

Title	Methodology development and application to analyse spatial transcriptomics data
Authors	Wang, Qingyue
Publication date	2025-12-31
Original Citation	Wang, Q. 2025. Methodology development and application to analyse spatial transcriptomics data. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2025, Qingyue Wang. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-10-07 17:14:42
Item downloaded from	https://hdl.handle.net/10468/18793

Ollscoil na hÉireann, Corcaigh
National University of Ireland, Cork



Methodology development and application to analyse spatial transcriptomics data

Thesis presented by

Qingyue Wang

for the degree of

Doctor of Philosophy

Supervisor:Dr. Jian Huang

Advisor:Dr. Michael Cronin

University College Cork

School of Mathematical Sciences

Head of School: Dr. Martin Kilian

2025

The future depends on what you do today.

Abstract

Spatial transcriptomics is a rapidly developing technology that enables high-throughput profiling of gene expression at defined spatial locations within intact tissues, thereby linking molecular measurements with the histological context. By enabling the study of cells in their native microenvironment, these technologies provide new opportunities to investigate spatial heterogeneity, cell-to-cell interactions, and the organization of complex biological systems. However, analytical challenges such as mixed-cell capture spots, sparse gene expression, and limited integration with proteomic data necessitate the development of dedicated statistical methodologies.

The overarching aim of this thesis is both methodological development and application. To address data sparsity and improve cell type identification, this thesis introduces a clustering-based deconvolution framework that explicitly incorporates spatial coherence to inform deconvolution strategies. In addition, the analytical scope of spatial transcriptomics is extended through the development of predictive modeling approaches to infer protein abundance from transcriptomic data, thereby offering a pathway toward spatial multi-omic integration. Large-scale benchmarking and systematic validation across multiple experimental datasets provide a rigorous evaluation of performance, generalizability, and computational trade-offs for a diverse set of statistical and machine learning models. Applications to real biological data illustrate the ability of these methods to delineate spatially organized domains, characterize protein-specific variability in predictability, and generate biologically interpretable hypotheses on tissue organization and cellular behavior.

In conclusion, this thesis advances the statistical foundations of spatial transcriptomics by introducing novel methodologies that address core analytical limitations, establishing frameworks for rigorous model evaluation, demonstrating how these approaches can yield biologically meaningful insights. Collectively, these contributions help transform spatial transcriptomics into a more robust and comprehensive platform for integrative and quantitative analysis in biomedical research.

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Digital signature of the candidate:

Qingyue Wang

Acknowledgement

The four-year doctoral journey has been a truly transformative experience, with the completion of this thesis marking an important milestone. This period has encompassed both moments of joy and accomplishment, as well as challenges and setbacks that tested resilience and perseverance. It has not only deepened my understanding of scientific research but also broadened my horizons and opened new opportunities for personal and professional growth. I am profoundly grateful to all those who have offered support, guidance, and encouragement along the way.

To **Dr. Jian Huang**, my main supervisor. You have been my guide throughout this doctoral journey, leading me onto the path of scientific research with patience and dedication. Your insightful guidance and encouragement have been invaluable in shaping both my academic progress and personal growth. Whenever I encountered difficulties, your thoughtful advice and constructive suggestions helped me overcome obstacles and move forward. Your broad vision and deep understanding of the field inspired me to explore new directions, make critical choices, and remain focused on the essential questions. Beyond academic mentorship, your support has instilled in me the confidence to pursue independent research and the determination to strive for excellence. For your guidance, wisdom, and encouragement, I am deeply grateful.

To **Dr. Trung Nghia Vu**, my main collaborator. Without your vision and leadership at the very beginning, this work would never have taken shape. You not only provided the foundation and direction for the project but also continuously inspired me with your creativity, perseverance, and dedication to scientific discovery. Throughout the years of collaboration, you offered invaluable guidance on both technical and conceptual levels, and your high standards of rigor motivated me to refine my ideas and strive for excellence. When I faced difficulties or moments of self-doubt, your encouragement and unwavering support helped me regain confidence and move forward. I will never forget your words, “*If you don’t believe in yourself, believe in us,*” which became a source of strength during the most challenging phases of this journey.

To **Professor Yudi Pawitan**, my academic grandfather. It has been a profound privilege to work with you, and I am sincerely grateful for your generous guidance and unwavering support. I feel fortunate to have studied within such a scholarly and inspiring research family, where your wisdom and mentorship have been a constant source of motivation. I am continually impressed by your intellectual agility and problem-solving ability, and I truly believe there is no question beyond your insight. Among the many memorable experiences, I will always treasure our trip to Vietnam and the joy of solving puzzles together, which reminded me of the excitement of curiosity and discovery.

To all members of my research family in Sweden:

To **Gulser Caliskan**, my sister. Thank you for your companionship and hospitality during my time in Sweden, which made my temporary stays far less lonely and much more enjoyable.

To **Yuying Li**, my sister and closest friend during my doctoral studies. I am deeply grateful for your constant help whenever I encountered difficulties and for the countless conversations that brought comfort and joy during moments of boredom.

To **Quang Thinh Trac**, my brother. Although most of our communication has taken place online, I have always felt your warmth and support whenever we met in person, and that has meant the world to me.

To my family:

To **Tinghao Liu**, my husband. Thank you for your constant love, patience, and support in life. You have been my strongest pillar throughout this journey, and I am deeply grateful to have always had you by my side. From Harbin to Cork, we have walked this path together, sharing both challenges and achievements, and growing side by side. I feel truly fortunate for the bond we share, and I look forward to a future in which we continue to support one another, striving together and shining in our respective fields.

To **my parents and parents-in-law**. Thank you for your unconditional support and care, which gave me the strength and peace of mind to focus on my studies.

To all my friends who have supported me in different ways during this journey. Thank you for your kindness, encouragement, and companionship, which have made this experience more enjoyable and meaningful.

Table of Contents

1	Overview of Thesis	1
1.1	Introduction	1
1.2	Main Contributions	2
1.3	Thesis Organization	4
2	Background	7
2.1	Evolution of spatial transcriptomics	7
2.2	Technologies for spatial transcriptomics	9
2.2.1	ROI selection technologies	10
2.2.2	<i>In situ</i> hybridization technologies	12
2.2.3	<i>In situ</i> sequencing technologies	13
2.2.4	Spatial barcoding technologies	15
2.3	Development trends and key challenges in spatial transcriptomics	16
2.3.1	Emerging trends	16
2.3.2	Limitations and challenges	18
2.4	Spatial multi-omics integration	20
2.5	Statistical perspective on spatial transcriptomics data	22
3	Literature Review	25
3.1	Computational strategies for cell type decomposition in spatial transcriptomics	25
3.1.1	General workflow of deconvolution approaches	27
3.1.2	Reference-based deconvolution models	30
3.1.3	Reference-free deconvolution models	32
3.1.4	Emerging regression-based approaches	33
3.1.5	Comparative evaluation and future perspectives	33
3.2	Inferring spatial protein abundance from transcriptomic profiles	34
3.2.1	Experimental platforms enabling paired RNA–protein measurements	35
3.2.1.1	Imaging-based multiplexed spatial proteomics	36
3.2.1.2	Sequencing-based spatial multi-omics	37
3.2.1.3	Commercial platforms for integrated RNA–protein profiling	37

3.2.2	Computational Strategies for Protein Inference	38
3.2.2.1	Image-Based Virtual Staining Approaches	38
3.2.2.2	Single-Cell RNA–Protein Prediction Models	38
3.2.2.3	Spatial Transcriptomics-Oriented Inference Frameworks	39
3.2.3	Current Challenges and Future Directions	40
3.3	Conclusion	41
4	A Cluster-Based Cell-Type Deconvolution of Spatial Transcriptomic Data	43
4.1	Introduction	44
4.2	Materials and methods	45
4.2.1	DECLUST pipeline	45
4.2.1.1	Identification of spatial clusters of spots from ST data	47
4.2.1.2	Discovery of marker genes from scRNA-seq data	51
4.2.1.3	Cell-type deconvolution using ordinary least squares	51
4.2.2	Datasets	53
4.2.2.1	Spatial transcriptomics data generation and preprocessing	53
4.2.2.2	Real datasets	54
4.2.2.3	Simulated ST datasets	55
4.2.3	Competing cell-type deconvolution methods	56
4.2.4	Performance metrics for simulation study	59
4.2.5	Cross-validation for real ST data	60
4.3	Results	61
4.3.1	Description of DECLUST methodology	61
4.3.2	DECLUST achieves high accuracy in simulated ST data	63
4.3.3	DECLUST performs well against other methods in the analysis of the real mouse brain ST data	66
4.3.4	DECLUST effectively preserves spatial pattern of cell-type marker genes in the real human ovarian cancer ST data	68
4.3.5	Diagnostic Assessment of Residual Heteroscedasticity and Distri- butional Assumptions	71
4.3.6	DECLUST exhibits computational efficiency in cell-type deconvo- lution	74
4.4	Discussion	75
5	A Comparative Study of Predictive Models for Spatial Protein Abundance from Transcriptomic Profiles	79
5.1	Introduction	80
5.2	Materials and methods	82

5.2.1	Overview of the benchmarking framework	82
5.2.2	Prediction models for evaluation	82
5.2.2.1	Linear regression	83
5.2.2.2	Tree-based ensemble models	84
5.2.2.3	Graph neural network models	85
5.2.3	Construction of spatial feature for non-GNN models	87
5.2.4	Validation strategies for the prediction models	88
5.2.4.1	Within-sample validation	88
5.2.4.2	Cross-sample validation	88
5.2.4.3	Cross-tissue validation	89
5.2.5	Performance metrics for evaluation	89
5.2.6	Datasets	90
5.2.6.1	Spatial multi-omics data generation	92
5.2.6.2	Data Processing	92
5.3	Results	92
5.3.1	Comparison of the prediction methods using within-sample cross-validation	92
5.3.2	External validation using samples from the same tissue type	98
5.3.3	External validation using samples from the different tissue types	101
5.3.4	Spatial protein predictions recover immune-rich regions in the HNSC tumor microenvironment and identify protein markers associated with immunotherapy response	105
5.3.5	Comparison of the prediction methods in terms of computational efficiency	108
5.4	Discussion	109
6	Discussion, Conclusions and Future Work	111
6.1	Discussion	111
6.2	Conclusion	112
6.3	Future Work	112
6.3.1	Sequencing-free spatial multi-omics	112
6.3.2	Deep learning for direct cell type deconvolution	113
6.3.3	Multi-modal data integration	113
6.3.4	Adaptive and generalizable frameworks	113
6.3.5	Protein-specific predictive modeling	114
6.3.6	Toward clinical translation	114
	Bibliography	115

A	Appendix: Simulation Results	143
B	Appendix: Spatial Visualization of Predicted Protein Abundance	147

List of Figures

2.1	Central dogma of molecular biology.	7
2.2	Timeline of spatial transcriptomics technologies.	9
2.3	Workflow of ROI selection via laser capture microdissection.	10
2.4	Workflow of <i>in situ</i> hybridization.	12
2.5	Workflow of <i>in situ</i> sequencing.	13
2.6	Workflow of spatial barcoding.	15
3.1	General workflow for spatial transcriptomics deconvolution.	28
4.1	Overview of the DECLUST method.	46
4.2	Overview of the proposed clustering approach of DECLUST.	47
4.3	Spatial expression of Ligand-Receptor pairs used to guide cell-type enrichment in Simulation 2.	57
4.4	Evaluating the contribution of the clustering step in DECLUST.	62
4.5	Comparison of DECLUST with other cell-type deconvolution methods in simulated datasets.	65
4.6	Comparative performance of spatial transcriptomics methods under two simulation settings across 10 replicated datasets in Simulation 1 and Simulation 2.	66
4.7	Cell-type deconvolution results for the real mouse brain ST data.	67
4.8	Cell-type deconvolution results for the real human ovarian cancer ST data.	69
4.9	Spatial expression patterns of marker genes for selected cell types across varying marker set sizes.	70
4.10	Gene-wise mean residuals versus predicted expression across clusters.	72
4.11	Spot-level residual distributions and normal Q–Q diagnostics for representative genes.	72
4.12	Computational time of the deconvolution methods in three ST datasets.	74
5.1	Overview of the benchmarking framework for methods predicting spatial protein abundance from spatial RNA expression profiles.	83
5.2	Spatial 10-fold cross-validation splits for within-sample evaluation.	93

5.3	Protein prediction performance in six spatial multi-omics datasets with within-sample validation (Spearman correlation).	94
5.4	Protein prediction performance with within-sample validation (Pearson correlation).	95
5.5	Protein prediction performance with within-sample validation (RMSE).	95
5.6	Within-sample prediction performance in Tonsil and Glioblastoma.	97
5.7	Within-sample prediction performance in HNSC and Breast Cancer.	98
5.8	Cross-sample distribution of total protein abundance.	99
5.9	Evaluation of the protein prediction methods using cross-sample validation.	100
5.10	Evaluation of the protein prediction methods using cross-sample validation.	101
5.11	Cross-tissue validation of the prediction methods across six samples.	101
5.12	Cross-tissue prediction performance in Tonsil and Breast Cancer.	103
5.13	Cross-tissue prediction performance in HNSC and Glioblastoma.	104
5.14	Deconvolution maps for a poorly immune-infiltrated HNSC sample.	105
5.15	Spatial characterization of tumor immune microenvironment features using predicted protein abundance.	106
5.16	Computational time across 10 models in the cross-tissue validation.	108
A.1	Accuracy of cell-type proportion predictions across multiple methods in Simulation 1.	143
A.2	Accuracy of cell-type proportion predictions across multiple methods in Simulation 2.	145
B.1	Cross-tissue prediction of PCNA and CD274 expression in breast cancer.	147
B.2	Cross-tissue prediction of VIM and CD27 expression in glioblastoma.	148
B.3	Cross-tissue prediction of EPCAM and HLA-DRA expression in head and neck squamous cell carcinoma (HNSC, A1).	149
B.4	Cross-tissue prediction of EPCAM and HLA-DRA expression in head and neck squamous cell carcinoma (HNSC, D1).	150
B.5	Cross-tissue prediction of CD8A and CD3E expression in human tonsil (sample 1).	151
B.6	Cross-tissue prediction of CD8A and CD3E expression in human tonsil (sample 2).	152

List of Tables

2.1	Commercially available spatial transcriptomics platforms.	10
2.2	Recent advances and emerging directions in spatial transcriptomics. . .	17
2.3	Current challenges and unresolved issues in spatial transcriptomics. . .	19
2.4	Spatial multi-omics integration technologies.	20
3.1	Overview of spatial transcriptomics deconvolution methods.	26
3.2	Overview of experimental platforms enabling paired RNA–protein spatial profiling.	36
4.1	Summary of competing cell-type deconvolution methods.	58
4.2	Gene-level residual summary statistics for representative genes used in spot-level diagnostics.	73
4.3	Comparison of five deconvolution methods by computational time (seconds). .	74
5.1	Pairwise comparison results between responders and non-responders at the cluster level.	107

Overview of Thesis

In this chapter, we begin with a brief introduction of spatial transcriptomics, highlighting its transformative role in biomedical research. Following this, we outline the key analytical challenges associated with spatial transcriptomics, including issues of data sparsity, spatial structure, and model evaluation. These challenges motivate the development of new statistical methodologies, which form the core contributions of this thesis. Finally, we present an outline of the thesis structure and summarize the main objectives guiding the research.

1.1 Introduction

Spatial transcriptomics is an innovative technology that enables the measurement of gene expression within its native spatial context in tissues. Unlike traditional transcriptomic approaches that dissociate cells and lose spatial information, spatial transcriptomics preserves the architecture of biological samples, allowing researchers to study how gene activity varies across different regions of a tissue. This spatial resolution is crucial for understanding cellular heterogeneity, tissue organization, and complex biological processes such as development, disease progression, and immune responses. As a result, spatial transcriptomics has become a powerful tool in biomedical research, with applications spanning oncology, neuroscience, and regenerative medicine.

Despite its transformative potential, spatial transcriptomics presents several analytical challenges. First, most current ST platforms, such as 10x Visium, lack single-cell resolution: each spot typically contains a mixture of different cell types [1, 2]. Conventional deconvolution methods operate at the spot level, but these analyses are hampered by sparse gene expression and by the neglect of spatial relationships across neighboring spots, ultimately reducing accuracy and robustness. Second, while transcriptomics captures gene expression, proteins are the primary effectors of cellular function. Spatial proteomics technologies remain technically limited and costly, resulting in most ST datasets lacking comprehensive protein measurements [3, 4, 5]. Inferring protein abundance directly from spatial transcriptomic data therefore represents an important computational challenge, one that could significantly expand the scope of spatial multi-omics research.

This thesis is motivated by these challenges and is organized around two central research directions: (1) the development of DECLUST, a cluster-based framework for spatial deconvolution that improves accuracy by aggregating spatially coherent spots into biologically meaningful domains; and (2) a systematic benchmarking of predictive models for protein abundance inference from ST data, evaluating the utility of diverse machine learning and graph-based approaches under multiple validation scenarios.

To address these directions, the work presented here is guided by three overarching questions:

1. What ways can spatial transcriptomics be utilized to uncover biologically meaningful insights, particularly through the accurate resolution of cell type composition within heterogeneous tissue environments?
2. Which statistical and computational strategies are most suitable for modeling spatial transcriptomic data, ensuring that both molecular features and spatial dependencies are faithfully represented?
3. How can transcriptomic profiles be systematically leveraged to predict spatial protein abundance, and what forms of feature representation best capture the regulatory and functional link between RNA and proteins?

1.2 Main Contributions

The work presented in this thesis is centered around two main contributions to the analysis of spatial transcriptomics and spatial multi-omics data. Each contribution is described in detail below.

1. **DECLUST: a cluster-based framework for cell type deconvolution.** This thesis introduces DECLUST, a novel method for accurate estimation of cell type composition in spatial transcriptomic (ST) data. Its main contributions are:
 - *Integration of spatial and molecular features.* DECLUST simultaneously incorporates spatial coordinates and gene expression profiles to cluster ST spots, ensuring that spatial domains are both biologically meaningful and spatially coherent.

- *Cluster-level deconvolution with pseudo-bulk profiles.* By aggregating expression across spatially coherent clusters, DECLUST mitigates sparsity and noise in ST data, thereby enhancing robustness in deconvolution.
- *Accurate identification of marker genes.* A two-statistic approach is applied to scRNA-seq reference data to select cell-type-specific marker genes, reducing noise and improving estimation precision.
- *Performance and robustness.* Evaluation on simulated and real datasets demonstrates that DECLUST consistently outperforms competing methods in terms of accuracy, robustness, and preservation of spatial structure.
- *Biological interpretability and extensibility.* Beyond deconvolution, the identified clusters may also act as proxies for tissue microenvironments or functional domains. Future extensions could incorporate additional modalities, including ligand–receptor interactions, proteomics, and histology, to further improve interpretability.

2. **Protein abundance prediction from spatial transcriptomics data.** The second contribution is a systematic benchmarking of computational methods for predicting protein abundance from RNA profiles in spatial multi-omics datasets. The major contributions are:

- *Large-scale model evaluation.* Ten predictive models, including linear regression, partial least squares (PLS), tree-based ensemble methods, and graph neural networks (GNNs), are comprehensively assessed across six datasets generated by Visium CytAssist CITE-seq.
- *Multiple validation strategies.* Model performance is systematically benchmarked using within-sample, cross-sample, and cross-tissue validation schemes, providing a broad perspective on robustness and generalizability.
- *Importance of spatial information.* Spatially augmented models, such as PLS-Spatial and boosting-based ensembles (LightGBM-Spatial, XGBoost-Spatial, CatBoost-Spatial), consistently outperform their non-spatial counterparts, demonstrating the value of incorporating spatial context into predictive modeling.

- *Model-specific trade-offs.* Gradient boosting models achieve the highest predictive accuracy but incur longer runtimes, whereas PLS offers competitive performance with extremely low computational cost. GNN models, although computationally efficient, show limited predictive accuracy, highlighting both their current shortcomings and their potential for future improvement.
- *Protein-specific predictability.* Performance heterogeneity across proteins is systematically characterized, enabling the identification of proteins with high or low predictability. This provides practical guidance for prioritizing targets in spatial protein inference.
- *Feasibility of RNA-based protein prediction.* By addressing the novel question of whether protein abundance can be inferred directly from spatial transcriptomic data, this study demonstrates the feasibility of RNA-based prediction and establishes it as a viable strategy for extending the scope of spatial multi-omics research.

1.3 Thesis Organization

This thesis is structured as follows:

Chapter 2 provides the necessary background for the thesis. It begins with a statistical perspective on the analysis of high-dimensional biological data, introduces the emergence of spatial transcriptomics, and describes the major experimental technologies including in situ hybridization, in situ sequencing, and spatial barcoding. The chapter also highlights the limitations of current platforms and discusses development trends such as spatial multi-omics integration.

Chapter 3 presents a literature review of computational approaches relevant to this thesis. The first part examines strategies for cell type deconvolution in spatial transcriptomics, including reference-based, reference-free, and regression-based frameworks. The second part reviews methods for inferring spatial protein abundance from transcriptomic profiles, outlining current experimental platforms, computational strategies, and their limitations. This chapter identifies methodological gaps that motivate the contributions in subsequent chapters.

Chapter 4 introduces a cluster-based framework for cell type deconvolution in spatial transcriptomics. The methodology integrates spatial clustering with pseudo-bulk expression generation and marker gene selection, aiming to mitigate data sparsity at the spot level while preserving spatial coherence. The chapter further evaluates the approach on both simulated and real datasets, providing comparisons with existing methods and discussing its performance characteristics and limitations.

Chapter 5 addresses the novel problem of predicting protein abundance from spatial transcriptomic data through a large-scale benchmarking study. Ten computational models, ranging from linear regression and ensemble learning to graph neural networks, are systematically assessed on multiple datasets using within-sample, cross-sample, and cross-tissue validation schemes. The analysis highlights the role of spatial feature integration, identifies trade-offs between different model classes, and characterizes protein-specific predictability.

Chapter 6 summarizes the thesis by consolidating its methodological and applied contributions. It discusses how spatially informed deconvolution enhances cell type inference, and how predictive modeling extends spatial transcriptomics toward multi-omic integration. The chapter concludes with perspectives on future directions, including sequencing-free approaches, multi-modal integration, adaptive frameworks, and clinical translation.

Background

This chapter introduces the evolution and technological landscape of spatial transcriptomics, followed by a statistical perspective on spatial transcriptomics data.

2.1 Evolution of spatial transcriptomics

Since the discovery of DNA's double-helical structure in 1953, remarkable progress has been made in understanding how genetic information is stored, transmitted, and expressed. Central to this framework is the principle that information flows from DNA to RNA, and ultimately to proteins (Figure 2.1)-a concept formalized as the central dogma of molecular biology [6, 7]. DNA encodes the blueprint of life, RNA conveys these instructions as a messenger, and proteins serve as the principal effectors of cellular function, carrying out diverse structural and regulatory roles essential for physiology. This continuum from DNA to RNA to proteins constitutes the fundamental logic of modern life sciences and provides the foundation upon which transcriptomics and proteomics are built.

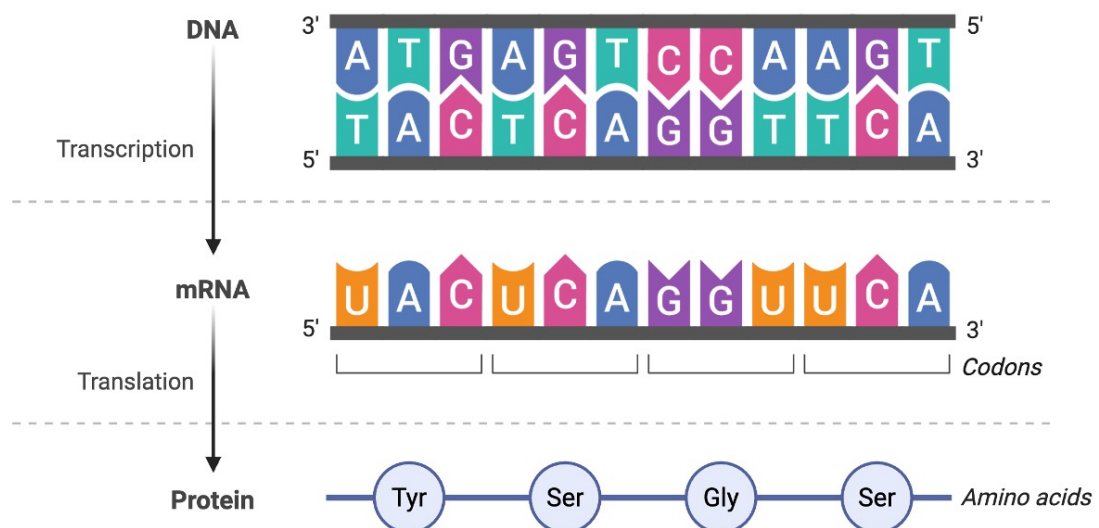


Fig. 2.1.: Central dogma of molecular biology.

Building on this foundation, transcriptomics has emerged as a powerful discipline dedicated to studying how the instructions embedded in DNA are executed through gene expression, providing dynamic snapshots of cellular activity across diverse biological contexts. Understanding gene expression is central to uncovering how cells, tissues, and organisms function [8]. Over the past decade, technological innovations have transformed our ability to measure transcriptomes with remarkable precision [9]. Transcriptomics has diversified into bulk and single-cell approaches, each with distinct advantages and limitations [10, 9]. Early studies relied on bulk RNA sequencing, which captures average expression across large groups of cells [11]. Although effective for identifying global transcriptional changes, bulk measurements obscure cellular heterogeneity and mask the contributions of distinct cell types, limiting interpretation in complex tissues.

The development of single-cell RNA sequencing (scRNA-seq) overcame this limitation by enabling transcriptome-wide profiling at single-cell resolution. This breakthrough revealed profound heterogeneity in cell types, dynamic states, and developmental trajectories that were invisible in bulk data [12, 13, 14, 15]. However, scRNA-seq requires dissociating tissues into suspensions, which eliminates the spatial context necessary for understanding how cells are arranged, how they interact with neighbors, and how they collectively shape tissue structure and disease progression, thereby losing critical information about tissue structure and microenvironmental context [16, 17].

Spatial transcriptomics (ST) emerged to bridge this gap by preserving the positional information of gene expression within intact tissues [18, 19], and was named Method of the Year 2020 by Nature Methods (published in 2021) [20]. By maintaining the spatial relationships among cells and their gene expression profiles, ST has transformed transcriptomic analysis—bridging molecular measurement with spatial localization. This enables investigations of tissue heterogeneity, developmental trajectories, and disease microenvironments in ways that were previously unattainable [21, 22]. By embedding expression data in its native spatial framework, these technologies enable molecular signatures to be directly mapped onto histological structures, tissue domains to be reconstructed, and cell–cell communication to be inferred within their natural microenvironments. The data produced are typically represented as spatially annotated expression matrices, where rows denote genes and columns correspond to spatial units such as capture spots, beads, or single cells. Each unit is linked to precise tissue coordinates, often in conjunction with histological imaging. This structure creates a powerful platform for statistical and computational analyses that jointly leverage gene expression and spatial organization, laying the foundation for new insights into tissue biology and pathology [23, 24].

2.2 Technologies for spatial transcriptomics

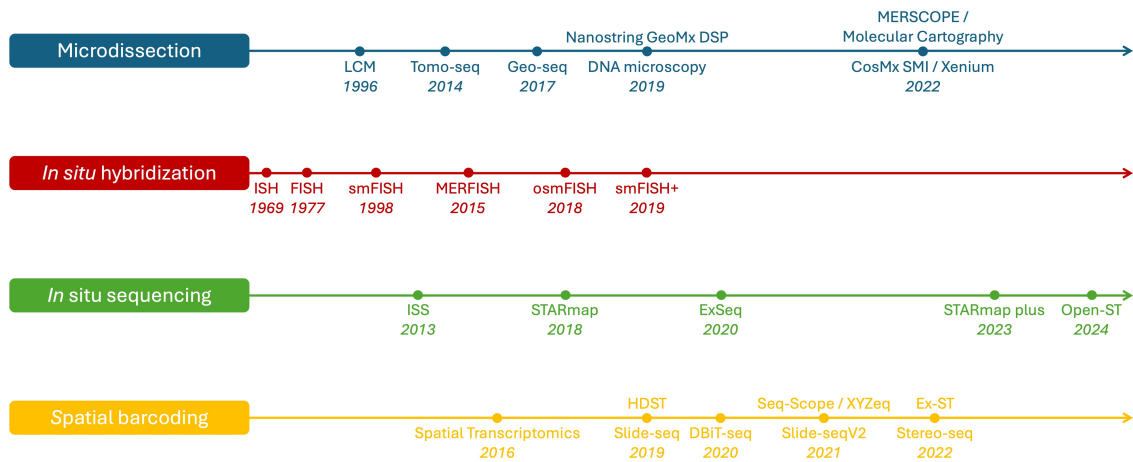


Fig. 2.2.: Timeline of spatial transcriptomics technologies. The development of spatial transcriptomics can be categorized into four main groups: ROI-based methods (blue), *in situ* hybridization methods (red), *in situ* sequencing methods (green), and spatial barcoding methods (yellow).

Spatial transcriptomics (ST) technology, which integrates two-dimensional (2D) and three-dimensional (3D) tissue structural context into molecular investigations, has advanced rapidly since the first spatial transcriptomics workflow in 2016 [25]. We group methods into four families—region-of-interest (ROI) selection [26], *in situ* hybridization [27], *in situ* sequencing [28, 29], and spatial barcoding [25] (Figure 2.2). Detailed workflows, historical milestones, and representative platforms for each family are provided in the following subsections.

In parallel with these methodological advances, several commercial platforms have been widely adopted in both research and translational applications. Representative examples are summarized in Table 2.1, which lists underlying techniques, compatible tissue types, and corresponding websites. Among them, 10x Genomics Visium has become a prevalent barcoding-based solution for large-scale transcriptome profiling, NanoString GeoMx DSP provides ROI-based profiling with flexible tissue compatibility, and Vizgen MERSCOPE enables high-resolution single-molecule imaging through MERFISH.

Tab. 2.1.: Commercially available spatial transcriptomics platforms.

Platform	Technique	Tissue	Website
10x Genomics Visium	Spatial barcoding	Fresh frozen, FFPE	https://www.10xgenomics.com/products/spatial-gene-expression
10x Genomics Visium CytAssist	Spatial barcoding + IF/antibody workflows	Fresh frozen, FFPE	https://www.10xgenomics.com/products/visium-cytassist
NanoString GeoMx DSP	ROI selection (DSP)	Fresh frozen, FFPE	https://nanosttring.com/products/geomx-digital-spatial-profiler/
Vizgen MERSCOPE	MERFISH (<i>in situ</i> hybridization)	Fresh frozen, FFPE	https://vizgen.com/products/
10x Genomics Xenium	Imaging-based RNA (targeted)	Fresh frozen, FFPE	https://www.10xgenomics.com/products/xenium-in-situ
NanoString CosMx SMI	Imaging-based RNA/protein (targeted)	Fresh frozen, FFPE	https://nanosttring.com/products/cosmx-spatial-molecular-imager/

By capturing spatial context, ST technologies enable mapping of tissue organization and function, shedding light on developmental programs and the molecular mechanisms of disease [30, 31, 32].

2.2.1 ROI selection technologies

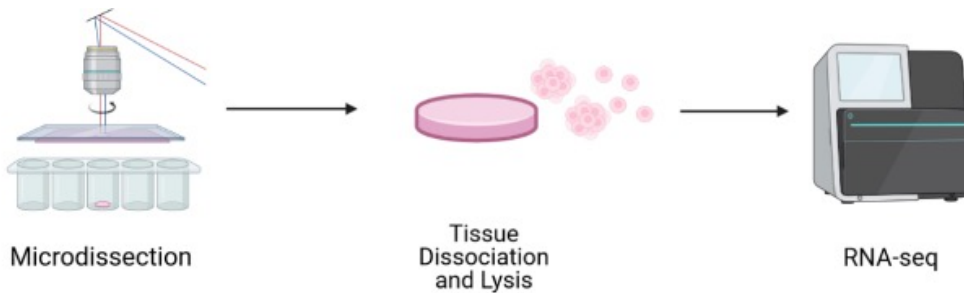


Fig. 2.3.: Workflow of ROI selection via laser capture microdissection. Defined tissue regions are excised, dissociated, and lysed, after which RNA is extracted and sequenced.

Regions of interest (ROIs) can be precisely defined and isolated from tissue sections through either physical dissection or light-based labeling strategies, enabling spatially resolved transcriptomic analysis. Traditional approaches such as laser capture microdissection (LCM) [33] employ a focused laser beam to excise and retrieve specific tissue regions, which are then dissociated and lysed for downstream molecular assays, including RNA-seq, DNA-seq, and epigenetic profiling. As shown in Figure 2.3, the microdissected regions undergo dissociation and lysis, followed by RNA extraction and sequencing, thereby generating transcriptomic profiles that retain spatial context. LCM provides high spatial precision and can, in principle, achieve near single-cell resolution, but the process is

technically demanding, time-consuming, and often compromises RNA integrity during capture, thereby restricting its scalability to larger tissue cohorts.

Building on this foundation, several technologies have extended ROI-based profiling beyond conventional LCM. Tomo-seq reconstructs gene expression atlases by sequential cryosectioning and RNA-seq, enabling systematic mapping of expression patterns across tissue axes [34]. Geo-seq integrates LCM with scRNA-seq protocols, offering high-resolution analysis of spatial heterogeneity and rare cellular populations within defined niches [35]. STRP-seq further advances this concept by slicing tissue sections into thin strips at multiple orientations and computationally reconstructing three-dimensional transcriptomic landscapes, thereby enhancing the capacity to capture tissue-wide spatial organization [36].

In contrast, light-based ROI labeling provides a non-invasive alternative that circumvents physical dissection. Approaches such as TIVA-tag [37] and photo-isolation chemistry (PIC) [38] employ photosensitive groups to trigger reverse transcription directly in selected regions, either in situ or in fixed samples, thereby reducing tissue manipulation. NanoString GeoMx applies photocleavable tags to RNA or protein probes, releasing them upon light activation for subsequent quantification across targeted panels, making it particularly suitable for large-scale cancer transcriptome profiling. More recently, methods such as Light-seq [39] and ZipSeq [40] use light-triggered attachment or release of DNA barcodes, which enables sequencing reads to be reassigned back to their spatial origins with high flexibility.

In conclusion, ROI selection technologies provide versatile tools to interrogate tissue architecture at scales ranging from individual cells to larger domains. Physical dissection approaches offer precise isolation and unbiased transcriptome coverage but are limited by low throughput, technical complexity, and potential RNA degradation. By contrast, light-based strategies enable higher scalability and reduced tissue handling, but often rely on predefined gene panels or barcoding schemes, which may restrict transcriptome completeness. Despite these trade-offs, ROI selection remains a cornerstone of spatial transcriptomics, particularly when precise interrogation of specific microanatomical regions or targeted spatial profiling is required.

2.2.2 *In situ* hybridization technologies

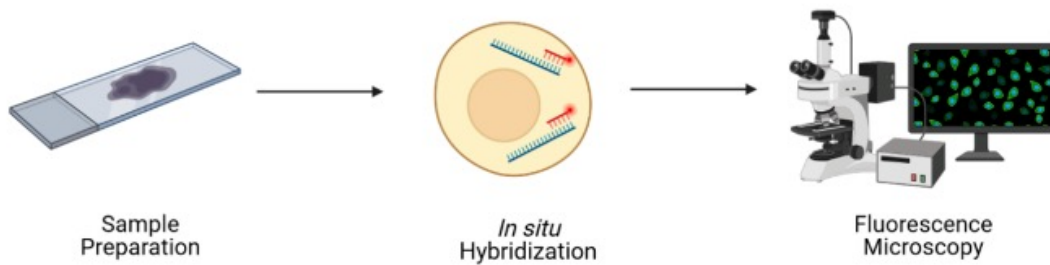


Fig. 2.4.: Workflow of *in situ* hybridization. Tissue samples are prepared and incubated with fluorescently labeled probes that hybridize to target transcripts *in situ*. Fluorescence microscopy is then used to visualize and quantify RNA molecules within their native spatial context.

In situ hybridization represents a class of spatial transcriptomic strategies that visualize RNA molecules directly within their native tissue context using complementary probes, rather than extracting them from tissue sections. As illustrated in Figure 2.4, fluorescently labeled probes hybridize to target transcripts *in situ*, and the resulting signals are imaged using fluorescence microscopy. This principle enables detection of gene expression patterns at cellular and even subcellular resolution, thereby linking molecular signatures with histological architecture.

The earliest and most widely adopted approach is single-molecule fluorescent *in situ* hybridization (smFISH) [41], which decomposes individual RNA molecules into sub-diffraction fluorescence spots and allows simultaneous detection of multiple transcripts. smFISH offers high sensitivity and is often considered the “gold standard” for RNA quantification, especially for validating candidate genes identified through bioinformatic analyses. However, its reliance on distinct fluorescence channels inherently limits multiplexing capacity, restricting detection to a small number of genes per experiment. To address this limitation, cyclic variants such as ouroboros smFISH (osmFISH) [42] have been developed, which repeat hybridization and imaging cycles to expand the number of detectable targets, enabling detection of low-abundance transcripts across larger tissue areas, albeit at the cost of throughput.

To improve multiplexing while mitigating optical crowding, sequential FISH (seqFISH) [43] was introduced. This approach assigns barcodes to transcripts through iterative rounds of hybridization and imaging, with fluorophores removed after each cycle while keeping

the RNA molecules in place. By combining temporal barcoding with probe recycling, seqFISH dramatically increases the number of detectable transcripts, making it possible to reconstruct spatial transcriptomic maps of complex tissues such as the brain. Nevertheless, repeated hybridization steps make the protocol time-intensive and susceptible to cumulative errors across cycles.

To overcome these drawbacks, multiplexed error-robust FISH (MERFISH) [44] introduced an error-robust combinatorial barcoding scheme, in which each RNA is assigned a binary code that can detect and in some cases correct errors. MERFISH achieves highly multiplexed measurements, initially capturing hundreds of genes simultaneously and later extended to transcriptome-scale profiling (up to 10,000 targets). Importantly, MERFISH removes only fluorophores rather than probes, thereby shortening hybridization cycles compared to seqFISH. This approach has been commercialized as Vizgen MERSCOPE, which supports both fresh frozen and formalin-fixed paraffin-embedded (FFPE) samples, enabling its application to a broad range of tissues.

In summary, *in situ* hybridization-based technologies provide powerful tools for visualizing RNA transcripts in their native spatial context, with smFISH offering high sensitivity and validation utility, seqFISH enabling large-scale multiplexing through barcoding, and MERFISH advancing to error-robust, transcriptome-wide applications. While these methods are limited by requirements for probe synthesis and remain relatively low throughput compared to sequencing-based approaches, they uniquely combine single-molecule resolution with direct imaging, making them indispensable for spatial transcriptomic studies.

2.2.3 *In situ* sequencing technologies

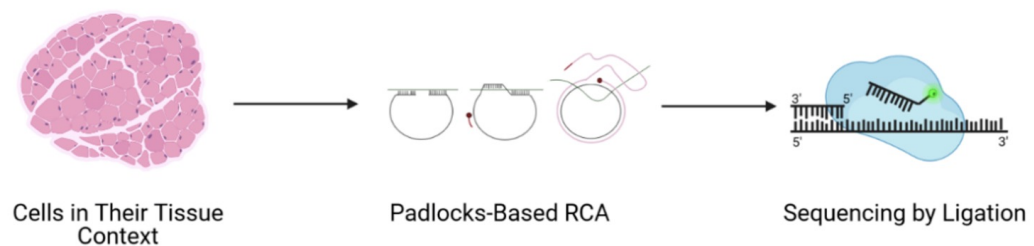


Fig. 2.5.: Workflow of *in situ* sequencing. Padlock probes hybridize to cDNA molecules in their tissue context, followed by rolling-circle amplification. The amplicons are subsequently decoded through sequencing by ligation, enabling spatially resolved transcriptome profiling.

In situ sequencing (ISS) enables targeted or untargeted sequencing of RNA molecules directly within preserved tissue architecture. Unlike traditional sequencing methods that require RNA extraction, ISS maintains the histological context, allowing precise mapping of gene expression patterns at cellular and subcellular resolution. As illustrated in Figure 2.5, the workflow typically involves hybridization of padlock probes to cDNA, rolling-circle amplification (RCA) to generate localized amplicons, and subsequent sequencing by ligation.

The original ISS protocol developed by Ke and colleagues [28] introduced padlock-based probes, which can operate in two modes. In gap-targeted sequencing, the probe contains a small gap that matches targeted bases in the cDNA; this gap is filled by DNA polymerization and sealed by ligation, ensuring high specificity. In barcode-targeted sequencing, the probe contains a unique barcode sequence, enabling simultaneous detection of multiple transcripts through ligation alone. Rolling-circle amplification of the circularized DNA produces amplicons that can be visualized and decoded by ligation-based sequencing chemistry. ISS has been successfully applied to human breast cancer samples to identify point mutations and profile multiplexed gene expression, demonstrating its robustness in both mutation detection and spatial transcriptomics. However, probe design requires prior knowledge of the transcripts of interest, limiting flexibility.

To overcome this limitation, fluorescent *in situ* sequencing (FISSEQ) was developed as a transcriptome-wide, untargeted strategy [45, 29]. FISSEQ enables unbiased detection of thousands of RNA species within intact tissues, thus capturing global expression landscapes. Despite this advantage, its sequencing depth remains low, and inefficiencies in reverse transcription and cDNA circularization constrain sensitivity. Building on FISSEQ, expansion sequencing (ExSeq) [46] enhances both resolution and coverage. By anchoring RNA molecules to an expandable hydrogel and physically enlarging the tissue before sequencing, ExSeq alleviates molecular crowding and achieves nanoscale precision. Targeted ExSeq enables spatial mapping of predefined gene panels, whereas untargeted ExSeq supports transcriptome-wide detection, including rare transcripts.

Another advancement is spatially resolved transcript amplicon readout mapping (STARmap) [47], which integrates hydrogel–tissue chemistry with ISS. STARmap can sequence RNA in intact 3D tissues; its improved version incorporates sequencing with error-reduction by dynamic annealing and ligation (SEDAL), mitigating misdecoding due to sequencing errors. More recently, modified STARmap protocols have demonstrated applicability to thick tissue blocks, expanding the scope of 3D spatial transcriptomics.

Overall, *in situ* sequencing-based technologies enable direct spatial analysis of RNA molecules while preserving tissue morphology. ISS offers high specificity but requires prior transcript knowledge; FISSEQ provides transcriptome-wide coverage at the expense of depth and sensitivity; ExSeq enhances spatial resolution and multiplexing through hydrogel expansion; and STARmap extends ISS into three-dimensional intact tissues with high accuracy. Despite limitations such as short read lengths (typically ~ 30 nucleotides), complex workflows, and dependence on imaging instrumentation, these approaches remain powerful tools for linking gene expression to cellular and tissue architecture.

2.2.4 Spatial barcoding technologies

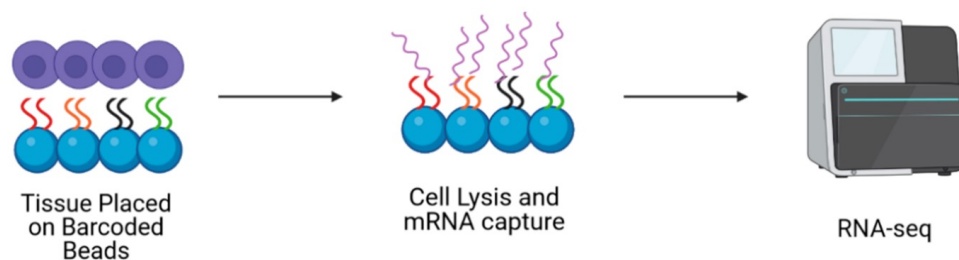


Fig. 2.6.: Workflow of spatial barcoding. Tissue is placed on a surface or bead array containing DNA barcodes. Following cell lysis, mRNAs hybridize to barcoded primers, and captured transcripts undergo reverse transcription and sequencing. The positional barcodes allow reconstruction of transcriptome maps within tissue context.

Spatial barcoding enables the parallel capture of gene expression profiles across tissue sections while preserving positional information. The principle is to place tissue slices onto surfaces carrying spatially indexed primers—typically glass slides or bead arrays—so that mRNA molecules released during tissue permeabilization are captured together with unique positional barcodes and unique molecular identifiers (UMIs). The captured transcripts are subsequently reverse transcribed, amplified, and sequenced, enabling array- or bead-indexed cDNA libraries for downstream position-aware quantification (Figure 2.6).

The earliest implementation of this approach was Spatial Transcriptomics (ST), introduced in 2016 [25]. In this method, tissue sections were mounted on microarray slides containing probes with poly-T tails, spatial barcodes, and UMIs. During permeabilization, cellular mRNAs diffused into defined capture spots of $\sim 100\ \mu\text{m}$, hybridized with the probes, and were reverse transcribed into cDNA. In 2019, 10x Genomics commercialized ST as Visium. Visium reduced the capture spot diameter to $55\ \mu\text{m}$ (with $\sim 100\ \mu\text{m}$ center-to-center spacing) and optimized protocols for both fresh-frozen and formalin-fixed paraffin-embedded (FFPE)

samples. Applications include large-scale tissue atlases, disease studies, and investigations of brain architecture, such as mapping gene expression across the human dorsolateral prefrontal cortex [48].

The pursuit of higher spatial resolution has driven the development of bead-based strategies. Slide-seq [49] employed randomly deposited DNA-barcoded beads ($\sim 10\ \mu\text{m}$). After sequencing the bead positions, tissues were placed onto the array, enabling near single-cell resolution mapping, although capture efficiency was limited by incomplete RNA transfer. Slide-seqV2 [50] addressed this with improved chemistry, achieving roughly an order-of-magnitude increase in sensitivity and robustness. High-Definition Spatial Transcriptomics (HDST) [51] used bead arrays with $\sim 2\ \mu\text{m}$ features, pushing toward subcellular resolution. Seq-Scope [52] further improved resolution to $0.5\text{--}0.8\ \mu\text{m}$ by generating dense high-definition map coordinate identifiers directly on sequencing chips, eliminating decoding “blind spots.” More recently, Stereo-seq [53] combined DNA nanoball (DNB) arrays with photolithography to achieve subcellular resolution across centimeter-scale tissue areas.

Despite these advances, trade-offs persist. Visium offers robust transcriptome-wide detection at moderate resolution, while ultrahigh-resolution methods such as Seq-Scope and Stereo-seq can suffer from data sparsity, noise, and increased computational burden. RNA capture efficiency remains a bottleneck, particularly for high-resolution bead-based methods. Moreover, aligning transcriptomic data with histology becomes increasingly complex at subcellular scales.

2.3 Development trends and key challenges in spatial transcriptomics

2.3.1 Emerging trends

The past decade has witnessed progress directed toward improving molecular sensitivity, spatial resolution, tissue compatibility, and integrative capacity with other molecular modalities. Collectively, these advances have transformed ST into a versatile platform for both fundamental biology and translational medicine. A structured overview of these emerging directions is provided in Table 2.2.

Tab. 2.2.: Recent advances and emerging directions in spatial transcriptomics.

Trend	Representative Platforms / Methods	Year	Key Features	Remaining Gaps
Molecular sensitivity and capture efficiency	Stereo-seq; Seq-Scope; Slide-seqV2; <i>in situ</i> poly(A) tailing	2019–2023	Increased UMI recovery at single-cell-sized areas; chemistry and permeabilization optimizations to reduce RNA loss	Still below scRNA-seq sensitivity; incomplete capture of degraded/non-polyadenylated RNAs
Spatial resolution	Slide-seqV2; HDST; Seq-Scope; Stereo-seq; expansion microscopy	2019–2023	From multi-cell spots to subcellular (0.5–0.8 μm); expansion increases effective resolution (Visium $\sim 20 \mu\text{m}$)	Signal sparsity at ultrahigh resolution; depth and diffusion constraints
Tissue compatibility and sample universality	Visium FFPE; Xenium; CosMx; fixation protocols for FFPE/PFA	2021–2024	Enables analysis of archived tissues; thousands of genes per spot in FFPE	Performance still lags fresh-frozen; fragmentation, crosslinking, autofluorescence
Spatial multi-omics integration	SM-Omics; Visium CytAssist; CosMx; Xenium; spatial CUT&Tag; spatial ATAC-seq	2021–2023	Joint profiling of RNA, proteins, chromatin; complementary biological interpretability	High cost and technical complexity; limited standardized multi-omics datasets
Computational analysis	Standardized workflows; graph-based and deep learning models	2023–2024	Segmentation, registration, cross-modal integration, denoising, deconvolution at scale	Reproducibility and interpretability; benchmarking across platforms and tissues

One important trajectory concerns the enhancement of capture sensitivity and molecular coverage. Although current ST platforms remain less sensitive than single-cell RNA sequencing methods, gradual improvements are evident. At a single-cell-sized area of approximately $100 \mu\text{m}^2$, Stereo-seq [53] has reported ~ 1450 unique molecular identifiers (UMIs) in the mouse olfactory bulb, Seq-Scope [52] has detected about ~ 848 UMIs in the mouse liver, and Slide-seqV2 [50] has achieved ~ 550 UMIs in the mouse embryo. By contrast, single-cell RNA sequencing platforms such as 10x Chromium [54] and DNBelab C4 [55] can recover approximately 40,000 UMIs (about 6000 genes) per cell, while Smart-seq2 [56] is capable of detecting up to 8000 genes. To narrow this gap, *in situ* poly(A) tailing seeks to capture degraded or non-polyadenylated RNAs, while surface chemistry and permeabilization optimizations reduce RNA loss [57]. Secondly, the advancement of spatial resolution has progressed from multi-cell spots to subcellular levels. Slide-seqV2 [50] achieves bead sizes of approximately $10 \mu\text{m}$, HDST [51] reduces feature size to around $2 \mu\text{m}$, and Seq-Scope [52] operates with pixel centers of $0.5\text{--}0.8 \mu\text{m}$, yielding an average density of 23–27 UMIs per μm^2 . Stereo-seq [53] further advances resolution through the use of patterned DNA nanoballs with diameters of about 220 nm. In parallel, expansion microscopy physically enlarges tissue within swellable hydrogels and, when coupled with barcoded arrays, has increased the effective resolution of Visium from $55 \mu\text{m}$ to approximately $20 \mu\text{m}$ while also improving RNA yield per region [58]. Theoretically, combining expansion with ultrahigh-density barcoding could approach

effective resolutions near 100 nm. The third trend lies in extending sample universality, which is essential for clinical adoption and archival studies. Protocols have been adapted for formalin-fixed paraffin-embedded (FFPE) specimens, as demonstrated by platforms such as Visium FFPE, Xenium, and CosMx, which are capable of capturing thousands of genes per spot in archived samples [59, 60]. Paraformaldehyde fixation is also under investigation as a means of preserving tissue morphology and reducing RNA diffusion. Nevertheless, fixed-tissue performance continues to lag behind that of fresh-frozen samples, underscoring the need for more robust protocols (see Section 2.3.2). A transformative direction is the integration with additional molecular readouts, which enables spatial multi-omics. This includes imaging-based RNA–protein co-detection approaches as well as sequencing-based assays of chromatin organization. These developments enrich interpretability and are discussed in detail in Section 2.4. A structured overview of these advances is provided in Table 2.2.

Concurrently, computational analysis has matured. High-resolution and multimodal datasets necessitate algorithms for cell segmentation, spatial registration, cross-modal integration, denoising, imputation, and deconvolution. Graph-based and deep learning models increasingly capture spatial dependencies, while standardized workflows enable comparative analyses across platforms and tissue types [61, 62].

2.3.2 Limitations and challenges

Despite rapid progress, spatial transcriptomics technologies still face fundamental challenges that limit their broader applicability and biological interpretability. Key obstacles and potential strategies for overcoming them are summarized in Table 2.3. A central issue is insufficient molecular capture relative to single-cell RNA sequencing [63, 2]. The reduced sensitivity lowers the ability to detect low-abundance transcripts and rare cellular states and increases dependence on RNA integrity, which is often compromised in clinical samples. Although strategies such as *in situ* poly(A) [57] tailing and optimized permeabilization protocols can extend the capture of degraded or non-polyadenylated RNAs, overall efficiency remains a major bottleneck for downstream analysis. This limitation directly affects applications such as cell type classification, trajectory inference, and the detection of rare populations. Another challenge lies in spatial resolution and its inherent trade-offs with throughput and sensitivity. Ultrahigh-resolution platforms such as Seq-Scope and Stereo-seq can resolve features at the submicron scale [52, 64], but these approaches require substantial sequencing depth and typically generate sparse count matrices with many undetected transcripts per pixel. By contrast, array-based methods such

as Visium yield more robust transcript capture at the expense of coarser resolution [65]. Expansion microscopy offers improved effective resolution by physically enlarging tissues within hydrogels, yet this approach introduces tissue distortion, workflow complexity, and alignment difficulties. Striking a balance among resolution, capture sensitivity, and scalability remains an unresolved issue that shapes the choice of technology for different biological applications.

Tab. 2.3.: Current challenges and unresolved issues in spatial transcriptomics.

Challenge Category	Example Platforms / Context	Main Issue	Impact	Possible Strategies
Insufficient molecular capture	Array- and bead-based ST (e.g., Visium, Slide-seqV2)	Lower sensitivity compared with scRNA-seq; reliance on RNA integrity	Missed low-abundance transcripts and rare states; reduced downstream power	In situ poly(A) tailing; optimized permeabilization and chemistry; improved capture probes
Resolution–throughput trade-off	Seq-Scope, Stereo-seq (submicron); Visium (coarser)	Ultrahigh resolution requires high depth and yields sparse matrices; coarser arrays lose single-cell detail	Affects segmentation, clustering, deconvolution; scaling costs	Hybrid designs; expansion microscopy; adaptive sampling; multi-scale modeling
Tissue universality (FFPE/PFA)	FFPE/PFA workflows; Xenium; CosMx; Visium FFPE	Fragmentation, crosslinking, autofluorescence; diffusion artifacts	Lower capture and image quality; limits clinical translation	Robust extraction and decrosslinking; improved probe chemistry; standardized QC for archived tissues
Multimodal integration barriers	RNA–protein (SM-Omics, CytAssist); chromatin assays (CUT&Tag, ATAC-seq)	Iterative cycles cause damage and signal loss; fluorophore crosstalk; assay trade-offs	Limited analyte breadth vs. resolution/tissue integrity; alignment across large areas	Fewer-cycle chemistries; spectral unmixing; joint assay co-design; spatial registration standards
Computational scalability and reproducibility	High-resolution and multimodal datasets	Storage and compute bottlenecks; sparsity; nonstandard formats and pipelines	Hinders comparability and clinical translation; model interpretability concerns	Community benchmarks; uncertainty-aware models; efficient graph/deep pipelines

Tissue compatibility further constrains the universality of current protocols. While fresh-frozen samples are generally well-suited for existing workflows, formalin-fixed paraffin-embedded and paraformaldehyde-fixed tissues remain problematic due to RNA fragmentation, chemical crosslinking, and autofluorescence. These factors collectively reduce capture efficiency and complicate downstream imaging. Although commercial protocols for archived tissues have improved feasibility, their performance still falls short of fresh-frozen standards, which limits translational potential given the predominance of FFPE repositories in clinical pathology [1, 66]. The integration of multiple molecular modalities within the same section, although highly promising, is technically demanding. Iterative cycles of staining, imaging, and sequencing can lead to tissue damage and signal loss, while fluorophore crosstalk and probe chemistry limit the number of analytes that can be assayed simultaneously. Consistently aligning transcriptomic, proteomic, and epigenomic

signals across large tissue areas remains challenging, and current implementations often require compromises between breadth of analytes, resolution, and tissue integrity [67].

Finally, computational challenges have become increasingly prominent as data complexity escalates. High-resolution datasets generate enormous volumes of information, creating bottlenecks in storage, preprocessing, and integration. Sparse capture at subcellular scales complicates segmentation, clustering, deconvolution, and cell–cell communication analysis. In addition, nonstandardized formats and pipelines hinder reproducibility and cross-study comparability. Although graph-based and deep learning frameworks are emerging to address these challenges, further progress is needed to ensure robustness, scalability, and interpretability in real-world applications. The widespread adoption of spatial transcriptomics is constrained by limitations in molecular capture, resolution–throughput trade-offs, restricted tissue universality, technical barriers to multimodal integration, and the growing computational burden [68]. Overcoming these obstacles through improved chemistry, standardized protocols, and scalable analytical frameworks will be critical for translating spatial data into reliable biological insight.

2.4 Spatial multi-omics integration

The complexity of biological systems cannot be fully captured by transcriptomics alone. Cellular identity and function arise from multiple molecular processes, including transcription, translation, chromatin regulation, and signaling. Consequently, integrating spatial transcriptomics (ST) with additional molecular readouts—spatial multi-omics—has emerged as a critical frontier for understanding tissue organization and disease mechanisms. A summary of these representative technologies is provided in Table 2.4.

Tab. 2.4.: Spatial multi-omics integration technologies.

Category	Modalities	Representative Platforms / Methods	Year	Resolution	Key Features	Limitations
Transcriptome Morphology / Protein context	+RNA + H&E / IF	Histology-guided ST overlays	2018–2019	Tissue / cell level	Provides morphological and protein context for transcriptomes	Limited multiplexing; not simultaneous capture
RNA + Protein (co-detection)	RNA + Protein	10x Genomics Visium CytAssist	2022	Spot level	Co-detection of RNA and protein on the same slide using antibody assays	High cost; technically demanding; limited datasets
Sequencing-based multi-omics	RNA + Protein	SM-Omics (Ståhl Lab, <i>Sci Adv</i>)	2021	Spot level	Extends ST to proteomics via antibody-derived tags	Antibody-dependent; moderate resolution

Continued on next page

Category	Modalities	Representative Platforms / Methods	Year	Resolution	Key Features	Limitations
Imaging-based multi-omics	RNA + Protein	NanoString CosMx SMI; 10x Genomics Xenium	2022	Subcellular	High-plex imaging: thousands of RNAs with dozens of proteins	Signal loss from iterative cycles; fluorophore crosstalk
Spatial epigenomics	RNA + Chromatin	Spatial CUT&Tag; Spatial ATAC-seq	2021–2022	Spot / cell level	Maps chromatin occupancy or accessibility alongside RNA	Early-stage; limited datasets; lack of standardization
Functional perturbation	RNA + Perturbation	Perturb-map (<i>Science</i>)	2021	Cell level	CRISPR-based perturbation mapping across microenvironments	Complex protocols; limited scalability
Lineage tracing	RNA + Clonal barcodes	Space-TREX (<i>Cell</i>)	2023	Cell level	Spatially resolved lineage tracing with barcoding	Requires sophisticated barcoding; still emerging
Translation-level profiling	RNA + Active translation	RIBOmap (<i>Nat Methods</i>)	2022	Spot / cell level	Captures spatial distribution of active protein synthesis	Nascent approach; few datasets available

Early efforts combined RNA measurements with histological and immunofluorescence imaging to contextualize transcriptomes within morphological and protein-level features. More advanced approaches enabled the simultaneous detection of RNA and protein in the same section. For example, sequencing-based methods such as SM-Omics [69] extend ST to proteomics using antibody-derived tags, while the Visium CytAssist [70] platform co-detects RNA and protein abundance on the same slide through antibody-based workflows, though with high cost and technical complexity. Imaging-based platforms, including CosMx SMI [71] and Xenium [72], allow subcellular profiling of thousands of RNA species together with dozens of proteins. These technologies provide complementary perspectives: transcriptomes reveal regulatory states, whereas proteins more directly reflect cellular phenotypes and functions.

Beyond RNA–protein co-detection, spatial multi-omics now includes epigenomic modalities. Extensions such as spatial CUT&Tag [73] and spatial ATAC-seq [74, 75] provide chromatin occupancy and accessibility information within tissue context, linking regulatory landscapes to spatially resolved transcriptional states. Other innovations combine ST with functional perturbations or lineage tracing: CRISPR-based Perturb-map [76] quantifies perturbation effects across microenvironments, while Space-TREX [77, 78] enables spatially resolved clonal dynamics. At the translation level, RIBOmap [79] captures the spatial distribution of active protein synthesis, extending multi-omics profiling beyond transcription and chromatin regulation.

Despite these advances, substantial challenges remain. Technical limitations include signal loss and tissue damage from iterative staining or sequencing cycles, fluorophore crosstalk

that constrains multiplexing, and probe-efficiency variability that complicates cross-modal comparisons. On the computational side, integrative frameworks are still evolving, with ongoing work to harmonize data formats, standardize pipelines, and develop statistical models for the joint analysis of RNA, protein, and epigenomic signals. Graph-based approaches and deep learning are promising but raise concerns about reproducibility and interpretability.

Looking forward, spatial multi-omics will play a central role in systems biology. By enabling simultaneous characterization of transcriptional states, protein abundances, chromatin organization, perturbation responses, lineage histories, and translation activity within intact tissue architecture, these technologies offer unprecedented opportunities to decode the molecular principles governing tissue function and disease progression. Importantly, such developments align with the demand for predictive models that link RNA profiles to protein abundances or other downstream phenotypes. A summary of these representative technologies is provided in Table 2.4. In this thesis, the second project leverages CytAssist datasets to develop and evaluate computational models for predicting protein abundance from spatial RNA profiles, illustrating how analytical approaches can extend the reach of spatial multi-omics when experimental data are costly or limited.

2.5 Statistical perspective on spatial transcriptomics data

Modern omics technologies—including bulk RNA sequencing, single-cell RNA sequencing (scRNA-seq), and spatial transcriptomics—yield datasets where the number of features (genes, etc.) far exceeds the number of observations (cells, spots, or samples). These datasets are inherently high-dimensional, sparse, and comprised of count-based measurements with heteroscedastic noise [80]. Such characteristics pose significant challenges for traditional statistical inference, often leading to issues such as overfitting, unstable parameter estimates, and difficulties in interpretation. Consequently, regularization strategies, low-rank approximations, and dimensionality reduction techniques are essential for stable and interpretable modeling [81, 82].

In scRNA-seq data, the problem of dropouts—instances where transcripts are undetected due to technical limitations—is widespread. When coupled with batch effects and variability in sequencing depth, this zero inflation distorts data distributions and undermines downstream analyses, such as clustering and differential expression. While some methods explicitly model zeros using zero-inflated distributions, in many modern UMI-based datasets

simpler count models suffice given appropriate normalization [83]. It is therefore critical to distinguish between technical zeros and genuine biological absence to ensure valid interpretation [84].

Spatial transcriptomics retains these challenges while introducing additional complexity through spatial structure. Expression measurements are collected at defined tissue locations—often aggregating signals from multiple cells—leading to spatial correlation and heterogeneity [85]. Therefore, statistical models for spatial transcriptomics must account for both count characteristics and spatial dependencies. In many practical scenarios, standard models augmented with spot-level covariates adequately describe most gene expression patterns. Nevertheless, more sophisticated frameworks, such as spatial random effects or Gaussian process models, are increasingly employed to capture fine-grained spatial variations [86].

To manage the combined challenges of high dimensionality, sparsity, noise, and spatial complexity, representation learning and deep learning methodologies have become popular. These techniques enable automatic extraction of meaningful low-dimensional features, data denoising, and multimodal integration. However, they also raise concerns regarding model interpretability, robustness, and the imperative for rigorous evaluation protocols and benchmarking [83, 87].

Literature Review

This chapter reviews computational strategies in spatial transcriptomics, focusing on cell type deconvolution and inference of spatial protein abundance. It highlights key methodological frameworks, experimental foundations, and future directions toward integrative spatial omics.

3.1 Computational strategies for cell type decomposition in spatial transcriptomics

Spatial transcriptomics (ST) technologies enable the quantification of mRNA expression while preserving the spatial organization of cells within tissue sections, offering a unique perspective for studying cellular identity and microenvironmental organization. At the tissue scale, ST enables the discovery of novel transcriptional patterns and the reconstruction of tissue-wide gene expression landscapes [88]. At a finer scale, it facilitates the investigation of cell–cell interactions and the relationship between intracellular states and extracellular environments, thereby advancing the understanding of cell function and disease mechanisms. Current ST technologies can be broadly divided into two categories. Image-based approaches, including *in situ* hybridization and *in situ* sequencing methods, provide high spatial resolution at the subcellular level but remain limited by the number of profiled genes, sensitivity, and technical complexity [89]. Sequencing-based approaches, in contrast, capture barcoded mRNA with nongene-specific probes, enabling transcriptome-wide profiling across diverse tissue types with greater scalability and accessibility.

Despite their transformative potential, sequencing-based ST technologies are constrained by spot-level resolution, where each spot often contains mixtures of multiple cells and blended cell types. This low resolution can obscure genuine transcriptional heterogeneity and distort cellular-level reconstruction of tissue architectures [90]. To address this challenge, computational strategies for cell type decomposition (also referred to as cellular deconvolution) have been developed to infer the proportion of constituent cell types within each spot. Accurate deconvolution not only refines the reconstruction of heterogeneous tissues but also facilitates downstream analyses such as cell–cell communication, trajectory inference, and disease characterization [91, 92, 93].

Over the past five years, numerous computational methods for spatial deconvolution have emerged, reflecting diverse methodological innovations. These methods can be broadly divided into two groups [94]: reference-based approaches, which rely on single-cell RNA sequencing (scRNA-seq) reference datasets, and reference-free approaches, which operate without external references. Within reference-based approaches, methodological categories include probabilistic models, nonnegative matrix factorization (NMF)-based methods, graph-based approaches, deep learning frameworks, and optimal transport formulations. Recently, regression-based methods have also been introduced, aiming to provide efficient solutions particularly for high-resolution or single-cell spatial data. Reference-free approaches, in parallel, have advanced through marker-based, latent factor, and topic-modeling frameworks, expanding applicability to contexts where no suitable scRNA-seq reference is available.

Representative methods across these categories are summarized in Table 3.1, highlighting their underlying principles, implementation frameworks, and application scenarios. Rather than benchmarking performance, this review emphasizes methodological distinctions, discusses advantages and limitations, and outlines emerging trends for next-generation deconvolution algorithms.

Tab. 3.1.: Overview of spatial transcriptomics deconvolution methods.

Method	Year	Category	Reference type	Implementation	Notes
stereoscope	2020	Probabilistic	With scRNA-seq	Python	Bayesian hierarchical model; robust but scalability to large datasets is limited
RCTD	2020	Probabilistic	With scRNA-seq	R	Provides high accuracy; sensitive to reference bias and requires careful normalization
DestVI	2021	Probabilistic	With scRNA-seq	Python	Captures continuous cell states; useful for complex tissues but computationally intensive
STRIDE	2021	Probabilistic	With scRNA-seq	Python	Flexible application to diverse data; adoption remains limited in practice
CARD	2021	Probabilistic	With scRNA-seq	R	Incorporates spatial correlation; enhances biological plausibility but increases complexity
SpatialDecon	2021	Probabilistic	With scRNA-seq	R	Designed for immune-rich tissues; robust to noise and widely applied in immunology studies
cell2location	2021	Probabilistic	With scRNA-seq	Python	High accuracy in complex tissues; computationally demanding and GPU support recommended
SPOTlight	2020	NMF-based	With scRNA-seq	R	Uses marker gene selection and NMF; interpretable but depends on curated marker panels

Continued on next page

Method	Year	Category	Reference type	Implementation	Notes
NMFReg	2021	NMF-based	With scRNA-seq	R	Combines regression and NMF; balances accuracy and interpretability
SpatialDWLS	2021	NMF-based	With scRNA-seq	R	Incorporates DWLS weighting; scales well to large datasets with improved stability
DSTG	2020	Graph-based	With scRNA-seq	Python	Leverages cell graph structures; enhances accuracy in heterogeneous microenvironments
SD ²	2022	Graph-based	With scRNA-seq	Python	Integrates scRNA-seq and ST graphs; improves resolution but requires complex preprocessing
GraphST	2023	Graph-based	With scRNA-seq	Python	Employs graph attention with spatial context; balances accuracy and scalability
STdGCN	2024	Graph-based	With scRNA-seq	Python	Deep graph convolutional framework; state-of-the-art spatial modeling performance
Tangram	2021	Deep learning	With scRNA-seq	Python	Aligns scRNA-seq with ST for near single-cell resolution; effective but resource-heavy
SpaOTsc	2020	Optimal transport	With scRNA-seq	Python	Implements theoretical OT; interpretable but sensitive to gene dropout
NovoSpaRc	2021	Optimal transport	With scRNA-seq	Python	Handles missing genes flexibly; suitable for developmental or perturbed systems
Berglund et al.	2020	Reference-free	No scRNA-seq	R	Early marker-based method; simple but limited resolution
SpiceMix	2021	Reference-free	No scRNA-seq	Python	Latent factor model; interpretable with moderate scalability
STdeconvolve	2022	Reference-free	No scRNA-seq	R	Uses topic modeling; effective for unsupervised exploration of cell types
SMART	2024	Reference-free	No scRNA-seq	Python	Strong performance across benchmarks; generalizable to diverse tissues
STsisal	2025	Reference-free	No scRNA-seq	Python	Scalable framework; actively developed with focus on large-scale datasets
Redeconve	2023	Regression-based	Single-cell resolution	R	Tailored for high-resolution ST; emphasizes cell-level precision
NLSDeconv	2025	Regression-based	With scRNA-seq	Python	Lightweight regression model; efficient and suitable for large datasets

3.1.1 General workflow of deconvolution approaches

The general workflow of deconvolution approaches in ST follows a structured pipeline that integrates spatially resolved transcriptomic data with computational models to infer cellular composition. As illustrated in Figure 3.1, the process begins with the acquisition of two primary data types: ST data, consisting of gene expression profiles linked to

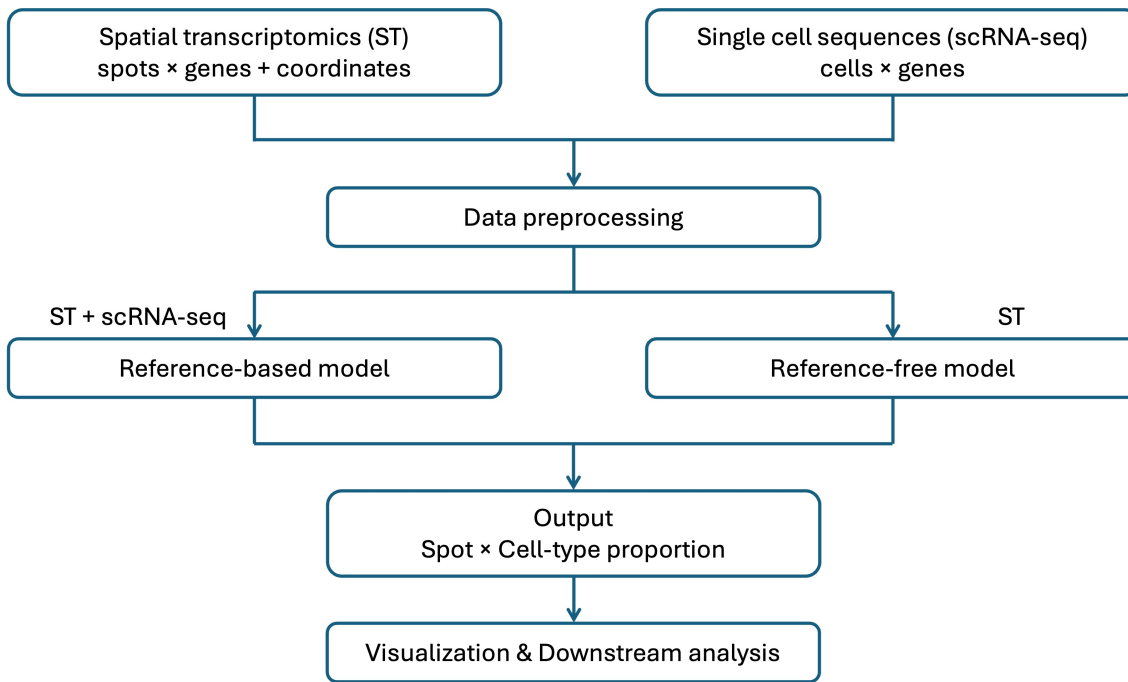


Fig. 3.1.: General workflow for spatial transcriptomics deconvolution. ST data and optionally scRNA-seq reference datasets undergo preprocessing (quality control, normalization, batch correction, and feature selection). Depending on the availability of references, deconvolution is performed using reference-based or reference-free models, resulting in estimated cell type proportions across spatial spots that enable downstream analyses.

spatial coordinates at the spot level, and optionally, scRNA-seq data, which provides high-resolution cell-type specific transcriptomic profiles.

Following data collection, preprocessing ensures data quality and compatibility across modalities [95, 23]. Typical steps include:

- **Quality control:** Removal of low-quality spots or cells, filtering of genes with very low expression, and exclusion of technical artifacts.
- **Normalization:** Transformation of raw counts into comparable expression levels across samples and platforms (e.g., library size normalization, log-transformation).
- **Batch correction:** Adjustment for technical or biological batch effects between ST sections or between ST and scRNA-seq datasets (e.g., Harmony, Seurat integration).

- **Feature selection:** Identification of informative genes, such as highly variable genes (HVGs), marker genes, or genes overlapping between ST and scRNA-seq datasets, to reduce noise and computational burden.
- **Dimensionality reduction (optional):** Techniques such as PCA can project expression data into lower dimensions to stabilize downstream modeling.

Depending on the availability of scRNA-seq references, deconvolution models fall into two broad categories [94]:

Reference-based models utilize scRNA-seq datasets as transcriptomic references. These models map spot-level ST profiles to annotated single-cell profiles to estimate cell type proportions. Their advantage lies in leveraging well-characterized cell-type signatures; however, performance depends heavily on the quality and biological relevance of the reference dataset.

Reference-free models, in contrast, rely solely on ST data to infer latent cellular components. Typically, they employ unsupervised approaches such as matrix factorization or probabilistic modeling to decompose spot-level expression into putative cell-type like profiles. While reference-free approaches avoid reference biases, interpretability can be limited and validation often requires external biological knowledge.

The primary output of both strategies is a matrix representing the estimated cell type proportions across spatial spots. This quantitative cell type map refines tissue heterogeneity and serves as a foundation for downstream analyses, including visualization of spatial organization, inference of cell–cell communication networks, reconstruction of developmental or disease-related trajectories, and characterization of pathological microenvironments [96].

Overall, this workflow underscores the central role of deconvolution in bridging spot-level ST with cellular-level insights, enabling a more accurate understanding of tissue architecture and function [97]. The following sections introduce representative deconvolution methods in detail.

3.1.2 Reference-based deconvolution models

Reference-based models represent the mainstream computational strategies for ST [98]. These approaches leverage scRNA-seq datasets as high-resolution references to infer the cellular composition of spatial spots. The core principle is to express each spot-level transcriptomic profile as a weighted mixture of cell-type specific signatures derived from scRNA-seq, with the weights corresponding to cell type proportions. By exploiting single-cell resolution, these models decompose bulk-like spot expression into biologically meaningful components, thereby refining tissue-level reconstruction. Their accuracy, however, depends critically on the quality, depth, and representativeness of the reference dataset.

Probabilistic models were among the earliest and remain widely adopted. Stereoscope [99] applies a generative Bayesian model in which each spot is a probabilistic mixture of reference-derived cell-type distributions, offering robustness but limited scalability. RCTD [100] employs a regression-based probabilistic framework that achieves high accuracy across diverse tissues, though it is sensitive to mismatched or incomplete references. DestVI [101] extends these ideas with a variational autoencoder, uniquely capturing continuous variation within cell types and accommodating intra-cell-type heterogeneity. CARD [102] incorporates spatial priors to exploit neighborhood information, yielding smoother deconvolution maps. SpatialDecon [103] is tailored to immune-oncology, proving useful in dissecting tumor microenvironments. cell2location [104] employs a hierarchical Bayesian framework for joint estimation across multiple tissue sections, achieving high-resolution maps but with substantial computational demands often requiring GPU acceleration. Collectively, probabilistic models demonstrate strong accuracy and versatility but are computationally intensive and reliant on suitable references.

NMF-based methods provide an alternative, approximating spot-level profiles as nonnegative linear combinations of cell-type signatures. SPOTlight [105] combines NMF with marker gene regression, producing interpretable results where each component corresponds to a reference-derived cell type. Its reliance on curated marker sets, however, can introduce subjectivity. NMFReg [49] integrates regression with NMF to improve stability, while SpatialDWLS [106] adapts dampened weighted least squares (DWLS) [107] for improved scalability. These approaches are computationally efficient and accessible, but they may oversimplify cellular variability and are less effective at capturing rare or transitional populations compared to probabilistic models.

Graph-based methods capture structural information inherent in both scRNA-seq and ST data. DSTG [108] introduced a graph neural network (GNN) that models similarity between spatial spots and reference single cells for semi-supervised learning. SD² [109] integrates graphs constructed separately from scRNA-seq and ST to improve robustness across modalities. GraphST [61] employs graph attention mechanisms to adaptively weigh neighborhood information, enhancing modeling of local context. The more recent STdGCN [62] incorporates spatial constraints into a deep GCN, providing state-of-the-art performance but at increased computational cost. Graph-based approaches are particularly valuable when tissue architecture is critical, though they remain sensitive to graph construction and parameterization.

Deep learning frameworks have also been applied, with Tangram [110] being a prominent example. Tangram formulates mapping between scRNA-seq and ST data as an optimization problem, assigning single-cell profiles to spatial coordinates to best reconstruct observed spot-level expression. This strategy achieves near single-cell resolution and can integrate multimodal data, such as imaging or proteomics. However, Tangram requires substantial computational resources and is prone to overfitting if the reference does not closely match the target tissue.

Optimal transport (OT) formulations provide a mathematically principled alternative. SpaOTsc [111] and NovoSpaRc [112] align scRNA-seq and ST distributions by minimizing transport costs under biological constraints. SpaOTsc emphasizes single-cell resolution alignment, whereas NovoSpaRc flexibly handles missing genes and partial overlaps. While elegant in theory, OT methods are computationally demanding and sensitive to parameter tuning, limiting their utility for very large datasets.

In summary, reference-based methods have substantially advanced the resolution and interpretability of ST. Probabilistic models dominate current applications due to their accuracy and flexibility, NMF-based methods emphasize efficiency and interpretability, graph-based and deep learning approaches highlight the integration of spatial context and multimodal data, and OT frameworks offer theoretical rigor. Their limitations—dependency on reference quality, sensitivity to batch effects, and computational burden—motivate future developments in scalability, multimodal integration, and more robust use of multiple references.

3.1.3 Reference-free deconvolution models

Reference-free models rely solely on ST data without requiring scRNA-seq references [113]. These methods aim to infer latent cellular components directly from spot-level expression, decomposing observed profiles into a set of underlying signatures and proportions. Their main motivation is independence from external references, which may be unavailable, incomplete, or biologically mismatched. However, validation is challenging, and interpretability often requires post hoc annotation with marker genes.

Early work by Berglund et al. [114] demonstrated a marker-based approach, using predefined marker sets to assign cell type composition without references. While conceptually simple, this method is limited by the need for curated marker lists and cannot easily capture novel or subtle states.

Subsequent approaches introduced statistical frameworks. SpiceMix [115] employs a latent factor model, representing spot-level expression as mixtures of hidden components with spatial regularization to improve biological plausibility. STdeconvolve [116] extends this idea with a topic-modeling approach based on latent Dirichlet allocation, treating spots as mixtures of topics that correspond to latent programs. These methods enable unsupervised discovery of cell-type like profiles, though interpretation still requires external knowledge.

Recent developments have improved scalability and accuracy. SMART [117] integrates matrix factorization with spatial regularization, achieving accuracy comparable to some reference-based methods. STsisal [118] introduces an efficient algorithm capable of decomposing tens of thousands of spots, enabling application to large tissues and clinical samples.

Overall, reference-free models provide flexibility, independence from external bias, and the capacity to reveal novel or tissue-specific populations absent from reference datasets. Remaining challenges include interpretability, hyperparameter sensitivity, and lack of direct ground truth. Promising directions include combining reference-free models with marker-based annotation, integrating multimodal data, and developing hybrid frameworks that flexibly incorporate reference information when available.

3.1.4 Emerging regression-based approaches

In addition to probabilistic, factorization, graph-based, and reference-free frameworks, a new category of regression-based approaches has emerged for ST deconvolution. These methods directly model the relationship between spot-level expression and candidate cell-type signatures using regression formulations, aiming for a balance between interpretability, computational efficiency, and accuracy. By exploiting the linearity of mixtures, regression-based models provide simple yet effective strategies that are attractive for large-scale and practical applications.

Redeconve [119] illustrates this trend, designed for high-resolution ST platforms approaching single-cell granularity. Unlike most methods assuming mixed spots, Redeconve applies a regression framework directly linking local expression to cell type abundances. Its main strengths are conceptual simplicity and computational tractability, though its reliance on high-resolution input limits broader applicability.

NLSDeconv [120] employs nonlinear least-squares regression with regularization, offering a lightweight and efficient solution for reference-based deconvolution. It achieves competitive accuracy while maintaining scalability, making it suitable for large datasets and resource-constrained settings.

In conclusion, these regression-based approaches highlight renewed interest in accessible, efficient, and interpretable deconvolution models. Their limitations include reliance on linear assumptions, reduced flexibility in modeling complex states, and limited validation across diverse contexts. Nevertheless, they represent promising complements to more complex frameworks, particularly for clinical and translational applications.

3.1.5 Comparative evaluation and future perspectives

The current landscape of ST deconvolution demonstrates a diversity of strategies, each with distinct strengths and limitations. Reference-based models remain the most widely applied, leveraging scRNA-seq references to infer cell type proportions with high accuracy and interpretability. They are particularly effective when comprehensive and context-matched references are available, enabling fine-grained reconstruction and downstream analyses such as cell–cell communication and microenvironmental characterization. Their limitations include dependence on reference quality, susceptibility to batch effects, and computational burden in large datasets.

Reference-free models offer an alternative when suitable references are unavailable. By relying solely on ST data, they avoid reference bias and enable discovery of novel or tissue-specific components. Their data-driven flexibility makes them valuable for exploratory analysis, yet interpretability and validation remain significant challenges.

Regression-based models emphasize simplicity and computational efficiency. By framing deconvolution as a regression problem, they are transparent, scalable, and accessible, providing practical solutions for large-scale or clinical applications. Their reliance on linear assumptions, however, constrains their flexibility in capturing complex transcriptional programs.

Future directions include hybrid models that integrate reference-based and reference-free strategies, adaptive use of reference information, and multimodal integration (e.g., spatial proteomics or imaging). Advances in scalability and computational efficiency will be critical for clinical translation, while rigorous benchmarking across tissues and platforms remains essential for establishing best practices.

In summary, deconvolution methods have made substantial progress in bridging the gap between spot-level ST and cellular-level resolution. Reference-based models dominate current practice, reference-free approaches provide unique opportunities for discovery, and regression-based methods emphasize accessibility and efficiency. Together, these complementary strategies advance the development of more comprehensive, interpretable, and practical frameworks for dissecting cellular heterogeneity in spatial transcriptomics.

3.2 Inferring spatial protein abundance from transcriptomic profiles

In multicellular systems, cellular behavior is profoundly influenced by spatial context, where local microenvironments determine how cells interact, differentiate, and respond to perturbations. Spatial transcriptomics (ST) has provided transformative insights by enabling high-throughput mRNA profiling while preserving tissue architecture. These advances have facilitated the discovery of gene expression patterns, the identification of spatially distinct cellular niches, and the reconstruction of tissue-level organization. Nevertheless, transcriptomic measurements alone are insufficient to capture the full spectrum of cellular functionality, as proteins represent the primary effectors of biological processes [121, 122].

Proteins play essential roles in regulating signaling cascades, mediating metabolic pathways, and maintaining structural integrity, thereby directly linking molecular states to cellular phenotypes [123, 124]. A growing body of evidence demonstrates that mRNA abundance only partially explains protein expression, with weak correlation observed at the single-cell level for many genes. This discordance arises from multiple biological and technical factors, including post-transcriptional regulation, variable translational efficiency, protein turnover, and stochastic noise in RNA processing and measurement. Moreover, temporal delays between transcription and translation further contribute to discrepancies between RNA and protein abundance [125]. As a result, mRNA readouts often provide an incomplete representation of the proteomic landscape within tissues.

Although several experimental technologies have been developed to directly measure spatial protein expression, they are constrained by substantial challenges [126, 127]. Imaging-based multiplexed assays are typically limited by antibody availability and panel size, while sequencing-based multi-omics methods remain technically demanding. Both categories of approaches are often associated with high costs and limited scalability, restricting their application in large-scale or routine studies. Consequently, the majority of spatial studies continue to rely exclusively on RNA measurements. Given these limitations, inferring spatial protein abundance from transcriptomic profiles has emerged as a critical research direction. Computational frameworks capable of predicting protein distributions from ST data offer the potential to complement experimental assays, extend the utility of transcript-centric datasets, and generate a more comprehensive view of tissue physiology. By bridging RNA and protein modalities, such approaches can facilitate deeper insights into cell–cell communication, disease mechanisms, and therapeutic response within spatially organized biological systems.

3.2.1 Experimental platforms enabling paired RNA–protein measurements

A range of experimental technologies has been established to measure RNA and proteins simultaneously in intact tissues, providing essential multimodal information for understanding cell states in spatial contexts. These methods differ in their underlying principles, resolution, scalability, and accessibility, and they collectively form the experimental foundation for computational inference of protein abundance from transcriptomic data. A comparative overview is presented in Table 3.2.

Tab. 3.2.: Overview of experimental platforms enabling paired RNA–protein spatial profiling.

Method	Year	Category	Principle	Notes
CODEX	2019	Imaging-based	Cyclic antibody staining with DNA barcodes	High multiplexing; subcellular resolution; limited by antibody panel size and high technical complexity
MIBI	2020	Imaging-based	Secondary ion mass spectrometry with isotope-tagged antibodies	High-resolution proteomic imaging; expensive instrumentation; limited throughput
CosMx	2022	Imaging-based	Multiplexed fluorescence in situ hybridization with antibody conjugates	Single-cell resolution; commercial availability; restricted panel size and high cost
SPOTS	2020	Sequencing-based	Oligo-tagged antibodies combined with spatial barcoding for RNA and proteins	Joint RNA–protein sequencing; requires custom microfluidics; limited adoption
SM-Omics	2020	Sequencing-based	Integration of transcriptome sequencing and antibody-derived tags	Flexible multi-omic profiling; constrained by antibody availability; technically demanding
Spatial-CITE-seq	2021	Sequencing-based	Spatial barcoding of RNA and antibody-derived tags in tissue	Extends CITE-seq to spatial domain; scalable in principle; limited accessibility
Stereo-CITE-seq	2023	Sequencing-based	High-density spatial barcoding chip capturing RNA and antibody tags	High spatial resolution; large field of view; technically demanding
Visium CytAssist	2022	Commercial platform	Tissue transfer onto Visium slides with whole-transcriptome RNA + targeted antibody–oligo panels	Standardized workflow; scalable and widely accessible; limited protein panel and coarse spot resolution

3.2.1.1 Imaging-based multiplexed spatial proteomics

Imaging-based approaches achieve direct protein detection through antibody labeling combined with high-resolution microscopy or ion-based readout, offering single-cell to subcellular resolution of tissue architecture. Among these, CODEX (CO-Detection by indEXing) [128, 129] employs DNA-barcoded antibodies in iterative cycles of hybridization, imaging, and stripping, enabling detection of dozens to more than one hundred proteins within the same section. Similarly, MIBI (Multiplexed Ion Beam Imaging) [130, 131] conjugates antibodies to isotopically pure metal reporters and visualizes them via secondary ion mass spectrometry, thereby circumventing the spectral overlap inherent to fluorescence-based techniques and achieving high multiplexing capacity with spatial resolution approaching 200 nm. These methods collectively provide unparalleled resolution and the ability to interrogate spatial proteomes with great detail. However, their utility is constrained by the dependence on validated antibody panels, potential artifacts from non-specific binding, and substantial technical and cost barriers. CODEX requires iterative imaging cycles and careful panel design, while MIBI demands highly specialized instrumentation with limited throughput. As a result, despite their power, imaging-based multiplexed proteomics remains most suitable for focused studies rather than large-scale or routine applications.

3.2.1.2 Sequencing-based spatial multi-omics

In parallel, sequencing-based methods extend CITE-seq principles into spatial contexts by coupling transcriptome profiling with antibody-derived oligonucleotide tags (ADTs). SPOTS (Spatial Protein and Transcriptome Sequencing) [132] pioneered this direction by integrating barcoded antibodies into spatial transcriptomic workflows, thereby enabling concurrent RNA and protein detection in fixed tissue sections. Building on this, SM-Omics [69] introduced refined library preparation strategies that improved capture efficiency and enhanced compatibility with multiple molecular layers.

Subsequent developments have further increased resolution and scalability. Spatial-CITE-seq [133] employed spatial barcodes to assign RNA and protein reads to defined tissue regions, thus markedly improving multiplexing capacity compared to imaging-based platforms. Stereo-CITE-seq [134] advanced this further through the use of high-density barcoding chips, achieving finer spatial resolution and a broader field of view, thereby approaching single-cell granularity across large tissue areas. Despite these advantages, sequencing-based approaches require complex microfluidic systems or custom chip designs, remain dependent on antibody panel availability, and are largely confined to specialized laboratories with the requisite technical expertise.

3.2.1.3 Commercial platforms for integrated RNA–protein profiling

Commercial solutions have sought to balance accessibility with performance. CosMx [71], developed by NanoString, provides highly multiplexed single-cell spatial profiling of both RNA and proteins through fluorescence imaging of oligonucleotide-conjugated antibodies. It offers high resolution and seamless integration with histopathological analysis, yet panel sizes remain restricted and operational costs are considerable. More broadly, imaging-based commercial platforms excel in resolution and visualization but face challenges in scalability and affordability.

In contrast, 10x Genomics Visium CytAssist [70] has emerged as the most widely adopted commercial solution for integrated RNA–protein profiling. CytAssist enables tissue transfer onto Visium slides, which allow unbiased whole-transcriptome capture in conjunction with antibody–oligonucleotide conjugates targeting selected proteins. The platform provides a standardized workflow, reproducibility, and an expanding repository of publicly available paired RNA–protein datasets. Nonetheless, its resolution is limited to spot level—each spot typically encompassing multiple cells—and protein panel sizes remain relatively small

compared to the breadth of the proteome. Despite these limitations, CytAssist currently represents the most accessible and scalable option, and its datasets have become critical benchmarks for computational method development.

3.2.2 Computational Strategies for Protein Inference

Although experimental technologies have enabled simultaneous measurement of RNA and proteins in situ, they remain costly, technically demanding, and limited in throughput. As a result, computational methods that infer protein abundance from transcriptomic profiles have gained increasing attention. These approaches can be grouped into three categories: image-based virtual staining methods, single-cell RNA–protein prediction models, and spatial transcriptomics-specific inference frameworks. Each category reflects a distinct strategy for leveraging available data to approximate proteomic information.

3.2.2.1 Image-Based Virtual Staining Approaches

In the imaging domain, deep learning models have been employed to computationally predict protein marker distributions from hematoxylin and eosin (H&E) stained tissue sections. These methods are conceptually analogous to image-to-image translation, where H&E morphology serves as input and protein stains are the output. Early work focused on predicting single protein markers, while more recent frameworks, such as Multi-V-Stain [135], have demonstrated the ability to predict multiple markers jointly, thereby preserving spatial coherence and improving biological relevance. These approaches highlight the feasibility of replacing expensive immunohistochemistry or immunofluorescence assays with virtual staining, potentially enabling large-scale proteomic mapping from routine histology slides. However, such methods are inherently morphology-driven and cannot exploit transcriptomic signals. Consequently, they are not directly applicable to ST data, where RNA information provides complementary insights into cellular states and regulatory processes.

3.2.2.2 Single-Cell RNA–Protein Prediction Models

The advent of multimodal single-cell technologies, such as CITE-seq and REAP-seq, has stimulated the development of computational approaches that learn systematic relationships between transcriptomic and proteomic layers. A central goal of these models is to impute protein abundance from RNA profiles, thereby extending the utility of scRNA-seq datasets. Several strategies have emerged to address this task. Neural network–based methods,

exemplified by cTP-net [136], achieve strong predictive performance by directly mapping high-dimensional transcriptomes to corresponding protein levels. Complementary to this, gene set testing approaches such as STREAK [137] emphasize interpretability by linking receptor abundance to biologically coherent gene modules, while embedding-based frameworks like sciPENN [138] introduce latent protein representations that facilitate flexible downstream applications. In parallel, regression and ensemble-based strategies, including scLinear [139] and SPIDER [140], provide computationally efficient alternatives that balance scalability with competitive accuracy.

These methods illustrate the feasibility of predicting protein expression from single-cell transcriptomes and have highlighted the extent to which post-transcriptional regulation modulates RNA–protein concordance. Importantly, however, they remain intrinsically cell-centric, neglecting the spatial context that shapes protein abundance in tissues. As such, while these models represent important methodological advances in single-cell analysis, their direct applicability to spatial transcriptomic data is limited without explicit incorporation of microenvironmental information.

3.2.2.3 Spatial Transcriptomics-Oriented Inference Frameworks

To overcome the limitations of cell-centric approaches, computational methods tailored specifically to spatial transcriptomics have recently emerged. Among them, the Dual-Graph Attention Network (DGAT) [141] represents an important proof of concept. DGAT constructs two complementary graphs: a molecular graph based on transcriptomic similarity across spots, and a spatial graph encoding tissue topology and neighborhood relationships. Through dual attention mechanisms, DGAT adaptively weights contributions from both graphs, allowing simultaneous propagation of molecular and spatial information. This design enables the model to predict protein distributions that are not only transcriptomically informed but also spatially coherent.

Benchmarking studies have shown that DGAT outperforms conventional baselines, including linear regression models, multilayer perceptrons, and standard graph neural networks, both in terms of prediction accuracy and in capturing biologically relevant spatial patterns. Despite these advances, current ST-specific frameworks face several challenges. First, their evaluation has been limited to a small number of publicly available paired RNA–protein ST datasets, raising concerns about their generalizability. Second, reliance on deep learning architectures introduces issues of interpretability, making it difficult to disentangle biological drivers from model artifacts. Third, the scarcity of large-scale multi-omic training

data restricts the ability of such methods to generalize across tissue types and experimental conditions.

Nonetheless, these computational strategies underscore the potential of integrative modeling to augment transcriptomic datasets with proteomic information. They highlight the importance of combining molecular signals with spatial context to achieve biologically meaningful predictions of protein abundance, while also illustrating the methodological and data-related challenges that remain to be addressed.

3.2.3 Current Challenges and Future Directions

Although considerable progress has been made in linking transcriptomic and proteomic information within spatial contexts, several persistent challenges continue to constrain both experimental and computational advances. A major limitation is the scarcity of large, high-quality paired RNA–protein spatial datasets. Existing resources typically encompass a limited number of proteins, are generated from restricted tissue types, and often vary in protocols, antibody panels, or preprocessing pipelines, complicating reproducibility and cross-study benchmarking. Expanding the availability of standardized multimodal datasets will therefore be essential to enable robust evaluation and improve the generalizability of computational frameworks.

Experimental platforms themselves also present significant barriers. Despite differences in design, most current methods share common limitations of high cost, technical complexity, and restricted antibody coverage, which together hinder scalability and routine application. Even commercial solutions such as CosMx and CytAssist, while providing greater accessibility and standardization, remain constrained in panel size and resolution. These shared constraints underscore why computational strategies are needed to complement experimental assays, extending the reach of transcript-centric datasets toward proteomic insights.

Beyond technical limitations, the biological complexity of RNA–protein relationships further complicates inference. Weak correlation between transcript and protein abundance arises from layers of regulation including mRNA stability, translational efficiency, protein turnover, and modulation by local microenvironments such as paracrine signaling and extracellular matrix interactions. These processes introduce nonlinear and context-dependent dynamics that cannot be fully captured by direct statistical mappings from RNA to protein.

Computational methods must therefore move beyond simple regression and incorporate biological priors or spatial context to capture these relationships more faithfully.

On the modeling side, questions of generalizability and interpretability remain pressing. Methods trained on specific tissues or experimental platforms often show reduced performance when applied to new datasets, reflecting sensitivity to both biological heterogeneity and technical noise. Furthermore, many of the most accurate frameworks rely on deep learning architectures, which, despite strong predictive power, function largely as black boxes and limit the extraction of mechanistic insights. Enhancing interpretability and ensuring transferability across tissues and conditions will be critical for building trust and broadening translational relevance.

Looking forward, several directions hold promise for advancing the field. The systematic generation and sharing of large-scale, multimodal datasets will provide the foundation for benchmarking and cross-tissue validation. Advances in machine learning, particularly interpretable graph-based and probabilistic approaches, may help bridge transcriptomic and spatial information while offering greater transparency. Hybrid strategies that combine limited experimental protein measurements with computational inference could alleviate cost barriers, while integrative frameworks that incorporate additional modalities such as histology, chromatin accessibility, or epigenetic states may yield richer biological insights. Ultimately, the maturation of computational spatial proteomics will depend not only on methodological innovation but also on integration into translational pipelines. By enabling therapy response prediction, spatial biomarker discovery, and detailed dissection of microenvironmental signaling, these frameworks have the potential to transform both basic biology and clinical research.

3.3 Conclusion

Spatial transcriptomics has transformed the study of tissue organization by enabling the measurement of gene expression within its native spatial context. However, the spot-level resolution of sequencing-based platforms necessitates computational deconvolution, while the limited coverage of proteomic assays has motivated inference of protein abundance from transcriptomic data. Together, these two methodological frontiers represent complementary strategies for bridging the gap between molecular readouts and cellular-level understanding. Deconvolution methods in spatial transcriptomics have evolved rapidly, encompassing a diverse array of computational frameworks, including probabilistic models,

matrix factorization techniques, graph-based approaches, deep learning architectures, and regression-based strategies. Each approach carries unique strengths and limitations, underscoring the need for hybrid frameworks that adaptively integrate multiple strategies.

At the same time, computational protein inference has emerged as a critical complement to RNA-centric analyses. While experimental platforms have enabled paired RNA–protein measurements, they remain constrained by cost, scalability, and antibody availability. Computational frameworks—ranging from image-based virtual staining, to single-cell RNA–protein mapping, to spatially aware deep learning models—aim to overcome these limitations. However, challenges persist in interpretability, generalizability, and data scarcity, highlighting the importance of generating standardized multimodal benchmarks and developing biologically informed algorithms.

In the future, progress will depend on three key directions: (i) systematic expansion of large-scale multimodal datasets that enable robust cross-tissue benchmarking, (ii) methodological innovation that balances predictive performance with interpretability and scalability, and (iii) integration of deconvolution and protein inference into translational pipelines for biomarker discovery, therapy response prediction, and clinical diagnostics. By advancing computational frameworks that bridge RNA and protein modalities while resolving cellular heterogeneity, the field is poised to deliver deeper mechanistic insights into tissue biology and disease.

A Cluster-Based Cell-Type Deconvolution of Spatial Transcriptomic Data

This chapter is based on the peer-reviewed article first-authored by Qingyue Wang, titled “DECLUST: A Framework for Cell Type Decomposition in Spatial Transcriptomics”, published in *Nucleic Acids Research* by Oxford University Press. The study presents a novel computational framework to estimate cell type composition from spatial transcriptomics data, addressing key challenges in resolution, interpretability, and benchmarking. The inclusion of this work reflects its central contribution to the thesis, both methodologically and conceptually. The content is reproduced under the terms of the Creative Commons Attribution (CC BY) license, with minor formatting adjustments to comply with thesis requirements.

Abstract

Spatial transcriptomics (ST) has emerged as an efficient technology for mapping gene expression within tissue sections, offering informative spatial context for gene activities. However, most current ST techniques suffer from low spatial resolution, where each spatial location often contains cells of various types. Deconvolution methods are used to resolve the cell mixture within the spots, but conventional approaches rely on spot-by-spot analyses, which are limited by low gene expression levels and disregard spatial relationships between spots, ultimately reducing performance. Here, we introduce DECLUST, a cluster-based deconvolution method to accurately estimate the cell-type composition in ST data. The method identifies spatial clusters of spots using both gene expression and spatial coordinates, hence preserving the spatial structure of the tissue. Deconvolution is subsequently performed on the aggregated gene expression of individual clusters, mitigating the challenges associated with low expression levels in individual spots. We evaluate DECLUST on simulated ST datasets from a human breast cancer tissue and two real ST datasets from human ovarian cancer and mouse brain. We compare DECLUST to current methods including CARD, GraphST, Cell2location, and Tangram. The results indicate that DECLUST not only maintains the spatial integrity of tissues but also outperforms

existing methods in terms of robustness and accuracy. In conclusion, DECLUST provides an effective and reliable approach for identifying cell-type compositions in ST data.

4.1 Introduction

Spatial transcriptomics (ST) enables the acquisition of transcriptome data while preserving the spatial context of cells within tissues, providing novel insights into cellular organization and micro-environmental interactions [142, 143, 85]. ST technologies can be broadly categorized into two main types: image-based and RNA-capture-based approaches. Image-based methods, such as in situ hybridization (ISH) [144, 145], single-molecule FISH (smFISH) [146], multiplexed error-robust fluorescence in situ hybridization (MERFISH) [44], and fluorescent in situ sequencing (FISSEQ) [29], provide subcellular resolution and enable direct visualization of individual mRNA molecules with high sensitivity. These methods are widely regarded as the gold standard for mRNA quantification. However, these methods are low throughput and the data analysis is complex and challenging. Their application is constrained by the relatively small number of genes, low sensitivity in mRNA detection, and labor-intensive protocols. In contrast, RNA-capture-based methods, such as laser capture microdissection combined with RNA sequencing (LCM + RNA-seq) [147] and array-based platforms like 10x Visium [25], allow for transcriptome-wide profiling across large tissue areas. These methods rely on spatial barcoding and next-generation sequencing to capture gene expression information, offering scalability and broader coverage compared to image-based techniques. However, they are limited by lower spatial resolution, as each capture spot typically contains multiple cells from different cell types [94], complicating the interpretation of cellular heterogeneity and intercellular interactions. The information of cell-type composition at spot level can improve the interpretation of intercellular interactions and support reconstructing the comprehensive landscape of cellular heterogeneity in tissues. This issue can be resolved by cell-type deconvolution based on the gene expression of spots.

Recent cell-type deconvolution methods can be broadly categorized into three main groups: 1) Probabilistic-based methods, which explicitly parameterize data distributions and utilize probability-based inference (e.g. Cell2location [104], RCTD [100], STRIDE [148], stereoscope [99]); 2) Matrix Factorization-based methods, which use techniques such as non-negative matrix factorization (NMF) and non-negative least squares (NNLS) to infer spot composition (e.g. CARD [102], SPOTlight [105], NMFreg [49], SpatialDWLS [106]); 3) Machine Learning-based methods, which apply machine learning models to

estimate cell-type composition (e.g. Tangram [110], GraphST [61]). Most deconvolution models do not adequately utilize the spatial information inherent in spatial transcriptomic data, thus limiting their accuracy. Some methods such as CARD and GraphST incorporate spatial information in their models, however they primarily perform deconvolution at the individual spots where inherently low gene expression levels in individual spots in ST data substantially decrease the performance of decomposition methods.

To address these challenges, we developed DECLUST, a cluster-based cell-type deconvolution method with a motivation that deconvolution could be more effective when applied to aggregated gene expression from groups of similar spots, where the signal is inherently stronger. DECLUST identifies spatial clusters of spots by integrating both gene expression profiles and spatial coordinates to preserve the spatial structure of the tissue. Deconvolution is then applied to the aggregated gene expression within each cluster, which overcomes the problem of low expression levels in individual spots.

We compared the performance of DECLUST with four other advanced cell-type deconvolution methods including CARD [102], GraphST [61], Cell2location [104], and Tangram [110]. These methods were selected because of their strong performance in previous benchmarking studies and their frequent use in the recent literature. By including a diverse range of top-performing approaches, we provide a comprehensive comparison across different modeling strategies using both simulated and real datasets.

4.2 Materials and methods

4.2.1 DECLUST pipeline

Figure 4.1 illustrates the DECLUST workflow, from the data input including ST data and reference scRNA-seq data to the final output of estimated cell-type proportions for individual spots in the ST data. The ST dataset consists of n spots, each has a vector of expressions of g_n genes and a 2-dimensional coordinate presenting the spatial location of the spot in the tissue image. We denote the gene expression matrix of the ST dataset as X ($n \times g_n$), including n rows of spots and g_n columns of genes. The scRNA-seq data contains a gene expression matrix, denoted as Z , including m cells and g_m genes. The cells are pre-annotated to belong to k distinct cell-types. We retain the top 5,000 most variable genes from each dataset, and the overlapping gene set (g genes) is kept for downstream analysis.

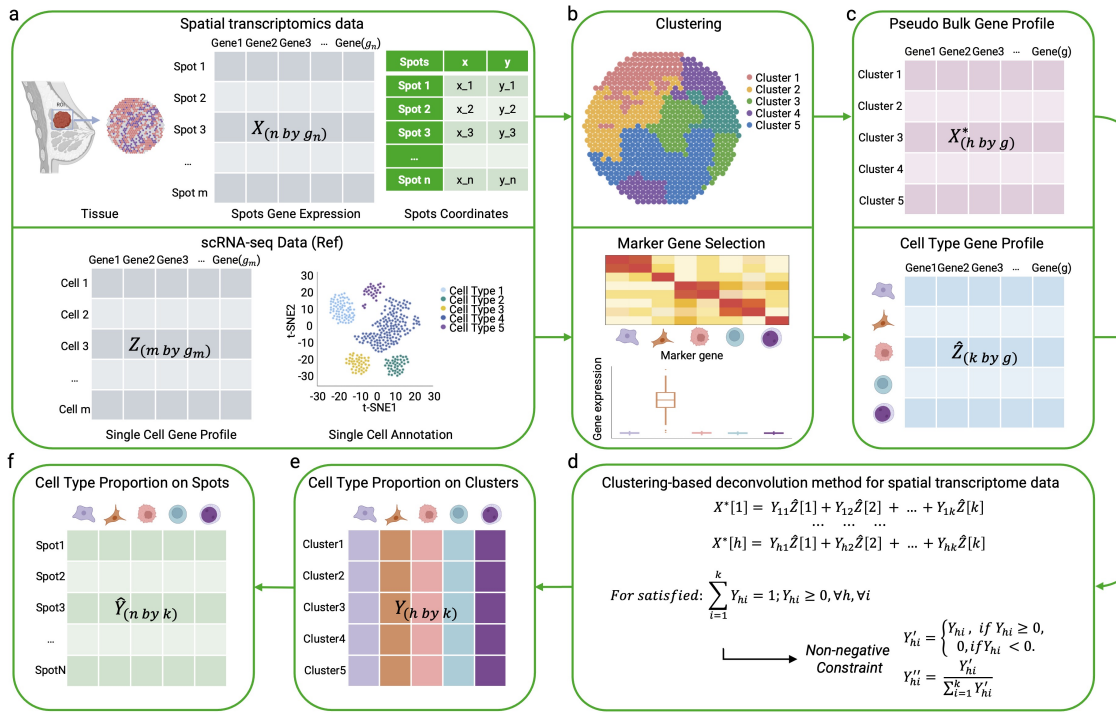


Fig. 4.1.: Overview of the DECLUST method. This flowchart illustrates the DECLUST for cell-type deconvolution from spatial transcriptomics data. (a) Input data of DECLUST consisting of spatial transcriptomics (ST) data and single-cell RNA sequencing (scRNA-seq) data derived from the same tissue sample for reference. (b) Data preprocessing involving two primary tasks: i) clustering spots of the ST data to identify spatial clusters, simultaneously considering both gene expression and spatial information, and ii) selecting cell-type-specific marker genes from the scRNA-seq data using a two-statistic method. (c) Preparation of gene profile matrices for deconvolution, including i) a pseudo-bulk gene profile matrix for the identified clusters and ii) a cell-type gene profile matrix for the cell-type-specific marker genes. (d) Deconvolution algorithm: the ordinary least squares (OLS) method with a non-negative constraint is employed to estimate the composition of cell-types within each spatial cluster. (e) Cluster-level cell-type proportion matrix which is the output of the deconvolution step. (f) Spot-level cell-type proportion matrix which is derived from the cluster-level cell-type proportion matrix.

In the subsequent step, DECLUST deploys a clustering method to identify spatial clusters of spots from the ST data and a two-statistic approach to discover marker genes of each cell-type before applying the ordinary least squares (OLS) algorithm with a non-negative constraint for deconvolution to predict cell-type proportions. The details of these steps will be described in the following sections.

4.2.1.1 Identification of spatial clusters of spots from ST data

This section will present an approach to identify spatial groups of spots based on both gene expression and spatial information, as depicted in Figure 4.2.

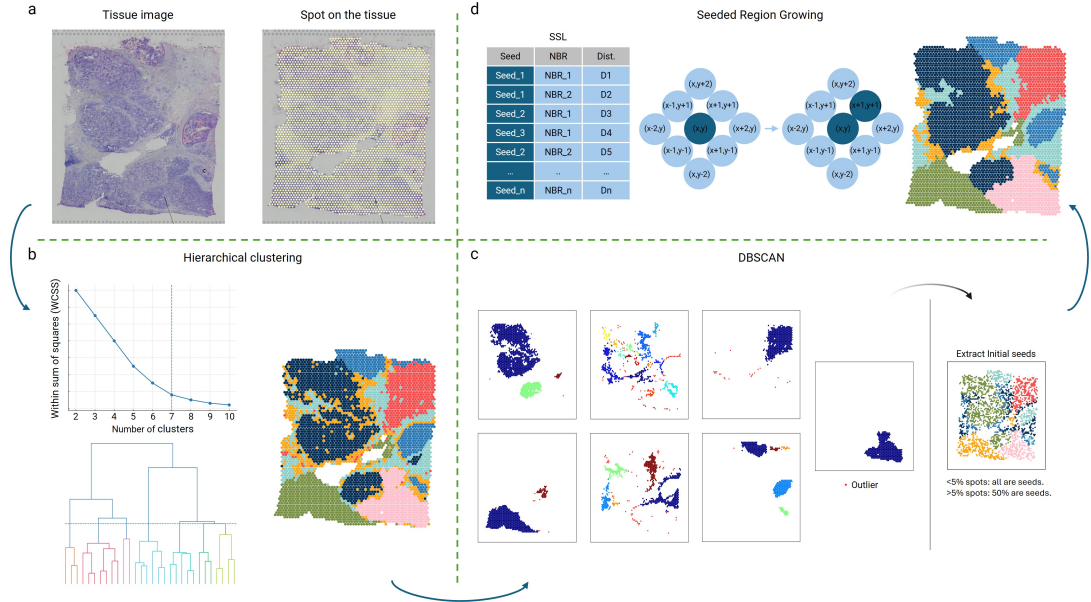


Fig. 4.2.: Overview of the proposed clustering approach of DECLUST. (a) The tissue images and the corresponding under-tissue spots are shown, serving as the starting spot for the analysis. (b) Hierarchical clustering is applied to the gene expression data to generate initial clusters, with the optimal number of clusters determined using the elbow method, as demonstrated by the within-cluster sum of squares (WCSS) plot. (c) DBSCAN (density-based spatial clustering of applications with noise) is applied to the initial clusters to identify spatial sub-clusters and outliers. (d) The seeded region growing (SRG) algorithm is used to refine the clusters, taking into account both gene expression similarity and spatial proximity of spots. Seeds are selected based on sub-cluster size and input into the SRG to finalize spot clusters, producing a segmentation that integrates both spatial and gene expression features.

Particularly, this method starts with applying the classical hierarchical clustering on gene expression to obtain initial spot clusters. Although the clusters derived based on gene expression contains rich biological signals, the clustering overlooks spatial continuity, leading to potentially fragmented clusters. To address this, the density-based spatial clustering of applications with noise (DBSCAN) [149] is applied to each of the clusters to identify their spatial sub-clusters. These sub-clusters not only capture the underlying spatial structure but also identify isolated spots that do not belong to any dense spatial group, which are treated as potential outliers. The incorrectly clustered spots may arise from the low sensitivity of gene expression profiles of individual spots, which could be improved in the next step. Finally, we generate the set of seeds from these sub-clusters, and input them

into the seeded region growing (SRG) algorithm. SRG explicitly grows regions based on both gene expression similarity and spatial proximity. Starting from the seeds identified by DBSCAN, SRG expands the regions in a controlled manner, allowing us to capture subtle spatial patterns while correcting for potential misassignments caused by noisy gene expression or spatial sparsity, resulting in the final clusters for the spots. For deconvolution, gene expression of the spots in each cluster are aggregated into a pseudo-bulk gene profile for the cluster, forming a cluster-gene expression matrix X^* including h clusters and g genes. Next sections describe further details of each step.

Hierarchical clustering of spots using gene expression

The hierarchical clustering is applied to the gene expression of spots to identify a set of initial spot clusters. The optimal number of clusters is determined using the elbow method [150, 151], which is based on the plot of the within-cluster sum of squares (WCSS) against the number of clusters. The "elbow" point, where further increases in the number of clusters result in minimal reductions in WCSS, indicates the optimal number of clusters.

The WCSS is calculated as follows:

$$\text{WCSS} = \sum_{j=1}^h \sum_{i=1}^{n_j} \|x_i^{(j)} - \mu_j\|^2 \quad (4.1)$$

where:

- h is the total number of clusters.
- n_j is the number of spots in cluster j .
- $x_i^{(j)}$ represents the gene expression profile of the i -th spot in cluster j .
- μ_j is the centroid (mean) of cluster j , representing the average gene expression profile.
- $\|x_i^{(j)} - \mu_j\|^2$ measures the squared Euclidean distance between a spot's gene expression profile and the cluster centroid.

The hierarchical clustering begins with each of the N spots treated as an individual cluster. Clusters are iteratively merged based on their proximity using the Ward linkage until the desired number of clusters is reached. The Ward linkage criterion, which minimizes the increase in total within-cluster variance when merging two clusters, is mathematically defined as:

$$d(A, B) = \frac{|A| + |B|}{T} \cdot \|\text{centroid}_A - \text{centroid}_B\|^2 \quad (4.2)$$

where:

- $d(A, B)$ is the distance between clusters A and B after merging.
- $|A|$ and $|B|$ are the sizes of clusters A and B , respectively.
- $T = |A| + |B|$, the total number of spots in both clusters.
- $\|\text{centroid}_A - \text{centroid}_B\|^2$ is the squared Euclidean distance between the centroids of clusters A and B .

The clusters identified by the hierarchical clustering provide a structure of the spots based on gene expression, but neglect the spatial relationships between spots. In the next section, for each cluster, DBSCAN algorithm will be applied on its spots to explore spatial sub-clusters based on their spatial relationship. Using these sub-clusters, a set of the pre-selected seeds will be collected for the SRG method at the last step.

Spatially sub-clustering of spots using DBSCAN and seed selection

To identify spatial sub-clusters of each initial cluster in the previous step, DBSCAN is applied to the coordinates of spots within the cluster. DBSCAN requires two parameters: ϵ , the maximum Euclidean distance between spots within the same sub-cluster, and minPts, the minimum number of spots required to form a sub-cluster. In practical implementation of DECLUST for ST data, we select $\epsilon = 4$ and minPts = 8.

Selection of seeds for the SRG method is based on the sub-cluster size:

- For a small sub-cluster with less than 5% of the total spots, none spots of the sub-cluster are designated as seeds.
- For a sub-cluster containing no less than 5% of the total spots, 50% of the spots in the sub-cluster are randomly selected as seeds, excluding the spots at the boundary of the tissue image.

Finally, each detected seed is assigned with a label by its corresponding sub-clusters. The set of seeds with their labels is inputted into the SRG algorithm in the next section to finalize the spot clusters for cell-type deconvolution.

In DECLUST, the choice of the parameters is empirically determined through extensive testing across multiple datasets. If the proportion of seeds from spots is too high, the spots are highly depending on the sub-cluster label which is determined by the hierarchical clustering in the previous step. In contrast, if the proportion too small, it means many spots are uncertain to belong to their current sub-clusters, which are later determined via the updating process by gene expression distance and/or their neighbors.

Seeded region growing method for spot segmentation to identify final spot clusters

In this section, we apply the SRG, an efficient segmentation method in image processing, to construct the spot clusters based on both the spatial relationship and gene expression similarity between spots. The algorithm utilizes the set of seeds and their labels identified from the DBSCAN in the previous step to construct clusters based on the minimum distance between spots and the seeds. To facilitate the search for the spot with minimum distance to the seeds, SRG constructs a sequentially sorted list (SSL). SRG defines an SSL as the set of spatial neighboring spots of the current seeds. The SSL is defined as a set of spatial neighboring spots of the current seeds, and is sorted by the gene-expression distance between the spots and their neighboring seeds [152].

SRG expands the seeds' regions iteratively based on the distance of neighboring spots stored in the SSL. At each iteration, the closest spot p is removed from the top of the SSL, then labelled based on its neighbors. Next, its unlabelled neighbors are added to the SSL. The process repeats until every spot is assigned with a label. The label of spot p is assigned based on the following criteria:

$$L(p) = \begin{cases} L(q) & \text{if } L(q) = L(r) \text{ for all neighboring spots } q, r \\ L_{\text{nearest}} & \text{if neighboring spots have different labels} \end{cases} \quad (4.3)$$

where $L(p)$ is the label assigned to the spot p , and L_{nearest} refers to the label of the closest neighboring region in terms of the gene expression profile.

The details of the SRG algorithm can be found in the related study [152]. SRG not only efficiently integrates spatial relationship and gene expression similarity of spots to generate clusters, but also automatically determines the number of clusters, enhancing the the overall segmentation accuracy and quality of ST data.

4.2.1.2 Discovery of marker genes from scRNA-seq data

We apply a two-statistic approach outlined in a recent study [153] to identify cell-type specific genes for k cell-types from scRNA-seq reference data, then use them as marker genes for the deconvolution step. We consider a cell-type specific marker gene if the expression of the gene is highly up-regulated in one cell-type, but its expression in the remaining cell-types is not statistically different. To do this, we employ two statistics: a robust t-test (T1) and a chi-square test (T2). T1 determines whether a given cell-type differs significantly from all others, while T2 assesses whether the remaining cell-types exhibit similar expression levels. A gene is considered cell-type-specific if T1 is large, indicating a statistically significant difference from other cell-types, and T2 is small, indicating no significant difference among the remaining cell-types. Statistical significance is controlled using p-values, with thresholds set at p for the $T1 < 0.05$ and p for the $T2 > 0.05$. Top t genes by t-statistics are selected as marker genes for each cell-type to build a reference matrix. In implementation, we keep top $t = 10$ genes for individual cell-types. The reference matrix $\hat{Z}(k \times g)$ is constructed from k cell-types and g marker genes. For a marker gene j and cell-type i , the value of $\hat{Z}_{i,j}$ is calculated by averaging gene expression of the marker gene across all single cells belonging to the cell-type in the single-cell data reference.

4.2.1.3 Cell-type deconvolution using ordinary least squares

The primary objective of the deconvolution task is to estimate the cell-type proportion matrix $Y(m \times j)$ based on the cluster-gene profile matrix $X^*(h \times g)$ and cell-type gene profile

matrix $\hat{Z}(k \times g)$, which are identified in the previous steps. The problem is formulated as a system of linear equations:

$$\begin{aligned} X^*[1] &= Y_{11}\hat{Z}[1] + Y_{12}\hat{Z}[2] + \cdots + Y_{1k}\hat{Z}[k], \\ X^*[2] &= Y_{21}\hat{Z}[1] + Y_{22}\hat{Z}[2] + \cdots + Y_{2k}\hat{Z}[k], \\ &\vdots \\ X^*[h] &= Y_{h1}\hat{Z}[1] + Y_{h2}\hat{Z}[2] + \cdots + Y_{hk}\hat{Z}[k]. \end{aligned} \tag{4.4}$$

where:

- $X^*[h]$ represents the gene expression profile for the h -th spatial cluster. Each row in the matrix X^* corresponds to a different spatial cluster, capturing the expression levels of a set of genes for that cluster.
- $\hat{Z}[k]$ represents the gene expression profile for the k -th cell-type. Each row in the matrix $\hat{Z}$ corresponds to a different cell-type, describing how strongly each gene is expressed in that particular cell-type.

The cell-type proportions for each cluster are estimated using ordinary least squares (OLS) with a non-negative constraint. The estimates are obtained as follows.

$$\begin{aligned} Y[1] &= \left(\hat{Z}^T \hat{Z}\right)^{-1} \hat{Z}^T X^*[1], \\ Y[2] &= \left(\hat{Z}^T \hat{Z}\right)^{-1} \hat{Z}^T X^*[2], \\ &\vdots \\ Y[h] &= \left(\hat{Z}^T \hat{Z}\right)^{-1} \hat{Z}^T X^*[h]. \end{aligned} \tag{4.5}$$

where each cell-type proportion vector $Y[i]$ of cluster i is constrained by:

$$\sum_{j=1}^k Y_{ij} = 1, \quad Y_{ij} \geq 0, \quad \forall i, \forall j. \quad (4.6)$$

To keep the non-negative constraint, each Y_{ij} is adjusted as follows:

$$Y'_{ij} = \begin{cases} Y_{ij}, & \text{if } Y_{ij} \geq 0 \\ 0, & \text{if } Y_{ij} < 0 \end{cases} \quad (4.7)$$

Then, the vector $Y'[i]$ is normalized to keep the sum of cell-type proportions equal 1:

$$Y''_{ij} = \frac{Y'_{ij}}{\sum_{j=1}^k Y'_{ij}}. \quad (4.8)$$

The resulting matrix Y , representing cell-type proportions at the cluster level, is then expanded into a matrix $\hat{Y}(n \times k)$ to achieve the cell-type proportions at spot level.

4.2.2 Datasets

4.2.2.1 Spatial transcriptomics data generation and preprocessing

The real spatial transcriptomics datasets analyzed in this chapter were generated using array-based spatial barcoding technologies, primarily the 10x Genomics Visium platform. This platform enables transcriptome-wide profiling while preserving spatial information by capturing mRNA molecules on spatially indexed spots arranged on a glass slide. Each capture spot has a diameter of approximately 55 μm and typically contains transcripts from multiple neighboring cells.

Briefly, fresh-frozen tissue sections were mounted onto Visium slides and stained for histological visualization. Following tissue permeabilization, released mRNA molecules hybridized to barcoded oligonucleotides containing spatial barcodes and unique molecular identifiers. Reverse transcription and library preparation were subsequently performed according to the manufacturer's protocol, and sequencing libraries were generated for downstream analysis. The resulting sequencing reads were processed using the Space

Ranger pipeline provided by 10x Genomics, which aligns reads to the reference genome, collapses unique molecular identifiers, and assigns gene expression counts to individual spatial spots based on their barcodes.

Raw gene expression matrices produced by Space Ranger were subjected to standard quality control and preprocessing prior to downstream analysis. Spots with low sequencing depth or low gene detection rates were removed to mitigate technical noise. Genes expressed in only a small number of spots were filtered out to reduce sparsity. The retained expression counts were normalized to account for differences in library size across spots and log-transformed to stabilize variance. Spatial coordinates associated with each spot were retained and used in subsequent clustering and deconvolution analyses. These preprocessing steps yielded spatially annotated expression matrices that serve as the input for the DECLUS pipeline.

4.2.2.2 Real datasets

Two real ST datasets generated by the 10x Genomics Visium Spatial Gene Expression platform [154] are used in this study.

Mouse brain ST dataset: Fresh frozen coronal sections from a male C57BL/6 mouse (>8 weeks old) were prepared using the 10x Genomics Visium Spatial Gene Expression technology. Tissue sections of 10 μm thickness were stained with Hematoxylin and Eosin (H&E) and imaged using a Nikon Ti2-E microscope. Spatial gene expression libraries were sequenced on an Illumina NovaSeq 6000 with paired-end reads (28×90 bp), achieving an average sequencing depth of 115,740 reads per spot. There were a total of 2,698 spots, the median of 27,198 UMI counts and 5,798 genes were detected per spot. This dataset is used to assess the accuracy of predicted cell-type proportion and how well the cell-type deconvolution methods keep the spatial relation between spots through cross-validation.

Single-cell reference data for the mouse brain ST dataset: The gene expression of this single-cell RNA sequencing dataset was derived from the primary visual cortex of 6–8-week-old C57BL6/J mice. The dataset, generated using the inDrops platform, included data from 16,188 cells and 25,187 genes, covering eight major cell-types. The details of this dataset can be found in the original article [155, 156].

Human ovarian cancer ST dataset: Formalin-fixed, paraffin-embedded (FFPE) tissue sections (5 μm) from a case of serous papillary carcinoma were sourced from iSpecimen

(Block 108906). Sections were processed using the Visium platform, stained, and imaged using a brightfield camera at 20x magnification. Libraries were sequenced on an Illumina NovaSeq 6000 with paired-end reads (28×75 bp), with a sequencing depth averaging 63,836 reads per spot. There were a total of 3,455 spots, the median of 28,927 UMI counts and 7,621 genes were detected per spot. We utilize this dataset to investigate the preservation of spatial pattern of cell-type marker genes from ST data.

Single-cell reference data for human ovarian cancer ST dataset: The gene expression of this dataset consisted of 26,906 genes and 18,299 cells, generated by 10x genomics single-cell 5' RNA sequencing for 9 patients with high-grade serous ovarian carcinoma (HGSOC). The dataset captured the tumor's cellular heterogeneity, including epithelial tumor cells, tumor-infiltrating immune cells, and various stromal components. More information about this dataset can be referred to its original study [157, 158].

4.2.2.3 Simulated ST datasets

In this study, we utilize the real human breast cancer ST dataset to simulate ST data for evaluation of the deconvolution methods. This dataset involved a Visium Spatial Gene Expression experiment on fresh frozen Invasive Ductal Carcinoma breast tissue (AJCC/UICC Stage IIA, ER-positive, PR-negative, Her2-positive), obtained from BioIVT Asterand. The tissue was 10 μ m thick, placed on Visium Gene Expression Slides, and imaged using a Nikon Ti2-E microscope. Sequencing was performed on an Illumina NovaSeq 6000 with a configuration of 28×120 bp and a depth of 150,392 reads per spot. A total of 3,798 spots were detected, with a median of 20,762 UMI counts per spot and a median of 6,026 genes per spot.

Simulation 1. We first apply DECLUSST on the dataset to predict cell-type compositions of spots, then use them to simulate the gene expression of spots using single cell reference data. The single-cell transcriptomics dataset comprises 100,064 cells across 9 major cell types, collected from 26 primary tumors representing three breast cancer subtypes: 11 ER+, 5 HER2+, and 10 TNBC patients. It contains expression data for 29,733 genes [159, 160]. Specifically, for each spot, we use the multinomial distribution to generate cell counts of each cell-type according to the cell-type proportions by DECLUSST and the total number of cells in the spot. We fix the number of cells per spot to 10, which is in the range value of the number of cells in the spots from 10x Visium technology. Next, we randomly selected single cells from the single-cell reference dataset using generated cell counts for each cell-type. Finally, the gene expression profiles of the selected cells are aggregated to

create a simulated gene expression profile of the spot. We use the same spot coordinates of the real ST data for the simulated ST dataset. Thus, the simulated ST dataset preserves the spatial structure observed in the original data while controlling cell-type proportions of spots.

We generate ten simulated datasets by repeating the procedure, allowing us to assess the variability of the deconvolution methods. The simulation 1 replicates are used to evaluate the variability of deconvolution methods, which is discussed in section Results.

Simulation 2. We have added Ligand-Receptor(LR) interactions in addition to Simulation 1 framework for generating ST datasets, which mimic real spatial transcriptomics data. We used **BayesSpace** [161], a method that incorporates both gene expression and spatial coordinates to detect spatial domains, to generate spatial clusters for real breast cancer ST dataset. Then cell-type proportions are inferred for each cluster using **CIBERSORT** [162], a widely used reference-based bulk deconvolution algorithm.

To mimic cell-type heterogeneity and co-localization within each cluster, we select a pair of dominant cell types and simulate enrichment for the spots of the cluster using LR interactions with a high probability, which are inferred from **CellChat** [163]. To make the LR enrichment specific for a pair of cell types, we select only single cells highly expressed in both ligand and receptor from the single cell reference dataset for the simulation. Specifically, single cells satisfying the expression of both ligand and receptor in the third quartile (top 25%) are used. In Simulation 2, we select two clusters for enrichment, including:

- Cluster 4, guided by the *CXCL12–CXCR4* interaction, we enriched *CAF* and *Plasmablast* cells.
- Cluster 7, guided by *TNF–TNFRSF1A*, we enriched *Myeloid* and *Endothelial* cells.

The spatial expression patterns of the selected ligand-receptor pairs are visualized in Figure 4.3 to illustrate their role in guiding cell-type-specific enrichment in Simulation 2.

4.2.3 Competing cell-type deconvolution methods

To benchmark our model, we compared it against four established cell-type deconvolution methods that utilize a variety of computational approaches. These methods, widely adopted

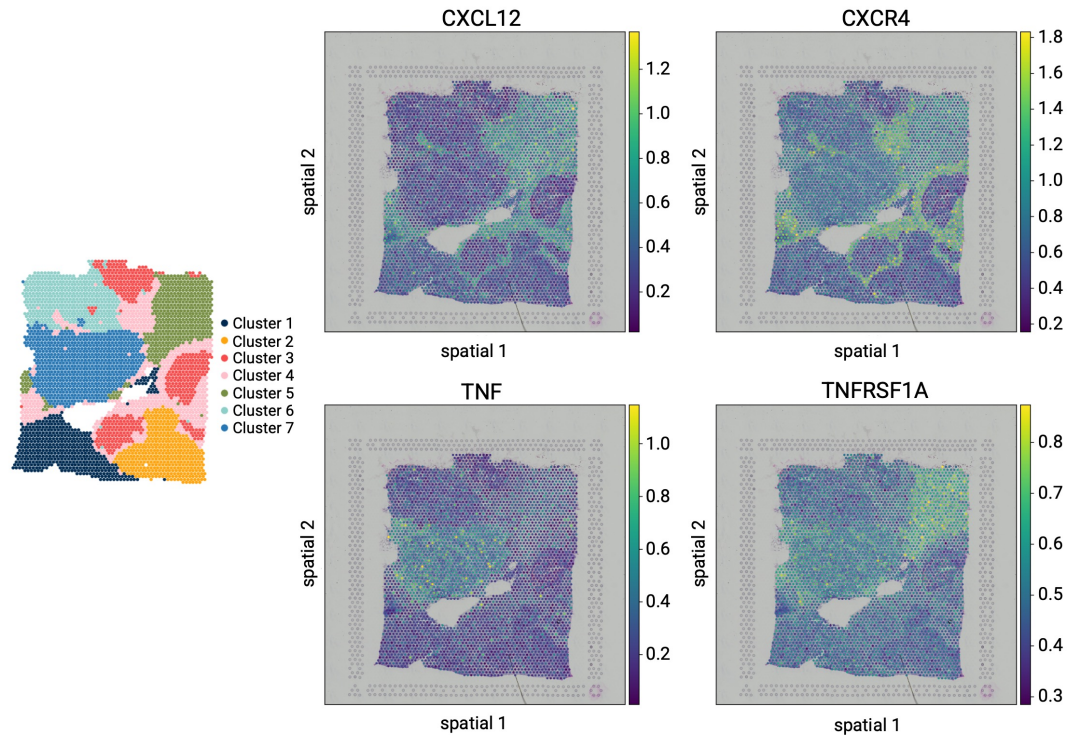


Fig. 4.3.: Spatial expression of Ligand-Receptor pairs used to guide cell-type enrichment in Simulation 2. The left panel displays the spatial clusters identified by BayesSpace in the simulated dataset, with each color representing a distinct cluster (Clusters 1–7). The right panels show the spatial expression patterns of four selected ligand-receptor genes: CXCL12, CXCR4, TNF, and TNFRSF1A. The ligand-receptor pair CXCL12–CXCR4 guides the enrichment of CAF and Plasmablast cells in Cluster 4, while TNF–TNFRSF1A guides the enrichment of Myeloid and Endothelial cells in Cluster 7. These spatial plots illustrate the localized expression patterns that support the targeted enrichment strategy.

for both bulk RNA-seq and spatial transcriptomic data, leverage different strategies such as matrix factorization, probabilistic inference, graph-based learning, and deep learning. Each of these methods aims to decompose mixed cellular signals into distinct cell-type-specific profiles. A detailed summary of these methods can be found in Table 4.1.

CARD employs non-negative matrix factorization to infer cell-type proportions by using highly variable genes and single-cell RNA-seq data as reference profiles. It incorporates spatial information of spatial transcriptomics into deconvolution, improving the accuracy of both proportion estimates and spatial maps of cell distributions.

Cell2location adopts a probabilistic model with variational Bayesian inference, assuming gene expression follows a negative binomial distribution. It relies on scRNA-seq reference

data for estimating cell-type-specific signatures and models the spatial expression as a combination of these signatures.

GraphST utilizes a graph neural network-based framework, constructing a spatial graph from highly variable genes and applying self-supervised contrastive learning. This method aligns spatial gene expression with reference scRNA-seq data, capturing local spatial patterns and producing deconvolved cell-type distributions.

Tangram illustrates a deep learning model to map scRNA-seq or snRNA-seq data onto spatial transcriptomics data. By maximizing the correlation between shared gene expression across datasets, Tangram estimates cell-type proportions while maintaining spatial coherence.

Tab. 4.1.: Summary of competing cell-type deconvolution methods.

Method	Language	Feature selection	Inference method	Implement
CARD	R	Highly variable genes	Non-negative matrix factorization model	CARD deconvolves bulk RNA-seq or spatial transcriptomics data using a non-negative matrix factorization model. It begins by identifying highly variable genes and uses single-cell RNA-seq data as a reference to extract cell-type-specific gene expression profiles. CARD then models the bulk or spatial data as a mixture of these profiles, estimating the proportions of different cell-types in each sample. For spatial data, it incorporates spatial information to improve the accuracy of deconvolution, producing both cell-type proportion estimates and spatial maps of cell-type distributions.
Cell2location	Python	No selection	Probabilistic, negative binomial distribution, variational Bayesian inference	Cell2location applies a hierarchical Bayesian framework, assuming gene expression counts follow a negative binomial distribution. It first estimates cell-type-specific signatures using external scRNA-seq data as a reference. The spatial expression count matrix is then modeled with a negative binomial distribution, and variational Bayesian inference is used to approximate the posterior distribution, producing parameter estimates for cell-type proportions accordingly.

Continued on next page

Method	Language	Feature selection	Inference method	Implement
GraphST	Python	Highly variable genes	Graph self-supervised contrastive learning	GraphST deconvolves spatial transcriptomics data using a graph-based, self-supervised learning framework. It starts by selecting highly variable genes and constructing a spatial graph where nodes represent spatial spots connected based on proximity. Each node is associated with gene expression data. A graph neural network (GNN) is trained using contrastive learning to capture spatial and gene expression patterns. The learned embeddings are then matched to reference single-cell RNA-seq data to estimate cell-type proportions in each spot, providing both deconvolved cell-type distributions and low-dimensional representations of the data.
Tangram	Python	Highly variable genes	Deep learning model	Tangram takes sc/snRNA-seq and spatial transcriptome data from the same region or tissue, requiring only a shared subset of common genes. It initially places sc/snRNA-seq expression profiles randomly in space, then computes an objective function to model the spatial correlation between genes in both datasets. Tangram iteratively rearranges the sc/snRNA-seq profiles to maximize the total spatial correlation of the shared genes.

4.2.4 Performance metrics for simulation study

In each simulated dataset, the true cell-type proportions of a spot is calculated by the ratio between cell-type's population and the total number of cells in that spot. We denote the true cell-type proportion matrix of the spots as S . After obtaining the estimated cell-type proportion matrix $\hat{Y}$ from a deconvolution method, we assess its performance by computing the Root Mean Square Error (RMSE) between $\hat{Y}$ and S as follows.

$$\text{RMSE} = \sqrt{\frac{1}{g} \sum_{j=1}^g (\hat{Y}_{ij} - S_{ij})^2}. \quad (4.9)$$

where $\hat{Y}_{ij}$ and S_{ij} are the estimated and true cell-type proportion of the i th spot for the j th cell-type, respectively. For the calculation of Mean Square Error (MSE), we exclude the squared root from the formula.

4.2.5 Cross-validation for real ST data

We aim to use cross-validation to evaluate performances of the deconvolution methods in both cell-type proportion estimation accuracy and in the preservation of spatial relation of spots. To do cross-validation, we first randomly divide the spots in each cluster to 10 folds, thus the spots in each fold are independent and spatially non-adjacent. Then, each fold in turn selected as a test set, the remains are used for training set, across clusters. Next, a deconvolution model is applied to the training set to estimate cell-type proportions of individual spots. For the spots in the test set, the cell-type proportions are predicted by the average values of the neighboring spots from the training set.

The rationale behind this is that the molecular profiles of neighboring spots in ST dataset are highly correlated, as shown in Figure 4.7b. Therefore, we can use the predicted cell-type proportions for the test set as the baseline for comparison of deconvolution methods. The result of a good deconvolution method is expected to keep well the spatial relation between spots. Of note, our intention with Figure 4.7b is not to claim the existence of local correlation as a novel finding. It is a model-independent base performance that can be utilized as a target to evaluate deconvolution methods in real ST data where ground-truth proportions are unavailable. Similar to DECLUST, other methods including CARD and GraphST do incorporate spatial information into their models. So, the information does not give DECLUST a special advantage.

Subsequently, we reconstruct the gene expression profile of the spot in the test set using a convolution process. The convolution is performed for the predicted proportions of cell-types and the average gene expression across all cells of the cell-types from the single-cell reference data. Finally, the performance of the deconvolution methods are assessed by comparing these convoluted gene expression profiles with the observed values at spot level in the test set. We use the following metrics to assess metrics to assess the cross-validated performance of the deconvolution methods.

Root Mean Square Error

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^n \sqrt{\frac{1}{g} \sum_{j=1}^g (y_{ij} - \tilde{y}_{ij})^2}. \quad (4.10)$$

where n is the number of spots in the test set, g is the number of genes, y_{ij} is the observed of gene j on spot i , $\tilde{y}_{ij}$ is the predicted expression level of gene j on spot i .

Pearson Correlation Coefficient

$$\text{PCC} = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^g (y_{ij} - \bar{y}_i)(\tilde{y}_{ij} - \bar{\tilde{y}}_i)}{\sqrt{\sum_{j=1}^g (y_{ij} - \bar{y}_i)^2} \cdot \sqrt{\sum_{j=1}^g (\tilde{y}_{ij} - \bar{\tilde{y}}_i)^2}}. \quad (4.11)$$

where the symbols defined are the same as those for calculating the RMSE value, added $\bar{y}_i$ is the observed mean expression level in the spot i , $\bar{\tilde{y}}_i$ is the predicted mean expression level in the spot i .

Spearman Correlation Coefficient

$$\text{Spearman} = \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{6 \sum_{j=1}^g (R_{y_{ij}} - R_{\tilde{y}_{ij}})^2}{g(g^2 - 1)} \right). \quad (4.12)$$

where $R_{y_{ij}}$ and $R_{\tilde{y}_{ij}}$ represent the rankings of the i th site and the j th gene in the observed and predicted gene expression levels, respectively.

4.3 Results

4.3.1 Description of DECLUST methodology

The pipeline of DECLUST to estimate cell-type composition from spatial transcriptomic (ST) data is presented in Figure 4.1. DECLUST highlights the value of spatial information for enhancing deconvolution performance. The motivation of DECLUST stems from the idea that deconvolution could be more effective when applied to aggregated gene expression from groups of similar spots, where the signal is inherently stronger. The key steps of DECLUST include 1) clustering spots using both gene expression profiles and spot spatial coordinates; 2) identifying cell-type-specific marker genes for reference from single-cell transcriptomics; and 3) estimating cell-type proportions using the gene expression profiles of the clusters and cell-type-specific marker genes. Although this strategy sacrifices some spatial granularity, it enhances signal robustness in regions with sparse expression. Importantly, the clusters are both transcriptionally similar and spatially coherent, consistent with prior observations that neighboring cells in tissues tend to form biologically meaningful domains [164]. As shown in Figure 4.4, the clustering step with the aggregation strategy makes a significant contribution to DECLUST's performance.

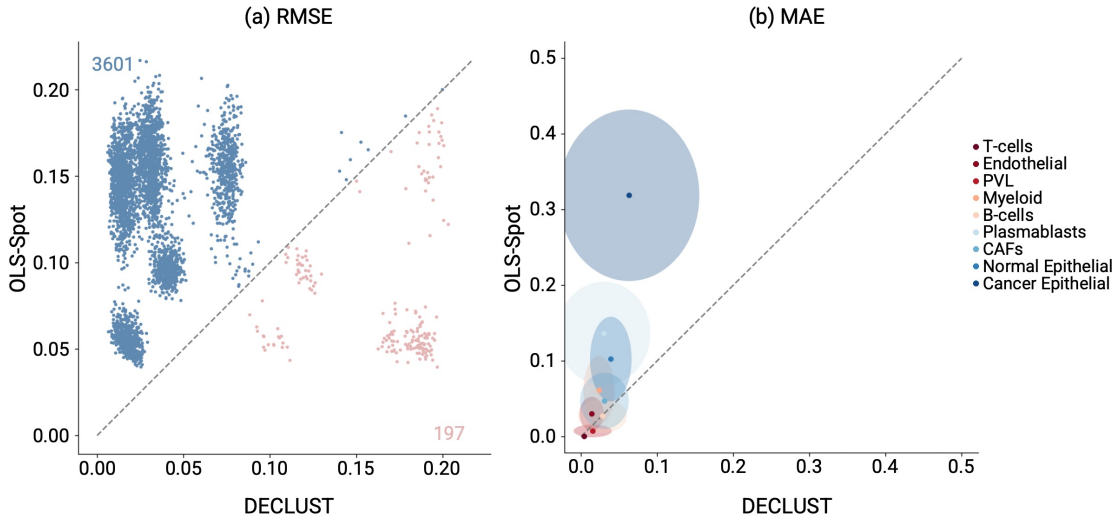


Fig. 4.4.: Evaluating the contribution of the clustering step in DECLUST.(Scatter plots illustrating the performance comparison between DECLUST and OLS deconvolution on simulated dataset 1. Left: Root Mean Squared Error (RMSE) per spatial spot. Each point represents a single spot, with the x-axis showing the error from DECLUST and the y-axis showing the error from OLS. The majority of points lie above the diagonal, indicating that DECLUST achieves lower RMSE than OLS on 3601 out of 3798 spots. This suggests that incorporating the clustering step before deconvolution consistently improves prediction accuracy at the spot level. Right: Mean Absolute Error (MAE) per cell type. Each point represents the average MAE for a specific cell type, and the size of each ellipse corresponds to the variability across spots. Most points fall below the diagonal, indicating lower MAE for DECLUST compared to OLS. Notably, larger improvements are observed for cell types with higher variability under OLS, such as Cancer Epithelial and Normal Epithelial cells, highlighting the robustness of the clustering-enhanced approach in handling heterogeneous populations.

The input for DECLUST in Figure 4.1a includes an ST dataset for deconvolution and single-cell RNA sequencing (scRNA-seq) data from the same tissue as a reference. The ST dataset consists of gene expression profiles and spatial coordinates of spots, and the scRNA-seq reference provides gene expression profiles along with corresponding cell-type annotations for individual cells. DECLUST starts with identification of spatial clusters of spots from the ST data using its proposed clustering method based on both gene expression profiles and spot spatial coordinates, Figure 4.1b. Marker genes of each cell-type are also discovered from the scRNA-seq data using a two-statistic approach. The outputs of these methods are presented in Figure 4.1c, including two matrices: a pseudo-bulk gene profile matrix (X^*) of spatial clusters of spots and a cell-type gene profile matrix ($\hat{Z}$) derived from the scRNA-seq reference. Subsequently, the matrices are entered into the ordinary least squares (OLS) algorithm [165, 166] with a non-negative constraint to estimate the cell-type proportion of individual spatial clusters, Figure 4.1d. The OLS output is presented by a

matrix Y in Figure 4.1e, which is later used to achieve the estimated cell-type proportions for individual spots, matrix $\hat{Y}$, in Figure 4.1f.

4.3.2 DECLUST achieves high accuracy in simulated ST data

The simulated ST datasets to evaluate DECLUST are generated based on a real human breast cancer ST data produced by 10x Visium [154]. Figure 4.5a shows that the simulated ST data closely resemble the real ST data and effectively preserve the original tissue structure. In the left-most column of Figure 4.5a, the image of the real ST data is shown at the top, while the next two images present the spot clusters of the simulated ST data in Simulation 1 and Simulation 2. The remaining columns of Figure 4.5a illustrate the similarity between the simulated and real ST data by comparing the spatial expression patterns of the top marker genes for four of the most dominant cell types: CAFs, myeloid cells, cancer epithelia, and endothelial cells, arranged from left to right. Because the real ST data usually lacks of the cell-type annotations from the ground truth, marker genes can serve as proxies for evaluating the spatial distribution of cell-types [167, 168, 48, 169]. The top marker genes, shown in Figure 4.5b, are identified from a publicly available scRNA-seq breast cancer dataset [159], satisfying that they are highly expressed in one cell-type but not showing statistically significant differences in the remaining cell-types.

Figure 4.5c presents the performance of DECLUST in estimation of cell-type proportions across 10 replicate datasets of Simulation 1 (top) and Simulation 2 (bottom). The cancer epithelial, the most abundant cell-type, has the largest mean squared error (MSE) of 0.013 in Simulation 2, while the MSEs of all other cell-types are less than 0.004, indicating the high accuracy of DECLUST. To further illustrate the accuracy of DECLUST in predicting cell-type proportions, Figures A.1 and A.2 in the Appendix provide scatter plots comparing the true cell-type proportions with the predicted proportions for each method across all 10 replicates in Simulation 1 and Simulation 2, respectively. The close alignment of points along the diagonal line in the DECLUST and CARD methods demonstrates their superior performance, while other methods, particularly GraphST, show greater variability and divergence from the true proportions. This visualization underscores the strong predictive accuracy of DECLUST across various simulated datasets.

Figure 4.5d shows that DECLUST also outperforms existing deconvolution methods including CARD, GraphST, Cell2location and Tangram in the simulated ST data. This figure presents the results of Simulation 1 (left) and Simulation 2 (right). As shown in the figure, DECLUST achieves the lowest root mean square error (RMSE) compared to

the other methods in both simulations, although its RMSE values in Simulation 2 are higher than Simulation 1. CARD follows as the runner-up with an RMSE slightly higher than DECLUST. The remaining methods—Cell2location, Tangram, and GraphST—show progressively higher RMSE values, with GraphST performing the worst.

Furthermore, the results of DECLUST are less varied across simulation replicates with the narrow interquartile range (IQR) in the box plots than the other methods, Figure 4.6. The narrow IQR indicates DECLUST's reliability for cell-type deconvolution from ST data, producing results with consistent accuracy.

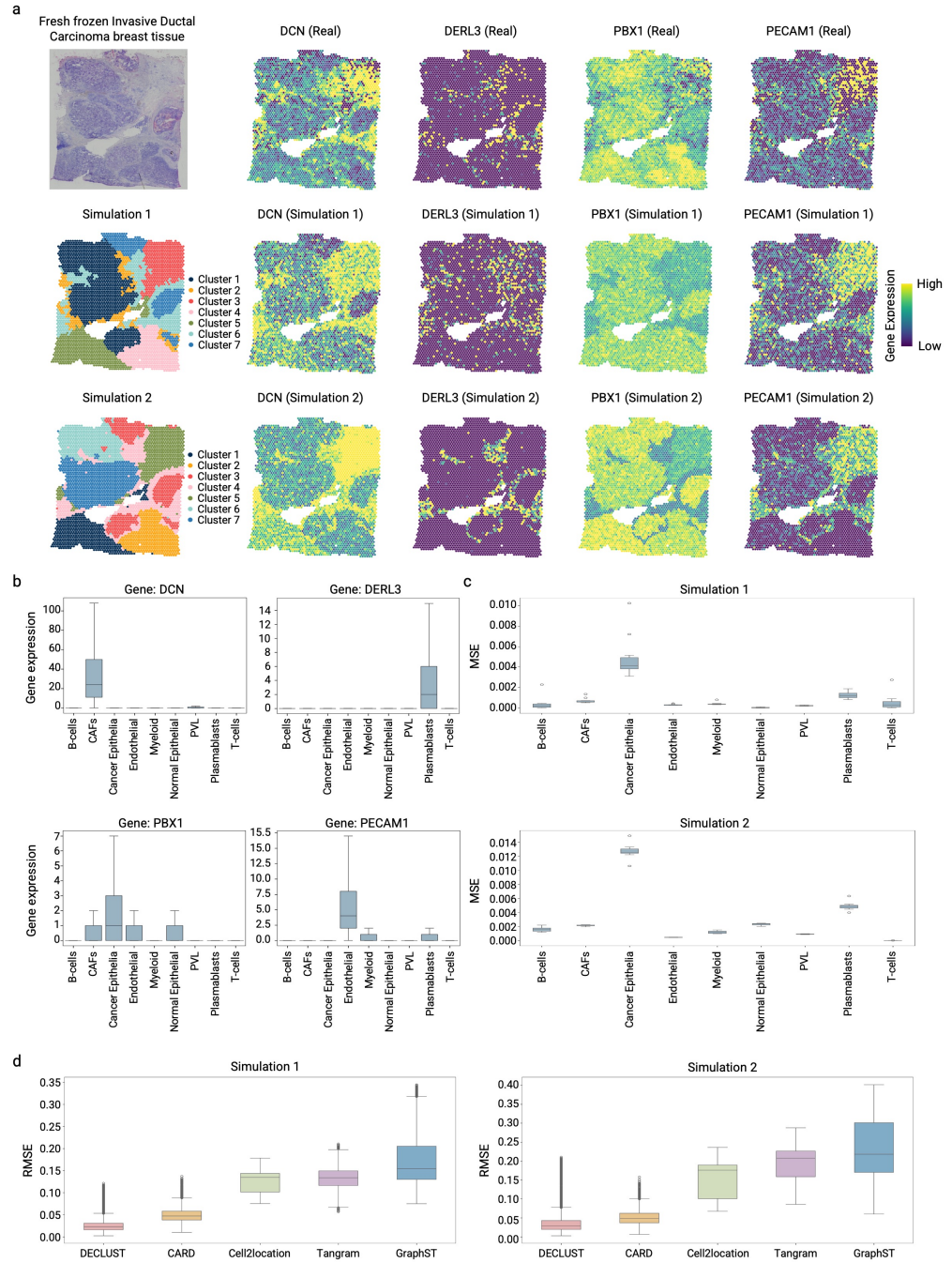


Fig. 4.5: Comparison of DECLUST with other cell-type deconvolution methods in simulated datasets.(a) Similarity between simulated and real ST data from fresh frozen breast IDC tissue. Top left: original H&E image. Middle and bottom left: DECLUST clustering results from Simulation 1 and 2 with seven clusters. Remaining panels: expression of four marker genes (DCN, TYROBP, PBX1, PECAM1) from CAFs, myeloid cells, cancer epithelia, and endothelial cells. Top row: real ST; second and third rows: Simulations 1 and 2. (b) Box plots of top marker gene expression across the four dominant cell types. (c) DECLUST performance by mean squared error (MSE) across 10 replicates for Simulation 1 (top) and 2 (bottom). (d) Comparison with CARD, Cell2location, Tangram, and GraphST in Simulations 1 (left) and 2 (right). Box plots show RMSE across all spots (median, quartiles, whiskers at 1.5 IQR).

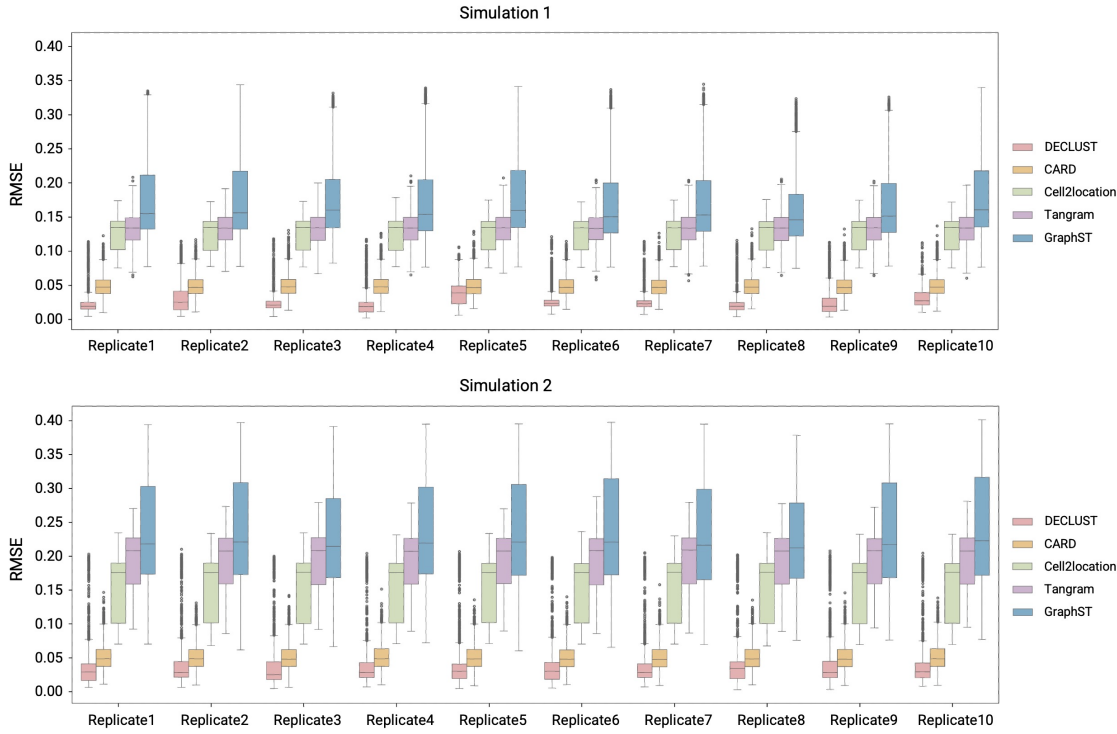


Fig. 4.6.: Comparative performance of spatial transcriptomics methods under two simulation settings across 10 replicated datasets in Simulation 1 and Simulation 2. Boxplots display the distribution of root mean square error (RMSE) for five deconvolution methods—DECLUST, CARD, Cell2location, Tangram, and GraphST—across 10 independent replicates for each simulation scenario. In both Simulation 1 and Simulation 2, DECLUST consistently demonstrates the lowest median RMSE values, with narrow interquartile ranges and fewer outliers, suggesting both high accuracy and robustness. CARD follows with slightly higher RMSEs but remains consistently lower than Cell2location, Tangram, and GraphST. Notably, GraphST shows the highest RMSE and largest variability across replicates, with median values exceeding 0.20 in most cases. These results indicate that DECLUST significantly outperforms existing methods in terms of deconvolution accuracy and consistency across diverse simulated conditions.

4.3.3 DECLUST performs well against other methods in the analysis of the real mouse brain ST data

Next we apply cross-validation to evaluate performance of DECLUST in the analysis of a real ST dataset of a mouse brain produced by 10x Visium [154]. The corresponding scRNA-seq dataset for the reference is collected from a recent study [156, 155], consisting of eight cell-types, which will be used for deconvolution. DECLUST suggests seven clusters for the mouse brain ST data, Figure 4.7a, closely aligning with the observed image structure. For example, the hippocampal region inside temporal lobe, surrounded by a dark

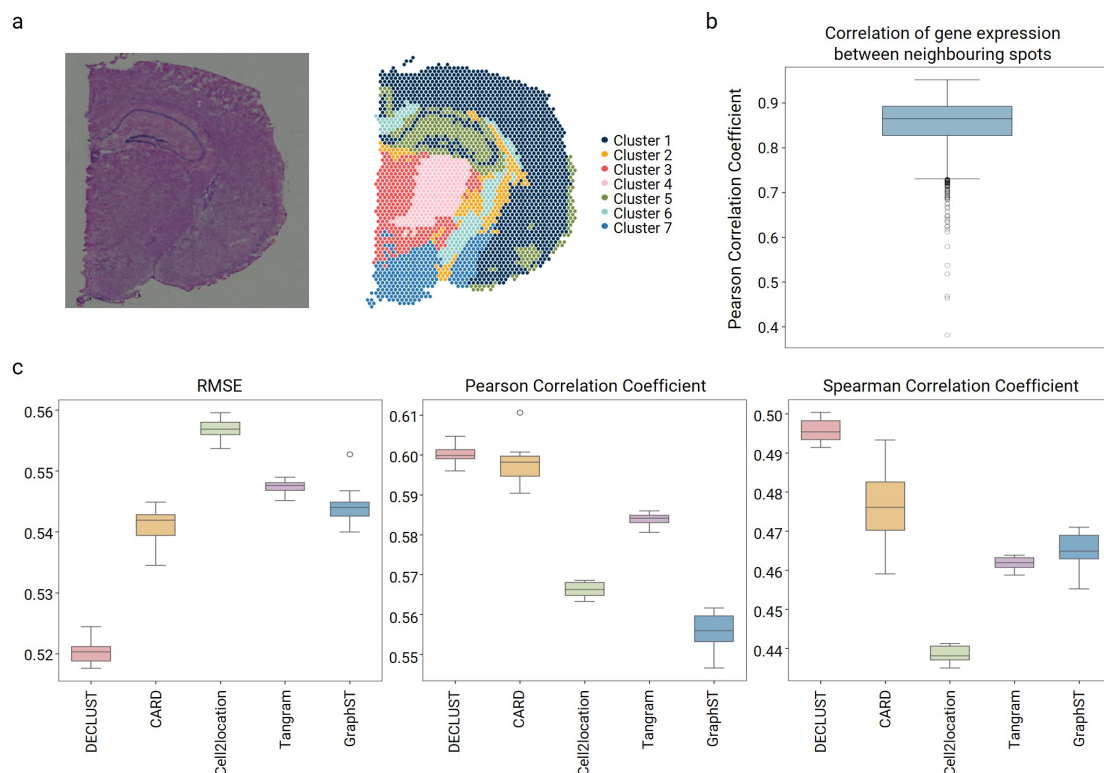


Fig. 4.7.: Cell-type deconvolution results for the real mouse brain ST data. (a) DECLUST clustering results reflect the observed structural features of the mouse brain coronal section. For example, cluster 5 accurately captures the region within the dark blue contour of the brain image, corresponding to the hippocampal region; (b) Pearson correlation coefficients between spots and their neighbors. The x-axis presents the spot ID and the y-axis indicates the value of Pearson correlation coefficient; (c) The 10-fold cross-validation results for five cell type deconvolution methods are presented in terms of RMSE, Pearson correlation coefficient, and Spearman correlation coefficient.

blue contour in the brain image at the left panel, is well captured by cluster 5 discovered by the DECLUST’s clustering method at the right panel.

The true cell-type proportions of individual spots are usually not available in a real ST data. However, we expect a high correlation of molecular profiles between spatially neighboring spots; this can be utilized to evaluate deconvolution methods. The neighbor of a spot is one of eight spots spatially adjacent with that spot in the tissue image. Indeed, Figure 4.7b shows that the gene expression profiles of spatially neighboring spots are highly correlated. In this figure, the boxplot represents the distribution of Pearson correlation coefficients between the observed gene expression of a spot and the predicted value, which is calculated by averaging the gene expression of its eight neighboring spots. The average Pearson correlation coefficient across all spots is 0.85, indicating that the molecular profile of

a spot can be well approximated from its neighbors. While this spatial correlation is biologically expected and has also been incorporated into models such as CARD and GraphST, DECLUS_T differs in that it explicitly aggregates spatially coherent spots prior to deconvolution. This strategy improves robustness by enhancing the signal-to-noise ratio and results in smoother transitions across spatial domains, which is reflected in the correlation-based performance comparison. We apply this idea to cell-type proportions as the basis for comparing deconvolution methods: A good deconvolution method should preserve the spatial correlation between neighboring spots.

The results of 10-fold cross-validation are evaluated using three performance metrics: Root Mean Square Error (RMSE), Pearson Correlation Coefficient (PCC), and Spearman Correlation Coefficient (SCC), as shown in Figure 4.7c. Consistent with the simulation study results, DECLUS_T and CARD outperforms the remaining methods in all performance metrics. Specifically, DECLUS_T obtains the median RMSE of 0.525 significantly higher than the follow-up method CARD (median RMSE=0.542, $p\text{-value}=9.3 \times 10^{-13}$). Similar result is observed for SCC, where DECLUS_T has higher the median $r=0.495$ in comparison with $r=0.475$ of CARD ($p\text{-value}=1.2 \times 10^{-5}$). The PCC results of DECLUS_T (median $r=0.60$) are comparable with those of CARD (median $r=0.598$). For the remaining deconvolution methods, Cell2location generally shows the poorest performance compared to Tangram and GraphST.

4.3.4 DECLUS_T effectively preserves spatial pattern of cell-type marker genes in the real human ovarian cancer ST data

We now utilize an ST dataset of a human ovarian cancer tissue [154] to compare the deconvolution methods on their preservation of spatial pattern of cell-type marker genes. Of note, as discussed in the simulation study, cell-type marker genes can serve as proxies for evaluation of the spatial distribution of cell-types. In the absence of ground-truth cell-type proportions at the spot level for real FFPE spatial transcriptomics data, direct quantitative assessment of deconvolution accuracy is not feasible. Instead, we adopt a proxy-based evaluation strategy by quantifying the spatial concordance between predicted cell-type proportions and the expression patterns of known cell-type marker genes, an approach commonly used in spatial transcriptomics studies when ground truth annotations are unavailable. Among the 22 different cell-types from the single-cell reference dataset of human ovarian cancer [157, 158], we selected three different cell-types for evaluation based on their sizes: tumor cells ($n=5,438$ cells, the most prevalent cell-type), mesangial

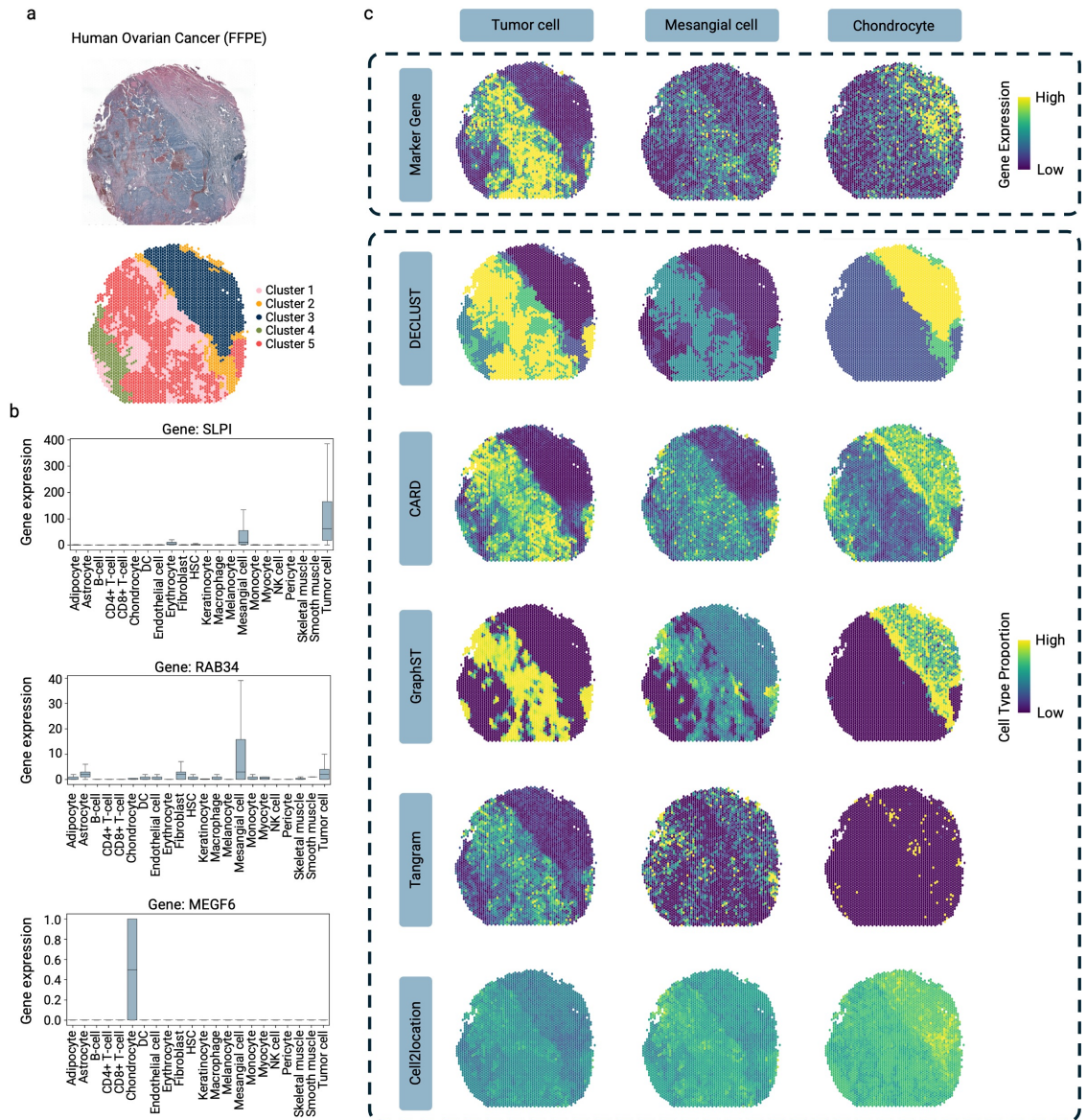


Fig. 4.8.: Cell-type deconvolution results for the real human ovarian cancer ST data. (a) Clustering results of DECLUST for the ST data of a human ovarian cancer tissue (FFPE). The top image shows the original H&E-stained FFPE ovarian cancer tissue, while the bottom panel shows five clusters discovered by DECLUST; (b) Box plots representing the expression levels of the marker genes including SLPI, RAB34, and MEGF6; (c) Spatial distribution of expression of marker genes and cell-type proportions in the human ovarian cancer ST data. The top box shows observed expression of three marker genes: SLPI, RAB34, and MEGF6, the corresponding to the results from DECLUST and other methods for tumor cells, mesangial cells, and chondrocytes in the bottom box.

cells (n=338 cells, a moderately abundant cell-type) and chondrocytes (n=4 cells, a rare cell-type).

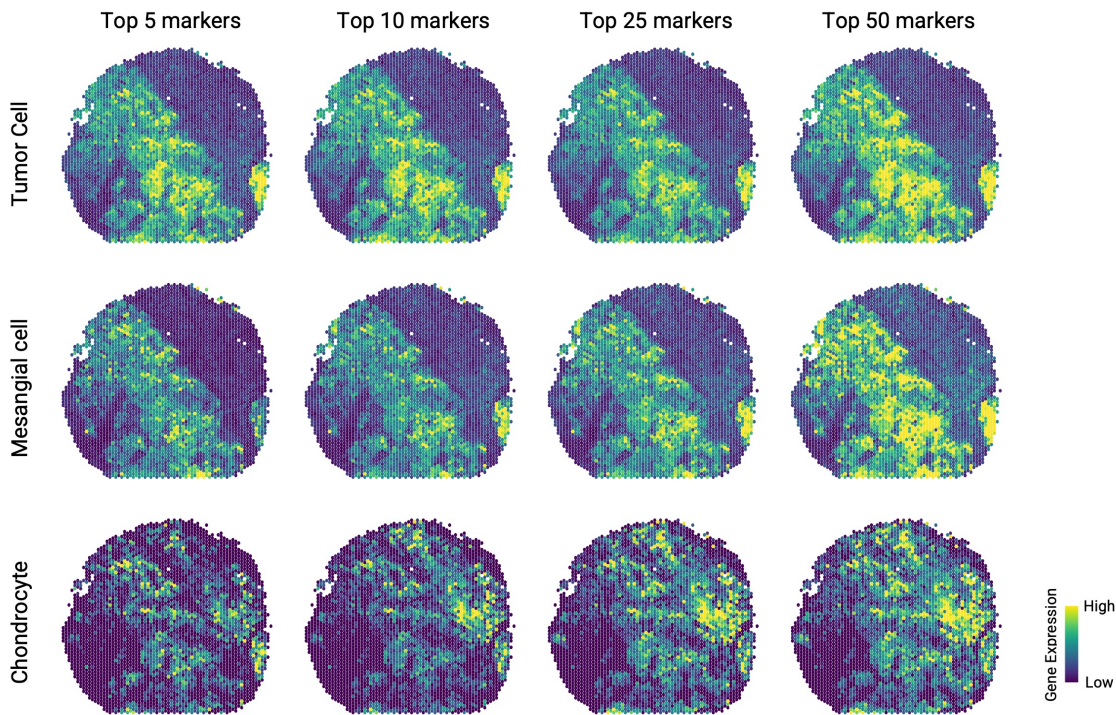


Fig. 4.9.: Spatial expression patterns of marker genes for selected cell types across varying marker set sizes. Heatmaps display the aggregated spatial expression of the top 5, 10, 25, and 50 marker genes for three representative cell types—tumor cells, mesangial cells, and chondrocyte—derived from real spatial transcriptomics (ST) data. Marker genes were ranked and selected based on the t-statistic, which measures differential expression between the target cell type and others. For each marker set size, normalized expression values of the selected genes were summed across spatial spots to produce a composite heatmap. The resulting spatial patterns are consistent across different numbers of markers, demonstrating coherent and robust localization signals for each cell type.

Figure 4.8a shows the image of the human ovarian cancer (top) and distribution of five clusters discovered by DECLUS (bottom). Figure 4.8c displays the spatial pattern of the three selected cell-types from the ovarian cancer tissue, each in one column. For each column, the heatmap in the first row displays the observed gene expression of the single marker gene of the cell-type discovered by DECLUS (Figure 4.8b). The corresponding heatmap at the second row shows the proportion of that cell-type predicted by DECLUS. The high concordance between two heatmaps across three cell-types demonstrates the DECLUS’s ability to accurately capture the spatial patterns of the cell-type marker genes. We further investigated the spatial expression of the top 5, 10, 25, and 50 marker genes for selected cell-types from the real spatial transcriptomics data. For each case, we selected the corresponding number of top marker genes based on prior knowledge, summed up their normalized expression values across spots, and plotted the resulting aggregated expression

as a heatmap. Each heatmap thus represents the spatial expression pattern derived from a distinct set of marker genes. As shown in Figure 4.9, tumor cells, mesangial cells, and chondrocytes exhibit consistent and coherent spatial patterns across different marker set sizes.

Next, we compare the results of DECLUST with other deconvolution methods, which are presented in Figure 4.8c. Specifically, the evaluation focuses on the consistency between predicted cell-type proportions and marker gene expression patterns across multiple cell types with different abundance levels, as well as the robustness of spatial patterns derived from aggregated marker gene sets of varying sizes. In the figure, each row indicates one cell-type, while each column refers to a deconvolution method. As shown in the figure, the cell-type proportions identified by both Tangram and Cell2location do not well reflect the expression of the top marker genes in all cell-types. CARD, GraphST, and Tangram all perform reasonably well for tumor cells, capturing the expression pattern of the marker gene *SLPI*. However, compared to DECLUST, Tangram tends to underestimate the overall proportion of tumor cells. For mesangial cells, GraphST and Tangram incorrectly predict elevated proportions on the right side of the tissue, whereas CARD highlights many spots with a high mesangial cell proportion in the bottom-left corner. For Chondrocyte cells, Tangram fails to detect this cell type, CARD overestimates their abundance in the bottom-left region, and GraphST underestimates their presence in the same area. Thus, DECLUST preserves better the spatial information of the cell-types in the ST data.

4.3.5 Diagnostic Assessment of Residual Heteroscedasticity and Distributional Assumptions

Beyond predictive accuracy and spatial pattern concordance, the stochastic properties of the regression residuals were examined to characterize the empirical error structure associated with the least-squares estimation procedure in DECLUST. All analyses were conducted using the breast cancer spatial transcriptomics dataset.

As model fitting is performed at the cluster level, residual diagnostics were first evaluated at this level of aggregation. Figure 4.10 presents gene-wise mean residuals plotted against mean predicted expression across clusters. Residuals corresponding to low predicted expression are predominantly concentrated near zero. With increasing predicted expression, both the magnitude and dispersion of residuals increase, and a subset of genes exhibits comparatively large positive residual values, indicating systematic under-prediction for certain highly expressed features. These observations indicate a dependence of residual variability on the

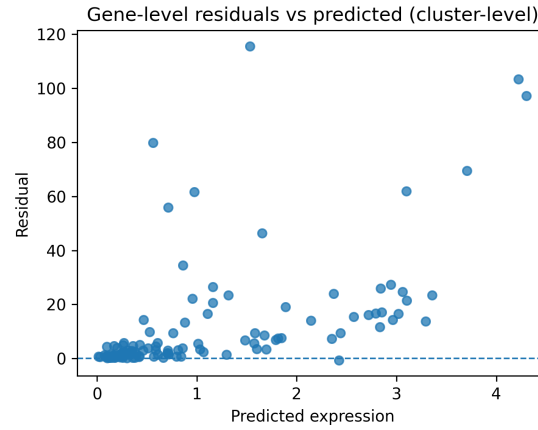


Fig. 4.10.: Gene-wise mean residuals versus predicted expression across clusters. Each data point represents a specific gene evaluated at the cluster level of aggregation.

scale of predicted expression and suggest non-constant variance (heteroscedasticity) across genes.

To further assess the ordinary least squares fit, fitted cluster-level predictions were projected to the spot level. For visualization, residuals were scaled gene-wise by their standard deviation to facilitate cross-gene comparison while preserving systematic bias (i.e., not mean-centered). Figure 4.11 displays histograms and normal Q–Q plots for four representative

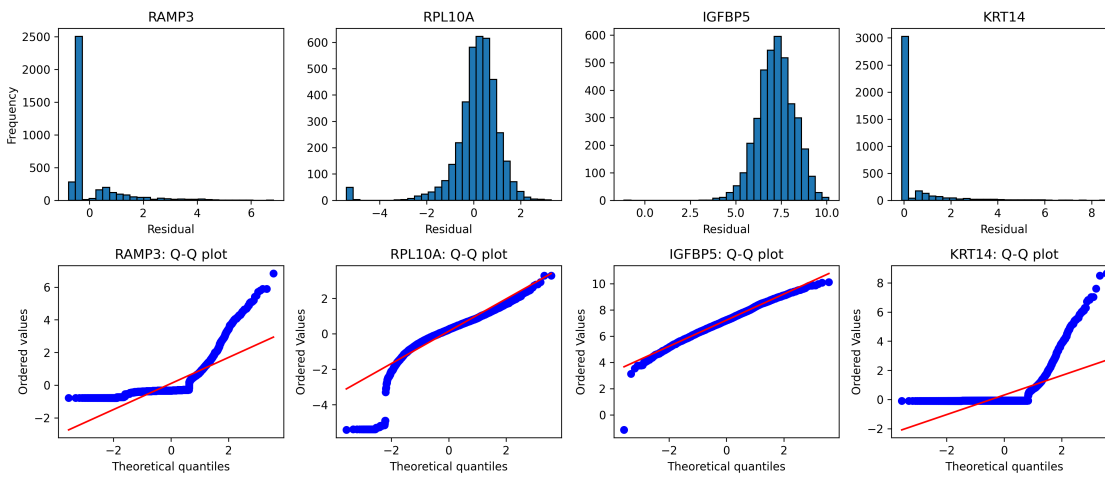


Fig. 4.11.: Spot-level residual distributions and normal Q–Q diagnostics for representative genes. The top row displays residual histograms, and the bottom row shows corresponding normal Q–Q plots. Residuals were scaled by their standard deviation to facilitate cross-gene comparison while preserving systematic bias.

genes (RAMP3, RPL10A, IGFBP5, and KRT14), selected based on residual summary statistics (Table 4.2) to represent distinct distributional regimes.

The empirical residual distributions exhibit substantial heterogeneity:

- **RAMP3 and KRT14:** These genes exhibit pronounced right-skewness and heavy-tailed behavior. The horizontal plateaus in the lower quantiles of the Q–Q plots correspond to zero-valued observations in the spatial data, where residuals are constrained by the predicted intensities.
- **RPL10A:** As a highly expressed gene, it shows significant tail departures with negative skewness (-2.257), indicating localized over-prediction despite its high mean expression.
- **IGFBP5:** The Q–Q plot indicates that the residuals are approximately normally distributed, with most points closely following the reference line, except for a single outlier in the lower tail.

Overall, the residual diagnostics reveal scale-dependent variability and gene-specific deviations from normality at the spot level. While ordinary least squares provides unbiased point estimates under mild regularity conditions, the observed heterogeneity suggests that weighted least squares (WLS) formulations may offer improved efficiency by accounting for gene-dependent variance structures. Incorporating weights derived from empirical variance estimates could mitigate the impact of heteroscedasticity and enhance robustness in future developments of the DECLUST framework.

Tab. 4.2.: Gene-level residual summary statistics for representative genes used in spot-level diagnostics.

Gene	Pred. Mean	Resid. Mean	Resid. Var.	Skewness	Kurtosis	Selection Rationale
RAMP3	0.120	0.033	0.104	2.557	7.202	Strong positive skewness; heavy-tailed with zero-inflation.
RPL10A	2.166	0.061	0.167	-2.257	10.328	High expression with pronounced negative tail deviations.
IGFBP5	0.575	3.092	0.184	-0.301	1.261	Systematic positive residual shift with near-normal shape.
KRT14	0.027	0.108	0.125	3.565	14.643	Extreme skewness and kurtosis driven by sparse expression.

4.3.6 DECLUST exhibits computational efficiency in cell-type deconvolution

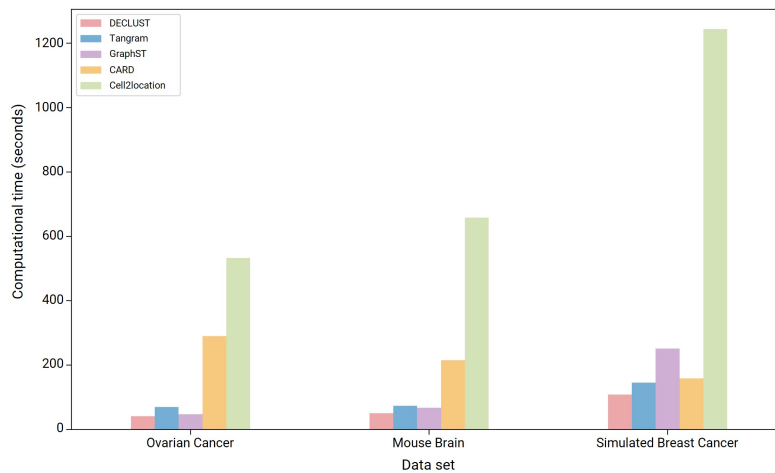


Fig. 4.12.: Computational time of the deconvolution methods in three ST datasets. The x-axis shows the results of each dataset. The y-axis presents computational time in seconds.

To compare the computational time between deconvolution methods, we measure the time starting from data inputting until achieving the cell-type proportions of individual spots. All methods are implemented on a server equipped with an Intel Core i7-12700H CPU and an NVIDIA RTX 3080Ti Laptop GPU. We remove parallel computation for all methods to measure their computational usage. Figure 4.12 displays the computational time (by seconds) of five methods for one simulated ST dataset and two real ST datasets. The details are provided in Table 4.3. For the real ST datasets, DECLUST, Tangram, and GraphST require relatively short computational time.

Tab. 4.3.: Comparison of five deconvolution methods by computational time (seconds).

	DECLUST	CARD	Cell2location	Tangram	GraphST
Breast Cancer	108.16	158.50	1242.88	145.13	251.25
Mouse Brain	50.39	214.75	657.82	73.26	67.65
Ovarian Cancer	41.19	289.75	532.25	69.82	46.96

Although the difference in computational time between DECLUST and GraphST on the real dataset appears modest, we emphasize that DECLUST consistently achieves lower runtime across all datasets. In addition, DECLUST does not rely on GPU acceleration or deep learning frameworks, making it a more lightweight and interpretable solution. Of note, most methods including GraphST, Tangram, and Cell2location require GPUs

for their computation. For the simulated ST dataset, Cell2location still has the highest computational cost, while the other methods show favorable computational efficiency.

4.4 Discussion

Recent advances in spatial transcriptomics have significantly enhanced our understanding of cell-type composition within tissue architecture [170]. However, due to technical constraints and high costs, most spatial transcriptomics (ST) data currently lack single-cell resolution, hindering the exploration of the cellular composition at specific tissue locations [171, 172]. We have developed DECLUST, an efficient cluster-based cell-type deconvolution of spatial transcriptomic data. We have shown that DECLUST performs well against existing ST deconvolution methods in both simulated and real datasets.

A key advantage of DECLUST lies in its ability to simultaneously cluster spatial data based on both gene expression profiles and spatial coordinates of spots. This spatial clustering approach preserves the spatial structure of the spot clusters. In contrast, conventional clustering methods, such as Seurat [173], and the Louvain method [174] primarily rely on gene expression data alone. As a result, the domains identified by these methods tend to be discontinuous, since they do not incorporate spatial information to recognize co-localized cells that might belong to the same domain. While GraphST improves upon this limitation by introducing spatial proximity through graph-based representation learning, which refines spatial clustering, it still faces challenges. Specifically, smooth boundaries are not always ensured, and the method demands longer computational times due to the need for multiple runs.

Furthermore, by aggregation of gene expression of spots in the same cluster for deconvolution, DECLUST can perform on the higher gene-expression signals as compared to the low gene expression levels of individual spots, consequently resulting in high accuracy. Another unique characteristic of DECLUST is the approach to identify cell-type specific marker genes. DECLUST aims to detect marker genes from scRNA-seq reference data that their expression is significantly upregulated in one cell-type while not statistically different in the remaining cell-types. Intuitively, this could decrease the noise for the deconvolution model, subsequently improving the estimation of cell-type composition.

In practice, the performance of DECLUST is influenced by the choice of clustering-related parameters, as these parameters determine the granularity and spatial coherence of the

resulting spot clusters. Parameters associated with hierarchical clustering, DBSCAN-based spatial sub-clustering, and the seeded region growing procedure may affect cluster size, boundary smoothness, and consequently the aggregated gene expression profiles used for deconvolution. DECLUST is nevertheless designed to mitigate excessive sensitivity to individual parameter choices through its multi-stage clustering strategy. By integrating transcriptional similarity with spatial continuity and allowing iterative refinement of cluster assignments, the framework reduces the impact of local parameter perturbations on global clustering structure. As a result, moderate variations in parameter values tend to produce incremental changes in cluster boundaries rather than drastic shifts in cluster composition, supporting stable downstream deconvolution performance. To facilitate practical use, recommended default parameter settings are provided. In this thesis, the number of initial clusters is determined using the elbow criterion, while DBSCAN parameters $\epsilon = 4$ and $\text{minPts} = 8$ are adopted to identify spatially coherent sub-clusters. These settings were empirically found to offer a reasonable balance between spatial smoothness and biological resolution for common 10x Visium datasets. While these defaults are not intended to be universally optimal, they serve as a practical starting point, and users may adjust parameters based on tissue heterogeneity, spot density, and specific analytical goals.

Besides its advantages, DECLUST also has some limitations. For example, similar to other traditional deconvolution methods using the reference from single-cell data, DECLUST cannot estimate the cell-types not existing in the single-cell reference. Careful selection of an appropriate scRNA-seq reference is therefore important for achieving accurate decomposition. To broaden applicability when matched tissue-specific scRNA-seq references are unavailable, a standardized reference signature derived from large publicly available scRNA-seq resources can also be provided. For example, a Tabula Sapiens-based reference covering 180 annotated cell types has been generated and released for public use (https://drive.google.com/file/d/1W3s9leAINu3huTyLUKafMf8zss5rwKFS/view?usp=drive_link). Nevertheless, constructing a study-specific reference is recommended whenever possible, because feature selection, particularly the identification of highly variable genes, can be tailored to the characteristics of the target spatial transcriptomics dataset, which may improve deconvolution accuracy. Furthermore, all analyses in this study are based on the ST data generated by 10x Visium, the most commonly used ST technology in the field.

The aggregation can potentially blur fine-grained biological variation across individual spots, and this trade-off between computational robustness and spot-level resolution must be carefully considered. However, the aggregation can still be a correct and biologically

reasonable with a proper aggregation strategy to group spots with similar biological profiles together, as what DECLUST is aiming for. In DECLUST, spots are not randomly aggregated, but they are already similar based on both spatial and expression features. Gene expression carries the biological information of spots, but individual spots often suffer from extremely sparse expression in spatial transcriptomics. Many genes have zero counts or near-zero, not because they are truly absent, but because of technical limitations such as capture efficiency and sequencing depth. Aggregating across biologically similar spots increases the signal-to-noise ratio, making it possible to detect real patterns. Furthermore, neighboring spots in ST often share highly similar biological profiles, because of, for example, coherent tissue structures such as a cortex layer or tumor clone, etc. Therefore, aggregation of these reflects true biology rather than distorting it.

Although the clustering step in DECLUST is currently designed to enhance deconvolution by incorporating spatial coherence, the resulting spatial domains may also serve as a basis for biological interpretation. This is conceptually related to the approach adopted by spacedeconv [175], which utilizes clustering to define and interpret tissue niches. While we have not explicitly assessed the biological relevance of the clusters in this study, future investigations may explore whether the identified clusters reflect tissue microenvironments or spatially organized functional regions. Such exploration could expand the utility of DECLUST beyond deconvolution. Future extensions of DECLUST could also incorporate additional information or modalities beyond gene expression, such as ligand–receptor interactions, spatially inferred cell–cell communication networks, epigenomic profiles, spatial proteomics, and histology-derived image features. These enhancements may further improve clustering accuracy and help capture complex patterns of intercellular communication and tissue microenvironment heterogeneity.

In summary, we have introduced DECLUST, a highly accurate, efficient, and robust tool for cell-type deconvolution from spatial transcriptomics. The results from the simulated and real ST datasets show that DECLUST performs well against other competing methods in both prediction accuracy and preservation of spatial structure in ST data.

A Comparative Study of Predictive Models for Spatial Protein Abundance from Transcriptomic Profiles

Proteins are central effectors of cellular function, and their spatial distribution provides critical information on tissue organization, cell-cell communication, and disease pathology. Understanding and predicting spatial protein abundance is therefore a key challenge in computational biology, particularly in the context of spatial transcriptomics data. This chapter presents a comparative study of 11 models of protein abundance prediction from transcriptomic profiles. The work originates from a manuscript in preparation, titled “A Comparative Study of Predictive Models for Spatial Protein Abundance”, first authored by Qingyue Wang, and intended for submission to *Nature Communications*. The manuscript is reproduced here in full to preserve its integrity and demonstrate the contribution of the candidate to peer-reviewed research.

Abstract

Proteins are central effectors of cellular function, and their spatial distribution yields critical insights into tissue organization, cell–cell communication, and disease pathology. However, spatial proteomics assays remain costly and technically challenging, leading most spatially resolved datasets to contain only RNA measurements. Computational prediction of protein abundance from transcriptomic profiles offers a cost-effective alternative, yet is hindered by spatial heterogeneity and the low abundance of both genes and proteins at the spot level.

This study presents a systematic evaluation of predictive models for estimating spatial protein abundance from spatial transcriptomics data. Eleven models, spanning linear regression, tree-based ensemble methods, and graph neural networks (GNNs), are assessed using two in-house and four public spatial multi-omics datasets from different tissues, generated by Visium CytAssist CITE-seq. The GNN architectures include Graph Attention Networks (GAT), GraphSAGE, and the more recent Dual Graph Attention Network (DGAT),

which jointly leverages feature and spatial graphs to enhance neighborhood modeling. Three validation strategies, including within-sample, cross-sample, and cross-tissue, and multiple performance metrics such as Spearman correlation, Pearson correlation, and root mean squared error are employed to assess model robustness and generalizability.

Results show that spatially augmented tree-based ensemble models, LightGBM, XGBoost, and CatBoost, consistently outperform non-spatial baselines across all evaluation settings, achieving median Spearman correlations above 0.6 in most datasets. In contrast, GNN models including GAT, GraphSAGE, and DGAT exhibit lower and more variable performance, indicating limited benefit from local neighborhood aggregation in the current framework.

This work provides the first large-scale benchmark of spatial protein prediction methods, delivering practical guidance for model selection and emphasizing the importance of spatial-feature integration. The findings demonstrate the potential of computational approaches to derive spatial protein abundance from spatial RNA expression.

5.1 Introduction

Proteins are the functional executors in cells and more directly reflect biological activity than RNA [176, 177]. Their spatial distribution determines cellular roles and interactions within the tissue microenvironment, while abnormal protein expression or localization serves as a key biomarker for disease subtype classification, immune profiling, and therapeutic stratification [178, 179]. Spatial proteomics technologies, such as Visium CytAssist [154] and CODEX [129], enable direct mapping of protein abundance in situ, but these methods remain costly, technically demanding, and low throughput. As a result, the majority of spatial omics datasets contain only transcriptomic measurements, leaving protein-level insights inaccessible for most available data.

Predicting protein abundance from RNA expression offers a practical and cost-effective strategy to expand the utility of spatial datasets. Such an approach can enable retrospective protein profiling in existing RNA-only datasets and facilitate integrative multi-omics analyses. However, the RNA–protein relationship is complex and influenced by multiple regulatory layers, including transcriptional and translational regulation, post-translational modifications, and protein degradation, leading to substantial RNA–protein mismatches [180, 181]. Spatial heterogeneity further complicates this relationship, as protein abundance is modu-

lated by local microenvironmental factors and cell–cell interactions[182, 183, 184]. Finally, spot-level RNA expression and protein abundance are usually low and subject to noise and high variability due to the small number of cells in each spot, challenging the prediction task.

A recent study proposes *sclinear*, a linear regression–based model for predicting protein abundance at single-cell resolution from non-spatial RNA expression datasets [139]. This model shows promising performance in capturing transcriptome–proteome relationships. Although it can be applied to spatial datasets, it disregards spatial context during training, limiting its ability to capture tissue-level patterns and microenvironmental dependencies intrinsic to spatial transcriptomics. Consequently, directly applying it to spatial data may lead to the loss of valuable spatial information and reduced prediction accuracy.

Graph neural networks (GNNs) have emerged as powerful tools for spatial omics analysis because they naturally encode spatial relationships through graph-based representations[61, 185, 186]. By modeling spatial spots as nodes connected by proximity-based edges, architectures such as Graph Attention Networks (GAT), GraphSAGE, and more recently Dual Graph Attention Networks (DGAT)[141] can integrate transcriptomic profiles with spatial context. In particular, DGAT extends the conventional attention mechanism by jointly leveraging both a feature graph and a spatial graph, thereby providing a richer representation of the underlying tissue microenvironment. Although GNNs have achieved notable success in spatial domain identification, cell–cell interaction analysis, and multimodal integration, their direct application to RNA-to-protein prediction in spatial datasets remains underexplored.

Beyond GNNs, tree-based ensemble models—particularly gradient boosting frameworks such as XGBoost, LightGBM, and CatBoost are widely recognized for their scalability and predictive power in structured data. However, conventional implementations of these models lack mechanisms to directly incorporate spatial information, which will be taken into account in this study.

Here, we present a systematic evaluation study of predictive models for spatial protein abundance using six spatial multi-omics datasets. We evaluate eleven model configurations: partial least squares regression (PLS)[187], three gradient-boosting ensemble models including XGBoost [188], LightGBM [189], CatBoost [190, 191], and two GNN architectures including GAT [192], GraphSAGE [193], DGAT [141], each tested both with and without spatial neighborhood augmentation. Model performance is systematically assessed using

various performance metrics such as Spearman correlation, Pearson correlation and root mean squared error, under three biologically relevant evaluation settings: (1) within-sample validation, (2) cross-sample validation within a tissue type, and (3) cross-tissue validation across distinct tissues. This benchmark provides quantitative insights into the role of spatial modeling in protein abundance prediction and offers practical guidance for future spatial multi-omics analyses.

5.2 Materials and methods

5.2.1 Overview of the benchmarking framework

Figure 5.1 presents the overall workflow for evaluating methods that predict spatial protein abundance from spatial transcriptomic profiles. Each spatial multi-omics dataset comprises n spatial locations (spots), where each spot is associated with: (1) a gene expression vector of dimension g , (2) a protein abundance vector of dimension p , and (3) a two-dimensional coordinate indicating its spatial position within the tissue section. The goal of these methods is to estimate protein abundance at spot level based on the corresponding gene expression data.

We denote the gene expression matrix as $\mathbf{X}$ ($n \times g$) and the protein abundance matrix as $\mathbf{Y}$ ($n \times p$). For each protein $i \in \{1, \dots, p\}$, we train an independent predictive model to learn the mapping $f^{(i)} : \mathbf{X} \rightarrow \mathbf{Y}^{(i)}$, where $\mathbf{Y}^{(i)}$ denotes the abundance of protein i across spatial locations.

5.2.2 Prediction models for evaluation

We select six prediction methods across three different categories including (1) a linear regression, (2) three tree-based ensemble methods and (3) two graph neural network (GNN) methods for the evaluation. Since the GNN methods inherently integrate neighborhood information into their models, we apply two settings - spatial and non-spatial - only to the four methods in the first two categories. Thus, results of a total of ten method configurations are reported. We summarize the three categories of predictive models below, with further details of each model and its training procedures.

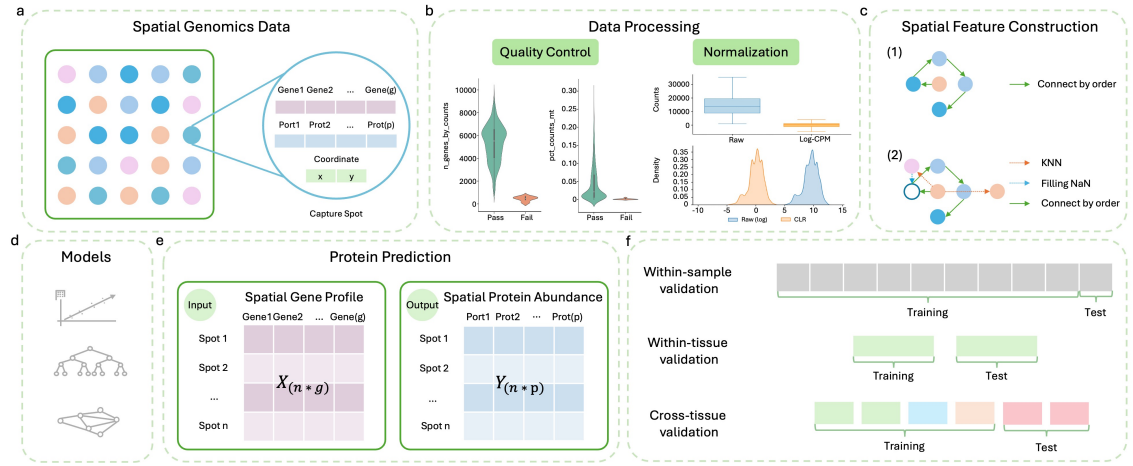


Fig. 5.1.: Overview of the benchmarking framework for methods predicting spatial protein abundance from spatial RNA expression profiles. (a) Spatial multi-omics datasets, comprising paired spatial transcriptomic (gene expression) and proteomic (protein abundance) measurements at matched spatial coordinates (spots). (b) Data preprocessing involves two main steps: (i) quality control, where low-quality spots are removed if fewer than 700 genes are detected or if mitochondrial UMIs exceed 30%, and genes detected in fewer than 2.5% of spots are excluded; and (ii) normalization, where RNA expression matrices are normalized to 10,000 UMIs per spot and log-transformed, while protein abundance (antibody-derived tag counts) is normalized using the centered log-ratio (CLR) transformation. (c) Spatial feature construction to incorporate spatial context into model inputs, including ordered neighbor connections, k-nearest neighbor (KNN) relationships, and missing value imputation. (d) Candidate predictive models encompass three categories: linear regression models, tree-based boosting models, and graph neural network (GNN) architectures. (e) Model inputs consist of spatial gene expression matrices $X_{n \times g}$, where n denotes the number of spots and g the number of genes, and outputs correspond to spatial protein abundance matrices $Y_{n \times p}$, where p denotes the number of proteins. (f) Model evaluation is conducted under three validation strategies: within-sample validation, cross-sample validation, and cross-tissue validation. Different colors denote different tissue types.

5.2.2.1 Linear regression

We use Partial Least Squares Regression (PLS) [187], which integrates principal component analysis with multiple linear regression. PLS is well suited for high-dimensional omics data due to its ability to handle multicollinearity effectively [194]. We build a separate PLS model for each protein for the prediction.

PLS

PLS models the relationship between the RNA expression matrix $\mathbf{X}$ ($n \times g$) and each protein abundance vector $\mathbf{Y}^{(i)}$ by projecting both predictors and responses into a shared latent space:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\top + \mathbf{E}, \quad \mathbf{Y}^{(i)} = \mathbf{T}\mathbf{q}^\top + \mathbf{F} \quad (5.1)$$

where $\mathbf{T}$ contains latent components that maximize the covariance between $\mathbf{X}\mathbf{w}$ and $\mathbf{Y}^{(i)}$, subject to $\|\mathbf{w}\| = 1$. The regression coefficients are computed as:

$$\hat{\boldsymbol{\beta}}_{\text{PLS}} = \mathbf{W}(\mathbf{P}^\top \mathbf{W})^{-1} \mathbf{q}, \quad \hat{\mathbf{Y}}^{(i)} = \mathbf{X} \hat{\boldsymbol{\beta}}_{\text{PLS}} \quad (5.2)$$

A separate PLS model was trained for each protein using 5 latent components.

5.2.2.2 Tree-based ensemble models

Tree-based gradient boosting models are powerful non-linear predictors that combine multiple decision trees through boosting [195, 196] that have been widely adopted in the bioinformatics field in recent years [197, 198, 199]. Three methods including XGBoost [188], LightGBM [189], and CatBoost [190, 191] are selected for evaluation as they are commonly used with computational efficiency and advanced performance in multiple general tasks. All models are trained using the mean squared error (MSE) loss to quantify the discrepancy between predicted and observed protein abundance.

All tree-based ensemble boosting models used in this study follow a shared formulation:

$$\hat{y}_j^{(i)} = \sum_{k=1}^K f_k(\mathbf{x}_j), \quad f_k \in \mathcal{F} \quad (5.3)$$

where f_k are regression trees, and $\mathbf{x}_j$ is the input for spot j . Model optimization minimizes the regularized squared error:

$$\mathcal{L}^{(i)} = \sum_{j=1}^n (y_j^{(i)} - \hat{y}_j^{(i)})^2 + \sum_{k=1}^K \Omega(f_k), \quad \Omega(f_k) = \gamma T_k + \frac{1}{2} \lambda \|\mathbf{w}_k\|^2 \quad (5.4)$$

All boosting models were trained separately for each protein with default hyperparameters, using a fixed random seed and full parallelism enabled.

XGBoost

XGBoost employs a second-order Taylor expansion of the loss and sparse-aware split finding. The objective at each iteration is approximated as:

$$\mathcal{L}^{(t)} \approx \sum_{j=1}^n \left[g_j f_t(\mathbf{x}_j) + \frac{1}{2} h_j f_t^2(\mathbf{x}_j) \right] + \Omega(f_t) \quad (5.5)$$

where g_j and h_j are the first and second derivatives of the loss function. The model was trained using the squared error loss.

LightGBM

LightGBM constructs trees using a leaf-wise growth strategy that selects splits with the largest loss reduction. It employs histogram-based feature binning and gradient-based one-side sampling (GOSS) to reduce training complexity and accelerate convergence. Default parameters were used with squared error loss.

CatBoost

CatBoost builds symmetric trees and uses permutation-driven ordered boosting to avoid overfitting and target leakage. While originally designed for categorical features, it also performs competitively on fully numerical input. The model was trained with root mean squared error (RMSE) as the objective and default parameter settings.

5.2.2.3 Graph neural network models

Graph Neural Networks (GNNs) [200] extend traditional neural networks to graph-structured data by enabling information exchange across neighboring nodes. This inductive bias makes GNNs particularly suitable for spatial omics analysis, where local tissue architecture and neighborhood context influence molecular phenotypes [201, 202, 87]. We evaluate three GNN variants: Graph Attention Networks (GAT) [192], GraphSAGE [193], and Dual-Graph Attention Network (DGAT) [141]. For each tissue section, we construct

a spatial graph by connecting each spot to its k nearest neighbors in Euclidean space. Node features are gene expression vectors. Models are trained end-to-end (per protein) to minimize mean squared error (MSE) in protein abundance prediction (learning rate = 0.001).

GAT

GAT updates node features via attention-weighted aggregation:

$$\mathbf{h}'_v = \sigma \left(\sum_{u \in \mathcal{N}(v)} \alpha_{vu} \mathbf{W} \mathbf{h}_u \right) \quad (5.6)$$

where the attention weights are:

$$\alpha_{vu} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^\top [\mathbf{W} \mathbf{h}_v \parallel \mathbf{W} \mathbf{h}_u]))}{\sum_{k \in \mathcal{N}(v)} \exp(\text{LeakyReLU}(\mathbf{a}^\top [\mathbf{W} \mathbf{h}_v \parallel \mathbf{W} \mathbf{h}_k]))} \quad (5.7)$$

We use two attention heads per layer, ELU activations, and dropout between layers.

GraphSAGE

GraphSAGE updates node embeddings via sampling-based neighborhood aggregation:

$$\mathbf{h}'_v = \sigma(\mathbf{W} \cdot \text{AGGREGATE}(\{\mathbf{h}_v\} \cup \{\mathbf{h}_u : u \in \mathcal{N}(v)\})) \quad (5.8)$$

Here, AGGREGATE is the mean function, σ is ReLU, and $\mathbf{W}$ is a learnable weight matrix. We stack two aggregation layers, followed by a linear regression head.

DGAT

We implemented DGAT that integrates two heterogeneous graphs—a spatial graph $\mathcal{G}_s$ and an expression graph $\mathcal{G}_e$ —to jointly capture spatial proximity and transcriptomic similarity. For each spot v , the embeddings from the two graphs are updated by separate GAT layers:

$$\mathbf{h}_v^{(s)} = \text{GAT}_s(\mathbf{h}_v, \mathcal{N}_s(v)), \quad \mathbf{h}_v^{(e)} = \text{GAT}_e(\mathbf{h}_v, \mathcal{N}_e(v)), \quad (5.9)$$

where $\mathcal{N}_s(v)$ and $\mathcal{N}_e(v)$ denote the neighbors of v in the spatial and expression graphs, respectively.

The two embeddings are combined via a learnable gating parameter $g \in (0, 1)$:

$$\mathbf{h}'_v = g \cdot \mathbf{h}_v^{(s)} + (1 - g) \cdot \mathbf{h}_v^{(e)}. \quad (5.10)$$

The final latent embedding $\mathbf{z}_v$ is fed into two decoder branches: a protein prediction head and an RNA reconstruction head:

$$\hat{\mathbf{y}}_v^{(\text{prot})} = f_{\text{prot}}(\mathbf{z}_v), \quad (5.11)$$

$$\hat{\mathbf{x}}_v^{(\text{rna})} = f_{\text{rna}}(\mathbf{z}_v), \quad (5.12)$$

where f_{prot} is a regression network outputting protein abundances, and f_{rna} reconstructs the original gene expression vector to regularize the latent space.

The overall training objective is a weighted combination of protein prediction loss and RNA reconstruction loss:

$$\mathcal{L} = \mathcal{L}_{\text{prot}} + \lambda_{\text{recon}} \mathcal{L}_{\text{rna}}, \quad (5.13)$$

with $\mathcal{L}_{\text{prot}} = \text{MSE}(\hat{\mathbf{y}}^{(\text{prot})}, \mathbf{y}^{(\text{prot})})$ and $\mathcal{L}_{\text{rna}} = \text{MSE}(\hat{\mathbf{x}}^{(\text{rna})}, \mathbf{x}^{(\text{rna})})$.

5.2.3 Construction of spatial feature for non-GNN models

While GNNs naturally incorporate spatial information through message passing over graphs, traditional prediction models such as PLS, XGBoost, LightGBM, and CatBoost do not account for spatial context unless explicitly encoded in the input features. This limitation leads to the loss of valuable spatial information embedded in tissue architecture, ultimately degrading prediction performance on spatial multi-omics data [203].

In this study, we construct spatial features for all non-GNN models (with spatial setting) by incorporating transcriptomic information from each spot's local neighborhood. Specifically, for each spot, we identify its four fixed neighbors along cardinal directions: west and east neighbors (horizontal axis), and north and south neighbors (vertical axis). These neighbors are identified using exact coordinate offsets from the target spot. If one or more neighbors

are missing due to irregular tissue shape or boundary effects, spatially closest unselected neighbors (based on Euclidean distance) are selected instead.

When using the spatial setting for non-GNN models, the RNA expression input of a spot consists of its original gene expression vector concatenated with the vectors from its four selected neighbors, arranged in a fixed spatial order: west, north, east, and south. This fixed ordering preserves the spatial configuration in the expanded feature representation, allowing models to implicitly capture directional spatial dependencies. This strategy enables traditional machine learning models to incorporate spatial structure in a systematic and reproducible way without relying on an explicit graph representation. The same spatial feature construction process is applied uniformly to both the training and testing datasets.

5.2.4 Validation strategies for the prediction models

We deploy three cross-validation approaches including within-sample validation, cross-sample validation, and cross-tissue validation, on the spatial multi-omics samples (see section 5.2.6) to assess the accuracy, robustness, and generalizability of the prediction methods.

5.2.4.1 Within-sample validation

Within-sample validation evaluates a model on the same data used for training, providing a measure of fit but often overestimating performance due to the lack of independence between training and test sets. This within-sample cross-validation approach is applied to individual samples. To ensure the preservation of spatial structure during model evaluation, we implement a spatially aware 10-fold cross-validation scheme. Specifically, k-means clustering is applied to the two-dimensional spatial coordinates of all spots of the sample to obtain 10 spatially contiguous groups. This ensures that each fold corresponds to a spatially localized region within the tissue. K-means clustering also preserves approximately equal-sized folds, facilitating balanced evaluation across all cross-validation splits. In each iteration, one spatial cluster is held out as the test set, while the remaining nine clusters serve as the training set.

5.2.4.2 Cross-sample validation

Cross-sample validation tests generalizability within the same tissue type by training on one biological sample and testing on another (from the same tissue), capturing variability from

technical and sample-specific differences. This approach is applied for the human tonsil and HNSC datasets (see section 5.2.6). Particularly, for human tonsil datasets, Tonsil 1 is used for training and Tonsil 2 for testing, then the roles are reversed. The same procedure is applied to the HNSC datasets. This procedure guarantees that all data samples are involved in the training and evaluation.

5.2.4.3 Cross-tissue validation

Cross-tissue validation challenges further the generalizability of the methods by training and testing across samples from different tissue types, assessing robustness to broader biological diversity and revealing the model's ability to handle unseen tissue contexts. This approach is applied to all six spatial multi-omics samples. In each round, all samples from healthy or diseased tissues were designated as the testing set, and the remaining type of tissues were used for training.

5.2.5 Performance metrics for evaluation

To quantitatively compare model performance across evaluation settings, three complementary metrics are computed for each protein-specific model: Spearman correlation, Pearson correlation, and root mean squared error (RMSE). These metrics are calculated between the predicted and measured protein abundances across all spots in the test set.

Formally, let $\hat{\mathbf{y}}^{(i)} = \{\hat{y}_1^{(i)}, \dots, \hat{y}_n^{(i)}\}$ denote the predicted abundance vector for protein i from model $f^{(i)}$, and $\mathbf{y}^{(i)} = \{y_1^{(i)}, \dots, y_n^{(i)}\}$ the corresponding measured abundance vector.

Spearman correlation. The Spearman rank correlation coefficient $\rho_s^{(i)}$ is used as the primary metric to assess the monotonic relationship between $\hat{\mathbf{y}}^{(i)}$ and $\mathbf{y}^{(i)}$. It is computed as:

$$\rho_s^{(i)} = 1 - \frac{6 \sum_{j=1}^n \left(r(\hat{y}_j^{(i)}) - r(y_j^{(i)}) \right)^2}{n(n^2 - 1)}, \quad (5.14)$$

where $r(\cdot)$ denotes the rank transformation. Spearman correlation is robust to non-linear dependencies and was reported for all within-sample, cross-sample, and cross-tissue validation experiments.

Pearson correlation. The Pearson correlation coefficient $\rho_p^{(i)}$ measures the linear association between predictions and observations:

$$\rho_p^{(i)} = \frac{\sum_{j=1}^n (\hat{y}_j^{(i)} - \overline{\hat{y}^{(i)}}) (y_j^{(i)} - \overline{y^{(i)}})}{\sqrt{\sum_{j=1}^n (\hat{y}_j^{(i)} - \overline{\hat{y}^{(i)}})^2} \sqrt{\sum_{j=1}^n (y_j^{(i)} - \overline{y^{(i)}})^2}}, \quad (5.15)$$

where $\overline{\hat{y}^{(i)}}$ and $\overline{y^{(i)}}$ denote the sample means. This metric is mainly used in cross-sample and cross-tissue evaluations to assess linear predictive accuracy.

Root mean squared error (RMSE). RMSE quantifies the average magnitude of prediction errors:

$$\text{RMSE}^{(i)} = \sqrt{\frac{1}{n} \sum_{j=1}^n (\hat{y}_j^{(i)} - y_j^{(i)})^2}. \quad (5.16)$$

Lower RMSE values indicate more accurate predictions in absolute terms and provide a complementary perspective to correlation-based metrics.

Statistical significance testing.

To evaluate the effect of incorporating spatial features in non-GNN methods, we compared each base model (non-spatial setting) with the same model in spatial setting using a two-sided Wilcoxon signed-rank test on the per-protein Spearman correlations. For each protein i , the paired difference is

$$z_i = \rho_{s,\text{spatial}}^{(i)} - \rho_{s,\text{base}}^{(i)}, \quad (5.17)$$

and the Wilcoxon statistic is

$$W = \sum_{i=1}^m R_i \cdot \text{sgn}(z_i), \quad (5.18)$$

where R_i is the rank of $|z_i|$ and m is the number of proteins considered.

5.2.6 Datasets

A total of six spatial multi-omics samples from the Visium CytAssist multi-omics platform are used in this study. These include four public tonsil1, tonsil2, breast cancer, and glioblastoma samples obtained from 10x Genomics webpage, and two in-house Head and Neck Squamous Cell Carcinoma (HNSC) samples A1 and D1.

Breast cancer: Formalin-fixed paraffin-embedded (FFPE) human breast cancer tissue (female, 50s) is prepared using the 10x Genomics Visium CytAssist Spatial Gene and Protein Expression technology. Tissue sections of 5 μ m thickness are stained with DAPI, Vimentin, and PCNA for immunofluorescence and imaged on an Olympus slide scanning microscope (20 $\times$ UPLXAPO objective, NA 0.8) with a Hamamatsu Orca FusionBT sCMOS camera. Spatial gene and protein expression libraries are sequenced on an Illumina NovaSeq with targeted depths of 35,000 reads per spot for gene expression and 20,000 reads per spot for protein expression. The dataset comprises 4,169 tissue-covered spots, with a median of 5,473 genes per spot, a mean of 44,009 reads per spot for gene expression, and a mean of 43,038 reads per spot for protein expression.

Glioblastoma: The sample is from a FFPE human glioblastoma tissue (female, 60s) and generated using the same protocol in the breast cancer sample. The dataset comprises 5,756 tissue-covered spots, with a median of 7,629 genes per spot, a mean of 68,906 reads per spot for gene expression, and a mean of 75,377 reads per spot for protein expression.

Tonsil 1: The same protocol is also performed for a FFPE human tonsil tissue (female, 30s). The dataset comprised 4,194 tissue-covered spots, with a median of 8,275 genes per spot, a mean of 111,046 reads per spot for gene expression, and a mean of 72,898 reads per spot for protein expression.

Tonsil 2: Similar to the Tonsil 1 sample, but this one is based on a FFPE human tonsil tissue of a male who is 25 years old. The dataset comprised 4,908 tissue-covered spots, with a median of 8,114 genes per spot, a mean of 75,598 reads per spot for gene expression, and a mean of 34,845 reads per spot for protein expression.

HNSC (A1): Based on a FFPE human head and neck squamous cell carcinoma tissue. The dataset comprised 486 tissue-covered spots, with a median of 6 genes per spot, a mean of 169,346 reads per spot for gene expression, and a mean of 130,024 reads per spot for protein expression.

HNSC (D1): Another FFPE human head and neck squamous cell carcinoma tissue sample. The dataset comprised 966 tissue-covered spots, with a median of 1,137 genes per spot, a mean of 81,833 reads per spot for gene expression, and a mean of 53,410 reads per spot for protein expression.

5.2.6.1 Spatial multi-omics data generation

The spatial multi-omics datasets analyzed in this chapter were generated using the 10x Genomics Visium CytAssist platform, which enables the joint measurement of spatial RNA expression and protein abundance on the same tissue section. This platform extends the standard Visium workflow by incorporating antibody-derived tags, allowing the simultaneous profiling of transcriptomic and proteomic signals while preserving spatial context at the spot level.

In brief, tissue sections were mounted onto Visium slides and processed according to the Visium CytAssist protocol. RNA transcripts were captured via spatially barcoded primers, while protein abundance was quantified using oligonucleotide-labeled antibodies targeting a predefined panel of proteins. Following reverse transcription and library preparation, separate sequencing libraries were generated for RNA and protein modalities. Sequencing reads were processed using the Space Ranger pipeline, which produces spatially resolved gene expression matrices and corresponding protein count matrices for each capture spot.

5.2.6.2 Data Processing

All datasets provide paired RNA expression and protein abundance of each spot. Low-quality spots are removed if they contain fewer than 700 detected genes or if more than 30% of unique molecular identifiers (UMIs) are mitochondrial. Genes detected in fewer than 2.5% of spots are excluded. RNA expression matrices are normalized to 10,000 UMIs per spot, log-transformed, scaled to unit variance per gene, and clipped at a maximum value of 10. Protein abundance (antibody-derived tag, ADT counts) is normalized using the centered log-ratio (CLR) transformation recommended by 10x Genomics. After preprocessing, the dataset contains 18,085 genes and 31 proteins across six samples for downstream analyses.

5.3 Results

5.3.1 Comparison of the prediction methods using within-sample cross-validation

In this section, we apply a spatial 10-fold cross-validation strategy for within-sample validation (details in section 5.2). Briefly, k-means clustering is applied to the spatial

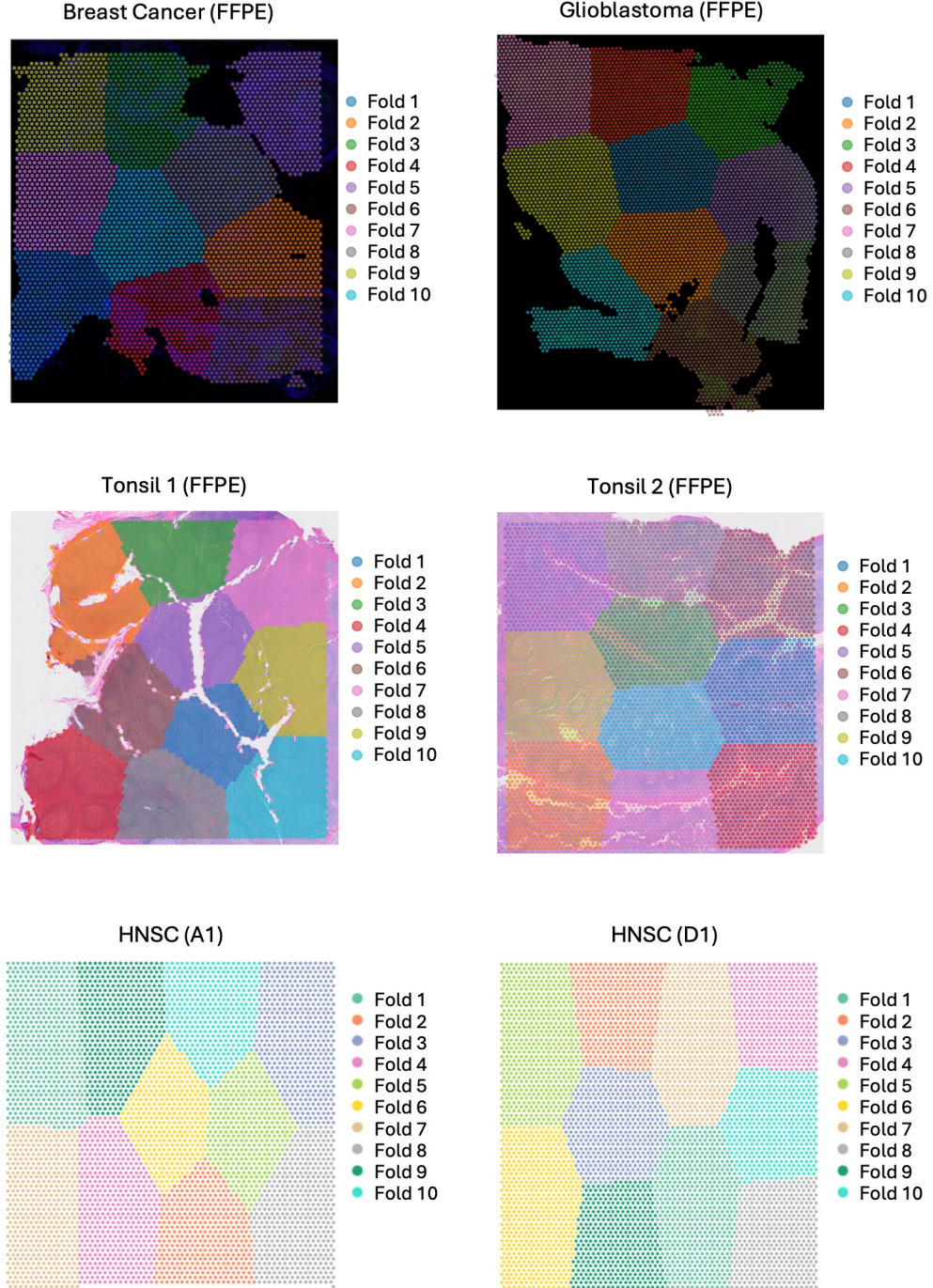


Fig. 5.2.: Spatial 10-fold cross-validation splits for within-sample evaluation. Illustration of the spatial partitioning strategy applied to spatial multi-omics datasets for within-sample evaluation. Capture spots were divided into ten spatially contiguous folds using K-means clustering on two-dimensional spatial coordinates, ensuring spatial separation between training and test sets in each fold. Colors represent fold assignments, with each fold serving once as the test set. This spatially aware design preserves tissue architecture and provides a more realistic estimate of model generalization compared to random splitting.

coordinates of spots to construct 10 folds of spots for the within-sample cross-validation. Fig. 5.2 presents the spatial arrangement of 10 folds for each dataset.

Fig. 5.3 shows that most models achieve moderate to high prediction accuracy across datasets, with spatially enhanced tree-based ensembles—LightGBM-Spatial, XGBoost-Spatial, and CatBoost-Spatial—consistently ranking among the top performers. In HNSC and tonsil samples, non-GNN methods with spatial setting significantly improves Spearman correlation ($p < 0.05$) relative to ones with non-spatial setting, underscoring the value of spatial context in protein abundance prediction.

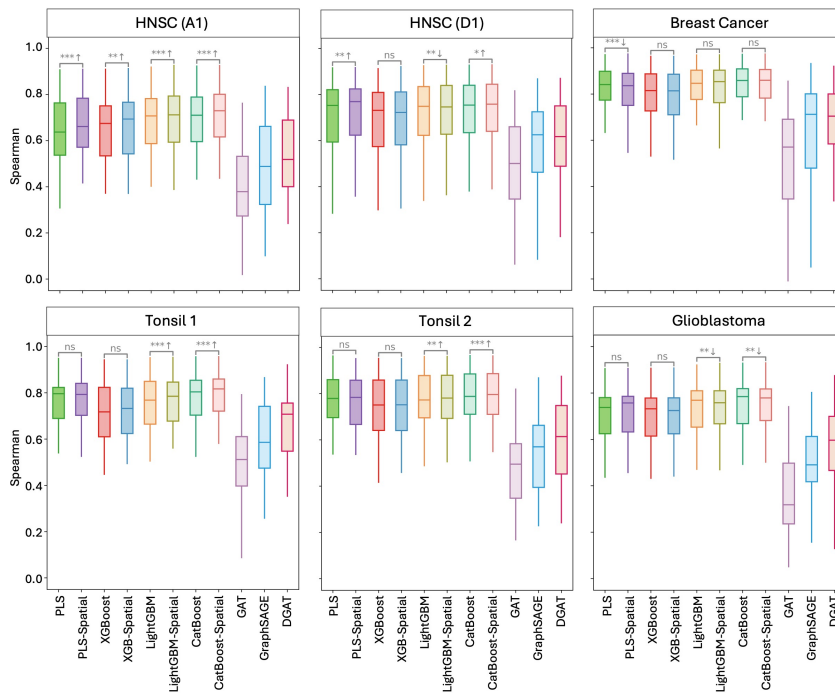


Fig. 5.3.: Protein prediction performance in six spatial multi-omics datasets with within-sample validation (Spearman correlation). Prediction accuracy of ten models for each dataset, evaluated using Spearman correlation between predicted and measured protein abundances. The two-tailed Wilcoxon rank-sum test is used for the difference between spatial setting and non-spatial setting of non-GNN methods ($p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***), *ns* indicating no significant difference). Arrows present the direction of statistical significance that upward (downward) arrows indicate the spatial setting is better (worse) than the non-spatial setting.

Figs. 5.4 and 5.5 present the Pearson correlation and RMSE results, respectively, which show the same trends as the Spearman correlation, further demonstrating the advantage of incorporating spatial context.

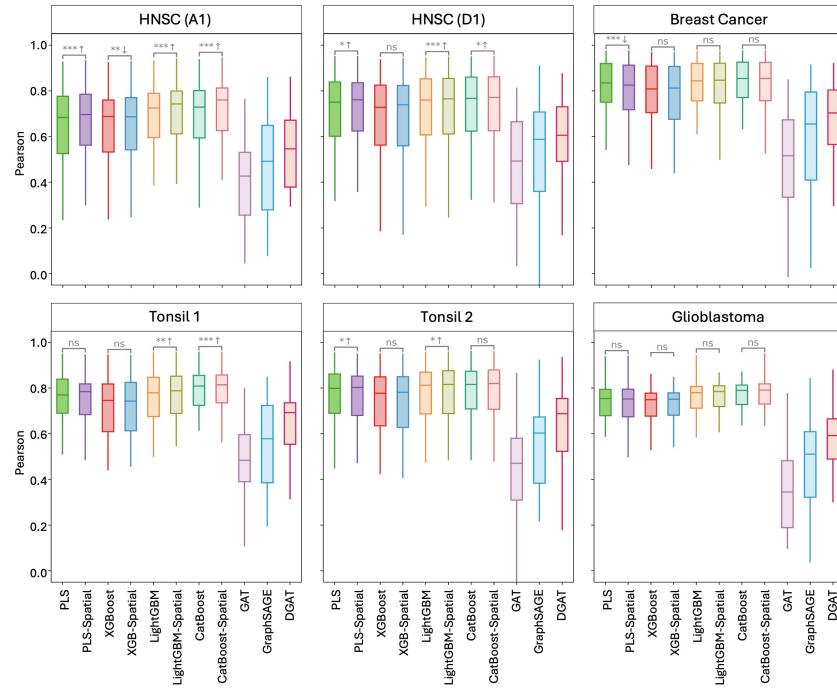


Fig. 5.4.: Protein prediction performance with within-sample validation (Pearson correlation).

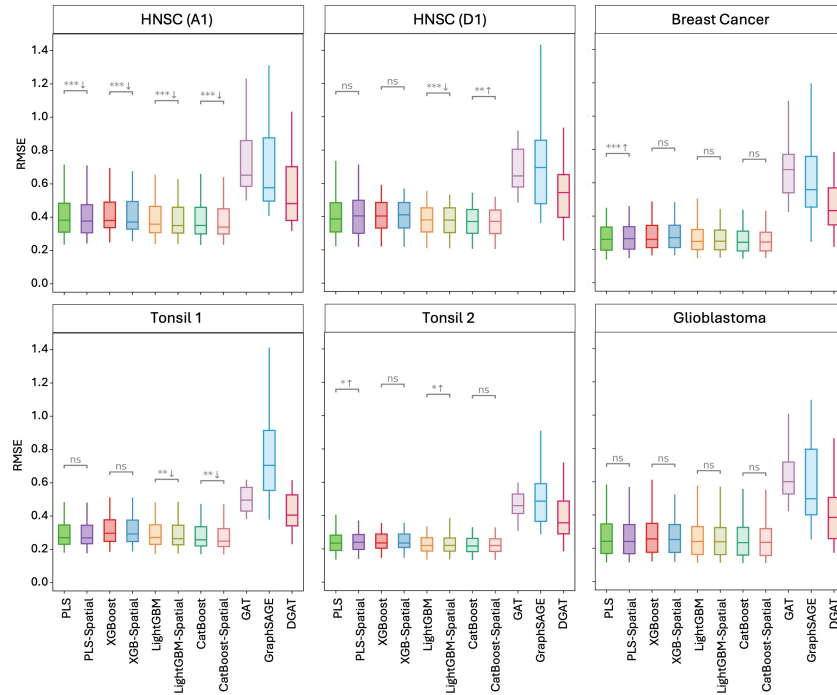


Fig. 5.5.: Protein prediction performance with within-sample validation (RMSE).

From the figure, methods of CatBoost, LightGBM and PLS (both spatial and non-spatial settings) are generally top performing methods, followed by XGBoost methods. GNN-based models (GAT, GraphSAGE and DGAT) show the worst performance with lower median Spearman correlations and larger variance. Furthermore, among the tree-based approaches, CatBoost methods are usually slightly better than LightGBM across datasets.

Protein-level results across all datasets are presented in Fig. 5.6 and Fig. 5.7. Spatially augmented tree-based ensemble models (LightGBM-Spatial, XGBoost-Spatial, CatBoost-Spatial) consistently ranked among the top performers, achieving median Spearman correlations above 0.70 for multiple proteins such as CD68, CD163, HLA-DRA, and CD3E in immune-rich tissues (e.g., tonsil and head and neck cancer). In contrast, proteins with low expression or weak spatial structure, including KRT5 and CEACAM8, exhibited uniformly poor prediction accuracy (median Spearman correlation < 0.30) across all models. Glioblastoma displayed more heterogeneous outcomes, with certain targets (e.g., CD31, PDGFRB) attaining moderate correlations (0.50–0.65), while others remained poorly predicted. Breast cancer datasets demonstrated strong predictive performance for epithelial and immune markers (e.g., EPCAM, CD3E), with several models exceeding a correlation of 0.75. In comparison, graph neural network models (GAT, GraphSAGE and DGAT) showed lower median performance (frequently < 0.50) and greater variance relative to spatially enhanced tree-based methods, particularly in datasets characterized by pronounced spatial clustering of protein abundance.

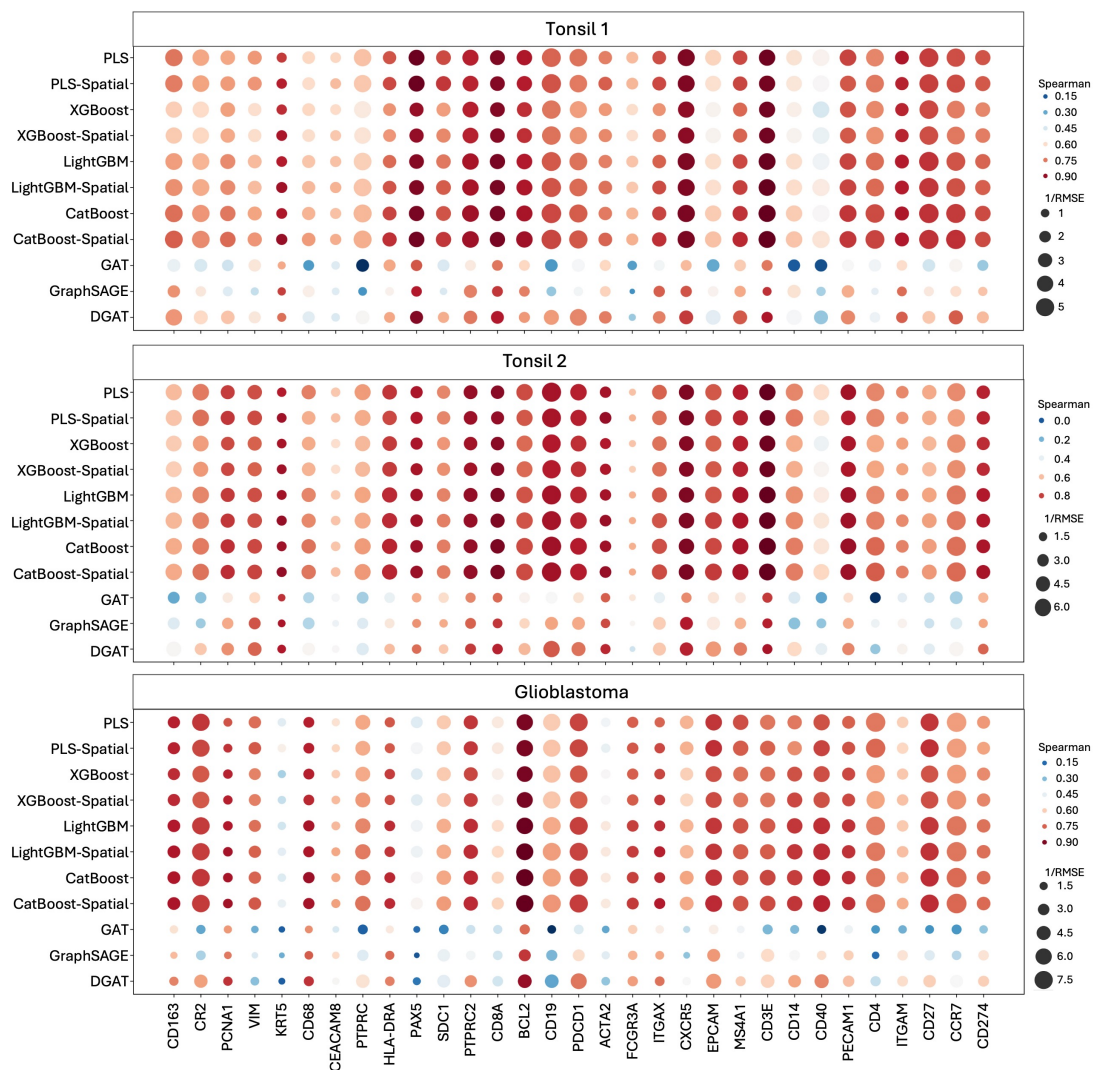


Fig. 5.6: Within-sample prediction performance in Tonsil and Glioblastoma. Protein-wise Spearman correlations for Tonsil 1, Tonsil 2, and Glioblastoma datasets, with dot size indicating inverse RMSE (larger dots denote better accuracy).

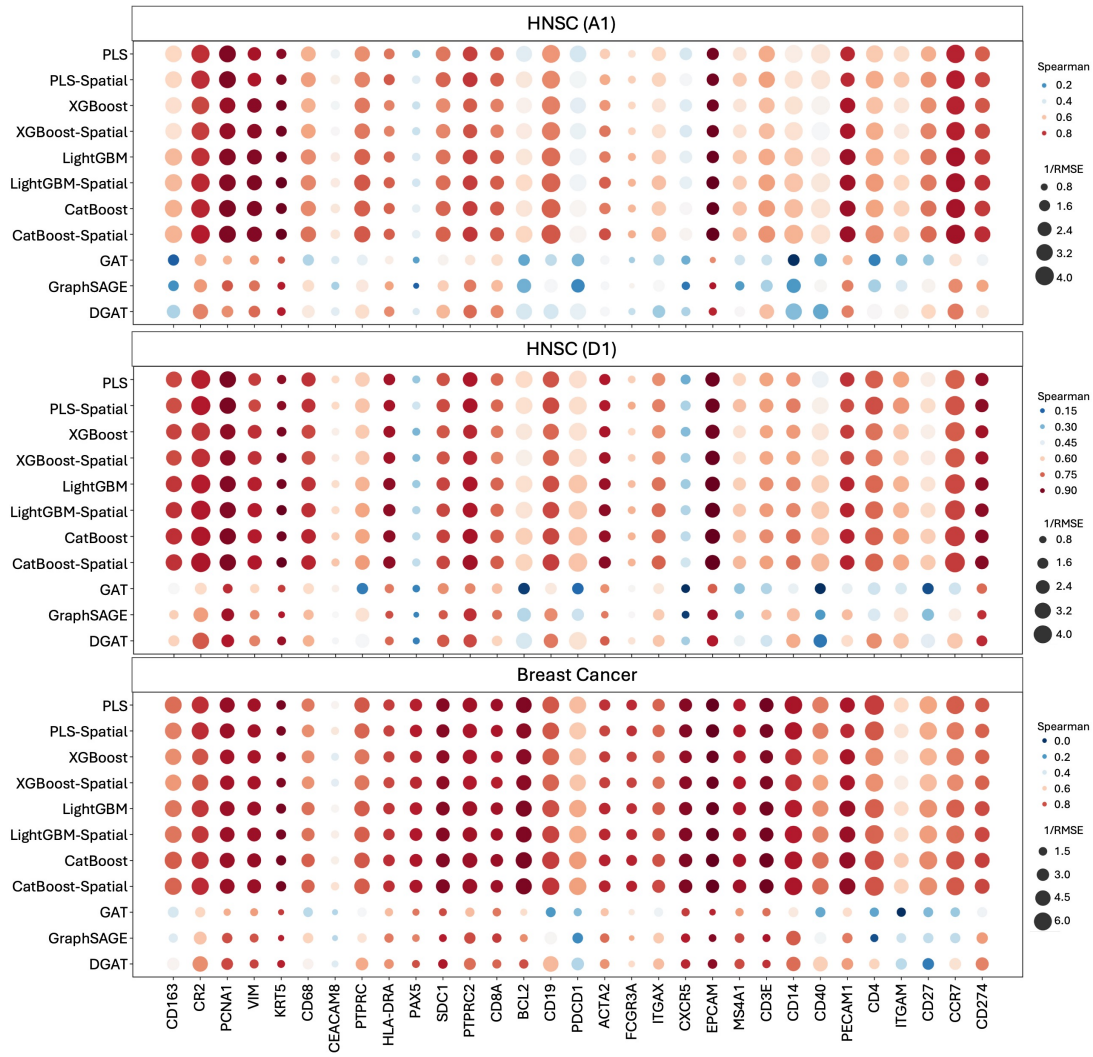


Fig. 5.7.: Within-sample prediction performance in HNSC and Breast Cancer. Protein-wise Spearman correlations for HNSC (A1), HNSC (D1), and Breast Cancer datasets, visualized with dot size reflecting inverse RMSE.

5.3.2 External validation using samples from the same tissue type

In this section, we apply cross-sample validation for two samples from the same tissue types including HNSC and human tonsil. Figure 5.8 compares the distribution of spot-wise total protein abundance between the two samples in each tissue type. As a result, despite differences in the number of spots, the overall abundance distributions are comparable between sample within each tissue type, indicating that large-scale distributional shifts do not confound cross-sample prediction.

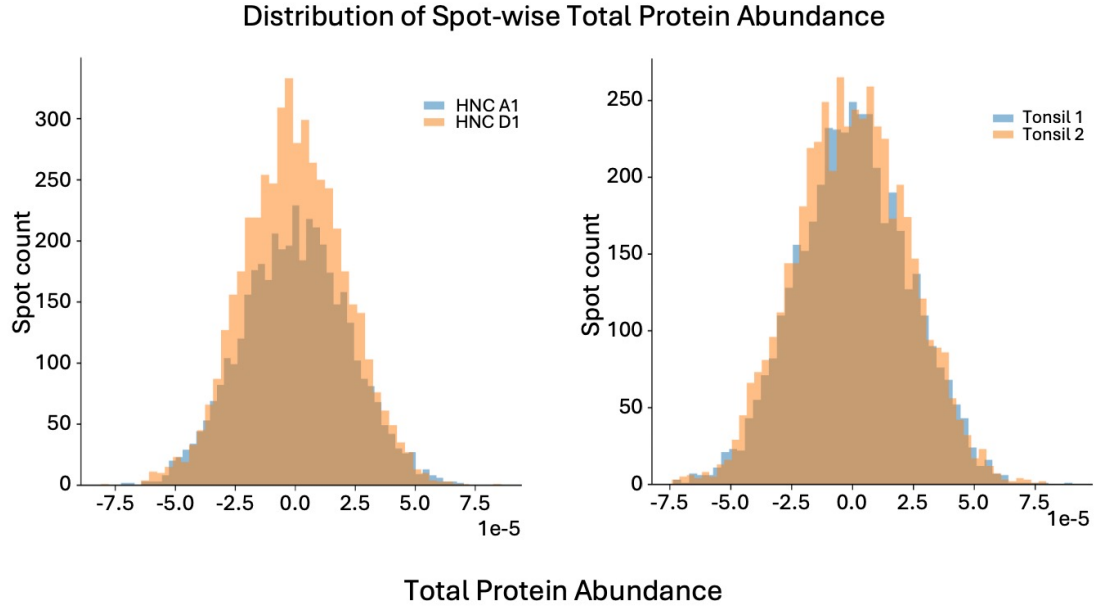


Fig. 5.8.: Cross-sample distribution of total protein abundance. Distribution of spot-wise total protein abundance for the source and target datasets in cross-sample validation.

Figure 5.9 presents the cross-sample prediction results for HNSC (A1), HNSC (D1), Tonsil 1, and Tonsil 2, where each sample is predicted using another sample from the same tissue (HNSC (D1), HNSC (A1), Tonsil 2, and Tonsil 1, respectively) as the training dataset. Consistent with the internal validation results, PLS, LightGBM, and CatBoost emerge as the top-performing methods, whereas GNN-based models such as GAT, GraphSAGE, and DGAT generally yield lower correlations (median Spearman often below 0.40) and higher RMSE values, reflecting their limited robustness.

Figure 5.9 further shows that models incorporating spatial features consistently achieve higher prediction accuracy compared to their non-spatial counterparts. For instance, in the HNSC (D1) sample, LightGBM-Spatial attains a median Pearson correlation of 0.68 and Spearman correlation of 0.65, with an RMSE of 0.21, compared to 0.54, 0.51, and 0.28 for the non-spatial LightGBM model. Similarly, in the Tonsil 1 sample, XGBoost-Spatial achieves a median Spearman correlation of 0.72 and Pearson correlation of 0.70, outperforming its non-spatial counterpart by approximately 0.12–0.15 in correlation and reducing RMSE by about 0.06. DGAT, however, does not provide notable improvements over GAT and GraphSAGE, with correlations remaining modest and RMSE values comparatively high, underscoring the limited effectiveness of GNN models in this cross-sample setting.

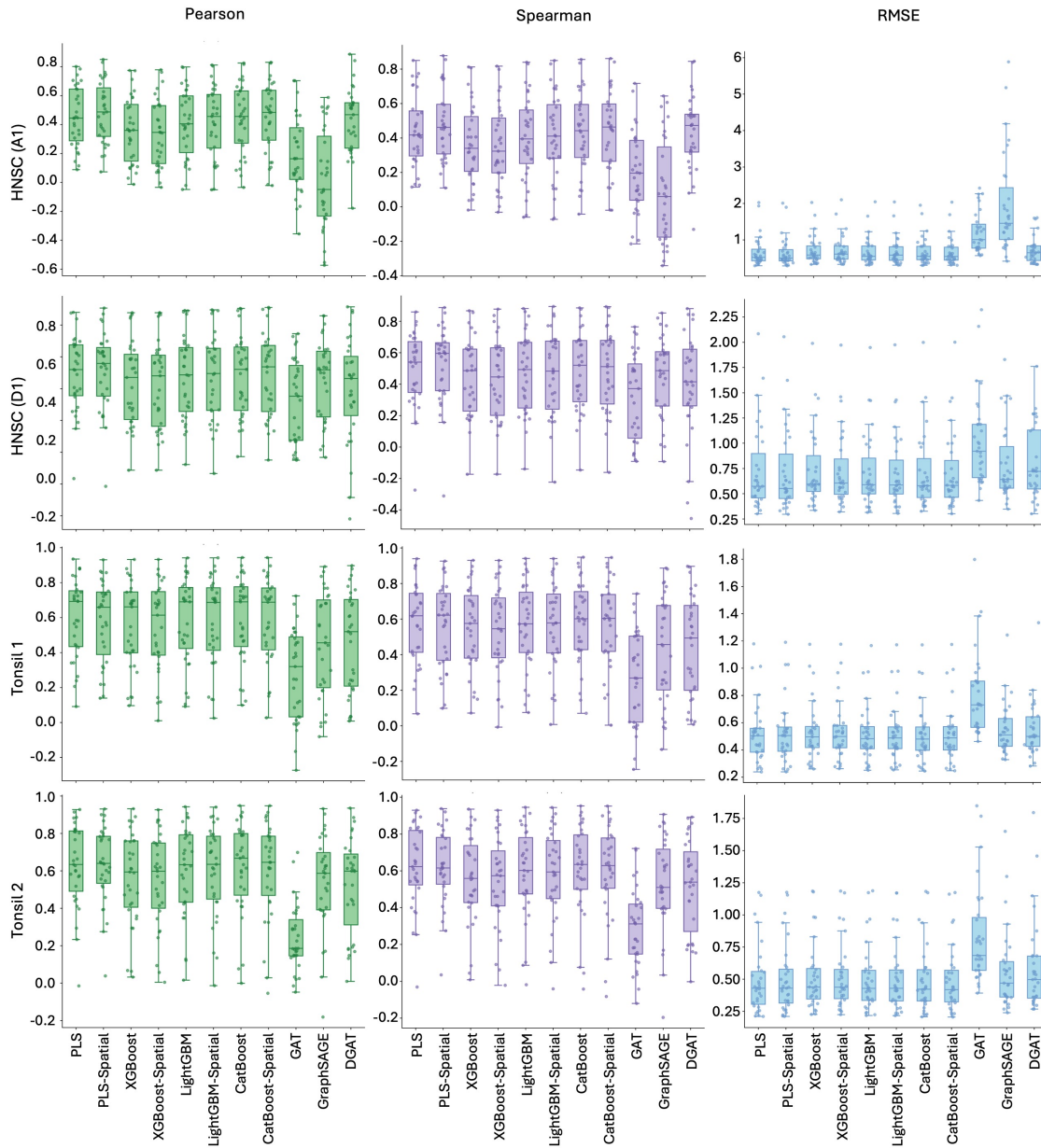


Fig. 5.9.: Evaluation of the protein prediction methods using cross-sample validation. Prediction performance on the HNSC (A1), HNSC (D1), Tonsil 1 and Tonsil 2 samples, evaluated using Pearson correlation, Spearman correlation, and root mean squared error (RMSE) across prediction methods.

Fig. 5.10 visualizes the prediction results using a protein-level prediction map for protein EPCAM in HNSC (D1) sample. EPCAM is an epithelial cell adhesion molecule highly expressed in carcinoma epithelial layers, serving as a key marker for tumor epithelial compartments. This protein is also among the good predictive proteins across methods. As shown in the figure, tree-based models both with spatial and non-spatial settings more

accurately reproduce the measured spatial patterns, PLS models miss some patterns, and GNN models often fail to recover fine-grained spatial structures.

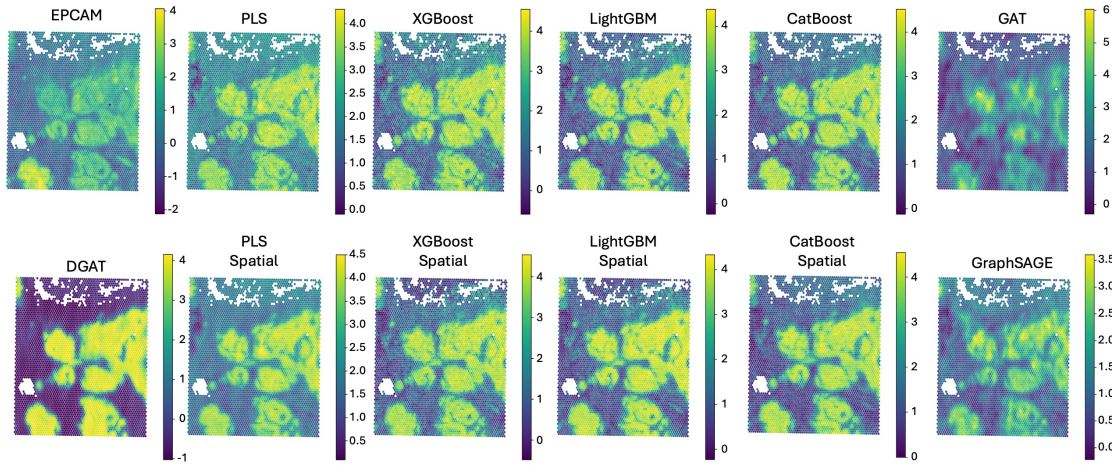


Fig. 5.10.: Evaluation of the protein prediction methods using cross-sample validation. Spatial prediction maps for one representative protein: EPCAM from the HNSC dataset. For each protein, the leftmost panel shows the measured distribution, followed by predictions from each model.

5.3.3 External validation using samples from the different tissue types

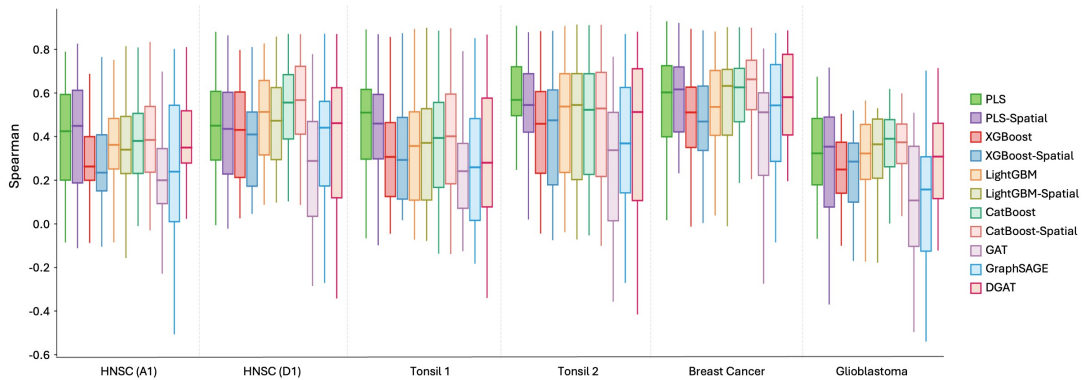


Fig. 5.11.: Cross-tissue validation of the prediction methods across six samples. Protein-wise performance on held-out datasets. Results are shown as boxplots of Spearman correlations, with panels ordered from left to right by the test dataset: HNSC (A1), HNSC (D1), Tonsil 1, Tonsil 2, Breast Cancer, and Glioblastoma. For each dataset, the results methods are in the following orders: PLS, PLS-Spatial, XGBoost, XGBoost-Spatial, LightGBM, LightGBM-Spatial, CatBoost, CatBoost-Spatial, GAT, GraphSAGE and DGAT.

In this section, we further assess the robustness and generalizability of the prediction models using cross-tissue validation. Figure 5.11 summarizes protein-wise performance for

each test dataset using the Spearman correlation across all proteins. PLS with both spatial and non-spatial settings are among top performing methods. CatBoost methods generally better than the other tree-based boosting method, while GNN methods still achieve worst performance across datasets.

We observe substantial variability in performance across tissues. For the breast cancer dataset, spatially augmented models achieve the strongest results, with CatBoost-Spatial reaching a median correlation of about 0.62 and interquartile ranges (IQRs) consistently above 0.50. The tonsil datasets yield more moderate results, with Tonsil 1 attaining the highest median correlation of approximately 0.50 with PLS and Tonsil 2 reaching 0.55. However, boosting and GNN models do not provide improved performance in these cases. DGAT shows similar patterns to the other GNN methods, with lower median correlations and broader variability, indicating limited cross-tissue predictive ability. In contrast, the HNSC datasets show lower performance, with median correlations of around 0.55 for both HNSC A1 and D1 using the best-performing models. Glioblastoma consistently demonstrates the weakest transferability, with median correlations below 0.40 for most models and IQRs frequently spanning zero, indicating unstable predictions across proteins. These differences suggest that cross-tissue prediction accuracy is strongly influenced by the degree of molecular similarity between the training and testing tissues.

To illustrate protein-level variability, Figure 5.12 and 5.13 present a heatmap of protein-wise Spearman correlations for the different samples. Results reveal marked performance differences across proteins, with certain markers (e.g., CD3E, PTPRC2, EPCAM) achieving correlations above 0.7 with multiple models, while others (e.g., CD40, ITGAM) are poorly predicted across all methods. This heterogeneity underscores that cross-tissue prediction accuracy depends not only on prediction models but also on predictive proteins. Tonsil datasets achieved the highest cross-tissue prediction accuracy, with median correlations up to 0.82 for Tonsil 1 and 0.80 for Tonsil 2 using spatially augmented tree-based models. HNSC datasets exhibited moderate performance (median correlations around 0.55–0.58 for the best models), while breast cancer showed slightly higher transferability than HNSC, with CatBoost-Spatial reaching 0.65 median correlation. Glioblastoma remained the most challenging tissue, with median correlations frequently below 0.30 and several proteins exhibiting near-zero or negative correlations across all models. Protein-level variability was observed in all tissues, with immune-related markers generally predicted more accurately than structural or stromal proteins. Figs B.1-B.6 in the Appendix illustrate, for each tissue type, the spatial distributions of two representative proteins alongside their model predictions, enabling visualization of tissue-specific prediction accuracy.

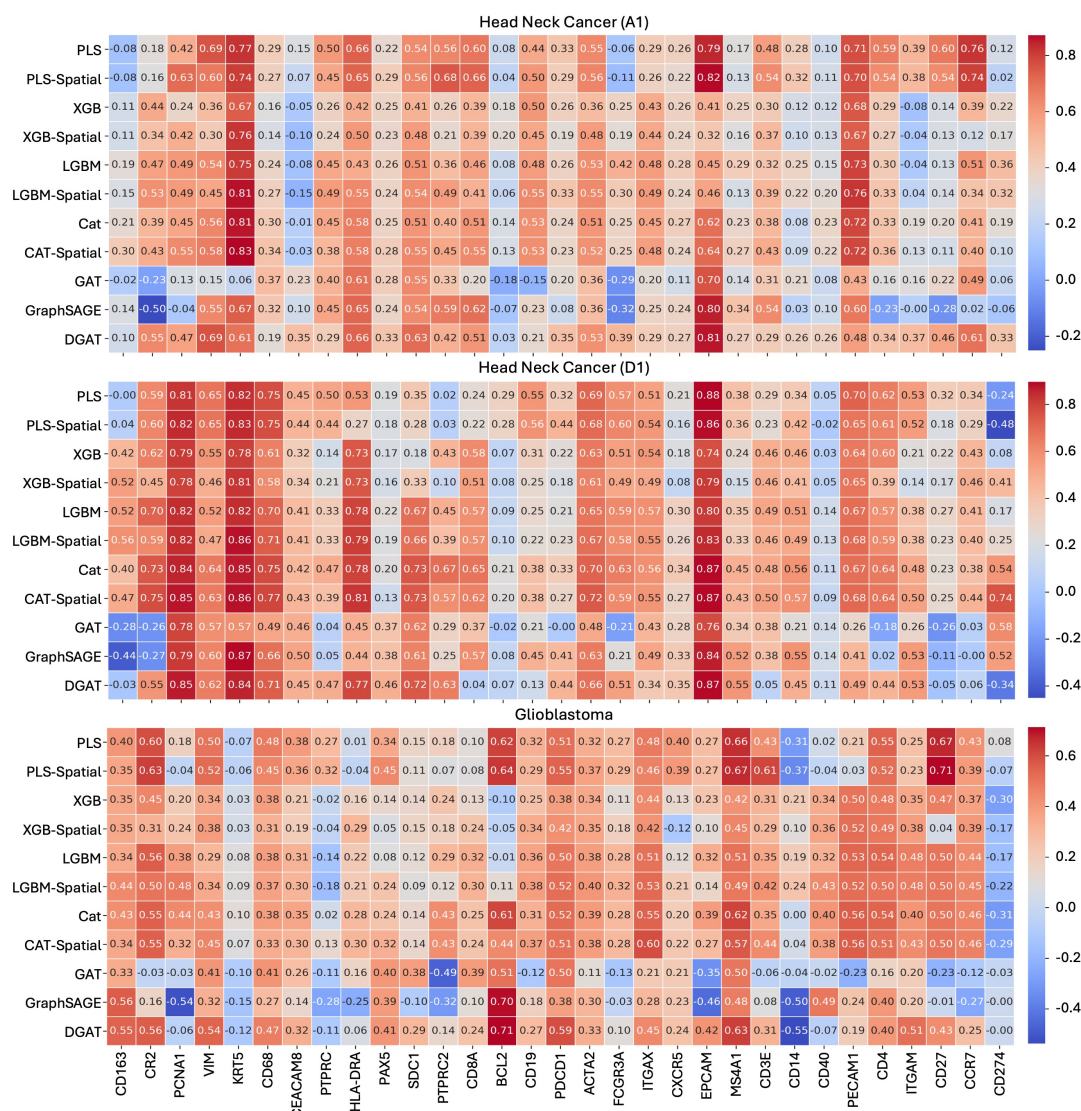


Fig. 5.13.: Cross-tissue prediction performance in HNSC and Glioblastoma. Heatmaps of protein-wise Spearman correlations for HNSC (A1), HNSC (D1), and Glioblastoma test sets, illustrating protein-level heterogeneity and differences in model transferability across tissues.

5.3.4 Spatial protein predictions recover immune-rich regions in the HNSC tumor microenvironment and identify protein markers associated with immunotherapy response

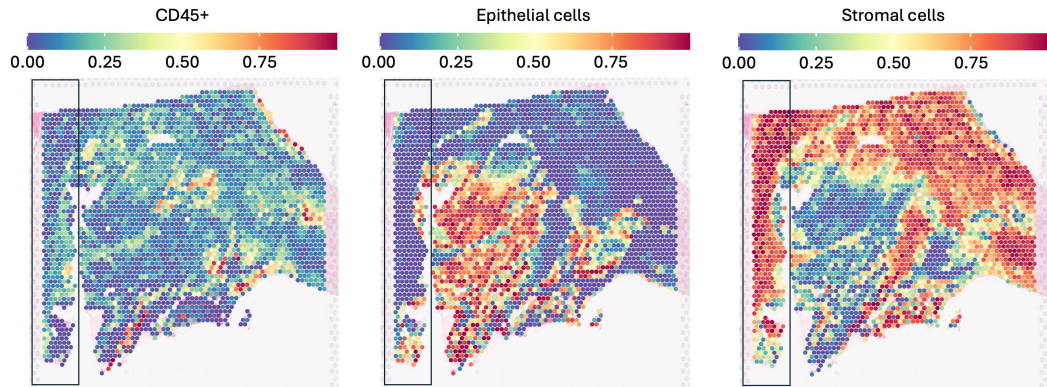


Fig. 5.14.: Deconvolution maps for a poorly immune-infiltrated HNSC sample. Spatial maps of cell type proportions estimated by SCDC deconvolution are shown for representative immune and tissue compartments, including CD45+ immune cells, epithelial cells, and stromal cells. Each spot is color-mapped to the estimated proportion (0-1) and overlaid on the corresponding tissue image. The boxed region highlights an area where immune infiltration is weak and spatial boundaries are less distinguishable using transcriptomic signals alone, motivating the integration of predicted protein features to recover previously undercharacterized macrophage-enriched domains.

To demonstrate the applications of spatial protein prediction models, we used training data from two HNSC Visium Spatial-CITESeq samples to predict protein abundance for 31 markers across 8 Visium spatial transcriptome (RNA alone) samples from a HNSC patient cohort treated with ICB immunotherapy.

First, we asked whether predicted proteins help us identify previously missing regions due to the lack of surface protein markers. In a poorly immune-infiltrated sample (Fig. 5.14 - with cell-type deconvolution maps), we used spatial clustering using three feature representations: RNA-only, predicted protein-only, and a combined RNA+protein feature space. The combined RNA+protein clustering identifies 9.7% more of macrophage-enriched spatial spots compared to the single data modality (Fig. 5.15a-b). Macrophage-enriched spots were defined as spatial locations with the highest macrophage proportion, as estimated by DECLUST deconvolution [203]. Specifically, the combination clustering recovered clusters 16 and 29, which are identified as macrophage-rich regions only in RNA- and protein-only analyses, respectively, among the highly enriched macrophage regions (clusters 1, 10, 16, 27, and 29) based on deconvolution. These clusters show high correlation (Pearson $r = 0.77$)

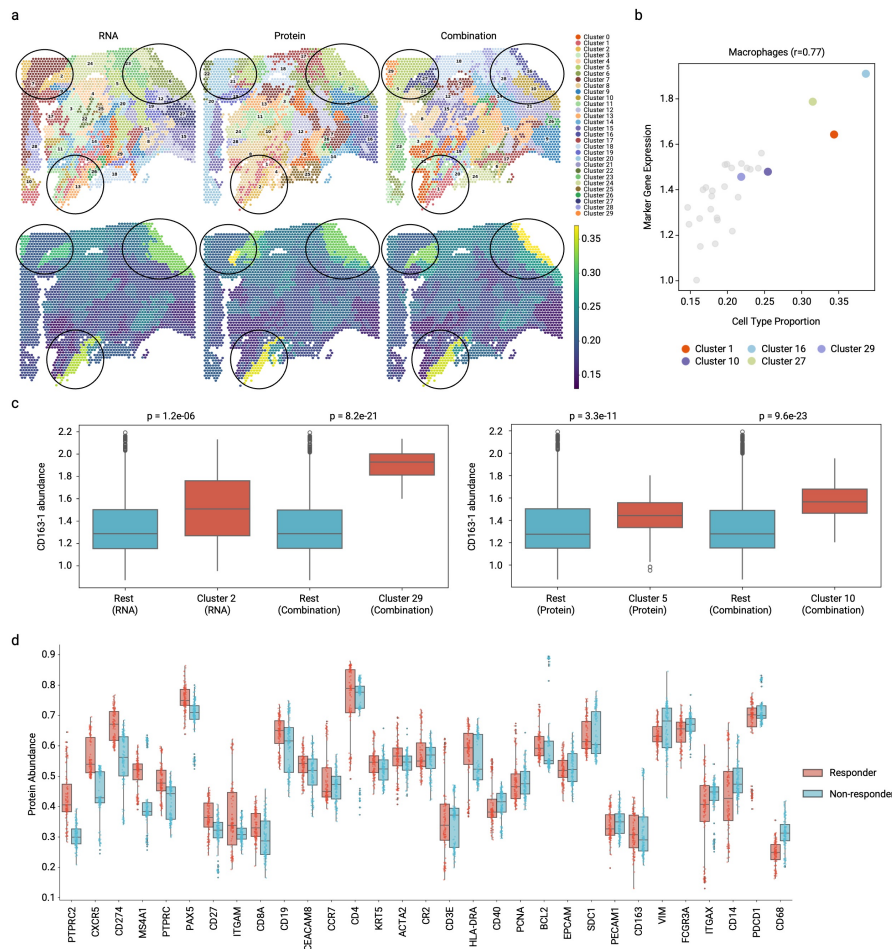


Fig. 5.15.: Spatial characterization of tumor immune microenvironment features using predicted protein abundance. (a) Spatial clustering of the HNSC sample using RNA-only, protein-only, and combined RNA+protein features. The circled regions highlight macrophage-enriched spatial domains. (b) For the combined-feature clustering, the relationship between cluster-level macrophage cell-type proportion and the mean expression of macrophage-associated marker genes is shown, with each point representing a spatial cluster. (c) Differential CD163 abundance between macrophage-enriched clusters and the remaining tissue across RNA-based, protein-based, and combined feature clustering. (d) Comparison of predicted protein abundance between responder and non-responder groups across immune-related markers.

in macrophage proportions estimated using deconvolution and macrophage gene signatures from scRNA-Seq references (**Fig. 5.15b**). Notably, we found that macrophage-enriched clusters identified in the RNA-only (Cluster 2) and protein-only (Cluster 5) settings are re-identified as clusters (29 & 10) with higher enrichment in macrophage protein marker, CD163, compared to the rest of the tissue when using the combined RNA+protein clustering ($p = 1.2 \times 10^{-6}$ vs. $p = 8.2 \times 10^{-21}$, and $p = 3.3 \times 10^{-11}$ vs. $p = 9.6 \times 10^{-23}$, **Fig. 5.15c**). These data confirmed the necessity of integrating complementary transcriptomic and

proteomic information, preserving distinctions captured by each modality alone, and producing more spatially coherent macrophage-rich region boundaries, thereby making inferred tissue domains more biologically interpretable.

Next, we evaluated predicted protein markers that differ between Responder and Non-Responder patients using a combination of multi-omics and RNA-only Visium samples, resulting in a total of 10 samples. Given the limited number of patients, statistical comparisons were performed at the cluster level to increase analytical power (**Fig. 5.15d**; **Table 5.1**), yielding 116 clusters from Responder samples and 138 clusters from Non-Responder samples. We found that Responder-associated clusters exhibit higher known “immunologically hot tumor” phenotypes, including higher CD45+ markers (PTPRC and PTPRC2; adjusted $p = 8.20 \times 10^{-14}$ and 1.48×10^{-30}), CD8+ T cell markers (CD8A; adjusted $p = 1.29 \times 10^{-7}$). In contrast, clusters in non-responder samples display elevated levels of macrophage- and myeloid-associated proteins, such as CD68, CD14, and ITGAX, marking immunosuppressive myeloid infiltration phenotypes as observed in anti-PD1 resistance. These observed differences indicate the feasibility of using predicted protein information derived from RNA gene expression to identify proteomic immunotherapeutic biomarkers and targets.

Tab. 5.1.: Pairwise comparison results between responders and non-responders at the cluster level.

Protein	p-value	Adjusted p-value (FDR)
PTPRC2	4.759552×10^{-32}	1.475461×10^{-30}
CXCR5	3.481524×10^{-23}	5.396362×10^{-22}
CD274	4.339586×10^{-21}	3.363179×10^{-20}
CD68	2.215664×10^{-21}	2.289519×10^{-20}
PAX5	7.358431×10^{-16}	4.562227×10^{-15}
MS4A1	2.739636×10^{-14}	1.213267×10^{-13}
PTPRC	1.586430×10^{-14}	8.196557×10^{-14}
BCL2	1.198689×10^{-7}	3.715937×10^{-7}
CD8A	3.736547×10^{-8}	1.287033×10^{-7}
CD27	6.657509×10^{-11}	2.579785×10^{-10}
ITGAM	3.314074×10^{-4}	7.338306×10^{-4}
CD19	3.227955×10^{-4}	7.338306×10^{-4}
CEACAM8	4.164777×10^{-4}	8.607206×10^{-4}
KRT5	1.527228×10^{-4}	4.304007×10^{-4}
ACTA2	5.182900×10^{-4}	1.004187×10^{-3}
FCGR3A	2.576045×10^{-4}	6.654784×10^{-4}
ITGAX	2.924025×10^{-3}	5.035821×10^{-3}
CD14	1.325004×10^{-3}	2.416183×10^{-3}
VIM	1.019148×10^{-2}	1.662820×10^{-2}
CD4	1.664853×10^{-2}	2.580521×10^{-2}
CD40	3.890557×10^{-2}	5.743204×10^{-2}
PDCD1	6.902013×10^{-2}	9.725564×10^{-2}

Continued on next page

Protein	p-value	Adjusted p-value (FDR)
HLA-DRA	1.209455×10^{-1}	1.630135×10^{-1}
PCNA	1.410351×10^{-1}	1.821703×10^{-1}
PECAM1	1.544639×10^{-1}	1.915352×10^{-1}
EPCAM	3.015942×10^{-1}	3.595931×10^{-1}
CD163	4.238029×10^{-1}	4.865886×10^{-1}
CCR7	4.966172×10^{-1}	5.498262×10^{-1}
CD3E	5.873657×10^{-1}	6.278737×10^{-1}
CR2	8.069796×10^{-1}	8.338789×10^{-1}
SDC1	8.537664×10^{-1}	8.537664×10^{-1}

5.3.5 Comparison of the prediction methods in terms of computational efficiency

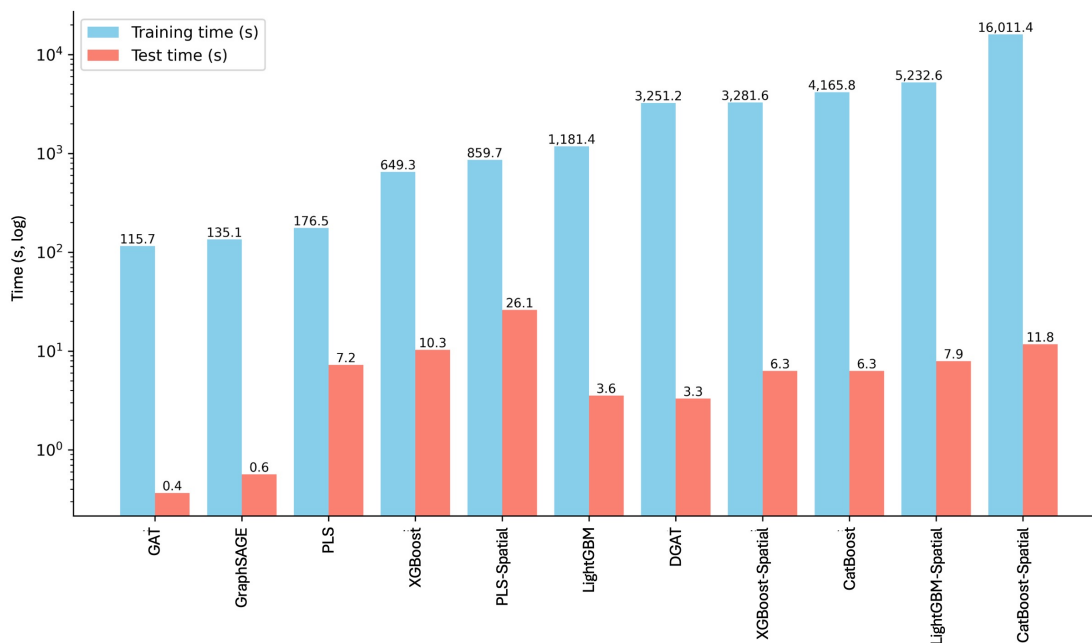


Fig. 5.16.: Computational time across 10 models in the cross-tissue validation. The x-axis represents the prediction models, while the y-axis shows the computational time in seconds on a logarithmic scale.

To compare the computational efficiency of different models is evaluated under the cross-tissue validation setting using the breast cancer dataset as the test set. All experiments are conducted on a Linux server equipped with 96 logical CPUs (48 physical cores) and no GPU acceleration. The computational time is measured from the start of the cross-tissue validation of a tool until its results are exported.

Figure 5.16 presents the training and testing times for the evaluated models on a logarithmic scale. The graph-based approaches (GAT and GraphSAGE) exhibit the shortest runtimes, requiring approximately 116 and 135 seconds for training and less than one second for testing, thereby representing the most computationally efficient methods. DGAT, in contrast, incurs a substantially higher training cost (3,281.6 seconds) but still achieves a relatively fast testing time of 6.3 seconds, positioning it between the lightweight graph-based methods and the more computationally intensive tree-based models. PLS also demonstrates relatively low computational demand (176.5 seconds for training and 7.2 seconds for testing), although the inclusion of spatial features increases its training and testing times substantially (859.7 and 26.1 seconds, respectively). In contrast, tree-based models require considerably longer training times, particularly when spatial features are incorporated. LightGBM-Spatial and CatBoost-Spatial show the highest computational costs, with training times of 5,232.6 and 16,011.4 seconds, respectively, though both maintain comparatively fast testing performance (7.9 and 11.8 seconds).

5.4 Discussion

This study provides a systematic comparison of predictive models for estimating spatial protein abundance from transcriptomic profiles, evaluated under within-sample, cross-sample, and cross-tissue validation schemes. The results show that 1) methods with spatial setting including PLS-Spatial, LightGBM-Spatial, XGBoost-Spatial, and CatBoost-Spatial generally perform better than the same methods with non-spatial setting, highlighting the importance of spatial information in protein prediction task, 2) PLS, LightGBM and CatBoost in both spatial and non-spatial settings are among top-performing methods, while GNN modes including GAT and GraphSAGE show the worst performance.

We observe that method performance in within-sample and cross-sample validation is comparable but drops substantially in cross-tissue validation, highlighting the challenges arising from biological and technical variation in samples from different tissues. In cross-tissue validation, prediction performance varies greatly across proteins but is consistent across methods, enabling the identification of proteins with high or low confidence for practical abundance inference.

Although GNNs are the fastest, they yield the lowest predictive performance among all methods. While boosting-based models deliver higher accuracy in many cases, they also incur significantly greater training times. In contrast, partial least squares (PLS) regression

offers extremely fast runtimes while still producing competitive results in certain scenarios. This efficiency makes PLS an appealing option when computational constraints outweigh the pursuit of maximal accuracy.

GNN models including GAT, GraphSAGE, and DGAT, which inherently encode spatial topology, consistently exhibit lower and more variable performance in our benchmarks. DGAT, despite its dual-graph attention mechanism designed to leverage both feature and spatial graphs, does not achieve substantial improvements over GAT or GraphSAGE in our evaluation, particularly under cross-sample and cross-tissue validation. These findings suggest that current GNN architectures for spatial multi-omics prediction require further optimization, potentially through architectural refinements and advanced feature engineering. Moreover, cross-sample assessment demonstrates potential utility for cost reduction in practical applications: laboratories could generate a limited number of spatial samples with both gene and protein expression, then extrapolate protein abundance in other, more economical RNA-only spatial transcriptomics datasets.

The biological application with HNSC samples from patients treated with immunotherapy demonstrates that predicted protein abundance carries a meaningful biological signal that complements transcriptomic information. Incorporating predicted protein features improves the spatial delineation of macrophage-enriched tumor regions and strengthens the correspondence between inferred cell-type composition and marker gene expression. Moreover, predicted protein profiles distinguish responder and non-responder immune states, with responders exhibiting higher abundance of T cell-associated markers and non-responders enriched for macrophage-associated signatures. These results indicate that predicted protein abundance enables refined characterization of immune microenvironments, even in datasets lacking direct proteomic measurements.

In summary, this study provide a systematic benchmark for evaluating spatial protein prediction methods in diverse validation scenarios. The results clarify trade-offs between model classes, emphasizing that spatially augmented tree ensembles currently provide the most favorable balance between accuracy and robustness, while GNNs hold untapped potential with targeted architectural improvements. Future research could explore hybrid models that combine the spatial representational power of GNNs with the predictive stability of gradient-boosting ensembles, as well as unified preprocessing pipelines to facilitate reproducible and fair method comparisons.

Discussion, Conclusions and Future Work

6.1 Discussion

This thesis addresses two key methodological challenges in spatial omics: enhancing cell type deconvolution from sparse spatial transcriptomic data and systematically evaluating computational frameworks for predicting protein abundance from RNA profiles. These two lines of research demonstrate how spatial context can be leveraged not only to improve cellular composition estimates but also to extend the predictive scope of transcriptomic data toward multi-omic integration.

The DECLUST framework highlights the importance of incorporating spatial coherence into computational pipelines. By aggregating gene expression across spatially coherent spots, this approach mitigates noise and sparsity, resulting in improved deconvolution accuracy and interpretability, outperforming several existing approaches in both simulated and real datasets. By explicitly incorporating spatial coordinates, DECLUST preserves tissue architecture and enables the identification of functional niches. However, its reliance on high-quality single-cell references and sensitivity to clustering parameters may limit its generalizability across diverse tissue types or low-quality datasets.

The benchmarking study of protein prediction models under within-sample, cross-sample, and cross-tissue validation settings further highlights the importance of spatial feature integration. Gradient boosting ensembles consistently outperform other methods in terms of predictive accuracy, while partial least squares regression offers a computationally efficient alternative. Graph neural networks, despite their theoretical suitability for spatial data, underperform in current implementations, suggesting the need for architectural refinement. Importantly, predictive performance varies across proteins and tissues, underscoring the context-dependent nature of RNA–protein relationships and the need for protein-specific modelling strategies.

Together, these studies illustrate complementary strategies for advancing spatial omics: one by improving the fidelity of cell type deconvolution through spatially informed clustering,

and the other by enabling transcriptome-based prediction of proteomic landscapes. Both emphasize the critical role of spatial information in transforming spatial transcriptomics from a descriptive tool into a predictive and integrative framework for studying tissue organization and molecular function.

6.2 Conclusion

This thesis presents methodological innovations that enhance the analytical power of spatial transcriptomics. Through the development of DECLUST and the systematic benchmarking of protein prediction models, it demonstrates the importance of incorporating spatial context into computational frameworks for improving both cell type deconvolution and protein abundance inference.

The findings underscore the value of spatial coherence, multi-modal integration, and biologically informed modelling in addressing the limitations of current spatial omics technologies. By addressing both the decomposition of cellular composition and the inference of downstream molecular modalities, this work lays the groundwork for advancing spatial omics toward more comprehensive and scalable applications in biology and medicine.

6.3 Future Work

6.3.1 Sequencing-free spatial multi-omics

To reduce cost and improve scalability, future models should explore sequencing-free approaches that infer molecular profiles directly from histological images. Deep learning approaches trained on paired image–transcriptome datasets could enable prediction of molecular profiles directly from tissue sections, bypassing the need for RNA sequencing. Key challenges include the collection of sufficiently large and diverse training datasets, as well as the development of models that can generalize across staining protocols, imaging platforms, and tissue types. Addressing these issues could transform spatial multi-omics into a more accessible and cost-efficient technology for routine use in research and clinical practice.

6.3.2 Deep learning for direct cell type deconvolution

Extending DECLUST into an image-guided framework could enable cell type inference without single-cell RNA references. Deep neural architectures, such as convolutional neural networks, vision transformers, or multimodal graph networks, could be developed to map tissue morphology to cellular composition without requiring single-cell RNA. Such models enable deconvolution in tissues where reference atlases are incomplete or unavailable, thereby broadening the applicability of spatial transcriptomics. Techniques such as visual attention maps, feature attribution methods, and multimodal embeddings can help establish transparent links between morphological cues and computational predictions, thereby increasing confidence in the results.

6.3.3 Multi-modal data integration

Future research should move beyond transcriptomics and histology to incorporate additional molecular modalities, including spatial proteomics, metabolomics, and ligand–receptor interactions. Multi-modal integration would provide a more comprehensive view of tissue organization and function, improving both interpretability and robustness. Hybrid computational frameworks that fuse histological features, RNA profiles, and protein measurements could bridge gaps between morphology and molecular activity, offering richer biological insights. Developing such integrative models will likely require both novel architectures capable of handling heterogeneous data and sophisticated data harmonization techniques to correct for differences in resolution and noise levels across modalities. Multi-modal learning could also reveal higher-order biological principles, such as the coordination between transcriptional programs, protein localization, and metabolic activity, which remain inaccessible to unimodal analyses.

6.3.4 Adaptive and generalizable frameworks

Both DECLUST and current protein prediction models require dataset-specific parameter tuning and remain sensitive to the quality of reference data. Future frameworks should emphasize adaptability, for example by developing clustering strategies that automatically adjust to dataset-specific spatial resolution, noise level, and heterogeneity. Generalizable models capable of transferring knowledge across tissues, conditions, or even species will be crucial for translational applications. Advances in domain adaptation and transfer learning may play a central role in achieving this goal. One possible direction is the creation of

universal pretrained models for spatial omics, analogous to large language models in natural language processing, which could be fine-tuned for specific tissues or research questions. Such developments would substantially reduce the barrier to entry for researchers and make spatial omics analyses more consistent across studies and laboratories.

6.3.5 Protein-specific predictive modeling

The benchmarking analysis highlights that not all proteins can be predicted with equal reliability from RNA data, and predictive performance also varies across tissues. Future efforts should focus on protein-specific modeling strategies that integrate prior biological knowledge, such as known transcriptional regulators, signaling pathways, or subcellular localization. Attention-based deep learning models may be particularly effective for emphasizing protein-relevant RNA features, while incorporating curated pathway information could further improve predictions. Another important direction will be the systematic characterization of why certain proteins are more predictable than others, which could lead to general design principles for RNA-to-protein inference. This line of research not only improves the accuracy of computational prediction but also deepens biological understanding of the transcriptional and post-transcriptional mechanisms that shape protein abundance in different tissue contexts.

6.3.6 Toward clinical translation

A long-term research direction is the translation of computational frameworks into clinical practice. Sequencing-free prediction models based on histology could be integrated into diagnostic workflows, providing rapid and cost-effective assessment of cellular composition or protein expression in patient tissue sections. Such approaches may support applications in oncology, pathology, and personalized medicine by enabling the identification of cellular niches, tumor microenvironment composition, or biomarker expression directly from routine slides. Ensuring robustness, interpretability, and regulatory compliance will be essential steps in moving from research tools to clinically deployable systems. Collaborations with clinicians and pathologists will be critical for validating models in real-world scenarios, while rigorous benchmarking on diverse patient cohorts will be necessary to ensure fairness and reproducibility. If successful, these developments could bridge the gap between computational innovation and clinical utility, making spatial omics a practical tool for precision healthcare.

Bibliography

- [1] Mei R. M. Du, Changqing Wang, Charity W. Law, Daniela Amann-Zalcenstein, Casey J. A. Anttila, Ling Ling, Peter F. Hickey, Callum J. Sargeant, Yunshun Chen, Lisa J. Ioannidis, Pradeep Rajasekhar, Raymond K. H. Yip, Kelly L. Rogers, Diana S. Hansen, Rory Bowden, and Matthew E. Ritchie. Benchmarking spatial transcriptomics technologies with the multi-sample SpatialBenchVisium dataset. *Genome Biology*, 26(1):77, March 2025.
- [2] Jongwon Lee and Minsu Yoo & Jungmin Choi. Recent advances in spatially resolved transcriptomics: challenges and opportunities. *BMB Reports*, 55(3):113–124, March 2022. Publisher: Korean Society for Biochemistry and Molecular Biology.
- [3] Mikko J. Lammi and Chengjuan Qu. Spatial Transcriptomics, Proteomics, and Epigenomics as Tools in Tissue Engineering and Regenerative Medicine. *Bioengineering*, 11(12):1235, December 2024.
- [4] Ritian Qin, Jiacheng Ma, Fuchu He, and Weijie Qin. In-depth and high-throughput spatial proteomics for whole-tissue slice profiling by deep learning-facilitated sparse sampling strategy. *Cell Discovery*, 11(1):21, March 2025. Publisher: Nature Publishing Group.
- [5] Beiyu Hu, Ruiqiao He, Kun Pang, Guibin Wang, Ning Wang, Wenzhuo Zhu, Xin Sui, Huajing Teng, Tianxin Liu, Junjie Zhu, Zewen Jiang, Jinyang Zhang, Zhenqiang Zuo, Weihua Wang, Peifeng Ji, and Fangqing Zhao. High-resolution spatially resolved proteomics of complex tissues based on microfluidics and transfer learning. *Cell*, 188(3):734–748.e22, February 2025. Publisher: Elsevier.

- [6] Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, August 1970. Publisher: Nature Publishing Group.
- [7] Matthew Cobb. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biology*, 15(9):e2003243, September 2017.
- [8] Daniel R. Larson, Robert H. Singer, and Daniel Zenklusen. A single molecule view of gene expression. *Trends in Cell Biology*, 19(11):630–637, November 2009.
- [9] Xinmin Li and Cun-Yu Wang. From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science*, 13(1):36, November 2021. Publisher: Nature Publishing Group.
- [10] Liangchen Yue, Feng Liu, Jionsong Hu, Pin Yang, Yuxiang Wang, Junguo Dong, Wenjie Shu, Xingxu Huang, and Shengqi Wang. A guidebook of spatial transcriptomic technologies, data resources and analysis approaches. *Computational and Structural Biotechnology Journal*, 21:940–955, January 2023.
- [11] Jana-Charlotte Hegenbarth, Giuliana Lezzocche, Leon J. De Windt, and Monika Stoll. Perspectives on Bulk-Tissue RNA Sequencing and Single-Cell RNA Sequencing for Cardiac Transcriptomics. *Frontiers in Molecular Medicine*, 2:839338, February 2022.
- [12] Jane Ru Choi, Kar Wey Yong, Jean Yu Choi, and Alistair C. Cowie. Single-Cell RNA Sequencing and Its Combination with Protein and DNA Analyses. *Cells*, 9(5):1130, May 2020. Publisher: Multidisciplinary Digital Publishing Institute.
- [13] Ashraful Haque, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1):75, August 2017.
- [14] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1):57–63, January 2009.
- [15] Sean K. Maden, Sang Ho Kwon, Louise A. Huuki-Myers, Leonardo Collado-Torres, Stephanie C. Hicks, and Kristen R. Maynard. Challenges and opportunities to computationally deconvolve heterogeneous tissue with varying cell sizes using single-cell RNA-sequencing datasets. *Genome Biology*, 24(1):288, December 2023.

- [16] Kyongho Choe, Unil Pak, Yu Pang, Wanjun Hao, and Xiuqin Yang. Advances and Challenges in Spatial Transcriptomics for Developmental Biology. *Biomolecules*, 13(1):156, January 2023.
- [17] Jorge A. Tzec-Interián, Daianna González-Padilla, and Elsa B. Góngora-Castillo. Bioinformatics perspectives on transcriptomics: A comprehensive review of bulk and single-cell RNA sequencing analyses. *Quantitative Biology*, 13(2):e78, 2025. [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qub2.78](https://onlinelibrary.wiley.com/doi/pdf/10.1002/qub2.78).
- [18] Fredrik Salmén, Patrik L. Ståhl, Annelie Mollbrink, José Fernández Navarro, Sanja Vickovic, Jonas Frisén, and Joakim Lundeberg. Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. *Nature Protocols*, 13(11):2501–2534, November 2018.
- [19] Kim Thrane, Hanna Eriksson, Jonas Maaskola, Johan Hansson, and Joakim Lundeberg. Spatially Resolved Transcriptomics Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous Malignant Melanoma. *Cancer Research*, 78(20):5970–5979, October 2018.
- [20] Vivien Marx. Method of the Year: spatially resolved transcriptomics. *Nature Methods*, 18(1):9–14, January 2021. Publisher: Nature Publishing Group.
- [21] Monika Piwecka, Nikolaus Rajewsky, and Agnieszka Rybak-Wolf. Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. *Nature Reviews Neurology*, 19(6):346–362, June 2023. Publisher: Nature Publishing Group.
- [22] Rohit Arora, Christian Cao, Mehul Kumar, Sarthak Sinha, Ayan Chanda, Reid McNeil, Divya Samuel, Rahul K. Arora, T. Wayne Matthews, Shamir Chandarana, Robert Hart, Joseph C. Dort, Jeff Biernaskie, Paola Neri, Martin D. Hycza, and Pinaki Bose. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nature Communications*, 14(1):5029, August 2023. Publisher: Nature Publishing Group.
- [23] Jun Du, Yu-Chen Yang, Zhi-Jie An, Ming-Hui Zhang, Xue-Hang Fu, Zou-Fang Huang, Ye Yuan, and Jian Hou. Advances in spatial transcriptomics and related data analysis strategies. *Journal of Translational Medicine*, 21(1):330, May 2023.

- [24] Luyi Tian, Fei Chen, and Evan Z. Macosko. The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41(6):773–782, June 2023. Publisher: Nature Publishing Group.
- [25] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, July 2016.
- [26] W. Meier-Ruge, W. Bielser, E. Remy, F. Hillenkamp, R. Nitsche, and R. Unsöld. The laser in the Lowry technique for microdissection of freeze-dried tissue slices. *The Histochemical Journal*, 8(4):387–401, July 1976.
- [27] C. Brown. In situ hybridization with riboprobes: an overview for veterinary pathologists. *Veterinary Pathology*, 35(3):159–167, May 1998.
- [28] Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, 10(9):857–860, September 2013. Publisher: Nature Publishing Group.
- [29] Je Hyuk Lee, Evan R. Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C. Ferrante, Richard Terry, Brian M. Turczyk, Joyce L. Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M. Church. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*, 10(3):442–458, March 2015. Publisher: Nature Publishing Group.
- [30] Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Archibald Enninfu, Cindy C. Guo, Toma Tebaldi, Di Zhang, Dongjoo Kim, Zhiliang Bai, Eileen Norris, Alisia Pan, Jiatong Li, Yang Xiao, Stephanie Halene, and Rong Fan. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell*, 183(6):1665–1681.e18, December 2020.
- [31] Xiaonan Fu, Li Sun, Runze Dong, Jane Y. Chen, Runglawan Silakit, Logan F. Condon, Yiing Lin, Shin Lin, Richard D. Palmiter, and Liangcai Gu. Polony gels

enable amplifiable DNA stamping and spatial transcriptomics of chronic pain. *Cell*, 185(24):4621–4633.e17, November 2022.

- [32] J. G. Gall and M. L. Pardue. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America*, 63(2):378–383, June 1969.
- [33] M. R. Emmert-Buck, R. F. Bonner, P. D. Smith, R. F. Chuaqui, Z. Zhuang, S. R. Goldstein, R. A. Weiss, and L. A. Liotta. Laser capture microdissection. *Science (New York, N.Y.)*, 274(5289):998–1001, November 1996.
- [34] Jan Philipp Junker, Emily S. Noël, Victor Guryev, Kevin A. Peterson, Gopi Shah, Jan Huisken, Andrew P. McMahon, Eugene Berezikov, Jeroen Bakkers, and Alexander van Oudenaarden. Genome-wide RNA Tomography in the zebrafish embryo. *Cell*, 159(3):662–675, October 2014.
- [35] Jun Chen, Shengbao Suo, Patrick P. Tam, Jing-Dong J. Han, Guangdun Peng, and Naihe Jing. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nature Protocols*, 12(3):566–580, March 2017.
- [36] Halima Hannah Schede, Christian G. Schneider, Johanna Stergiadou, Lars E. Borm, Anurag Ranjak, Tracy M. Yamawaki, Fabrice P. A. David, Peter Lönnerberg, Maria Antonietta Tosches, Simone Codeluppi, and Gioele La Manno. Spatial tissue profiling by imaging-free molecular tomography. *Nature Biotechnology*, 39(8):968–977, August 2021.
- [37] Ditte Lovatt, Brittani K. Ruble, Jaehee Lee, Hannah Dueck, Tae Kyung Kim, Stephen Fisher, Chantal Francis, Jennifer M. Spaethling, John A. Wolf, M. Sean Grady, Alexandra V. Ulyanova, Sean B. Yeldell, Julianne C. Griepenburg, Peter T. Buckley, Junhyong Kim, Jai-Yoon Sul, Ivan J. Dmochowski, and James Eberwine. Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nature Methods*, 11(2):190–196, February 2014.
- [38] Mizuki Honda, Shinya Oki, Ryuichi Kimura, Akihito Harada, Kazumitsu Maehara, Kaori Tanaka, Chikara Meno, and Yasuyuki Ohkawa. High-depth spatial transcriptome analysis by photo-isolation chemistry. *Nature Communications*, 12(1):4416, July 2021.

- [39] Jocelyn Y. Kishi, Ninning Liu, Emma R. West, Kuanwei Sheng, Jack J. Jordanides, Matthew Serrata, Constance L. Cepko, Sinem K. Saka, and Peng Yin. Light-Seq: light-directed in situ barcoding of biomolecules in fixed cells and tissues for spatially indexed sequencing. *Nature Methods*, 19(11):1393–1402, November 2022. Publisher: Nature Publishing Group.
- [40] Kenneth H. Hu, John P. Eichorst, Chris S. McGinnis, David M. Patterson, Eric D. Chow, Kelly Kersten, Stephen C. Jameson, Zev J. Gartner, Arjun A. Rao, and Matthew F. Krummel. ZipSeq: barcoding for real-time mapping of single cell transcriptomes. *Nature Methods*, 17(8):833–843, August 2020.
- [41] Joseph G. Gall. The Origin of In Situ Hybridization - a Personal History. *Methods (San Diego, Calif.)*, 98:4–9, April 2016.
- [42] Simone Codeluppi, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods*, 15(11):932–935, November 2018.
- [43] Eric Lubeck, Ahmet F. Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single cell in situ RNA profiling by sequential hybridization. *Nature methods*, 11(4):360–361, April 2014.
- [44] Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, April 2015.
- [45] Je Hyuk Lee, Evan R. Daugharthy, Jonathan Scheiman, Reza Kalhor, Joyce L. Yang, Thomas C. Ferrante, Richard Terry, Sauveur S. F. Jeanty, Chao Li, Ryoji Amamoto, Derek T. Peters, Brian M. Turczyk, Adam H. Marblestone, Samuel A. Inverso, Amy Bernard, Prashant Mali, Xavier Rios, John Aach, and George M. Church. Highly multiplexed subcellular RNA sequencing in situ. *Science (New York, N.Y.)*, 343(6177):1360–1363, March 2014.
- [46] Shahar Alon, Daniel R. Goodwin, Anubhav Sinha, Asmamaw T. Wassie, Fei Chen, Evan R. Daugharthy, Yosuke Bando, Atsushi Kajita, Andrew G. Xue, Karl Marrett, Robert Prior, Yi Cui, Andrew C. Payne, Chun-Chen Yao, Ho-Jun Suk, Ru Wang, Chih-Chieh Jay Yu, Paul Tillberg, Paul Reginato, Nikita Pak, Songlei Liu, Sukanya

- Punthambaker, Eswar P. R. Iyer, Richie E. Kohman, Jeremy A. Miller, Ed S. Lein, Ana Lako, Nicole Cullen, Scott Rodig, Karla Helvie, Daniel L. Abravanel, Nikhil Wagle, Bruce E. Johnson, Johanna Klughammer, Michal Slyper, Julia Waldman, Judit Jané-Valbuena, Orit Rozenblatt-Rosen, Aviv Regev, IMAXT Consortium, George M. Church, Adam H. Marblestone, and Edward S. Boyden. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science (New York, N.Y.)*, 371(6528):eaax2656, January 2021.
- [47] Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science (New York, N.Y.)*, 361(6400):eaat5691, July 2018.
- [48] Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, Matthew N. Tran, Zachary Besich, Madhavi Tippi, Jennifer Chew, Yifeng Yin, Joel E. Kleinman, Thomas M. Hyde, Nikhil Rao, Stephanie C. Hicks, Keri Martinowich, and Andrew E. Jaffe. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, March 2021.
- [49] Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, March 2019.
- [50] Robert R. Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L. Marshall, Daniela J. Di Bella, Paola Arlotta, Evan Z. Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology*, 39(3):313–319, March 2021.
- [51] Sanja Vickovic, Gökçen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernández Navarro, Joshua Gould, Gabriel K. Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L. Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16(10):987–990, October 2019. Publisher: Nature Publishing Group.

- [52] Chun-Seok Cho, Jingyue Xi, Yichen Si, Sung-Rye Park, Jer-En Hsu, Myungjin Kim, Goo Jun, Hyun Min Kang, and Jun Hee Lee. Microscopic examination of spatial transcriptome using Seq-Scope. *Cell*, 184(13):3559–3572.e22, June 2021.
- [53] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, Xin Wang, Huifang Lu, Xi Chen, Xing Liu, Xin Huang, Zhao Li, Yan Hong, Yujia Jiang, Jian Peng, Shuai Liu, Mengzhe Shen, Chuanyu Liu, Quanshui Li, Yue Yuan, Xiaoyu Wei, Huiwen Zheng, Weimin Feng, Zhifeng Wang, Yang Liu, Zhaohui Wang, Yunzhi Yang, Haitao Xiang, Lei Han, Baoming Qin, Pengcheng Guo, Guangyao Lai, Pura Muñoz-Cánoves, Patrick H. Maxwell, Jean Paul Thiery, Qing-Feng Wu, Fuxiang Zhao, Bichao Chen, Mei Li, Xi Dai, Shuai Wang, Haoyan Kuang, Junhou Hui, Liquan Wang, Ji-Feng Fei, Ou Wang, Xiaofeng Wei, Haorong Lu, Bo Wang, Shiping Liu, Ying Gu, Ming Ni, Wenwei Zhang, Feng Mu, Ye Yin, Huanming Yang, Michael Lisby, Richard J. Cornall, Jan Mulder, Mathias Uhlén, Miguel A. Esteban, Yuxiang Li, Longqi Liu, Xun Xu, and Jian Wang. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21, May 2022. Publisher: Elsevier.
- [54] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, January 2017. Publisher: Nature Publishing Group.
- [55] Chuanyu Liu, Tao Wu, Fei Fan, Ya Liu, Liang Wu, Michael Junkin, Zhifeng Wang, Yeya Yu, Weimao Wang, Wenbo Wei, Yue Yuan, Mingyue Wang, Mengnan Cheng, Xiaoyu Wei, Jiangshan Xu, Quan Shi, Shiping Liu, Ao Chen, Ou Wang, Ming Ni, Wenwei Zhang, Zhouchun Shang, Yiwei Lai, Pengcheng Guo, Carl Ward, Giacomo Volpe, Lei Wang, Huan Zheng, Yang Liu, Brock A. Peters, Jody Beecher, Yongwei Zhang, Miguel A. Esteban, Yong Hou, Xun Xu, I.-Jane Chen, and Longqi Liu. A portable and cost-effective microfluidic system for massively parallel single-cell

transcriptome profiling, November 2019. Pages: 818450 Section: New Results.

- [56] Jiarui Ding, Xian Adiconis, Sean K. Simmons, Monika S. Kowalczyk, Cynthia C. Hession, Nemanja D. Marjanovic, Travis K. Hughes, Marc H. Wadsworth, Tyler Burks, Lan T. Nguyen, John Y. H. Kwon, Boaz Barak, William Ge, Amanda J. Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K. Shalek, Alexandra-Chloé Villani, Aviv Regev, and Joshua Z. Levin. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology*, 38(6):737–746, June 2020.
- [57] David W. McKellar, Madhav Mantri, Meleana M. Hinchman, John S. L. Parker, Praveen Sethupathy, Benjamin D. Cosgrove, and Iwijn De Vlamincx. Spatial mapping of the total transcriptome by in situ polyadenylation. *Nature Biotechnology*, 41(4):513–520, April 2023.
- [58] Yuhang Fan, Žaneta Andrusivová, Yunming Wu, Chew Chai, Ludvig Larsson, Mengxiao He, Liqun Luo, Joakim Lundeberg, and Bo Wang. Expansion spatial transcriptomics. *Nature Methods*, 20(8):1179–1182, August 2023.
- [59] Huan Wang, Ruixu Huang, Jack Nelson, Ce Gao, Miles Tran, Anna Yeaton, Kristen Felt, Kathleen L. Pfaff, Teri Bowman, Scott J. Rodig, Kevin Wei, Brittany A. Goods, and Samouil L. Farhi. Systematic benchmarking of imaging spatial transcriptomics platforms in FFPE tissues. *bioRxiv*, page 2023.12.07.570603, December 2023.
- [60] Pengfei Ren, Rui Zhang, Yunfeng Wang, Peng Zhang, Ce Luo, Suyan Wang, Xiaohong Li, Zongxu Zhang, Yanping Zhao, Yufeng He, Haorui Zhang, Yufeng Li, Zhidong Gao, Xiuping Zhang, Yahui Zhao, Zhihua Liu, Yuanguang Meng, Zhe Zhang, and Zexian Zeng. Systematic Benchmarking of High-Throughput Subcellular Spatial Transcriptomics Platforms, December 2024. Pages: 2024.12.23.630033 Section: New Results.
- [61] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, Li Zeng, Huazhu Fu, Min Wu, Lina Hsiu Kim Lim, Longqi Liu, and Jinmiao Chen. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14(1):1155, March 2023.

- [62] Yawei Li and Yuan Luo. STdGCN: spatial transcriptomic cell-type deconvolution using graph convolutional networks. *Genome Biology*, 25(1):206, August 2024.
- [63] Sophia K. Longo, Margaret G. Guo, Andrew L. Ji, and Paul A. Khavari. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature reviews. Genetics*, 22(10):627–644, October 2021.
- [64] Xiaomeng Wan, Jiashun Xiao, Sindy Sing Ting Tam, Mingxuan Cai, Ryohichi Sugimura, Yang Wang, Xiang Wan, Zhixiang Lin, Angela Ruohao Wu, and Can Yang. Integrating spatial and single-cell transcriptomics data using deep generative models with SpatialScope. *Nature Communications*, 14(1):7848, November 2023. Publisher: Nature Publishing Group.
- [65] Haoyu Wang, Peng Cheng, Juan Wang, Hongzhi Lv, Jie Han, Zhiyong Hou, Ren Xu, and Wei Chen. Advances in spatial transcriptomics and its application in the musculoskeletal system. *Bone Research*, 13:54, May 2025.
- [66] Tuan Vo, P Prakrithi, Kahli Jones, Sohye Yoon, Pui Yeng Lam, Yung-Ching Kao, Ning Ma, Samuel X Tan, Xinnan Jin, Chenhao Zhou, Joanna Crawford, Shaun Walters, Ishaan Gupta, Peter H Soyer, Kiarash Khosrotehrani, Mitchell S Stark, and Quan Nguyen. Assessing spatial sequencing and imaging approaches to capture the molecular and pathological heterogeneity of archived cancer tissues. *The Journal of Pathology*, 265(3):274–288, March 2025.
- [67] Julienne L. Carstens, Santhoshi N Krishnan, Arvind Rao, Anna G. Sorace, Erin H. Seeley, Sammy Ferri-Borgogno, and Jared K Burks. Spatial Multiplexing and Omics. *Nature reviews. Methods primers*, 4(1):54, 2024.
- [68] Shuangfang Fang, Bichao Chen, Yong Zhang, Haixi Sun, Longqi Liu, Shiping Liu, Yuxiang Li, and Xun Xu. Computational Approaches and Challenges in Spatial Transcriptomics. *Genomics, Proteomics & Bioinformatics*, 21(1):24–47, February 2023.
- [69] S. Vickovic, B. Lötstedt, J. Klughammer, S. Mages, Å Segerstolpe, O. Rozenblatt-Rosen, and A. Regev. SM-Omics is an automated platform for high-throughput spatial multi-omics. *Nature Communications*, 13(1):795, February 2022.
- [70] Olivia Habern. Simplifying spatial transcriptomics protocols with Visium CytAssist.

- [71] CosMx SMI - Single-Cell Imaging, 2025.
- [72] Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A. Morrison, Michelli F. Oliveira, Jordan T. Sicherman, Andrew Kohlway, Jawad Abousoud, Tingsheng Yu Drennon, Seayar H. Mohabbat, and Sarah E. B. Taylor. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, December 2023. Publisher: Nature Publishing Group.
- [73] Yanxiang Deng, Marek Bartosovic, Petra Kukanja, Di Zhang, Yang Liu, Graham Su, Archibald Enniful, Zhiliang Bai, Gonalo Castelo-Branco, and Rong Fan. Spatial-CUT&Tag: Spatially resolved chromatin modification profiling at the cellular level. *Science (New York, N.Y.)*, 375(6581):681–686, February 2022.
- [74] Yanxiang Deng, Marek Bartosovic, Sai Ma, Di Zhang, Yang Liu, Xiaoyu Qin, Graham Su, Mina L. Xu, Stephanie Halene, Joseph E. Craft, Gonalo Castelo-Branco, and Rong Fan. Spatial-ATAC-seq: spatially resolved chromatin accessibility profiling of tissues at genome scale and cellular level, June 2021. Pages: 2021.06.06.447244 Section: New Results.
- [75] Yanxiang Deng, Marek Bartosovic, Sai Ma, Di Zhang, Petra Kukanja, Yang Xiao, Graham Su, Yang Liu, Xiaoyu Qin, Gorazd B. Rosoklija, Andrew J. Dwork, J. John Mann, Mina L. Xu, Stephanie Halene, Joseph E. Craft, Kam W. Leong, Maura Boldrini, Gonalo Castelo-Branco, and Rong Fan. Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature*, 609(7926):375–383, September 2022. Publisher: Nature Publishing Group.
- [76] Maxime Dhainaut, Samuel A. Rose, Guray Akturk, Aleksandra Wroblewska, Sebastian R. Nielsen, Eun Sook Park, Mark Buckup, Vladimir Roudko, Luisanna Pia, Robert Sweeney, Jessica Le Berichel, C. Matthias Wilk, Anela Bektesevic, Brian H. Lee, Nina Bhardwaj, Adeeb H. Rahman, Alessia Baccarini, Sacha Gnjatich, Dana Pe’er, Miriam Merad, and Brian D. Brown. Spatial CRISPR genomics identifies regulators of the tumor microenvironment. *Cell*, 185(7):1223–1239.e20, March 2022.
- [77] Xiaoli Zhang, Yirui Huang, Yajing Yang, Qi-En Wang, and Lang Li. Advancements in prospective single-cell lineage barcoding and their applications in research.

- [78] Michael Ratz, Leonie von Berlin, Ludvig Larsson, Marcel Martin, Jakub Orzechowski Westholm, Gioele La Manno, Joakim Lundeberg, and Jonas Frisén. Clonal relations in the mouse brain revealed by single-cell and spatial transcriptomics. *Nature Neuroscience*, 25(3):285–294, March 2022. Publisher: Nature Publishing Group.
- [79] Hu Zeng, Jiahao Huang, Jingyi Ren, Connie Kangni Wang, Zefang Tang, Haowen Zhou, Yiming Zhou, Hailing Shi, Abhishek Aditham, Xin Sui, Hongyu Chen, Jennifer A. Lo, and Xiao Wang. Spatially resolved single-cell translomics at molecular resolution. *Science (New York, N.Y.)*, 380(6652):eadd3067, June 2023.
- [80] Muhammad Sohail Sajid, Sadia Ghazanfer, Hafiz Muhammad Rizwan, Muhammad Haziq Bajwa, Nimra Anwar, and Muhammad Saqib. Introduction to Omics Technologies. In *Omics Approaches in Veterinary Parasitology*. CRC Press, 2024. Num Pages: 25.
- [81] Wei Vivian Li and Jingyi Jessica Li. Modeling and analysis of RNA-seq data: a review from a statistical perspective. *Quantitative Biology*, 6(3):195–209, 2018. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1007/s40484-018-0144-7>.
- [82] Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-Dimensional Statistics with a View Toward Applications in Biology. *Annual Review of Statistics and Its Application*, 1(Volume 1, 2014):255–278, January 2014. Publisher: Annual Reviews.
- [83] Shuang Ge, Shuqing Sun, Huan Xu, Qiang Cheng, and Zhixiang Ren. Deep learning in single-cell and spatial transcriptomics data analysis: advances and challenges from a data science perspective. *Briefings in Bioinformatics*, 26(2):bbaf136, April 2025.
- [84] Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biology*, 23(1):31, January 2022.
- [85] Cameron G. Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraf Haque. An introduction to spatial transcriptomics for biomedical research.

Genome Medicine, 14(1):68, June 2022.

- [86] Zexian Zeng, Yawei Li, Yiming Li, and Yuan Luo. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biology*, 23(1):83, March 2022.
- [87] Jonathan Karin, Reshef Mintz, Barak Raveh, and Mor Nitzan. Interpreting single-cell and spatial omics data using deep neural network training dynamics. *Nature Computational Science*, 4(12):941–954, December 2024.
- [88] Yue You, Yuting Fu, Lanxiang Li, Zhongmin Zhang, Shikai Jia, Shihong Lu, Wenle Ren, Yifang Liu, Yang Xu, Xiaojing Liu, Fuqing Jiang, Guangdun Peng, Abhishek Sampath Kumar, Matthew E. Ritchie, Xiaodong Liu, and Luyi Tian. Systematic comparison of sequencing-based spatial transcriptomic methods. *Nature Methods*, 21(9):1743–1754, September 2024. Publisher: Nature Publishing Group.
- [89] Namyoung Jung and Tae-Kyung Kim. Spatial transcriptomics in neuroscience. *Experimental & Molecular Medicine*, 55(10):2105–2115, October 2023.
- [90] Hyun Ju Lim, Ye Wang, Anton Buzdin, and Xinmin Li. A practical guide for choosing an optimal spatial transcriptomics technology from seven major commercially available options. *BMC Genomics*, 26(1):47, January 2025.
- [91] Zixuan Cang, Yanxiang Zhao, Axel A. Almet, Adam Stabell, Raul Ramos, Maksim V. Plikus, Scott X. Atwood, and Qing Nie. Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nature Methods*, 20(2):218–228, February 2023. Publisher: Nature Publishing Group.
- [92] Duy Pham, Xiao Tan, Brad Balderson, Jun Xu, Laura F. Grice, Sohye Yoon, Emily F. Willis, Minh Tran, Pui Yeng Lam, Arti Raghubar, Priyakshi Kalita-de Croft, Sunil Lakhani, Jana Vukovic, Marc J. Ruitenber, and Quan H. Nguyen. Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nature Communications*, 14(1):7739, November 2023. Publisher: Nature Publishing Group.
- [93] Yang Jin, Yuanli Zuo, Gang Li, Wenrong Liu, Yitong Pan, Ting Fan, Xin Fu, Xiaojun Yao, and Yong Peng. Advances in spatial transcriptomics and its applications in cancer research. *Molecular Cancer*, 23(1):129, June 2024.

- [94] Haoyang Li, Juexiao Zhou, Zhongxiao Li, Siyuan Chen, Xingyu Liao, Bin Zhang, Ruochi Zhang, Yu Wang, Shiwei Sun, and Xin Gao. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nature Communications*, 14(1):1548, March 2023.
- [95] Muiz Khan, Suzan Arslanturk, and Sorin Draghici. A comprehensive review of spatial transcriptomics data alignment and integration. *Nucleic Acids Research*, 53(12):gkaf536, July 2025.
- [96] Lucie C. Gaspard-Boulinc, Luca Gortana, Thomas Walter, Emmanuel Barillot, and Florence M. G. Cavalli. Cell-type deconvolution methods for spatial transcriptomics. *Nature Reviews Genetics*, pages 1–19, May 2025. Publisher: Nature Publishing Group.
- [97] Mengnan Cheng, Yujia Jiang, Jiangshan Xu, Alexios-Fotios A. Mentis, Shuai Wang, Huiwen Zheng, Sunil Kumar Sahu, Longqi Liu, and Xun Xu. Spatially resolved transcriptomics: a comprehensive review of their technological advances, applications, and challenges. *Journal of Genetics and Genomics*, 50(9):625–640, September 2023.
- [98] Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, Ke Liu, Xuetong Zhao, Jun Lin, Linzhao Cheng, Falai Chen, Tian Xue, and Kun Qu. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, 19(6):662–670, June 2022. Publisher: Nature Publishing Group.
- [99] Alma Andersson, Joseph Bergenstråhle, Michaela Asp, Ludvig Bergenstråhle, Aleksandra Jurek, José Fernández Navarro, and Joakim Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology*, 3(1):1–8, October 2020.
- [100] Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526, April 2022.
- [101] Romain Lopez, Baoguo Li, Hadas Keren-Shaul, Pierre Boyeau, Merav Kedmi, David Pilzer, Adam Jelinski, Ido Yofe, Eyal David, Allon Wagner, Can Ergen, Yoseph

- Addadi, Ofra Golani, Franca Ronchese, Michael I. Jordan, Ido Amit, and Nir Yosef. DestVI identifies continuums of cell types in spatial transcriptomics data. *Nature Biotechnology*, 40(9):1360–1369, September 2022.
- [102] Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*, 40(9):1349–1359, September 2022.
- [103] Patrick Danaher, Youngmi Kim, Brenn Nelson, Maddy Griswold, Zhi Yang, Erin Piazza, and Joseph M. Beechem. Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nature Communications*, 13(1):385, January 2022. Publisher: Nature Publishing Group.
- [104] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40(5):661–671, May 2022.
- [105] Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu, Ivo Gut, and Holger Heyn. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9):e50, May 2021.
- [106] Rui Dong and Guo-Cheng Yuan. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biology*, 22(1):145, May 2021.
- [107] Daphne Tsoucas, Rui Dong, Haide Chen, Qian Zhu, Guoji Guo, and Guo-Cheng Yuan. Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, 10(1):2975, July 2019.
- [108] Qianqian Song and Jing Su. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in Bioinformatics*, 22(5):bbaa414, September 2021.
- [109] Haoyang Li, Hanmin Li, Juexiao Zhou, and Xin Gao. SD2: spatially resolved transcriptomics deconvolution through integration of dropout and spatial information. *Bioinformatics (Oxford, England)*, 38(21):4878–4884, October 2022.

- [110] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11):1352–1362, November 2021.
- [111] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, 11:2084, April 2020.
- [112] Noa Moriel, Enes Senel, Nir Friedman, Nikolaus Rajewsky, Nikos Karaiskos, and Mor Nitzan. NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*, 16(9):4177–4200, September 2021.
- [113] Jiawen Chen, Weifang Liu, Tianyou Luo, Zhentao Yu, Minzhi Jiang, Jia Wen, Gaurav P. Gupta, Paola Giusti, Hongtu Zhu, Yuchen Yang, and Yun Li. A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. *Briefings in Bioinformatics*, 23(4):bbac245, July 2022.
- [114] Emelie Berglund, Jonas Maaskola, Niklas Schultz, Stefanie Friedrich, Maja Marklund, Joseph Bergenstråhle, Firas Tarish, Anna Tanoglidi, Sanja Vickovic, Ludvig Larsson, Fredrik Salmén, Christoph Ogris, Karolina Wallenborg, Jens Lagergren, Patrik Ståhl, Erik Sonnhhammer, Thomas Helleday, and Joakim Lundberg. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nature Communications*, 9(1):2419, June 2018. Publisher: Nature Publishing Group.
- [115] Benjamin Chidester, Tianming Zhou, Shahul Alam, and Jian Ma. SpiceMix enables integrative single-cell spatial modeling of cell identity. *Nature Genetics*, 55(1):78–88, January 2023. Publisher: Nature Publishing Group.
- [116] Brendan F. Miller, Feiyang Huang, Lyla Atta, Arpan Sahoo, and Jean Fan. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nature Communications*, 13(1):2339, April 2022.

- [117] Chen Xi Yang, Don D. Sin, and Raymond T. Ng. SMART: spatial transcriptomics deconvolution using marker-gene-assisted topic model. *Genome Biology*, 25(1):304, December 2024.
- [118] Yinghao Fu, Leqi Tian, and Weiwei Zhang. STsisal: a reference-free deconvolution pipeline for spatial transcriptomics data. *Frontiers in Genetics*, 16:1512435, 2025.
- [119] Zixiang Zhou, Yunshan Zhong, Zemin Zhang, and Xianwen Ren. Spatial transcriptomics deconvolution at single-cell resolution using Redeconve. *Nature Communications*, 14(1):7930, December 2023. Publisher: Nature Publishing Group.
- [120] Yunlu Chen, Feng Ruan, and Ji-Ping Wang. NLSDeconv: an efficient cell-type deconvolution method for spatial transcriptomics data. *Bioinformatics*, 41(1):btae747, December 2024.
- [121] Jiawei Li, Yi Zhang, Cheng Yang, and Ruiming Rong. Discrepant mRNA and Protein Expression in Immune Cells. *Current Genomics*, 21(8):560–563, December 2020.
- [122] Johan Reimegård, Marcel Tarbier, Marcus Danielsson, Jens Schuster, Sathishkumar Baskaran, Styliani Panagiotou, Niklas Dahl, Marc R. Friedländer, and Caroline J. Gallant. A combined approach for single-cell mRNA and intracellular protein expression analysis. *Communications Biology*, 4(1):624, May 2021. Publisher: Nature Publishing Group.
- [123] Mengyao Wu, Huihui Tao, Tiantian Xu, Xuejia Zheng, Chunmei Wen, Guoying Wang, Yali Peng, and Yong Dai. Spatial proteomics: unveiling the multidimensional landscape of protein localization in human diseases. *Proteome Science*, 22:7, September 2024.
- [124] Takashi Semba and Takatsugu Ishimoto. Spatial analysis by current multiplexed imaging technologies for the molecular characterisation of cancer tissues. *British Journal of Cancer*, 131(11):1737–1747, December 2024. Publisher: Nature Publishing Group.
- [125] Peter Horvath and Fabian Coscia. Spatial proteomics in translational and clinical research. *Molecular Systems Biology*, 21(6):526–530, April 2025.

- [126] John W. Hickey, Elizabeth K. Neumann, Andrea J. Radtke, Jeannie M. Camarillo, Rebecca T. Beuschel, Alexandre Albanese, Elizabeth McDonough, Julia Hatler, Anne E. Wiblin, Jeremy Fisher, Josh Croteau, Eliza C. Small, Anup Sood, Richard M. Caprioli, R. Michael Angelo, Garry P. Nolan, Kwanghun Chung, Stephen M. Hewitt, Ronald N. Germain, Jeffrey M. Spraggins, Emma Lundberg, Michael P. Snyder, Neil L. Kelleher, and Sinem K. Saka. Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. *Nature methods*, 19(3):284–295, March 2022.
- [127] Beiyu Hu, Junjie Zhu, and Fangqing Zhao. The evolving landscape of spatial proteomics technologies in the AI age. *Fundamental Research*, December 2024.
- [128] Wilson Kuswanto, Garry Nolan, and Guolan Lu. Highly multiplexed spatial profiling with CODEX: bioinformatic analysis and application in human disease. *Seminars in Immunopathology*, 45(1):145–157, 2023.
- [129] Sarah Black, Darci Phillips, John W. Hickey, Julia Kennedy-Darling, Vishal G. Venkataraman, Nikolay Samusik, Yury Goltsev, Christian M. Schürch, and Garry P. Nolan. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nature Protocols*, 16(8):3802–3835, August 2021.
- [130] Kausalia Vijayaragavan, Bryan J. Cannon, Dmitry Tebaykin, Marc Bossé, Alex Baranski, J. P. Oliveria, Syed A. Bukhari, Dunja Mrdjen, M. Ryan Corces, Erin F. McCaffrey, Noah F. Greenwald, Yari Sigal, Diana Marquez, Zumana Khair, Trevor Bruce, Mako Goldston, Anusha Bharadwaj, Kathleen S. Montine, R. Michael Angelo, Thomas J. Montine, and Sean C. Bendall. Single-cell spatial proteomic imaging for human neuropathology. *Acta Neuropathologica Communications*, 10:158, November 2022.
- [131] Method of the Year 2024: spatial proteomics. *Nature Methods*, 21(12):2195–2196, December 2024. Publisher: Nature Publishing Group.
- [132] Nir Ben-Chetrit, Xiang Niu, Ariel D. Swett, Jesus Sotelo, Maria S. Jiao, Caitlin M. Stewart, Catherine Potenski, Paulius Mielinis, Patrick Roelli, Marlon Stoeckius, and Dan A. Landau. Integration of whole transcriptome spatial profiling with protein markers. *Nature Biotechnology*, 41(6):788–793, June 2023.

- [133] Yang Liu, Marcello DiStasio, Graham Su, Hiromitsu Asashima, Archibald Enniful, Xiaoyu Qin, Yanxiang Deng, Pino Bordignon, Marco Cassano, Mary Tomayko, Mina Xu, Stephanie Halene, Joseph E. Craft, David Hafler, and Rong Fan. Spatial-CITE-seq: spatially resolved high-plex protein and whole transcriptome co-mapping. *Research Square*, pages rs.3.rs-1499315, March 2022.
- [134] Sha Liao, Yang Heng, Weiqing Liu, Jinqiong Xiang, Yong Ma, Ligen Chen, Xiuwen Feng, Dongmei Jia, Diyan Liang, Caili Huang, Jiajun Zhang, Min Jian, Kui Su, Mei Li, Yulin-Han Loh, Ao Chen, and Xun Xu. Integrated Spatial Transcriptomic and Proteomic Analysis of Fresh Frozen Tissue Based on Stereo-seq, April 2023. Pages: 2023.04.28.538364 Section: New Results.
- [135] Sonali Andani, Boqi Chen, Joanna Ficek-Pascual, Simon Heinke, Ruben Casanova, Bettina Sobottka, Bernd Bodenmiller, Viktor Koelzer, and Gunnar Ratsch. Multi-V-Stain: Multiplexed Virtual Staining of Histopathology Whole-Slide Images. October 2023.
- [136] Zilu Zhou, Chengzhong Ye, Jingshu Wang, and Nancy R. Zhang. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications*, 11(1):651, January 2020.
- [137] Azka Javaid and Hildreth Robert Frost. STREAK: A supervised cell surface receptor abundance estimation strategy for single cell RNA-sequencing data using feature selection and thresholded gene set scoring. *PLoS computational biology*, 19(8):e1011413, August 2023.
- [138] Justin Lakkis, Amelia Schroeder, Kenong Su, Michelle Y.Y. Lee, Alexander C. Bashore, Muredach P. Reilly, and Mingyao Li. A multi-use deep learning method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and imputation. *Nature machine intelligence*, 4(11):940–952, November 2022.
- [139] Daniel Hanhart, Federico Gossi, Maria Anna Rapsomaniki, Marianna Kruithof-de Julio, and Panagiotis Chouvardas. ScLinear predicts protein abundance at single-cell resolution. *Communications Biology*, 7(1):1–7, March 2024.
- [140] Ruoqiao Chen, Jiayu Zhou, and Bin Chen. Imputing abundance of over 2,500 surface proteins from single-cell transcriptomes with context-agnostic zero-shot

deep ensembles. *Cell Systems*, 15(9):869–884.e6, September 2024.

- [141] Haoyu Wang, Brittany Cody, and Hatice Ulku Osmanbeyoglu. DGAT: A Dual-Graph Attention Network for Inferring Spatial Protein Landscapes from Transcriptomics, July 2025.
- [142] Nicola Crosetto, Magda Bienko, and Alexander van Oudenaarden. Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics*, 16(1):57–66, January 2015.
- [143] Ed Lein, Lars E. Borm, and Sten Linnarsson. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*, 358(6359):64–69, October 2017.
- [144] R. Lehmann and D. Tautz. In situ hybridization to RNA. *Methods in Cell Biology*, 44:575–598, 1994.
- [145] Ellen Jensen. Technical Review: In Situ Hybridization. *The Anatomical Record*, 297(8):1349–1353, 2014.
- [146] Andrea M. Femino, Fredric S. Fay, Kevin Fogarty, and Robert H. Singer. Visualization of Single RNA Transcripts in Situ. *Science*, 280(5363):585–590, April 1998.
- [147] Susanne Nichterwitz, Geng Chen, Julio Aguila Benitez, Marlene Yilmaz, Helena Storvall, Ming Cao, Rickard Sandberg, Qiaolin Deng, and Eva Hedlund. Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nature Communications*, 7(1):12139, July 2016.
- [148] Dongqing Sun, Zhaoyang Liu, Taiwen Li, Qiu Wu, and Chenfei Wang. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Research*, 50(7):e42, April 2022.
- [149] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231. AAAI Press, 1996. event-place: Portland, Oregon, USA.
- [150] Robert L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, December 1953.

- [151] Congming Shi, Bingtao Wei, Shoulin Wei, Wen Wang, Hai Liu, and Jialei Liu. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, 2021(1):31, February 2021.
- [152] Yudi Pawitan and Jian Huang. Constrained clustering of irregularly sampled spatial data. *Journal of Statistical Computation and Simulation*, 73(12):853–865, December 2003.
- [153] Trung Nghia Vu, Setia Pramana, Stefano Calza, Chen Suo, Donghwan Lee, and Yudi Pawitan. Comprehensive landscape of subtype-specific coding and non-coding RNA transcripts in breast cancer. *Oncotarget*, 7(42):68851–68863, September 2016.
- [154] Genomics 10x. 10x Genomics Visium.
- [155] Sinisa Hrvatin, Daniel R. Hochbaum, M. Aurel Nagy, Marcelo Cicconet, Keiramarie Robertson, Lucas Cheadle, Rapolas Zilionis, Alex Ratner, Rebeca Borges-Monroy, Allon M. Klein, Bernardo L. Sabatini, and Michael E. Greenberg. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nature Neuroscience*, 21(1):120–129, January 2018.
- [156] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastri, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018.
- [157] Joonha Kwon, Junho Kang, Areum Jo, Kayoung Seo, Dohyeon An, Mert Yakup Baykan, Jun Hyeong Lee, Nayoung Kim, Hye Hyeon Eum, Sohyun Hwang, Ji Min Lee, Woong-Yang Park, Hee Jung An, Hae-Ock Lee, Jong-Eun Park, and Jung Kyoony Choi. Single-cell mapping of combinatorial target antigens for CAR switches using logic gates. *Nature Biotechnology*, 41(11):1593–1605, November 2023.
- [158] Huiram Kang, Sohyun Hwang, Haeyoun Kang, Areum Jo, Ji Min Lee, Jung Kyoony Choi, Hee Jung An, and Hae-Ock Lee. Altered tumor signature and T-cell profile after chemotherapy reveal new therapeutic opportunities in high-grade serous ovarian carcinoma. *Cancer Science*, 115(3):989–1000, 2024.

- [159] Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warriar, Florent Ginhoux, Ewan Millar, Joseph E. Powell, Stephen R. Williams, X. Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M. Perou, and Alexander Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53(9):1334–1347, September 2021.
- [160] Michael Papanicolaou, Amelia L. Parker, Michelle Yam, Elysse C. Filipe, Sunny Z. Wu, Jessica L. Chitty, Kaitlin Wyllie, Emmi Tran, Ellie Mok, Audrey Nadalini, Joanna N. Skhinas, Morghan C. Lucas, David Herrmann, Max Nobis, Brooke A. Pereira, Andrew M. K. Law, Lesley Castillo, Kendelle J. Murphy, Anais Zaratzian, Jordan F. Hastings, David R. Croucher, Elgene Lim, Brian G. Oliver, Fatima Valdes Mora, Benjamin L. Parker, David Gallego-Ortega, Alexander Swarbrick, Sandra O'Toole, Paul Timpson, and Thomas R. Cox. Temporal profiling of the breast tumour microenvironment reveals collagen XII as a driver of metastasis. *Nature Communications*, 13(1):4587, August 2022.
- [161] Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam, Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, Jason H. Bielas, and Raphael Gottardo. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39(11):1375–1384, November 2021.
- [162] Aaron M. Newman, Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, Mohammad S. Esfahani, Bogdan A. Luca, David Steiner, Maximilian Diehn, and Ash A. Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7):773–782, July 2019.
- [163] Suoqin Jin, Christian F. Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V. Plikus, and Qing Nie. Inference and analysis of cell-cell communication using CellChat. *Nature Communications*,

12(1):1088, February 2021.

- [164] Yuzhe Li, Jinsong Zhang, Xin Gao, and Qiangfeng Cliff Zhang. Tissue module discovery in single-cell-resolution spatial transcriptomics data via cell-cell interaction-aware cell embedding. *Cell Systems*, 15(6):578–592.e7, June 2024.
- [165] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. Linear Methods for Regression. In Trevor Hastie, Jerome Friedman, and Robert Tibshirani, editors, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pages 41–78. Springer, New York, NY, 2001.
- [166] Bozena Zdaniuk. Ordinary Least-Squares (OLS) Model. In Alex C. Michalos, editor, *Encyclopedia of Quality of Life and Well-Being Research*, pages 4515–4517. Springer Netherlands, Dordrecht, 2014.
- [167] Amy Ralston and Kenna Shaw. Gene Expression Regulates Cell Differentiation. *Nature Education*, 1(1):127–131, 2008.
- [168] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, June 2019.
- [169] Michaela Asp, Stefania Giacomello, Ludvig Larsson, Chenglin Wu, Daniel Fürth, Xiaoyan Qian, Eva Wärdell, Joaquin Custodio, Johan Reimegård, Fredrik Salmén, Cecilia Österholm, Patrik L. Ståhl, Erik Sundström, Elisabet Åkesson, Olaf Bergmann, Magda Bienko, Agneta Månsson-Broberg, Mats Nilsson, Christer Sylvén, and Joakim Lundeberg. A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell*, 179(7):1647–1660.e19, December 2019.
- [170] Ran Zhou, Gaoxia Yang, Yan Zhang, and Yuan Wang. Spatial transcriptomics in development and disease. *Molecular Biomedicine*, 4(1):32, October 2023.
- [171] Chongyue Zhao, Zhongli Xu, Xinjun Wang, Shiyue Tao, William A MacDonald, Kun He, Amanda C Poholek, Kong Chen, Heng Huang, and Wei Chen. Innovative super-resolution in spatial transcriptomics: a transformer model exploiting histology images and spatial gene expression. *Briefings in Bioinformatics*, 25(2):bbae052, March 2024.

- [172] Sanjay Jain and Michael T. Eadon. Spatial transcriptomics in health and disease. *Nature Reviews Nephrology*, 20(10):659–671, October 2024.
- [173] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, May 2015.
- [174] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008.
- [175] Constantin Zackl, Maria Zopoglou, Reto Stauffer, Markus Ausserhofer, Marieke E. Ijsselsteijn, Gregor Sturm, Noel Filipe da Cunha Carvalho de Miranda, and Francesca Finotello. spacedeconv: deconvolution of tissue architecture from spatial transcriptomics, October 2024.
- [176] Jing Wang, Zihao Ma, Steven A. Carr, Philipp Mertins, Hui Zhang, Zhen Zhang, Daniel W. Chan, Matthew J. C. Ellis, R. Reid Townsend, Richard D. Smith, Jason E. McDermott, Xian Chen, Amanda G. Paulovich, Emily S. Boja, Mehdi Mesri, Christopher R. Kinsinger, Henry Rodriguez, Karin D. Rodland, Daniel C. Liebler, and Bing Zhang. Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction *. *Molecular & Cellular Proteomics*, 16(1):121–134, January 2017.
- [177] George L. Gabor Miklos and Ryszard Maleszka. Protein functions and biological contexts. *PROTEOMICS*, 1(2):169–178, 2001.
- [178] Jinsen Zhang, Rui Sun, Yingying Lyu, Chaxian Liu, Ying Liu, Yuan Feng, Minjie Fu, Peter Jih Cheng Wong, Zunguo Du, Tianming Qiu, Yi Zhang, Dongxiao Zhuang, Zhiyong Qin, Yu Yao, Wei Zhu, Tiannan Guo, Wei Hua, Hui Yang, and Ying Mao. Proteomic profiling of gliomas unveils immune and metabolism-driven subtypes with implications for anti-nucleotide metabolism therapy. *Nature Communications*, 15(1):10005, November 2024.
- [179] Hui Zhang, Tao Liu, Zhen Zhang, Samuel H. Payne, Bai Zhang, Jason E. McDermott, Jian-Ying Zhou, Vladislav A. Petyuk, Li Chen, Debjit Ray, Shisheng Sun, Feng Yang, Lijun Chen, Jing Wang, Punit Shah, Seong Won Cha, Paul Aiyetan, Sunghee Woo, Yuan Tian, Marina A. Gritsenko, Therese R. Clauss, Caitlin Choi, Matthew E.

Monroe, Stefani Thomas, Song Nie, Chaochao Wu, Ronald J. Moore, Kun-Hsing Yu, David L. Tabb, David Fenyö, Vineet Bafna, Yue Wang, Henry Rodriguez, Emily S. Boja, Tara Hiltke, Robert C. Rivers, Lori Sokoll, Heng Zhu, Ie-Ming Shih, Leslie Cope, Akhilesh Pandey, Bing Zhang, Michael P. Snyder, Douglas A. Levine, Richard D. Smith, Daniel W. Chan, Karin D. Rodland, Steven A. Carr, Michael A. Gillette, Karl R. Klauser, Eric Kuhn, D. R. Mani, Philipp Mertins, Karen A. Ketchum, Ratna Thangudu, Shuang Cai, Mauricio Oberti, Amanda G. Paulovich, Jeffrey R. Whiteaker, Nathan J. Edwards, Peter B. McGarvey, Subha Madhavan, Pei Wang, Daniel W. Chan, Akhilesh Pandey, Ie-Ming Shih, Hui Zhang, Zhen Zhang, Heng Zhu, Leslie Cope, Gordon A. Whiteley, Steven J. Skates, Forest M. White, Douglas A. Levine, Emily S. Boja, Christopher R. Kinsinger, Tara Hiltke, Mehdi Mesri, Robert C. Rivers, Henry Rodriguez, Kenna M. Shaw, Stephen E. Stein, David Fenyö, Tao Liu, Jason E. McDermott, Samuel H. Payne, Karin D. Rodland, Richard D. Smith, Paul Rudnick, Michael Snyder, Yingming Zhao, Xian Chen, David F. Ransohoff, Andrew N. Hoofnagle, Daniel C. Liebler, Melinda E. Sanders, Zhiao Shi, Robbert J. C. Slebos, David L. Tabb, Bing Zhang, Lisa J. Zimmerman, Yue Wang, Sherri R. Davies, Li Ding, Matthew J. C. Ellis, and R. Reid Townsend. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*, 166(3):755–765, July 2016.

- [180] Alexander Franks, Edoardo Airoidi, and Nikolai Slavov. Post-transcriptional regulation across human tissues. *PLOS Computational Biology*, 13(5):e1005535, May 2017.
- [181] Qian Zhong, Xina Xiao, Yijie Qiu, Zhiqiang Xu, Chunyu Chen, Baochen Chong, Xinjun Zhao, Shan Hai, Shuangqing Li, Zhenmei An, and Lunzhi Dai. Protein post-translational modifications in health and diseases: Functions, regulatory mechanisms, and therapeutic implications. *MedComm*, 4(3):e261, 2023.
- [182] Si-yu Jing, He-qi Wang, Ping Lin, Jiao Yuan, Zhi-xuan Tang, and Hong Li. Quantifying and interpreting biologically meaningful spatial signatures within tumor microenvironments. *npj Precision Oncology*, 9(1):68, March 2025.
- [183] Yanfen Xu, Xi Wang, Yuan Li, Yiheng Mao, Yiran Su, Yize Mao, Yun Yang, Weina Gao, Changying Fu, Wendong Chen, Xueting Ye, Fuchao Liang, Panzhu Bai, Ying Sun, Shengping Li, Ruilian Xu, and Ruijun Tian. Multimodal single cell-resolved spatial proteomics reveal pancreatic tumor heterogeneity. *Nature Communications*,

15(1):10100, November 2024.

- [184] Andreas Heindl, Sidra Nawaz, and Yinyin Yuan. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Laboratory Investigation*, 95(4):377–384, April 2015.
- [185] Lei Cao, Chao Yang, Luni Hu, Wenjian Jiang, Yating Ren, Tianyi Xia, Mengyang Xu, Yishuai Ji, Mei Li, Xun Xu, Yuxiang Li, Yong Zhang, and Shuangfang Fang. Deciphering spatial domains from spatially resolved transcriptomics with Siamese graph autoencoder. *GigaScience*, 13:giae003, January 2024.
- [186] Michael Ritter, Christina Blume, Yiheng Tang, Areeba Patel, Bhuvic Patel, Natalie Berghaus, Jasim Kada Benotmane, Jan Kueckelhaus, Yahaya Yabo, Junyi Zhang, Elena Grabis, Giulia Villa, David Niklas Zimmer, Amir Khriesh, Philipp Sievers, Zaira Seferbekova, Felix Hinz, Vidhya M. Ravi, Marcel Seiz-Rosenhagen, Miriam Ratliff, Christel Herold-Mende, Oliver Schnell, Juergen Beck, Wolfgang Wick, Andreas von Deimling, Moritz Gerstung, Dieter Henrik Heiland, and Felix Sahm. Spatially resolved transcriptomics and graph-based deep learning improve accuracy of routine CNS tumor diagnostics. *Nature Cancer*, 6(2):292–306, February 2025.
- [187] Hervé Abdi. Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics*, 2(1):97–106, 2010.
- [188] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794. Association for Computing Machinery, 2016. event-place: New York, NY, USA.
- [189] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [190] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 6639–6649. Curran Associates Inc., 2018. event-place: Red Hook, NY, USA.

- [191] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. CatBoost: gradient boosting with categorical features support, October 2018. arXiv: 1810.11363 [cs].
- [192] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks, February 2018. arXiv: 1710.10903 [stat].
- [193] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [194] Giuseppe Palermo, Paolo Piraino, and Hans-Dieter Zucht. Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. *Advances and Applications in Bioinformatics and Chemistry*, 2:57–70, May 2009.
- [195] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967, March 2021.
- [196] Andreea Anghel, Nikolaos Papandreou, Thomas Parnell, Alessandro De Palma, and Haralampos Pozidis. Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms, January 2019. arXiv: 1809.04559 [cs].
- [197] Rohit Shukla and Tiratha Raj Singh. AlzGenPred - CatBoost-based gene classifier for predicting Alzheimer’s disease using high-throughput sequencing data. *Scientific Reports*, 14(1):30294, December 2024.
- [198] Aleksandr Kotov, Andrei Zinovyev, and Anne-Helene Monsoro-Burq. scEvoNet: a gradient boosting-based method for prediction of cell state evolution. *BMC Bioinformatics*, 24(1):83, March 2023.
- [199] Ping Xuan, Chang Sun, Tiangang Zhang, Yilin Ye, Tonghui Shen, and Yihua Dong. Gradient Boosting Decision Tree-Based Method for Predicting Interactions Between Target Genes and Drugs. *Frontiers in Genetics*, 10:459, 2019.
- [200] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The Graph Neural Network Model. *IEEE Transactions on Neural*

Networks, 20(1):61–80, January 2009.

- [201] Eduard Chelebian, Christophe Avenel, and Carolina Wählby. Combining spatial transcriptomics with tissue morphology. *Nature Communications*, 16(1):4452, May 2025.
- [202] Paul Kiessling and Christoph Kuppe. Spatial multi-omics: novel tools to study the complexity of cardiovascular diseases. *Genome Medicine*, 16(1):14, January 2024.
- [203] Qingyue Wang, Parth Khatri, Huy Q Dinh, Jian Huang, Yudi Pawitan, and Trung Nghia Vu. A cluster-based cell-type deconvolution of spatial transcriptomic data. *Nucleic Acids Research*, 53(14):gkaf714, August 2025.

Appendix: Simulation Results

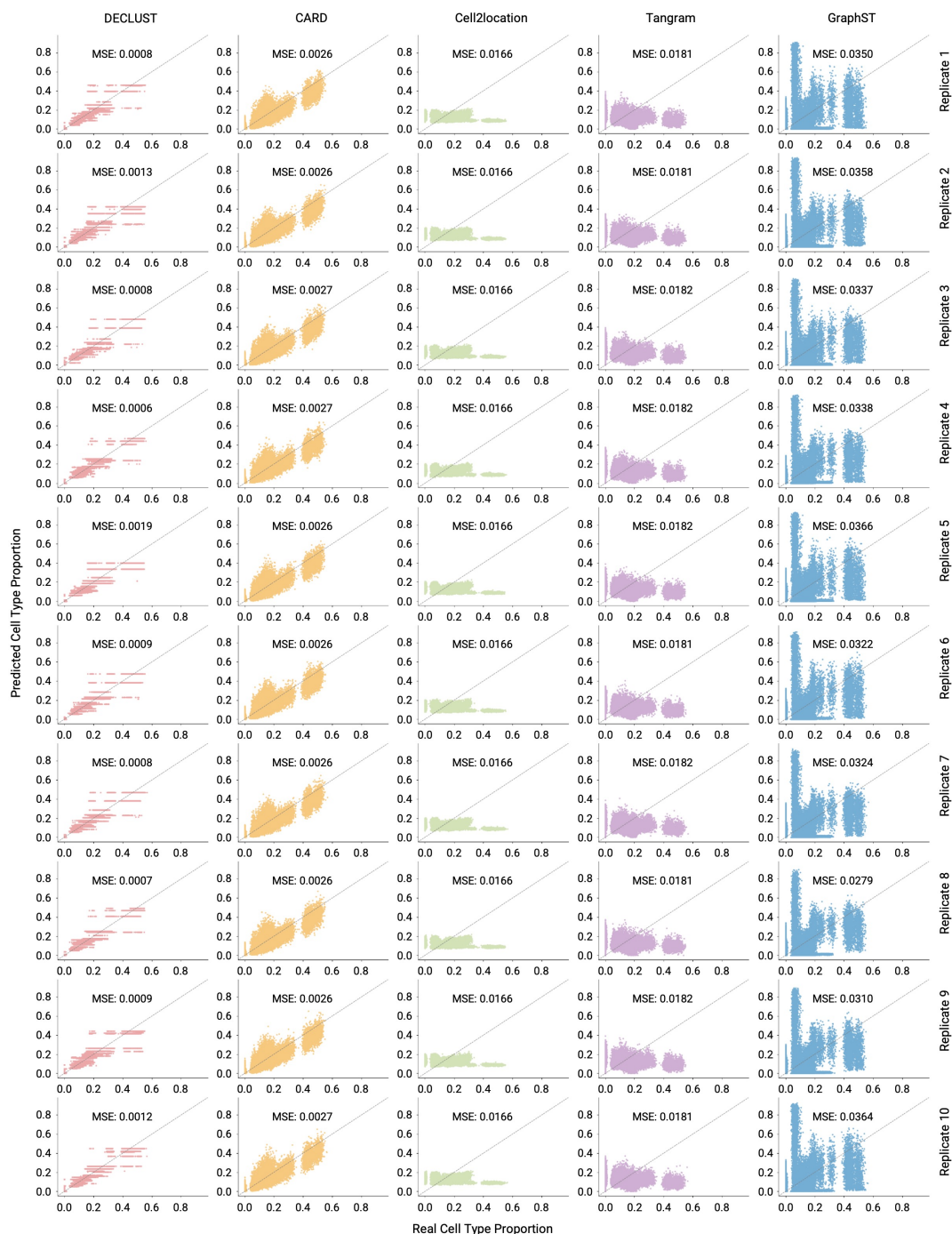


Fig. A.1.: Accuracy of cell-type proportion predictions across multiple methods in Simulation 1.

In Fig.A.1, each scatterplot compares the predicted and true cell type proportions for DECLUST, CARD, Cell2location, Tangram, and GraphST across 10 independent replicates under the Simulation 1 framework. Points closer to the diagonal line indicate more accurate predictions. DECLUST achieves the highest agreement with the ground truth, with mean squared error (MSE) values ranging from 0.0006 to 0.0019, indicating exceptional precision. CARD also performs well, with consistent MSEs around 0.0026–0.0027. In contrast, Cell2location and Tangram show moderate performance, each producing stable but higher MSEs (0.0166 and 0.0181, respectively). GraphST displays the largest prediction error and spread, with MSEs ranging from 0.0279 to 0.0366 and predictions frequently diverging from the diagonal. These results underscore the superior accuracy of DECLUST in settings that align with its modeling assumptions, highlighting its strength in capturing true cell type proportions under idealized simulation conditions.

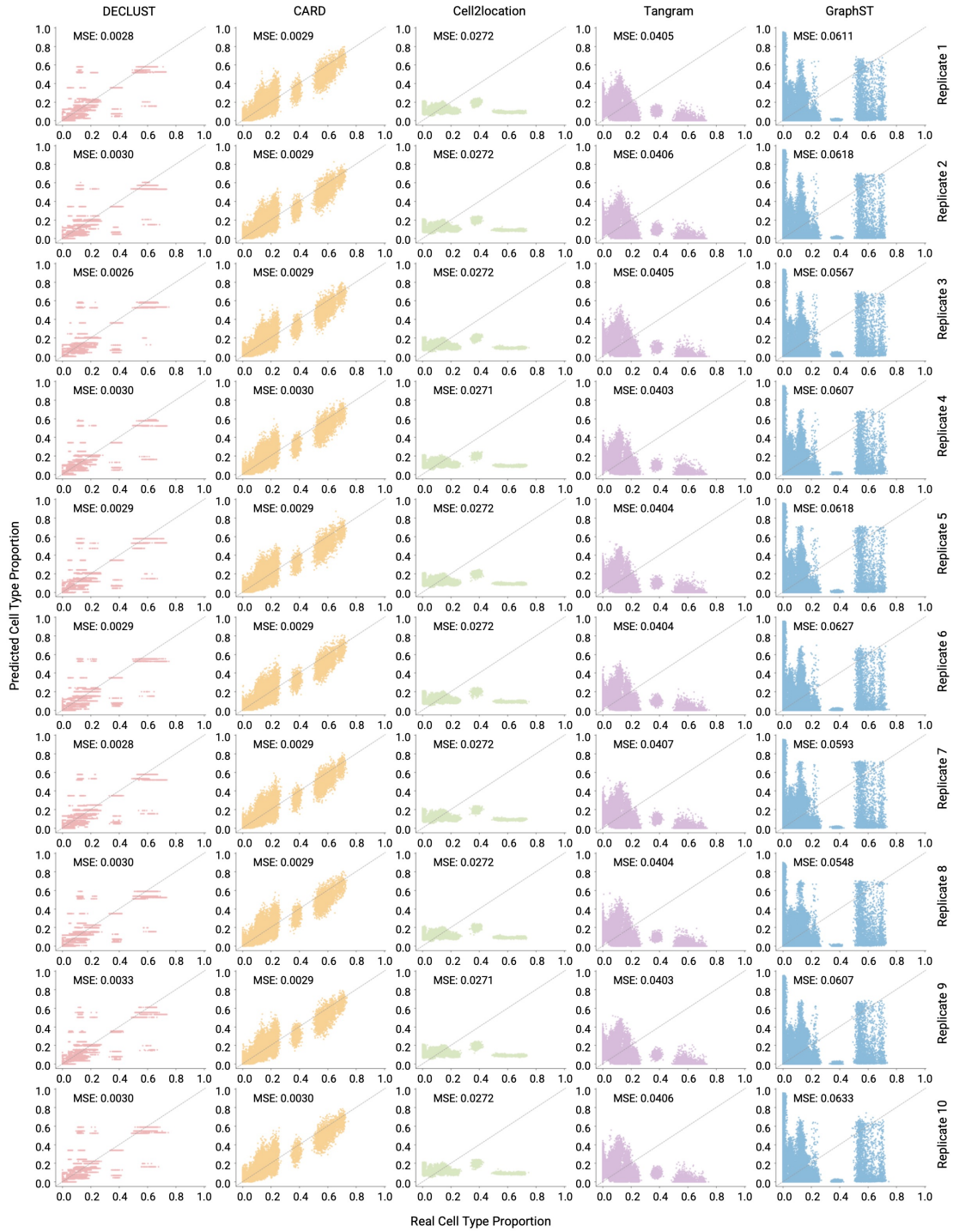


Fig. A.2.: Accuracy of cell-type proportion predictions across multiple methods in Simulation 2.

In Fig. A.2, scatterplots compare the predicted and true cell type proportions for five deconvolution methods—DECLUST, CARD, Cell2location, Tangram, and GraphST—across 10 replicates datasets in the Simulation 2 setting. Each point represents a capture spot, and the

dashed diagonal line indicates perfect agreement between predicted and real proportions. DECLUS and CARD show close alignment with the diagonal, with notably low mean squared error ($MSE \approx 0.0028\text{--}0.0033$), indicating high accuracy in cell type prediction. Cell2location and Tangram exhibit moderate deviations, with MSE values around 0.0272 and 0.0404, respectively. GraphST displays the largest discrepancies, with MSEs frequently exceeding 0.06 and greater dispersion from the diagonal. These results further confirm the strong predictive performance of DECLUS under Simulation 2, even in a more challenging scenario with alternative assumptions.

Appendix: Spatial Visualization of Predicted Protein Abundance

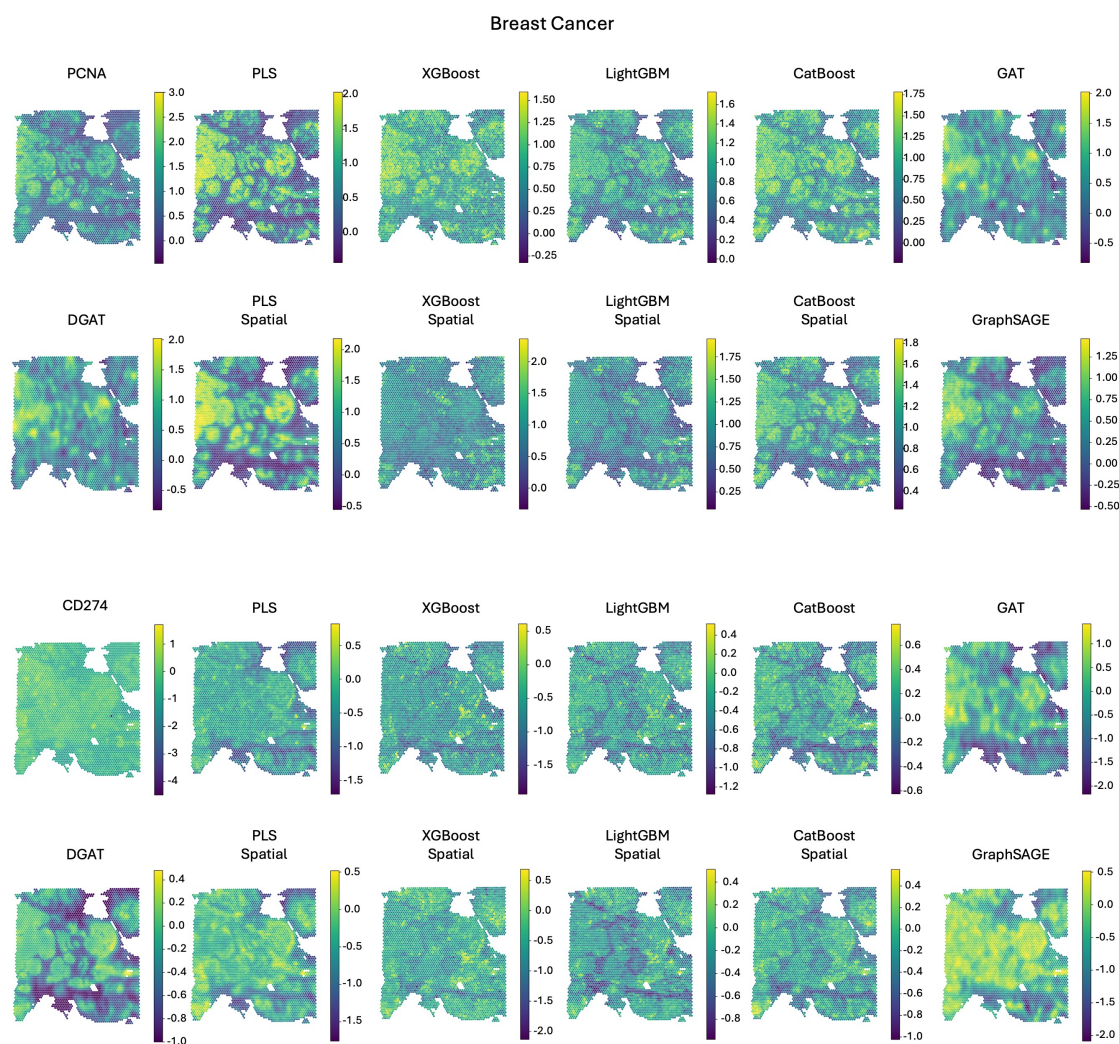


Fig. B.1.: Cross-tissue prediction of PCNA and CD274 expression in breast cancer. PCNA2, a nuclear marker associated with cell proliferation, displayed distinct high-expression clusters corresponding to tumor regions. Boosting-based models, particularly when incorporating spatial features, effectively reconstructed these clusters, whereas non-spatial GNNs exhibited reduced local contrast. CD274 (PD-L1), with a more diffuse expression pattern, was predicted with moderate accuracy across models, though spatially informed approaches provided improved delineation of expression gradients compared with their non-spatial counterparts.

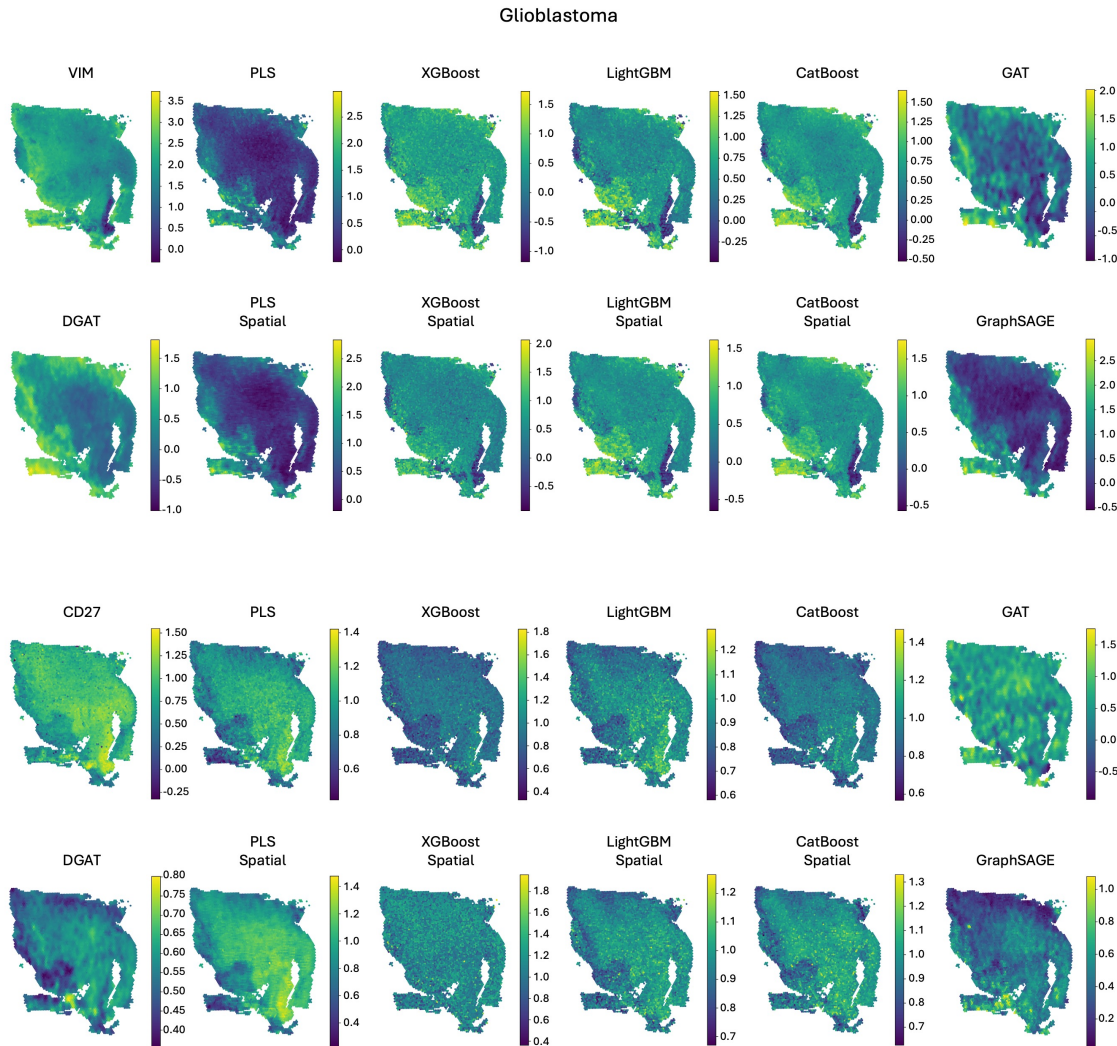


Fig. B.2.: Cross-tissue prediction of VIM and CD27 expression in glioblastoma. VIM (vimentin), a cytoskeletal marker of mesenchymal phenotype, showed broad expression across tumor and peritumoral regions. Spatially informed boosting models achieved the most accurate reconstruction, capturing both regional gradients and high-intensity zones, whereas GNN predictions tended to under-represent finer spatial details. CD27, expressed on activated T cells, exhibited a localized distribution that was better preserved by models integrating spatial features, while non-spatial methods produced more homogeneous but less accurate predictions.

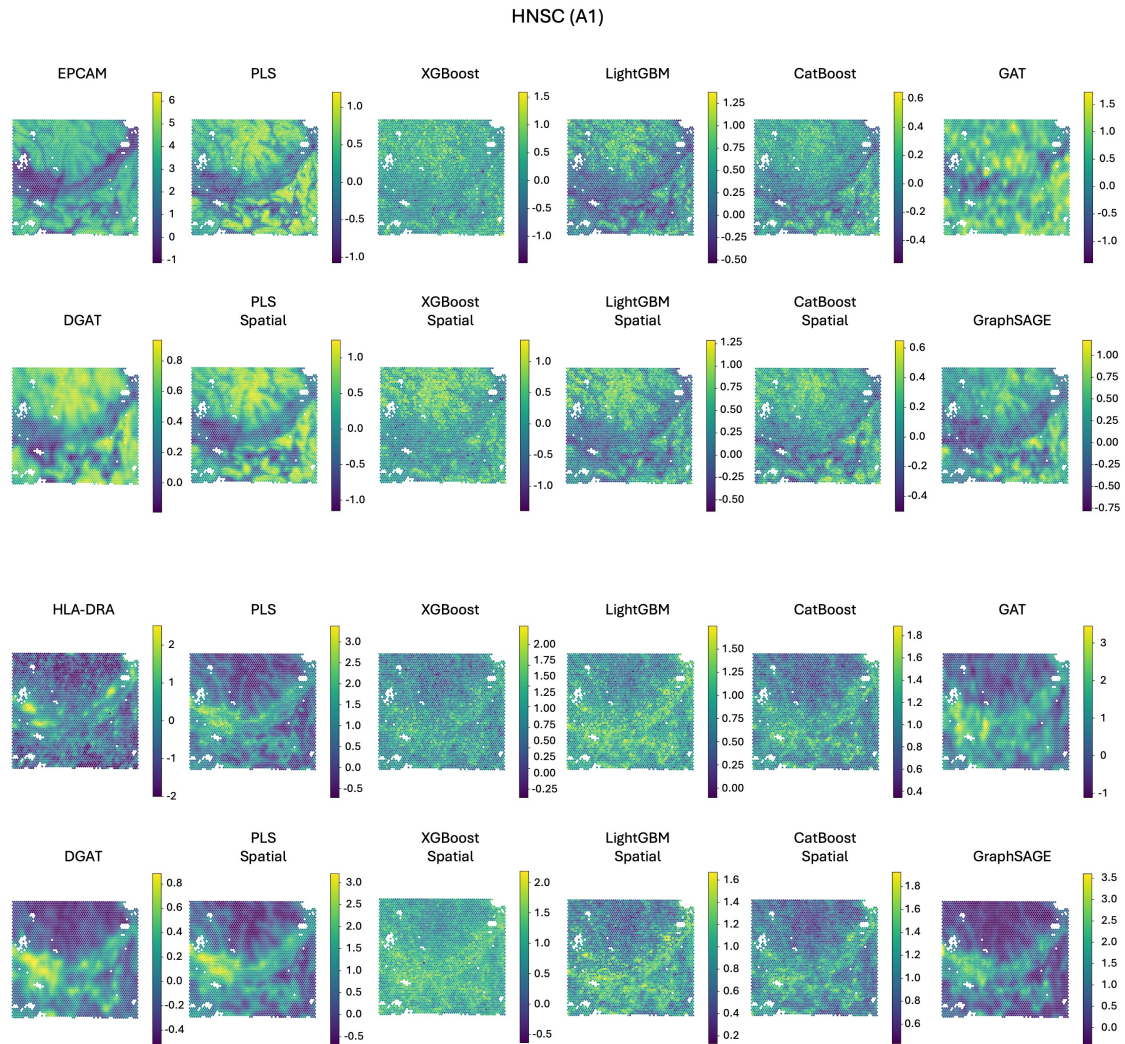


Fig. B.3.: Cross-tissue prediction of EPCAM and HLA-DRA expression in head and neck squamous cell carcinoma (HNSC, A1). EPCAM, an epithelial cell adhesion marker, displayed sharply localized expression in tumor epithelial regions. Spatially enhanced boosting models produced highly accurate reconstructions, closely matching the ground truth, while non-spatial models showed blurring at tissue boundaries. HLA-DRA, an MHC class II molecule, had a more heterogeneous expression pattern across immune cell-rich areas; spatial models outperformed non-spatial ones in identifying discrete high-expression zones, reducing over-smoothing effects.

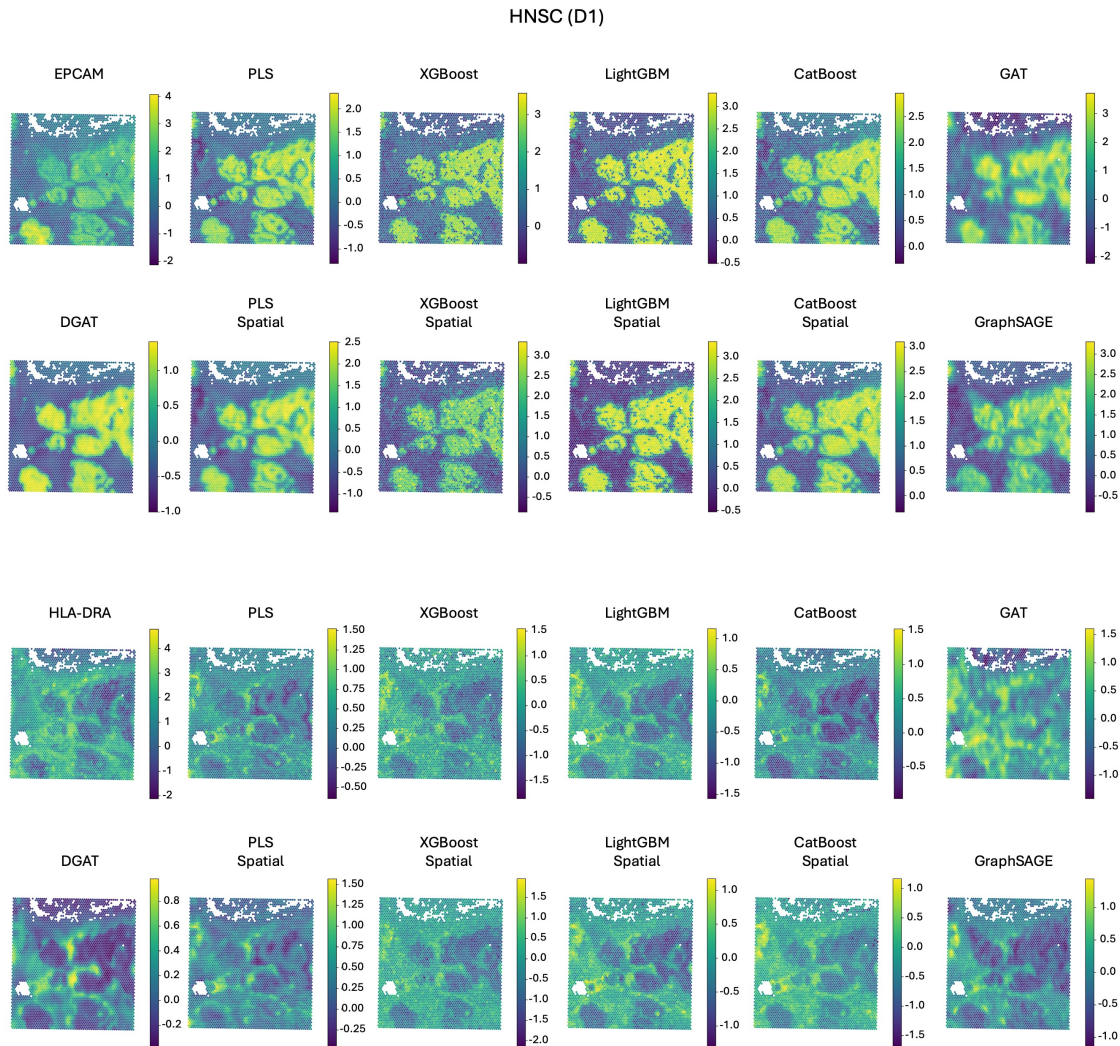


Fig. B.4.: Cross-tissue prediction of EPCAM and HLA-DRA expression in head and neck squamous cell carcinoma (HNSC, D1). For EPCAM, strong epithelial localization was accurately recapitulated by most boosting-based models, particularly with spatial feature integration, whereas GNNs without explicit spatial inputs yielded less distinct regional contrasts. HLA-DRA exhibited a patchy distribution linked to immune infiltrates; spatially informed models demonstrated superior delineation of these regions, reducing background noise and better preserving local intensity variations compared with non-spatial predictions.

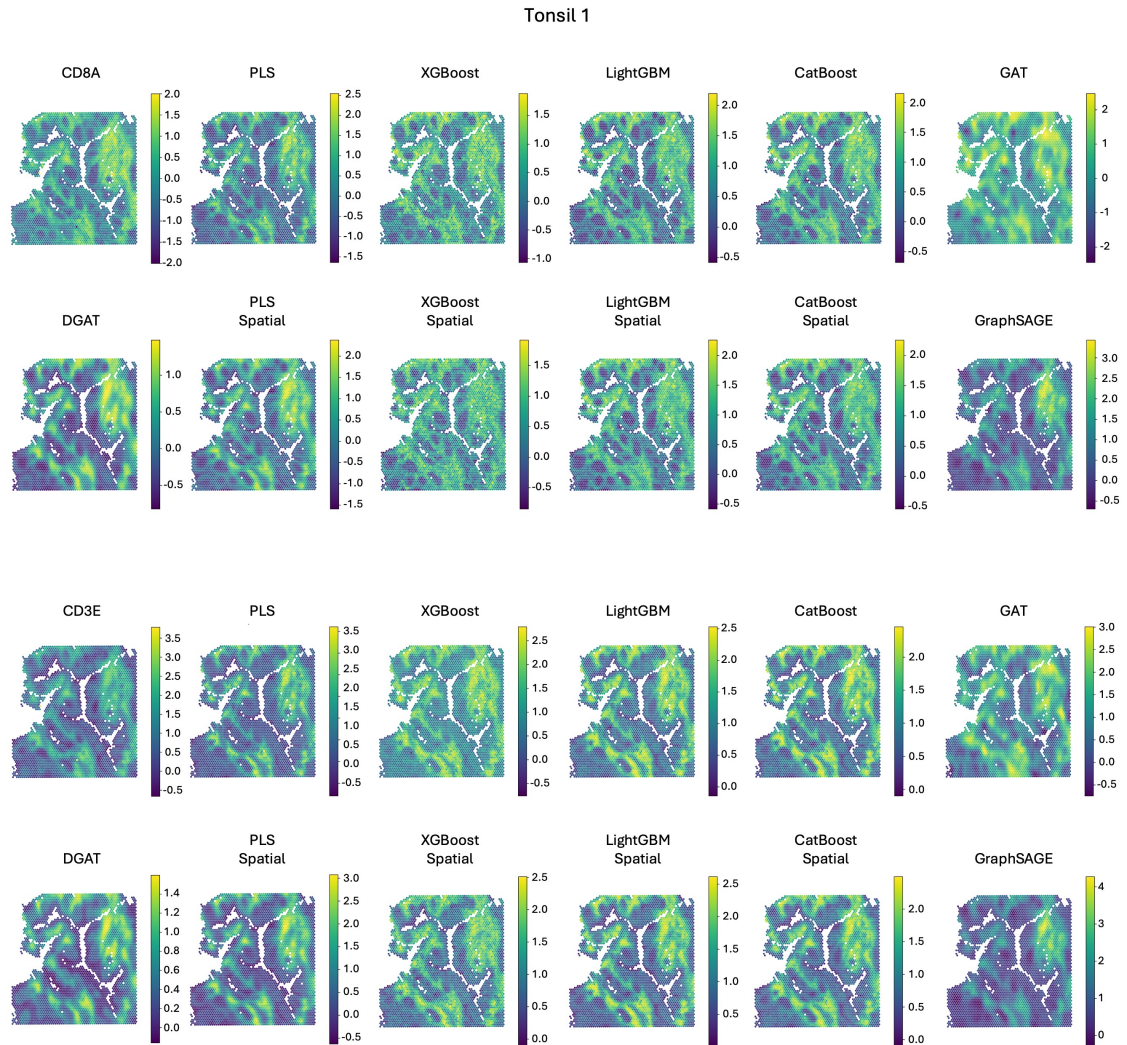


Fig. B.5.: Cross-tissue prediction of CD8A and CD3E expression in human tonsil (sample 1). CD8A, a marker of cytotoxic T lymphocytes, exhibited a sharply localized distribution within lymphoid regions, forming distinct boundaries between high- and low-expression zones. Boosting-based models with spatial feature integration most accurately captured these patterns, while non-spatial GNN predictions appeared smoother with reduced boundary sharpness. CD3E, a pan-T cell marker, showed a broader distribution across lymphoid compartments; spatially informed models provided enhanced reconstruction of intra-tissue variation, whereas non-spatial methods tended to underrepresent localized expression peaks.

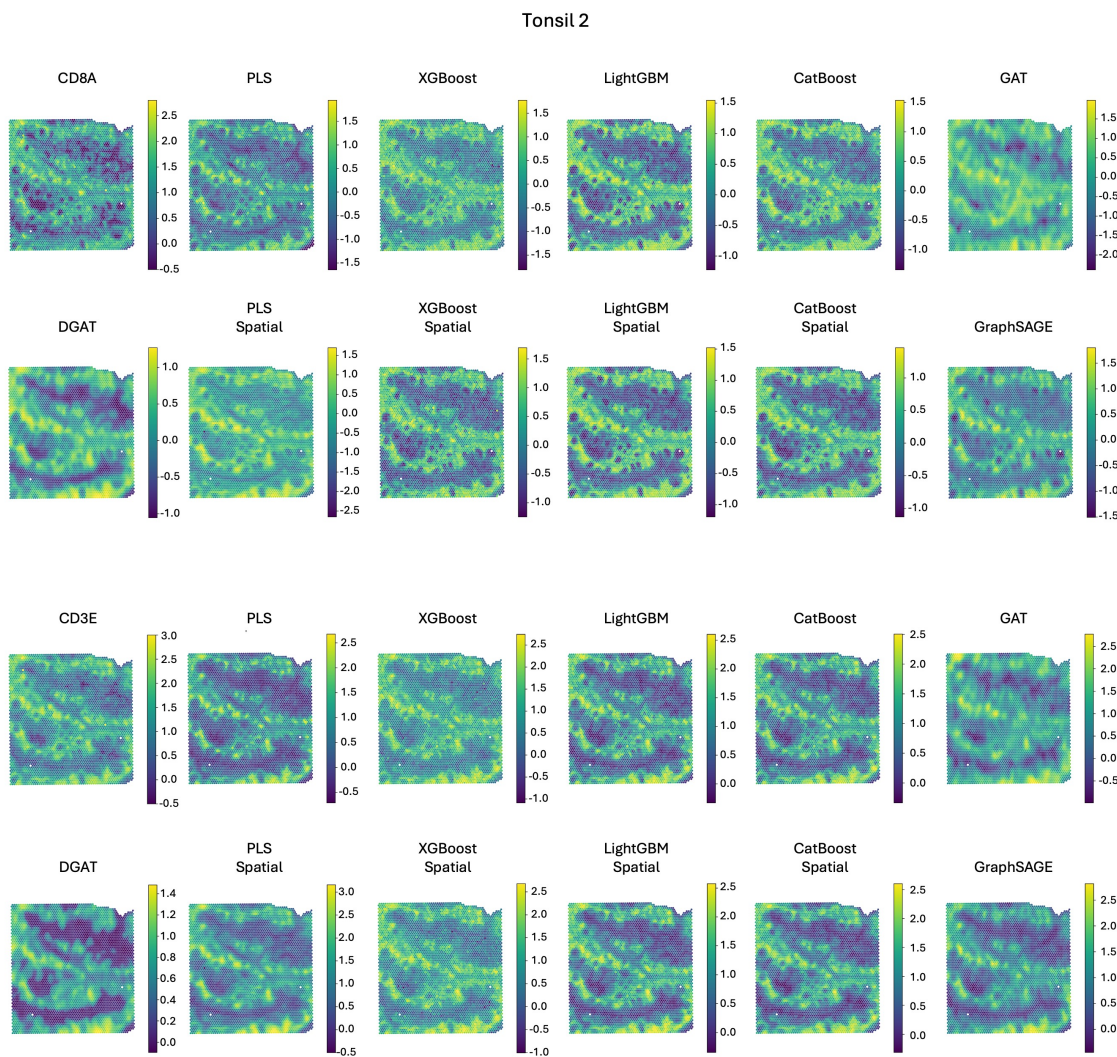


Fig. B.6.: Cross-tissue prediction of CD8A and CD3E expression in human tonsil (sample 2). CD8A displayed a punctate expression pattern concentrated in discrete immune cell clusters within lymphoid follicles. Spatially enhanced boosting models produced clearer delineation of these high-expression regions compared with their non-spatial counterparts, which tended to blur localized intensity peaks. CD3E exhibited a broader yet structured distribution across the lymphoid tissue, with spatially informed models capturing both inter- and intra-follicular variation more effectively. GNN-based models generated smoother reconstructions but occasionally underestimated sharp spatial transitions present in the ground truth.

