Analysis of Multimodal Data Integrating Genomics and Spatial Information

Ruize Jia

A thesis
submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE AND SYSTEMS

University of Washington

2025

Committee:

Ka Yee Yeung

Ling-Hong Hung

Program Authorized to Offer Degree:

Computer Science and Systems

University of Washington

Abstract

Analysis of Multimodal Data Integrating Genomics and Spatial Information

Ruize Jia

Chair of the Supervisory Committee:
Professor Ka Yee Yeung
School of Engineering and Technology

The advancement of modern computational technologies has paved the way for the processing of high-dimensional data. Benefiting from this, current research in the bioinformatics field has been utilizing sophisticated deep learning architecture to understand the complex human biological system. Spatial Transcriptomics (ST), a popular recent technology, generates multimodal data by integrating imaging data as spatial context, thereby providing extra spatial information to the high-dimensional gene expression profiles. This multimodal, high-dimensional data demands analysis from methods that jointly analyze both the sequencing and imaging data. One critical task that those methods are trying to improve is Spatial Domain identification (SDI). If SDI is well executed, the identified spatial domain can offer valuable biological insights, ultimately facilitating improved diagnosis and treatment strategies for diseases. This project will delve into spatial transcriptomics and its associated analyses to improve our capability of interpreting such data.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Background

In the evolving field of bioinformatics, analyzing high-dimensional data is crucial to unlocking complex biological mechanisms. Among the latest technologies, Spatial Transcriptomics (ST) stands out as it combines genomics data with spatial information, offering a more comprehensive view of spatial structure and cellular interactions within tissues. This thesis focuses on the analysis of the multimodal ST data and develops methods to facilitate the interpretation of ST data. Definitions of basic terminology and concepts are outlined in Section 1.1.1. An overview of ST technology is provided in Section 1.1.2, followed by Section 1.1.3, which illustrates the three stages in ST analysis. Subsequent sections (Section 1.1.4, 1.1.5, and 1.1.6) facilitate understanding of the relationship between spatial resolution and cell deconvolution, as well as the benefits of ST analysis and Spatial Domain Identification (SDI).

1.1.1 Terminology and Concepts

Transcriptomics is the study of all RNA molecules in a cell [37]. This field has witnessed revolutionary progress due to the rapid development of single-cell analysis, which has enabled deeper insights into developmental and cancer biology.

Bulk-RNA sequencing (RNA-seq) is the prevalent profiling approach, providing a view of gene expression across an entire sample. However, it lacks the resolution to capture the gene expression profile of individual cells within a tissue.

In contrast, **single cell RNA sequencing (scRNA-seq)** is a technology that profiles the transcriptome of individual cells [32]. This advancement allows quantitative analysis of the molecular activity, revealing the underlying phenotypic diversity of cells within a tissue

[5]. Computational analysis of scRNA-seq data is essential for deriving biological insights about the cell types and states, crucial for understanding tissue architecture at single-cell resolution from such high-dimensional data.

However, current methods of scRNA-seq require dissociating cells from tissues, thereby losing potentially valuable spatial information for inferring the cell types and states within tissues. Therefore, **Spatial Transcriptomic (ST)** technology has been developed to provide a spatially resolved, high-dimensional assessment of gene transcription [76].

Cell Segmentation, a downstream task in ST analysis, distinguishes the boundaries of individual cells, assigns cell labels to observed molecules, and separates intercellular backgrounds. [57]

Spatial Domain Identification (SDI), another downstream task in ST analysis, involves accurately identifying regions within tissue that exhibit similar spatial expression patterns [19].

Cell Type Deconvolution is a downstream task in ST analysis, and refers to methods that derive cell type-specific signals from heterogeneous mixture data [40].

Imputation involves estimating missing gene expression values in datasets [43].

Differential Expression, in the context of ST, refers to identifying genes that exhibit different expression patterns within specific cell types across tissues [11].

Trajectory Inference is a biological inference task in ST analysis that infers the progression of cellular states or differentiation processes within tissues [63].

Cell-Cell Interaction Analysis studies the communication and signaling processes between different cell types within tissues[45].

1.1.2 An Overview of Spatial Transcriptomics (ST)

It is a well-known fact that tissues in the heart, brain, and tumor consist of numerous cells and different cell types. These cells are not randomly positioned in space but are highly organized. Modern medical diagnostics involve procedures like biopsy, where fresh tissue samples from patients are taken, sequenced to advance understanding of the disease by examining the

genetic information of the tissue. As a traditional gene expression profiling method, bulk RNA sequencing blends all cells into a smoothie, providing a general taste of all the 'fruits'. Subsequently, single-cell RNA sequencing is developed to improve this "blending" approach by separating each "fruit" and then tasting it individually [34]. Although this improved approach provides more fine-grained results, the structural/spatial information is missing due to the isolation process [82].

To address this problem, Spatial Transcriptomics (ST) technology has been developed. It is like presenting all fruit in a palette, not only can we taste each fruit, but the spatial arrangement is also preserved. [39] It sequences tissue's gene expression profile with a spatial layout of the cells, adding a spatial dimension to the traditional single-cell experiments. This additional information reveals tissue microenvironment and heterogeneity, enables a more nuanced observation towards tissues. For clinical applications, ST enables optimal post-biopsy assessment and treatment by providing additional spatial information to determine the spread and stage of the tumor cells.[38]

1.1.3 A Data-Centric View of ST Analysis

Figure 1.1 is a recreation of Fig. 3 from "An Introduction to Spatial Transcriptomics for Biomedical Research" by Williams et al. [76] It summarizes the three stages in the analysis of Spatial Transcriptomics: upstream analysis, downstream analysis, and biological inference. Raw imaging and sequencing data are generated from an ST experiment. The first stage is upstream analysis, which converts the raw data into structured data, then performs quality check and normalization. The structured data for imaging and sequencing data are typically represented by spatial coordinates for each spatial spot and a gene expression matrix. Each spatial spot represents a small area on the tissue slide where gene expression was measured. Several commercial upstream analysis tools are available for obtaining ST data, such as Space Ranger from 10x Genomics [2], GeoMx Digital Spatial Profiler (DSP) from NanoString [33], and Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH) from Vizgen [15].

Several packages (including Scanpy [77], Squidpy [55], Seurat [10], Giotto [20], etc.) are designed to process the ST data, fetching the upstream outputs and reorganizing them to a standardized data structure for researchers familiar with Python and/or R. In the downstream analysis stage, data scientists can benefit from this standardized data structure, allowing them to focus on developing new methods to perform various tasks such as cell segmentation [57], SDI [19], and cell type deconvolution [40]. The last stage of ST analysis is biological inference, where a robust predictive model from the downstream analysis may lead to biological discoveries such as finding new pathways, cell-cell interaction, and spatially differential expressed genes.

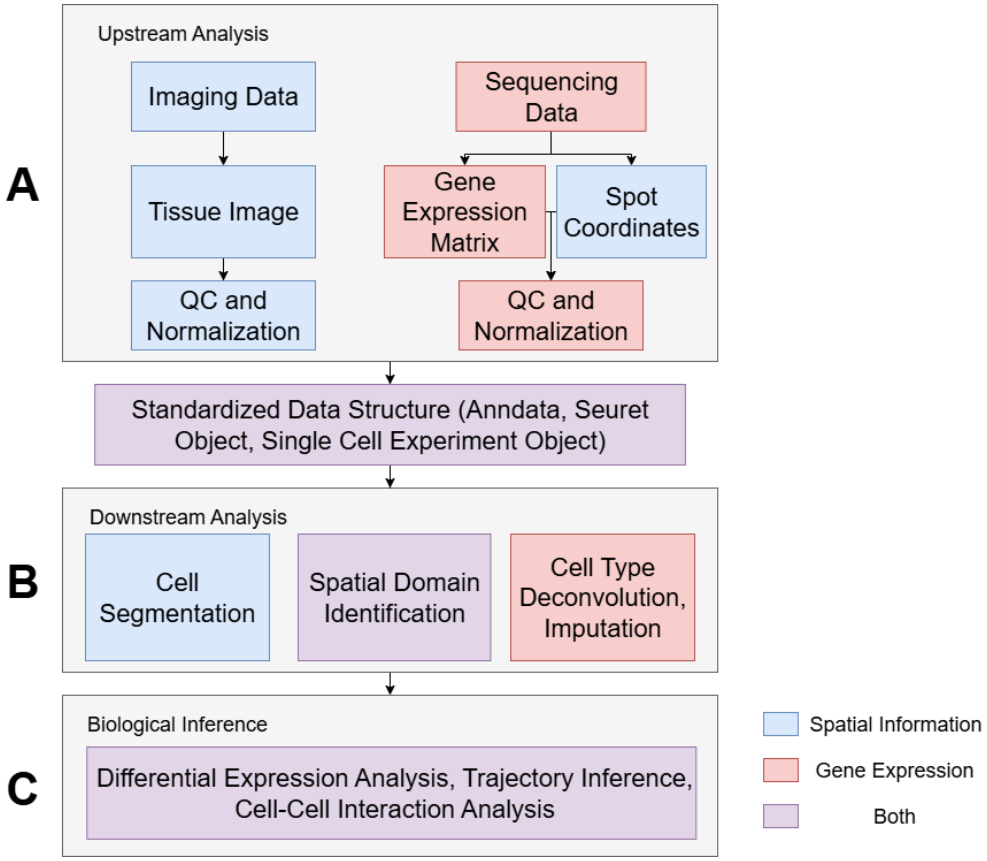


Figure 1.1: **A Data-centric View of Spatial Transcriptomics Analysis.** This figure, recreated from Fig. 3 of "An Introduction to Spatial Transcriptomics for Biomedical Research" by Williams et al. [76], presents the three stages of analyzing ST data: **A. Upstream Analysis.** This stage processes the raw data consisting of both imaging and sequencing data. A standardized data object can then be constructed using the output from the upstream analysis. **B. Downstream Analysis.** In this stage, tasks can focus on various modalities. Specifically, tasks like cell segmentation and visualization only require imaging data, whereas cell type deconvolution and imputation involve sequencing data. Spatial domain identification are the most important tasks, as they incorporate both spatial information and gene expression profiles. **C. Biological inference.** In this stage, models trained during the downstream analysis can perform inference tasks to reveal biological insights such as cell-cell interaction, differential expression, Trajectory Inference, etc.

1.1.4 Resolution of ST Datasets and Cell Type Deconvolusion

In an ST experiment (e.g., data generation using the 10X Genomics Visium Platform [28]), a tissue is placed on a microarray where each spot is barcoded before sequencing. Although this type of sequencing method can detect a high number of genes, each spot does not necessarily correspond to a single cell but is better described as cell mixtures of 1 to 10 cells [6]. This lack of single-cell resolution hampers the study of gene expression variation and spatial organization specific to a cell type [76].

Therefore, in the downstream stage, several reference-based, supervised deconvolution techniques have been developed to estimate the fraction of cell types within an ST spot or pixel to meet this challenge. Among them, SPOTlight3 seeds a non-negative matrix factorization using cell-type marker genes obtained from a single-cell RNA-sequencing (scRNA-seq) reference [21]. Robust Cell Type Decomposition (RCTD) constructs a probabilistic model representing the relative contributions of each cell type to the observed gene counts in each pixel, using a scRNA-seq reference and the cell-type-specific mean expression of marker genes [13]. Additionally, some reference-free tools like STdeconvolve have been developed, which rely on training a model on numerous ST datasets [50]. However, these techniques do not consider spatial information but solely rely on the gene expression data. Accurate assignment of the cell itself may be obtained from the cell types of nearby cells. In some cases, this information may help identify novel cell types by analyzing neighboring characteristics. Therefore, it is important to consider neighboring spots while constructing the workflow for spatial transcriptomics [9].

Apart from the 10X Visium approach, fluorescent in situ hybridization (FISH)-based ST methods like MERFISH can display subcellular resolution but lack genome-scale gene coverage. [47] A combined analysis of Visium and MERFISH data could be beneficial for performing downstream analyses that provide higher resolution and also broad gene coverage.

1.1.5 ST Analysis: Inferring Biological Insights

Analyzing ST data provides invaluable insights towards our knowledge of biological functions through biological inference, also facilitates the diagnoses and treatment for diseases. Data scientists could contribute to the downstream analysis by designing novel models for various tasks. Specifically, ST data helps researchers link gene expression of cells with their spatial distribution to better understand tissue micro-environment and biological progress. [66] For example, the laminar organization of the human cerebral cortex is especially related to its biological functions, in which cells residing within different cortical layers often differ in expressions, morphology and physiology [19].

1.1.6 The Objective: Spatial Domain Identification

Our objective is to identify biological regions within tissues, also known as spatial domain identification (SDI). SDI aims to recognize distinct regions in the tissue that vary genetically and spatially. SDI serves as a fundamental downstream task in ST. First, it reveals the tissue microenvironment and helps clinicians identify potential tumor regions, thus facilitating disease treatment. Second, most ST data do not come with labels and therefore require human annotations; however, manual annotation is slow and subjective, whereas SDI automates this process. Finally, SDI enables other downstream analytic tasks such as cell deconvolution, cell–cell communication, and trajectory inference.

1.1.7 ST Data for Our SDI Analysis: 10X Visium

10x Visium ST data are widely used for the task of spatial domain identification (SDI). As shown in Figure 1.2, a 10X Visium experiment produces two types of data that together span three modalities. For the imaging data, a slide scanner or microscope captures a hematoxylin and eosin (H&E) whole-slide image (WSI) of the tissue section. For the sequencing data, a sequencer records the spatial coordinates and gene expression profiles on the capture array. In total, this yields three modalities of data: gene expression, spot locations, and histology

images.

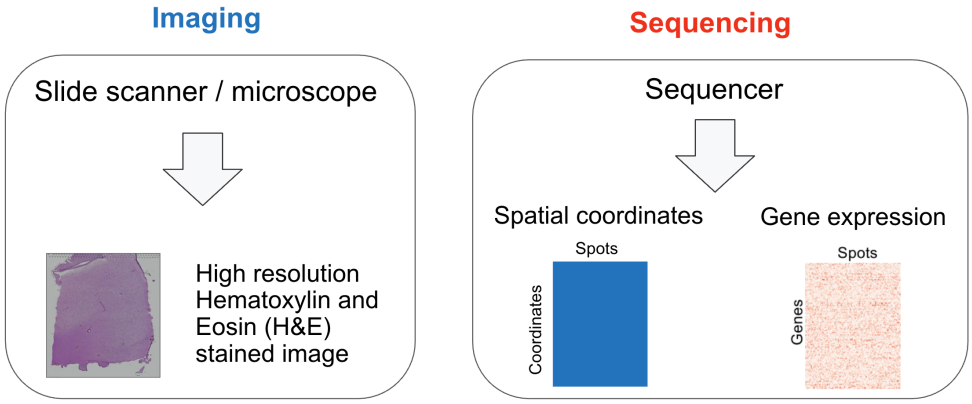


Figure 1.2: A schematic overview of data types and modalities produced by a 10x Visium spatial transcriptomics experiment. Imaging data contain high-resolution hematoxylin and eosin (H&E) whole-slide images captured by a slide scanner or microscope. Sequencing data consist of spot-level spatial coordinates and gene expression profiles from the sequencer's capture array.

1.1.8 An Example of 10X Visium ST Dataset: A Human Dorsolateral Prefrontal Cortex (DLPFC)

To illustrate 10x Visium data, we present an example from the Human Dorsolateral Prefrontal Cortex (DLPFC) dataset [60]. This publicly available dataset from the Spatial Transcriptomics human DLPFC pilot study by the Lieber Institute [60], accessible through GitHub at `https://github.com/LieberInstitute/HumanPilot`, is widely used as a reference for evaluating the performance of SDI. It contains 12 brain tissue samples with expert-annotated cortical layers serving as ground truth. Figure 1.3 illustrates Sample 151673. The imaging data consist of the H&E histology image, while the sequencing data include (left) a spot coordinate matrix with 3,639 spots, and (right) a gene expression matrix of the same spots across 36,601 genes. The expression matrix is both high-dimensional and sparse, with

most entries being zero. To make it suitable for downstream analysis, variable-gene selection and dimensionality reduction are typically applied to mitigate the curse of dimensionality.

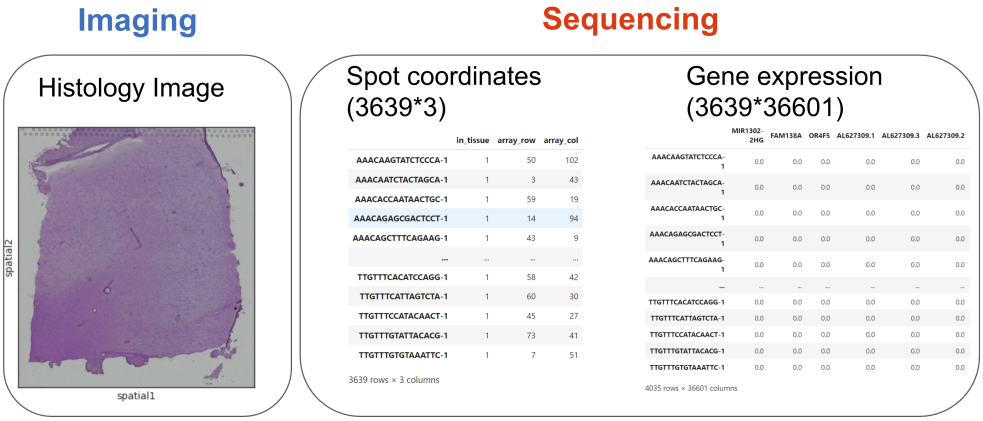


Figure 1.3: An example (sample 151673) of 10x Visium spatial transcriptomics data from the Human Dorsolateral Prefrontal Cortex (DLPFC) dataset. The imaging data consist of the H&E histology image, whereas the sequencing data include a spot coordinate matrix (3,639 spots * 2 coordinates + 1 boolean value) and a gene expression matrix (3,639 spots × 36,601 genes).

1.1.9 SDI Methods: A Preview

In recent years, both statistical methods (BayesSpace [83], Giotto [20], stLearn [58], etc.) and machine learning methods (SpaGCN [35], STAGATE [19], DeepST [81], GraphST [47], etc.) are proposed to tackle this task.

In Chapter 2, we discuss SDI methods in the literature, focusing on their approaches to the abstraction and augmentation of gene expression profiles, the extraction of spatial information, and the tokenization of histology images. In Chapter 3, we present our experimental design for an improved method that achieves better SDI performance. In Chapter 4, we evaluate this improved method on a reference dataset to demonstrate its robustness. Finally, in Chapter 5, we summarize our findings and discuss the conclusions.

Chapter 2

RELATED WORK

As introduced in Chapter 1, two types of data are collected in spatial transcriptomics: imaging data and sequencing data. Together, these yield three modalities: gene expression, spot locations, and histology images. In this chapter, we identify key aspects of the SDI workflow that influence performance, with a particular focus on cell/spot representation learning and clustering. The following sections discuss these aspects in detail and highlight representative methods from the literature.

In Section 2.1, we present clustering methods and assess their performance for SDI. Sections 2.2, 2.3, and 2.4 separately investigate the contributions of three modalities (gene expression, spatial coordinates, and histology images) to enhancing SDI outcomes. Section 2.5 illustrates the assessment metrics for SDI. Finally, Section 2.6 introduces STAIG, a state-of-the-art method that we extend to achieve improved SDI performance.

2.1 Clustering

Clustering is a crucial component of SDI, as most SDI workflows rely on clustering spots to partition the tissue into distinct regions. Unlike single-cell RNA sequencing, where clustering is typically performed directly on the gene expression matrix, applying the same approach in ST without spatial or histological information often fails to produce accurate spatial clusters. In ST, clustering usually incorporates two or more modalities, such as gene expression, spot coordinates, and histology images.

There are multiple strategies for preparing ST data for clustering. For example, as illustrated in Figure 2.1, one common workflow constructs a spot representation matrix that integrates all three modalities. Clustering is then applied to this representation, producing

assignments of spatial spots to domains. These domains can be visualized on the histology image and subsequently interpreted for downstream analyses.

In summary, clustering multi-modal ST data is not trivial, and the following subsections describe common clustering methods used in SDI.

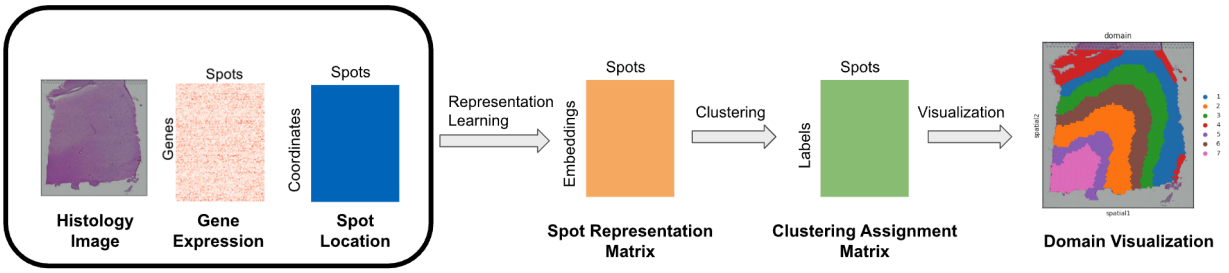


Figure 2.1: An example of SDI workflow. A spot representation matrix is learned from three modalities: gene expression, spot locations, and histology images. Clustering is then performed on this spot representation to generate a cluster assignment matrix, which can be visualized as spatial domains. The clustering visualization shown here corresponds to Sample 151673 from the Human Dorsolateral Prefrontal Cortex (DLPFC) dataset [60].

2.1.1 Non-spatial Clustering

Non-spatial clustering refers to traditional clustering methods designed for single-cell datasets that do not consider the spatial and histological context of the data. There are many clustering methods in the literature. For example, k-means is a partitional clustering method that minimizes within-cluster variance by iteratively reassigning points to the nearest centroid [46]. As another example, the Louvain algorithm groups nodes to maximize modularity, a measure of how well a network is divided into dense clusters with sparse interconnections. It uses a two-phase iterative process to refine its cluster assignments: node movement and network aggregation [7]. Leiden [71] is a popular community detection method used as the default clustering algorithm in bioinformatics tools such as Scanpy [77] and Seurat [10]. It improves the Louvain algorithm by guaranteeing well-connected communities, refining par-

titions to avoid suboptimal merges, and achieving faster convergence. Mclust [62] is a tool designed for model-based clustering that assumes the data are distributed as multivariate Gaussian. It uses the Bayesian Information Criterion (BIC) that represents the log likelihood and a penalty for model parameters to select the best model and number of clusters. These non-spatial methods directly cluster on the gene expression matrix, and they generally perform poorly when applied directly to spatial transcriptomics data because they do not leverage spatial and histological information[80].

2.1.2 Iterative Spatial Clustering

Iterative spatial clustering typically involves two main steps: cluster initialization and iterative refinement. It often begins with a non-spatial clustering step, where initial assignments are generated based on gene expression features alone. Spatial information is then incorporated into the refinement process to improve the clusters, encouraging spatial or topological smoothness. The iteration usually continues until a predefined convergence criterion is met, such as minimal changes in assignments or reaching a maximum number of iterations. Both statistical and deep learning tools are commonly used in this setting, including Expectation–Maximization (EM), Markov chain Monte Carlo (MCMC), and Deep Embedded Clustering (DEC).

Specifically, spatialLeiden [53] is a variant of the Leiden algorithm that incorporates spatial proximity into the quality function, ensuring clusters are both well-connected in the network and spatially contiguous. It initializes cluster assignments on a dimension-reduced gene expression matrix and then iterates using a spatial K-nearest neighbors (KNN) graph constructed on the spatial embeddings generated by Leiden multiplex to ensure spatial smoothness.

BayesSpace [83] is another statistical method that applies iterative spatial clustering. Similar to SpatialLeiden, it initializes clusters from the gene expression matrix without spatial information. The key difference is that BayesSpace applies super-resolution to ST data and uses Mclust [62] for clustering. In its iterative refinement step, BayesSpace defines a

spatial prior with a Potts model and updates cluster assignments using MCMC until convergence.

From the deep learning perspective, Attention-guided variational graph autoencoders (AVGN) [41] initialize clusters on a spatially enhanced spot-level representation using the Louvain algorithm, then iteratively refine them by feeding the Leiden [71] clustering assignments back into the model. This process continues until the model's cluster assignments stabilize.

2.1.3 Spatial Clustering after Spot Representation Learning

Spot representation learning provides another strategy for spatial clustering. By integrating multiple modalities, it produces representations that capture genetic, spatial, and histological information. This approach is also the most widely adopted in SDI.

To construct a spot representation matrix, methods typically employ Graph Neural Networks (GNNs). Methods such as stLearn [58], SpaGCN [35], DeepST [81], GraphST [47], Spatial transcriptomics analysis via image-aided graph (STAIG) [79], and spatially embedded deep representation (SEDR) [78]all use an adjacency matrix to represent spatial neighborhoods among spots. The adjacency matrix is generally constructed using the k-nearest neighbors (KNN) algorithm. Gene expression data are assigned as node embeddings for each spot, and these embeddings are then aggregated through the network during GNN training.

The output of these methods is a spot-by-embedding matrix, where each row corresponds to a spot and each column to an embedding dimension. This learned representation is then used for clustering. In terms of clustering strategy, these methods are broadly similar, as they all perform clustering on the learned spot representation matrix. Their main differences lie in how they integrate the three modalities of ST, which are discussed in the following subsections.

2.2 Abstraction and Augmentation of Gene Expression Profiles

The raw gene expression profile in ST is represented as a high-dimensional, sparse cell-by-gene matrix. For example, a typical dataset may contain around 3,500 cells and up to 30,000 genes [60]. To reduce both sparsity and dimensionality, researchers often select highly variable genes (HVGs) by measuring gene variability, typically reducing the gene count to around 3,000 [47, 78, 79, 41]. After HVG filtering, dimension reduction methods are applied to avoid the curse of dimensionality, lower computational costs, and improve clustering quality.

Moreover, in ST, the gene expression data often require further enhancement with spatial and histological context to achieve better SDI performance. In the following subsections, we describe methods that reduce the dimensionality of the gene expression matrix as well as approaches that augment gene expression profiles with features from other modalities.

2.2.1 Principal Component Analysis (PCA)

One of the simplest approaches to reduce the dimensionality of the gene expression matrix is Principal Component Analysis (PCA), an unsupervised method. PCA is computationally efficient, but when applied to biological data it may capture batch effects from experiments, since it prioritizes statistical variability rather than biologically meaningful features.

Despite this limitation, PCA is widely used as a preprocessing step for gene expression profiles. For example, spatialLeiden [53], BayesSpace [83], AVGN [41], and SpaGCN [35] all implement PCA to reduce the dimensionality of the gene expression matrix.

2.2.2 Methods Designed for Single-cell Sequencing

Many methods originally developed for single-cell sequencing could be applied directly to ST. However, similar to non-spatial clustering, these methods only focus on the gene expression matrix, ignoring the spatial and histological modalities inherent in ST.

For instance, scVI [48], a widely used variational inference framework for scRNA-seq,

projects the gene expression matrix to a latent space to generate representations for each cell while correcting for batch effects. Although effective in single-cell contexts, scVI does not account for spatial proximity or histological structure in ST data. As a result, applying it directly to ST can lead to suboptimal identification of spatial domains.

2.2.3 Deep Neural Networks (DNN)

Deep neural networks (DNNs) are widely used in ST because they can both reduce the dimensionality of high-dimensional gene expression data and augment it with spatial and histological context. Their ability to learn complex, nonlinear representations makes them powerful for abstracting sparse gene expression profiles and integrating multiple modalities into a unified spot-level representation for clustering.

A simple multi-layer perceptron (MLP) can learn such a representation by projecting high-dimensional data into lower dimensions, assigning different weights to each dimension for aggregation. Variational autoencoders (VAEs) offer a more robust approach by adding a reconstruction loss together with a Kullback–Leibler (KL) divergence regularization term for the representation.

Several methods use autoencoders to reduce the dimension of the gene expression matrix. For instance, SEDR [78] and DeepST [81] both adopt a Variational Graph Autoencoder (VGAE) to learn spot representations from spatial graphs constructed via KNN. Additionally, AVGN [41] also uses a VGAE; however, the input is a PCA-reduced gene expression matrix.

2.2.4 Foundation Models

Foundation models are neural networks trained in an unsupervised manner on massive datasets to perform a wide variety of tasks [54]. These models can accept and generate data from multiple modalities. A familiar example is ChatGPT, which is itself a foundation model. Such models have excellent zero-shot capabilities, meaning they can perform

tasks without task-specific training. They can also be fine-tuned later to specialize in certain applications.

Foundation models for genetic data, such as scGPT [18] and scFoundation [31], are powerful tools for reducing the dimensionality of the gene expression matrix. These models are trained on large-scale gene expression datasets to infer biologically meaningful, batch-effect–corrected, low-dimensional representations. However, current SDI methods that leverage foundation models with gene expression data are still immature and not robust enough to achieve strong SDI performance. For example, scGPT-spatial [75], which fine-tunes scGPT with spatial decoders, often suffers from hallucination, meaning the model is prone to generating false information, thereby degrading the quality of its spot representations.

2.2.5 Methods with Spatial Correction/Augmentation

There are also methods that enhance the gene expression matrix with spatial information using spot adjacency. GraphST [47] and STAIG [79] both construct an adjacency matrix and use graph convolution networks to aggregate gene expression profiles over the graph neighborhood, then use contrastive learning on graph convolutional networks to refine the gene expression with information from spatial and histological modalities. Similarly, STAGATE [19] uses a graph attention network to incorporate neighborhood information from the spatial KNN graph, thereby refining the gene expression matrix using the attention score learned from the spatial neighborhood.

Banksy [67], on the other hand, applies an azimuthal Gabor filter to extract the mean and global gradient of gene expression, and then concatenates the original and augmented gene expression data for spatial enhancement.

2.3 Extraction of Spatial Information

Spatial information can be derived from spot coordinates. This information is important for identifying spatial neighborhoods and characterizing the tissue microenvironment. Proper modeling of spot coordinates can contribute to more accurate SDI. The following subsec-

tions discuss statistical methods (e.g., the Potts model, Azimuthal Gabor filter) and machine learning approaches (e.g., multi-layer perceptrons, adjacency matrices) that illustrate different strategies for extracting and incorporating spatial information.

2.3.1 Potts Model

The Potts model is a statistical physics model that generalizes the Ising model to more than two possible states (interacting spins on a crystalline lattice) [59]. This model provides a statistical framework for defining spatial neighborhoods.

An application of the Potts model in SDI is BayesSpace [83], which incorporates the Potts model to define a spatial prior within its Bayesian framework. This spatial prior encourages spatially proximate spots to share the same cluster label. The clustering assignments are then optimized spatially using Markov chain Monte Carlo (MCMC) simulation. As a result, BayesSpace produces spatial domains with improved spatial smoothness that better align with tissue structure.

2.3.2 Azimuthal Gabor Kernel

In image and signal processing, a filter (kernel) is a mathematical function that is convolved with an input matrix (usually an image) to process or extract spatial features, such as blurring, sharpening, and edge detection [8].

The Gabor filter is a widely used filter, which excels at capturing spatial frequency and orientation information, thus particularly effective for detecting textures or repeated patterns in images [26]. The azimuthal Gabor filter extends this concept by introducing directional sensitivity, extracting gradient-like features radiating from a central point, thereby modeling spatial variations across multiple directions [72].

Banksy [67] incorporates spatial information into gene expression data using an azimuthal Gabor filter, which generates a spatial gradient matrix from spatial coordinates. This captures spatial patterns in gene expression profile and provides higher-level spatial context for each spot.

2.3.3 Deep Neural Networks (DNN)

As discussed in Subsection 2.2.3, DNNs are effective for both abstracting and augmenting gene expression data. They also excel at capturing spatial information by learning from the spatial neighborhood. In particular, graph neural networks (GNNs) are a common approach for modeling spatial relationships between spots. GNNs use an adjacency matrix to define the spatial neighborhood, which we describe in more detail in Subsection 2.3.4.

Examples of SDI methods that leverage multi-layer perceptrons (MLPs) to extract spatial information include stLearn [58], which uses an MLP to learn a morphological feature matrix from image tiles. These features are then used to calculate pairwise distance matrices, which define edge weights in the distance graph for the GNN. Another example is GASTON [17], which employs an MLP-based transcoder to generate spot representations that map spatial coordinates onto gene expression. These representations are then used to construct a directional graph structure for identifying topographic boundaries.

2.3.4 Adjacency Matrix in Graph Neural Networks (GNNs)

Graph neural networks (GNNs) are a type of deep neural network (DNN) designed to operate on graph-structured data. They can perform tasks such as node classification, edge prediction, graph classification, and regression. In a graph, vertices (nodes) and edges are represented as node embeddings and an adjacency matrix, which together serve as input to the GNN. A GNN is equipped with learnable weights that iteratively adjust the graph structure by either updating node embeddings based on edge weights or updating edge weights based on node embeddings. Through this iterative process, the model outputs new node or graph representations that can be used for various tasks.

The adjacency matrix in a GNN is a node-by-node matrix that encodes connectivity between nodes. If an edge exists between two nodes, the corresponding entry contains a value representing the edge weight. In this way, the adjacency matrix models the spatial neighborhood.

GNN-based methods such as stLearn [58], SpaGCN [35], GraphST [47], STAGATE [19], SEDR [78], AVGN [41], and STAIG [79] construct an adjacency matrix by running a k-nearest neighbors (KNN) algorithm on the Euclidean distance matrix of the spot coordinates. In this graph, spots are represented as vertices, spatial neighborhoods as edges, and gene expression profiles or histology images as node embeddings. However, the way edge weights are defined and how node embeddings are aggregated differ across methods.

Specifically, SpaGCN [35] leverages a Graph Convolutional Network (GCN) to integrate gene expression, spatial coordinates, and histology images, learning shared weights to define local neighborhoods for all spots. GraphST [47] improves upon SpaGCN by incorporating contrastive learning between spots, encouraging spatially close neighbors to generate similar node representations. STAGATE [19] implements a Graph Attention Network (GAT), which assigns learnable attention weights to each node instead of using uniform weights, thereby better capturing spatial relationships at the spot level. SEDR [78] employs a Variational Graph Autoencoder (VGAE) to generate spot-level embeddings and subsequently applies Deep Embedded Clustering (DEC) with k-means initialization to refine cluster assignments. AVGN [41] incorporates histology images into the node embeddings before passing them through a GCN. Finally, STAIG [79] differs in that it does not use histology images as node information; instead, it computes cosine similarity between BYOL [29] embeddings of image patches to estimate probabilities for random edge dropping, then applies contrastive learning on the updated graph to refine spot representations.

2.4 Tokenization of Histology Image

Histology images (Hematoxylin and Eosin (H&E) stained images in 10X Visium datasets) serve as an important modality in ST data, providing a morphological view of tissue structure. The final clustering assignments are often visualized on the histology image. However, these images may contain visual artifacts, such as air bubbles, which can mislead downstream modeling. To address this, many methods apply image preprocessing techniques, such as blurring or single-channel filtering, to reduce noise. Another challenge arises from

the irregular alignment of spatial spots, which makes it difficult to extract image patches that align consistently for direct use in computer vision models such as Convolutional Neural Networks (CNNs) and Transformers. Common strategies for handling histology images include image tiling and patching.

After tiling or patching, tokenizers are used to convert images into embeddings. This reduces the dimensionality of the image data into lower-dimensional latent representations. These embeddings are later incorporated into SDI methods to support the modeling of spatial neighborhoods. Approaches to tokenization vary. Early methods relied on handcrafted features such as gradient filters. These methods are simple and less powerful at capturing complex spatial textures. However, they are computationally efficient, interpretable, and do not require large training datasets. Modern approaches use deep learning–based feature extractors such as convolutional and transformer-based networks. However, these models may struggle and overfit when trained on small datasets. To address this, self-supervised learning and foundation models have emerged as powerful tokenizers. They produce embeddings that generalize well even on small datasets without severe overfitting.

In the following subsections, we discuss how SDI methods leverage tiling or patching for spot representation learning, also list some common appraoches for generating embeddings. It is also worth noting that high-resolution histology data are currently available only from the 10x Visium platform [28] as part of the experiment output. Some SDI methods are designed for multiple ST platforms and therefore lack the ability to leverage this modality.

2.4.1 Image Tiling

Image tiling is a common technique in image processing that divides a large image into a fixed n * n grid of smaller tiles. This approach allows computer vision models (especially transformers) to treat images as sequences, and it also reduces the computational cost compared to processing the entire image at once. Each tile can be mapped back to spatial spots, capturing the microenvironment around them. However, this method has limitations. First, multiple spots may share the same tile, which reduces the variability captured from

the histology image modality. Second, the number of tiles is limited by the computational complexity of transformers, which typically scales quadratically ($O(n^2)$) with the number of tokens (tiles).

An example of an SDI method that implements tiling is AVGN [41], which divides a histology image into a 9 * 9 grid of tiles. Each tile is tokenized into an image embedding using ImageNet-pretrained models. These embeddings are then mapped to spatial spots and concatenated with the gene expression data to form node embeddings, where each spot's histology information corresponds to a specific location within the 9 * 9 grid. As a result, the image tiles are directly used in training the spot representation matrix.

2.4.2 Image Patching

In computer vision, patching is another way to extract localized information from an image. Patches are essentially cropped regions centered on specific points of interest. Unlike tiling, which uniformly divides the entire image into a fixed grid, patching is more flexible. The image patches can overlap, and the number of patches depends on the regions of interest present in the image. In the context of SDI, patching usually refers to cropping image regions based on the spatial spot coordinates, with each patch having a predefined size. Each patch captures the local microenvironment around a spot within a specified area.

STAIG [79] is an SDI method that utilizes patching. It captures local-level features by cropping image patches centered on each spot coordinate, with a size of 2.5 times the spot diameter. These patches are then used to determine the spatial relationships between neighboring spots. As discussed in Subsection 2.3.4, STAIG computes embeddings of the image patches using Bootstrap Your Own Latent (BYOL) [29] and compares the patch embeddings between spots to decide whether an edge should be established. We examine STAIG in detail in Subsection 2.6.

2.4.3 Self-supervised Models

Self-supervised learning is a powerful approach for tokenizing images, as it learns representations directly from unlabeled data. This makes it especially suitable for spatial transcriptomics, where datasets are often small and lacking human-provided annotations. Many early self-supervised methods are contrastive, refining embeddings by comparing positive and negative pairs. More recent approaches, such as Bootstrap Your Own Latent (BYOL) [29] and masked autoencoders, show that meaningful representations can also be learned without explicit negatives. Specfically, BYOL learns image embeddings by comparing two augmented views of the same image. The key idea is to train an online network to predict the representation produced by a target network. This process is updated through exponential moving averages rather than gradient descent.

2.4.4 Foundation Models

Similar to the discussion in Section 2.2, vision-based foundation models are powerful tools for abstracting and augmenting information without the need for labeled data or large training datasets. For histology images, UNI-2 [16] is a foundation model trained on a huge collection of pathology images. It is capable of performing various tasks such as image segmentation, classification, biomarker prediction, and disease diagnosis. For SDI, UNI-2 provides strong denoising capabilities thanks to its pretraining on pathology images, allowing raw images to be directly converted into embeddings without preprocessing (e.g., blurring or single-channel filtering). Overall, UNI-2 can generate biologically meaningful representations from histology images.

2.5 Assessment Metrics

When evaluating SDI performance, we compare the predicted regions with expert-annotated the ground truth labels. It is often easy to distinguish between good and poor identifications, but differentiating between visually similar results can be challenging. Therefore, a

quantitative measure of SDI performance is necessary. In this section, we introduce two statistical measures: the Adjusted Rand Index (ARI) [36] and Normalized Mutual Information (NMI) [69]. As shown in Figure 2.2, we illustrate the difference between a good and a poor clustering assignment, along with their corresponding quantitative measurements.

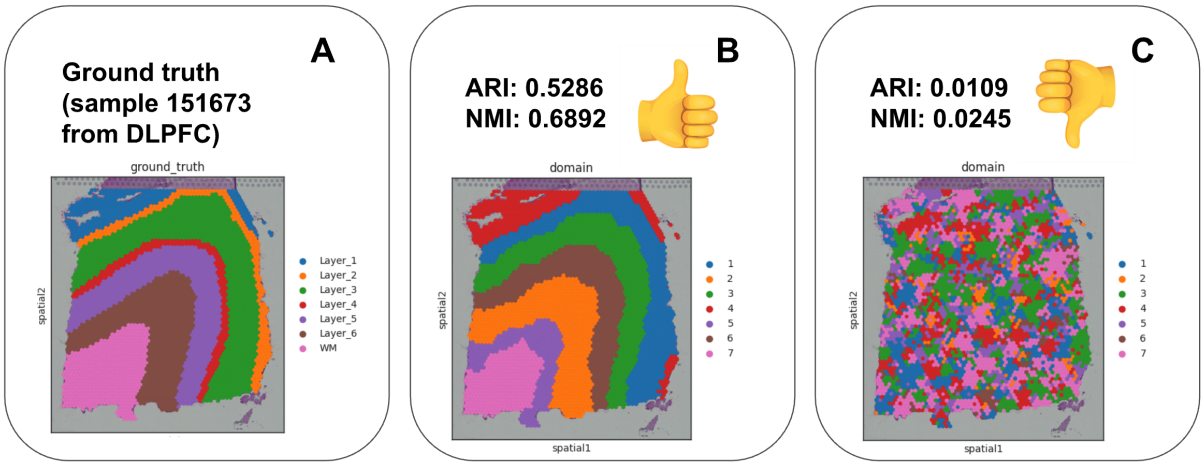


Figure 2.2: Ground truth and two clustering assignment visualizations for the same sample, 151673 from the DLPFC reference dataset. In panel A, we have the ground truth labels visualized on the histology image. In panel B, you see a relatively good clustering result. The domains align well with the known cortical layers, and the boundaries appear coherent and consistent with the ground truth. Additionally, the ARI and NMI are relatively high. In panel C, you see an extreme example of a bad clustering result. For this case, the spot representation matrix is randomly shuffled, and so is the clustering assignment. As a result, the clusters appear fragmented and do not follow the underlying tissue structure, so the identified regions fail to correspond to the actual cortical layers. The ARI and NMI values are very low in this case. This explains why we use metrics like ARI and NMI: they provide a quantitative way to distinguish between meaningful clustering results and poor ones.

2.5.1 Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) [36] is a measure of similarity between two clustering results, adjusted for random chance. A clustering result or ground truth can be represented as a partition in which each object is assigned to a single group or cluster. ARI can be used to quantify the agreement between two partitions, such as evaluation of how well a clustering result aligns with another clustering result or with the ground truth. Suppose X and Y denote the two partitions being compared. In Formula 2.1 [36],, N is the total number of data points, N_{ij} is the number of points shared between cluster i of X and cluster j of Y, N_i is the size of cluster i in X, and N_j is the size of cluster j in Y. The numerator measures agreements between the two partitions beyond what would be expected by chance, while the denominator normalizes this value. The ARI ranges from –1 to 1, where 1 indicates perfect agreement, 0 indicates performance no better than random, and negative values indicate performance worse than random.

In the context of SDI, the ARI is computed by comparing all pairs of spots between the predicted clustering and the ground truth labels, checking whether each pair is assigned to the same cluster in both cases. A high ARI value indicates that the estimated domains closely match the biological regions from the ground truth.

$$\mathrm{ARI}(X,Y) = \frac{\sum_{i,j}\binom{N_{ij}}{2} - \left[\sum_i \binom{N_i}{2}\sum_j \binom{N_j}{2}\Big/\binom{N}{2}\right]}{\frac{1}{2}\left[\sum_i \binom{N_i}{2} + \sum_j \binom{N_j}{2}\right] - \left[\sum_i \binom{N_i}{2}\sum_j \binom{N_j}{2}\Big/\binom{N}{2}\right]} \tag{2.1}$$

2.5.2 Normalized Mutual Information (NMI)

The Normalized Mutual Information (NMI) is another clustering evaluation metric derived from information theory. It measures the mutual information between the predicted clustering and the ground truth. As shown in Formulas 2.2 [69], 2.3 [64], and 2.4 [65], the NMI is defined using the mutual information (MI) [64] between two partitions and their entropies. MI quantifies the amount of information shared between the predicted and true clusters, while entropy [65] measures the uncertainty within each clustering. Here, similar to the

mathematical definition of ARI, X and Y denote the two clusterings being compared, while $x \in X$ and $y \in Y$ represent individual clusters within each partition.

NMI normalizes MI with respect to the entropies of both partitions, ensuring that values range between 0 and 1. An NMI of 1 indicates perfect correlation between the two clustering results, while an NMI close to 0 suggests no meaningful relationship.

In the context of SDI, similar to ARI, NMI provides another quantitative way to evaluate cluster quality, with higher values indicating better performance.

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{2.2}$$

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{2.3}$$

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{2.4}$$

2.6 An Improved SDI Method based on STAIG

We reviewed several state-of-the-art (SOTA) methods and found that Spatial transcriptomics analysis via image-aided graph contrastive learning (STAIG) [79] outperforms the others, both according to the literature and based on our own reproduced results. As shown in Figure 2.5, STAIG delivers superior SDI performance on slide 151673 from the DLPFC dataset [60]. This strength is further validated across all twelve DLPFC samples, where STAIG achieves higher ARI and NMI values compared to other methods, as shown in Figure 2.4. In addition to its strong performance, STAIG addresses all data modalities discussed in the previous sections, including gene expression profiles, spatial information, and histology images (Figure 2.3).

Specifically, STAIG [79] uses a graph convolutional network (GCN) to integrate all three modalities of ST. For spot coordinates, STAIG transforms them into an adjacency matrix and constructs a k-nearest neighbors (KNN) graph, where nodes (spots) are connected if

they are spatially close. The gene expression matrix is directly used as node embeddings for all spots. For histology images, STAIG adopts image patching by cropping square regions centered at each spot coordinate. Each patch is then tokenized through the self-supervised Bootstrap Your Own Latent (BYOL) framework [29], producing patch embeddings. These embeddings are used to refine spatial relationships by randomly dropping neighboring edges based on pairwise cosine similarity.

All three modalities contribute to the training of the GCN: node embeddings (gene expression profiles) are aggregated according to edge connections defined by spot coordinates and refined by histology images. Contrastive learning is applied to encourage connected nodes (spots) to generate similar representations, while unconnected nodes produce dissimilar ones. The GCN outputs a spot representation matrix, on which clustering is performed to generate spatial domain labels. These domains are then visualized on the histology image. The complete STAIG workflow is illustrated in Figure 2.3.

Building on its strengths, we choose STAIG as our baseline model because it achieves the best performance among current SDI methods and incorporates gene expression, spatial information, and histology images into its workflow. Building on this foundation allows us to fine-tune these three aspects accordingly. In Chapter 3, we present the contributions made to STAIG that lead to improved SDI performance.

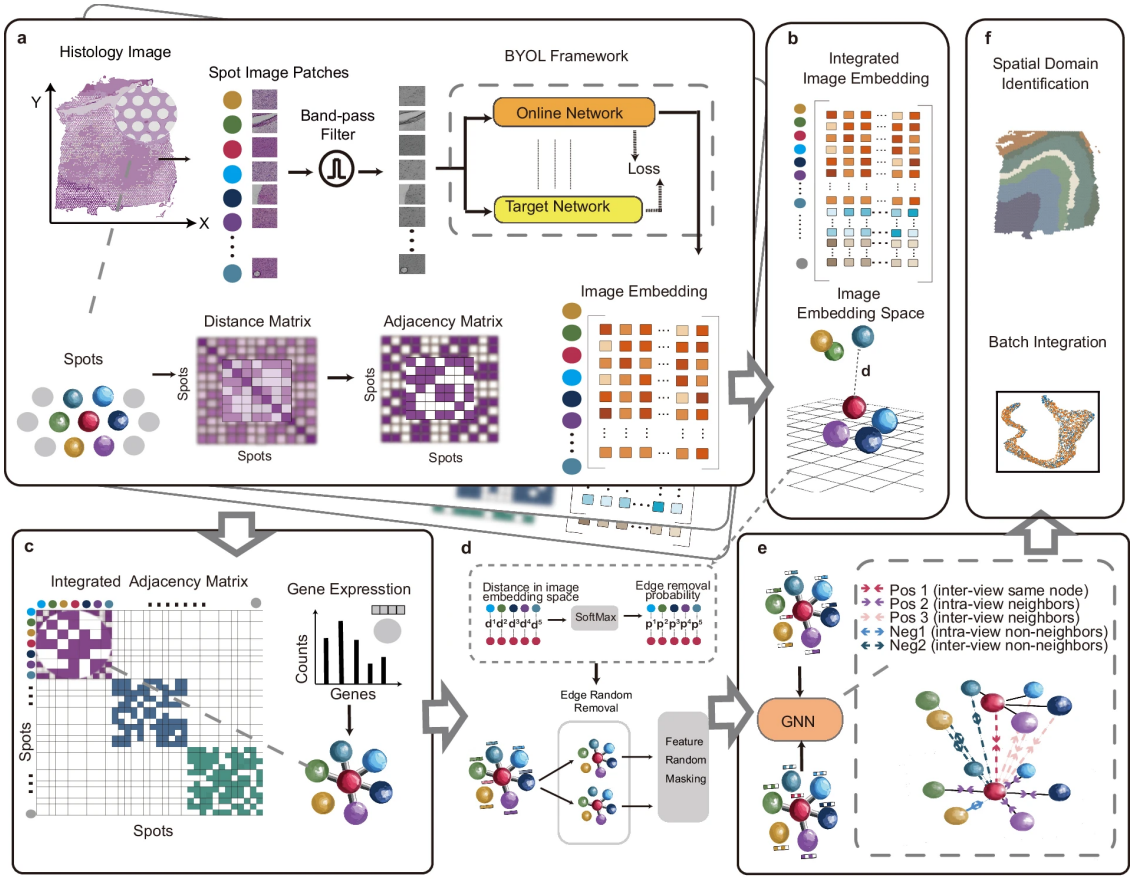


Figure 2.3: Workflow of STAIG created by Yang et al., from Fig. 1 of the original paper [79]. The description is adapted from Yang et al. [79]. a. STAIG starts with spatial coordinates, gene expression profiles, and histology images. The coordinates are used to construct the adjacency matrix, the gene expression profiles are used as node embeddings, and image patches cropped around the spot coordinates are used to generate latent image embeddings via BYOL after band-pass filtering. b. STAIG also supports multiple slides by vertically merging them. c. The adjacency matrix is simplified using KNN selection, preserving only neighboring spot relationships to build edges. d. For each edge, the cosine similarity between the image embeddings of its connected vertices is calculated. This similarity is then converted into a probability of random removal using a SoftMax function. e. A GCN aggregates node information through a contrastive learning process, generating similar node representations for neighboring spots. f. The node representations generated by the GCN are then used for clustering in SDI.

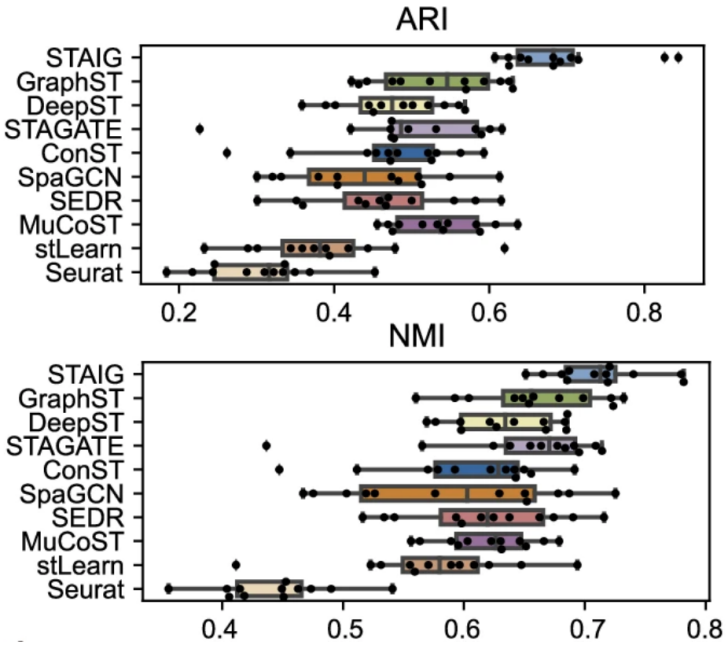


Figure 2.4: A comparison of SOTA methods's SDI performance across all twelve samples from the reference dataset DLPFC, measured in ARI and NMI. This is Fig. 2a from the STAIG paper by Yang et al. [79].

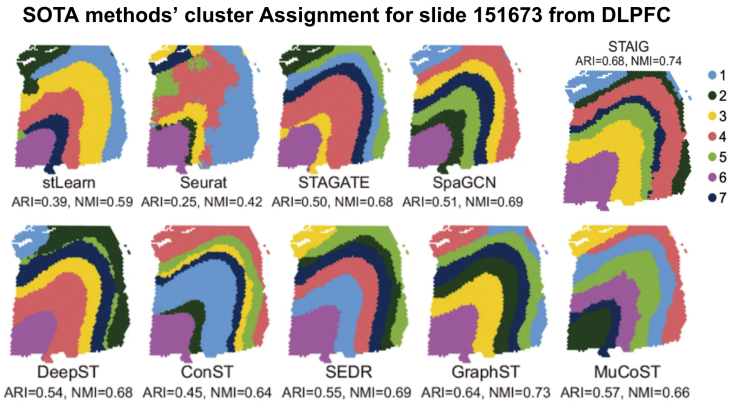


Figure 2.5: A visualization of SOTA methods's clustering assignment for slide 151673 from DLPFC dataset, with both ARI and NMI listed. The STAIG achieves the highest ARI and NMI among the methods shown. This is Fig. 2c from the STAIG paper by Yang et al. [79].

Chapter 3

EXPERIMENTAL DESIGN

In this chapter, we provide details of our contributions to the STAIG method [79]. We describe the dataset used in Section 3.1, the strategies adopted for improving the SDI performance of STAIG in Sections 3.2, 3.3, and 3.6, and potential future directions for further enhancing SDI in Section 3.7.

3.1 Dataset: Human Dorsolateral Prefrontal Cortex (DLPFC)

As discussed in Section 1.1.8, our improved method also performs SDI on the DLPFC dataset [60] generated from the 10x Genomics Visium platform [28]. This dataset consists of twelve samples, comprising four replicates of tissue slices from each of three subjects. It includes three modalities: histology images, gene expression profiles, and spot coordinates. The histology image is a high-resolution Hematoxylin and Eosin (H&E)-stained image. The gene expression matrix is a high-dimensional, sparse matrix that records gene expression counts for each spatial spot. The spot coordinate matrix contains the two-dimensional spatial coordinates for each spot, along with a boolean value indicating whether the spot is under tissue.

For example, one sample (slide 151673) contains a gene expression matrix with 3,639 rows (spots) * 33,538 columns (genes), a histology image of size 7,966 (width) * 8,758 (height), and a spot coordinate matrix with 3,639 rows (spots) * 3 columns (2D coordinates and a boolean value), as shown in Figure 1.3. The ground-truth is provided by expert annotation from the Lieber Institute, cluster visualization of this sample is shown in Figure 3.1.

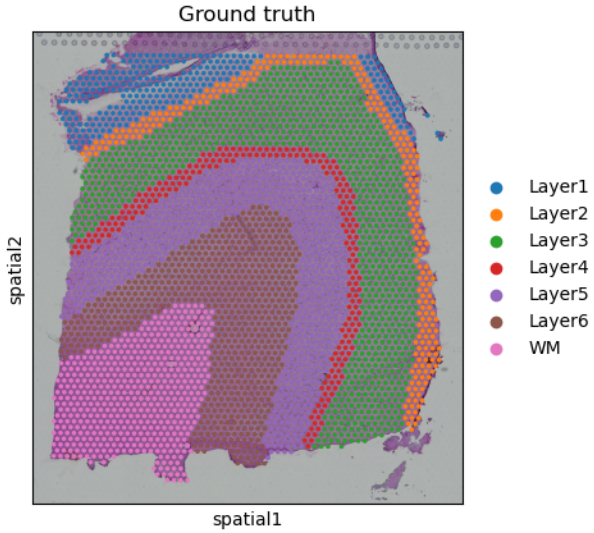


Figure 3.1: The ground truth visualization of slide 151673 from the DLPFC dataset (twelve slides in total), generated using Scanpy.

3.2 An Overview of Our SDI Method

Our proposed SDI method is summarized in Figure 3.2. We leverage two data modalities: histology image and spatial coordinates. Our contributions include optimization of the processing of histology images and clustering.

First, we introduce flexible patch sizes instead of relying on a fixed patch size, reducing the impact of artifacts in histology data on representation learning. Second, we adopt the UNI-2 [16] foundation model as the tokenizer, replacing the BYOL [29] framework. Unlike BYOL, UNI-2 offers stronger denoising and better generalizability, producing biologically meaningful embeddings from histology images. Third, we incorporate Moran's I assessment in the final clustering step to determine the optimal configuration for handling histology images.

Details on patch size and image embeddings are provided in Section 3.3 and Section 3.4, the optimized spatial clustering strategy is discussed in Section 3.6.

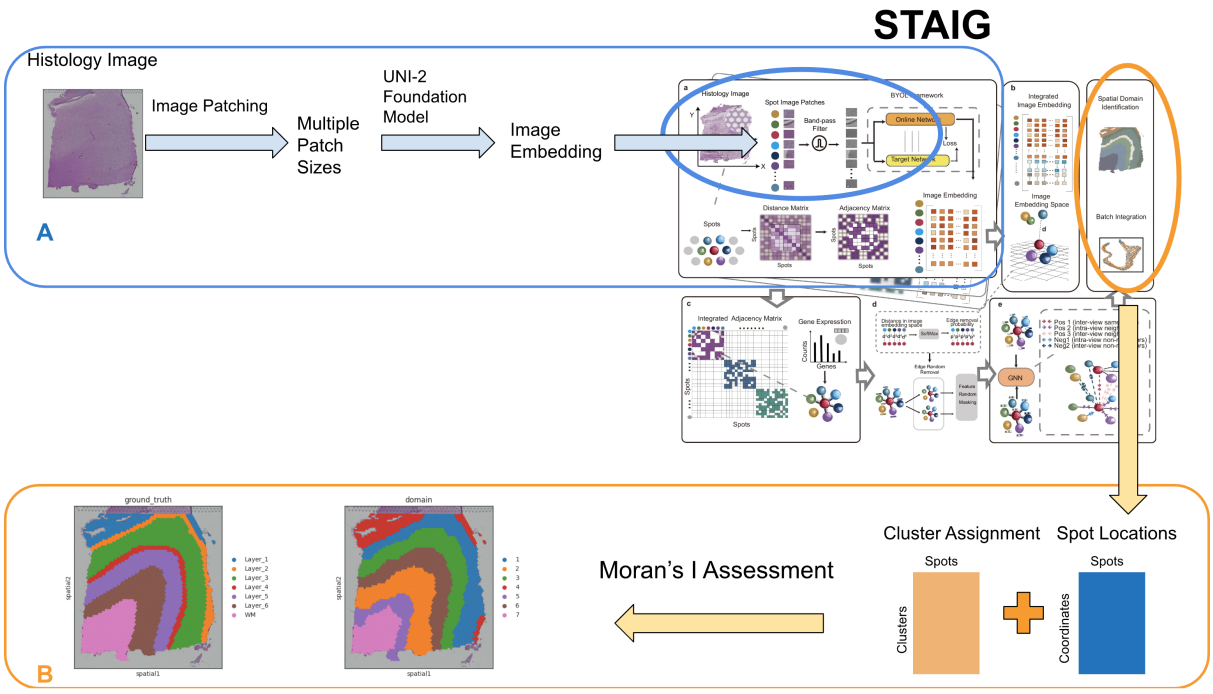


Figure 3.2: Overview of our proposed method. **A. Tokenization of Histology Images.** We adjust the patch diameter to 1.0 and 1.5 times the spot diameter and use the UNI-2 foundation model to tokenize these patches, replacing the Bootstrap Your Own Latent (BYOL) framework [29] utilized in the original STAIG. **B. Moran's I-Assisted Clustering.** We utilize spot locations and cluster assignments to calculate Moran's I for assessing spatial autocorrelation, thereby selecting the representation from the optimal configuration of patch size and tokenizer for clustering.

3.3 Patch Size and Artifacts in the Histology Image

The STAIG method uses an image patch size of 3.5 times the spot diameter. However, we find this setting to be suboptimal. Patches near the tissue edges at this scale often contain a large portion of background, and artifacts such as air bubbles in the slide (see panel C in Figure 3.3) introduce noise into the spot-level representations [79]. This is critical because image patch embeddings define spot-level pairwise relationships. When bubbles or

background dominate a patch, they can create spurious edges, making patches near bubbles or tissue edges appear artificially similar in the embedding space.

As shown in panel A of Figure 3.3, we visualize the cosine similarity of pairwise patch embeddings with a patch size of 3.5 times the spot diameter. Specifically, we construct KNN graphs based on spatial adjacency, then color the edge weights according to cosine similarity, where darker purple indicates higher similarity. In this visualization, spots surrounding bubbles or tissue edges form connections with high similarity. This is detrimental, because higher cosine similarity directly affects the edge-dropping probability later in the STAIG workflow, which is critical for spatial neighborhood modeling.

To mitigate this issue, we reduce the patch size to 1.0 and 1.5 times the spot diameter. Panels B of Figure 3.3 illustrates results using a patch size of 1.0 times the spot diameter. We observe that the spurious purple connections around bubbles and tissue edges disappear. This adjustment improves the quality of pairwise relationships between spots by reducing false edge formation, thereby avoiding irrelevant connections and allowing morphologically similar spots to form meaningful connections.

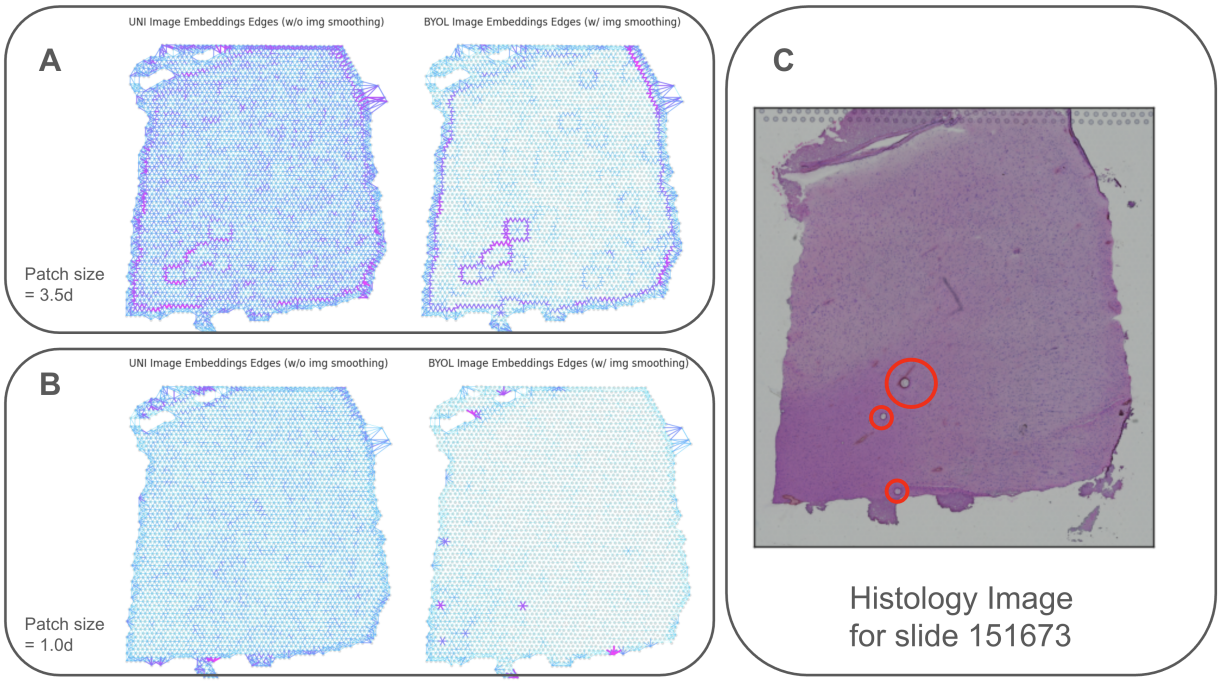


Figure 3.3: **A.** KNN graphs constructed using UNI [16] (left) and BYOL [29] (right) patch embeddings, with patches extracted at 3.5 times the spot diameter. Edges with high cosine similarity between embeddings are highlighted in purple. **B.** Same visualization with patches extracted at 1.0 times the spot diameter. **C.** Original histology image for slide 151673, with artifacts circled in red.

3.4 Image Embeddings from a Foundation Model: UNI-2

In addition to modifying the patch size, we replace the original BYOL [29] tokenization method with UNI-2 [16]. UNI-2, developed by the Mahmood Lab at Harvard Medical School, is a pathology foundation model capable of extracting more biologically meaningful features from image patches [16]. As discussed in Section 2.4.4, UNI-2 produces pathologically relevant embeddings while remaining resistant to noise.

Another advantage of UNI-2 is that it does not require image preprocessing, which is mandatory in BYOL. This allows UNI-2 to learn directly from histology images without ar-

tificially applied filters, thereby reducing the risk of misinterpreting histological information.

3.5 Configurations of Patch Sizes and Tokenizers

We experiment with image patches of 1.0 and 1.5 times the spot diameter, along with the 3.5-diameter patches originally used in STAIG. For each patch size, we apply both the BYOL pipeline and the UNI-2 pipeline for tokenizing the image patches, resulting in five configurations: (1) 1.0-diameter patches with BYOL embeddings, (2) 1.0-diameter patches with UNI-2 embeddings, (3) 1.5-diameter patches with BYOL embeddings, (4) 1.5-diameter patches with UNI-2 embeddings, and (5) 3.5-diameter patches with BYOL embeddings (default STAIG setting).

Our clustering process selects the optimal configuration for each sample in the DLPFC dataset. The detailed selection procedure is described in Section 3.6.

3.6 Automatic Patch Size and Tokenizer Selection: Moran's I

We achieve automatic configuration of patch size and tokenizer selection using Moran's I [52], a statistical measure of spatial autocorrelation that evaluates how well clustering assignments preserve spatial structure. Specifically, Moran's I quantifies the degree to which neighboring spots share similar cluster labels. In Formula 3.1 [52], N denotes the number of spots, W is the sum of spatial weights, w_{ij} represents the spatial weight between spots i and j, x_i is the cluster label of spot i, and $\bar{x}$ is the mean of all labels. The numerator measures the similarity of cluster assignments between neighboring spots, weighted by spatial proximity, while the denominator normalizes this value by the overall variance.

The value of Moran's I ranges from -1 to 1. A value close to 1 indicates strong positive spatial autocorrelation, meaning nearby spots tend to belong to the same cluster, which is the desired outcome for SDI. A value near 0 indicates randomness or absence of spatial structure, while negative values imply dispersion, where neighboring spots are deliberately different, which is not desirable in this context.

As illustrated in Figure 3.4, Moran's I provides a measure of spatial clustering quality

without requiring ground-truth labels. High-quality clustering yields large Moran's I values, reflecting strong alignment between clusters and anatomical layers. For moderate clustering, Moran's I decreases, indicating weaker spatial consistency. For poor clustering, Moran's I is very low, suggesting fragmented or random assignments. Therefore, Moran's I allows us to select the optimal clustering configuration by assessing spatial coherence directly, without relying on ground-truth labels.

$$I = \frac{N}{W} \cdot \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{3.1}$$

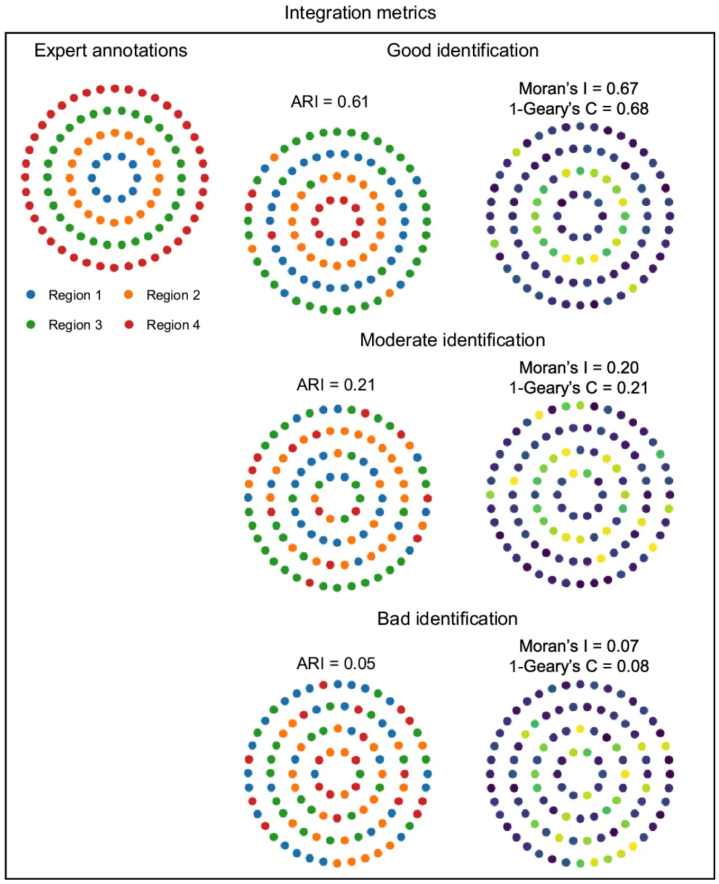


Figure 3.4: Examples of good, moderate, and bad identification of spatial regions evaluated using ARI and Moran's I. This is Figure 4 from "Combining Spatial Transcriptomics with Tissue Morphology" by Chelebian et al.[14]. Integration metrics measure the agreement of expert annotations with the domains defined jointly by morphology and spatial transcriptomics via the adjusted Rand index (ARI). It is common to also define spatially variable genes that represent the identified domains and measure their degree of spatial autocorrelation with Moran's I or Geary's C [14].

3.7 Future Works

Our proposed SDI method currently benefits primarily from 10X Visium data, as Visium provides high-resolution histology images. However, further improvements are needed for

the processing of gene expression profiles, enabling broader applicability across other ST platforms in addition to 10X Visium. Subsection 3.7.1 outlines directions for improving the abstraction of gene expression profiles, Subsection 3.7.2 and Subsection 3.7.1 discuss additional datasets for evaluating the robustness of our method.

3.7.1 Augmentation of Gene Expression Profile: Local and Global Views

STAIG [79] applies highly variable gene (HVG) selection (3,000 genes) to the gene expression matrix and uses the result directly as the node embedding for each spot without augmentation. However, this raw representation captures only a local view of each spot, neglecting the broader microenvironment and global context of gene expression.

To address this limitation, we can generate augmented matrices that capture both local and global information within a tissue slice. These representations enrich the gene expression profiles by incorporating the spatial microenvironment around each spot. Potential strategies include mean averaging of neighboring spots, Gaussian filtering, and azimuthal Gabor filtering as implemented in Banksy [67].

3.7.2 Joint Analysis of Visium and MERFISH Data

We can also utilize data from both the 10X Genomics Visium [28] and Vizgen MERFISH [15] platforms. These two types of ST sequencing platforms are publicly accessible and complement each other in terms of gene coverage and spatial resolution. The Visium platform provides spots larger than a single cell but offers high gene coverage in the expression profile. In contrast, MERFISH achieves subcellular resolution but lacks genome-scale gene coverage. By jointly analyzing datasets from both platforms on the same tissue type, we can gain better insights into cellular characteristics and tissue composition at higher resolution.

Currently, there are 109 public datasets (65 human samples and 44 mouse samples) available on the Visium platform [1], and 23 public datasets (10 human samples and 13 mouse samples) available on the MERFISH platform [74]. Additionally, the Allen Brain Cell

Atlas (ABC Atlas) [25] contains mouse brain ST data generated with both 10X Visium and MERFISH, which facilitates our joint analysis.

3.7.3 Simulated Data

Due to the limited availability of ST data, we can include simulated data in our evaluation. Recently, methods such as scDesign3 [68] and scMultisim [42] have emerged as tools for generating realistic single-cell and spatial omics data with ground truth labels. In contrast, most publicly available datasets lack ground-truth annotations.

Chapter 4

EVALUATION

In this chapter, we evaluate our proposed SDI method against the baseline STAIG [79] using all twelve samples from the DLPFC [60] reference dataset. We report Moran's I values across all tested configurations of patch sizes and tokenizers, illustrating how the automatic selection process identifies the optimal setting for each sample. The evaluation is carried out using two clustering metrics, ARI and NMI, and is further supported by visualizations of clustering assignments overlaid on histology images. Together, these results provide both quantitative and qualitative evidence of the improvements introduced by our method, demonstrating stronger SDI capability compared to the baseline.

4.1 Automatic Selection of Patch Sizes and Tokenizers with Moran's I

Table 4.1 shows the Moran's I values for all configurations to identify the best setting for each sample in the DLPFC dataset. For each sample, we select the configuration (shown in bold font) with the maximum Moran's I.

4.2 ARI and NMI across all Samples on the DLPFC Dataset

We present ARI and NMI values across the DLPFC dataset in Table 4.2. To better illustrate the improvements of our method over STAIG, we also provide two box plots, one for ARI and one for NMI. As shown in Figure 4.1, our method achieves a higher mean and median, along with lower dispersion, indicating more consistent performance across samples.

Table 4.1: Moran's I values across DLPFC samples for different configurations.

Sample	**BYOL, 3.5d**	**BYOL, 1d**	**BYOL, 1.5d**	**UNI-2, 1d**	**UNI-2, 1.5d**
151507	**0.8825**	0.8716	0.8723	0.8792	0.8742
151508	0.8186	0.8623	0.8629	0.8584	**0.8670**
151509	0.8958	**0.9029**	0.8961	0.8943	0.8936
151510	0.8738	0.8464	0.8755	0.8742	**0.8853**
151669	0.8541	0.8577	0.8536	**0.9108**	0.8593
151670	0.8575	0.8599	**0.8696**	0.8451	0.8552
151671	0.9162	0.9012	0.8928	**0.9178**	0.8842
151672	0.8758	**0.9187**	0.8788	0.8937	0.9077
151673	0.8958	**0.9029**	0.8961	0.8943	0.8936
151674	0.8958	**0.9029**	0.8961	0.8943	0.8936
151675	0.8606	**0.8748**	0.8726	0.8483	0.8595
151676	**0.8648**	0.8459	0.8641	0.8513	0.8609

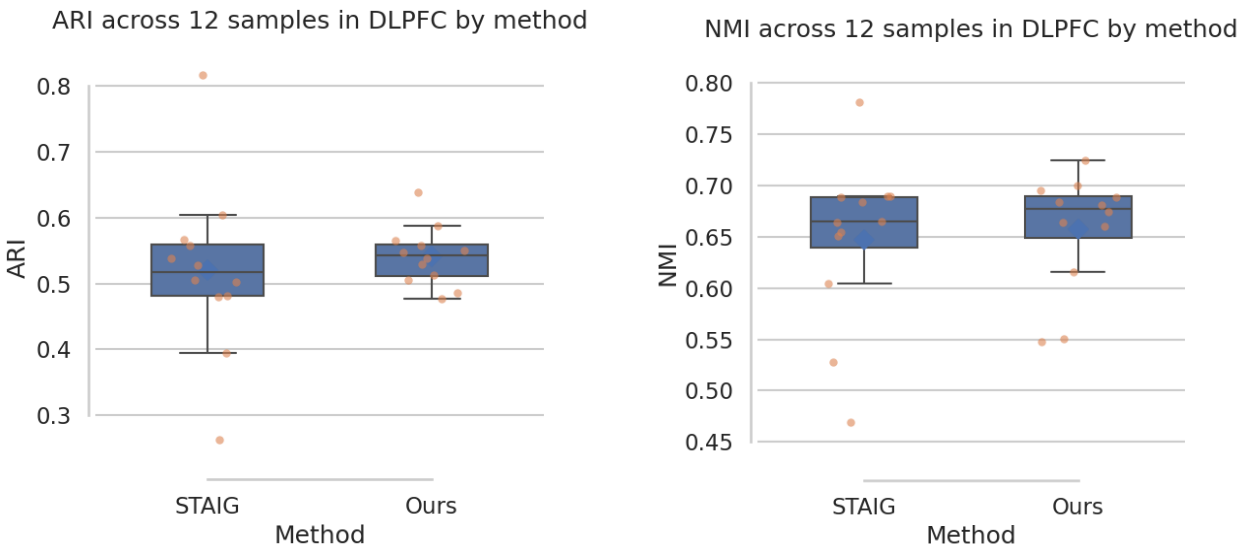
Table 4.2: Comparison of STAIG and our method using ARI (left) and NMI (right) across all DLPFC samples.

Sample	STAIG	Ours
151507	0.5666	0.5666
151508	0.3954	**0.4573**
151509	0.5029	**0.5503**
151510	0.4821	**0.5218**
151669	**0.6043**	0.4767
151670	0.2633	**0.5090**
151671	0.5052	**0.6121**
151672	**0.8168**	0.5295
151673	**0.5286**	0.5165
151674	0.5379	**0.5425**
151675	0.4801	**0.5247**
151676	0.5577	0.5577

(a) Adjusted Rand Index (ARI).

Sample	STAIG	Ours
151507	0.6897	0.6897
151508	0.5283	**0.6238**
151509	0.6549	**0.6741**
151510	0.6505	**0.6523**
151669	**0.6046**	0.5479
151670	0.4690	**0.5492**
151671	0.6642	**0.7054**
151672	**0.7815**	0.6603
151673	**0.6892**	0.6730
151674	**0.6882**	0.6484
151675	0.6649	**0.6667**
151676	0.6843	0.6843

(b) Normalized Mutual Information (NMI).



(a) ARI across samples.

(b) NMI across samples.

Figure 4.1: Comparison of clustering performance between STAIG and our method on all twelve DLPFC samples using ARI (a) and NMI (b). Our method achieves a higher mean, higher median, and less spread in both plots.

4.3 Visualization of Clustering Result across all Twelve Samples from the Reference Dataset DLPFC

In this section, we present a qualitative view of SDI performance through domain visualizations (Figure 4.2). Specifically, we compare each region identified by our method with the ground truth and the baseline method. Overall, our method demonstrates better recognition of spatial boundaries that align more closely with the biological regions (cortical layers) in the dataset.

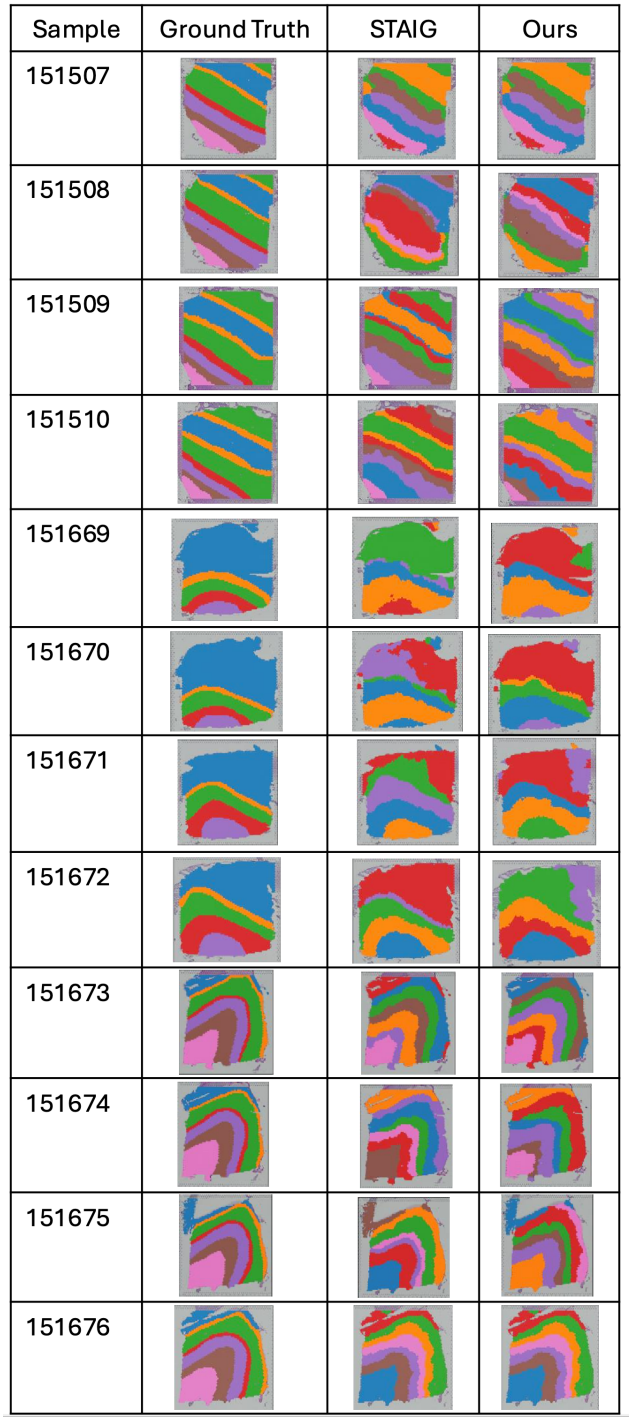


Figure 4.2: Visualization of SDI performance across the DLPFC dataset, with clustering labels overlaid on histology images. For each sample, the ground truth labels, the clustering results from STAIG, and the results from our method are shown.

Chapter 5

CONCLUSION

5.1 Conclusion

In this thesis, we improve the state-of-the-art SDI method STAIG [79] by implementing automatic patch size selection using Moran's I [14] and introducing the pathology foundation model UNI-2 [16]. These enhancements mitigate STAIG's sensitivity to artifacts and reliance on handcrafted configurations.

Our evaluation provides both quantitative and qualitative evidence of improvement. On the DLPFC dataset [60], our method achieves higher ARI and NMI scores compared to the baseline. Visualizations of clustering assignments further confirm that the predicted spatial domains align more closely with known cortical layers.

Nonetheless, some limitations remain. Our method currently requires high-resolution histology images to show improvements over the baseline. Moreover, foundation models like UNI-2 introduce higher computational costs compared to lightweight tokenizers.

Overall, our work demonstrates the potential of integrating foundation models into SDI pipelines and presents Moran's I as a practical tool for spatial clustering assessment without ground-truth labels. These contributions enable more accurate identification of spatial domains, ultimately facilitating the understanding of tissue microenvironment and disease diagnosis.

BIBLIOGRAPHY

[1] 10x Genomics. 10x genomics dataset, 2024.

[2] 10x Genomics. 10x genomics space ranger 2.1.0, 2024.

[3] 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):565, 2020.

[4] Lyla Atta and Jean Fan. Computational challenges and opportunities in spatially resolved transcriptomic data analysis. *Nature Communications*, 12(1):5283, 2021.

[5] Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, pages 1–19, 2023.

[6] Source BioScience. 10x visium spatial transcriptomics, 2024.

[7] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. 2008.

[8] Wilhelm Burger and Mark J. Burge. *Principles of Image Processing: Fundamental Techniques*. Springer, 2009.

[9] Darren J Burgess. Spatial transcriptomics coming of age. *Nature Reviews Genetics*, 20(6):317–317, 2019.

[10] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[11] Dylan M Cable, Evan Murray, Vignesh Shanmugam, Simon Zhang, Michael Z Diao, Haiqi Chen, Evan Macosko, Rafael A Irizarry, and Fei Chen. Cell type-specific differential expression for spatial transcriptomics. *bioRxiv*, 2021.

[12] 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, February 2021.

[13] 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, 2022.

[14] Eduard Chelebian, Christophe Avenel, and Carolina Wählby. Combining spatial transcriptomics with tissue morphology. *Nature Communications*, 16(1), May 2025.

[15] 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, 2015.

[16] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024.

[17] Uthsav Chitra, Brian J. Arnold, Hirak Sarkar, Kohei Sanno, Cong Ma, Sereno Lopez-Darwin, and Benjamin J. Raphael. Mapping the topography of spatial gene expression with interpretable deep learning. *Nature Methods*, 22(2):298–309, January 2025.

[18] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11, 2024.

[19] Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1):1739, 2022.

[20] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology*, 22:1–31, 2021.

[21] 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–e50, 2021.

[22] Long et al. Tutorial 1: 10x visium - graphst 1.1 documentation, 2024.

[23] Zhen Fan, Runsheng Chen, and Xiaowei Chen. Spatialdb: a database for spatially resolved transcriptomes. *Nucleic acids research*, 48(D1):D233–D237, 2020.

[24] Shuangsang Fang, Mengyang Xu, Lei Cao, Xiaobin Liu, Marija Bezulj, Liwei Tan, Zhiyuan Yuan, Yao Li, Tianyi Xia, Longyu Guo, et al. Stereopy: modeling comparative and spatiotemporal cellular heterogeneity via multi-sample spatial transcriptomics. *bioRxiv*, pages 2023–12, 2023.

[25] Allen Institute for Brain Science. Allen brain cell atlas, 2024.

[26] D. Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93(26):429–441, November 1946.

[27] 10X Genomics. Human breast cancer and mouse brain tissue sections datasets, 2024.

[28] 10X Genomics. Visium spatial whole transcriptome discovery in the tissue context, 2024.

[29] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[30] Hemalatha Gunasekaran, K Ramalakshmi, A Rex Macedo Arokiaraj, S Deepa Kanmani, Chandran Venkatesan, and C Suresh Gnana Dhas. Analysis of dna sequence classification using cnn and hybrid models. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

[31] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, June 2024.

[32] 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):1–12, 2017.

[33] Sharia Hernandez, Rossana Lazcano, Alejandra Serrano, Steven Powell, Larissa Kostousov, Jay Mehta, Khaja Khan, Wei Lu, and Luisa M Solis. Challenges and opportunities for immunoprofiling using a spatial high-plex technology: the nanostring geomx® digital spatial profiler. *Frontiers in Oncology*, 12:890410, 2022.

[34] Chiara Herzog. Single cell rna sequencing - a technique has come of age, 2024.

[35] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.

[36] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[37] National Cancel Institute. Nci dictionary of cancer terms, 2024.

[38] Da Hyun Kang, Yoonjoo Kim, Ji Hyeon Lee, Hyeong Seok Kang, and Chaeuk Chung. Spatial transcriptomics in lung cancer and pulmonary diseases: A comprehensive review. *Cancers*, 17(12):1912, June 2025.

[39] Anni Kivinen and Aliisa Tiihonen. From smoothies to fruit tarts: The development of spatial transcriptomics, 2022.

[40] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology*, 40(5):661–671, 2022.

[41] Lixin Lei, Kaitai Han, Zijun Wang, Chaojing Shi, Zhenghui Wang, Ruoyan Dai, Zhiwei Zhang, Mengqiu Wang, and Qianjin Guo. Attention-guided variational graph autoencoders reveal heterogeneity in spatial transcriptomics. *Briefings in Bioinformatics*, 25(3), March 2024.

[42] Hechen Li, Ziqi Zhang, Michael Squires, Xi Chen, and Xiuwei Zhang. scmultisim: simulation of single-cell multi-omics and spatial data guided by gene regulatory networks and cell–cell interactions. *Nature Methods*, 22(5):982–993, April 2025.

[43] Zhuliu Li, Tianci Song, Jeongsik Yong, and Rui Kuang. Imputation of spatially-resolved transcriptomes by graph-regularized tensor completion. *PLoS computational biology*, 17(4):e1008218, 2021.

[44] Teng Liu, Zhao-Yu Fang, Xin Li, Li-Ning Zhang, Dong-Sheng Cao, and Ming-Zhu Yin. Graph deep learning enabled spatial domains identification for spatial transcriptomics. *Briefings in Bioinformatics*, 24(3):bbad146, 2023.

[45] Zhaoyang Liu, Dongqing Sun, and Chenfei Wang. Evaluation of cell-cell interaction methods by integrating single-cell rna sequencing data with spatial information. *Genome Biology*, 23(1):1–38, 2022.

[46] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

[47] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14(1):1155, 2023.

[48] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[49] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.

[50] 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, 2022.

[51] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C Devlin, Maayan Baron, Cristina H Hajdu, Diane M Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature biotechnology*, 38(3):333–342, 2020.

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

[53] Niklas Müller-Bötticher, Shashwat Sahay, Roland Eils, and Naveed Ishaque. Spatialleiden: spatially aware leiden clustering. *Genome Biology*, 26(1), February 2025.

[54] NVIDIA. What are foundation models?, 2025. Accessed: 2025-08-25.

[55] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, Isaac Virshup, et al. Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, 19(2):171–178, 2022.

[56] Brenda Pardo, Abby Spangler, Lukas M Weber, Stephanie C Page, Stephanie C Hicks, Andrew E Jaffe, Keri Martinowich, Kristen R Maynard, and Leonardo Collado-Torres. spatiallibd: an r/bioconductor package to visualize spatially-resolved transcriptomics data. *BMC genomics*, 23(1):434, 2022.

[57] Viktor Petukhov, Rosalind J Xu, Ruslan A Soldatov, Paolo Cadinu, Konstantin Khodosevich, Jeffrey R Moffitt, and Peter V Kharchenko. Cell segmentation in imaging-based spatial transcriptomics. *Nature biotechnology*, 40(3):345–354, 2022.

[58] Duy Pham, Xiao Tan, Jun Xu, Laura F Grice, Pui Yeng Lam, Arti Raghubar, Jana Vukovic, Marc J Ruitenberg, and Quan Nguyen. stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv*, pages 2020–05, 2020.

[59] R. B. Potts. Some generalized order-disorder transformations. *Proceedings of the Cambridge Philosophical Society*, 48:106–109, 1952.

[60] 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, 2019.

[61] 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, 2019.

[62] Luca Scrucca, Chris Fraley, T. Brendan Murphy, and Raftery Adrian E. *Model-Based Clustering, Classification, and Density Estimation Using mclust in R.* Chapman and Hall/CRC, March 2023.

[63] Lulu Shang and Xiang Zhou. Spatially aware dimension reduction for spatial transcriptomics. *Nature Communications*, 13(1):7203, 2022.

[64] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[65] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[66] Xuejing Shi, Juntong Zhu, Yahui Long, and Cheng Liang. Identifying spatial domains of spatially resolved transcriptomics via multi-view graph convolutional networks. *Briefings in Bioinformatics*, 24(5):bbad278, 2023.

[67] Vipul Singhal, Nigel Chou, Joseph Lee, Yifei Yue, Jinyue Liu, Wan Kee Chock, Li Lin, Yun-Ching Chang, Erica Mei Ling Teo, Jonathan Aow, Hwee Kuan Lee, Kok Hao Chen, and Shyam Prabhakar. Banksy unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nature Genetics*, 56(3):431–441, February 2024.

[68] Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun, and Jingyi Jessica Li. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*, pages 1–6, 2023.

[69] Alex Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[70] Yuhao Tao, Xiaoang Sun, and Fei Wang. Bigatae: a bipartite graph attention auto-encoder enhancing spatial domain identification from single-slice to multi-slices. *Briefings in Bioinformatics*, 25(2):bbae045, 2024.

[71] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), March 2019.

[72] Vladimír Ulman. Filtering with anisotropic 3d gabor filter bank. In *Proceedings of the 11th International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1–13. 2010.

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[74] Vizgen. Data release program, 2024.

[75] Chloe Wang, Haotian Cui, Andrew Zhang, Ronald Xie, Hani Goodarzi, and Bo Wang. scgpt-spatial: Continual pretraining of single-cell foundation model for spatial transcriptomics. February 2025.

[76] Cameron G Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1):1–18, 2022.

[77] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[78] Hang Xu, Huazhu Fu, Yahui Long, Kok Siong Ang, Raman Sethi, Kelvin Chong, Mengwei Li, Rom Uddamvathanak, Hong Kai Lee, Jingjing Ling, Ao Chen, Ling Shao, Longqi Liu, and Jinmiao Chen. Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Medicine*, 16(1), January 2024.

[79] Yitao Yang, Yang Cui, Xin Zeng, Yubo Zhang, Martin Loza, Sung-Joon Park, and Kenta Nakai. Staig: Spatial transcriptomics analysis via image-aided graph contrastive learning for domain exploration and alignment-free integration. *Nature Communications*, 16(1), January 2025.

[80] Zhiyuan Yuan, Fangyuan Zhao, Senlin Lin, Yu Zhao, Jianhua Yao, Yan Cui, Xiao-Yong Zhang, and Yi Zhao. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21(4):712–722, March 2024.

[81] Yuansong Zeng, Rui Yin, Mai Luo, Jianing Chen, Zixiang Pan, Yutong Lu, Weijiang Yu, and Yuedong Yang. Identifying spatial domain by adapting transcriptomics with histology through contrastive learning. *Briefings in Bioinformatics*, 24(2):bbad048, 2023.

[82] Zexian Zeng, Yawei Li, Yiming Li, and Yuan Luo. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome biology*, 23(1):1–23, 2022.

[83] Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah EB Taylor, Paul Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology*, 39(11):1375–1384, 2021.