Contrastive Structural Transformers with Graph Corruption for Spatial Transcriptomics

Thesis by
Ahmad Farhan

In Partial Fulfillment of the Requirements for the
Degree of
MSc. Artificial Intelligence

TEESSIDE UNIVERSITY
Middlesbrough, United Kingdom

January 15, 2025

Ahmad Farhan
ORCID: 0009-0008-1951-6986

ABSTRACT

Spatial transcriptomics techniques are revolutionizing our ability to understand tumour microenvironments by enabling the analysis of gene expression within its spatial context. Unlike conventional sequencing methods which require complete tissue dissociation, ST enables the examination of gene expression in situ, which preserves the spatial relationships among cells Current graph-based spatial transcriptomics methods, especially for domain identification, often suffer from over-smoothing, where deeper network layers produce increasingly homogeneous node representations and reduce model expressiveness.

To overcome this challenge, we introduce ContraST, a self-supervised contrastive learning-based computational framework that encodes structural information of graphs into transformers to maintain distinct and informative node representations. Our model integrates a graph corruption strategy which randomizes gene expression representations across nodes while preserving the original graph structure. By modifying the readout function of Deep Graph Infomax with a global special node, it captures the graph's spatial relationships and gene expression patterns, enhancing clustering accuracy and spatial resolution in transcriptomics analysis.

ContraST has been rigorously evaluated on a range of human and mouse tissue datasets, including Stereo-seq and Slide-seq samples of the brain, pancreatic cancer, and lymph node tissues. Our method achieves groundbreaking performance on three pivotal tasks, including spatial transcriptomics deconvolution at cell-type resolution, spatial domain identification, and multisample integration. We show that ContraST improves the precision and interpretability of spatial transcriptomics and sets a new benchmark for downstream tissue analysis. The code and processed datasets are publicly available at Angione-Lab/contraST.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

C h a p t e r 1

INTRODUCTION

"We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried."

— Alan Turing, "Computing Machinery and Intelligence", 1950

In molecular and systems biology, we seek to understand how genes, proteins, and other molecular actors come together to drive cellular and organismal function. This field merges diverse data sources (e.g., transcriptomes, single-cell RNA-seq profiles, spatial proteomics, histopathological images) and applies computational methods to uncover hidden patterns and interactions. Guided by Alan Turing's pioneering insights, one of the aims in artificial intelligence is to give machines the ability to "see" and interpret these data as readily as we do. In systems biology, we face problems that require not only the detection of molecular signals but also an appreciation of how they are organized across spatial and temporal scales. A human pathologist can glance at a tissue and identify distinct domains or subtle changes that mark disease, yet a computer needs intelligent algorithms to analyze, interpret, and understand such interconnected information. This is of great importance, for misleading interpretations can be fatal.

Cells in a multicellular organism tend to remain grouped and act as clusters, and the dynamics of these clusters are critical to the health of tissues and organisms. Spatial transcriptomics, an innovative technique in molecular biology, measures gene expression in tissue without disrupting the natural cellular organization. Unlike single cell RNA sequencing, it preserves the spatial arrangement of cells, providing information about location-dependent cellular function and interaction.

Challenges. In spatial transcriptomics, each spot on the tissue gathers signals from multiple cells that may overlap. There are three major challenges in ST i.e. spatial domain identification, cell-type deconvolution and multisample integration. Goal of Spatial domain identification is to detect clusters of spots in tissue regions that appear to share similar molecular patterns (both gene expressions and spatial coordinates). However, noise and partial overlaps among neighboring areas make this a difficult computational effort. A second task cell-type deconvolution is to separate signals from distinct cell types when a spot includes more than one cell, which demands reference data and algorithms able to disentangle multiple profiles without discarding the local arrangement. A third task merges data from multiple experiments.

Small differences in protocols or equipment can hide important biological patterns or inflate artifacts. Approaches that combine high-capacity computation with spatial organization can limit errors that might otherwise distort research or clinical decisions. Recent advances in ST such as Visium and Slide-seqcan generate tens of thousands of spatially resolved spots across thousands of genes. We need intelligent computational methods with enough capacity to handle the scale and complexity of these tasks.

Remaining challenges. Despite constant advances in ST, these current graph-based approaches in this domain suffer from a significant drawback, i.e., over-smoothing in deeper layers of graph neural networks. This issue obscures the distinctions among cell types and spatial regions, leading to inaccurate identification of spatial domains, especially for complex tissue samples. Also, Current methods often rely on static graph constructions from spatial data that cannot adapt to varying cell densities or heterogeneous tissue structures. Primarily, the rigid connectivity patterns in these methods frequently fail to learn topological properties of tissue samples, which inhibit their utility in discussed downstream tasks.

Research Question. Can a contrastive learning framework with graph corruption and structural encoding in transformers enhance node representations in ST by reducing over-smoothing and improve spatial domain identification and cell-type resolution?

Outline of contributions. In this dissertation, we address a persistent difficulty in spatial transcriptomics and deep learning in general, which is oversmoothing in deeper layers of graph neural networks. Our work proposes a framework founded on contrastive learning, combining a graph corruption approach with an adaptive

scheme for data processing. Structural details are encoded within transformers, employing varied strategies to uphold homology in advanced layers. A corruption step (see Section 5) randomizes selected gene expression profiles while retaining original connectivity, promoting broader generalization and sustaining distinctive representations of spots. We also introduce a dynamic process for constructing graphs based on Gaussian Mixture Models and k-Nearest Neighbors to adjust connectivity based on local density. Our framework, Contrastive Structural Transformers with Graph Corruption for Spatial Transcriptomics (ContraST), is tested across core tasks in spatial transcriptomics and shows significant performance gains on spatial domain identification and cell-type deconvolution. Our main contributions are listed below:

- *Contrastive learning framework with graph corruption.* Encourage diverse and informative node representations which allows to build deep graph neural network. This reduces oversmoothing, maintains structural integrity and high dimensional non-linear relationships.
- *Transformer based model that embeds graph homology.* Integrate spatially guided graph information into transformers by introducing encoding techniques that improve spot representations.
- *Dynamic graph construction method via GMM and kNN:* Analyze the density variations in the spatial distribution of spots to construct the initial graphs for ST tissues data.
- *Evaluate ContraST on core tasks in ST:* Compare performance against current approaches on three downstream tasks including spatial domain identification, cell-type resolution, and multi-sample integration.
- *Aaptability of the framework across diverse tissue datasets:* Empirically prove the utility for human and mouse samples using different Stereo-seq and Slide-seq datasets.

Chapter 2

RELATED WORK

Spatial domain identification is the primary focus of SRT as it allows the systematic characterization of the spatial organization of tissues through gene expression mapping. Methods have progressively evolved to integrate spatial and transcriptional data with increasing expressivity. Early strategies such as in situ hybridization (ISH) (Papouchado et al., 2010) and spatially barcoded reverse transcription primers (Jayaraman et al., 2023) established the foundation for maintaining spatial context in gene expression studies. Building upon these techniques, modern computational methods incorporate spatial dependencies to refine domain segmentation.

A recent technique, Giotto (Del Rossi et al., 2022) uses a Markov Random Field framework to link spatially adjacent capture spots. By learning the spatial relationships between spots, Giotto (Dries et al., 2021) achieves more contiguous and biologically meaningful clustering. BayesSpace applies a Bayesian framework to spatial transcriptomics data, leveraging neighborhood information to refine clustering and better align identified domains with spatial continuity. Recent experimental methods such as sequential FISH (seqFISH and seqFISH+), MERFISH, Slide-seq, and Stereo-seq (Shah et al., 2018; M. Zhang et al., 2021; Rodriques et al., 2019; Wei et al., 2022) have further improved resolution for spatial transcriptomics and gene throughput for detailed transcriptome mapping across diverse tissues. Methods like Spa-GCNJ. Hu et al., 2021 as well ass stLearn (Pham et al., 2020) integrate morphological data and histology images with gene expression data. This improves clustering accuracy by considering both spatial and transcriptional similarities.

Cell-type deconvolution. Current techniques such as STAGATE, SEDR (Xu et al., 2024), SCAN-IT (Cang et al., 2021), and RESEPT (Y. Chang et al., 2022) learn a low dimensional latent representation through autoencoders for gene expression and spatial information, and segment domains through embedding clustering. SpaceFlow (Ren et al., 2022), and GraphST (Long et al., 2023) utilize graph corruption and contrastive learning techniques to improve the generalisation and accuracy of spatial domain identification. Recent techniques use Graph Attention Networks (Velickovic et al., 2017) and Graph Convolutional Networks (S. Zhang et al., 2019) can suffer from oversmoothing (**li2023improves**), where the feature representations

of nodes across the entire graph become indistinguishably similar. This results in distinct spatial spots having nearly identical gene expression profiles, losing the ability to capture the subtle differences crucial for identifying distinct spatial domains. Additionally, oversmoothing blurs the boundaries between different spatial clusters, making it difficult to distinguish regions with different biological characteristics and leading to poor spatial resolution in the analysis.

Multi-sample Integration. A significant challenge in spatial transcriptomics is the need to analyze extensive tissue areas that exceed the capture capabilities of current technologies. Achieving single-cell resolution remains difficult as current platforms either have larger capture spots or suffer from high dropout rates. Computational methods for cell-type deconvolution, such as stereoscope (Andersson et al., 2020), RCTD (Cable et al., 2022), CARD(Cable et al., 2022), cell2location (Kleshchevnikov et al., 2022), and LIGER (Longo et al., 2021), often ignore spatial information, limiting their accuracy. These methods optimize a matrix of cell-type composition without fully leveraging the spatial context. Although cell2location also calculates gene expressions corresponding to different cell types for each location, spatially informed deconvolution methods are still needed. Moreover, traditional batch removal methods for scRNA-seq, like Harmony (Korsunsky et al., 2019) and scVI (Lopez et al., 2018), do not incorporate spatial information, limiting their effectiveness for spatial transcriptomics data. While STAGATE can analyze multiple slices, it struggles with batch effect removal. Therefore, there is a critical need for spatially informed integration methods to enhance cell-type deconvolution and the accuracy of spatial transcriptomics analyses.

Chapter 3

BIOLOGICAL SYSTEMS

3.1 Cell, the building block of life

Cells constitute the fundamental structural and functional units that underlie all biological systems, providing a delineated setting for orchestrating essential molecular processes. By regulating the passage of ions, nutrients, and signaling factors through a selectively permeable plasma membrane, cells preserve a tightly controlled intracellular environment important for metabolic homeostasis. Within this confined microenvironment, organelles, including mitochondria and ribosomes, facilitate energy metabolism for sustaining core cellular activities. The nucleus being one of the most important organelle, safeguards genomic integrity and modulates transcriptional programs that govern cell cycle progression, adaptive responses, and intercellular communication. Although cells share a conserved set of structural components, the emergence of tissue- and organ-level complexity arises from spatially and functionally specialized cell populations, which collectively create higher-order architecture and coordination. Through these tightly regulated processes, multicellular organisms achieve remarkable functional heterogeneity while maintaining a coherent framework for survival and adaptation.

Within eukaryotic systems, the process of converting genomic instructions into biologically active macromolecules proceeds through tightly orchestrated phases. First, transcription selectively copies discrete segments of DNA into messenger RNA (mRNA) that, once synthesized, migrates to ribosomal assemblies. There, the prescribed amino acid sequence is polymerized to yield proteins that provide critical structural, enzymatic, or regulatory functions. Notably, certain RNA transcripts do not encode proteins but instead act as intricately regulated effectors that modulate transcriptional or post-transcriptional events. By fine-tuning mRNA abundance or stability, such noncoding RNAs enable dynamic adjustment of proteomic outputs in response to the cell's functional demands or shifting environmental inputs.

This comprehensive flow of information, from DNA to RNA and subsequently to protein—is often formalized as the central dogma. DNA serves as the long-term repository of hereditary material, which is transcribed into RNA molecules, some of which (mRNAs) then serve as direct templates for polypeptide synthesis. Although

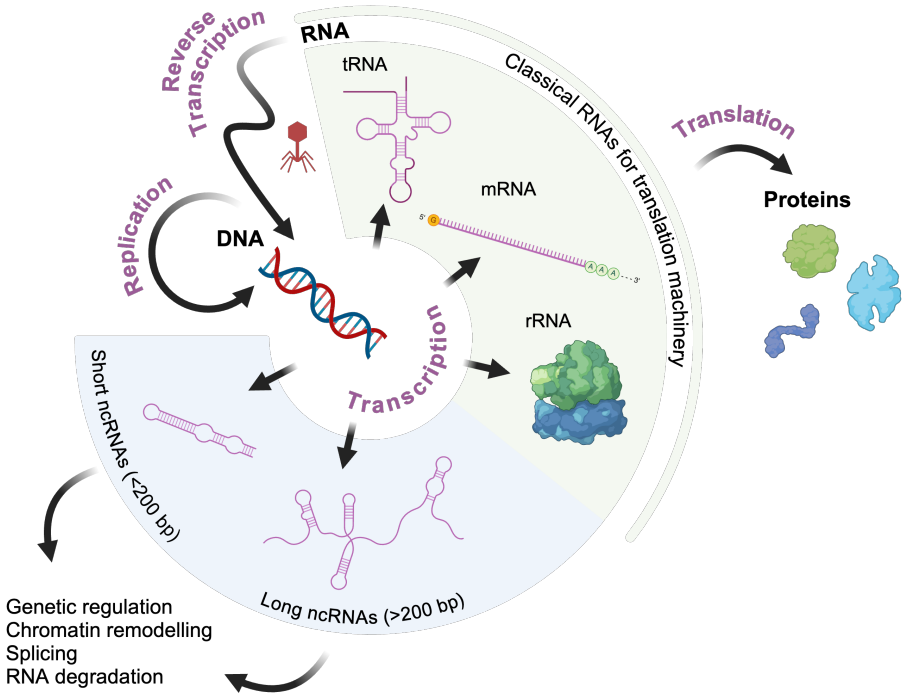

Figure 3.1: central dogma of molecular biology

this framework might appear linear, cells incorporate additional checkpoints, including splicing mechanisms that rearrange exonic elements, epigenetic alterations that influence DNA accessibility, and post-translational modifications that customize protein activity. Nevertheless, the core principle of the central dogma remains pivotal for describing how encoded information is translated into cellular function. Perturbations at any point—ranging from replication inaccuracies to errors in RNA processing—can disrupt key molecular circuits and destabilize normal homeostatic processes.

Within eukaryotic nuclei, DNA is packaged into linear chromosomes, each a dynamic chromatin assembly that modulates access to genetic loci. Collectively, these chromosomes constitute the organism's genome, encapsulating the full repository of heritable information required for development, physiological homeostasis, and adaptive responses. Although virtually every somatic cell in a multicellular system harbors an identical genomic blueprint, distinct cellular lineages activate or suppress discrete genomic regions in accordance with specialized functions. Thus, muscle cells preferentially express contractile apparatus-associated genes, whereas neurons prominently engage synaptic transmission-related transcriptional programs.

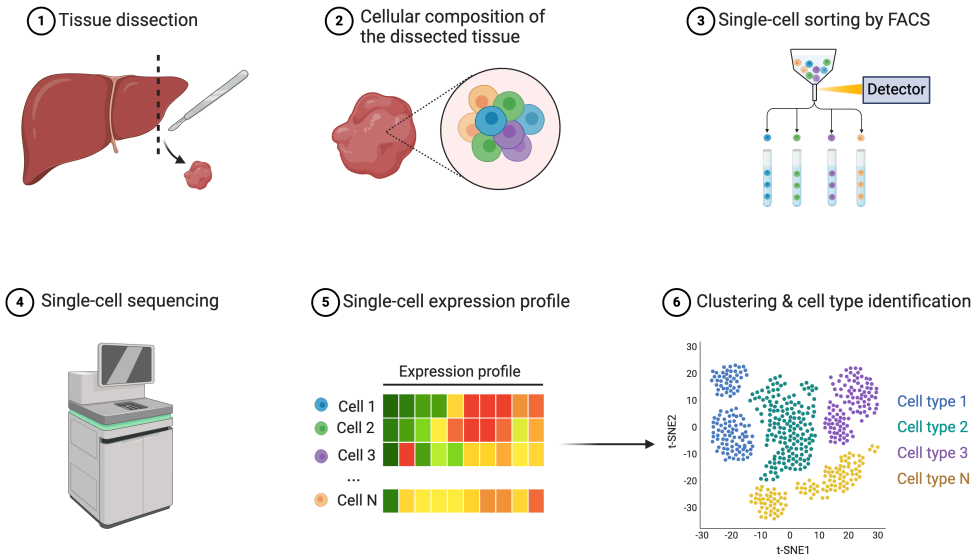

Figure 3.2: scRNA-sequencing

By delineating which genomic segments are transcriptionally active or quiescent in diverse cellular contexts, investigators elucidate both the mechanisms of lineage commitment and how deviations in these finely tuned gene networks can precipitate pathological states. In turn, such insights clarify why certain cells lose their canonical expression patterns and contribute to disease, highlighting the pivotal role of orchestrated gene regulation in maintaining organismal integrity.

3.2 Gene Expression Profiling

Within contemporary biology, understanding how genes are switched on or off—both within single cells and across entire tissues—is important for understanding core developmental pathways and disease causes. Referred to as gene expression, this coordinated process leads to the the transcription of DNA into RNA and, when relevant, the translation of select RNAs into functional proteins. By measuring the abundance of these RNA transcripts, we can capture a real-time readout of a gene's activation status, thereby revealing the underlying molecular circuits that shape cellular states, govern tissue organization, and predispose to pathological conditions.

From Bulk RNA Sequencing to Single-Cell Resolution

Initial methodologies for measuing gene expression predominantly utilized bulk RNA sequencing (bulk RNA-seq), where RNA is concurrently extracted from extensive cell populations or entire tissue specimens. This technique produces an averaged transcriptomic profile which obscures essential cell-to-cell variability and potentially hide the presence of rare cellular subpopulations. As a result, important transcriptional distinctions and heterogeneous cellular states remain undetected within bulk analyses.

To alleviate these issues, single-cell RNA sequencing (scRNA-seq) was developed which makes the isolation possible and allows individual transcriptomic profiling of single cells. This improved resolution has fundamentally transformed our capacity to explore cellular heterogeneity, identify novel cell types, and characterize dynamic cellular states. In spite of that, scRNA-seq typically involves tissue dissociation, which can damage the native spatial organization of cells within their microenvironments. While single-cell datasets explain transcriptional diversity with high precision, they inherently forfeit spatial contextual information, thereby limiting the ability to correlate gene expression profiles with specific anatomical locations. This trade-off emphasize the vital importance of integrative approaches that combine single-cell resolution with spatial context for comprehensive understanding of cellular function and tissue architecture.

Spatial Transcriptomics

Spatial transcriptomics solves the limitations inherent in traditional gene expression profiling methods by quantifying transcripts while maintaining their spatial context within intact tissue architectures. Unlike conventional sequencing methods like bulk RNA or scRNA-seq, which require complete tissue dissociation, ST enables the examination of gene expression in situ, which preserves the spatial relationships among cells. Spatial transcriptomics platforms are broadly classified into two primary categories based on the methods used to obtain gene expression profiles. *image-based ST* is the first type of Spatial transcriptomics, such as in situ hybridization (ISH), MERFISH, seqFISH, and other fluorescence-based techniques, which rely on microscopy to detect targeted RNA molecules within tissue sections. These methods often provide very high spatial resolution, even down to subcellular levels. However, sophisticated multiplexing schemes are required to mitigate the limitation of the number of genes they can profile simultaneously.

Image-based SRT techniques provide higher spatial resolution but have limited gene throughput and are often being confined to specific target genes. The processes involved in multiplexing and image acquisition can also require considerable effort and time which makes them less suitable for large scale studies. To resolve these scalibilty issues, another category of SRT called *Sequencing-based SRT*, advances gene expression profiling by capturing RNA transcripts directly within their spatial contexts on a solid surface. One widely recognized example is the Visium platform by 10x Genomics, which builds on an approach called spatial transcriptomics. Visium uses glass slides with microarrays of capture regions, each containing millions of oligonucleotides encoded with spatial barcodes, unique molecular identifiers (UMIs), and polyT sequences to capture polyadenylated transcripts. This method has reduced spot diameters from 100 µm to 55 µm and increased the number of spots per capture area from about 1,000 to up to 14,000 which significantly improve both resolution and scalability.

Another recent example of sequencing-based SRT is Stereo-seq. It improves spatial transcriptomics with DNA nanoballs (DNBs) carrying spatial barcodes on a solid surface. The method achieves a 2 µm spot size which results in ultra high resolution transcriptome maps across entire tissue sections. This thesis uses datasets from Visium and Stereo-seq to study spatial gene expression and tissue structure across different biological systems.

Chapter 4

DEEP LEARNING

Shift Toward Computational Methods in Spatial Transcriptomics. Using analytical pipelines for ST such as basic clustering or dimensionality reduction techniques have provided deep understanding of tissue organization and gene expression but they often rely on rigid parametric assumptions or simplified distance metrics. These constraints become obvious as newer platforms (e.g., Visium (Genomics, n.d.), and Stereo-seq (A. Chen et al., 2022)) generate tens of thousands of spatially resolved spots across thousands of genes and provide supplementary data streams like histology images or protein markers. The multidimensional nature of these datasets, combined with ever-growing throughput, poses scalability issues that result in inadequate analyses or overlooked biological signals. Consequently, there is a growing drive to adopt more advanced computational strategies capable of integrating these multimodal data sources and learning patterning across different tissue architectures. Deep learning (DL) frameworks—by virtue of their hierarchical feature extraction, capacity for large-scale data handling, and flexibility in modeling nonlinear dependencies—are well-positioned to meet these demands.

Definition and Historical Context. Deep learning, a subset of representation learning, is a computational framework in which a cascade of parameterized layers is trained to approximate increasingly involuted functions (**goodfellow2016deep**). The goal is to learn the underlying relationships in observations by pattern mining which makes it ideal for applications where data is complex, noisy and high dimensional. It has been highly utilised in some of the important tasks in computational biology like Cell fate identification, perturbation response prediction, and a few important downstream tasks in Spatial transcriptomics like spatial domain identification, cell type deconvolution. Below is a brief overview of some of the foundational models used in contraST framework.

Multilayer Perceptron

Multilayer Perceptron (MLP) are the quintessential foundational building blocks of deep learning architectures. Given a set of input features, MLP calculates the weighted combination of these features and weights, where the weights represent the importance of each feature. Following this linear combination, usually a non-

linearity, known as an activation function, is applied to increase the expressivity of the network. For a network of L layers, the activations of the ℓ-th layer $h^{(\ell)}$ are computed from the previous layer $h^{(\ell-1)}$ using the following transformations:

$$z^{(\ell)} = W^{(\ell)} h^{(\ell-1)} + b^{(\ell)}, \tag{4.1}$$

$$h^{(\ell)} = \sigma\left(z^{(\ell)}\right), \tag{4.2}$$

where $W^{(\ell)}$ and $b^{(\ell)}$ are the weight and bias vector for layer ℓ, respectively, and $\sigma(\cdot)$ is a non linearity such as ReLU or GeLU. First layer $h^{(0)}$ typically represents the input features, while the final layer $h^{(L)}$ is usually a classifier or regressor that outputs the network logits.

Autoencoders

Autoencoders are a special case of neural networks that reconstruct the input by learning a low-dimensional latent representation h. They consist of two main components. An encoder function $h = f_\theta(x)$, which compresses the original input x into a latent space. A inverse counterpart decoder $\hat{x} = g_\phi(h)$, which reconstructs x from this low-dimensional embedding. Here, h typically satisfies $h \ll x$ which means that the latent representation has a much lower dimensionality than the input. If an autoencoder simply learns to set $g_\phi(f_\theta(x)) = x$ everywhere, it fails to be useful.

Through this optimisation, autoencoders can capture the latent code of the data distribution that can be used in downstream computational biology tasks such as SDI, counterfactual prediction and perturbation response estimation. During training, autoencoders minimize a reconstruction loss $\mathcal{L}_{\text{rec}}$ that penalises the difference between input $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$. A common choice for this loss is the mean squared error (MSE):

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)} \right\|_2^2,$$

Here N denotes total samples in the training set. The parameters of both the encoder (θ) and the decoder (ϕ) are updated via standard stochastic gradient descent optimization to minimize $\mathcal{L}_{\text{rec}}$. This restricts it to learn a latent representation h that preserves only the important information of the data.

Self-Attention

Self-attention learns a convex combination of inputs based on their neighborhood to capture relationships and semantic similarity. It can model long-range dependencies by estimating how important each input is for any other input with $O(1)$ complexity per layer to attend to a single input. Given an input matrix $X \in \mathbb{R}^{n \times d}$, where each row represents a feature vector, self-attention computes three linear transformations to produce the query (Q), key (K), and value (V) matrices:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V, \quad W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}.$$

A common form of self-attention is Scaled Dot Product Attention (SDPA), where each row of Q is compared with every row of K through dot products to compute similarity scores. The softmax function:

$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

maps these scores into a probability distribution with a range of $[0, 1]$, which is used to form convex combinations of the rows in V. This results in a new representation that captures the interactions for a single input with respect to its neighbors.

The Transformer Architecture

Originally introduced for sequence-to-sequence modeling tasks in natural language processing (Vaswani, 2017), the Transformer architecture generalizes beyond text processing and has since proven effective in diverse domains. It stacks multiple layers, each consists of the following components:

Multi-Head Attention heads operate on distinct linear projections of Q, K, and V, capturing different aspects of the data. Their outputs are concatenated and projected:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\big(\text{head}_1, \ldots, \text{head}_h\big)W^O, \tag{4.3}$$

$$\text{head}_i = \text{Attention}\big(QW_i^Q,\ KW_i^K,\ VW_i^V\big)$$

$$W_i^Q, W_i^K, W_i^V, W^O$$

are trainable parameters.

Multilayer-perceptron A position-wise, fully connected sublayer applies a non-linear transformation to each element independently, enhancing representational capacity beyond simple weighted averaging.

Residual Connections and Layer Normalization Each sublayer is wrapped with residual shortcuts, and outputs are normalized to stabilize training, allowing deep stacks of attention blocks to converge more effectively.

To incorporate ordering or spatial structure, the Transformer additionally includes positional information by adding or concatenating learned or sinusoidal embeddings to the input at each layer. This technique ensures that the architecture can distinguish between different positions—even though the attention mechanism itself is indifferent to index ordering. Beyond language tasks, this general-purpose design has shown remarkable promise as a backbone for computer vision, speech analysis, and, as explored in this thesis, spatial transcriptomics.

Chapter 5

CONTRAST

Contrastive Structural Transformers

We introduce ContraST, a self supervised contrastive learning-based computational framework which is designed to address the issue of oversmoothing in graph-based spatial transcriptomics methods Long et al., 2023. It does so through its spatial transcritomics transformer-based encoder (STFormer) which incorporates different structure information through spot connectivity and spatial context encodings. Through its self-attention mechanism, STFormer captures and reinforces the structural and spatial relationships and enables the encoding of local and global contextual information. This mechanism forces the learned representations for spots to capture spatial relationships and cell-type-specific patterns which improves clustering accuracy and provides robustness against homogeneity in deeper network layers. The ability of STFormer in ContraST is validated in multiple fundamental downstream tasks in spatial transcriptomics analysis including identification of spatial domain, multi-sample integration and deconvolution at cell-type resolution, where the learned contextual embeddings are crucial for accurately analyzing spatial patterns and molecular signatures in the data. In all tasks, we used the gene expressions of spots to construct an implicit neighborhood graph, connecting spots that are spatially close to each other.

We consider a spatial transcriptomics dataset D consisting of a graph $G = (V, E)$, where each spot $v_i \in V$ is associated with a gene expression vector $x_i \in \mathbb{R}^n$. The encoder $E_s^{\phi} : \mathbb{R}^n \to \mathbb{R}^d$, learns intermediate latent representations for each spot. STFormer is a graph transformer Ying et al., 2021 based encoder that uses spatial and centrality encodings to learn distinct and informative node representations which mitigates the issue of oversmoothing in traditional graph-based spatial transcriptomics approaches. We adopt a graph corruption method wherein a corrupted version of the graph, G_c, is created by permuting the learned gene expression representations x_i among the spots while preserving the original spatial graph's adjacency matrix. This corruption strategy simulates variations in gene expressions across different spatial spots and reflects the variability observed in biological tissues where similar cell types often share local spatial contexts.

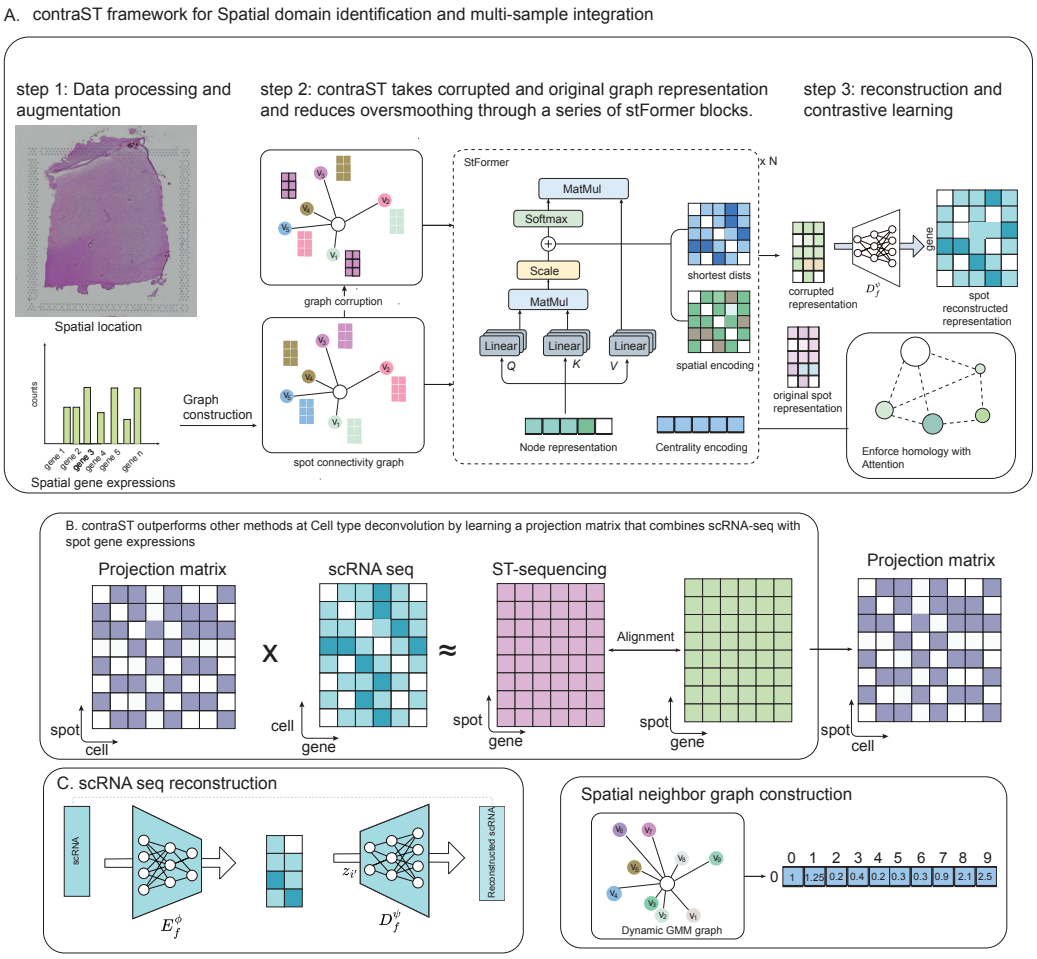

Figure 5.1: Overview of the ContraST framework for spatial domain identification and multi-sample integration. (A) Data processing, augmentation, and graph construction. (B) Cell type deconvolution using a projection matrix that integrates scRNA-seq and ST-seq data. (C) scRNA-seq reconstruction aligned with spatial transcriptomics. (D) Adaptive spatial neighbor graph construction with a dynamic Gaussian Mixture Model (GMM).

The STFormer encoder is shared across the original graph G and the corrupted graph G_c enable learning of intermediate latent representations h_i and $h_{i'}$ for the real and corrupted spots. The encoder shared between the two graphs permits the model to view real local neighborhood context in G and context altered in G_c. In this manner, the formulation guarantees that the learned representations indeed capture spot-specific local neighborhood information in addition to the variation from spot to spot. Our main aim is to minimize the distance between the embeddings of the same spot in G and G_c while at the same time maximizing the distance between the embeddings of distinct spots across original and corrupted graph. Specifically,

positive pairs are $(h_i, h_{i'})$ where h_i is in the original graph G and $h_{i'}$ is the same spot in the corrupted graph G_c, while negative pairs are $(h_i, h_{j'})$ where $h_{j'}$ is a different spot in G_c.

$$\min_{h_i, h_{i'}} \|h_i - h_{i'}\|, \quad \max_{h_i, h_{j'}} \|h_i - h_{j'}\| \quad \forall i \neq j \tag{5.1}$$

This process pulls the embeddings of each spot towards its correct gene expression context and push them away from the shuffled,incorrect ones provided by G_c. As a result, spatially nearby spots have similar embeddings while non-adjacent ones retain different embeddings. This mechanism increases the discriminator power and makes the embeddings sensitive to the fine variations among the spatial transcriptomics data and leading to better performances of spatial analyses such as clustering and deconvolution. Inspired by earlier work Ying et al., 2021, STFormer integrates structural information into transformers through the use of the following encodings.

Adaptive Graph Construction in Spatial Transcriptomics

Spatial transcriptomics leverages tissue organization to identify similar cell states in close proximity, enabling the mapping of tissue architecture. To fully leverage this spatial information, we use an undirected neighborhood graph $G = (V, E)$, where V represents the spots and E the edges connecting them based on spatial proximity. The graph is characterized by an adjacency matrix A where $A_{ij} = 1$ if spot j is a neighbor of spot i, and $A_{ij} = 0$ otherwise. The traditional method selects k-nearest neighbors based on Euclidean distances to define neighbors.

To refine this approach, we employ a Gaussian Mixture Model (GMM) to adaptively determine k, the number of neighbors for each spot, by analyzing the density variations in the spatial distribution of spots. The GMM fits a mixture of Gaussian distributions to the Euclidean distance data derived from initial neighbor computations, thereby modeling the probability distribution of spot distances as follows:

$$p(x) = \sum_{c=1}^{n} \pi_c \mathcal{N}(x|\mu_c, \Sigma_c)$$

where x represents the distance data, π_c are the mixture weights, and $\mathcal{N}(x|\mu_c, \Sigma_c)$ are the component Gaussian distributions with means μ_c and covariances Σ_c. Each spot i is assigned to the most probable component. The number of neighbors k_i, is determined by the number of spots associated with the same component and the

adjacency matrix A update:

$$k_i = \sum_{j=1}^{N_{\text{spot}}} \mathbf{1}_{\{c(j)=c(i)\}}, \quad A_{ij} = \begin{cases} 1 & \text{if } j \in N_i(k_i) \\ 0 & \text{otherwise} \end{cases}$$

This adaptive method enhances the graph's ability to reflect the true biological and spatial complexity of the tissue and improves the accuracy of subsequent downstream analyses. The dynamic adjustment of k based on local density estimates from the GMM allows for a more flexible network structure which is capable of capturing variations across different tissue types and conditions.

Spot Connectivity Encoding

Connectivity encoding assigns each spot two additional real-valued vectors, $z_{\deg^-}(v_i)$ and $z_{\deg^+}(v_i)$, computed from the adjacency matrix and are defined to be the in degree and out degree of the spot, respectively. The in-degree $\deg^-(v_i)$ is the total incoming connections and reflects the number of other spots that influence v_i. The out-degree $\deg^+(v_i)$ represents the total outgoing connections, capturing the number of other spots influenced by v_i. The initial representation of a spot v_i is then updated by incorporating these connectivity embeddings into its gene expression vector x_i:

$$h_i^{(0)} = x_i + z_{\deg^-}(v_i) + z_{\deg^+}(v_i) \tag{5.2}$$

Here, $h_i^{(0)}$ is the new representation of the spot v_i, which now includes both the gene expression data and the centrality information. This augmentation is particularly important for ST data as it allows the model to capture not only the molecular profile of each spot but also its implicit positional significance within the tissue. By integrating centrality measures, the model can more effectively identify key spots that are central to certain tissue functions or structures from those that are more peripheral.

Spatial Context Encoding

We implemented this spatial encoding to address the non-Euclidean spatial arrangement of spatial transcriptomics, where tissue spots are connected in an irregular, tissue-specific manner. By modifying the corresponding causal mask within the graph, we can more effectively model the spatial dependencies and proximities between spots. This approach mitigates the oversmoothing problem by capturing the

spatial relationships which ensures each node's representation maintains its uniqueness by reflecting the true topological distance rather than collapsing into overly similar representations through repeated convolution operations. For any pair of spots v_i and v_j, the spatial encoding function $\phi(v_i, v_j)$ is defined by the SPD between them. This encoding modifies the attention mechanism with information about the relative proximity and connectivity of spots. The attention mechanism is updated to include this spatial encoding as a bias term:

$$A_{ij} = \frac{(QW_Q)(KW_K)^T}{\sqrt{d}} + b_{\phi(v_i,v_j)} \tag{5.3}$$

where Q and K are the query and key matrices, W_Q and W_K are their corresponding weights, d is the dimensionality, and $b_{\phi(v_i,v_j)}$ is a learnable scalar indexed by $\phi(v_i, v_j)$.

Additionally, we modify the connectivity mask within the attention mechanism. This mask assigns an attention score of $-\infty$ to spots that are more than k hops apart which limits the attention mechanism to consider only the relevant spots. The masked attention scores are computed as:

$$A_{ij} = \begin{cases} \frac{(QW_Q)(KW_K)^T}{\sqrt{d}} + b_{\phi(v_i,v_j)}, & \text{if } \phi(v_i, v_j) \leq k \\ -\infty, & \text{if } \phi(v_i, v_j) > k \end{cases} \tag{5.4}$$

This integration of SPD-based biases and connectivity masking enhances the model's ability to focus on relevant contexts. This refined attention mechanism addresses the inherent challenges of balancing the preservation of local spatial context with the incorporation of relevant distant context. This approach is particularly useful for spatial domain identification, as it enables the model to accurately cluster both closely situated and faraway regions with similar gene expression profiles which improves the overall representational power and sensitivity to the tissue's spatial structure.

Graph-Level Representation with Dynamic Readout Node

In traditional GNNs, the readout function aggregates node features to generate a graph-level representation, often by summing or averaging the features of all nodes. However, this approach can be limited as it treats all nodes equally and may miss out on capturing the global context effectively, especially in the presence of noisy or

corrupted data. In STFormer, we enhance this traditional readout function through a special global node, referred to as [Readout]. This node is later used for contrastive learning by modifying Deep Graph Infomax (DGI) .Unlike static readout functions, this node dynamically updates its representation by aggregating information from all spots in the graph, effectively capturing the global context. This dynamic updating allows the [Readout] node to weigh contributions from different spots based on their structural encoding, avoiding the equal treatment problem of traditional readout. The [Readout] node serves as a global context aggregator, integrating information from all spots to summarize the entire graph. This node updated its representations using only the original graph data to maintains the integrity of the global context which is free from the influence of noise.

The connections between the [Readout] node and each spot are treated as special and non-physical connections. The shortest path from each node to the [Readout] node is set to 1, simulating a virtual connection. We assign distinct learnable scalars to these connections in the spatial encoding, allowing STFormer to better distinguish different regions of the graph. Our use of the Readout node with self-attention naturally handles graph-level aggregation and propagation without over-smoothing, a common problem when simply adding a supernode in traditional GNNs. The self-attention mechanism allows each node to attend to all others, effectively simulating a graph-level Readout function without losing the unique features of individual nodes.

As the Readout node connects to all other nodes (spots) in the graph. Its representation at layer $L + 1$ is updated as follows:

$$h_{[\text{Readout}]}^{(L+1)} = \sum_{j \in N([\text{Readout}])} A_{[\text{Readout}]j} V_j$$

where $N([\text{Readout}])$ denotes the set of all spots connected to [Readout]. The attention weights $A_{[\text{Readout}]j}$ are computed using the self-attention mechanism, incorporating both feature similarity and spatial encoding from the original graph. The [Readout] node can essentially represent an average readout. By setting $W_Q = W_K = 0$, the bias terms of Q and K to $T1$, and W_V to the identity matrix, where T is much larger than the scale of b_ϕ

5.1 Spatial Domain Identification via contrastive learning

We utilize a self-supervised learning strategy inspired by Deep Graph Infomax (Veličković et al., 2018) to maximize mutual information between local (node-level)

representations and the global context of the graph, adapted for spatial transcriptomics. Unlike DGI, which relies on a separate readout function for aggregating global information, our model directly utilizes node-level representations from both the original and corrupted graphs.

Representations of nodes for the original graph, $\tilde{h}_i$, and for the corrupted graph, $\tilde{h}_{i'}$, are derived from the last layer of our transformer-based model, STFormer. This adaptation is critical for capturing local spatial contexts effectively within the graph structure.

To distinguish between positive and negative examples for contrastive learning, we implement two discriminators, each configured as a Multi-Layer Perceptron (MLP). The first discriminator, $D_p^{\psi} : \mathbb{R}^d \rightarrow \mathbb{R}^d$, evaluates the similarity between node representations $\tilde{h}_i$ and their local context g from the original graph, forming positive pairs. The second discriminator, $D_n^{\psi} : \mathbb{R}^d \rightarrow \mathbb{R}^d$, assesses the similarity between the corrupted node representations $\tilde{h}_{i'}$ and their corresponding local contexts g', forming negative pairs.

Our loss aims to maximize the mutual information between positive pairs and minimize it for negative pairs as follows:

$$L_{\text{CL}}^{\text{original}} = \frac{1}{2N}\left(\sum_{i=1}^{N} \log D_p^{\psi}(\tilde{h}_i, g) + \sum_{i=1}^{N} \log(1 - D_n^{\psi}(\tilde{h}_{i'}, g'))\right) \tag{5.5}$$

Symmetric Contrastive Loss: To further enhance model stability and balance, we adopt a category of contrastive loss called symmetric loss $L_{\text{CL}}^{\text{corrupt}}$ for the corrupted graph. This loss is defined as follows:

$$L_{\text{CL}}^{\text{corrupt}} = \frac{1}{2N}\left(\sum_{i=1}^{N} \log D_p^{\psi}(\tilde{h}_{i'}, g') + \sum_{i=1}^{N} \log(1 - D_n^{\psi}(\tilde{h}_i, g))\right) \tag{5.6}$$

Both $L_{\text{CL}}^{\text{original}}$ and $L_{\text{CL}}^{\text{corrupt}}$ ensure that the representations are consistent and discriminative for both the original and corrupted graphs. This approach not only prevents overfitting but also enhances the generalizability of the learned representations across different spatial structures and configurations.

By implementing both the original and symmetric contrastive losses, our model aims to ensure that the spatial representations it learns are robust, discriminative, and capable of capturing essential spatial relationships within the data. This dual-loss strategy optimizes the network's performance by refining the similarity metric

within and across transformed graph representations, leading to improved accuracy and stability in identifying distinct spatial domains in transcriptomic data.

Reconstruction Loss

In addition to the mutual information maximization objective, we introduced a reconstruction loss to further refine the node representations. The latent representations h_i obtained from the encoder are fed into a multi-layer decoder to reconstruct the raw gene expression profiles. The decoder is an inverse counterpart of the encoder, ensuring effective mapping back to the original feature space.

To account for differences in node accessibility, we introduce a normalization term Z_i, defined as the sum of the node features that are accessible by each spot. The decoder is defined as follows:

$$h_i^{(0)} = \frac{h_i}{\sum_{j \in R(i)} g_j^{SP}} \tag{5.7}$$

where $R(i)$ represents the set of nodes accessible by spot i and g_j^{SP} denotes the spatial features of node j. We divide the initial node representations h_i obtained from the encoder by this normalization term to ensure uniform scaling across different nodes.

$$h_i^{(t+1)} = \sigma\left(h_i^{(t)} W_d^{(t)} + b_d^{(t)}\right) \tag{5.8}$$

where $h_i^{(t+1)}$ is the reconstructed gene expression profiles at the $(t+1)$-th layer, and $h_i^{(0)}$ is the squashed output representation from the encoder. $W_d^{(t)}$ is a trainable matrix and $b_d^{(t)}$ is bias vector at layer t, respectively, which are shared by all nodes in the graph. The final reconstructed output $\hat{x}_i$ is obtained after passing through all the layers of the decoder.

We also minimize the reconstruction loss of expression profiles as follow:

$$L_{\text{rec}}^{\text{original}} = \sum_{i=1}^{N} \|x_i - \hat{x}_i\|_F^2 \tag{5.9}$$

where x_i is the input and $\hat{x}_i$ is the reconstructed gene expression profiles for spot i, respectively, and $\|\cdot\|_F^2$ denotes the Frobenius norm. The reconstruction loss ensures that the learned representations retain the essential information required to accurately reconstruct the original gene expression profiles, thereby enhancing the overall quality and interpretability of the embeddings.

Multitask Loss Function

To ensure robust learning, we train the model using a multitask loss function that incorporates both the mutual information maximization objective and the reconstruction loss.

$$L_{\text{total}} = \alpha L_{\text{CL}}^{\text{original}} + \beta L_{\text{CL}}^{\text{corrupt}} + \gamma L_{\text{rec}}^{\text{original}} \tag{5.10}$$

where $L_{\text{CL}}^{\text{original}}$ is the contrastive learning loss for the original graph, $L_{\text{CL}}^{\text{corrupt}}$ is the symmetric loss for the corrupted graph, $L_{\text{rec}}^{\text{original}}$ is the reconstruction loss, and α, β, and γ are hyperparameters that balance the contribution of each term. This multitask objective ensures that the learned representations are both informative for mutual information maximization and accurate in reconstructing the original gene expression profiles.

5.2 Deconvolution at cell-type resolution

We begin by training an auto-encoder to learn latent representations from scRNA-seq data. Given the spatial transcriptomic data of spots × genes and single cell × genes, we identify the common genes between the two datasets. We consider a scRNA-seq dataset $D = \{(\mathbf{x}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^n$ represents the n-dimensional gene expressions per cell. Given a gene expression sample $\mathbf{x}_i \in \mathbb{R}^n$, the encoder $E_{\phi_g} : \mathbb{R}^n \to \mathbb{R}^{2d}$ is a Multilayer Perceptron (MLP) that learns the $\mathbb{R}^d$ latent representation. The decoder $D_g : \mathbb{R}^d \to \mathbb{R}^n$ is an MLP that reconstructs the gene expression profiles.

For the corresponding spatial transcriptomics dataset $D = \{(\mathbf{s}_i)\}_{i=1}^{N}$, where $\mathbf{s}_i \in \mathbb{R}^n$ represents the gene expressions for spot i, we use the overlapping genes between the scRNA-seq and spatial transcriptomics datasets to ensure consistency in the analysis. We use the STFormer encoder E_s to learn the latent representations $\mathbf{Z}_s$. We then combine $\mathbf{Z}_c$ and $\mathbf{Z}_h$ to obtain $\mathbf{Z}_{sc}$, which represents the cell type abundance of each ST spot using a separate network as a projection matrix $M \in \mathbb{R}^{N_{\text{cell}} \times N_{\text{spot}}}$. Each element M_{ij} in the matrix M represents the probability that cell i maps to spot j. To ensure these probabilities are meaningful, we enforce a constraint that the sum of the probabilities for all cells corresponding to any given spot must equal 1. Specifically, for each spot j, the sum of M_{ij} over all cells i is equal to 1, i.e., $\sum_{i=1}^{N_{\text{cell}}} M_{ij} = 1$.This constraint ensures that each spot has a normalized distribution of cell types, making the cell-type compositions for each spot directly comparable.

$$L_{\text{map}} = \alpha \sum_{i=1}^{N_{\text{spot}}} \sum_{j \in N_i} \log \frac{\exp(\text{sim}(\mathbf{h}_i^0, \mathbf{h}_j)/\tau)}{\sum_{p \neq i} \exp(\text{sim}(\mathbf{h}_i^0, \mathbf{h}_p)/\tau)} + \beta \|\mathbf{H}_s - \mathbf{H}_s^0\|^2$$

Here *sim* measures the similarity, τ is temperature, and α and β are weighting coefficients.

Chapter 6

EXPERIMENTS

To evaluate the performance of contraST on the downstream task of spatial domain identification, we used the human dorsolateral prefrontal cortex (DLPFC) (Genomics, n.d.) dataset which contains 12 tissue slices prepared with the 10x Visium platform. The dataset contains spatial gene expression data in which each slice contains 3,460 to 4,790 spots and expression profiles for 33000 genes. The tissue slices are annotated into regions representing the cortical layers of the DLPFC and the white matter. This data provides a snapshot of gene expression across layers involved in cognitive and executive functions for the study of molecular differences within the cortical structure. Spots in DLPFC are spatially organized to maintain their relative positions in the tissue which preserves the architecture needed for accurate spatial clustering.

We also evaluated contraST on a mouse brain 10x Visium dataset (Genomics, n.d.), in which the tissue sections were split into anterior and posterior slices. These slices were then "stitched" together to test how well contraST could align adjacent tissue sections horizontally. Using the Allen Mouse Brain Atlas for reference, contraST showed improved resolution of fine-grained structures (e.g., distinct hippocampal regions) compared with existing methods.

In addition, we tested contraST on a human lymph node 10x Visium (Genomics, n.d.) dataset containing germinal centers. The tissue includes mixed B-cell subtypes (e.g., cycling B cells, dark zone and light zone germinal center B cells) in tightly organized regions. By mapping single-cell RNA-seq data onto the spatial spots, contraST demonstrated strong localization of these B-cell subtypes, aligning well with known germinal center annotations and offering more sharply defined boundaries than alternative methods.

Data preprocessing

For spatial clustering, we began by processing spatial transcriptomics datasets with both gene expression counts and spatial coordinates. Raw gene expression counts were log-transformed to stabilize variability and normalized by library size to account for differences in sequencing depth. We then standardized the data by cen-

tering it to zero mean and scaling it to unit variance. From this processed data, we identified the top 3000 HVGs, which were used as input features for clustering.

For cell-type composition analysis, scRNA-seq data was preprocessed similarly. Raw counts were log-transformed, normalized for library size, and standardized to ensure consistency. Highly variable genes were identified, and the top 3000 HVGs were selected to align with the processed features from spatial transcriptomics data. To integrate spatial transcriptomics and scRNA-seq datasets, we matched the HVGs common to both datasets. This shared feature set served as input for the modeling framework which allow consistent representation of spatial spots and single cells.

Baseline Methods

In order to examine the performance of contraST, we performed a series of evaluations against three established baselines i.e. spaGCN, BayesSpace, and conST. Each of these methods represents a different approach to handling spatial gene expression data, and together they provide a well-rounded framework for comparison.The hyperparameters for each method was selected from the code and tutorials provided by the authors.

spaGCN applies graph convolutional operations to spatial location information, gene expressions and histological images for Spatial domain identification. Along with learning spatial relationships among spots or cells in a tissue, spaGCN supports tissue integration. Similar to contraST, SpaGCN refines cluster boundaries through learned adjacency patterns by treating each spot as a node in a graph.

BayesSpace applies a Bayesian model combined with a Markov random field to analyze spatial transcriptomics data. It integrates spatial information with gene expression data to identify spatially distinct clusters. The Bayesian model is used to estimate probabilistic distributions of gene expression for each spot. Whereas, The Markov random field imposes spatial smoothness by modeling the relationships and dependencies between neighboring spots. Through this approach, BayesSpace detects region-specific variations and latent structures in SRT datasets.

conST combines neighborhood information, gene expression, and spatial morphology for spatial domain identification. Similar to contraST, this method uses a multimodal contrastive learning framework between similar and opposite pairs of spots. Spot-level features, such as gene expression profiles, are combined with spatial patterns by constructing a neighborhood graph that defines the spatial relationships between spots. The graph calculates adjacency based interactions to learn

spatial proximity and enforce local structure.

6.1 Spatial domain identification

Figure 1 shows the comparison of contraST and baseline methods on slice 151673 of the LIBD DLPFC dataset for the task of spatial domain identification.

Receiver Operating Characteristic for Spatially Variable Genes. We analyzed the ability of contraST to identify spatially variable genes (SVGs), which are key for defining spatial domains. The ROC curve was used to measure the true positive rate against the false positive rate for each method. contraST consistently performed better than the other methods, showing higher true positive rates at all thresholds. This result show the effectiveness of encoding structural and graph information inside transformers as it leads to distinct genes representations. In contrast, methods such as BayesSpace, SpaGCN, and conST showed lower performance which hints towards possible oversmoothing resulting in a limited ability to distinguish genes contributing to spatial structure.

cell type convolution(Figure 6.1B) quantifies the accuracy of predicted spatial domain boundaries by measuring the spatial deviation from the ground truth annotation. Smaller values indicate more accurate boundary detection. We evaluated this metric across all 12 DLPFC slices, where contraST achieved the smallest mean shifting distance compared to other methods. This demonstrates its ability to capture precise boundaries between cortical layers and white matter.

We further evaluated the spatial clustering accuracy using the Adjusted Rand Index (ARI), a widely used metric for comparing predicted clusters against annotated regions. contraST achieved the highest ARI score of 0.67, outperforming BayesSpace (ARI = 0.56), SpaGCN (ARI = 0.59), and conST (ARI = 0.42). Visual inspection of the clustering results revealed that contraST produced clusters that closely aligned with the ground truth annotations, maintaining the correct layer thickness and sharp boundaries between layers. Importantly, contraST was the only method capable of accurately delineating Layer 6 from white matter, a boundary that competing methods struggled to resolve.

BayesSpace and SpaGCN partially captured the cortical layers but failed to maintain consistent layer thickness and produced mixed clusters in certain regions. conST performed poorly, with layers incorrectly clustered and significant mixing of white matter with cortical regions. These results highlight the importance of incorporating graph-based structural information to effectively model spatial relationships and

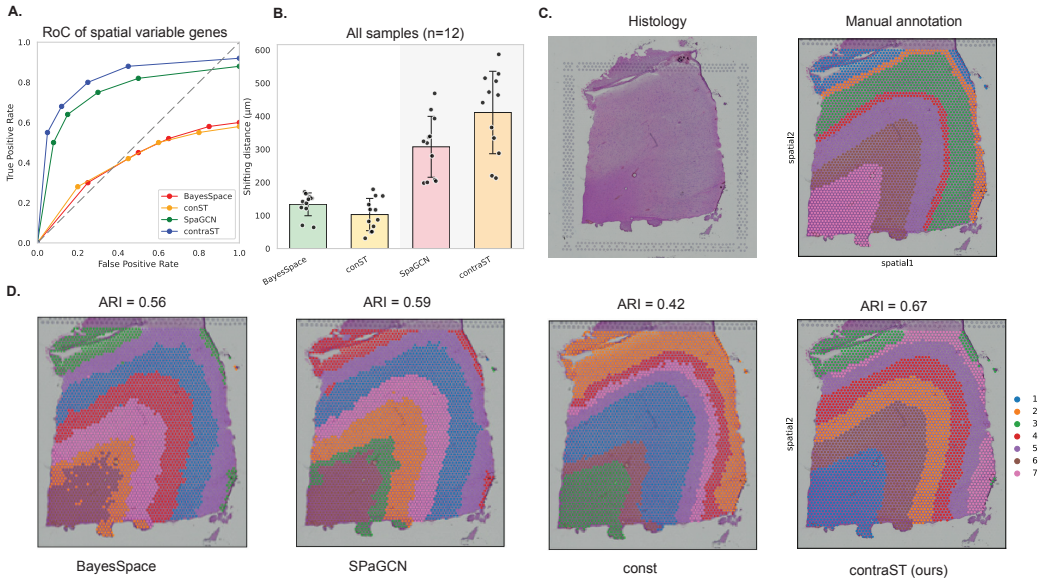

Figure 6.1: Comparison of contraST and baseline methods on slice 151673 of the LIBD DLPFC dataset for spatial domain identification. (A) Receiver Operating Characteristic (ROC) curve for spatially variable genes (SVGs) shows contraST outperforms baseline methods, demonstrating its ability to encode structural and graph information for improved SVG detection. (B) Shifting distance metric across all samples (n=12) quantifies the accuracy of predicted spatial domain boundaries, with contraST achieving the smallest mean deviation. (C) Histology image and manual annotations of spatial domains for reference. (D) Spatial domain identification results for BayesSpace, SPaGCN, conST, and contraST, with contraST achieving the highest Adjusted Rand Index (ARI = 0.67).

prevent the oversmoothing effects observed in other methods.

The combination of ROC analysis, shifting distance metrics, and ARI scores demonstrates the effectiveness of contraST in capturing spatial transcriptomic patterns with high accuracy. By integrating the graph structure into the transformer architecture, contraST addresses key limitations of existing methods, including oversmoothing and inaccurate boundary delineation. These results emphasize its suitability for spatial clustering tasks in high-resolution spatial transcriptomic datasets.

6.2 Multisample Integration

One of the major tasks in spatial transcriptomics (ST) is integrating samples that come from different experimental conditions or that exceed the capacity of a single ST capture slide. A key challenge is the batch effect, which can introduce bias into the analyses. Consequently, it is important to remove batch effects and align similar

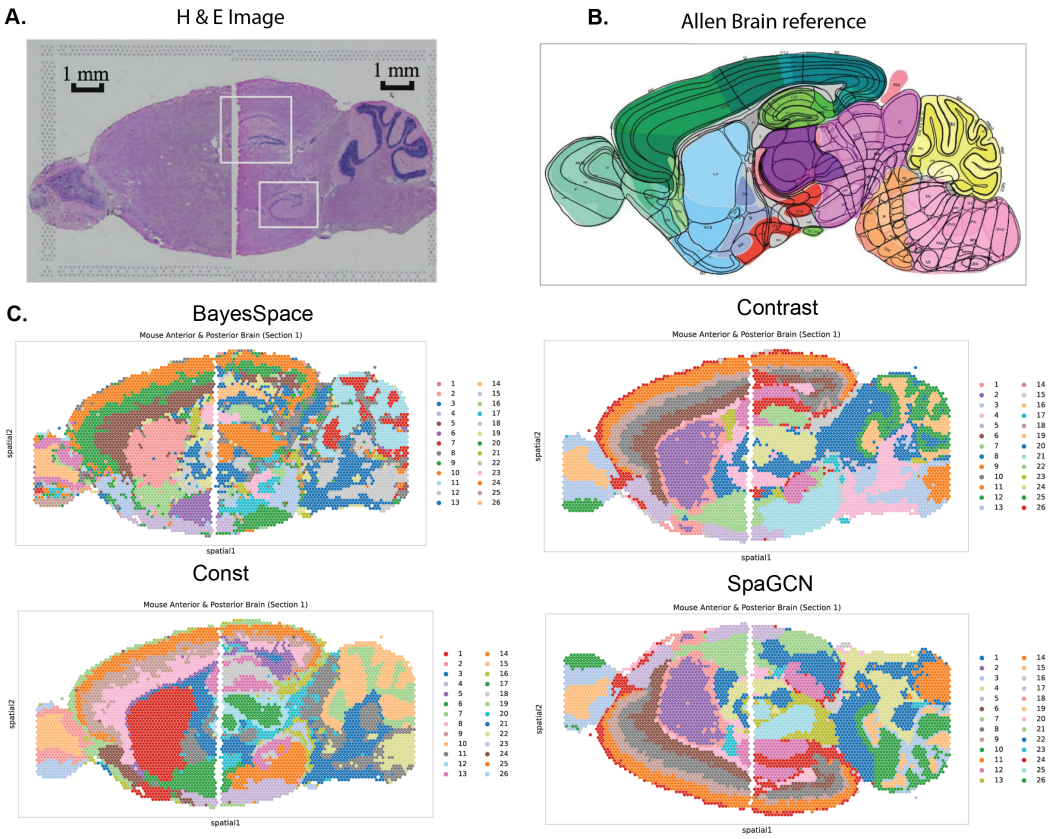

Figure 6.2: Sample integration and spatial domain identification in two mouse brain tissue sections (anterior and posterior) from 10x Visium. (A) image of the tissue slices. (B) Allen Mouse Brain Atlas reference regions. (C) Comparison of spatial domain identification results across BayesSpace, SpaGCN, conST, and contraST. ContraST outperforms other methods by accurately aligning slices and capturing fine-grained tissue boundaries, particularly in the cerebral cortex and corpus callosum.

tissue regions across slices.

Here, we evaluated the performance of contraST on two mouse brain tissue sections from 10x Visium, each split into anterior and posterior slices. We aligned the tissue samples using the PASTE algorithm and performed spatial domain identificatoin using stformer. We then compared contraST against SpaGCN, BayesSpace and Const.

Using the Allen Mouse Brain Atlas as a reference to examine how each method's clusters align with known anatomical regions, we found that baseline methods produced more fragmented clusters especially where the two slices meet and was unable to accurately capture fine-grained tissue boundaries. on the other hand,

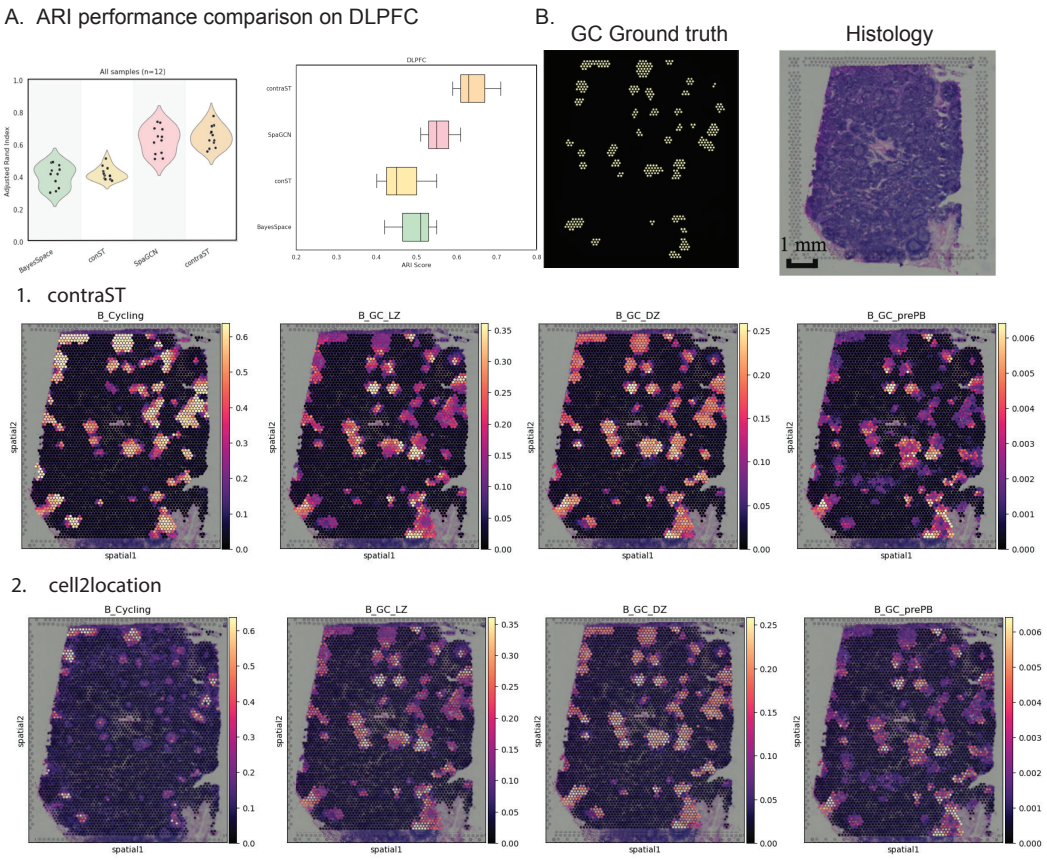

Figure 6.3: Comparison of contraST and cell2location on cell type deconvolution using the Human lymph node ST dataset. (A) Adjusted Rand Index (ARI) performance comparison on DLPFC samples, highlighting contraST's superior clustering accuracy. (B) Ground truth annotation and histological reference of germinal center (GC) regions. (C) Spatial probability maps for B-cell subtypes (cycling B cells, dark zone germinal center B cells, and pre-plasma B cells). ContraST produces sharper, more localized signals that align with ground truth, whereas cell2location results are more diffuse, demonstrating contraST's higher accuracy in detecting fine-grained spatial patterns.

contraST performed exceptionally well at learning subtle tissue boundaries, as seen particularly in the cerebral cortex and corpus callosum areas.

6.3 cell type deconvolution

Cell type Deconvolution is the process of identifying the types of cells and their proportions in each spot of spatial transcriptomics sample. As in ST, each spot can contain multiple cells, this means that gene expression profile is actually an average of multiple cells inside it. To learn this mapping, we use scRNA-seq data where we

already know the cell types and their expression profiles. To benchmark contraST on cell type deconvolution, we use Human lymph node ST dataset from Visium. On this dataset, the scRNA-seq reference data consists of 34 cell types from multiple human secondary lymphoid organ studies.

We use this reference dataset to learn cell-to-spot probability matrix which learns how likely each cell type is in each spot. In this figure, we compare the performance of contraST with cell2location focusing on B-cell types e.g. cycling B cells, dark zone germinal center B cells and naive B cells. for both methods, we show the spatial probability for each subtype. The results show that contraST is often able to provide stronger localized signals (bright patches) which align with the ground truth. Cell2location, is able to identify the cells but appear diffuse in certain areas. This shows that contraST can detect B-cell subtypes with higher confidence as compared to cell2location.

Chapter 7

DISCUSSION

As shown in this dissertation, Spatial transcriptomics is a subtle technique that allows us to sequence gene expressions in the presence of spatial context. Despite its potential, ST encounters technological limitations that computational approaches like ContraST can effectively address. The constrained size of ST capture areas necessitates the use of multiple adjacent tissue slides to encapsulate a more comprehensive spatial distribution of cell types. ContraST's utility was demonstrated through its capability to deconvolute complex environments within breast cancer tissues, analyzing both cancer cells and the immune milieu.

The distinguishing features of ContraST that contribute to its superior performance include the use of a transformer-based encoder, the STFormer, which integrates spatial and structural context through its self-attention mechanism. This mechanism enhances the capture and reinforcement of spatial relationships and structural information, leading to more informative and discriminative spot representations. While discussed methods like SpaGCN and STAGATE also utilize graph neural networks to integrate gene expression with spatial data, they often fall short in clustering performance and lack robust batch integration capabilities. ContraST, on the other hand, employs self-supervised contrastive learning to enhance latent representation learning and preserve local spatial context effectively.

ContraST is designed to be user-friendly and versatile across different experimental platforms. It has been validated on technologies such as 10x Visium, Slide-seqV2, and Stereo-seq, and is planned to be extended to other platforms like MERFISH and Nanostring CosMx SMI. Engineered to be computationally efficient, ContraST is capable of processing large datasets rapidly—a testament to this is its performance on the largest dataset tested, involving about 100,000 spots from an E14.5 mouse embryo, which was processed in just 30 minutes on a high-specification computing setup. Looking forward, we anticipate that future advancements in ST technologies will provide even greater resolution and data complexity, for which ContraST is well-prepared.

Finally, the integration capabilities of ContraST, both horizontally across tissue samples and vertically through serial tissue slides, enable it to identify biologically

coherent spatial domains effectively and accurately, aligning them across samples while mitigating batch effects. This robust integration underscores ContraST's potential to transform the landscape of spatial transcriptomics, making it a pivotal tool in the exploration of complex tissue structures and their underlying biological functions.

Future work will aim to extend ContraST's application to newer technologies and integrate additional modalities, such as spatial proteomics and histological imaging. Enhancing its capabilities for real-time processing and longitudinal studies will enable ContraST to capture dynamic tissue changes over time. By validating its utility in clinical settings, ContraST could facilitate tasks including identification of tumor microenvironments, immune cell interactions, and biomarkers, supporting personalized medicine. These efforts will further position ContraST as a important tool in advancing spatial transcriptomics research and its application in understanding tissue organization and disease mechanisms.

(Domcke et al., 2020; Han et al., 2020; Sikkema et al., 2023; Srivatsan et al., 2020; Lotfollahi, Susmelj, et al., 2021; Lotfollahi, Klimovskaia Susmelj, et al., 2023; Hetzel et al., 2022; Gehring, Hwee Park, et al., 2020; Alghamdi et al., 2021) Z.-J. Cao and Gao, 2022; Wu et al., 2022; Sachs et al., 2020; Al-Lazikani, Banerji, and Workman, 2012; Datlinger et al., 2017; Norman et al., 2019; Schmidt et al., 2022; Hagai et al., 2018; Sabour, Frosst, and Hinton, 2017; Luo, Kang, and Schönhuth, 2023; Mazzia, Salvetti, and Chiaberge, 2021; Choi et al., 2019; Zrimec et al., 2022; Dincer, Janizek, and S.-I. Lee, 2020; McGinnis et al., 2019; Stoeckius et al., 2018; Gehring, J. H. Park, et al., 2018; Russkikh et al., 2020; Lotfollahi, Naghipourfar, et al., 2020; Angelini and Costa, 2014; Saillard et al., 2021; Zarour, 2016; D. Huang et al., 2018; Z. Zhang et al., 2015; Adler et al., 2007; Hui Chen et al., 2023; Yaribeygi et al., 2022; Feron, 2009; Kathagen et al., 2013; Sukumar, Kishton, and Restifo, 2017; Yin et al., 2019; Vaughn and Haviland, 2021; Ruterbusch et al., 2020; L. Zhou and Littman, 2009; Hughes et al., 2000; Bonzanni et al., 2013; Eckschlager et al., 2017; Kusaczuk et al., 2016; J. King, Patel, and Chandrasekaran, 2021; Ganai, 2016

BIBLIOGRAPHY

Adler, Henric S et al. (2007). "Activation of MAP kinase p38 is critical for the cell-cycle–controlled suppressor function of regulatory T cells". In: *Blood* 109.10, pp. 4351–4359.

Alghamdi, Norah et al. (2021). "A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data". In: *Genome research* 31.10, pp. 1867–1884.

Andersson, Alma et al. (2020). "Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography". In: *Communications biology* 3.1, p. 565.

Angelini, Claudia and Valerio Costa (2014). "Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems". In: *Frontiers in cell and developmental biology* 2, p. 51.

Bonzanni, Nicola et al. (2013). "Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model". In: *Bioinformatics* 29.13, pp. i80–i88.

Cable, Dylan M et al. (2022). "Robust decomposition of cell type mixtures in spatial transcriptomics". In: *Nature biotechnology* 40.4, pp. 517–526.

Cang, Zixuan et al. (2021). "SCAN-IT: Domain segmentation of spatial transcriptomics images by graph neural network". In: *BMVC: Proceedings of the British Machine Vision Conference. British Machine Vision Conference*. Vol. 32. NIH Public Access.

Cao, Zhi-Jie and Ge Gao (2022). "Multi-omics single-cell data integration and regulatory inference with graph-linked embedding". In: *Nature Biotechnology* 40.10, pp. 1458–1466.

Chang, Yuzhou et al. (2022). "Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning". In: *Computational and Structural Biotechnology Journal* 20, pp. 4600–4617.

Chen, Ao et al. (2022). "Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays". In: *Cell* 185.10, 1777–1792.e21.

Chen, Hui et al. (2023). "The regulatory effects of second-generation antipsychotics on lipid metabolism: Potential mechanisms mediated by the gut microbiota and therapeutic implications". In: *Frontiers in Pharmacology* 14, p. 1097284.

Choi, Jaewoong et al. (2019). "Attention routing between capsules". In: *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pp. 0–0.

Datlinger, Paul et al. (2017). "Pooled CRISPR screening with single-cell transcriptome readout". In: *Nature methods* 14.3, pp. 297–301.

Del Rossi, Natalie et al. (2022). “Analyzing spatial transcriptomics data using Giotto”. In: *Current protocols* 2.4, e405.

Dincer, Ayse B, Joseph D Janizek, and Su-In Lee (2020). “Adversarial deconfounding autoencoder for learning robust gene expression embeddings”. In: *Bioinformatics* 36.Supplement_2, pp. i573–i582.

Domcke, Silvia et al. (2020). “A human cell atlas of fetal chromatin accessibility”. In: *Science* 370.6518, eaba7612.

Dries, Ruben et al. (2021). “Giotto: a toolbox for integrative analysis and visualization of spatial expression data”. In: *Genome biology* 22, pp. 1–31.

Eckschlager, Tomas et al. (2017). “Histone deacetylase inhibitors as anticancer drugs”. In: *International journal of molecular sciences* 18.7, p. 1414.

Feron, Olivier (2009). “Pyruvate into lactate and back: from the Warburg effect to symbiotic energy fuel exchange in cancer cells”. In: *Radiotherapy and oncology* 92.3, pp. 329–333.

Ganai, Shabir Ahmad (2016). “Histone deacetylase inhibitor givinostat: the small-molecule with promising activity against therapeutically challenging haematological malignancies”. In: *Journal of Chemotherapy* 28.4, pp. 247–254.

Gehring, Jase, Jong Hwee Park, et al. (2020). “Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins”. In: *Nature biotechnology* 38.1, pp. 35–38.

Gehring, Jase, Jong Hwee Park, et al. (2018). “Highly multiplexed single-cell RNA-seq for defining cell population and transcriptional spaces”. In: *BioRxiv*, p. 315333.

Genomics, 10x (n.d.). *Visium Spatial Gene Expression Solution*. `https://www.10xgenomics.com/products/spatial-gene-expression`. Accessed: 2025-01-06.

Hagai, Tzachi et al. (2018). “Gene expression variability across cells and species shapes innate immunity”. In: *Nature* 563.7730, pp. 197–202.

Han, Xiaoping et al. (2020). “Construction of a human cell landscape at single-cell level”. In: *Nature* 581.7808, pp. 303–309.

Hetzel, Leon et al. (2022). “Predicting cellular responses to novel drug perturbations at a single-cell resolution”. In: *Advances in Neural Information Processing Systems* 35, pp. 26711–26722.

Hu, Jian et al. (2021). “SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network”. In: *Nature methods* 18.11, pp. 1342–1351.

Huang, Dengliang et al. (2018). “GLI2 promotes cell proliferation and migration through transcriptional activation of ARHGEF16 in human glioma cells”. In: *Journal of Experimental & Clinical Cancer Research* 37.1, pp. 1–17.

Hughes, Timothy R et al. (2000). “Functional discovery via a compendium of expression profiles”. In: *Cell* 102.1, pp. 109–126.

Jayaraman, Sahana et al. (2023). “Barcoding intracellular reverse transcription enables high-throughput phenotype-coupled T cell receptor analyses”. In: *Cell Reports Methods* 3.10.

Kathagen, Annegret et al. (2013). “Hypoxia and oxygenation induce a metabolic switch between pentose phosphate pathway and glycolysis in glioma stem-like cells”. In: *Acta neuropathologica* 126, pp. 763–780.

King, Jacob, Maya Patel, and Sriram Chandrasekaran (2021). “Metabolism, HDACs, and HDAC inhibitors: A systems biology perspective”. In: *Metabolites* 11.11, p. 792.

Kleshchevnikov, Vitalii et al. (2022). “Cell2location maps fine-grained cell types in spatial transcriptomics”. In: *Nature biotechnology* 40.5, pp. 661–671.

Korsunsky, Ilya et al. (2019). “Fast, sensitive and accurate integration of single-cell data with Harmony”. In: *Nature methods* 16.12, pp. 1289–1296.

Kusaczuk, Magdalena et al. (2016). “Molecular and cellular effects of a novel hydroxamate-based HDAC inhibitor–belinostat–in glioblastoma cell lines: a preliminary report”. In: *Investigational New Drugs* 34, pp. 552–564.

Al-Lazikani, Bissan, Udai Banerji, and Paul Workman (2012). “Combinatorial drug therapy for cancer in the post-genomic era”. In: *Nature biotechnology* 30.7, pp. 679–692.

Long, Yahui et al. (2023). “Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST”. In: *Nature Communications* 14.1, p. 1155.

Longo, Sophia K et al. (2021). “Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics”. In: *Nature Reviews Genetics* 22.10, pp. 627–644.

Lopez, Romain et al. (2018). “Deep generative modeling for single-cell transcriptomics”. In: *Nature methods* 15.12, pp. 1053–1058.

Lotfollahi, Mohammad, Anna Klimovskaia Susmelj, et al. (2023). *Predicting cellular responses to complex perturbations in high-throughput screens*.

Lotfollahi, Mohammad, Mohsen Naghipourfar, et al. (2020). “Conditional out-of-distribution generation for unpaired data using transfer VAE”. In: *Bioinformatics* 36.Supplement_2, pp. i610–i617.

Lotfollahi, Mohammad, Anna Klimovskaia Susmelj, et al. (2021). *Compositional perturbation autoencoder for single-cell response modeling*. BioRxiv. Cold Spring Harbor Laboratory.

Luo, Xiao, Xiongbin Kang, and Alexander Schönhuth (2023). "Predicting the prevalence of complex genetic diseases from individual genotype profiles using capsule networks". In: *Nature Machine Intelligence* 5.2, pp. 114–125.

Mazzia, Vittorio, Francesco Salvetti, and Marcello Chiaberge (2021). "Efficient-capsnet: Capsule network with self-attention routing". In: *Scientific reports* 11.1, p. 14634.

McGinnis, Christopher S et al. (2019). "MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices". In: *Nature methods* 16.7, pp. 619–626.

Norman, Thomas M et al. (2019). "Exploring genetic interaction manifolds constructed from rich single-cell phenotypes". In: *Science* 365.6455, pp. 786–793.

Papouchado, Bettina G et al. (2010). "Silver in situ hybridization (SISH) for determination of HER2 gene status in breast carcinoma: comparison with FISH and assessment of interobserver reproducibility". In: *The American journal of surgical pathology* 34.6, pp. 767–776.

Pham, Duy et al. (2020). "stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues". In: *BioRxiv*, pp. 2020–05.

Ren, Honglei et al. (2022). "Identifying multicellular spatiotemporal organization of cells with SpaceFlow". In: *Nature communications* 13.1, p. 4076.

Rodriques, Samuel G et al. (2019). "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution". In: *Science* 363.6434, pp. 1463–1467.

Russkikh, Nikolai et al. (2020). "Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis". In: *Bioinformatics* 36.20, pp. 5076–5085.

Ruterbusch, Mikel et al. (2020). "In vivo CD4+ T cell differentiation and function: revisiting the Th1/Th2 paradigm". In: *Annual review of immunology* 38, pp. 705–725.

Sabour, Sara, Nicholas Frosst, and Geoffrey E Hinton (2017). "Dynamic routing between capsules". In: *Advances in neural information processing systems* 30.

Sachs, Stephan et al. (2020). "Targeted pharmacological therapy restores β-cell function for diabetes remission". In: *Nature Metabolism* 2.2, pp. 192–209.

Saillard, Margaux et al. (2021). "Impact of immunotherapy on CD4 T cell phenotypes and function in cancer". In: *Vaccines* 9.5, p. 454.

Schmidt, Ralf et al. (2022). "CRISPR activation and interference screens decode stimulation responses in primary human T cells". In: *Science* 375.6580, eabj4008.

Shah, Sheel et al. (2018). "Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH". In: *Cell* 174.2, pp. 363–376.

Sikkema, Lisa et al. (2023). “An integrated cell atlas of the lung in health and disease”. In: *Nature Medicine*, pp. 1–15.

Srivatsan, Sanjay R et al. (2020). “Massively multiplex chemical transcriptomics at single-cell resolution”. In: *Science* 367.6473, pp. 45–51.

Stoeckius, Marlon et al. (2018). “Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics”. In: *Genome biology* 19.1, pp. 1–12.

Sukumar, Madhusudhanan, Rigel J Kishton, and Nicholas P Restifo (2017). “Metabolic reprograming of anti-tumor immunity”. In: *Current opinion in immunology* 46, pp. 14–22.

Vaswani, A (2017). “Attention is all you need”. In: *Advances in Neural Information Processing Systems*.

Vaughn, Nicole and David L Haviland (2021). “Acly promotes metabolic reprogramming and induction of IRF4 during early CD8+ T cell activation”. In: *Cytometry Part A* 99.8, pp. 825–831.

Velickovic, Petar et al. (2017). “Graph attention networks”. In: *stat* 1050.20, pp. 10–48550.

Veličković, Petar et al. (2018). “Deep graph infomax”. In: *arXiv preprint arXiv:1809.10341*.

Wei, Xiaoyu et al. (2022). “Single-cell Stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration”. In: *Science* 377.6610, eabp9444.

Wu, Yulun et al. (2022). *Predicting Cellular Responses with Variational Causal Inference and Refined Relational Information*.

Xu, Hang et al. (2024). “Unsupervised spatially embedded deep representation of spatial transcriptomics”. In: *Genome Medicine* 16.1, p. 12.

Yaribeygi, Habib et al. (2022). “Mechanistic view on the effects of SGLT2 inhibitors on lipid metabolism in diabetic milieu”. In: *Journal of Clinical Medicine* 11.21, p. 6544.

Yin, Zhongping et al. (2019). “Targeting T cell metabolism in the tumor microenvironment: an anti-cancer therapeutic strategy”. In: *Journal of Experimental & Clinical Cancer Research* 38, pp. 1–10.

Ying, Chengxuan et al. (2021). “Do transformers really perform badly for graph representation?” In: *Advances in neural information processing systems* 34, pp. 28877–28888.

Zarour, Hassane M (2016). “Reversing T-cell dysfunction and exhaustion in cancer”. In: *Clinical cancer research* 22.8, pp. 1856–1864.

Zhang, Meng et al. (2021). “Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH”. In: *Nature* 598.7879, pp. 137–143.

Zhang, Si et al. (2019). “Graph convolutional networks: a comprehensive review”. In: *Computational Social Networks* 6.1, pp. 1–23.

Zhang, Zhanguang et al. (2015). “DNAM-1 controls NK cell activation via an ITT-like motif”. In: *Journal of Experimental Medicine* 212.12, pp. 2165–2182.

Zhou, Liang and Dan R Littman (2009). “Transcriptional regulatory networks in Th17 cell differentiation”. In: *Current opinion in immunology* 21.2, pp. 146–152.

Zrimec, Jan et al. (2022). “Controlling gene expression with deep generative design of regulatory DNA”. In: *Nature communications* 13.1, p. 5099.