishart (IW) prior for $\Sigma_k$ [114], i.e.,

$$\mu_k | \Sigma_k \sim MN(\nu_0, \Sigma_k/\tau_0)$$

$$\Sigma_k \sim IW(\eta_0, \Phi_0).$$

This conjugate setting leads to analytically tractable posterior distributions on $\mu_k$ and $\Sigma_k$. Here, $\nu_0$, $\tau_0$, $\eta_0$, and $\Phi_0$ are fixed hyperparameters. We set $\nu_0$ to be the empirical mean vector over all spots and $\tau_0 = 0.01$ to provide a weak prior information so that the data itself would dominate the estimation of $\mu_k$. We set the degree of freedom parameter $\eta_0 = P' + 1$, controlling the informative strength, and the scale matrix $\Phi_0$ to be the identity matrix. Let $n_k = \sum_{i=1}^N I(z_i = k)$ and $\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^N (I(z_i = k)y_i)$, the closed-form posterior distributions are

$$\mu_k | \Sigma_k, Y \sim MN(\nu_k, \Sigma_k/\tau_k),$$

$$\Sigma_k | Y \sim IW(\eta_k, \Phi_k),$$

where $\tau_k = \tau_0 + n_k$, $\eta_k = \eta_0 + n_k$, $\nu_k = (\nu_0\tau_0 + n_k\bar{y}_k)/(\tau_0 + n_k)$, and $\Phi_k = \Phi_0 + \sum_{i=1}^N I(z_i = k)(y_i - \bar{y}_k)(y_i - \bar{y}_k)^T + \frac{n_{mk}}{\tau_0 + n_k}(\bar{y}_k - \nu_0)(\bar{y}_k - \nu_0)^T$.

Suppose we choose PCA to perform an orthogonal projection of the scaled and normalized SRT molecular profiling data, we can further set all off-diagonal entries in $\Sigma_k$ to be zero,
i.e., $\sigma_{kp'} = 0, \forall p \neq p'$. In this case, the multivariate normal model can be decomposed into a product of $P'$ independent normal model,

$$
y_i \mid z_i = k, \mu_k, \sigma_k^2 \sim \prod_{p=1}^{P'} N(\mu_{kp}, \sigma_{kp}^2)
$$

, where $\sigma_k^2 = (\sigma_{k1}^2, \ldots, \sigma_{kP'}^2)^T$ is the set of diagonal entries in $\Sigma_k$. The conjugate setting for each dimension becomes a normal-inverse-gamma (IG) distribution [114],

$$
\mu_{kp} \mid \sigma_{kp}^2 \sim N(0, \sigma_{kp}^2/\tau_0), \quad \sigma_{kp}^2 \sim IG(\eta_0/2, \Phi_0/2),
$$

resulting in the closed-form posteriors

$$
\mu_{kp} \mid \sigma_{kp}^2, y_p \sim N(0, \sigma_{kp}^2/\tau_k), \quad \sigma_{kp}^2 \mid y_p \sim IG(\eta_k/2, \Phi_k/2),
$$

where $\tau_k = \tau_0+n_k, \eta_k = \eta_0+n_k$, and $\Phi_k = \Phi_0+\sum_{i=1}^{N} I(z_i = k)(y_{ip} - \bar{y}_{kp})^2 + (n_k \tau_0)/((\tau_0+n_k)\bar{y}_{kp}^2)$. One standard way of setting a weakly informative IG prior is to choose small values of both parameters, such as $\eta_0/2 = \Phi_0/2 = 0.1$.

**Modeling the image profile $V$:** We use a multinomial sub-component for modeling the number of cells with different types $v_i$ within spot $i$ and its expanded area:

$$
v_i \mid z_i = k, \omega_k \sim \text{Multi}(m_i, \omega_k),
$$

where $m_i = \sum_{q=1}^{Q} v_{iq}$ is the total number of cells observed within the area and $\omega_k = (\omega_{k1}, \ldots, \omega_{kQ})^T$ is defined on a $Q$-dimensional simplex (i.e., $\omega_{kq} > 0, \forall q$ and $\sum_{q=1}^{Q} \omega_{kq} = 1$), representing the underlying relative abundance of cell types in histology-based spatial domain $k$. Of particular note is that $\omega_1, \ldots, \omega_K$ are the parameters of key interest in iIMPACT, because it can be used to interpret or even define the identified histology-based spatial domains. For example, if a histology-based spatial domain is heavily dominated by cell type
\( q \), i.e., \( \omega_{kq} \gg \omega_{kq'} \), \( \forall q' \), then it could be named after cell type \( q \). Note that cell type abundance is assumed to be homogeneous across the same histology-based spatial domain. For computational efficiency, we specify a Dirichlet prior setting for \( \omega_k \), i.e.,

\[
\omega_k \sim \text{Dirichlet}(\alpha_0),
\]

where \( \alpha_0 = (\alpha_{01}, \ldots, \alpha_{0Q})^T \), \( \alpha_{0q} \in \mathbb{R}^+ \) are fixed hyperparameters. This conjugate setting leads to an analytically tractable posterior distribution on

\[
\omega_k | V \sim \text{Dirichlet}(\alpha_k)
\]

with each entry \( \alpha_{kq} = \alpha_{0q} + \sum_{i=1}^{N} [I(z_i = k)v_i] \). We recommend \( \alpha_{01} = \cdots = \alpha_{0Q} = 0.5 \) or 1 for a non or weakly informative setting.

**Incorporating the geospatial profile \( G \):** To utilize the available spatial information in the geospatial profile, we employ a Markov random field prior [58, 115] on the histology-based spatial domain indicator \( z \), encouraging neighboring spots to be clustered into the same histology-based spatial domain:

\[
\pi(z_i = k | z_{-i}) \propto \exp \left( d_k + f \sum_{i' = 1}^{N} [g_{ii'} I(z_{i'} = k)] \right)
\]

where \( z_{-i} \) denotes the set of all entries in \( z \) excluding the \( i \)-th one, the hyperparameters \( d = (d_1, \ldots, d_N)^T \) control the number of spots belonging to each of the \( K \) histology-based spatial domains and \( f \in \mathbb{R}^+ \) controls the spatial dependency or smoothness. Note that if a spot has no neighbors, the above prior distribution reduces to a multinomial distribution, \( z_i \sim \text{Multi}(N, \exp(d)/\sum_{k=1}^{K} \exp(d_k)) \). Although the larger the \( f \), the smoother the pattern of spatial domains, careful determination of \( f \) is required. This is because a large value of \( f \) may lead to a phase transition problem (i.e., all spots are assigned to the same histology-based spatial domain) [76]. In this paper, we choose \( d_1 = \cdots = d_K = 1 \) and \( f = 1 \) by default,
as this setting performs very well in the simulation study and yields reasonable results in our real data analysis.

**Data likelihood:** iIMPACT integrates the molecular, image, and geospatial profiles to partition the whole domain into $K$ biologically meaningful spatial domains. Because the low-dimensional molecular profile $Y$ and AI-reconstructed image profile $V$ are generated from different sources, they are conditionally independent of each other. Thus, we define the mixture component

$$f_k(x_i = \{y_i, v_i\}|z_i = k, \theta_k = \{\mu_k, \Sigma_k, \omega_k\}) = \text{MN}(y_i; \mu_k, \Sigma_k) \left[\text{Multi}(v_i; m_i, \omega_k)\right]^w,$$

where the tuning parameter $w \in [0, 1]$ controls the image profile's contribution to the clustering process, with respect to that of the molecular profile. Parameterizing the data likelihood above by decreasing $w$ will result in a flatter multinomial distribution, thus downplaying the role of the image profile. When $w = 0$, iIMPACT will not depend on any cell type abundance information. We conducted a sensitivity analysis to search for the best choice of $w$ for both 10x Visium and STARmap data (see Figure B.2 in Appendix B). Our result suggests setting $w = 0.05$ and 0.5 for 10x Visium and STARmap data, respectively. Note that in addition to the SRT platform and application, we should also consider the image and molecular profiles’ dimensionalities (i.e., $Q$ and $P'$) to determine the value of $w$ with some degree of caution.

3.2.4. A generalized linear regression model to detect domain-specific spatially variable genes

To test if each gene is differentially expressed among those identified histology-based spatial domains in Stage I of iIMPACT, we use a generalized linear regression model, where the response variable is gene expression counts, and the predictor variables are the histology-based spatial domain indicators. In particular, we assume that all read counts from a gene
$j$ across different spots indexed by $i$ are from an NB distribution:

$$c_{ij} \sim \text{NB}(s_i \lambda_{ij}, \psi_j), \ i = 1, \ldots, N$$

where $s_i$ is the size factor of spot $i$, $\psi_j$ is the over-dispersion parameter of gene $j$, and $\lambda_{ij}$ is the underlying normalized expression level for gene $j$ at spot $i$. We further use the canonical link,

$$\ln(\lambda_{ij}) = \alpha_{jk} + \beta_{jk} x_{ik},$$

which is typically used in the Poisson and NB regression models. Here, $x_{ik} = I(z_i = k)$ is a binary indicator. If spot $i$ is assigned to histology-based spatial domain $k$ in Stage I of iIMPACT, then $x_{ik} = 1$; otherwise, $x_{ik} = 0$. Thus, we can interpret the intercept $\alpha_{jk}$ as the baseline expression level of gene $j$ in the whole domain excluding histology-based spatial domain $k$, and the slope $\beta_{jk}$ as the differential expression level of gene $j$ in histology-based spatial domain $k$ as a shift from the baseline. With this modeling framework, SVGs, which are differentially expressed in a given histology-based spatial domain $k$ compared with all other domains, can be identified via testing the null hypothesis $H_0 : \beta_{jk} = 0$ versus the alternative $H_\alpha : \beta_{jk} \neq 0$. For those genes whose resulting adjusted p-values are less than a significance level (e.g., 0.05), we define them as domain-$k$-specific spatially variable genes. To control the false discovery rate, the Benjamini and Hochberg method [85] needs to be applied to adjust p-values. The above NB regression model is fitted via the function `glm.nb` in the $R$ package MASS [116].
3.3. Model Fitting

3.3.1. MCMC algorithms

We give the full posterior distribution of the Bayesian finite mixture model as,

\[
\pi(z, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \omega_1, \ldots, \omega_K | Y, V) \propto \prod_{k=1}^{K} \prod \{i: z_i = k\} \text{MN}(y_i; \mu_k, \Sigma_k) \text{Multi}(v_i; m_i, \omega_k)^w \times \pi(z) \times \prod_{k=1}^{K} \text{MN}(\mu_k; \nu_0, \Sigma_k/\tau_0) \text{IW}(\Sigma_k; \eta_0, \Phi_0) \times \prod_{k=1}^{K} \text{Dir}(\omega_k; \alpha_0)^w
\]

To identify histology-based spatial domains, the posterior distribution of \(z_i\) will be of direct interest to us, given by

\[
\pi(z_i = k | \cdot) \propto \text{MN}(y_i; \mu_k, \Sigma_k) \times [\text{Multi}(v_i; m_i, \omega_k)]^w \times \exp(d_k + f \sum_{i' = 1}^{N} [g_{ii'} I(z_{i'} = k)]).
\]

The individual quantities of all possible values of \(z_i\) are first computed and then summed to find the normalization constant \(e = \sum_{k=1}^{K} \pi(z_i = k | \cdot)\). A new value of \(z_i\) can be drawn from a multinomial distribution \(\text{Multi}(1, (\pi(z_i = 1 | \cdot)/e, \ldots, \pi(z_i = K | \cdot)/e)^T)\). For any particular domain-specific parameters, i.e., \(\mu_k, \Sigma_k, \omega_k\), we only require the partial data likelihood in estimating its posterior density as detailed before. Since the posterior conditional distributions for all parameters are in closed form, it is straightforward to use a Gibbs sampler, a type of Markov chain Monte Carlo (MCMC) algorithm, to obtain a sequence of observations approximated from the multivariate distribution \(\pi(z, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \omega_1, \ldots, \omega_K | Y, V)\). The details of the MCMC algorithm is in Appendix B.1. Consequently, the posterior inference can be made by post-processing the MCMC samples, such as \(\{z^{(1)}, \ldots, z^{(U)}\}\) and
\{\omega_{k}^{(1)}, \ldots, \omega_{k}^{(U)}\}, \text{ where } u \text{ indexes the MCMC iteration and } U \text{ is the total number of iterations after burn-in.}

In any finite mixture model, the invariance of the likelihood under permutation of the cluster labels \(z\) may result in an identifiability problem, leading to symmetric and multimodal posterior distributions with up to \(K!\) copies of each genuine model. What is worse, it will also complicate inference on other parameters. To address this issue, we impose an order restriction on the posterior samples of parameters \(\omega_{1}, \ldots, \omega_{K}\) based on a given cell type \(q\). In particular, at each iteration \(u\), we relabel \(z\) and switch all the related domain-specific parameters of the MCMC outputs to satisfy the constraint \(\omega_{kq}^{(u)} > \omega_{k'q}^{(u)}\) for cluster indicator \(k < k'\). In other words, the first histology-based spatial domain has the largest proportion of cell type \(q\), while histology-based spatial domain \(K\) has the smallest proportion of cell type \(q\).

3.3.2. Posterior inference

**Identifying histology-based spatial domains and interactive zones**: Our primary interest lies in identifying histology-based spatial domains via making inferences on the spatial domain indicator vector \(z\). Here we apply the mode estimates 54 based on the marginal probabilities \(\pi(z_i = k | \cdot) \approx 1/U \sum_{u=1}^{U} I(z_i^{(u)} = k)\). The estimate of \(\hat{z}_i\) can be obtained by selecting the highest value:

\[
\hat{z}_i = \text{argmax}_{k} \pi(z_i = k | \cdot) \approx \text{argmax}_{k} I(z_i^{(u)} = k).
\]

Uncertainty quantification is one advantage of the proposed Bayesian finite mixture model. For example, if the marginal probability of assigning spot \(i\) to histology-based spatial domain \(k\) is considerably high, e.g., \(\pi(z_i = k | \cdot) \geq 0.9\), then we are confident about the assignment. However, if some marginal probabilities are almost equivalent or there is no significant mode
for a spot, e.g., \( \pi(z_i = k | \cdot) < 0.9, \forall k \), then we tend not to assign the spot to any histology-based spatial domains. Instead, we define the spot as the boundary spot, and the resulting connected area as the interactive zone.

**Interpreting and defining histology-based spatial domains:** The domain-specific relative abundance of cell types \( \omega_1, \ldots, \omega_K \) are another group of parameters of interest in our model, because it can be used to interpret or even define the identified histology-based spatial domains. We use the posterior mean as the estimate,

\[
\hat{\omega}_{kq} = \mathbb{E}[\omega_{kq}|\cdot] \approx \frac{1}{U} \sum_{u=1}^{U} \omega_{kq}^{(u)},
\]

averaging over all its MCMC samples. Additionally, the credible interval for each \( \omega_{kq} \) can be approximated by its post-burn-in MCMC sample quantiles. Note that the MCMC samples can also be used to approximate any other quantity of interest that analytical solution is impossible, e.g., \( \pi(\omega_{kq} > \omega_{k'q}|\cdot) \) for some \( k, k', \) and \( q \).

### 3.4. Simulation Study

#### 3.4.1. Data generative model

The simulated data were generated based on the \( K = 5 \) histology-based spatial domains identified by iIMPACT on the human breast cancer dataset (Figure 3.3B). The posterior means of those domain-specific relative abundances of cell types \( \hat{\omega}_k \)'s and mean vectors of low-dimensional representation of gene expression \( \hat{\mu}_k \)'s are given in Table B.4. Using this real data information, we generated the cell type abundance for each spot from a multinomial distribution

\[
v_i | z_i = k \sim \text{Multi}(m_i, \hat{\omega}_k),
\]
where the size parameter $m_i$ was also obtained from real data. For generating high-dimensional gene expression counts, we first projected the $P'$-dimensional domain-specific mean vectors $\mu_k$’s on to the original basis, denoted by a $P$-dimensional vector $\hat{\mu}_k$. To mimic the excess zeros and over-dispersion, we sampled each gene expression count $c_{ij}$ from a zero-inflated negative binomial (ZINB) distribution,

$$c_{ij} | z_i = k \sim \pi_i I(c_{ij} = 0) + (1 - \pi_i) \text{NB}(s_i \exp(\tilde{\mu}_{kj}), \psi_j), \quad (3.1)$$

where the size factors $s_i$’s were sampled from a log-normal distribution with mean 0 and standard deviation 0.2, i.e., $s_i \sim \text{LN}(0, 0.2)$. The dispersion parameters $\psi_j$’s were sampled from an exponential distribution, i.e., $\psi_j \sim \text{Exp}(\lambda_{\psi})$, with two choices of the rate parameter $\lambda_{\psi} = 0.1$ or 0.2, corresponding to low and high variability. The false zero proportion parameters $\pi_i$’s were set to be 0.3 or 0.5, corresponding to low and high sparsity. For each of the four scenarios in terms of $\lambda_{\psi}$ and $\pi_i$, we independently repeated the above steps to generate 10 replicated simulated datasets.

3.4.2. iIMPACT and competing method settings

We chose the number of reduced dimensions as $P' = 3$ in the PCA step for obtaining the low-dimensional representation of gene expression levels $Y$. The number of histology-based spatial domains was fixed at $K = 5$. We followed the recommended prior setting, as detailed in the METHODS section. As for the MCMC algorithm, we ran four independent MCMC chains with $U = 10,000$ iterations, discarding the first half as burn-in. We started each chain from a model by randomly drawing all parameters from their prior distributions. The results reported were obtained by pooling the MCMC outputs from the four chains after. We compared the performance of iIMPACT on spatial domain identification with four current state-of-the-art methods, Louvain [26,117], stLearn [31], BayesSpace [29], and SpaGCN [24].
We used the default setting of each competing method, as suggested by the authors. The number of spatial domains was fixed as $K = 5$ for all methods.

3.4.3. Model performance

We quantified the clustering performance via the widely used adjusted Rand index (ARI). It ranges from $-1$ to $1$, with higher values indicating greater consistency between the clustering results and the ground truth. The results are shown in Figure 3.2. iIMPACT substantially outperformed all other methods, exhibiting the highest average ARI, under all four scenarios, which highlights the benefit of integrating the cell type abundance information into the spatial domain identification process. SpaGCN also demonstrated superior performance compared to other competing methods, leveraging its ability to effectively utilize histology information. Conversely, stLearn, despite its capability to incorporate histology images, had unsatisfactory clustering accuracy and performed similarly to Louvain, a non-spatial clustering method. BayesSpace had a large variance among replicates since it might fail to converge for some replicates. Comparing the performance under low and high variability settings, iIMPACT was robust to the level of over-dispersion of gene expression counts due to the normalization and dimensionality reduction procedures employed before the clustering model. However, it suffered from reduced clustering accuracy under high sparsity settings.

3.5. Real Data Analysis

3.5.1. Case study on human breast cancer 10x Visium dataset

We applied iIMPACT to analyze an SRT dataset from a human breast cancer study. This dataset includes 2,518 spots and 17,651 genes. The gene expression was measured on a section of human breast with invasive ductal carcinoma via the 10x Visium platform, along with annotation from pathologists that was used as the ground truth for algorithm
Figure 3.2: Simulation study: The boxplots of ARIs achieved by iIMPACT, SpaGCN, BayesSpace, stLearn, and Louvain under different scenarios in terms of sparsity and variability settings.

comparison purposes (H&E-stained image with five annotated tissue regions in Figure 3.3A). After applying HD-Staining [100] to the histology image of breast cancer tissue, we identified 156,235 cells within seven categories: macrophage, ductal epithelium, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell (Detailed information in Figure B.4 in Appendix B).

Firstly, we compared the five spatial domains defined by iIMPACT, SpaGCN, and BayesSpace, with manually annotated domains by pathologists. We found that iIMPACT achieved the highest consistency with the manual annotation (See Figure 3.3B. Adjusted Rand Index (ARI) = 0.634), while BayesSpace mistakenly clustered the tumor region into two groups. It is worth noting that SpaGCN took only the image RGB values instead of detailed histology information, which might contribute to the unsatisfactory performance on segmenting the non-tumor regions (ARI = 0.520). In contrast, the better performance of iIMPACT suggests the advantage of integrating both molecular and image profiles in the clustering analysis of SRT data.
Figure 3.3: Real data analysis on human breast cancer dataset: A. H&E-stained image of the tissue section with ground truth labels from pathologists. B. Spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, with the number of clusters to be five. C. Estimates (posterior means and credible intervals) of domain-specific relative abundance of cell types for the seven cell types observed in the AI-reconstructed histology image. D. Interactive zones (black asterisk spots) defined by iIMPACT. E. Identified interactive zones (black asterisk spots) and other boundary areas of tumor and its adjacent Domain 3, and boxplots of gene expression richness for spots in the interactive zone and other boundaries. F. Gene enrichment analysis between SVGs detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and known breast cancer genes from the COSMIC database. G. Spatial expression patterns of two example SVGs, COX6C and ELF3, that were only detected by iIMPACT.
Secondly, iIMPACT is able to define each individual histology-based spatial domain through its underlying relative abundance of cell types parametrized by the Bayesian multinomial-normal mixture model (Figure 3.3C). In contrast, SpaGCN and BayesSpace, despite their good capabilities in identifying spatial domains, currently lack the ability to effectively integrate cell type information and interpret the identified domains in a biologically meaningful way. For example, as detailed in Figure 3.3C, the proportion of tumor cells is higher in domain 1 (green spots in Figure 3.3B) than in other domains, indicating that domain 1 is the tumor region. This inference is consistent with tumor regions in the manual annotation. Domain 2 (blue) and domain 5 (red) have a similar proportion of stromal cells, while the proportion of lymphocytes in domain 2 is higher than in domain 5. The difference in the relative abundance of cell types may indicate the functional difference between these two domains. These examples confirm that iIMPACT is able to provide biological interpretation of spatial domains.

Thirdly, iIMPACT can identify the interactive zones among histology-based spatial domains (Figure 3.3D). Interactive zones are spots with higher uncertainty on domain allocation, which potentially have higher diversity in cell type abundance and heterogeneity in gene expression compared with neighboring spots with unambiguous domain definition. We calculated the gene expression richness, defined as the percentage of genes with non-zero read counts, for each spot in the tumor-immune interactive zone and other tumor-immune boundaries. We observed statistically significant differences among these comparisons (Figure 3.3E), implying that the identified zones are connected areas between tumor and immune domains with a high level of heterogeneity in gene expression and complex cellular interactions. By further comparing the gene expressions for these groups, we found several known cancer or immune genes with high expression in the interactive zones (e.g., GREM1 [118]), suggesting the possible tumor-immune interactions in these zones.
Finally, we asked whether the domain-specific SVGs defined by iIMPACT are more consistent with biological knowledge than those from other algorithms, which is an independent evaluation step frequently used for validating the clustering approaches on single-cell and spatial profiling data [19,20,24]. We focused on the tumor-domain specific SVGs defined by iIMPACT, SpaGCN, SpatialDE, and SPARK, respectively, and performed the enrichment analysis by comparing tumor-domain SVGs defined by these four methods with the known breast cancer gene set defined in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database. As summarized in Figure 3.3F, the tumor-domain SVGs detected by iIMPACT showed higher overlap with the known breast cancer gene set than that of SpaGCN, SpatialDE, and SPARK, respectively, including two example genes that can only be detected by iIMPACT (Figure 3.3G): COX6C (estimated coefficient $\hat{\beta} = 0.829$, with adjusted p-value $= 2.89 \times 10^{-197}$, a known biomarker for the identification of hormone-responsive breast cancer [119], and ELF3 ($\hat{\beta} = 0.930$, with adjusted p-value $= 4.82 \times 10^{-91}$, an epithelial-specific gene that is a novel therapeutic target of breast cancer and has been amplified in early breast cancer [120]. These results confirm that iIMPACT-defined SVGs align better with existing biological knowledge.

### 3.5.2. Case study on human prostate cancer 10x Visium dataset

To evaluate the performance of iIMPACT in different tissue types, we studied another SRT dataset from a human prostate cancer study, which includes 4,371 spots and 17,651 genes. The gene expression was measured on a section from invasive carcinoma of human prostate via the 10x Visium platform. We applied HD-Staining to analyze the histology image of this tissue (Figure 3.4A). 259,257 cells were segmented and classified into six categories: macrophage, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell (Detailed information in Figure B.5).
Figure 3.4: Real data analysis on human prostate cancer dataset: A. H&E-stained image of the tissue section. B. Spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to be five. C. Estimates (posterior means and credible intervals) of domain-specific relative abundance of cell types for the six cell types observed in the AI-reconstructed histology image. D. Interactive zones (black asterisk spots) defined by iIMPACT. E. Identified interactive zones (black asterisk spots) and other boundary areas of Domain 2 and Domain 3, and boxplots of gene expression richness for spots in the interactive zone and other boundaries. F. Gene enrichment analysis between SVGs detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and the known prostate cancer genes from the COSMIC database. G. Spatial expression patterns of two example SVGs, EIF3E and TBL1XR1, that were only detected by iIMPACT.
We confirmed that iIMPACT outperformed BayesSpace and SpaGCN in spatial domain detection, assuming there are five spatial domains. As shown in Figure 3.4B, these three methods could identify the domain (marked in green) with a high proportion of tumor cells, compared with the spatial distribution of tumor cells (Figure B.5C). Interestingly, iIMPACT could distinguish histology-based spatial domains with different red blood cell proportions (Figure 3.4B, yellow region vs. red region). We further compared the results of three methods with the manual annotation based on three morphologically distinguish regions: tumor, stroma and partially atrophic changes, and stroma (Figure B.6 in Appendix B). We observed that iIMPACT achieved the highest consistency with the manual annotation (ARI = 0.659).

To demonstrate the interpretability of iIMPACT, we characterized the domain-specific relative abundance of cell types in Figure 3.4C. We observed that domain 1 has a higher proportion of tumor cells than other domains, indicating that it is probably the tumor domain. Comparing domain 2 with domain 3, we observed that they have a similar proportion of tumor cells, but domain 2 has a higher proportion of immune cells (i.e., lymphocyte and macrophage), implying the heterogeneity of immune composition within tumors.

In addition, interactive zones can also be defined by iIMPACT (Figure 3.4D). By checking the interactive zones of domain 2 and 3 and calculating the gene expression richness, we observed a clear trend between the interactive zones and the surrounding boundaries, indicating the unique characteristics of interactive zones (Figure 3.4E). We further found that gene \textit{DNAJC5} [121] expressed higher on the identified interactive zones, implying its potential relationship with the intermediate areas of immune cell distribution.

We also compared iIMPACT, SpaGCN, SpatialDE, and SPARK in detecting biologically meaningful SVGs in this prostate cancer dataset. We confirmed that, for tumor-domain (domain 1) specific SVGs, iIMPACT outperformed SpaGCN, SpatialDE, and SPARK in detecting known prostate cancer genes from the COSMIC database (Figure 3.4F), illustrating...
that iIMPACT could detect SVGs that are biologically relevant. These iIMPACT-defined SVGs in tumor domains have experimental evidence to support their functional relevance to the development of prostate cancer. For example, as shown in Figure 3.4G, gene *EIF3E* ($\hat{\beta} = 0.480$, with adjusted p-value $= 1.62 \times 10^{-77}$), which is associated with increased cell cycle progression and motility in prostate cancer [122], and *TBL1XR1* ($\hat{\beta} = 0.243$, with adjusted p-value $= 2.96 \times 10^{-54}$), which displays an oncogene role for prostate cancer cell proliferation [123].

3.5.3. Case study on human ovarian cancer 10x Visium dataset

The third NGS-based SRT dataset is from a section of human ovarian tumor tissue. This dataset includes 3,455 spots and 17,651 genes. The gene expression was measured on a section of serous papillary carcinoma from human ovarian via the 10x Visium platform, with the H&E-stained image shown in Figure 3.5A, HD-Staining model segmented and classified 211,746 cells in six categories: macrophage, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell (Detailed information in Figure B.7 in Appendix B). By utilizing the cell type abundance information from the histology image, we observed that iIMPACT had better performance on spatial domain identification. By comparing the clustering results of three methods (iIMPACT, BayesSpace, and SpaGCN) with the annotated tumor and benign domains for this SRT dataset, we observed a remarkable concordance between the clustering results obtained from iIMPACT and the pathologist’s annotations (ARI = 0.967, see Figure B.8 in Appendix B). In addition, iIMPACT could identify the domain (marked in green) with a high proportion of tumor cells, which has a high consistency with the tumor region annotated by the pathologist (Figure B.8) and the region with a high amount of tumor cells (Figure B.7C), while the other two methods could not confirm it as a tumor domain (Figure 3.5B).
Figure 3.5: Real data analysis on human ovarian cancer dataset: A. H&E-stained image of the tissue section. B. Spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to be five. C. Estimates (posterior means and credible intervals) of domain-specific relative abundance of cell types for the six cell types observed in the AI-reconstructed histology image. D. Interactive zones (black asterisk spots) defined by iIMPACT. E. Identified interactive zones (black asterisk spots) and other boundary areas of tumor and its adjacent Domain 5, and boxplots of gene expression richness for spots in interactive zone and other boundaries. F. Gene enrichment analysis between SVGs detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and the known ovarian cancer genes from the COSMIC database. G. Spatial expression patterns of two example SVGs, \( BCL6 \) and \( CHD4 \), that were only detected by iIMPACT.
iIMPACT could also distinguish domains with different red blood cell proportions. Figure 3.5C shows the estimation of relative abundance of cell types for the five histology-based spatial domains. Domain 1 has a higher proportion of tumor cells than other domains, indicating that it is likely to be the tumor domain. We further examined the interactive zones (Figure 3.5D) and compared the interactive zone between domain 1 and 5 with other boundary spots (Figure 3.5E). A significant difference in gene expression richness between boundary spots and the interactive zone was observed. Furthermore, we found that gene *TTLL5* [124] and *CLEC12A* [125] have higher expression on the interactive zone between domain 1 and 5, which may infer their potential relationship with the tumor-immune interaction.

We further detected SVGs using iIMPACT, and then queried tumor-region SVGs with the known ovarian cancer gene set defined by the COSMIC database. We observed that iIMPACT-defined ovarian cancer SVGs showed a higher overlap with the known ovarian cancer gene set than that of SpaGCN, SpatialDE, and SPARK (Figure 3.5F). Moreover, we explored these ovarian cancer SVGs only defined by iIMPACT and found that many of them possess compelling experimental evidence substantiating their functional relevance to ovarian cancer. For example, our list included gene *BCL6* ($\hat{\beta} = 0.247$, with adjusted p-value $= 3.22 \times 10^{-37}$), which displays pro-oncogenic activity in ovarian cancer [126], and *CHD4* ($\hat{\beta} = 0.220$, with adjusted p-value $= 4.33 \times 10^{-35}$), which is associated with apoptosis mediated by cisplatin in ovarian cancer cells [127] (Figure 3.5G).

### 3.5.4. Case study on mouse visual cortex STARmap dataset

To demonstrate iIMPACT is also able to analyze data from imaging-based SRT platforms, we applied iIMPACT to a STARmap dataset [14]. This dataset was generated from mouse visual cortex, including hippocampus, corpus callosum, and the neocortical layers. In total, 1,020 genes were measured in 1,207 cells with 15 cell types. The layer structure and cell
As shown in Figure 3.6A, iIMPACT displayed the most accurate clustering results with the known layer structure (ARI = 0.592). We also noticed that implementing iIMPACT on a lower resolution level (grids in Figure 3.6A) might reduce the influence of noise, thus making the clustering result more robust. We also leveraged iIMPACT to identify the interactive zones (Figure 3.6B). The majority areas of identified interacting areas were boundaries between two adjacent layers. We found these iIMPACT-defined SVGs are frequently functionally relevant to visual cortex (Figure 3.6C). For example, we observed gene Depotor ($\hat{\beta} = 0.360$, with adjusted p-value = 0.0320), which is highly expressed and functions in a significant portion of corticostriatal and callosal neurons, located in the middle and superficial portions of layer 5 (L5) [128], and Vamp1 ($\hat{\beta} = 0.363$, with adjusted p-value = 0.00608), which is ubiquitously expressed and functioned in layer III pyramidal neurons in higher-order areas [129] (Figure 3.6D).

3.6. Discussion

In this paper, we presented iIMPACT, a two-stage statistical method that integrates histology images and molecular profiles. The first stage is a Bayesian finite normal-multinomial mixture model for identifying histology-based spatial domains. Compared with the alternative methods, it fully leverages cellular-level information from histology images to improve clustering performance and increase interpretability. The second stage is a NB regression model for detecting domain-specific SVGs. From both simulation study and real data analysis, we demonstrated that iIMPACT had higher accuracy in identifying spatial domains than published state-of-the-art methods due to the integration of histopathology images in iIMPACT. In addition, iIMPACT is versatile in analyzing both NGS-based and imaging-based SRT techniques, and therefore have broad impacts in the SRT field. Furthermore, iIMPACT has good biological interpretability to characterize histology-based spatial domains. For
Figure 3.6: Real data analysis on mouse visual cortex STARmap data: A. Layer structure of the tissue section from the original study. Spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to seven (the number of layers). Manually added square lattice grid when fitting iIMPACT is displayed with dashed lines. B. Interactive zones (black asterisk spots) defined by iIMPACT. C. Gene enrichment analysis between SVGs detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and genes functionally relevant to visual cortex for five layers. D. Spatial expression patterns of two example SVGs, Deptor and Vamp3, that were only detected by iIMPACT.
example, the inferred domain-specific cell-type compositions are consistent with curated annotations, and the interactive zones emphasize the areas with highly heterogeneous cell-type composition and gene expression compared with surroundings. Compared with other SVG detection methods, iIMPACT-defined SVGs are more enriched of known functional genes, confirming that iIMPACT could provide a better understanding of both cellular spatial organization and functional gene landscape of developmental and diseased tissues. Last but not least, compared with other methods, we also confirmed that iIMPACT is computationally efficient (Table B.2 in Appendix B).

There are several important future extensions for iIMPACT. First, improvement of nuclei segmentation and classification methods might further improve the performance of iIMPACT and therefore will be our focus in the near future. Second, the number of histology-based spatial domains has to be pre-specified when implementing the current version of iIMPACT. To automatically estimate the number of spatial domains, we plan to replace the proposed Bayesian finite mixture model with a Bayesian nonparametric model, such as the Dirichlet process mixture model [130] or mixture of finite mixture model [131, 132]. Third, cell-cell interaction information can be incorporated into iIMPACT to improve the accuracy of histology-based spatial domain identification and increase the model interpretability. These future directions could potentially further boost the performance and interpretability of iIMPACT.
Chapter 4
BayesDeep: A Bayesian Hierarchical Model for Gene Expression Reconstruction at Single-cell Resolution

4.1. Background

Understanding the spatial distribution of transcript expression offers valuable insights into biological function and histopathology [133]. The advent of single-cell RNA sequencing (scRNA-seq) techniques has profoundly transformed our understanding of gene expression regulation in different cell lineages or types. However, the process of tissue dissociation in scRNA-seq results in the loss of the spatial context of gene expression, which is essential for comprehending the cell-cell and cell-environment interaction mechanism [37]. Thanks to the recent technological breakthroughs in spatially resolved transcriptomics (SRT), it is now feasible to explore the spatial organization of cells, alongside their molecular and even morphological features. Next-generation sequencing (NGS)-based spatial molecular profiling platforms, such as spatial transcriptomics (ST) [15] and 10x Visium [134] (i.e., an improved platform by 10x Genomics), utilize spatial barcodes to capture RNA molecules, synthesize their complementary DNA molecules, and subsequently perform sequencing. These techniques facilitate measuring expression for the whole genome across thousands of spatial locations, referred to as ‘spots’, on the tissue section. The development of SRT technology has provided valuable clinical and biological insights into various areas, including tumor heterogeneity, brain function, and sepsis pathophysiology [26,135,136].

One of the primary constraints inherent to NGS-based SRT techniques is their spatial resolution. Gene expression in these platforms is measured on an array of spots, with a typical spot diameter of 100 µm for ST and 55 µm for 10x Visium [29]. Consequently,
each spot’s area may encompass a heterogeneous population of cells, ranging from a few to as many as 200 cells, depending on the platform and the type of biological tissue being analyzed. Furthermore, SRT techniques solely measure gene expression within the area of those spots, thus covering part of the total cells observed within the entire tissue section. To illustrate, the area covered by measured spots accounts for approximately 1/3 of the total tissue region for the 10x Visium platform and around 20% for ST platforms (see Figure C.1 in the Appendix C for a detailed explanation). The inherent limitation in spatial resolution and the scope of regions could result in substantial information loss when analyzing SRT data. Such limitations may, in turn, impact the applicability of these techniques in facilitating the discovery of deeper insights in various biomedical studies, specifically in the context of investigating gene expression patterns at the cellular level.

Several computational methods have been developed to enhance the spatial resolution of SRT data. For instance, BayesSpace [29] employs a Bayesian approach to improve gene expression at the sub-spot resolution. However, despite this advancement, sub-spots generated by BayesSpace may still encompass multiple cells, yielding only a modest improvement in spatial resolution. Furthermore, BayesSpace lacks the capability to predict gene expression out of the spot regions. In parallel, a series of approaches, such as TESLA [34], XFuse [33], ST-Net [35], and HisToGene [36], have been developed to reconstruct gene expression at super-resolution. These deep-learning-based techniques utilize molecular information from SRT data and morphological information from the paired histology image. Notably, while these methods enhance gene expression prediction at the pixel level, which may complicate interpretation, none directly account for cellular spatial organization in their analyses.

In response, we introduce a novel Bayesian methodology, BayesDeep, for deeply resolving gene expression for all ‘real’ cells by integrating the molecular profile from SRT data and the morphological information extracted from the paired histology image. Particularly, BayesDeep builds upon a regularized negative binomial regression model with grouped ob-
servations. The response variable is the spot-resolution gene expression measurements in terms of counts, while the explanatory variables are a range of cellular features extracted from the paired histology image, including cell type and nuclei shape descriptors. Following the regression coefficient estimation, BayesDeep makes predictions on the gene expression of all cells that exhibit these cellular features, regardless of within or beyond spot regions. The regularization is achieved by employing a spike-and-slab prior distribution to each regression coefficient to improve model robustness. We validated the accuracy of gene expression prediction at both spot and single-cell resolution using simulated and real SRT datasets. Additionally, we demonstrated that the comprehensive single-cell-resolution spatial molecular profiles characterized by BayesDeep enable in-depth investigations on cell-type specific differential expression analysis, cell-cell communication analysis, and pseudo-time analysis. Overall, our findings highlight the effectiveness of BayesDeep in reconstructing molecular profiles at the single-cell resolution from NGS-based SRT data only at the spot resolution, paving the way for advanced research in pseudotime and cell-cell communication analysis.

4.2. Model

In this section, we first generally introduce the idea of BayesDeep, then define the spot-level molecular and geospatial profiles from NGS-based SRT techniques (e.g., spatial transcriptomics and the improved 10x Visium platform) and the image profile from the matching AI-reconstructed histology image. Then we discuss how to construct the spatial relationship between spots and cells. After that, we detail the statistical model used in reconstructing the single-cell-resolution gene expressions. Table C.1 in the Appendix C summarizes all notations introduced in this chapter.
Figure 4.1: Flowchart of the proposed BayesDeep: A. BayesDeep integrates the spot-resolution molecular profile $Y$ from NGS-based SRT data, the single-cell-resolution image profile $X$ from the paired AI-reconstructed histology image, and the spot-cell geospatial profile $G$ to recover gene expression at the single-cell resolution $\Theta$. B. The hierarchical formulation of the BayesDeep model, based on a Bayesian regularized negative binomial regression model with grouped observations. C. BayesDeep estimates the association between the single-cell-resolution molecular and image profiles $B$ and predicts the single-cell-resolution molecular profile $\Theta$. D. Several downstream analyses can be enhanced based on the availability of the single-cell-resolution molecular profile $\Theta$, including identifying distinct cell populations, elucidating the process of tumorigenesis via pseudotime analysis, and exploring the mechanisms of cell-cell communication.
4.2.1. Overview

BayesDeep builds upon a Bayesian negative binomial regression model to recover gene expression at the single-cell resolution from NGS-based SRT data. The schematic diagram is depicted in Figure 4.1.

BayesDeep integrates three distinct modalities from a standard NGS-based SRT experiment: the molecular, image, and geospatial profiles (see Figure 4.1A). The molecular profile refers to the spot-resolution gene expression data denoted by an \( N \)-by-\( P \) count matrix \( Y \), where \( N \) is the number of spots and \( P \) is the number of genes. The image profile corresponds to the detailed morphological context of the paired histology image in terms of a set of cellular features. We use an \( M \)-by-\( L \) design matrix \( X \) to denote the image profile, where \( M \) is the number of all observed cells and \( L \) is the number of cellular features, which may include cell types, nuclei shape characteristics, and any other relevant explanatory features that can be quantified at large scale. The geospatial profile reveals the spatial relationship between the \( N \) spots and \( M \) cells, which can be defined by an \( N \)-by-\( M \) binary matrix \( G \), with one signifying a cell is within the barcoded area of a spot.

The spot-resolution gene expression matrix \( Y \), along with the single-cell-resolution morphological features of those cells within spot regions, serve as a reference for recovering the single-cell-resolution gene expression of all \( M \) cells, whether within or beyond spot regions. The model is specified in Figure 4.1B. We first modeled the observed read count for a specific gene within a spot using a negative binomial (NB) distribution. Then, the underlying spot-resolution relative gene expression in the NB mean is assumed to be the average of single-cell-resolution relative expression across all cells within the spot. Next, we considered the logarithm of each cell’s relative expression as a linear combination of covariates that includes a scalar of one for the intercept and \( L \) measurable explanatory variables that pertain to that cell. A spike-and-slab prior model is applied for each covariate coefficient. On one hand, this feature selection scheme improves the robustness of our model. On the other
hand, the corresponding coefficient matrix $B$ uncovers significant associations between gene expression and cellular characteristics, illustrated in Figure 4.1C, thereby potentially offering valuable biological insights. With the reconstructed single-cell-resolution spatial molecular profile $\Theta$, we can undertake several pivotal downstream analyses, as depicted in Figure 4.1D. These analyses allow for the differentiation of cell populations, the exploration of tumorigenesis through pseudotime analysis, and the dissection of ligand-receptor signaling pathways vital for cell-cell communication.

4.2.2. Data preparation

BayesDeep takes spatial gene expression data from NGS-based SRT techniques, along with the location information as input. It also requires single-cell level image and geospatial profiles extracted from the histology image for the corresponding tissue section based on some nuclei identification techniques.

The spot-resolution molecular profile: We denote the SRT molecular profile (i.e., gene expression measurements in terms of counts) as an $N$-by-$P$ matrix $Y = [y_{ij}]_{N \times P}$, where each entry $y_{ij} \in \mathbb{N}$ represents the read count for gene $j$ ($j = 1, \ldots, P$, with $P$ being the total number of genes) observed at spot $i$ ($i = 1, \ldots, N$, with $N$ being the total number of spots). These $N$ spots are regularly arrayed on a two-dimensional square or triangle lattice, with their spatial coordinates given by an $N$-by-2 matrix $T^Y = [t^Y_{ir}]_{N \times 2}$, $t_{ir} \in \mathbb{N}$, where each row $t^Y_i = (t^Y_{i1}, t^Y_{i2})$ records the x and y-coordinates of spot $i$ in the designated domain.

The single-cell-resolution image profile: We denote the image profile, derived from the histology image paired with the SRT data, as an $M$-by-$L$ matrix $X = [x_{ml}]_{M \times L}$, where each entry $x_{ml} \in \mathbb{R}$ represents a measurement for explanatory variable $l$ ($l = 1, \ldots, L$, with $L$ being the total number of explanatory variables) observed for cell $m$ ($m = 1, \ldots, M$, with $M$ being the total number of cells). Our study utilizes $L$ single-cell-resolution morphological
features as the explanatory variables, which are extracted using the histology-based digital (HD)-staining model \[100\]. Specifically, the HD-staining model is a deep-learning model based on the mask regional convolutional neural network (Mask R-CNN) architecture \[110\], which is trained to segment the nucleus of various cell types, such as immune, tumor, and stromal cells. HD-staining then computes ten geometric shape features for each identified cell nucleus, including filled area, net area, convex area, extent, perimeter, solidity, eccentricity, major axis length, minor axis length, and orientation. The definitions of these shape features can be found in Table S3 of a recently published paper \[137\]. Supposing there are \(Q\) different types of cells coded in \(Q\) dummy variables, the total number of explanatory variables is \(L = Q + 10\). The HD-staining model was initially trained using histology images from lung adenocarcinoma patients in the National Lung Screening Trial (NLST) study, wherein nuclei of six different cell types were manually annotated by well-experienced pathologists. Although the original model was trained by using data specific to lung cancer, it has been improved and validated to adapt to histology images from various cancer types, including breast cancer, prostate cancer, and other carcinomas. To represent cell nuclei locations, we use an \(M\)-by-2 matrix \(T^X = [t^X_{mr}]_{M \times 2}\), where each row \(t^X_m = (t^X_{m1}, t^X_{m2}) \in \mathbb{N}^2\) records the x and y-coordinates of cell \(m\) in the designated domain.

**The spot-cell geospatial profile:** In the context of NGS-based SRT technologies, spots are defined as circular regions comprising barcoded mRNA capture probes, where gene expression is quantified within a given tissue section \[15\]. The SRT data provides both the physical spot diameter \(d\) and its corresponding length in pixels. Assuming that the x and y-coordinates in spot-resolution geospatial profile \(T^Y\) represent the locations of spot centers, then we can identify whether cell \(m\) is located within spot \(i\) by a direct evaluation of the condition \(\sqrt{(t^X_{m1} - t^Y_{i1})^2 + (t^X_{m2} - t^Y_{i2})^2} \leq d/2\). This allows us to construct the spot-cell geospatial profile, denoted by an \(N\)-by-\(M\) binary matrix \(G = [g_{im}]_{N \times M}\), which reflects whether a given cell \(m\) is within the barcoded area of a given spot \(i\) (i.e., \(g_{im} = 1\)) or not (i.e., \(g_{im} = 0\)). Notably, the coverage of measured areas in NGS-based SRT techniques is relatively...
limited, encompassing approximately 38% of the overall tissue section. As a result, most cells do not fall within the boundaries of any of the defined spots. In our real data analysis, for instance, 97,135 out of $M = 156,115$ cells for the human breast cancer 10x Visium data, and 252,765 out of $M = 352,818$ cells for the human prostate cancer 10x Visium data, are not covered by the measured area of any spots. This limitation in the scope of measured regions can lead to substantial information loss when analyzing SRT data. This motivates us to develop BayesDeep to address this challenge and enhance the analysis of SRT data.

4.2.3. A Bayesian negative binomial regression model with grouped observations

BayesDeep is essentially a negative binomial regression model with regularization for handling grouped observations under the Bayesian framework. Its primary objective is to utilize the spot-level molecular profile, single-cell-resolution image profile, and spot-cell geospatial profile to reconstruct the single-cell-resolution molecular profile, which can be represented by an $M$-by-$P$ matrix $\Theta = [\theta_{mj}]_{M \times P}$, where each entry $\theta_{mj} \in \mathbb{R}^+$ is the predictive relative expression for gene $j$ within cell $m$. It is worth noting that BayesDeep focuses on reconstructing relative gene expression at the single-cell resolution for one gene at a time (i.e., a specific column vector in $\Theta$), which allows for efficient parallel processing of multiple genes.

**Modeling the spot-resolution molecular profile:** We start by modeling the over-dispersed spot-resolution gene expression matrix $Y$ using a negative binomial (NB) distribution. NB-based models are widely used for analyzing sequence count data [21, 62, 138, 139] due to their ability to accommodate inherent over-dispersion. Specifically, we postulate that each observed read count follows an NB distribution:

$$y_{ij}|\lambda_{ij}, \phi_j \sim \text{NB}(s_i\lambda_{ij}, \phi_j)$$
Here the NB($\nu, \phi$) distribution is parameterized in terms of its mean $\nu \in \mathbb{R}^+$ and dispersion $1/\phi \in \mathbb{R}^+$, with the probability mass function given by $\frac{\Gamma(y+\phi)}{y! \Gamma(\phi)} \left( \frac{\phi}{\nu+\phi} \right)^\phi \left( \frac{\nu}{\nu+\phi} \right)^y$. This parameterization provides the flexibility to characterize an unknown mean-variance structure, with the variance calculated $\nu + \nu^2 / \phi$. A small value of $\phi$ indicates a high variance to mean ratio (i.e., $(\nu + \nu^2 / \phi) / \nu = 1 + \nu / \phi$), while a large value approaching infinity reduces the NB distribution to a Poisson distribution with the same mean and variance. The NB mean is further decomposed of two multiplicative components, 1) the size factor, denoted as $s_i$, and 2) the spot-resolution relative expression for gene $j$ observed at spot $i$, denoted as $\lambda_{ij}$. Such a multiplicative characterization of the NB or Poisson mean is typical in both the frequentist [140–142] and the Bayesian [143,144] literature when modeling sequence count data. The set of size factors is represented as $s = [s_i]_{N \times 1}$, capturing a range of biological and technical variabilities across samples, such as reverse transcription efficiency, amplification/dilution efficiency, and sequencing depth [145]. To ensure identifiability between these two parameters, we set $s_i$ proportional to the summation of the total read counts across all genes at spot $i$ [20]. With a constraint of $\prod_{i=1}^N s_i = 1$, we compute $s_i = \sum_{j=1}^P y_{ij} / (\prod_{i=1}^N (\sum_{j=1}^P y_{ij}))$. It is noteworthy that alternative methods for estimating $s$ are available, such as setting each $s_i$ to be proportional to the upper-quartiles of non-zero read counts across all genes at spot $i$ [68], or even modeling $s$ through a Dirichlet process mixture model with mean constraint [146].

To complete the prior model specification, we place a common gamma prior on all dispersion parameters, denoted as $\phi = [\phi_j]_{P \times 1}$, that is, $\phi_j \sim \text{Gamma}(a_{\phi}, b_{\phi})$, where $a_{\phi}$ and $b_{\phi}$ are fixed hyperparameters. We recommend choosing small values, such as $a_{\phi} = b_{\phi} = 0.1$, to maintain a weakly informative setting.

**Linking the spot- and single-cell-resolution molecular profiles:** We assume that the spot-resolution relative expression for gene $j$ (i.e., $\lambda_{ij}$) can be derived as an average of single-cell-resolution relative expression (i.e., $\theta_{mij}$’s) across all cells within the given spot $i$. 
This relationship can be formally expressed as:

\[ \lambda_{ij} = \frac{1}{\sum_{m=1}^{M} g_{im}} \sum_{m=1}^{M} (g_{im} \theta_{mj}) \]

where the indicator variable \( g_{im} \) in the spot-cell geospatial profile \( G \) takes the value one if cell \( m \) is located within the barcoded area of spot \( i \), and zero otherwise. The denominator in the above formula, \( \sum_{m=1}^{M} g_{im} \), represents the total number of cells within spot \( i \) observed in the paired histology images. From an alternative perspective, this assumption can be interpreted as the spot-resolution gene expression being an average of cell-type-specific gene expression within the same spot, weighted by the respective cell-type proportions. Notably, this foundational assumption finds validations in various cell-type deconvolution algorithms designed for SRT data [39,147,148]. From a statistical viewpoint, this assumption enables the application of a regression model tailored to accommodate grouped observations [149,150].

**Linking the single-cell-resolution molecular and image profiles:** In our pursuit of predicting the single-cell-resolution gene expression \( \theta_{mj} \)'s, we leverage morphological information encompassing cell type and nuclei shape features of each cell \( m \). Specifically, we adopt a linear model to capture the relative expression for gene \( j \) within cell \( m \):

\[ \log(\theta_{mj}) = \beta_{0j} + \sum_{l=1}^{L} \beta_{lj} x_{ml} \]

The choice of a log-link function is driven by the necessity for maintaining \( \theta_{mj} \) values in the positive domain, thereby ensuring a positive NB mean. In this formulation, \( \beta_{0j} \) is the baseline expression for gene \( j \) shared by all cells. Note that \( \exp(\beta_{0j}) \) can also be interpreted as a scaling factor that adjusts for gene-specific effects. As previously introduced, we use the \( M \)-by-\( L \) matrix \( X \) to present observations from \( L \) explanatory variables extracted from the histology image, including cell type information, cell nuclei shape descriptors, and other important single-cell-resolution measurements relevant to the analysis. Given the coding of \( Q \)
cell types into $Q$ dummy variables, we added a constraint $\sum_{q=1}^{Q} \beta_{qj} = 0$ to avoid identifiability issues arising from the sum of the components. Consequently, $\exp(\beta_{lj})$ that corresponds to cell type $Q$, represents the mean relative expression for gene $j$ across all cells. Each column in the $L$-by-$P$ coefficient matrix $B = [\beta_{lj}]_{L \times P}$, denoted as $\mathbf{\beta}_j = [\beta_{lj}]_{L \times 1}$ with $\beta_{lj} \in \mathbb{R}$, describes the effect of $L$ explanatory variables on the logarithm of the relative expression across cells for gene $j$. Hence, we can use the coefficient matrix $B$ to explore the association between the image and molecular profiles at the single-cell resolution. In practice, for a specific gene $j$, it is likely that only a limited number of explanatory variables account for gene expression. For instance, the expression of some housekeeping genes exhibits small variability among different tissues, cell types, or samples [151]. Under such circumstances, their corresponding coefficients $\beta_{lj}$’s tend to be zeros. On the other hand, some pairs of explanatory variables might be highly correlated, leading to an inflation of regression coefficients and potentially harming prediction performance. To prevent over-fitting and reduce the potential multi-collinearity, we employ a regularization mechanism by specifying a spike-and-slab prior [152] on each $\beta_{lj}$. This prior is a mixture of distributions:

$$\beta_{lj} \sim (1 - \gamma_{lj})\delta_0(\beta_{lj}) + \gamma_{lj}\text{Norm}(0, \sigma_\beta^2)$$

with the spike component $\delta_0(\beta_{lj})$ being a point mass distribution at $\beta_{lj} = 0$, while the slab component being a normal distribution centered at zero. If the auxiliary binary variable $\gamma_{lj} = 1$, then the probability of $\beta_{lj} = 0$ is zero, indicating that explanatory variable $l$ is relevant for explaining the relative expression for gene $j$. Conversely, $\gamma_{lj} = 0$ restricts that $\beta_{lj} = 0$, indicating explanatory variable $l$ has no contribution to the gene $j$’s expression. This prior setting enables us to identify significant associations between gene expression and explanatory variables, which, in our case, are the morphological features. The variance of the slab component $\sigma_\beta^2 \in \mathbb{R}^+$ is a fixed hyperparameter set to 1. We complete the model specification by placing an independent Bernoulli distribution on $\gamma_{lj}$, i.e., $\gamma_{lj} \sim \text{Bern}(\pi_\gamma)$,
where $\pi_\gamma \in (0, 1)$ is a fixed hyperparameter that indicates the percentage of explanatory variables included in the final model a priori. We set $\pi_\gamma = 0.5$ to incorporate relatively weak information.

4.3. Model Fitting

The model parameter space consists of 1) the dispersion parameter $\phi$ that accounts for the over-dispersion commonly observed in gene expression data, 2) the coefficient matrix $B$ that quantifies the relationship between gene expression as measured in SRT data and the morphological features extracted from the paired histology image, and 3) the $L$-by-$P$ selection matrix $\Gamma = [\gamma_{lj}]_{L \times P}$ that indicates the significant association in the coefficient matrix $B$. The complete data likelihood can be written as:

$$
f(Y|\phi, B, \Gamma) = \prod_{j=1}^{P} \prod_{i=1}^{N} \text{NB}\left( y_{ij}; s_i \sum_{m=1}^{M} \frac{1}{g_{im}} \sum_{m=1}^{M} \left( g_{im} \exp \left( \sum_{l=1}^{L} \beta_{lj} x_{ml} \right) \right), \phi_j \right) \quad (4.1)$$

With the prior specifications detailed above, the full posterior distribution can be written as:

$$
\pi(\phi, B, \Gamma|Y) = f(Y|\phi, B, \Gamma) \times \pi(B|\Gamma) \times \pi(\Gamma) \times \pi(\phi)
$$

$$
= \prod_{j=1}^{P} \prod_{i=1}^{N} \text{NB}\left( y_{ij}; s_i \sum_{m=1}^{M} \frac{1}{g_{im}} \sum_{m=1}^{M} \left( g_{im} \exp \left( \sum_{l=1}^{L} \beta_{lj} x_{ml} \right) \right), \phi_j \right)
$$

$$
\times \prod_{j=1}^{P} \prod_{l=1}^{L} \left( ((1 - \gamma_{lj})\delta_0(\beta_{lj}) + \gamma_{lj}\text{Norm}(\beta_{lj}; 0, \sigma_{\beta}^2))\text{Bern}(\gamma_{lj}; \pi_\gamma) \right)
$$

$$
\times \prod_{j=1}^{P} \text{Gamma}(\phi_j; a_\phi, b_\phi) \quad (4.2)
$$
4.3.1. MCMC algorithms

We explore the posterior distribution via a Markov chain Monte Carlo (MCMC) algorithm based on stochastic search variable selection [153, 154]. Specifically, we iteratively update each parameter using a Metropolis-Hasting (MH) algorithm. We note that this algorithm is sufficient to guarantee ergodicity for our model. Full details regarding the MCMC algorithm implementation are available in the Appendix C.

4.3.2. Posterior inference

We obtain posterior inference through post-processing the MCMC samples following the burn-in phase. Let \( \phi_j^{(u)} \), \( \beta_{ij}^{(u)} \) and \( \gamma_{lj}^{(u)} \) denote the posterior samples in the \( u \)-th iteration after burn-in, where \( u = 1, \ldots, U \). Our primary focus is on the selection of the important explanatory variables for each gene \( j \), via the selection matrix \( \Gamma \). An effective approach to summarize the posterior distributions of these binary parameters is by computing the marginal posterior probability of inclusion (PPI):

\[
PPI(\gamma_{lj}) = \pi(\gamma_{lj} = 1 | \cdot) \approx \frac{1}{U} \sum_{u=1}^{U} \gamma_{lj}^{(u)}
\]

Then, we set \( \hat{\gamma}_{lj} = 1 \) if its corresponding PPI exceeds a pre-specified threshold. When choosing the threshold, we recommend either opting for a threshold of 0.5, leading to a median model, or following a procedure that controls the expected Bayesian false discovery rate [155]. For each dispersion parameter \( \phi_j \) and each coefficient \( \beta_{ij} \), we estimate them by calculating their posterior means through averaging over all their respective MCMC samples after burn-in,

\[
\hat{\phi}_j = \mathbb{E}[\phi_j | \cdot] \approx \frac{1}{U} \sum_{u=1}^{U} \phi_j^{(u)}
\]
and

\[ \hat{\beta}_{lj} = \mathbb{E}[\beta_{lj} | \cdot] \approx \frac{1}{U} \sum_{u=1}^{U} \beta_{lj}^{(u)} \]

Additionally, a quantile estimation or credible interval for each parameter of interest can be obtained from MCMC samples.

### 4.3.3. Predictive inference

Our primary goal is to reconstruct gene expression at the single-cell resolution by estimating the matrix \( \Theta = [\theta_{mj}]_{M \times P} \). On the basis of the MCMC samples on the coefficient matrix \( B = [\beta_{lj}]_{L \times P} \), we predict each \( \theta_{mj} \) by Monte Carlo simulation. Specifically, at each iteration \( u \) after burn-in, we compute the relative gene expression for gene \( j \) within cell \( m \) as follows:

\[
\theta_{mj}^{(u)} = \exp \left( \beta_{0j}^{(u)} + \sum_{l=1}^{L} \beta_{lj}^{(u)} x_{ml} \right)
\]

Subsequently, we sample the gene expression for gene \( j \) within spot \( i \) using an NB distribution,

\[
y_{ij}^{(u)} \sim \text{NB}(s_i \left( \sum_{m=1}^{M} (g_{im} \theta_{mj}^{(u)}) / \left( \sum_{m=1}^{M} g_{im} \right) \right), \phi_j^{(u)})
\]

Consequently, both single-cell-resolution relative gene expression \( \hat{\theta}_{mj} \) and spot-resolution gene expression \( \hat{y}_{ij} \) can be estimated by summarizing their corresponding MCMC samples. For instance, their predictive means can be approximated as follows:

\[
\hat{\theta}_{mj} = \frac{1}{U} \sum_{u=1}^{U} \theta_{mj}^{(u)} \quad \text{and} \quad \hat{y}_{ij} = \left\lceil \frac{1}{U} \sum_{u=1}^{U} y_{ij}^{(u)} \right\rceil
\]

where \( \lceil \cdot \rceil \) denotes the ceiling function.
4.4. Simulation Study

4.4.1. Model validation on simulated data

We designed a simulation study to validate the accuracy of our BayesDeep. We selected a connected region of \( N = 500 \) spots from the human breast cancer 10x Visium data (displayed as green circles in Figure 4.2A) to generate the spot-resolution molecular profile. The single-cell-resolution image profile \( X \), including cell types and nuclei shape features, and the locations of cells, were from the nuclei identification results by HD-staining for the paired histology image of the SRT data. As introduced in section 4.2 and observed in real SRT data analysis, the covariate coefficient matrix \( B \) is highly sparse, indicating that many explanatory variables do not contribute to gene expression for most genes. To replicate this condition, we generated the coefficients, ensuring nearly half were zeros, reflecting the observed zero-inflation. Based on the model assumption of BayesDeep, we first generated single-cell-resolution relative gene expression matrix \( \Theta \) on the selected region, and then the spot-resolution gene expression matrix \( Y \) on all spots within. The data-generating procedure is detailed in the Appendix C.

We evaluated the performance of BayesDeep in recovering the single-cell-resolution molecular profile \( \Theta \), against two other methods, TESLA and Gaussian process (GP), by measuring the Pearson correlation coefficient \( \rho \in [-1, 1] \) between the logged estimated and actual gene expression at the single-cell resolution, \( \hat{\theta}_{mj} \)’s and \( \theta_{mj} \)’s. Using relative gene expression allows us to discern which cells exhibit high gene expression and which demonstrate lower expression. A high positive \( \rho \) value signifies the successful model outcome and robust overall performance. As depicted in Figure 4.2A, the results demonstrate that BayesDeep, when employing regularization, significantly surpasses the comparative methods, with a median \( \rho \) of 0.928. In contrast, TESLA and GP failed to reconstruct the single-cell-resolution molecular profile \( \Theta \), as indicated by their respective \( \rho \) values approaching zero. To further examine
the efficacy of the regularization approach, we computed the root mean square error (RMSE) between the estimated and actual covariate coefficients, $\hat{\beta}_{lj}$’s and $\beta_{lj}$’s, with lower RMSE values denoting more precise estimations. Specifically, in scenarios where regularization was omitted, a standard normal prior was applied in place of the spike-and-slab prior to each covariate coefficient $\beta_{lj}$. These results demonstrate that the incorporation of regularization significantly refines the precision of estimations for non-zero coefficients and substantially improves accuracy for truly zero coefficients (Figure C.2 in the Appendix C). These findings support the implementation of a regularized prior within our BayesDeep framework.

4.4.2. Model validation on real spatially resolved transcriptomics data with spot resolution

In this study, we utilized the human breast cancer and human prostate cancer 10x Visium data in the real data analysis to demonstrate the efficacy of BayesDeep in restoring gene expression within manually masked spots. For each SRT data, we randomly selected 100 genes with non-zero read counts in at least 50% spots. We assessed BayesDeep against TESLA and GP, through two distinct masking scenarios for spot-resolution gene expression imputation. In scenario 1, 40% of spots were randomly masked, while in scenario 2, we divided the domain into two halves and masked all spots on one side (refer to Figure 4.2B, with masked spots marked in orange and unmasked spots marked in green). The second scenario aimed to challenge the model’s predictive power over a separate region that does not overlap with those masked spots. The gene expression of those remaining unmasked spots $Y$ served as the main input of each method. For BayesDeep, we also incorporated the morphological information $X$, detailed in section 4.2, from the corresponding histology image.

Figure 4.2B shows the correlation coefficient $\rho$ for comparing observed and imputed spot-resolution gene expression in the masked spots within the human breast cancer 10x Visium
Figure 4.2: Overview of model validation, including the validation settings and evaluation results in terms of the Pearson correlation coefficients $\rho$ between the actual and predicted gene expression for BayesDeep, TESLA, and Gaussian Process (GP), respectively. The validation is stratified into three distinct data: A. Simulated data at the spot resolution; B. Human breast cancer 10x Visium data at the spot resolution; C. Mouse visual cortex STARmap data at the single-cell resolution.
data. BayesDeep achieved a median correlation of 0.585 in scenario 1, significantly higher than that of the competing methods (TESLA: $\rho = 0.358$; GP: $\rho = 0.391$). In scenario 2, the performance of BayesDeep is comparable to scenario 1, whereas TESLA and GP failed to impute gene expression for the entirely masked half, as evidenced by negligible correlation values. This failure is attributed to their spatial dependence assumption in reconstructing gene expression, resulting in inadequate performance when the masked and observed regions do not intersect. Figure C.3 in Appendix C shows similar results through the analysis on human prostate cancer 10x Visium data. These results show the capability of BayesDeep to predict gene expression at the spot resolution, both in the same and new regions, by leveraging single-cell-resolution morphological features.

4.4.3. Model validation on real spatially resolved transcriptomics data with single-cell resolution

We extended our study to assess the performance of BayesDeep on the prediction of single-cell-resolution gene expression when ground truth is available at the single-cell resolution. We validated the model using the mouse visual cortex STARmap data [14], which was collected by an imaging-based SRT platform. It comprises measurements of 1,020 genes across $M = 1,207$ cells, categorized into $Q = 15$ distinct cell types distributed across seven layers labeled in the original study. In our analysis, we filtered out genes that are expressed in less than 30% cells or the highest read count is less than ten. With these filtering criteria, we kept $P = 77$ genes for the following analysis. We then constructed the spot-resolution molecular profile $\mathbf{Y}$ by overlaying a square lattice grid across the entire domain, with each square unit representing a 'spot' of 750 $\times$ 750 pixels, as illustrated in Figure 4.2C. This yielded a total of $N = 105$ spots, each of which contains more than one cell, covering the entire domain without any inter-spot gaps. The read count $y_{ij}$ of gene $j$ for spot $i$ was calculated by aggregating the read counts from all cells within that spot. Owing to the
absence of a corresponding histology image for the data, we utilized the cell type and layer information provided along with the SRT data in the original study as explanatory variables $X$ to inform BayesDeep.

Figure 4.2C visualizes the correlation coefficient $\rho$ between the imputed and true relative expression across all cells for all genes. BayesDeep achieved a median correlation of approximately 0.336, similar to TESLA (median $\rho = 0.331$) but significantly outperforming GP (median $\rho = 0.065$). Additionally, Figure 4.2C illustrates the actual versus predicted relative expression for four representative genes, demonstrating a notable agreement between the actual and predicted gene expression patterns. These findings affirm the capability of BayesDeep to accurately reconstruct the expression patterns of the underlying cells directly from spot-resolution gene expression.

4.5. Real Data Analysis

4.5.1. Case study on human breast cancer 10x Visium dataset

We applied BayesDeep to reconstruct the single-cell-resolution molecular profiles for the SRT data from a human breast cancer study. The data includes $N = 2,518$ spots and 17,651 genes. The gene expression was measured on a section of human breast with invasive ductal carcinoma via the 10x Visium platform, along with annotation from pathologists as a reference (H&E-stained image with five annotated tissue regions in Figure 4.3A). After applying HD-Staining [100] to the histology image of breast cancer tissue, we identified $M = 156,235$ cells within seven categories: macrophage, ductal epithelium, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell. Table C.2 in Appendix C provided ten shape features used in this study. We applied BayesDeep to reconstruct the single-cell-resolution molecular profiles on the top $P = 2,000$ highly variable genes, compared its
performance to a competing method, TESLA, and further utilized the generated higher-resolution molecular profiles on several downstream analysis to reveal biological insights.

The gene expression at the single-cell resolution, generated by BayesDeep, offers a more detailed view of the spatial transcriptome landscape within the cellular environment. To demonstrate the capacity of BayesDeep to enhance our understanding of gene expression in regions with high cell type heterogeneity, we chose both a tumor and an immune-related region as illustrative examples (Figure 4.3A). For the example tumor region, we observed two distinct small areas densely populated with tumor cells. We further examined and presented the original spot-resolution expression from SRT data and reconstructed expression using BayesDeep and TESLA of three breast cancer-related genes (Figure 4.3B): First of all, $AZGP1$, which is primarily expressed in breast epithelial cells, plays a multifaceted role associated with cancer cachexia, carcinogenesis, and tumor differentiation [156]. Secondly, $MUCL1$, a breast-specific gene predominantly expressed in breast cancer, serves as a vital biomarker for tumor progression and metastasis [157, 158]. Last but not least, $TOP2A$ serves as a notable proliferation marker, demonstrating high expression in various subtypes of breast cancers [159]. At spot resolution (ST data in Figure 4.3B), the detailed expression pattern is relatively difficult to observe due to the low resolution. However, the expression reconstructed by BayesDeep for these three example genes distinctly exhibit the expression pattern consistent with the distribution of tumor cells, i.e., high expression on tumor cells and low expression on other cells. Conversely, for the gene expression reconstructed by TESLA, no similar expression pattern is observed in the two tumor areas for all three genes, indicating its limitation in recovering the comprehensive cellular expression. Violin plots in Figure 4.3B show the higher expression on tumor cells and ductal epithelium for BayesDeep, which aligns with the gene functions and their association with breast cancer. In contrast, the predicted expression from TESLA shows no difference among different cell types. Another observed region in Figure 4.3A is an immune region predominantly comprised of lymphocytes, which covers around 56 spots in SRT data. In this region, a large number of lymphocytes
Figure 4.3: Real data analysis on the human breast cancer 10x Visium data: A. Manual annotation and nuclei identification results by applying HD-staining to the paired histology image; and the two selected tumor and immune regions for an illustrative comparison between BayesDeep and TESLA. B. The actual expression from ST data at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected tumor region for genes AZGP1, MUC1, and TOP2A. Violin plots display gene expression on the selected tumor region across different cell types. C. The actual expression from ST data at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected immune region for genes CORO1A, IGH2, and TRBC1. Violin plots display gene expression on the selected immune region across different cell types.
encircle a small tumor area. We selected three example genes related to immune functions for validation purpose. *CORO1A* is an identified immunity gene signature in breast cancer cohort studies [160]. *IGHG2* encodes the constant region of the immunoglobulin gamma-2 heavy chain, which enables antigen binding activity and immunoglobulin receptor binding activity, and is involved in several processes, such as the activation of immune response [161]. *TRBC1* encodes T cell receptor β chain constant region 1, which is partially expressed in subsets of T cells [162, 163]. At the spot-resolution, no obvious expression patterns can be observed due to the limited resolution (Figure 4.3C). However, through single-cell-resolution reconstruction via BayesDeep, we observed that these three genes exhibited high expression in the surrounding immune area. Again, gene expression recovered by TESLA show no difference among various cell types and spatial regions. In summary, BayesDeep-predicted single-cell-resolution gene expression exhibit significantly stronger associations with the cell type information, unveiling detailed gene expression patterns not captured by the ST data alone.

In addition, BayesDeep quantifies the association between gene expression and the cellular features extracted from the histology image via the estimation of the covariate coefficient matrix $B$. In Figure 4.4A, the heatmap displays the coefficients of seven distinct cell types and ten nuclei shape descriptors for the top 2,000 highly variable genes. It is noticed that a substantial number of the estimated covariate coefficients $\beta_{ij}$’s are non-zero and vary across cell types for most genes, which highlights unique expression patterns for different cell types. On the other hand, certain genes are characterized by all $\beta_{ij}$’s being zero, implying homogenous expression across cell types. In contrast, the covariate coefficients of those shape descriptors excluding solidity are notably sparse, with approximately 87.0% of those coefficients being zero. There exists an association between the gene expression and the solidity of cell nucleus for around 50.1% of genes. It is important to note that the orientation covariate (i.e., the third column from right) is a negative control. This covariate is expected to be independent of gene expression, given that the measurement of nuclear orientation...
Figure 4.4: Real data analysis (continued) and downstream analysis on the human breast cancer 10x Visium data: A. Heatmap of the covariate coefficient matrix $\mathbf{B}$ estimated by BayesDeep, indicating the association between gene expression and morphological features extracted from the paired histology image. B. Cell population analysis on the BayesDeep- and TESLA-generated single-cell-resolution gene expression. C. Pseudotime analysis on tumor cells extracted from the cell population analysis. D and E. Cell-cell interactions inferred from the BayesDeep- and TESLA-generated single-cell-resolution gene expression, and their overlap.
relies on the tissue’s placement on the slide. We observed that the estimated coefficients for the orientation covariate are exactly zero for all genes, indicating the validity of BayesDeep results.

To demonstrate that BayesDeep-reconstructed single-cell-resolution gene expression data can improve downstream analysis, we conducted a comparative analysis between BayesDeep and TESLA across three tasks: (i) distinguishing different cell populations, (ii) explicating tumorigenesis through pseudotime analysis, and (iii) elucidating cell-cell communication patterns. First, we employed Seurat for clustering and dimensional reduction analysis on the BayesDeep-generated single-cell-resolution gene expression. The results demonstrated that cells belonging to the same cell type co-localized well on the UMAP plot and exhibited clear separation among distinct cell types (Figure 4.4B left panel). In contrast, we performed a similar analysis on the gene expression generated by TESLA. However, the results showed that TESLA was unable to effectively separate many cell types (Figure 4.4B right panel). These findings underscore the capability of BayesDeep in distinguishing different cell populations. Second, we employed Moncle3 [164, 165] to perform a pseudotime analysis on tumor cells extracted from BayesDeep and TESLA results. In such analysis, we incorporated CD44 and CD24 as breast cancer stem cell markers, as previously reported [166–168]. We designated the cells with the highest CD44 and CD24 expression as the root of the trajectory to assess cancer progression over pseudotime (Figure C.4 in Appendix C). Predicted results based on BayesDeep indicated that cancer cells from with earlier pseudotime are predominantly enriched in the conventional tumor region, whereas those from later pseudotime are primarily concentrated within the necrotic tumor regions [169] (Figure 4.4C, left panel). This trajectory pattern is consistent with the timing order of tumor evolution. In contrast, the pseudotime derived from TESLA-generated predictive results did not show a biologically meaningful trend (Figure 4.4C, right panel). Overall, our results demonstrate the consistent superiority of BayesDeep in explicating tumorigenesis through the pseudotime analysis. Third, because BayesDeep can identify various cell types more accurately
than TESLA, we hypothesized that BayesDeep could offer more details of cell-cell communications among these cell types, in comparison with TESLA. To corroborate this hypothesis, we applied CellChat [45] to define the cell-cell communication landscape based on the cell clusters identified by pathologist annotations. Both BayesDeep and TESLA can achieve abundant cell-cell interactions but BayesDeep can reveal more cellular communications than TESLA (Figure 4.4D). Interestingly, the majority \(\frac{1,001}{1,486} = 67.3\%\) of cell-cell communications revealed by BayesDeep overlapped with those identified by TESLA (Figure 4.4D). However, some detailed cell-cell interactions revealed by BayesDeep and TESLA are different (Figure 4.4E): for example, BayesDeep could identify more immune-related, especially macrophage-related cellular communications than TESLA. In general, consistent with our expectation, BayesDeep can reveal more cell-cell communications than TESLA.

4.5.2. Case study on human prostate cancer 10x Visium dataset

To assess the adaptability of BayesDeep across diverse tissue types, we conducted analysis utilizing another SRT data derived from human prostate cancer tissue. The data includes \(N = 4,371\) distinct spots for 17,651 genes. The gene expression measurement was obtained from a section of invasive carcinoma within the human prostate (H&E-stained image with six annotated tissue regions in Figure 4.5A), utilizing the 10x Visium platform. Subsequently, we applied the HD-Staining technique to identify nuclei on the histology image of this tissue. This image analysis process led to the segmentation of a total of \(M = 352,818\) individual cells, which were systematically categorized into six classes: macrophage, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell. We utilized BayesDeep to reconstruct high-resolution molecular profiles for top 2,000 highly variable genes at the single-cell resolution. We then assessed its performance against the TESLA method and leveraged the resulting detailed molecular profiles in several downstream analyses.
Figure 4.5: Real data analysis on the human prostate cancer 10x Visium data: A. Manual annotation and nuclei identification results by applying HD-staining to the paired histology image; and the two selected tumor and immune regions for an illustrative comparison between BayesDeep and TESLA. B. The actual expression at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected tumor region for genes *ADGRF1*, *SPON2*, and *TMEFF2*. Violin plots display gene expression on the selected tumor region across different cell types. C. The actual expression at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected immune region for genes *CD24*, *CD47*, and *CXCR4*. Violin plots display gene expression on the selected immune region across different cell types.
To examine the gene expression reconstructed at the single-cell resolution by BayesDeep and TESLA, we selected two example regions with high cell-type mixture - a tumor and an immune-related region (in Figure 4.5A). For the example tumor region depicted in Figure 4.5A, we observe that tumor cells have circular patterns around each empty region. We examined the expression of three prostate cancer-related genes (Figure 4.5B). ADGRF1 is an adhesion-G protein-coupled receptor and has an important function in cancer [170]. Aberrant expression and mutation of G protein-coupled receptors and their signaling partners, G proteins, have been well documented in many forms of cancer [171]. SPON2 (tumor cell-derived spondin 2) is an extracellular matrix glycoprotein and overexpression of SPON2 has been shown to promote tumor cell migration [172]. TMEFF2 encodes a transmembrane protein containing an epidermal growth factor-like motif and two follistatin domains, which is highly expressed in primary samples of prostate cancer [173]. At the spot resolution in Figure 4.5B, the expression patterns of these genes remain obscure. However, the gene expression reconstructed by BayesDeep shows strong expression patterns. These three example genes have higher expression in tumor cells. In contrast, the gene expression reconstructed by TESLA does not display differential expression on tumor cells. The violin plots in Figure 4.5B provide evidence of the differential expression among cell types for BayesDeep, in line with the three example genes’ cancer-related functions. For another selected region, as depicted in Figure 4.5A, we find an immune region in the center primarily composed of lymphocytes. For our analysis, we selected three immune-related genes. CD24 is a cell surface glycosyl-phosphatidylinositol–anchored protein expressed on a variety of cell types, including developing T and most B lymphocytes [174]. CD47 is an immunoglobulin superfamily pentatransmembrane protein that is ubiquitously expressed in hematopoietic cells, including lymphocytes [175]. CXCR4 is the receptor for the CXC chemokine stromal-derived-factor-1, which has essential functions on immunological functions and T lymphocyte trafficking [176]. At the spot resolution, differential expression in the small immune region is not readily observable since the region is only covered by around five spots. However,
through single-cell-resolution reconstruction with BayesDeep, we can clearly see that these three genes exhibit high expression within the immune area. And violin plots show their high expression within immune cells, such as lymphocytes or macrophage. In contrast, TESLA fails to detect gene expression changes that occur only in a small region.

Figure 4.6: Real data analysis (continued) and downstream analysis on the human prostate cancer 10x Visium data: A. Heatmap of the covariate coefficient matrix $B$ estimated by BayesDeep, indicating the association between gene expression and morphological features extracted from the paired histology image. B. Cell population analysis on the BayesDeep- and TESLA-generated single-cell-resolution gene expression. C. Pseudotime analysis on tumor cells extracted from the cell population analysis. D and E. Cell-cell interactions inferred from the BayesDeep- and TESLA-generated single-cell-resolution gene expression, and their overlap.
The inference of covariate coefficients in BayesDeep reflects the correlation between single-cell-resolution molecular profiles and extracted cell characteristics from histology images. Figure 4.6A shows a heatmap of coefficients for six cell types and ten nuclei shape covariates among the top 2,000 highly variable genes. Most genes exhibit non-zero coefficients for at least one cell type, indicating the differential expression of these cell types. For nuclei shape features, the estimated coefficients for nuclei shape features are highly sparse, with around 90.4% being zero. The estimated coefficients on the covariate for negative control, orientation, are exactly zero for all genes, further validating the efficiency of BayesDeep.

Furthermore, we compared BayesDeep and TESLA on their impact of downstream analysis. First, through clustering and dimensional reduction analysis of the prostate cancer single-cell-resolution results generated by BayesDeep and TESLA, we found that the BayesDeep UMAP plot, but not the TESLA UMAP plot, revealed clear clustering results for various cell types (Figure 4.6B). Second, in the pseudotime analysis of prostate cancer, we used \textit{ITGA6} and \textit{ALCAM} as prostate cancer stem cell markers, as previously reported [166–168] (Figure C.5 in Appendix C). Trajectory analysis of the prostate cancer cells, extracted from BayesDeep results, revealed an even distribution of cancer cells with varying pseudotime throughout the prostate tumor. In contrast, TESLA results indicated a preference for prostate cancer cells with earlier pseudotime to be enriched at the tumor periphery (Figure 4.6C). Third, we assessed the ability of BayesDeep in identifying cell-cell communication of prostate cancer, in comparison with TESLA. BayesDeep identified 281 ligand-receptor interacting pairs, which was more than twice the number identified by TESLA (Figure 4.6D). In summary, these findings demonstrated BayesDeep could outperform TESLA in multiple downstream analysis tasks and therefore might provide deeper insights into the underlying molecular mechanisms that regulate tumor microenvironment and tumorigenesis.
4.6. Discussion

Here we introduced BayesDeep, a Bayesian model for the reconstruction of gene expression at the single-cell resolution for NGS-based SRT data. BayesDeep links the cellular morphological information extracted from histology images and the spot-resolution molecular profiles derived from SRT data, enabling the inference of single-cell-resolution molecular profiles. Specifically, BayesDeep employs a negative binomial distribution to model spot-resolution gene expression and considers the latent normalized gene expression as an average of single-cell-resolution gene expression across all cells within a given spot. Our findings demonstrated the superior performance of BayesDeep compared to other competing methods for gene expression reconstruction in both simulation study and real SRT data.

Furthermore, by applying BayesDeep to two real SRT data, we unveiled detailed expression patterns that cannot be captured at the spot resolution, which offers invaluable insights for further exploration of the subsequent biomedical research. Firstly, cell populations can be separated to characterize different cell types or subpopulations based on the predicted single-cell-resolution molecular profiles, which can be further applied to marker gene identification and automated annotation [177]. Secondly, pseudotime analysis on the reconstructed high-resolution molecular profiles helps to understand the temporal progression of cellular states or trajectories within the tumor microenvironment, which further aids in unraveling the dynamic nature of tumors, providing valuable information on how cells change and evolve over time [178]. Last but not least, reconstructed molecular profiles at single-cell resolution enable the examination of cell-cell communication patterns, which further contributes to a deeper understanding of cellular behavior, intercellular signaling dynamics, and their implications in clinical studies [45]. In summary, the downstream analysis indicates the significance of BayesDeep in exploring the tumor microenvironment and advancing the study of molecular biology.
There are several important future extensions for BayesDeep. First, the performance of BayesDeep highly relies on the selection of cellular features included in the model. On the simulated data where the actual gene expression is artificially generated and thus totally determined by the included covariates, BayesDeep has the ability to reconstruct single-cell-resolution gene expression highly consistent with the actual expression. However, for the model validation on mouse visual cortex STARmap data, the limited cellular characteristic data available may constrain a comprehensive reconstruction of gene expression, thus resulting in relatively low consistency with the actual gene expression. Therefore, the improvement of nuclei segmentation and classification methods might offer more accurate biological and morphological characteristics for cells and, in turn, may further promote the performance of BayesDeep. Second, it is necessary to optimize the computational efficiency of BayesDeep for large-scale studies, especially for larger data that are becoming increasingly common with the advancement of SRT technologies. In addition, there is potential for the BayesDeep model to integrate external data or other types of omics data, such as scRNA-seq data, to enhance gene expression inference. The integration of other data types could provide a more comprehensive understanding of the cellular environment and the complex interactions within cells. Last but not least, we anticipate that BayesDeep could be extended to generate molecular profiles for three-dimensional tissue regions. We have displayed the ability of BayesDeep to extrapolate gene expression in a region with a similar cellular environment but without SRT data. With the available three-dimensional image profiles, BayesDeep can be adjusted to build molecular profiles at single-cell resolution on three-dimensional tissue sections, and further provide more detail and context to cellular interactions and tissue architecture. These future directions could potentially further boost the performance and generalizability of BayesDeep.
CHAPTER 5
Conclusion and Future Directions

This thesis introduces three statistical frameworks developed for the analysis of spatially resolved transcriptomics (SRT) data. These models consider several key aspects: 1) the inherent characteristics of SRT data, encompassing high dimensionality, sparsity, and overdispersion observed in sequencing counts of gene expressions; 2) the spatial dependencies within geospatial profiles; and 3) the integration of multiple sources of information, such as geospatial profiles and the corresponding histology image. Leveraging Bayesian models, these frameworks address distinct challenges in SRT data analysis, including 1) the detection of spatially variable genes; 2) the identification of spatial domains; and 3) the enhancement of gene expression resolution. In all, the proposed Bayesian models provide more powerful tools for SRT analysis and help our understanding of the functional and structural foundations of tissue architecture.

In Chapter 2, our proposed method, BOOST-MI, has shown great promise in detecting the spatially variable genes, as illustrated in both simulation and two real data analyses. Our model framework can be naturally extended to other spatial dependency analysis scenarios. The current version of BOOST-MI is an energy-based Ising model, which requires gene expression to be measured on a lattice grid. To overcome these potential limitations, we are developing a generalized energy-based framework, BOOST-HMI to model gene expression measured from imaging-based SRT platforms, accommodating the irregular spatial distribution of measured cells. The Bayesian model applies a zero-inflated negative binomial mixture model to dichotomize the raw count data. Additionally, we incorporate a geostatistical mark interaction model with a generalized energy function, where the interaction
parameter is used to identify the spatial pattern. Auxiliary variable MCMC algorithms were employed to sample from the posterior distribution with an intractable normalizing constant. This proposed method is able to identify genes with novel and strong spatial patterns for SRT data generated from various platforms. Moreover, we can generalize BOOST-MI to allow heterogeneous spatial patterns, which is more practical as the measured areas for SRT techniques are increasing. The computational cost of BOOST-MI can be further reduced to improve the model scalability, so that BOOST-MI can be implemented to SRT datasets resolved by newly developed technologies with large sample size.

In Chapter 3, we introduce iIMPACT, a multi-stage Bayesian approach that includes a finite mixture model to identify and define spatial domains based on AI-reconstructed histology images and the spatial context of gene expression measurements. Additionally, a negative binomial regression model is employed to detect domain-specific spatially variable genes. One potential future direction is to eliminate the dimensionality reduction step and incorporate feature selection in the mixture model, enabling the automatic selection of genes relevant to domain determination. We present BayesCafe, a zero-inflated negative binomial mixture model designed to cluster spots or cells based on their molecular profiles [63]. To enhance interpretability, we utilize a feature selection mechanism to provide a low-dimensional summary of the SRT molecular profile, focusing on discriminating genes that contribute to the clustering result. We further integrate the SRT geospatial profile via a Markov random field prior. This proposed method attains high domain identification accuracy and offers additional biological insights by generating a list of discriminating genes through the feature selection process. Moreover, to automatically estimate the number of spatial domains, we intend to replace the proposed Bayesian finite mixture model with a Bayesian nonparametric model, such as the Dirichlet process mixture model [130] or the mixture of finite mixture model [131,132]. These future directions hold the potential to further enhance the scalability and interpretability of iIMPACT.
Due to the low-resolution limitation of existing NGS-based SRT techniques, in Chapter 4, we propose BayesDeep, a Bayesian hierarchical model combining cellular spatial information for the precise reconstruction of gene expressions at the single-cell resolution for NGS-based SRT data. Several future extensions for BayesDeep should be considered. Firstly, enhancing nuclei segmentation and classification methods could provide more precise and specific biological and morphological characteristics for cells, potentially enhancing BayesDeep’s performance. Secondly, the computational costs in the current version are suboptimal, and future improvements may involve implementing advanced Bayesian computation methods to reduce these costs. In addition, there is potential for the BayesDeep model to integrate external data or other types of omics data, such as scRNA-seq data, to enhance gene expression inference. The integration of other data types could provide a more comprehensive understanding of the cellular environment and the complex interactions within cells.
Appendix A
Appendix of Chapter 2

A.1. Full Details of Data Normalization

Normalization is critical to the analysis of sequence count data that suffer from various sequence artifacts and bias. We provide eight normalization methods in two categories.

Table A.1: List of size factors used for normalizing sequence count data.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>$s_i \propto 1$</td>
</tr>
<tr>
<td>TSS</td>
<td>$s_i \propto Y_i$</td>
</tr>
<tr>
<td>Q75 [68]</td>
<td>$s_i \propto q_i^{0.75}$</td>
</tr>
<tr>
<td>RLE [69]</td>
<td>$s_i \propto \text{median}<em>j \left{ y</em>{ij} / \sqrt[#j]{\prod_{j'=1}^{#j} y_{ij'}} \right}$</td>
</tr>
<tr>
<td>TMM [70]</td>
<td>$s_i \propto Y_i \cdot \exp \left( \frac{\sum_{j \in G^<em>} \psi_j(i,r) M_j(i,r)}{\sum_{j \in G^</em>} \psi_j(i,r)} \right)$</td>
</tr>
</tbody>
</table>

Note 1: $q_i^x$ is defined as the $x$-th sample quantile of all the counts in sample $i$, i.e., there are $xp$ features in sample $i$ whose $y_{ij}$’s are less than $q_i^x$.

Note 2: The $M$-value $M_j(i,r) = \log(y_{ij}/Y_i)/\log(y_{rj}/Y_r)$ and $A$-value $A_j(i,r) = (\log y_{ij}/Y_i + \log y_{rj}/Y_r)/2$ are the ratio and average of log-scaled counts between sample $i$ and the reference sample $r$, respectively. $G^*$ denote a subset of genes whose $M$-values are not within the upper and lower 30% of all $M$-values and $A$-values are not within the upper and lower 5% of all $A$-values. The weight $\psi_j(i,r) = \frac{Y_i - y_{ij}}{y_{ij} Y_i} + \frac{Y_r - y_{rj}}{y_{rj} Y_r}$ is the inverse of the approximate asymptotic variances.

The first type is based on size factor estimation. Let $s_i$ be the size factor of sample $i$, capturing all nuisance effects. Each relative gene expression level can be computed as $\tilde{y}_{ij} = y_{ij}/s_i$. If the main interest is in the absolute gene expression level, then all $s_i$’s are set to the same value (e.g., $s_1 = \ldots = s_n = 1$); otherwise, we compute $s_i$’s directly from the gene expression count data. The simplest way is to set $s_i \propto Y_i = \sum_{j=1}^p y_{ij}$, i.e., the total number of counts across all genes in each sample (known as sequencing depth or library size). Note that SPARK [20] suggests this normalization, namely total sum scaling (TSS).
Table A.2: List of variance-stabilizing transformations (VST) used for normalizing sequence count data.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-VST [179]</td>
<td>( g(y_{ij}, \phi) = \sinh^{-1} \sqrt{y_{ij}/\phi} )</td>
</tr>
<tr>
<td>N-VST [180]</td>
<td>( g(y_{ij}, \phi) = \sinh^{-1} \sqrt{y_{ij}+3/8 \phi^{-3/4}} )</td>
</tr>
<tr>
<td>log-VST [19,179]</td>
<td>( g(y_{ij}, \phi) = \log(y_{ij} + \phi/2) )</td>
</tr>
</tbody>
</table>

Note 1: \( \sinh^{-1}(y) = \log(y + \sqrt{1 + y^2}) \).

Note 2: \( \phi \) can be estimated via a non-linear model: \( s_j^2 = \bar{y}_j + \bar{y}_j/\phi, j = 1, \ldots, p \), where \( \bar{y}_j \) and \( s_j^2 \) denote the sample mean and variance of all the counts belonging to gene \( j \), i.e., \( (y_{ij}, \ldots, y_{nj}) \).

In practice, we could consider other estimations on \( s_i \)'s, which have been introduced for mitigating the influence of extremely low and high counts when analyzing bulk RNA-seq data, such as upper-quartiles (Q75) [68], relative log expression (RLE) [69], and weighted trimmed mean by M-values (TMM) [70]. Table A.1 provides the definitions of the above size factor estimations. The size factor estimation is usually combined with some constraint, such as \( \prod_{i=1}^{n} s_i = 1 \).

The other type of normalization method is based on variance-stabilizing transformation (VST), which aims to transform a random variable with a negative binomial distribution into one with an approximately normal distribution. There are three options: Naïve, Anscombe [179], and logarithm, namely N-VST, A-VST, and log-VST, all of which can be abstracted as \( \tilde{y}_{ij} = g(y_{ij}, \phi) \), where \( g \) is the transformation-specific function (see Table A.2) and \( \phi \) is the dispersion parameter estimated from the count data. Then, the relative gene expression levels are further adjusted for the log-scale total read counts, i.e., \( \log Y_i \), via a linear regression model. Note that SpatialDE [19] employs the log-VST normalization before fitting the geostatistical model.

We conducted a sensitivity analysis to investigate how different normalization methods affect the SV gene identification. We simulated ten replicated datasets following the data generating process described in Section 2.4. We include only the scenario with the medium
zero-inflation setting (30% false zeros) and the two real spatial patterns (MOB II and BC). We assessed the model performance in terms of the area under the curve (AUC). The result is summarized in Figure A.1. We found that the methods based on size factor estimation significantly outperformed the VST-based methods. BOOST-MI was robust to the four size factor-based normalization methods. We conducted the analysis of variance (ANOVA) test on all pairs of the four size factor-based normalization methods. All $p$-values were greater than 0.05, confirming no significant difference among the four choices.

A.2. Full Details of Data Dichotomization

After correcting for sequence artifacts and bias, we denoise the relative expression levels by partitioning all spots into two groups. This step outputs the suitable data type required in the subsequent analysis and makes BOOST-MI more robust in the face of over-dispersion and zero-inflation.

There is no consensus on the dichotomization of spots based on either absolute or relative expression level. BinSpect [23] suggests allocating those spots with the top 30% relative expression levels to the high-expression group and the remaining to the low-expression group. Meanwhile, it also considers $k$-means ($k = 2$) as an alternative to avoid choosing a hard percentage rank cutoff.
We provide two choices of clustering methods. The first one is similar in spirit to BinSpect-k-means, searching the \( \xi_j \) corresponding to the minimum within-cluster sum of squares,

\[
\arg \min_{\xi_j} \sum_{k=0}^{\frac{1}{2}} \sum_{i=1}^{n} I(\xi_{ij} = k)(\tilde{y}_{ij} - m_k)^2,
\]

where \( I(\cdot) \) denotes the indicator function, \( m_k = \sum_{i=1}^{n} \tilde{y}_{ij} I(\xi_{ij} = k) / n_k \) and \( n_k = \sum_{i=1}^{n} I(\xi_{ij} = k) \) are the sample mean and size of each group. However, before applying k-means, we first exclude those spots whose \( \tilde{y}_{ij} \)'s are larger than \( \tilde{q}_{0.75}^j + 3(\tilde{q}_{0.75}^j - \tilde{q}_{0.25}^j) \), where \( \tilde{q}_{x}^j \) is defined as the \( x \)-th sample quantile of the relative expression levels of gene \( j \), and then removes those spots with \( y_{ij} = 0 \). Note that in the context of box-and-whisker plotting, a data point is defined as an extreme outlier if it stands outside this limit. We directly allocate those discarded spots in the first and second steps to the low and high-expression groups, respectively. Compared with BinSpect, this additional preprocessing leads to a more robust performance when excessive zeros and outliers are presented.

In addition to k-means that is implicitly based on pairwise distances between relative expression levels, we propose to estimate \( \xi_j \) via fitting a two-component Gaussian mixture model (GMM) with unequal variances,

\[
\tilde{y}_{ij} | \xi_{ij}, \mu_j, \sigma_j^2 \sim (1 - \xi_{ij})N(\mu_{j0}, \sigma_{j0}^2) + \xi_{ij}N(\mu_{j1}, \sigma_{j1}^2),
\]

subjecting to \( \mu_{j0} < \mu_{j1} \). Here \( \mu_j = (\mu_{j0}, \mu_{j1}) \) and \( \sigma_j^2 = (\sigma_{j0}^2, \sigma_{j1}^2) \) are the group means and variances that need to be estimated. To ensure the dichotomized expression levels are of the best quality to perform the subsequent modeling, we implement the above filtering steps as well.

To evaluate the performance between the two choices, we conducted a sensitivity analysis. We simulated ten replicated datasets following the data generating process described in Section 2.4. We include only the scenario with the medium zero-inflation setting (30%
false zeros) and the two real spatial patterns (MOB II and BC). We assessed the model performance in terms of AUC. The result is summarized in Figure A.2. BOOST-MI was robust to the two Dichotomization methods. We conducted the pairwise t-test. The resulting p-value = 0.585 and 0.850 for MOB II and BC pattern, respectively, confirming no significant difference among the two choices.

![Figure A.2: Simulation study: The boxplots of AUCs achieved by different dichotomization methods in BOOST-MI.](image)

**A.3. Full Details of MCMC Algorithms**

Our primary interest lies in the identification of SV gene via quantifying the posterior probability $\Pr(\theta_j < 0 \mid \cdot)$ while adjusting the abundance effect characterized by $\omega_j$. To serve this purpose, a double Metropolis-within-Gibbs sampler is designed. Our inferential strategy allows for simultaneously estimating the abundance parameters $\omega_j = (\omega_{j0}, \omega_{j1})$, which reveal the abundance of the two expression levels, and the interaction parameter $\theta_j$, which captures their spatial correlation. We give the full details of our MCMC algorithm as below.

**Update of the interaction parameter $\theta_j$:** To update $\theta_j$ within each iteration, we first simulate a new sample $\theta_j'$ from $\pi(\theta_j)$ using the MH algorithm starting with $\theta_j$. Then, we generate an auxiliary variable $\xi_j'$ through $m$ MH updates starting with the current state
ξ_j based on the new value θ'_j and accept it with probability min(1, R_{θ_j}), where

\[
R_{θ_j} = \frac{Pr(ξ'_j|ω_j, θ_j)Pr(ξ_j|ω_j, θ_j)}{Pr(ξ_j|ω_j, θ_j)Pr(ξ'_j|ω_j, θ'_j)} π(θ'_j)
\]

\[
= \frac{C(ω_j, θ_j) exp \left( -\sum_{i=1}^{n} ω_j ξ_{ij} - θ_j \sum_{i \sim j} I(ξ_{ij} \neq ξ'_{ij}) \right)}{C(ω_j, θ'_j) exp \left( -\sum_{i=1}^{n} ω_j ξ_{ij} - θ'_j \sum_{i \sim j} I(ξ_{ij} \neq ξ'_{ij}) \right)} \times \frac{exp \left( -\frac{(θ_j - μ)^2}{2τ_j^2} \right)}{exp \left( -\frac{(θ'_j - μ)^2}{2τ_j^2} \right)}
\]

As a result, the intractable normalizing constants C(ω_j, θ_j) and C(ω_j, θ'_j) can be canceled out. If the auxiliary variable ξ'_j is accepted, we update θ_j to θ'_j; otherwise, we keep the value of θ_j. No improvement in the performance was noticed beyond m = 5 in both simulation and application studies of this paper.

**Update of the abundance parameter ω_j:** Just as in updating θ_j, we use the DMH algorithm to update ω_{j0}. Specifically, we first simulate a new sample ω'_j = (ω'_{j0}, 1) from π(ω_{j0}) using the MH algorithm starting with ω_{j0}. Then, we generate an auxiliary variable ξ'_j through m MH updates starting with the current state ξ_j based on the new value ω'_j and
accept it with probability \( \min(1, R_{\omega_j}) \), where

\[
R_{\omega_j} = \frac{\Pr(\xi_j' | \omega_j, \theta_j) \Pr(\xi_j | \omega_j', \theta_j) \pi(\omega_{j0})}{\Pr(\xi_j | \omega_j, \theta_j) \Pr(\xi_j' | \omega_j', \theta_j) \pi(\omega_{j0})}
\]

\[
= \frac{\exp \left( -\sum_{i=1}^{n} \omega_{ij} \xi_{ij} - \theta_j \sum_{i \sim i'} I(\xi_{ij} \neq \xi_{ij}') \right)}{\exp \left( -\sum_{i=1}^{n} \omega_{ij}' \xi_{ij}' - \theta_j \sum_{i \sim i'} I(\xi_{ij}' \neq \xi_{ij}') \right)} \times \frac{\exp \left( -\frac{(\omega_{j0} - \mu_\omega)^2}{2\tau_\omega^2} \right)}{\exp \left( -\frac{(\omega_{j0}' - \mu_\omega)^2}{2\tau_\omega^2} \right)}
\]

As a result, the intractable normalizing constants \( C(\omega_j, \theta_j) \) and \( C(\omega_j', \theta_j) \) can be canceled out. If the auxiliary variable \( \xi_j' \) is accepted, we update \( \omega_{j0} \) to \( \omega_{j0}' \); otherwise, the value of \( \omega_{j0} \) remains the same.

### A.4. Sensitivity Analysis

Figure A.3: Simulation study: The heatmaps of averaged AUCs and MCCs achieved by different BOOST-MI hyperparameters \( \tau_\omega \) and \( \tau_\theta \).

We conducted a sensitivity analysis to investigate the sensitivity of BOOST-MI to the choice of \( \tau_\omega \) and \( \tau_\theta \). We applied BOOST-MI to each of the ten replicated datasets under the
scenario with the low zero-inflation setting (10% false zeros) and MOB II pattern. We varied values of $\tau_\theta$ from $1/4$ to 2 and $\tau_\omega$ from 1 to 100. We chose five values for each hyperparameter, resulting in 25 combinations. We assessed the model performance in terms of both AUC and MCC, where the latter was based on a Bayes factor (BF) threshold of 150. The result is summarized in Figure A.3, clearly indicating that BOOST-MI was not sensitive to the choices of hyperparameters.

A.5. Supporting Figures and Tables

Figure A.4: The scatter plot of the Moran’s I based on the dichotomized data against that based on the normalized data over all genes in the mouse olfactory bulb (MOB) and human breast cancer (BC) datasets.

Figure A.5: The histogram of low-expression spot proportions over all genes in the mouse olfactory bulb (MOB) and human breast cancer (BC) datasets.
Figure A.6: Real data analysis: The hierarchical clustering dendrograms of SV genes detected by BOOST-MI, SPARK, and BinSpect from the (a) mouse olfactory bulb and (b) human breast cancer datasets. Each color represents a unique spatial pattern defined in Figure 2.4 and 2.6.
Table A.3: The key notations of the proposed BOOST-MI model.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Support</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>$n \in \mathbb{N}$</td>
<td>The number of spots</td>
</tr>
<tr>
<td>$p$</td>
<td>$p \in \mathbb{N}$</td>
<td>The number of genes</td>
</tr>
<tr>
<td>$Y = [y_{ij}]_{n \times p}$</td>
<td>$y_{ij} \in \mathbb{N}$</td>
<td>The gene expression count table with $y_{ij}$ indicating the read count for gene $j$ observed at spot $i$</td>
</tr>
<tr>
<td>$T = [t_{id}]_{n \times 2}$</td>
<td>$t_{id} \in \mathbb{N}$</td>
<td>The $x$ and $y$ coordinates of spots on the 2D lattice.</td>
</tr>
<tr>
<td>$\tilde{Y} = [\tilde{y}<em>{ij}]</em>{n \times p}$</td>
<td>$\tilde{y}_{ij} \in \mathbb{R}$</td>
<td>The normalized gene expression level matrix.</td>
</tr>
<tr>
<td>$s = [s_i]_{n \times 1}$</td>
<td>$s_i \in \mathbb{R}^+$</td>
<td>The size factor.</td>
</tr>
<tr>
<td>$\Xi = [\xi_{ij}]_{n \times p}$</td>
<td>$\xi_{ij} \in {0, 1}$</td>
<td>The dichotomized gene expression level matrix with $\xi_{ij} = 0$ and 1 indicating the low and high expression level, respectively, for gene $j$ observed at spot $i$.</td>
</tr>
<tr>
<td><strong>GMM parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M = [\mu_{jk}]_{p \times 2}$</td>
<td>$\mu_{j0}, \mu_{j1} \in \mathbb{R}$</td>
<td>$\mu_{j0}$ and $\mu_{j1}$ are the means of the low and high expression level, respectively, for gene $j$.</td>
</tr>
<tr>
<td>$\Sigma = [\sigma^2_{jk}]_{p \times 2}$</td>
<td>$\sigma^2_{j0}, \sigma^2_{j1} \in \mathbb{R}^+$</td>
<td>$\sigma^2_{j0}$ and $\sigma^2_{j1}$ are the variance of the low and high expression level, respectively, for gene $j$.</td>
</tr>
<tr>
<td><strong>MI parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta = [\theta_j]_{p \times 1}$</td>
<td>$\theta_j \in \mathbb{R}$</td>
<td>$\theta_j$ characterizes the spatial correlation between the low and high expression levels for gene $j$.</td>
</tr>
<tr>
<td>$\Omega = [\omega_{jk}]_{p \times 2}$</td>
<td>$\omega_{j0} \in \mathbb{R}$ and $\omega_{j1} = 1$</td>
<td>$\omega_{j0}$ and $\omega_{j1}$ characterize the abundances of the low and high expression levels, respectively, for gene $j$.</td>
</tr>
<tr>
<td>$\tau^2_\theta$</td>
<td>$\tau^2_\theta \in \mathbb{R}^+$</td>
<td>The variance of the normal prior for all $\theta_j$’s.</td>
</tr>
<tr>
<td>$\tau^2_\omega$</td>
<td>$\tau^2_\omega \in \mathbb{R}^+$</td>
<td>The variance of the normal priors for all $\omega_{j0}$’s.</td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\phi = [\phi_j]_{p \times 1}$</td>
<td>$\phi_j \in \mathbb{R}^+$</td>
<td>$\phi_j$ is the estimated dispersion for gene $j$ used in the VST-based normalization method.</td>
</tr>
<tr>
<td>$W = [w_{ii'}]_{n \times n}$</td>
<td>$w_{ii'} \in {0, 1}$</td>
<td>The spatial weight matrix used in the Moran’s I calculation.</td>
</tr>
<tr>
<td>$I(\cdot)$</td>
<td>$I(\cdot) \in {0, 1}$</td>
<td>The indicator function</td>
</tr>
</tbody>
</table>

Note: GMM is Gaussian mixture model; MI is modified Ising model; VST is variance-stabilizing transformation.
Table A.4: Simulation study: Implementation details of BOOST-MI and the competing methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Normalization</th>
<th>Log-transformation</th>
<th>Dichotomization</th>
<th>Kernel</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOOST-MI</td>
<td>TSS</td>
<td>No</td>
<td>GMM</td>
<td></td>
<td>$\tau_\theta = \frac{1}{3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\tau_\omega = 2.5$</td>
</tr>
<tr>
<td>BOOST-GP</td>
<td>TSS</td>
<td>Yes</td>
<td></td>
<td>SE</td>
<td>$a_\pi = b_\pi = 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$a_\phi = b_\phi = a_l = b_l = 0.001$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$a_\omega = 0.1$ and $b_\omega = 1.9$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$a_\sigma = 3$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No need to specify the length-scale parameter $l$</td>
</tr>
<tr>
<td>BinSpect-k-means</td>
<td>TSS</td>
<td>Yes</td>
<td>$k$-means</td>
<td></td>
<td>Build the spatial network via $k$-nearest neighbors ($k = 4$)</td>
</tr>
<tr>
<td>BinSpect-rank</td>
<td>TSS</td>
<td>Yes</td>
<td>70% quantile</td>
<td></td>
<td>Build the spatial network via $k$-nearest neighbors ($k = 4$)</td>
</tr>
<tr>
<td>SPARK</td>
<td>TSS</td>
<td>Yes</td>
<td></td>
<td>SE</td>
<td>Pre-determine five length-scale parameters $l$'s based on the data</td>
</tr>
<tr>
<td>SpatialDE</td>
<td>log-VST</td>
<td>Yes</td>
<td></td>
<td>SE</td>
<td>No need to specify the length-scale parameter $l$</td>
</tr>
</tbody>
</table>

Note: TSS is total sum scaling; log-VST is variance-stabilizing transformation based on logarithm; GMM is Gaussian mixture model; SE is squared exponential.
Table A.5: Simulation study: The averaged MCCs (standard deviations) achieved by BOOST-MI, Ising, BOOST-GP, BinSpect, SPARK, and SpatialDE under different scenarios in terms of spatial pattern and zero-inflation setting.

<table>
<thead>
<tr>
<th></th>
<th>Spot</th>
<th>Linear</th>
<th>GP</th>
<th>MOB I</th>
<th>MOB II</th>
<th>BC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low zero-inflation (10% false zeros)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOOST-MI</td>
<td>0.519(0.071)</td>
<td><strong>0.853</strong>(0.054)</td>
<td><strong>0.343</strong>(0.102)</td>
<td><strong>0.642</strong>(0.103)</td>
<td>0.626(0.086)</td>
<td>0.531(0.108)</td>
</tr>
<tr>
<td>Ising</td>
<td>0.168(0.063)</td>
<td>0.185(0.071)</td>
<td>0.088(0.067)</td>
<td>0.381(0.132)</td>
<td>0.162(0.045)</td>
<td>0.123(0.101)</td>
</tr>
<tr>
<td>BinSpect-rank</td>
<td>0.341(0.176)</td>
<td>0.418(0.167)</td>
<td>0.318(0.135)</td>
<td>0.333(0.172)</td>
<td><strong>0.632</strong>(0.080)</td>
<td>0.639(0.111)</td>
</tr>
<tr>
<td>BinSpect-km</td>
<td>0.061(0.117)</td>
<td>0.036(0.102)</td>
<td>0.051(0.133)</td>
<td>0.041(0.158)</td>
<td>-0.019(0.035)</td>
<td>-0.021(0.032)</td>
</tr>
<tr>
<td>BOOST-GP</td>
<td>0.551(0.118)</td>
<td>0.726(0.124)</td>
<td>0.327(0.092)</td>
<td>0.015(0.078)</td>
<td>0.571(0.102)</td>
<td>0.586(0.104)</td>
</tr>
<tr>
<td>SPARK</td>
<td><strong>0.624</strong>(0.091)</td>
<td>0.768(0.053)</td>
<td>0.298(0.080)</td>
<td>0.000(0.000)</td>
<td>0.488(0.116)</td>
<td><strong>0.652</strong>(0.083)</td>
</tr>
<tr>
<td>SpatialDE</td>
<td>0.128(0.136)</td>
<td>0.497(0.102)</td>
<td>0.050(0.127)</td>
<td>0.000(0.000)</td>
<td>0.184(0.203)</td>
<td>0.276(0.197)</td>
</tr>
<tr>
<td><strong>Medium zero-inflation (30% false zeros)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOOST-MI</td>
<td>0.237(0.155)</td>
<td><strong>0.604</strong>(0.077)</td>
<td>0.175(0.113)</td>
<td><strong>0.568</strong>(0.095)</td>
<td><strong>0.501</strong>(0.157)</td>
<td>0.398(0.071)</td>
</tr>
<tr>
<td>Ising</td>
<td>0.108(0.038)</td>
<td>0.091(0.060)</td>
<td>0.070(0.051)</td>
<td>0.106(0.140)</td>
<td>0.065(0.054)</td>
<td>0.068(0.091)</td>
</tr>
<tr>
<td>BinSpect-rank</td>
<td>0.067(0.111)</td>
<td>0.242(0.055)</td>
<td><strong>0.248</strong>(0.114)</td>
<td>0.306(0.096)</td>
<td>0.478(0.139)</td>
<td><strong>0.567</strong>(0.112)</td>
</tr>
<tr>
<td>BinSpect-km</td>
<td>0.049(0.126)</td>
<td>-0.004(0.076)</td>
<td>-0.017(0.064)</td>
<td>0.010(0.065)</td>
<td>-0.027(0.098)</td>
<td>-0.031(0.030)</td>
</tr>
<tr>
<td>BOOST-GP</td>
<td><strong>0.254</strong>(0.133)</td>
<td>0.487(0.115)</td>
<td>0.113(0.020)</td>
<td>0.000(0.000)</td>
<td>0.239(0.117)</td>
<td>0.387(0.136)</td>
</tr>
<tr>
<td>SPARK</td>
<td>0.055(0.048)</td>
<td>0.342(0.164)</td>
<td>0.074(0.076)</td>
<td>0.000(0.000)</td>
<td>0.065(0.113)</td>
<td>0.058(0.124)</td>
</tr>
<tr>
<td>SpatialDE</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
</tr>
<tr>
<td><strong>High zero-inflation (50% false zeros)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOOST-MI</td>
<td>0.046(0.108)</td>
<td>0.175(0.158)</td>
<td>0.001(0.088)</td>
<td><strong>0.249</strong>(0.166)</td>
<td><strong>0.368</strong>(0.107)</td>
<td><strong>0.288</strong>(0.139)</td>
</tr>
<tr>
<td>Ising</td>
<td>-0.009(0.091)</td>
<td>0.004(0.084)</td>
<td>-0.022(0.087)</td>
<td>0.000(0.000)</td>
<td>0.010(0.093)</td>
<td>0.020(0.063)</td>
</tr>
<tr>
<td>BinSpect-rank</td>
<td>0.046(0.122)</td>
<td>0.075(0.141)</td>
<td><strong>0.044</strong>(0.126)</td>
<td>0.021(0.100)</td>
<td>0.095(0.097)</td>
<td>0.197(0.102)</td>
</tr>
<tr>
<td>BinSpect-km</td>
<td>0.046(0.089)</td>
<td>0.008(0.108)</td>
<td>-0.006(0.041)</td>
<td>0.024(0.096)</td>
<td>0.018(0.078)</td>
<td>-0.044(0.063)</td>
</tr>
<tr>
<td>BOOST-GP</td>
<td><strong>0.085</strong>(0.152)</td>
<td><strong>0.198</strong>(0.146)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.070(0.145)</td>
<td>0.183(0.119)</td>
</tr>
<tr>
<td>SPARK</td>
<td>0.000(0.000)</td>
<td>0.024(0.076)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.048(0.101)</td>
</tr>
<tr>
<td>SpatialDE</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
<td>0.000(0.000)</td>
</tr>
</tbody>
</table>
Table A.6: Real data analysis on the mouse olfactory bulb (MOB) dataset: List of 60 SV genes that had an attraction pattern with a positive interaction parameter in the Ising model.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Bayes Factor BF</th>
<th>$2 \times \ln(BF)$</th>
<th>Total Raw Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rc3h2</td>
<td>Inf</td>
<td>Inf</td>
<td>1174</td>
</tr>
<tr>
<td>Trib2</td>
<td>Inf</td>
<td>Inf</td>
<td>503</td>
</tr>
<tr>
<td>Med21</td>
<td>Inf</td>
<td>Inf</td>
<td>290</td>
</tr>
<tr>
<td>Nup210</td>
<td>Inf</td>
<td>Inf</td>
<td>163</td>
</tr>
<tr>
<td>Rsad1</td>
<td>Inf</td>
<td>Inf</td>
<td>94</td>
</tr>
<tr>
<td>Arap1</td>
<td>Inf</td>
<td>Inf</td>
<td>86</td>
</tr>
<tr>
<td>Zfp938</td>
<td>9999.000</td>
<td>18.420</td>
<td>190</td>
</tr>
<tr>
<td>Trim8</td>
<td>6665.667</td>
<td>17.609</td>
<td>809</td>
</tr>
<tr>
<td>Mrps18b</td>
<td>2856.143</td>
<td>15.914</td>
<td>192</td>
</tr>
<tr>
<td>Rbm15b</td>
<td>2499.000</td>
<td>15.647</td>
<td>151</td>
</tr>
<tr>
<td>Fam20c</td>
<td>2221.222</td>
<td>15.412</td>
<td>1603</td>
</tr>
<tr>
<td>Pdik1l</td>
<td>2221.222</td>
<td>15.412</td>
<td>143</td>
</tr>
<tr>
<td>Ccnh</td>
<td>951.381</td>
<td>13.716</td>
<td>1261</td>
</tr>
<tr>
<td>Ercc3</td>
<td>951.381</td>
<td>13.716</td>
<td>239</td>
</tr>
<tr>
<td>Vapa</td>
<td>868.565</td>
<td>13.534</td>
<td>4017</td>
</tr>
<tr>
<td>Dgke</td>
<td>713.286</td>
<td>13.140</td>
<td>393</td>
</tr>
<tr>
<td>Zfp248</td>
<td>713.286</td>
<td>13.140</td>
<td>109</td>
</tr>
<tr>
<td>Fen1</td>
<td>644.161</td>
<td>12.936</td>
<td>88</td>
</tr>
<tr>
<td>Amigo2</td>
<td>624.000</td>
<td>12.872</td>
<td>445</td>
</tr>
<tr>
<td>Mfap3</td>
<td>605.061</td>
<td>12.811</td>
<td>208</td>
</tr>
<tr>
<td>Katnal1</td>
<td>539.541</td>
<td>12.581</td>
<td>204</td>
</tr>
<tr>
<td>Tkt</td>
<td>525.316</td>
<td>12.528</td>
<td>518</td>
</tr>
<tr>
<td>Plekhn3</td>
<td>525.316</td>
<td>12.528</td>
<td>317</td>
</tr>
<tr>
<td>Ofd1</td>
<td>525.316</td>
<td>12.528</td>
<td>97</td>
</tr>
<tr>
<td>Mrps24</td>
<td>511.821</td>
<td>12.476</td>
<td>813</td>
</tr>
<tr>
<td>Wdr3</td>
<td>511.821</td>
<td>12.476</td>
<td>218</td>
</tr>
<tr>
<td>Pold3</td>
<td>499.000</td>
<td>12.425</td>
<td>227</td>
</tr>
<tr>
<td>Rcbtb1</td>
<td>486.805</td>
<td>12.376</td>
<td>552</td>
</tr>
<tr>
<td>Snx21</td>
<td>433.783</td>
<td>12.145</td>
<td>199</td>
</tr>
<tr>
<td>Fgf14</td>
<td>407.163</td>
<td>12.018</td>
<td>158</td>
</tr>
</tbody>
</table>

114
<table>
<thead>
<tr>
<th>Gene</th>
<th>Value1</th>
<th>Value2</th>
<th>Value3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nop58</td>
<td>376.358</td>
<td>11.861</td>
<td>248</td>
</tr>
<tr>
<td>Zfp9</td>
<td>337.983</td>
<td>11.646</td>
<td>144</td>
</tr>
<tr>
<td>Bbs5</td>
<td>326.869</td>
<td>11.579</td>
<td>249</td>
</tr>
<tr>
<td>Eif3c</td>
<td>284.714</td>
<td>11.303</td>
<td>1123</td>
</tr>
<tr>
<td>Chpf2</td>
<td>265.667</td>
<td>11.164</td>
<td>126</td>
</tr>
<tr>
<td>Ruvbl1</td>
<td>262.158</td>
<td>11.138</td>
<td>470</td>
</tr>
<tr>
<td>Zfp846</td>
<td>262.158</td>
<td>11.138</td>
<td>103</td>
</tr>
<tr>
<td>Umps</td>
<td>231.558</td>
<td>10.890</td>
<td>128</td>
</tr>
<tr>
<td>Pcnxl4</td>
<td>226.273</td>
<td>10.843</td>
<td>286</td>
</tr>
<tr>
<td>Cstf2</td>
<td>214.054</td>
<td>10.732</td>
<td>780</td>
</tr>
<tr>
<td>Pias2</td>
<td>209.526</td>
<td>10.690</td>
<td>920</td>
</tr>
<tr>
<td>Cog4</td>
<td>209.526</td>
<td>10.690</td>
<td>308</td>
</tr>
<tr>
<td>Dnalc1</td>
<td>205.186</td>
<td>10.648</td>
<td>616</td>
</tr>
<tr>
<td>Cnih3</td>
<td>205.186</td>
<td>10.648</td>
<td>287</td>
</tr>
<tr>
<td>Phyhipl</td>
<td>203.082</td>
<td>10.627</td>
<td>1488</td>
</tr>
<tr>
<td>Hnrnpf</td>
<td>203.082</td>
<td>10.627</td>
<td>819</td>
</tr>
<tr>
<td>Ncoa2</td>
<td>195.078</td>
<td>10.547</td>
<td>953</td>
</tr>
<tr>
<td>Atg13</td>
<td>195.078</td>
<td>10.547</td>
<td>728</td>
</tr>
<tr>
<td>1700025G04Rik</td>
<td>187.679</td>
<td>10.469</td>
<td>794</td>
</tr>
<tr>
<td>2310003H01Rik</td>
<td>184.185</td>
<td>10.432</td>
<td>94</td>
</tr>
<tr>
<td>Pin1</td>
<td>175.991</td>
<td>10.341</td>
<td>913</td>
</tr>
<tr>
<td>Nceh1</td>
<td>174.439</td>
<td>10.323</td>
<td>871</td>
</tr>
<tr>
<td>Lamtor5</td>
<td>172.913</td>
<td>10.306</td>
<td>356</td>
</tr>
<tr>
<td>Tmem42</td>
<td>171.414</td>
<td>10.288</td>
<td>157</td>
</tr>
<tr>
<td>Anks3</td>
<td>169.940</td>
<td>10.271</td>
<td>384</td>
</tr>
<tr>
<td>Crot</td>
<td>167.067</td>
<td>10.237</td>
<td>481</td>
</tr>
<tr>
<td>Tbc1d23</td>
<td>161.602</td>
<td>10.170</td>
<td>399</td>
</tr>
<tr>
<td>Armcx2</td>
<td>157.730</td>
<td>10.122</td>
<td>594</td>
</tr>
<tr>
<td>Nabp1</td>
<td>156.480</td>
<td>10.106</td>
<td>111</td>
</tr>
<tr>
<td>Rasgrp2</td>
<td>154.039</td>
<td>10.074</td>
<td>170</td>
</tr>
</tbody>
</table>
Appendix B
Appendix of Chapter 3

B.1. Full Details of MCMC Algorithms

According to the model description in the Section 3.2, the full data likelihood of the proposed Bayesian finite normal-multinomial mixture model is given as follows.

\[
\begin{align*}
f(Y, V | z, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \omega_1, \ldots, \omega_K) &= K \prod_{k=1}^{K} \prod_{i=1}^{N} I(z_i = k) f(y_i, v_i | z_i = k; \mu_k, \Sigma_k, \omega_k),
\end{align*}
\]

where \( I(\cdot) \) denotes the indicator function and

\[
\begin{align*}
f(y_i, v_i | z_i = k; \mu_k, \Sigma_k, \omega_k) &= f(y_i | z_i = k; \mu_k, \Sigma_k) f(v_i | z_i = k; \omega_k)^w \\
&= \text{MN}(y_i; \mu_k, \Sigma_k) \text{Multi}(v_i; m_i, \omega_k)^w \\
&= (2\pi)^{-P'/2} |\Sigma_k|^{-1/2} \exp \left( -\frac{1}{2} (y_i - \mu_k) ^\top \Sigma_k^{-1} (y_i - \mu_k) \right) \left( \frac{m_i!}{\prod_{q=1}^{Q} v_{iq}! \prod_{q=1}^{Q} \omega_{kq}^{v_{iq}}} \right)^w \\
&\propto |\Sigma_k|^{-1/2} \exp \left( -\frac{1}{2} (y_i - \mu_k) ^\top \Sigma_k^{-1} (y_i - \mu_k) \right) \left( \prod_{q=1}^{Q} \omega_{kq}^{v_{iq}} \right)^w.
\end{align*}
\]

We assume an independent prior structure 1) between the normal and multinomial sub-components; and 2) among their parameters belonging to different spatial domains. Thus,
the joint distribution of priors for parameters is given as follows.

\[
\pi(z, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \omega_1, \ldots, \omega_K) \\
= \pi(z) \prod_{k=1}^{K} \pi(\mu_k, \Sigma_k) \prod_{k=1}^{K} \pi(\omega_k) \\
= \pi(z) \prod_{k=1}^{K} \pi(\mu_k | \Sigma_k) \prod_{k=1}^{K} \pi(\Sigma_k) \prod_{k=1}^{K} \pi(\omega_k) \\
= \pi(z) \prod_{k=1}^{K} \pi(\mu_k | \Sigma_k) \pi(\Sigma_k) \prod_{k=1}^{K} \pi(\omega_k).
\]

We assign the Markov random field (MRF) prior for histology-based spatial domain indicator \(z\) as

\[
\pi(z_i | z_{-i}) \propto \exp \left( d_k + f \sum_{i'=1, i' \neq i}^{N} g_{ii'} I(z_{i'} = k) \right),
\]

We assign the conjugate priors for other parameters, listed as follows, so that the Gibbs sampler can be applied for posterior sampling.

\[
\mu_k | \Sigma_k \sim \text{MN}(\nu_0, \Sigma / \tau_0) \\
\text{or equivalently, } \pi(\mu_k | \Sigma_k) = (2\pi / \tau_0)^{-P'/2} |\Phi_0|^{-1/2} \exp \left( -\frac{\tau_0}{2} (\mu_k - \nu_0)^\top \Phi_0^{-1} (\mu_k - \nu_0) \right),
\]

\[
\Sigma_k \sim \text{IW}(\eta_0, \Phi_0) \\
\text{or equivalently, } \pi(\Sigma_k) = \frac{|\Phi_0|^{|\eta_0|/2}}{2^{|\eta_0|P'/2} \Gamma_{P'}(|\eta_0|/2)} |\Sigma_k|^{-|\eta_0|+P'/2} \exp \left( -\frac{1}{2} \text{tr} (\Phi_0 \Sigma_k^{-1}) \right),
\]

and

\[
\omega_k \sim \text{Dir}(\alpha_0) \text{ or equivalently, } \pi(\omega_k) = \frac{\Gamma \left( \sum_{q=1}^{Q} \alpha_{0q} \right)}{\prod_{q=1}^{Q} \Gamma(\alpha_{0q})} \prod_{q=1}^{Q} \omega_k^{\alpha_{0q}-1},
\]

where \(\Gamma_{P'}(\cdot)\) and \(\Gamma(\cdot)\) denote the \(P'\)-dimensional and univariate gamma function.
We recommend a weakly informative prior setting by choosing the MRF hyperparameters \( d_1 = \ldots = d_K = 1 \) and \( f = 1 \), the multivariate normal hyperparameters \( \nu_0 = \frac{1}{N} \sum_{i=1}^{N} y_i \), \( \tau_0 = 0.01 \), \( \eta_0 = P' + 1 \), and \( \Phi_0 = I_{P' \times P'} \) (i.e., the \( P' \)-by-\( P' \) identity matrix), and the multinomial hyperparameters \( \alpha_{01} = \ldots = \alpha_{0Q} = 1 \).

The full posterior distribution of the proposed Bayesian normal-multinomial mixture model is given in the following formula.

\[
\pi(z, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \omega_1, \ldots, \omega_K | Y, V) \propto \\
f(Y, V | z, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \omega_1, \ldots, \omega_K) \pi(z, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \omega_1, \ldots, \omega_K)
\]

Posterior sampling is employed by the MCMC algorithm. Our primary interest lies in identifying histology-based spatial domains and the interactive zone via inferring the spatial domain indicator vector \( z \), and in characterizing domain-specific relative abundance of cell types via inferring \( \omega_1, \ldots, \omega_K \). Since we use conjugate priors on all model parameters \( \{z, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \omega_1, \ldots, \omega_K\} \), their conditional distributions are all in closed form and easy to sample from. Consequently, we can rely on the Gibbs sampler, an MCMC algorithm for obtaining a sequence of observations approximated from a multivariate probability distribution when direct sampling is difficult. To be specific, we perform the following steps sequentially at each MCMC iteration after a random initialization.

**Update the histology-based spatial domain indicator \( z \):** We update \( z_1, \ldots, z_N \) sequentially. To allocate spot \( i \) to one of the \( K \) histology-based spatial domains, we sample \( z_i \) from a single-drawing multinomial distribution,

\[
z_i | \sim \text{Multi}(1, (\pi(z_i = 1 | \cdot)/e, \ldots, \pi(z_i = K | \cdot)/e))
\]
where
\[
\pi(z_i = k|\cdot) \propto f(y_i, v_i | z_i = k, \mu_k, \Sigma_k, \omega_k) \pi(z_i = k | z_{-i})
\]
\[
\propto |\Sigma_k|^{-1/2} \exp \left( -\frac{1}{2} (y_i - \mu_k)^\top \Sigma_k^{-1} (y_i - \mu_k) \right) \left( \prod_{q=1}^Q \omega_{k,q} \right)^w 
\]
\[
\exp \left( d_k + f \sum_{i' = 1, i' \neq i}^N g_{i'i} I(z_{i'} = k) \right)
\]
and the normalization constant \( e = \sum_{k=1}^K \pi(z_i = k|\cdot) \).

**Update the domain-specific relative abundance of cell types \( \omega_k \)'s:** We update \( \omega_1, \ldots, \omega_K \) sequentially. For each histology-based spatial domain \( k \), we draw a sample of \( \omega_k \) from a Dirichlet distribution,
\[
\omega_k|\cdot \sim \text{Dir}(\alpha_k),
\]
where the concentration parameters \( \alpha_k = (\alpha_{k1}, \ldots, \alpha_{kQ}) \) with each entry \( \alpha_{kq} = \alpha_{0q} + \sum_{i=1}^N I(z_i = k)v_{iq} \). Note that the last term \( \sum_{i=1}^N I(z_i = k)v_{iq} \) denotes the total number of cells with type \( q \) observed in histology-based spatial domain \( k \).

**Update the domain-specific low-dimensional representation of gene expression mean \( \mu_k \)'s:** We update \( \mu_1, \ldots, \mu_K \) sequentially. For each histology-based spatial domain \( k \), we draw a sample of \( \mu_k \) from a multivariate normal distribution,
\[
\mu_k|\cdot \sim \text{MN}(\nu_k, \Sigma_k/\tau_k),
\]
where \( \nu_k = (\nu_0\tau_0 + n_k\bar{y}_k)/(\tau_0 + n_k) \) and \( \tau_k = \tau_0 + n_k \). Note that \( n_k = \sum_{i=1}^N I(z_i = k) \) is the number of spots allocated to histology-based spatial domain \( k \) and \( \bar{y}_k = \frac{1}{n_k} \sum_{i=1}^N I(z_i = k)y_i \) denotes the average low-dimensional gene expression value over all the spots allocated to histology-based spatial domain \( k \).
If PCA is chosen to reduce the dimension of the SRT molecular profile, then we can further set $\Sigma_k$ to an $P'\times P'$ diagonal matrix due to orthogonality among principal components. In this special case, we can draw each entry in $\mu_k$ independently,

$$
\mu_{kj} \mid \sim N(\nu_{kj}, \sigma_{kj}^2 / \tau_k),
$$

where $\nu_{kj} = (\nu_0 \tau_0 + n_k \bar{y}_{kj}) / (\tau_0 + n_k)$.

Update the domain-specific covariance matrix of the low-dimensional representation of gene expression $\Sigma_k$'s: We update $\Sigma_1, \ldots, \Sigma_K$ sequentially. For each histology-based spatial domain $k$, we draw a sample of $\Sigma_k$ from an inverse-Wishart distribution,

$$
\Sigma_k \mid \sim IW(\eta_k, \Phi_k),
$$

where $\eta_k = \eta_0 + n_k$ and $\Phi_k = \Phi_0 + \sum_{i=1}^{N} I(z_i = k)(y_{ij} - \bar{y}_k)(y_{ij} - \bar{y}_k)^\top + \frac{\tau_0 \eta_k}{\tau_0 + n_k} (\bar{y}_k - \nu_0)(\bar{y}_k - \nu_0)^\top$.

For the case when using PCA, the inverse-Wishart prior reduces to an inverse-gamma prior, $\sigma_{kj}^2 \sim IG(\eta_0/2, \phi_0/2)$. Thus, we only need to draw each diagonal entry in $\Sigma_k$ independently,

$$
\sigma_{kj}^2 \mid \sim IG(\eta_k/2, \phi_k/2),
$$

where $\phi_k = \phi_0 + \sum_{i=1}^{N} I(z_i = k)(y_{ij} - \bar{y}_{kj})^2 + \frac{\tau_0 \eta_k}{\tau_0 + n_k} (\bar{y}_{kj} - \nu_{0j})^2$.

### B.2. Sensitivity Analysis

We first conducted a sensitivity analysis to investigate how the number of dimensions reduced affects iIMPACT’s performance in histology-based spatial domain identification. In particular, we varied the number of top principal components (PC) $P'$ kept in the molecular profile from 2 to 15 and computed the ARI between the ground truth and the MAP estimate of the histology-based spatial domain indicator $\hat{z}$. Take human breast cancer data as example, Figure B.1 summarizes the achieved ARI against different choices of $P'$, along
with the cumulative variance explained by the PCs, indicating that iIMPACT was relatively robust to the choice of number of PCs and the best performance occurred when $P' = 3$.

Figure B.1: Sensitivity analysis: ARIs achieved by iIMPACT when setting the image profile weight $w$ to be zero, and proportion of variance explained under different number of leading principal components in PCA.

Then, another sensitivity analysis was performed to investigate the sensitivity of iIMPACT to the choice of the image profile weight $w$, which controls the image profile’s contribution to the spatial domain identification result. In particular, we varies $w$ in

$$\{0.01, 0.02, 0.05, 0.10, 0.25, 0.50, 0.75, 1.00\}$$

and computed the ARI between the ground truth and the MAP estimate of the histology-based spatial domain indicator $\hat{z}$. Figure B.2 summarizes the achieved ARI against different values of $w$ for both the human breast cancer and mouse visual cortex STARmap data. It is interesting to know that iIMPACT was very sensitive to the choice of $w$. The best performance occurred when setting $w = 0.05$ for the human breast cancer data from NGS-
based SRT platforms (e.g., 10x Visium) and $w = 0.5$ for the mouse visual cortex data from imaging-based SRT techniques (e.g., STARmap). Therefore, to guard against under or over-fitting, the value of $w$ should be chosen with some degree of caution.

Figure B.2: Sensitivity analysis: ARIs achieved by iIMPACT clustering under different choices of the image profile weight $w$ on A. human breast cancer data; and B. mouse visual cortex STARmap data.
B.3. Supporting Figures and Tables

Figure B.3: Demonstration of geometric representations of spatial distribution of spots, and definition of spot expanded area in the 10x Visium and classical ST technologies.
Figure B.4: Nuclei identification results from the HD-Staining model for human breast cancer data: A. The number of nuclei identified for seven different nuclei classes; B. Histogram of number of cells in each spot expanded area; C. Spatial distribution of spot-level cell type abundance for seven nuclei classes.
Figure B.5: Nuclei identification results from the HD-Staining model for human prostate cancer data: A. The number of nuclei identified for six different nuclei classes; B. Histogram of number of cells in each spot expanded area; C. Spatial distribution of spot-level cell type abundance for six nuclei classes.
Figure B.6: Histology image of the tissue section with manually annotated labels from pathologists, and histology-based spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to be 3, for human prostate cancer data.
Figure B.7: Nuclei identification results from the HD-Staining model for human ovarian cancer data: A. The number of nuclei identified for six different nuclei classes; B. Histogram of number of cells in each spot expanded area; C. Spatial distribution of spot-level cell type abundance for six nuclei classes.
Figure B.8: Histology image of the tissue section with manually annotated labels from pathologists, and histology-based spatial domains detected by iIMPACT, SpaGCN, and BayesSpace, setting the number of clusters to be 2, for human ovarian cancer data.
Table B.1: Summary of the four real datasets analyzed in the paper.

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Technology</th>
<th>Organism</th>
<th>Tissue</th>
<th>Disease</th>
<th>Number of genes</th>
<th>Sample size</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human breast cancer data</td>
<td>ST (10x Visium)</td>
<td>Human</td>
<td>Breast</td>
<td>Ductal carcinoma in situ, invasive carcinoma</td>
<td>17943</td>
<td>2518</td>
<td>10x Genomics</td>
</tr>
<tr>
<td>Human prostate cancer data</td>
<td>ST (10x Visium)</td>
<td>Human</td>
<td>Prostate</td>
<td>Adenocarcinoma, invasive carcinoma</td>
<td>17943</td>
<td>4371</td>
<td>10x Genomics</td>
</tr>
<tr>
<td>Human ovarian cancer data</td>
<td>ST (10x Visium)</td>
<td>Human</td>
<td>Ovarian</td>
<td>Serous papillary carcinoma</td>
<td>17943</td>
<td>3455</td>
<td>10x Genomics</td>
</tr>
<tr>
<td>Mouse visual cortex STARmap data</td>
<td>STARmap</td>
<td>Mouse</td>
<td>Brain-visual cortex</td>
<td>-</td>
<td>1207</td>
<td>1020</td>
<td>[14]</td>
</tr>
</tbody>
</table>

Table B.2: Running time (in minutes) of iIMPACT, BayesSpace and SpaGCN on the four real datasets analyzed in the paper.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>iIMPACT</th>
<th>BayesSpace</th>
<th>SpaGCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human breast cancer data</td>
<td>1.78</td>
<td>18.30</td>
<td>1.83</td>
</tr>
<tr>
<td>Human prostate cancer data</td>
<td>2.67</td>
<td>27.99</td>
<td>2.27</td>
</tr>
<tr>
<td>Human ovarian cancer data</td>
<td>2.24</td>
<td>23.49</td>
<td>1.90</td>
</tr>
<tr>
<td>Mouse visual cortex STARmap data</td>
<td>1.09</td>
<td>13.86</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Table B.3: The key notations of the proposed iIMPACT.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Support</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>$N \in \mathbb{N}$</td>
<td>Number of spots</td>
</tr>
<tr>
<td>$P$</td>
<td>$P \in \mathbb{N}$</td>
<td>Number of genes</td>
</tr>
<tr>
<td>$P'$</td>
<td>$P' \in \mathbb{N}, P' \ll P$</td>
<td>Number of reduced dimensions</td>
</tr>
<tr>
<td>$Q$</td>
<td>$Q \in \mathbb{N}$</td>
<td>Number of cell types</td>
</tr>
<tr>
<td>$K$</td>
<td>$K \in \mathbb{N}$</td>
<td>Number of histology-based spatial domains</td>
</tr>
<tr>
<td>$C = [c_{ij}]_{N \times P}$</td>
<td>$c_{ij} \in \mathbb{N}$</td>
<td>The gene expression count table with $c_{ij}$ being the read count for gene $j$ observed at spot $i$</td>
</tr>
<tr>
<td>$Y = [y_{ij}]_{N \times P'}$</td>
<td>$y_{ij} \in \mathbb{R}$</td>
<td>The low-dimensional gene expression table with $y_{ij}$ being the relative expression on dimension $j$ at spot $i$</td>
</tr>
<tr>
<td>$V = [v_{iq}]_{N \times Q}$</td>
<td>$v_{iq} \in \mathbb{N}$</td>
<td>The cell type abundance table with $v_{iq}$ being the number of cells with cell type $q$ at spot $i$ and its expanded area</td>
</tr>
<tr>
<td>$m = [m_i]_{N \times 1}$</td>
<td>$m_i \in \mathbb{N}$</td>
<td>The total number of cells at spot $i$ and its expanded area</td>
</tr>
<tr>
<td>$T = [t_{ir}]_{N \times 2}$</td>
<td>$t_{ir} \in \mathbb{N}$</td>
<td>The $x$ and $y$ coordinates of spots</td>
</tr>
<tr>
<td>$G = [g_{ir}]_{N \times N}$</td>
<td>$g_{ir} \in {0,1}$</td>
<td>The adjacent matrix with $g_{ir} = 1$ indicating spot $i$ and spot $i'$ are neighbors</td>
</tr>
<tr>
<td>$s = [s_i]_{N \times 1}$</td>
<td>$s_i \in \mathbb{R}^+$</td>
<td>Size factors</td>
</tr>
<tr>
<td><strong>Model parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Omega = [\omega_{kq}]_{K \times Q}$</td>
<td>$\omega_{kq} \in [0,1]$</td>
<td>With $\omega_{kq}$ being the relative abundance of cell type $q$ in histology-based spatial domain $k$</td>
</tr>
<tr>
<td>$M = [\mu_{kj}]_{K \times P'}$</td>
<td>$\mu_{kj} \in \mathbb{R}$</td>
<td>With $\mu_k$ being the mean vector of the multivariate normal subcomponent for histology-based spatial domain $k$</td>
</tr>
<tr>
<td>$\Sigma = [\Sigma_k]_{K \times P' \times P'}$</td>
<td>$\sigma_{kjj'} \in \mathbb{R}^+$</td>
<td>With $\Sigma_k$ being the covariance matrix of the multivariate normal subcomponent for histology-based spatial domain $k$</td>
</tr>
<tr>
<td>$z = [z_i]_{N \times 1}$</td>
<td>$z_i \in {1, \ldots, K}$</td>
<td>Histology-based spatial domain indicator</td>
</tr>
<tr>
<td>$\psi = [\psi_{ij}]_{P' \times 1}$</td>
<td>$\psi_{ij} \in \mathbb{R}^+$</td>
<td>The gene-specific dispersion parameter</td>
</tr>
<tr>
<td><strong>Hyperparameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha = [\alpha_q]_{Q \times 1}$</td>
<td>$\alpha_q \in \mathbb{R}^+$</td>
<td>Parameters in the Dirichlet priors for $\Omega$.</td>
</tr>
<tr>
<td>$\eta_0$</td>
<td>$\eta_0 \in (P' - 1, +\infty)$</td>
<td>Parameters in the inverse Wishart priors for $\Sigma$.</td>
</tr>
<tr>
<td>$\Phi_0$</td>
<td>$P' \times P'$</td>
<td>Positive definite matrix</td>
</tr>
<tr>
<td>$\nu_0 = [\nu_j]_{P' \times 1}$</td>
<td>$\nu_j \in \mathbb{R}$</td>
<td>Parameters in the inverse Wishart priors for $\Sigma$.</td>
</tr>
<tr>
<td>$\tau$</td>
<td>$\tau \in \mathbb{R}^+$</td>
<td>The mean parameters of the normal priors for $M$.</td>
</tr>
<tr>
<td>$d = [d_k]_{K \times 1}$</td>
<td>$d_k \in \mathbb{R}^+$</td>
<td>The MRF first-order hyperparameter.</td>
</tr>
<tr>
<td>$f$</td>
<td>$f \in \mathbb{R}^+$</td>
<td>The MRF second-order hyperparameter</td>
</tr>
<tr>
<td><strong>Tuning parameters</strong></td>
<td>$w$</td>
<td>$w \in [0,1]$</td>
</tr>
</tbody>
</table>
Table B.4: Posterior means of parameters (domain-specific relative abundances of cell types $\hat{\Omega}$ and means of low-dimensional representation of gene expression $\hat{M}$) for human breast cancer data, which are applied in generating the simulated data.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Stromal cell</th>
<th>Red blood cell</th>
<th>Lymphocyte cell</th>
<th>Tumor cell</th>
<th>Karyorrhexis Ductal epithelium</th>
<th>Macrophage</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>0.4209</td>
<td>0.0905</td>
<td>0.0492</td>
<td>0.3407</td>
<td>0.0415</td>
<td>0.0562</td>
</tr>
<tr>
<td>D2</td>
<td>0.5221</td>
<td>0.1045</td>
<td>0.3346</td>
<td>0.0174</td>
<td>0.0161</td>
<td>0.0045</td>
</tr>
<tr>
<td>D3</td>
<td>0.2551</td>
<td>0.5202</td>
<td>0.1855</td>
<td>0.0166</td>
<td>0.0141</td>
<td>0.0037</td>
</tr>
<tr>
<td>D4</td>
<td>0.3875</td>
<td>0.2802</td>
<td>0.0498</td>
<td>0.2207</td>
<td>0.0384</td>
<td>0.0168</td>
</tr>
<tr>
<td>D5</td>
<td>0.5327</td>
<td>0.1888</td>
<td>0.2502</td>
<td>0.0066</td>
<td>0.0203</td>
<td>0.0011</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain</th>
<th>PC 1</th>
<th>PC 2</th>
<th>PC 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>-26.2131</td>
<td>-2.3886</td>
<td>-0.0389</td>
</tr>
<tr>
<td>D2</td>
<td>-2.6583</td>
<td>-5.3326</td>
<td>1.2944</td>
</tr>
<tr>
<td>D3</td>
<td>-6.6993</td>
<td>-2.6915</td>
<td>4.0473</td>
</tr>
<tr>
<td>D4</td>
<td>-19.0341</td>
<td>10.3570</td>
<td>1.1856</td>
</tr>
<tr>
<td>D5</td>
<td>-18.0996</td>
<td>1.3260</td>
<td>-1.8454</td>
</tr>
</tbody>
</table>
C.1. Full Details of MCMC Algorithms

The model parameter space \((\phi, B, \Gamma)\) consists of 1) the NB dispersion parameters \(\phi\) that accounts for the over-dispersion commonly observed in gene expression count data, 2) the coefficient matrix \(B\) that quantifies the relationship between gene expression as measured in SRT data and the morphological features extracted from the paired histology image, and 3) the selection matrix \(\Gamma\) that indicates the significant association in the coefficient matrix \(B\). The full likelihood of the model is as follows:

\[
 f(Y|\phi, B, \Gamma) = \prod_{j=1}^{P} f(y_j|\phi_j, \beta_j, \gamma_j) \\
 = \prod_{j=1}^{P} \prod_{i=1}^{N} f(y_{ij}|\phi_j, \beta_j) \\
 = \prod_{j=1}^{P} \prod_{i=1}^{N} \text{NB}\left(y_{ij}; s_i, \frac{\sum_{m=1}^{M} g_{im} \exp\left(\sum_{l=1}^{L} \beta_{lj} x_{ml}\right)}{\sum_{m=1}^{M} g_{im}}, \phi_j\right). 
\]
With the prior specifications detailed in Methods Section of the manuscript, we give the full posterior distribution as follows:

$$
\pi(\phi, B, \Gamma | Y) = f(Y | \phi, B, \Gamma) \times \pi(B | \Gamma) \pi(\Gamma) \pi(\phi)
$$

$$
= f(Y | \phi, B, \Gamma) \times \left( \prod_{j=1}^{P} \prod_{l=1}^{L} \pi(\beta_{lj} | \gamma_{lj}) \right) \left( \prod_{j=1}^{P} \prod_{l=1}^{L} \pi(\gamma_{lj}) \right) \left( \prod_{j=1}^{P} \pi(\phi_{j}) \right)
$$

$$
= \prod_{j=1}^{P} \prod_{i=1}^{N} \text{NB} \left( y_{ij}; s_{i} \sum_{m=1}^{M} g_{im} \exp \left( \sum_{l=1}^{L} \beta_{lj} x_{ml} \right), \phi_{j} \right)
$$

$$
\times \prod_{j=1}^{P} \prod_{l=1}^{L} \left( 1 - \gamma_{lj} \right) \delta_{0}(\beta_{lj}) + \gamma_{lj} \mathcal{N}(\beta_{lj}; 0, \sigma_{\beta}^{2})
$$

$$
\times \prod_{j=1}^{P} \prod_{l=1}^{L} \text{Bern}(\gamma_{lj}; \pi_{\gamma})
$$

$$
\times \prod_{j=1}^{P} \text{Ga}(\phi_{j}; a_{\phi}, b_{\phi})
$$

(3.2)

MCMC algorithm was conducted for the posterior sampling. Specifically, we performed the following steps sequentially at each MCMC iteration after a random initialization.

**Update of over-dispersion parameter $\Phi$:** We update each $\phi_{j}$ ($j = 1, \ldots, P$) independently by using the Metropolis-Hasting algorithm. We first propose a new $\phi_{j}^{*}$ from the log-normal distribution, i.e., $\ln(\phi_{j}^{*}) \sim \mathcal{N}(\ln(\phi_{j}), \tau_{\phi}^{2})$, and accept the proposed value with probability $\min(1, r_{\phi})$, where

$$
\tau_{\phi}^{2} = \frac{\prod_{i=1}^{N} f(y_{ij} | \phi_{j}^{*}, \beta_{j}, \gamma_{j}; s_{i}, X, G) \times \pi(\phi_{j}^{*}) \times J(\phi_{j}^{*}; \phi_{j})}{\prod_{i=1}^{N} f(y_{ij} | \phi_{j}, \beta_{j}, \gamma_{j}; s_{i}, X, G) \times \pi(\phi_{j}) \times J(\phi_{j}^{*}; \phi_{j})}
$$

**Joint update of covariate coefficient $B$ and its underlying indicator $\Gamma$:** We update each $\beta_{lj}$ and $\gamma_{lj}$ ($l = 1, \ldots, L$ and $j = 1, \ldots, P$) sequentially by using an independent Metropolis-Hasting algorithm. A between-model step is implemented first to jointly update $\beta_{lj}$ and $\gamma_{lj}$. We use an add-delete algorithm, where we change the value of $\gamma_{lj}$. For the add
case, i.e., changing from \( \gamma_{lj} = 0 \) to \( \gamma_{lj}^* = 1 \), we propose \( \beta_{lj}^* \) from the normal distribution \( N(0, \tau_\beta^2) \). For the delete case, i.e., \( \gamma_{lj} = 1 \rightarrow \gamma_{lj}^* = 0 \), we set \( \beta_{lj}^* = 0 \). We finally accept the proposed \( \gamma_{lj}^* \) with probability \( \min(1, r_\gamma) \), where

\[
\begin{align*}
    r_\gamma &= \frac{\prod_{i=1}^N f(y_{ij}|\beta_{ij}, \gamma_{lj}, \phi_j; s, X, G) \times \pi(\beta_{ij}|\gamma_{lj}) \times \pi(\gamma_{lj}) \times J(\beta_{ij}; \beta_{ij}^*) \times J(\gamma_{lj}; \gamma_{lj}^*)}{\prod_{i=1}^N f(y_{ij}|\beta_{ij}, \gamma_{lj}, \phi_j; s, X, G) \times \pi(\beta_{ij}|\gamma_{lj}) \times \pi(\gamma_{lj}) \times J(\beta_{ij}; \beta_{ij})}
\end{align*}
\]

Further update of \( \beta_{lj} \) when \( \gamma_{lj} = 1 \): A within-model step is followed to further update \( \beta_{lj} \) that corresponds to \( \gamma_{lj} = 1 \). We first propose a new \( \beta_{lj}^* \) from the normal distribution \( N(\beta_{lj}, (\sigma_\beta/2)^2) \) and then accept the proposed value with probability \( \min(1, r_\beta) \), where

\[
\begin{align*}
    r_\beta &= \frac{\prod_{i=1}^N f(y_{ij}|\beta_{ij}^*, \gamma_{lj}, \phi_j; s, X, G) \times \pi(\beta_{ij}^*|\gamma_{lj}) \times \pi(\gamma_{lj}) \times J(\beta_{ij}; \beta_{ij}^*)}{\prod_{i=1}^N f(y_{ij}|\beta_{ij}, \gamma_{lj}, \phi_j; s, X, G) \times \pi(\beta_{ij}|\gamma_{lj}) \times \pi(\gamma_{lj}) \times J(\beta_{ij}; \beta_{ij})}
\end{align*}
\]

C.2. Simulated Data Generative Model

In our simulation study, we aimed to validate the accuracy of our model. We selected \( N = 500 \) adjacent spots from the human breast cancer 10x Visium data (see Figure 2A in the manuscript) with known spot locations \( T^Y \). The cell locations \( T^X \) and the covariate matrix \( X \), including cell types and nuclei shape features, were obtained from the nuclei identification results through HD-staining on the paired histology image. As detailed in Methods Section of the manuscript, in real SRT data, typically only a subset of covariates correlate with gene expression for most genes. To replicate this scenario, we generated coefficients with high sparsity. Specifically, for cell-type-related explanatory variables, we assumed all corresponding coefficients to be non-zero and normally distributed as \( \beta_{ij} \sim N(1, 1) \). In contrast, for nuclei-shape-related explanatory variables, we produced highly sparse coefficients as follows:

\[
\beta_{ij} \sim 0.9I(\beta_{ij} = 0) + 0.1[0.5\text{Unif}(-1, -0.5) + 0.5\text{Unif}(0.5, 1)].
\]
Consequently, the coefficient matrix $B$ exhibited 90% sparsity. We then generated single-cell-resolution expression levels in line with our model assumption $\theta_j = \exp(X\beta_j)$. Spot-resolution relative expression levels for gene $j$ (i.e., $\lambda_j$) were calculated as $\lambda_j = (G\theta_j)/(G1)$, where $G$ is derived from spot and cell locations (i.e., $T^Y$ and $T^X$), and the denominator $G1$ represents the number of cells covered by each spot. To simulate observed spot-resolution gene expression counts, we followed the process outlined in [20] and [106]. Specifically, we generated the size factor $s_i \sim \text{Unif}(0.5, 1.5)$ for $i = 1, \ldots, N$ and dispersion parameter $\phi_j \sim \text{Exp}(0.1)$ for $j = 1, \ldots, P$, and subsequently, each read count $y_{ij} \sim \text{NB}(s_i\lambda_{ij}, \phi_j)$. This data generation process was repeated for $P = 300$ genes. Finally, with the generated $Y$ and the observed $X$ and $G$, we examined our model’s ability in reconstructing the single-cell-resolution molecular profile $\Theta$.

C.3. Supporting Figures and Tables

Figure C.1: Geometric representations of a barcoded spot (in red) and its expanded region (in green) on the spatial transcriptomics (ST) and the improved 10x Visium platforms. For ST, the spot diameter is $55\mu m$ with a center-to-center distance of $100\mu m$ between two adjacent spots. For 10x Visium, these measures are $100\mu m$ and $200\mu m$, respectively. The percentage of area covered by ST barcoded spots can be approximated by the ratio of the red circle area to the green square, calculated as $\frac{\pi \times (100\mu m)^2}{(200\mu m)^2} = 19.63\%$. Similarly, the percentage of area covered by 10x Visium barcoded spots can be approximated by the ratio of the red circle to the green hexagon, calculated as $\frac{\pi \times \left(\frac{55\mu m}{2}\right)^2}{\frac{3}{\sqrt{3}} \times \left(\frac{100\mu m}{2}\right)^2} = 27.43\%$. 
Figure C.2: Overview of model validation for the simulated data: Evaluation results in terms of the root mean square error (RMSE) between the actual and estimated covariate coefficient $\beta_{lj}$’s for BayesDeep with and without regularization, respectively.

Figure C.3: Overview of model validation for the human prostate cancer 10x Visium data, including the validation settings and evaluation results in terms of the Pearson correlation coefficients $\rho$ between the actual and predicted gene expression counts ($y_{ij}$’s vs. $\hat{y}_{ij}$’s) for BayesDeep, TESLA, and Gaussian Process (GP), respectively.
Figure C.4: Downstream analysis on the human breast cancer 10x Visium data: A. BayesDeep-generated single-cell-resolution expression of the breast cancer stem cell markers CD44 depicted on the UMAP. B. BayesDeep-generated single-cell-resolution expression of the breast cancer stem cell markers CD24 depicted on the UMAP. C. Pseudotime analysis on the UMAP.
Figure C.5: Downstream analysis on the human prostate cancer 10x Visium data: A. BayesDeep-generated single-cell-resolution expression of the prostate cancer stem cell markers ITGA6 depicted on the UMAP. B. BayesDeep-generated single-cell-resolution expression of the prostate cancer stem cell markers ALCAM depicted on the UMAP. C. Pseudotime analysis on the UMAP.
Table C.1: A list of key notations of the proposed BayesDeep.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Support</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>$\mathbb{N}$</td>
<td>The number of spots</td>
</tr>
<tr>
<td>$P$</td>
<td>$\mathbb{N}$</td>
<td>The number of genes</td>
</tr>
<tr>
<td>$M$</td>
<td>$\mathbb{N}$</td>
<td>The number of cells</td>
</tr>
<tr>
<td>$L$</td>
<td>$\mathbb{N}$</td>
<td>The number of explanatory variables</td>
</tr>
<tr>
<td>$Y = [y_{ij}]_{N \times P}$</td>
<td>$y_{ij} \in \mathbb{N}$</td>
<td>The gene expression measurement table with each entry $y_{ij}$ being the read count for gene $j$ observed at spot $i$</td>
</tr>
<tr>
<td>$s = [s_i]_{N \times 1}$</td>
<td>$s_i \in \mathbb{R}^+$</td>
<td>The spot-specific size factors</td>
</tr>
<tr>
<td>$T^Y = [t^Y_{ir}]_{N \times 2}$</td>
<td>$(t^Y_{i1}, t^Y_{i2}) \in \mathbb{R}^2$</td>
<td>The spot location with each row $(t^Y_{i1}, t^Y_{i2})$ being the $x$ and $y$ coordinates of spot $i$</td>
</tr>
<tr>
<td>$X = [x_{ml}]_{M \times L}$</td>
<td>$x_{ml} \in \mathbb{R}$</td>
<td>The covariate matrix</td>
</tr>
<tr>
<td>$T^X = [t^X_{mr}]_{M \times 2}$</td>
<td>$(t^X_{m1}, t^X_{m2}) \in \mathbb{R}^2$</td>
<td>The cell location with each row $(t^X_{m1}, t^X_{m2})$ being the $x$ and $y$ coordinates of cell $m$</td>
</tr>
<tr>
<td>$G = [g_{im}]_{N \times M}$</td>
<td>$g_{im} \in {0, 1}$</td>
<td>The spot-cell spatial relationship matrix with $g_{im} = 1$ indicating that cell $m$ locates within spot $i$ and $g_{im} = 0$ otherwise</td>
</tr>
<tr>
<td><strong>Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\phi = [\phi_j]_{P \times 1}$</td>
<td>$\phi_j \in \mathbb{R}^+$</td>
<td>The gene-specific negative binomial dispersion parameter</td>
</tr>
<tr>
<td>$\Lambda = [\lambda_{ij}]_{N \times P}$</td>
<td>$\lambda_{ij} \in \mathbb{R}^+$</td>
<td>The relative gene expression at the spot resolution</td>
</tr>
<tr>
<td>$\Theta = [\theta_{mj}]_{M \times P}$</td>
<td>$\theta_{mj} \in \mathbb{R}^+$</td>
<td>The relative gene expression at the single-cell resolution</td>
</tr>
<tr>
<td>$B = [\beta_{lj}]_{L \times P}$</td>
<td>$\beta_{lj} \in \mathbb{R}$</td>
<td>The covariate coefficient matrix with each entry $\beta_{lj}$ signifying the association between explanatory variable $l$ and gene $j$</td>
</tr>
<tr>
<td>$\Gamma = [\gamma_{lj}]_{L \times P}$</td>
<td>$\gamma_{lj} \in {0, 1}$</td>
<td>The selection matrix with each entry $\gamma_{lj} = 1$ indicating the corresponding covariate coefficient $\beta_{lj} \neq 0$ while $\gamma_{lj} = 0$ indicating $\beta_{lj} = 0$</td>
</tr>
<tr>
<td><strong>Hyper-parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_\phi$</td>
<td>${0} \cup \mathbb{R}^+$</td>
<td>The shape parameter in the gamma prior for each $\phi_j$</td>
</tr>
<tr>
<td>$b_\phi$</td>
<td>${0} \cup \mathbb{R}^+$</td>
<td>The rate parameter in the gamma prior for each $\phi_j$</td>
</tr>
<tr>
<td>$\sigma^2_\beta$</td>
<td>$\mathbb{R}^+$</td>
<td>The variance parameter of the slab component in the spike-and-slab prior for each $\beta_{lj}$</td>
</tr>
<tr>
<td>$\pi_\gamma$</td>
<td>$(0, 1)$</td>
<td>The parameter in the Bernoulli prior for each $\gamma_{lj}$</td>
</tr>
</tbody>
</table>
Table C.2: A list of explanatory variables utilized by BayesDeep in real data analysis, including cell type classification and nuclei shape features extracted from the paired histology image by HD-staining

<table>
<thead>
<tr>
<th>Cell type</th>
<th>human breast cancer 10x Visium dataset</th>
<th>human prostate cancer 10x Visium dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lymphocyte</td>
<td>Lymphocyte</td>
<td></td>
</tr>
<tr>
<td>Macrophage</td>
<td>Macrophage</td>
<td></td>
</tr>
<tr>
<td>Necrosis</td>
<td>Necrosis</td>
<td></td>
</tr>
<tr>
<td>Red blood cell</td>
<td>Red blood cell</td>
<td></td>
</tr>
<tr>
<td>Stromal cell</td>
<td>Stromal cell</td>
<td></td>
</tr>
<tr>
<td>Tumor cell</td>
<td>Tumor cell</td>
<td></td>
</tr>
<tr>
<td>Ductal epithelium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morphological</td>
<td>Area</td>
<td>Area</td>
</tr>
<tr>
<td>feature</td>
<td>Convex area</td>
<td>Convex area</td>
</tr>
<tr>
<td></td>
<td>Eccentricity</td>
<td>Eccentricity</td>
</tr>
<tr>
<td></td>
<td>Extent</td>
<td>Extent</td>
</tr>
<tr>
<td></td>
<td>Filled area</td>
<td>Filled area</td>
</tr>
<tr>
<td></td>
<td>Major axis length</td>
<td>Major axis length</td>
</tr>
<tr>
<td></td>
<td>Minor axis length</td>
<td>Minor axis length</td>
</tr>
<tr>
<td></td>
<td>Orientation</td>
<td>Orientation</td>
</tr>
<tr>
<td></td>
<td>Perimeter</td>
<td>Perimeter</td>
</tr>
<tr>
<td></td>
<td>Solidity</td>
<td>Solidity</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


[38] A. Wagner, A. Regev and N. Yosef, Revealing the vectors of cellular identity with single-cell genomics, Nature biotechnology 34 (2016) 1145–1160. 6


[40] R. Dong and G.-C. Yuan, SpatialDWLS: accurate deconvolution of spatial transcriptomic data, Genome biology 22 (2021) 145. 6

[42] V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li et al., Cell2location maps fine-grained cell types in spatial transcriptomics, Nature biotechnology 40 (2022) 661–671. 6


[61] P. Zhao, J. Zhu, Y. Ma and X. Zhou, *Modeling zero inflation is not necessary for spatial transcriptomics, Genome Biology* 23 (2022) 118. 9


[71] F. Zhou, K. He, Q. Li, R. S. Chapkin and Y. Ni, *Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization, Biostatistics* 23 (2022) 891–909. 15

[73] Y. Wang, B. Song, S. Wang, M. Chen, Y. Xie, G. Xiao et al., De-noising spatial transcriptomics data based on position and image information, 2021. 10.1101/2021.11.03.467103. 15


[75] W. Lenz, Contribution s t o the s t y l e understanding of the magnetic properties in solid bodies, Physical Z 21 (1920) 613–615. 17


[86] B. W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure 405 (1975) 442–451. 25


X. Jiang, G. Xiao and Q. Li, A bayesian modified ising model for identifying spatially variable genes from spatial transcriptomics data, Statistics in Medicine 41 (2022) 4647–4665. 38, 135


P. Kontkanen, P. Myllymaki and H. Tirri, Constructing bayesian finite mixture models by the EM algorithm, in ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT, Citeseer, 1996. 45


U. Neckmann, C. Wolowczyk, M. Hall, E. Almaas, J. Ren, S. Zhao et al., GREM1 is associated with metastasis and predicts poor prognosis in ER-negative breast cancer patients, Cell Communication and Signaling 17 (2019) 1–17. 58

C. Wang, J. Lv, C. Xue, J. Li, Y. Liu, D. Xu et al., Novel role of COX6c in the regulation of oxidative phosphorylation and diseases, Cell Death Discovery 8 (2022) 336. 59

[121] M. A. Moses, Y. S. Kim, G. M. Rivera-Marquez, N. Oshima, M. J. Watson, K. E. Beebe et al., Targeting the Hsp40/Hsp70 chaperone axis as a novel strategy to treat castration-resistant prostate cancer, Cancer research 78 (2018) 4022–4035. 61

[122] Y. Gao, J. Teng, Y. Hong, F. Qu, J. Ren, L. Li et al., The oncogenic role of EIF3D is associated with increased cell cycle progression and motility in prostate cancer, Medical Oncology 32 (2015) 1–8. 62

[123] G. Daniels, Y. Li, L. L. Gellert, A. Zhou, J. Melamed, X. Wu et al., TBLR1 as an ar coactivator selectively activates AR target genes to inhibit prostate cancer growth, Endocrine-related cancer 21 (2014) 127. 62


[125] R. Chatterjee and U. Chatterji, CLEC12A: A promise target for cancer therapy, Archives of Clinical and Medical Case Reports 6 (2022) 706–714. 64

[126] Y.-Q. Wang, M.-D. Xu, W.-W. Weng, P. Wei, Y.-S. Yang and X. Du, BCL6 is a negative prognostic factor and exhibits pro-oncogenic activity in ovarian cancer, American journal of cancer research 5 (2015) 255. 64


[134] 10x Genomics, Datasets: Spatial gene expression, 2020. 68

[136] L. Zhang, D. Chen, D. Song, X. Liu, Y. Zhang, X. Xu et al., Clinical and translational values of spatial transcriptomics, Signal Transduction and Targeted Therapy 7 (2022) 111. 68


[139] M. Zhang, Q. Li and Y. Xie, A bayesian hierarchical model for analyzing methylated rna immunoprecipitation sequencing data, Quantitative Biology 6 (2018) 275–286. 75


[141] D. M. Witten, Classification and clustering of sequencing data using a Poisson model, . 76


[145] J. Lu, Y. Sheng, W. Qian, M. Pan, X. Zhao and Q. Ge, scRNA-seq data analysis method to improve analysis performance, IET nanobiotecnology (2023) . 76


[161] B. G. De Jong, H. IJspeert, L. Marques, M. Van Der Burg, J. J. Van Dongen, B. G. Loos et al., Human IgG2-and IgG4-expressing memory B cells display enhanced molecular and phenotypic signs of maturity and accumulate with age, Immunology and cell biology 95 (2017) 744–752. 89


[164] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse et al., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, Nature biotechnology 32 (2014) 381–386. 91


