INTEGRATIVE LEARNING OF CELLULAR SYSTEMS AND NETWORKS

By

Julian D. Venegas

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computational Mathematics, Science, and Engineering—Doctor of Philosophy

2024

ABSTRACT

Advances in omics technologies have led to an abundance of comprehensive biomolecular information of biological systems, down to single-cell resolution. With omics data, biologists can gain a deeper understanding of the complex-hierarchical networks that constitute an organism. To this end, deep learning methods are often applied to assist in discovering meaningful patterns and relationships from omics data. Though deep learning methods can offer high performance on many complex tasks, some challenges arise with omics-based tasks: (1) Omics data are often high-dimensional with low-sample size and/or high levels of sparsity, with complex dependency structures between and within omics data types. (2) There is an imbalance of annotated data across different species and environments. These difficulties make desirable the integration of omics data across different modalities, group samples, and platforms, as well as environments and species.

This thesis examines, builds and implements approaches to address these challenges through Integrative Learning techniques, which I use as a general term to encompass techniques that incorporate multiple sources of related data for improved learning (e.g., transfer learning, multi-task learning, and multi-modal data integration). In this work I highlight and address these challenges in different omics-based tasks.

In addition to applied methods development, I provide probabilistic and mathematical frameworks that underpin many of these applied problems in omics analysis.

Lastly, I showcase some of my more current experiments in deconvolution. Though these experiments are prelimary, often through toy examples, they demonstrate some possible future directions I want to consider.

I dedicate this thesis to my father, mother, wife, and three sons.

ACKNOWLEDGEMENTS

First, I must thank my advisor and committee chair, Dr. Yuying Xie. He has been a consistent source of support since he welcomed me to his lab. I most appreciate him offering me strong guidance in research, balanced with freedom to pursue my own interests; it has been a pleasant experience. I would also like to thank my co-advisor, Dr. Shin-Han Shiu. He served a key role in applied projects, offering me insights and direction through the lens of a domain expert in biology.

I would also like to thank my other committee members, Dr. Frederi Viens and Dr. Longxiu Huang for their feedback, advice and support. All of my committee members have provided me with a great example of scholarship and professionalism, balanced with humanity.

I also want to thank the CMSE department and the College of Engineering more broadly. I was always made to feel welcome and supported by all. In particular, I'd like to thank Heather Williams for her administrative support, and Dr. Katy Colbry for guiding me towards being awarded the NSF Graduate Research Fellowship, among other funds.

Lastly, thank you to my family—my father, mother, wife, and three sons. You've supported, encouraged, and inspired me throughout my studies; I am forever grateful.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Complex cellular networks with dynamic functional states have a pivotal role in the hierarchy of networks and structures that make up an organism McManus et al. (2015). This makes the study of cells and cellular networks integral to understanding all of biology. This has motivated advances in high-throughput single-cell sequencing and imaging technologies have enabled the collection of massive biomolecular data at single-cell resolution. Indeed, single-cell sequencing technologies often generate tens of thousands to millions of samples/cells per study Svensson et al. (2018). This massive and complex data make deep learning methods an attractive approach to analyze cell behavior and cellular networks.

Deep learning methods, which have consistently shown cutting-edge performance in various big data applications Pouyanfar et al. (2018); Dong et al. (2021), have fertile new ground for research that pushes the frontiers of biological science.

The first developments in single-cell sequencing technology began with complementary DNA (cDNA) Eberwine et al. (1992); Brady et al. (1990). A major breakthrough would come decades later, through the development of single-cell RNA sequencing (scRNA-seq) methods Tang et al. (2009). scRNA-seq methods have given scientists the ability to analysis biology with single cell resolution, i.e. at the building blocks of life. This has served a pivotal role in changing how we study biology, and has lead to great advancements across many fields of biology and science more broadly. Since then, there has been further developments in next-generation sequencing platforms, with over one hundred currently existing single-cell sequencing technologies Wang and Navin (2015); Wen and Tang (2022). We now have technologies that measure a wide-array of cell features, e.g. DNA sequences and epigenetic features, methylation and chromatin accessibility, RNA expression, and profiles of surface proteins. Building on these developments, recent omics technology advances offer multi-feature capability, and additional ancillary features such as spatial location via spatial transcriptomics.

These advances in sequencing technologies have facilitated great advances on the bioinformatics front, through high throughput methods Svensson et al. (2018); Kolodziejczyk et al. (2015) which provide (1) higher resolution from bulk tissues to individual cells (and sub-cellular levels) and (2) a massive volume of open data. To illustrate this, consider bulk tissue RNAseq (bulk RNA-seq) Li and Wang (2021) vs scRNA-seq data. With bulk RNA-seq we quantify expression as an aggregate from a group of cells (cell pool). Here, cellular heterogeneity is lost, and we have only a single expression sample comprised of multiple cells. With single-cell sequencing technologies we quantify expression for each individual cell. So if we consider a cell pool consisting of 20 cells, we get a single sample from bulk RNA-seq and 20 samples from scRNA-seq. In real sequencing experiments, this small delta becomes a great delta, where a single experiment with scRNA-seq technologies can generate tens of thousands up to millions of samples. These high-resolution and high-throughput omics technologies have provided a massive amount of rich biological data to analyse and explore.

Naturally, as with data explosions across other scientific domains, this has fostered advancements in data analysis and machine learning. In particular, deep learning methods have come to the forefront in many big data applications Pouyanfar et al. (2018); Dong et al. (2021), and omics data analysis is no exception. Deep learning applications in omics have provided biologists with powerful in silico tools to supplement and inform their in vitro and in vivo experiments. Conversely, domain knowledge gained from in vitro and in vivo experiments can inform deep learning models through *integrative learning* methods. This feedback loop between domain experts and advanced in silico learning methods will serve as a vehicle we drive to new frontiers in our understanding of biology, and thereby, physical life.

Despite the success of single-cell data in numerous applications, difficulties arise due to the complexity of the data which requires advanced analysis pipelines with a number of steps. Single-cell data preprocessing includes many stages of data pruning, normalization, and often challenging machine learning tasks like batch effect correction, data imputation,

or dimensionality reduction. Moreover, specialized types of single-cell data require further processing such as multimodal data integration and cell type deconvolution for spatial transcriptomics. These steps are crucial to facilitate downstream tasks ranging from clustering and cell annotation, disease prediction, identifying gene coexpression networks, to the identification of developmental trajectories of cells transitioning between states Lähnemann and et al. (2020).

For tasks with clear evaluation metrics, deep learning often achieves top performance against other classical machine learning techniques Muzio et al. (2021). Deep learning can uniquely leverage its diverse architectures to capture networks of interdependencies between genes that alter other genes' expression levels Bansal et al. (2007), and cells that communicate with other cells through mechanisms like ligand-receptor pairs Li et al. (2022b). Due to the richness of deep learning architectures and the customization of hyper-parameters and loss functions, deep learning models can be more readily tailored to particular tasks in single-cell analysis compared to shallow-classical machine learning methods. Deep learning has already become a staple in omics analysis, but there is still fertile ground in deep learning applications to problems in omics data analysis.

In this work, I review and showcase some major problems in omics data analysis, with applications of deep learning on omics data. In the chapter 1, I give background of omics data and technologies, focusing on single-cell, bulk-seq and spatial transcriptomics data. Further, I put scRNA-seq data into a probabilistic framework that underpins many methods in scRNA-seq analysis. In Chapter 3 I apply deep learning frameworks and methods to predicting plant stress response from DNA sequences. There, I emphasize the utility of integrative learning through transfer learning and multi-task learning. In Chapter 4 I develop and apply deep graph learning methods for cell type deconvolution, provide benchmarking experiments, and a method for generating large-scale cell type deconvolution benchmarking datasets, with ground truth labels. In chapter 5 I showcase a few more (early) works in deconvolution, emphasizing probabilistic frameworks for deconvolution. Lastly, in chapter 6 I give some

concluding remarks about experimental results, developments, and future directions.

CHAPTER 2

BACKGROUND

We begin this chapter with an overview of omics data, with an emphasis on single-cell data and spatial transcriptomics data. Here, we will discuss developments in omics technologies, as well as structure and select problems in omics data analysis. We then develop a probabilistic framework for scRNA-seq count data, which serves as an underpinning assumption in many analytical methods for tasks involving scRNA-seq data.

The first section - Omics Data - is in part derived from my contributions to the survey paper Molho et al. (2024), and new developments paper Ding et al. (2024b).

2.1 Omics Data

At a high-level, we study biology to understand life through internal (unseen or micro-scale) dynamics, Further, to understand the levers that control those dynamics. To understand the dynamics and controls of any system, it is most helpful to create causal maps, i.e., mapping cause and effect. To do this, we need output or outcomes that we can observe through our senses. Of course, dynamic systems are complex, with controls defined by non-linear interactions of components across both horizontal and vertical scales. To understand the dynamics of life, then, it helps to understand biological systems across varying levels. This allows us to better understand causal maps controlling biological systems.

In biological systems, one of the most tangible levels we can observe outcomes is with phenotypes. On the other hand, genotypes are the microscopic basic building blocks of organisms. This makes the task of understanding causal maps from genotypes to phenotypes essential to understanding biological systems. Further, any map from genome to phenome begins with the transcriptome, making trancriptome analysis an essential goal for biologists Houle et al. (2010).

While the genotypes of cells within an organism are nearly identical, only a small subset of the total gene pool is expressed at any given moment in time. That is, cellular networks have dynamic functional states, which lead to variations in RNA transcribed from DNA

across cells, i.e., transcriptomic variations. A major control of these dynamics are gene regulatory networks, as they regulate gene activity. In short, the transcriptome is largely defined by gene regulatory networks. Moreover, variations in the transcriptome highlight and reinforce the notion of cellular heterogeneity. Thus, being able to capture and analyze omics data of individual cells can help us account for cellular heterogeneity, which in turn help us understand gene regulatory networks and other major drivers of biological systems. Most gladly, there has been significant developments in single-cell technologies that provide researchers with ever increasing information at the cellular level, including transcriptomics, genomics and epigenomics data.

While bulk sampling methods can access and take transcriptomics measurements of cell pools within a tissue, they lack the capability to capture the heterogeneity and stochasticity of the cells that make up the bulk sample. Further, even with a homogeneous bulk mixture, we lose granularity in the aggregate signal measured from the mixture. On the other hand, single-cell technologies measure signals from individual cells, thus giving more granularity to study cell heterogeneity. This provides a way to isolate individual cells and their influence on upstream biological functions Goldman et al. (2019); Kulkarni et al. (2019); Stegle et al. (2015); Nguyen et al. (2018). In this section we review developments of single-cell and other omics technologies over time, with a focus on transcriptomics technologies. We summarize these technological advances chronologically in Figure 2.1.

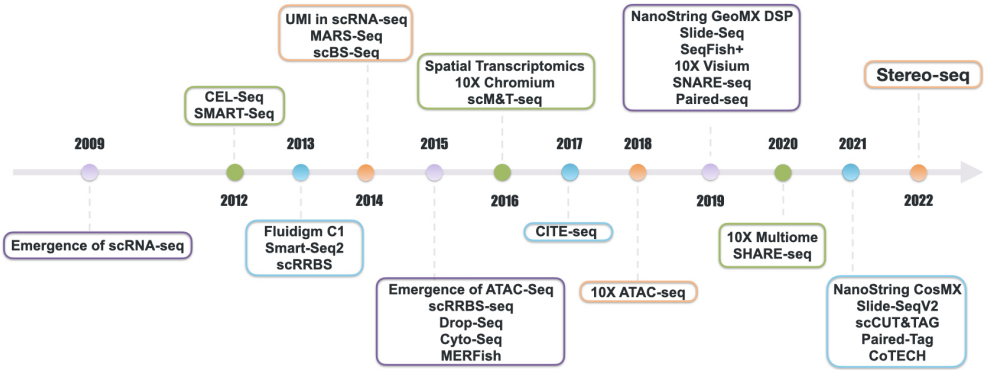


Figure 2.1 Timeline of major developments in single-cell technologies.

2.1.1 Single-cell Technologies

The first developments in sequencing technology can be traced back to James Eberwine et al. Eberwine et al. (1992) and Iscove et al.Brady et al. (1990), who first expanded complementary DNAs (cDNAs). Since then, modest yet consistent advancements culminated in a significant breakthrough in 2009, when single-cell RNA sequencing (scRNA-seq) was created Tang et al. (2009). The ideas underpinning scRNA-seq have since provided the way for various single-cell technologies for a broader range of target measures within a cell. Some examples include technologies that target DNA methylation Guo et al. (2013) (2013), protein and DNA accessibility (2015), and histone modifications Bartosovic et al. (2021) (2021). Beyond sequencing technologies, scRNA-seq has facilitated advancements in quantitative methods and understanding. With scRNA-seq data, researchers have made strides in essential tasks for deeper biological understanding, such as cell type segmentation, classification, and cell type expression profiling, to name a few.

Structurally, single-cell omics data is put into matrix form, with measured signals (e.g., RNA transcripts) as columns and cells as rows. Some example features could be some measures from accessible DNA regions in ATAC-seq data, genes in scRNA-seq data, and proteins in CITE-seq data. Here in Figure 2.2, we show a simple example of a single-cell omics data matrix.

For scRNA-seq, isolation of the cell is the first step for obtaining transcriptome information. Naturally, scRNA-seq and other single-cell technologies depend first on isolating individual cells. This process often distinguishes the different single-cell technologies, i.e., how they perform cell isolation. The earliest technologies tended to be low throughput and achieved cell isolation through serial dilution or robotic micromanipulation Brehm-Stecher and Johnson (1990). More recently, technologies that use microfluidic methods to isolate the cell have provided higher throughput capabilities Whitesides (2006). A key microfluidic based cell isolation method uses microdroplets Thorsen et al. (2001). Here, water droplets are uniformely disperesed in a medium of oil, which allows cells to isolate into the droplets.

While commercial microfluidic platforms like Fluidigm C1, ICELL8, and Chromium can benefit from high throughput, they face the challenge of high cost and often the requirement of uniform cell size in the sample. Once a cell is separated and lysed, messenger RNAs in this cell are reverse transcripted into more stable cDNAs with a unique cell barcode. The cDNAs are then amplified via Polymerase Chain Reaction (PCR) for better data capture before sequencing, which tends to introduce bias due to the uneven amplification efficiency. Therefore, besides the unique barcodes, the cDNA molecules in a cell are also given a Unique Molecular Identifier (UMI) to correct the amplification bias by collapsing the reads with the same UMI into one read. After debiasing, sequence reads are mapped to the genome and are grouped into genes for the creation of a count matrix Wang and Navin (2015).

In addition to measuring RNA transcripts, some single-cell technologies can also measure chromatin accessibility of a cell's chromosome. Eukaryotic genomes are hierarchically packaged into chromatin Kornberg (1974), and this packaging plays a central role in gene regulation Kornberg and Lorch (1974). Buenrostro et al. created a means for sampling the epigenome at the single-cell level through the Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq) Buenrostro et al. (2013) in 2013 . ATAC-seq allows the identification of accessible DNA, i.e. the nucleosome-free regions of the genome Hendrickson et al. (2018). DNA accessibility within the genome can be used to identify regulatory elements in different cell types which cause the activation or repression of gene expression Thurman et al. (2012). scATAC-seq produces a count matrix with a number of reads per open chromatin regions, which lead to very large matrices with hundreds of thousands of regions. Furthermore the data is known to be very sparse, where it is common to have the non-zero entries make up less than 3% of the data Li et al. (2021).

While single-cell sequencing techniques for transcriptome measurements have seen great growth, single-cell proteomics methods have been developing at a slower pace. This makes a gap in omics data analysis, since proteomic data are essential to understand how genes respond to environmental changes. Moreover, proteins are basic functional units for many

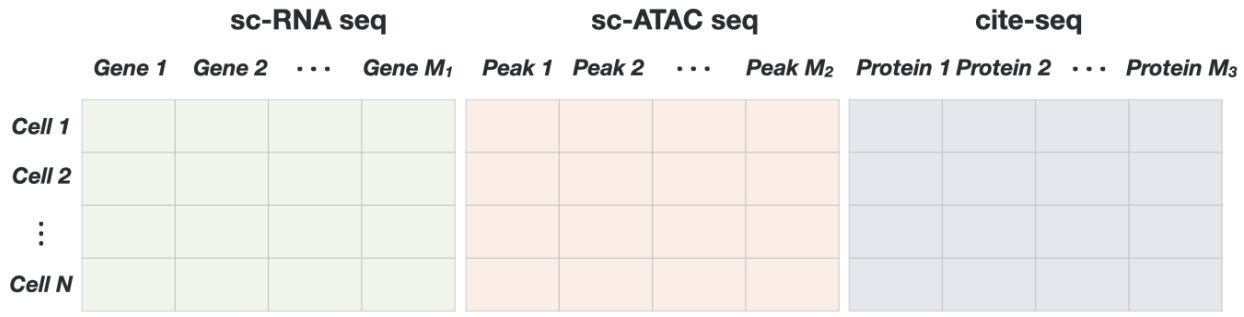


Figure 2.2 An illustration of data matrices produced by single-cell technologies.

cellular processes, which makes proteomics data essential to analyzing and understanding cellular behavior. Unlike most sequencing technologies, which have a standard process, proteomic measurements are often bespoke and designed for specific applications Vistain and Tay (2021). However, some technologies have made significant strides in capturing protein information of cells and combining this with mRNA measurements. Specifically, Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq), simultaneously sequences mRNA and measures the surface proteins on a cell Stoeckius et al. (2017). The method can sample over 1,000 genes and 80 proteins per cell, but like many other sequencing techniques, it suffers from high noise. In addition, CITE-seq is incapable of detecting intracellular proteins Baron and Yanai (2017).

2.1.2 Single-Cell Spatial Transcriptomics

Single-cell technologies that capture transcriptomic, proteomic, or epigenetic information do so with great precision but with the loss of spatial information of the cells within the tissues. However, the cells' relative locations within tissue is critical to understanding normal development and disease pathology. With spatial transcriptomic technologies, researchers are able to measure transcriptomics and leverage the spatial information or relative locations of cells in a tissue for better performing downstream tasks Crosetto et al. (2015); Moor and Itzkovitz (2017); Wang et al. (2018); Marx (2021); Asp et al. (2020); Waylen et al. (2020); Teves and Won (2020). For example, motivated by the fact that a pair of ligand and receptor with closer distance are easier to bind, HoloNet Li et al. (2022b) builds up a directed graph

based on the expression of ligand–receptor gene lists and the physical distance between the sender cell and receiver cell to represent cell–cell communication events. However, the early generations of spatially resolved profiling technologies are not at the single-cell resolution but instead sampled in groups called 'spots', which capture several cells. It requires additional work to determine the cell type proportion in spots, a process called cell type deconvolution. Alternatively, many cell imaging platforms provide RNA spatial information at the cellular and subcellular level, but the individual cells must be identified through cell segmentation methods.

Major technologies or platforms for spatial transcriptomics include multiplexed error-robust fluorescence in-situ hybridization (MERFISH), sequential fluorescence ISH (seqFISH+), Slide-Seq, Visium by 10x Ståhl et al. (2016), GeoMx Digital Spatial Profiler (DSP) Merritt et al. (2020) by NanoString,and CosMx Spatial Molecular Imager (SMI) by NanoString. MERFISH Moffitt and Zhuang (2016), first introduced in 2015, is a single-molecule-fluorescence-in-situ-hybridization (smFISH)-based technology that can be applied to fresh-frozen samples to provide subcellular resolution. While traditionally the procedure of these smFISH-based technologies is complex, a number of commercialized platforms have emerged recently, such as Vizgen, Rebus Esper, Molecular Cartography, and Resolve Biosciences Moses and Pachter (2022), which allow more convenient sequencing of spatial transcriptomic at a lower cost. As an alternative to MERFISH, seqFish+ Lubeck et al. (2014); Shah et al. (2016); Eng et al. (2019) employs 'pseudocolor' as a combination of colors to increase the amount of detectable transcripts Rao et al. (2021a).

Beyond early in-situ hybridization methods, a number of sequence-based technologies have emerged. Closely related to scRNA-seq technologies, these sequencing-based methods barcode RNAs such that each read can be mapped to its corresponding spatial location through the associated barcodes. The rest of the sequencing read is mapped to the genome to identify the transcript of origin, collectively generating a gene-expression matrix. Stahl et al. Ståhl et al. (2016) first proposed this method, which has been adapted by commercial

platforms such as 10X Visium. 10x Visium fixes spatially barcoded oligos to each spot in a capture slide (area $6.5mm^2$), with the barcoding done through DNA extension and reverse transcription for formalin fixed paraffin embedded tissues (FFPE) and fresh frozen tissues respectively. In particular, the 10x Visium expression slide contains 4 capture slides, each with area 6.5 mm^2 where fresh frozen or FFPE tissues are placed. Each of the capture slides contain a grid of approximately 5000 barcoded spots that are $55\mu m$ in diameter with a center-to-center distance of $100\mu m$ between any two adjacent spots. On average, there are $1-10$ cells in each of these spots, and $\sim 18,000$ unique genes in human ($\sim 20,000$ in mouse) can be quantified. Another major sequencing-based technology is Slide-Seq, which captures mRNA by placing barcoded beads on slides, which achieves a high resolution of 10 micron. Technological innovations further improved sequencing resolution in recent years. For instance, high-definition spatial transcriptomics (HDST) Rodriques et al. (2019a) uses wells rather than slides, whereas built upon Slide-Seq, Slide-seqV2 Stickels et al. (2020) raised the resolution to near-cellular level while reaching RNA capture efficiency of roughly 50% of scRNA-sequencing. Finally, spatio-temporal enhanced resolution omics sequencing (Stereo-seq)Chen et al. (2022) deposits barcoded DNA nanoballs in patterned arrays to achieve single-cell resolution while maintaining high sensitivity.

While 10x Visium and Slide-Seq do not profile at cellular resolution, Nanostring's GeoMx DSP is capable of cellular resolution through user-drawn profiling regions. Geomx DSP uses PC-linker to link barcodes via antibodies to proteins and RNA for identification. The spatial regions of interest (ROI) on the tissue are flexible and can be user-defined, or with pre-defined layouts (such as a square grid). During imaging, the DSP barcodes from each ROI are UV-cleaved and collected for sequencing, and the spatial information is recorded. Due to the flexibility of the ROI definitions, the ROIs can be a range of sizes, from a single-cell or hundreds of cells. The RNA assay can quantify $> 18,000$ target genes, and the protein assay can quantify > 96 proteins.

Though GeoMX can produce cellular-resolution sequencing, its scalability is limited. The

most recent platform, CosMx Spatial Molecular Imager (SMI) Lewis et al. (2022), is able to profile consistently at single cell, and even subcellular resolution. CosMx SMI follows much of the initial protocol as GeoMx DSP, with barcoding and ISH hybridization. However, the SMI instrument performs 16 cycles of automated cyclic readout, and in each cycle the set of barcodes (readouts) are UV-cleaved and removed. These cycles of hybridization and imaging yield spatial resolved profiling of RNA ($>$ 980 target genes) and protein ($>$ 80 validated proteins) at single-cell ($\sim 10\mu m$) and subcellular ($\sim 1\mu m$) resolution.

Multiplex imaging technologies have significantly advanced higher spatial resolution for single-cell profiling. Spatially resolved transcriptomic data, along with corresponding imaging data, enables single-cell or even subcellular analysis on both spatially morphological and pixel resolution information. Recently, antibody-based multiplexed imaging methods have dominated the multiplexing approaches, as they can capture cellular organization and tissue phenotypic heterogeneity at the protein level. They utilize various protein markers for cellular identification. Immunohistochemistry (IHC)Coons et al. (1942), first reported in 1942, is one of the most commonly used multiplexed imaging methods. It uses appropriately-labeled antibodies to bind specifically to their target antigens in situ (in the original site), which can be better captured by current light or fluorescence microscopy. Due to the limited protein readouts, methods including multiplexed immunofluorescence (MxIF) Gerdes et al. (2013) and cyclic immunofluorescence (CyCIF) Lin et al. (2015, 2018) were proposed to add more new antibodies in multiple rounds of staining. Another imaging platform, Co-Detection by IndeXing (CODEX)Goltsev et al. (2018), is designed for up to 40 proteins using cyclic detection of DNA-indexed antibody panels. Imaging mass cytometry (IMC)Giesen et al. (2014) is an evolutionary technology that leverages mass spectroscopy to obtain images from tissues with 40+ labels simultaneously. This vastly reduces data noise and enhances the multiplex capability. Multiplexed ion beam imaging (MIBI)Keren et al. (2019) is also performed by imaging tissues with secondary ion mass spectrometry based on metal-labeled antibodies. These multiplexed imaging tools provide high-dimensionality imaging assays at

the single-cell level and enable analyzing and understanding of the single-cell function and tissue structure.

2.1.3 Spatial & Bulk Deconvolution

In addition to omics data, spatial transcriptomics technologies provide spatial location data of samples. The spatial resolution capabilities now range from multi-cell pools or bulk samples, to single-cell and even sub-cellular levels. This information gives us yet another aspect to study the functional dynamics in biological systems. Along with cellular heterogeneity, we can now study spatial heterogeneity and the cellular composition of tissues (Molho et al., 2024; Rao et al., 2021b; Fan et al., 2023). With spatial transcriptomics, we can analyze omics data within a spatial context, which can offer deeper insight into cellular interactions (Tian et al., 2023; Raredon et al., 2023), and cell type localization under varying conditions. For example, we can study how cell types are organized in cancerous or diseased tissue (Williams et al., 2022; Rao et al., 2021b). Of course, to analyze this most readily and objectively requires single-cell resolution. With lower resolution (multi-cell pools or spots), each omics data point is an aggregate of the multiple cells within the captured region. For example, with RNA we would have an aggregate expression of the cells in the cell-pool, which heterogeneous in many cases. While this does provide us with some spatial context, it does not readily allow for studying the spatial distribution of cells and thus any downstream analysis that is dependent on cell type composition. While there are options with single-cell resolution, the more affordable options in spatial transcriptomics are not single-cell resolution, naturally. Some popular lower resolution options include 10X Visium (Maynard et al., 2021), and Slide-seq (Stickels et al., 2021; Rodriques et al., 2019b).

With this tradeoff in mind, it is desirable to make best use of the non single-cell resolution spatial transcriptomics data. Though spatial information is not captured at single-cell resolution, we can use the spatial omics data together with robust reference single-cell data to deconvolute the aggregated omics data in terms cell type composition. Indeed, this task is called cell type deconvolution (Ding et al., 2024a; Molho et al., 2024), which is the task of

quantifying the cell type composition of bulk mixtures (in this context, spatial mixtures) by decomposing bulk mixture omics measurements by cell type proportion. The simplest idea to accomplish this is through non-negative matrix factorization (NMF). Here, we take a cell type expression profile derived from some set of robust single-cell reference data, and regress onto the bulk mixture expression data, with a non-negative constraint on the coefficients. These non-negative coefficients (after normalization) are taken as the cell type composition estimates. This is fairly intuitive, as the bulk mixture expression is an aggregate of the multiple cells in the mixture and so finding the cell type composition (coefficients) that best match a standard cell type expression profile should give a decent estimate of the true cell type compositions. In recent years there has been further development in cell type deconvolution, with many methods building on this basic idea of NMF. Here are some methods and brief descriptions (further description found in the Deconvolution chapter):

SPOTlight (Elosua-Bayes et al., 2021a) essentially applies a seeding to NMF regression to deconvolve bulk mixtures with reference scRNA-seq data. Stereoscope (Andersson et al., 2020) is a Bayesian model that integrates information from both single-cell and spatial transcriptomics data to estimate the probability of each cell type at each location within the tissue sample. Cell2location (Kleshchevnikov et al., 2022a) puts bulk expression into a Bayesian hierarchical framework with a spatially determined prior on the cell-type compositions. Outside of these shallow-classical methods that build on the basic idea from NMF, there has been developments in deep learning-based methods (Molho et al., 2024) fro cell type deconvolution. Tangram (Biancalani et al., 2021) spatially aligns reference scRNA-seq data and identifies spatially co-expressed gene modules, from which it infers the presence of different cell types in a tissue sample. DSTG (Song and Su, 2021a) creates synthetic mixtures by sampling reference scRNA-seq data and map the synthetic and real bulk mixtures to a common domain. A graph is then constructed by mutual-nearest neighbors in this domain and taken as input through graph-based convolutional networks, which directly estimate the cell type proportion.

Currently, a road block in the way of making significant advancements in cell type deconvolution methods is the relatively limited amount of data with ground truth cell type compositions. At its core, cell type deconvolution reduces to an inverse problem of sorts, as we are trying to recover cell type composition from an aggregate signal over all cells in a mixture. The only way to get cell type labeled data is with single-cell resolution, so synthetic experiments must be done to create cell type labeled mixture data. In most cases, such labeled data is small and taken from non-human tissue. Moreover, creating synthetic mixtures does not always reflect the heterogeneity and organization of cell types within real tissues, and hence real biological systems. For example, Lulu Yan et al. (Yan and Sun, 2023) provide three deconvolution benchmark datasets from mouse tissues, which contains at most 80,000 cells. Bin Li et al. (Li et al., 2022a) provide 32 synthetic deconvolution datasets, taken from scRNA-seq reference data.

Further developments in methods and benchmark datasets in cell type deconvolution can provide a way to better leverage high-throughput bulk omics data for downstream analyses that depend on cell type composition. Of course, this is strengthened further with spatial omics data, as we can then use spatial information together with cell type composition. An important area that could greatly benefit from these developments is in immuno-oncology, to better understand tumor cell organization, and their interactions with the immune system. Indeed, tumors grow from cell proliferation, making the study of immune cell composition and organization an important tool for understanding cancer and developing better therapies (Sturm et al., 2019). Towards this end, in the Cell Type Deconvolution section we develop a spatial transcriptomic benchmark dataset from samples that include TME (human), a deconvolution method, and set of benchmarking exercises.

2.2 Probabalistic Framework for scRNA-seq Data

2.2.1 scRNA-seq Count Models

Cell type composition relies on cell type annotation of cells from scRNA-seq data, either directly or as a reference for deconvolution of bulk data. Many approaches to cell type anno-

tation rely on scRNA count models to estimate the annotations. Below is small progression of scRNA count models that add parameters to incorporate technical effects, which is often needed to deal with technical effects across studies (batches) and platforms. We are given d genes, N single-cells, K cell-types

Observed

$$X \in \mathbb{R}^{N \times d} = (x_{i,j}) \text{ - observed single-cell expression}$$

$$\widetilde{X} = (\tilde{x}_{i,j}) \in \mathbb{R}^{K \times d} \text{ - estimate of mean cell-type profiles } \mu$$

Unobserved

$$\mu = (\mu_{i,j}) \in \mathbb{R}^{K \times d} \text{ - True mean cell-type profiles}$$

$$c(i) \text{ - cell-type of cell } i \in [1, N]$$

Goal: Predict true cell-type $c(i)$ of each cell $i \in [1, N]$ from $\{X, \widetilde{X}\}$

$$X_{i,j} \sim NB(\mu = \mu_{c(i),j}, size = \theta) \equiv Poisson(\Gamma) \tag{2.1}$$

$$\Gamma = Gamma(\theta, \theta/\mu_{c(i),j})$$

Technical effects model

$$X_{i,j} \sim Poisson(b_i) + NB(\mu = s_i \mu_{c(i),j}, size = \theta) \tag{2.2}$$

Technical effects of observation

(i) Not all mRNA transcripts get detected $\rightarrow$ Detection efficiency scale factor s_i

(ii) Off-target binding $\rightarrow$ Background counts additive factor b_i

(iii) Gene-specific detection efficiencies

Simplified technical effects model

$$X_{i,j} \sim NB(\mu = b_i + s_i \widetilde{X}_{c(i),j}, size = \theta) \tag{2.3}$$

Simplifying assumption to avoid discrete distribution convolution. Same mean model as (2.2), but higher variance.

2.2.2 scRNA-seq Count Pre-processing

Forgiving the minor deviation in symbol definition from the previous section, we define

$$X = [X_{gm}] \in \mathbb{R}^{D \times N} \text{ raw expression : } D \text{ genes} \times N \text{ samples}$$
$$\overline{X} = \frac{1}{N} X1 = [\overline{x}_g] \in \mathbb{R}^{D \times 1} \text{ mean expression : } D \text{ genes over } N \text{ samples}$$

and we outline some standard pre-processing steps applied to scRNA-seq expression data.

1. Select $d \leq D$ candidate genes G with highest mean expression

$$G = \underset{G' \subset [1,D], |G'|=d}{arg\ max} \left\{ \sum_{G'} \overline{x}_g \right\}$$

2. For all d candidate genes $g, g' \in G$, where $g \neq g'$:

(i) Compute log-transformed expression ratios $L = [L_{gg'}] \in \mathbb{R}^{N \times d(d-1)}$

$$L_{gg'} = \left[log_2\left(\frac{x_{gm}}{x_{g'm}}\right) \right]_{m=1}^{N} \in \mathbb{R}^{N \times 1}$$

(ii) Compute pair-wise variations $V = [V_{gg'}] \in \mathbb{R}^{d \times (d-1)}$

$$V_{gg'} = SE(L_{gg'}), \text{ where}$$

$$SE^2(L_{gg'}) = \frac{1}{N-1} \sum_{m=1}^{N} \left(log_2\left(\frac{x_{gm}}{x_{g'm}}\right) - \overline{L}_{gg'} \right)^2, \text{ where } \overline{L}_{gg'} = \frac{1}{N} 1^T L_{gg'}$$

or, in matrix form

$$SE^2(L_{gg'}) = \frac{1}{N-1} \Big(L_{gg'} - \overline{L}_{gg'} 1 \Big)^T \Big(L_{gg'} - \overline{L}_{gg'} 1 \Big) = \frac{1}{N-1} L_{gg'}^T \Big(I_N - \frac{1}{N} 11^T \Big) L_{gg'}$$

(iii) Compute gene-stability measures $M = [M_g] \in \mathbb{R}^{d \times 1}$ - arithmetic mean of pairwise variations

$$M_g = \frac{1}{d-1} \sum_{g' \neq g} V_{gg'}$$

3. Determine $d^* \leq d$ housekeeper/reference genes G^*: genes with the lowest gene-stability measure (low $\implies$ more stable)

$$G^* = \underset{G' \subset G, |G'|=d^*}{arg\ min} \left\{ \sum_{G'} M_g \right\}$$

Normalizing factors: $\widetilde{F} = \overline{F} * F \in \mathbb{R}^{N\times 1}$ where

$$F = \big[F_m\big]_{m=1}^{N}, \;\; F_m = \Big(\prod_{G^*} x_{gm}\Big)^{1/d^*}$$

$$\overline{F} = \frac{1}{N} 1^T F$$

Normalized expression matrix: $\widetilde{x} = X diag(\widetilde{F}) \in \mathbb{R}^{D\times N}$

$$median(X) = [median(\mathbf{X}_c)]_{c=1}^{K} \in \mathbb{R}^{D\times K}\text{: median normalized expression}$$

Then re-scale for equal median expression across cells

$$\widetilde{X} = median(X) diag\big(M(X)\big), \text{ where } M(X) = \big(\tfrac{median(\mathbf{x}_1)}{median(\mathbf{x}_c)}\big)_{c=1}^{K}$$

The background and normalized read counts can then be computed from negative control probes.

$$\{P_l\}_{l=1}^{L}: \text{ probe pools}$$

$$\{R_l\}_{l=1}^{L}, R_l \subset P_l: \text{ negative control (nc) probes in pool } P_l$$

$$\widetilde{X}_{R_l} \in \mathbb{R}^{|R_l|\times N}: \text{ normalized read counts of nc probes (from } \widetilde{X})$$

CHAPTER 3

PLANT STRESS RESPONSE

The chapter is comprised of work on a collaborative project between Dr. Xie's and Dr. Shiu's labs. Here, I have reproduced original experiments I worked on with Dr. Yuning Hao and Dr. Runze Su. I give them both credit for the original experiments, and figures 3.4-3.7, 3.9-3.10. In part, the reproductions were done to make further inquiries and developments. Particularly, I contributed inquiries and experiments in stress type grouping and alternative model architectures.

3.1 Introduction

A central problem in molecular plant biology is to understand how plants respond to various abiotic and biotic stressors (e.g. heat waves, drought, and pest infestations) at the molecular level. As stresses trigger changes in gene expression levels, differential expression analysis is a key tool to understand stress response in plants. The goal of such analyses is to determine motifs that are

A main component of gene expression regulation is through the binding of transcription factors to specific sequences of DNA called regulatory elements (motifs). For this reason, an avenue of research has been to identify these transcription factors and the respective regulatory motifs, in order to predict gene expression responses Uygun et al. (2017); Wilkins et al. (2016). However, identifying individual regulatory motifs, such as transcription factor binding sites (TFBS), is only a small part of the complex process of gene regulation. Indeed, gene regulation processes also depend on the location, orientation, quantity and co-localization of regulatory motifs. These dependencies form the structures that modulate gene regulation, and these structures form what is called regulatory grammar Weingarten-Gabbay and Segal (2014).

Understanding of regulatory grammar by computational modeling of these complex dependencies has thus become a hot area of bioinformatics research. Many advancements towards modeling complex regulatory grammar have come from deep sequence learning models,

traditionally used in natural language processing.

One of the early deep learning models developed to account for the sequential dependencies was DeepSea Zhou and Troyanskaya (2015). This was done by using convolutional neural networks (CNN), from which motifs and local dependencies were learned, ultimately used for functional-variant prediction.

Building on the DeepSea model, Quang and Xie developed DanQ Quang and Xie (2016), which couples the CNN with a recurrent neural network (RNN), namely a bi-directional long short-term memory network (LSTM) Hochreiter and Schmidhuber (1997). The LSTM component helps identify long-range dependencies [9], and hence co-localization dependencies. As the LSTM is bi-directional, it learns these features on both the forward and reverse ordering of sequences (hence orientation).

These developments of deep sequence learning models are easily tailored and applied to our problem of interest: predicting plant stress response from DNA sequences. Building on these ideas, we propose DeepCAT, a convolutional self-attention architecture to predict plant stress response from DNA sequences [1]. DeepCAT consists of 3 layers. The first is a convolutional layer which converts DNA base-pairs to a numerical sequence, identifying key predictive motifs and local dependencies. The second layer is self-attention, which captures key predictive co-localization dependencies. Lastly, a fully-connected (FC) layer to output prediction scores of gene up-regulation under different abiotic and biotic stresses.
A few properties of DeepCAT yield advantages over popular learning models. First, the self-attention Vaswani et al. (2017) method has been shown to capture long-range sequence dependencies beyond the capabilities of RNNs Vaswani et al. (2017). Another property of self-attention is that it does not impose a strict order on how a sequence is processed Vaswani et al. (2017), which is advantageous since the ordering of base-pairs may not always be a factor in gene expression. With these advantages we hypothesize that DeepCAT outperforms many other popular learning models in predicting plant stress response. Also, that Deep-

[1]DeepCAT is an adaptation of another model our lab has proposed: CANEE, a convolutional self-attention architecture to analyze the function of non-coding DNA sequences

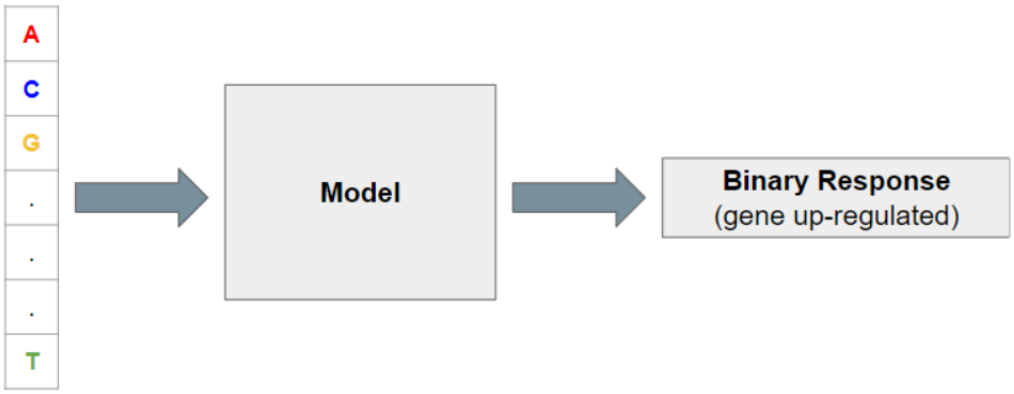


Figure 3.1 Basic DNA up-regulation prediction model.

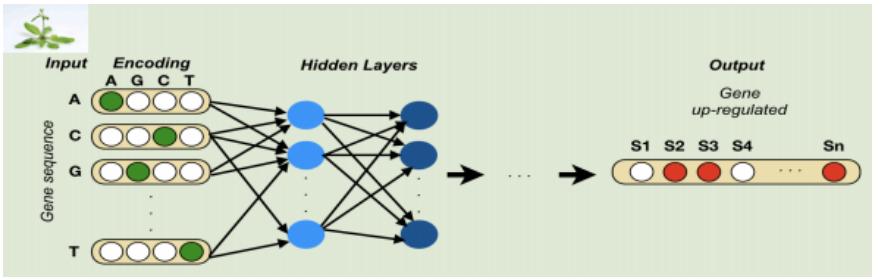


Figure 3.2 High level pipeline of DeepCat.

CAT extracts sequential features that can identify known and potentially novel transcription factor binding motifs (TFBMs) and their interactions. We demonstrate this on the problem of predicting gene expression of *Arabidopsis thaliana* (A.thaliana) in response to 57 abiotic and biotic stress conditions.

3.2 Data and Problem Statement

Given raw arabidopsis thaliana DNA sequence data, the objective is to predict gene up-regulation under 57 environmental stress conditions. Specifically, to predict if an arabidopsis gene was up-regulated or not in shoot tissue under each of 36 abiotic (e.g. cold, heat, osmotic) and 21 biotic (e.g. 71 Pseudomonas syringae, bacterial flagellin) stress conditions.

Gene expression and sequence data of 20,799 A. thaliana genes each consisting of 3,200-bp (covering promoter and 5' UTR) were downloaded from the AtGenExpress database and processed as in Uygun et al. (2017). Genes with a log2 fold change ≥ 1 were considered up-regulated. In particular, the preprocessed and normalized expression data from

AtGenExpress was used to calculate log2 fold change between stress and control conditions using Limma in the R environment. Genes with a log2 fold change ≥ 1 were considered up-regulated.

DNA sequences were pulled for each gene from TAIR10. Particularly, the sequences were taken from 1-kilobase (kb) upstream and 500-base pairs (bp) downstream the transcription start site and 500-bp upstream and 1-kb downstream the transcription stop site. These sequences were then one-hot encoded, with each sequence converted into a 3200x4 binary matrix. The columns correspond to A,C,G,T, and rows correspond to the position in the DNA sequence, with each row containing a single 1 in one column and zeros in the remaining columns. Genes were randomly assigned according to a training-validation-test split of 70-10-20.

3.3 Model Architecture and Training

DeepCAT consists of 3 main modules: (1) CNN, (2) self-Attention and (3) FC/output.

CNN Module

A 1D convolutional layer and a max-pooling layer makes up the CNN module. Suppose the input of the convolutional layer has shape (N, I, L) and the output is (N, O), then the 1D convolution is given by:

$$Conv1D(X_{N_m,O_j}) = ReLU(Bias(O_j) + \sum_{k=0}^{I-1} W_{O_j,k} \star X_{N_m,k}),$$

where for batch size N, input sequence dimension I, output element dimension O, and input sequence length L. The subsequent max-pooling is given by:

$$Output(N_i, C_j, k) = \max_{m=0,1,...,kernel\,size-1} input(N_i, C_j, k+m),$$

where input value is of size (N, C, L).

Self Attention Module

First, positional encoding is applied to enable the model to capture relative positional information, and thus potential order structure. This consists of applying a positional embedding

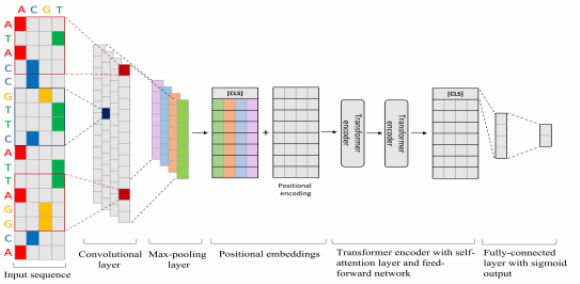


Figure 3.3 DeepCAT architecture.

of the sequences that are output from the CNN module. We apply the sinusoidal embedding as in Vaswani et al. (2017).

The self-attention mechanism, as in Vaswani et al. (2017), has three factors: query Q, key K and value V. Assuming the input is X, the formulation can be expressed as:

$$Q_i = W_Q X_i$$

$$K_i = W_K X_i$$

$$V_i = W_K X_i$$

$$S_{i,j} = \frac{Q_i \cdot K_j}{\sqrt{d}}$$

$$Score_{i,j} = \frac{exp(S_{i,j})}{\sum_k exp(S_{i,k})}$$

$$output_i = \sum_j S_{i,j} V_j$$

Here, $W_{(\cdot)}$ represents a weight matrix. The weights corresponding to each of these factors, and hence the factors themselves, are learned during the training process.

The output from the self-attention network is then passed to the FC module.

Fully-Connected Output Module

We apply a single FC layer, giving weighted scores for each of the 57 stress types. We then apply a sigmoid output layer to obtain probability scores for gene up-regulation. We trained DeepCAT by minimizing the average multi-task binary cross-entropy loss in mini-batches of size 50 using the Adam optimizer Kingma and Ba (2014). All the weights and biases were initialized with Xavier (uniform) Glorot and Bengio (2010) and zero values respectively. For

model regularization purposes, we applied dropout with rate of 0.1 in attention layers.

Validation data was used to determine an optimal number of training iterations. Namely, we use an early-stopper to stop the training process if the validation loss does not decrease for a set number of epochs (default 5), thus keeping the model that performs best on the validation set.

In all of our experiments we trained DeepCAT with settings: 320 convolutional kernels/filters, kernel dimension 26, pooling dimension 13, and used 4 attention heads.

Our implementation was with PyTorch, and our experiments (training and testing) were ran on NVIDIA K80 GPU.

3.4 Experiments

Using the fully trained models, performance was measured on the testing data (70-10-20 train-valid-test split). We used two metrics: the Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) and the Precision Recall-Area Under the Curve (PR-AUC). For overall comparison purposes we averaged the PR-AUC and ROC-AUC across the 57 stress types.

In each of our experiments we compare DeepCAT against a few well-known shallow and deep learning models. The shallow models consisted of Support Vector Machines (SVM) and Random Forest. The deep learning model we compared against was essentially the DanQ model Quang and Xie (2016), with the modification of the output layer to give plant response probability scores for the 57 different stress types. We chose this deep learning model for comparison, as it has a similar structure as DeepCAT, and has performed well on a different but similar problem with human DNA data. We first evaluated baseline performances of our standard DeepCAT model and the other learning models. To improve on these baseline results, we experimented with transfer learning strategies.

3.4.1 Transfer Learning - experimentally verified and pre-learned sources

The main idea of transfer learning Tan et al. (2018) is to leverage existing knowledge in one setting to learn a task in a different but related setting. Here we injected existing

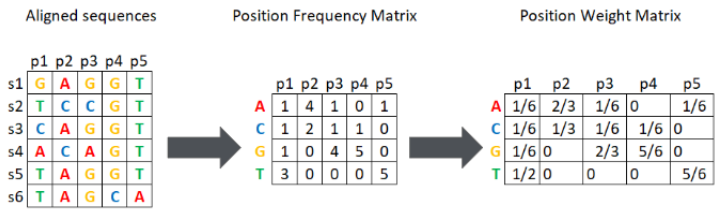


Figure 3.4 PFM construction process.

knowledge in two ways. One was through experimentally verified information. The other was information learned from a model with a much larger data set. As the kernels in the CNN layer of DeepCAT act as DNA motif finders, we experimented initializing the kernels with known A.thaliana TFBMs. That is, we initialize with the resulting position weight matrices (PWM). To do this, consider a set of n aligned sequences of length m. Then we construct the position frequency matrix (PFM) of size $4 \times m$ with the counts of each base over the sequences. We then apply row-wise normalization to get the $4 \times m$ position probability matrix, as each column-position sums to 1. Finally, we compute the PWM with log odds from the PFM against a background model. In the simplest case, we use a uniform background model, i.e. each base is equally likely to occur in a sequence. Moreover, the DanQ model used a massive amount of human gene data ($>$ 4 million), so we also experimented initializing the kernels with the kernel weights learned in the DanQ model. We found that implementing these transfer learning methods in DeepCAT lead to better performances across nearly all 57 stresses (Figures 3.5 and 3.6).

3.4.2 Multi-task learning - Stress Grouping

Our previous results are based on learning all 57 tasks (57 stress responses) simultaneously. Hence we are in a multi-task learning (MTL) setting. In this setting, the learning process seeks the shared representation (feature mapping of the sequences) that best predicts response to each of the 57 stress types. This can be problematic, however, because different stress types may induce very different underlying regulatory mechanisms, and finding a good shared representation may not be possible. The expectation is that in an MTL setting, learning stresses with similar underlying regulatory mechanisms will benefit from

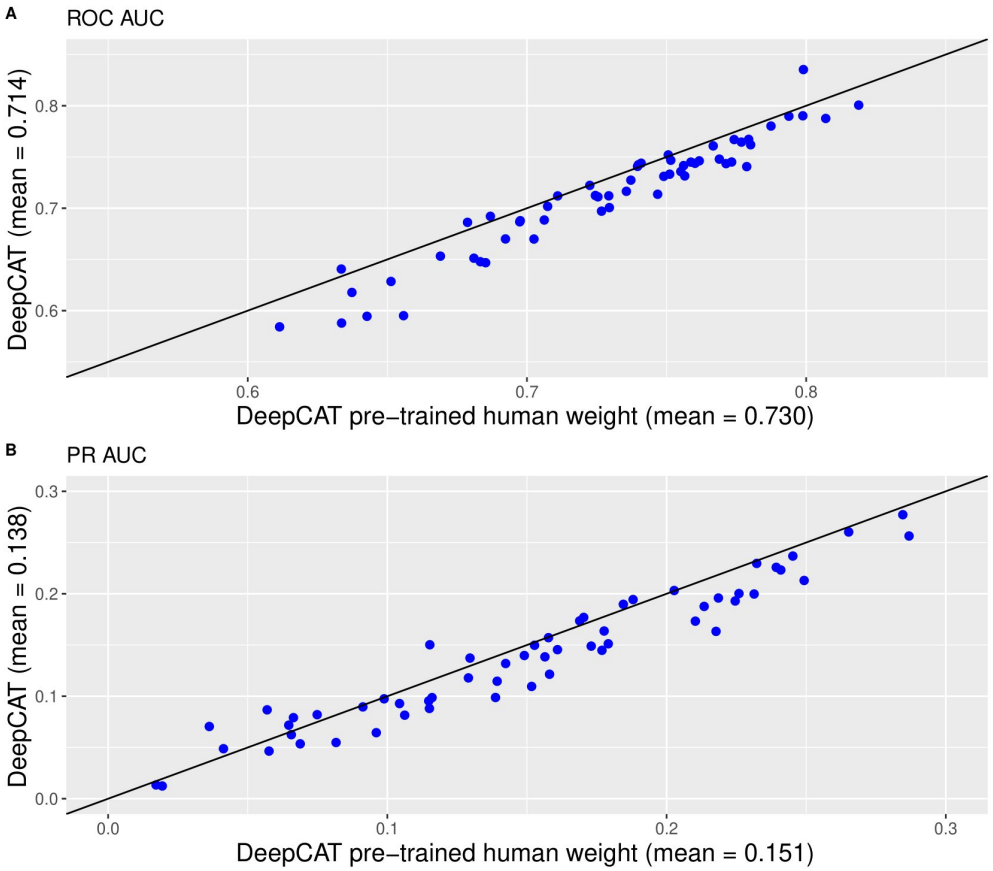


Figure 3.5 Performance of DeepCAT with kernels initialized from weights learned from the DanQ human model - updated original figure from Dr. Yuning Hao.

each other, while stresses with very different underlying regulatory mechanisms may hinder performance.

We tested this hypothesis by finding related stresses through k-means (k=3) clustering of the stress responses, and trained the three models separately. We also paired this a transfer learning scheme by initializing the convolutional kernels with known A.thaliana TFBMs. Overall we found that both of these experiments lead to better performances, with the latter yielding the best performance (Figure 3.8). In figure 3.7 we show the clustering hierarchy of the stress types.

3.5 Motif Learning and Location Dependencies

An interesting result is that the DeepCAT model can be interpreted as a motif learner, by a translation of the kernels in the convolution layer to positional weight matrices B. Alipanahi

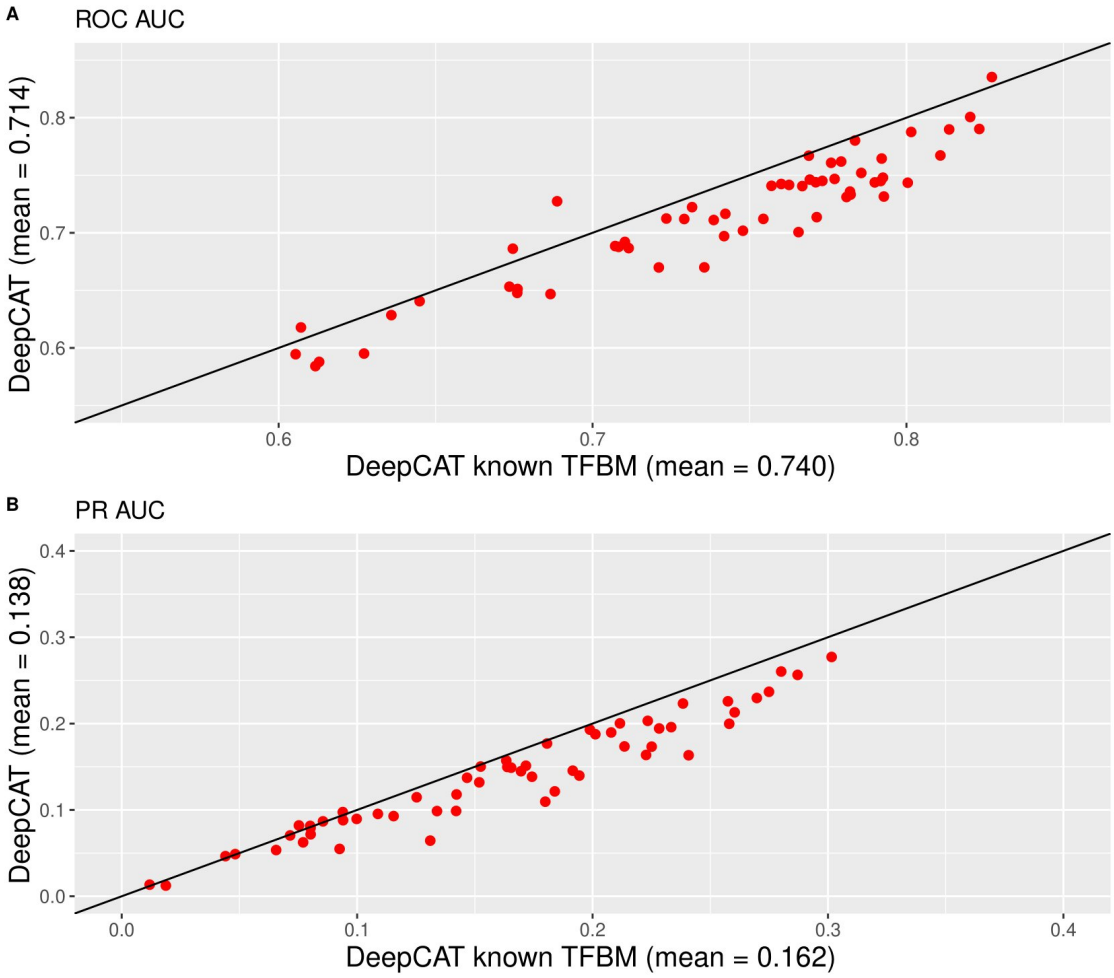


Figure 3.6 Performance of DeepCAT with kernels initialized from experimentally verified TFBMs - updated original figure from Dr. Yuning Hao.

and Frey (2015). We aligned these to known motifs from the DAP-seq and CIS-BP databases using TOMTOM software (see meme-suite.org/meme/tools/tomtom). Of the 319 motifs learned by our model, 114 significantly match known motifs ($E < 0.1$); a threshold of 0.05 was used for p-value to measure the similarities. Figure A.7 shows this process, Figures A.8-A.11 show resulting learned TFBMs, which are all found in the appendix.

Analyzing the scores in the attention module, we found interactions of motifs at different positions. From Figure A.11 we can see that the attention model identifies interactions between base-pairs at long ranges, and thus identifies long-range co-localization dependencies.

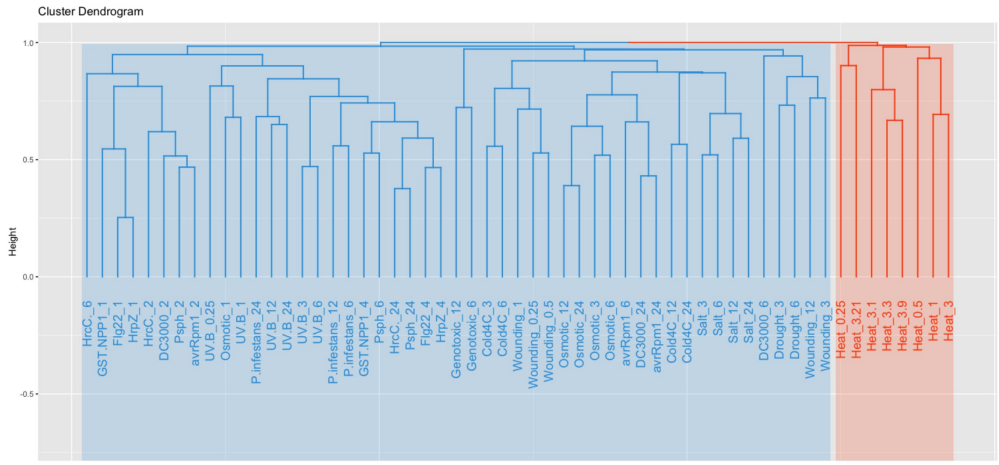


Figure 3.7 Clustering hierarchy of the stress types from k-means clustering. Three large clusters are identifiable, with red highlighted stresses being heat related - original figure from Dr. Yuning Hao.

Figure 3.8 Performance of DeepCAT with known TFBM initialized kernels, and the clustered response multi-task model (also with TFBM initialized kernels) - original figure from Dr. Yuning Hao.

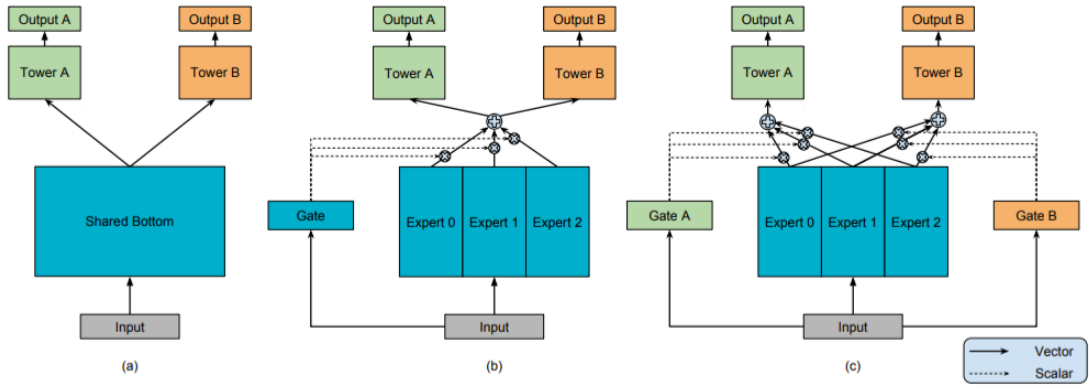


Figure 3.9 Task-specific architectural variations.

3.6 Additional Experiments and Future directions

A direction I explored, and continue to explore, for mitigating negative transfer between stress types is through the addition of task-specific layers. In its most basic form, the idea is to compose the model in two modules. The first module is a shared module, where the parameters are shared between all the tasks. This is the set-up of the basic multi-task architecture. The second module consists of task-specific silos or towers. These serve as the output layers where each task has a dedicated silo, and parameters are not shared between the tasks. These architectures can be improved by incorporating gating mechanisms that limit the amount of shared information each silo receives as input, as well as incorporating a mixture of experts. At this point in time I have only done preliminary bench-marking exercises with these architectures and have not yet seen significant improvement above our standard DeepCAT architecture.

Nonetheless, with DeepCAT we have shown how deep sequence learning and other learning mechanisms, such as grouped learning and transfer learning, can move us towards solving the problem of plant stress response prediction. Moreover, these methods are able to learn and extract key motifs and long-range motif interactions, which are important components of understanding regulatory grammar, and hence gene regulation.

CHAPTER 4

CELL TYPE DECONVOLUTION

4.1 Introduction

Quantifying cell type composition (proportion) is an important tool to better understand and characterize diseases and other abnormalities by differences in cell type composition between experimental groups Karagiannis et al. (2022). With single-cell transcriptomics data, this can be done by annotating the cell type of each cell. Cell type annotation is often done computationally by comparing expression patterns in a cell with reference cell type expression profiles with maximum likelihood approaches. Segmentation based methods are also used, as well as expert validation (when the sample size is small).

While advances in high-throughput transcriptomics technologies have facilitated this task by providing single-cell resolution, bulk sequencing is significantly less labor intensive and costly Jin and Liu (2021). Estimating cell type composition in bulk samples offers a more economical approach, and provides a way to make use of a wealth of public bulk data. A more recent motivating factor for estimating cell type composition in bulk data is due to developments in spatial transcriptomics, as doing so provides a way understand cell type composition in a spatial context within a tissue. Moreover, spatial transcriptomics technologies measure gene expression of small spatially tagged regions, but most platforms do not have single-cell resolution.

An issue with bulk sampling of biomolecular information, such as bulk RNA-seq, is that information is averaged across a cell pool that is often heterogeneous. This makes it difficult to deconfound cell type composition from differences in the molecular profile between experimental groups Repsilber et al. (2010). *Cell type deconvolution* is the task of estimating and deconfounding the composition of cell types in bulk samples of biomolecular information. This is a type of inverse problem, as we are trying to determine the signal of individual cell types from aggregated readings across multiple cell types. Solving this task then requires some transfer learning approaches, using single-cell data as a reference (transfer source) and

bulk data as a query (transfer target). Typically, gene expression data are used, though other data such as protein expression Okendo et al. (2022) or DNA methylation have also been used Singh et al. (2021).

4.2 Problem Formulation

The input of the task of cell type deconvolution consists of three components: 1) bulk gene expression (to be deconvoluted), 2) reference scRNA-seq, 3) (optional) spot coordinates to indicate the location within the tissue of each spot. Note additional data, such as histology images can be leveraged as well.

The problem is formulated as follows. We're given bulk expression data $Y \in \mathbb{R}^{d \times n}$ where each cell-pool $i \in [1, n]$ is composed of a mixture of cell types $[1, K]$. Then for each cell-pool $i \in [1, n]$, we wish to construct an estimator $\hat{a}_i \in \Delta^{K-1}$ of the true cell type composition $a_i \in \Delta^{K-1}$, where Δ^{K-1} is the regular K-simplex

$$\Delta^{K-1} = \{x \in \mathbb{R}^K : \sum_{k=1}^{K} x_k = 1, x_k \geq 0 \text{ for } k = 1, 2, ..., K\} \tag{4.1}$$

We are also given some reference scRNA-seq expression data $X \in \mathbb{R}^{d \times N}$ with one-hot labeled cell types $C \in \mathbb{R}^{N \times K}$. Then construct the estimator of cell type compositions for the n cell-pools by some function

$$\widehat{B} = F(Y, X, C) \in \mathbb{R}^{K \times n} \tag{4.2}$$

If the spatial information (2D or 3D coordinates in the given tissue)

$$S = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix} \in \mathbb{R}^{n \times m},\ m \in \{2, 3\}$$

of the cell-pools are available, then we may incorporate this into the estimator of cell type compositions for the n cell-pools by some function

$$\widehat{B} = F(Y, S, X, C) \in \mathbb{R}^{K \times n} \tag{4.3}$$

We can further generalize this setup to account for multi-batch data.

Reference: N reference single-cells, K cell-types

Target: T experiments/batches of n_t cell-pools

Observed

$$Y(t) \in \mathbb{R}^{d \times n_t} = (y(t)_{i,j}) \text{ - bulk mRNA counts}$$

$$X \in \mathbb{R}^{d \times N} = (x_{i,j}) \text{ - reference single-cell mRNA counts}$$

$$c(i) \text{ - cell type of reference cell } i \in [N]$$

$$\widetilde{X} = (\tilde{x}_{i,j}) \in \mathbb{R}^{d \times K} \text{ - estimate of mean cell-type profiles } \mu$$

Unobserved

$$\mu = (\mu_{i,j}) \in \mathbb{R}^{d \times K} \text{ - True mean cell type profiles}$$

$$B(t) = (\beta(t)_{i,j}) \in \mathbb{R}^{K \times n_t} \text{ - cell type abundances of cell-mixtures}$$

$$\beta(t)_{\cdot,j} \in \Delta^{K-1} = \{x \in \mathbb{R}^K : \sum_{k=1}^{K} x_k = 1, x_k \geq 0 \text{ for } k = 1, 2, ..., K\}$$

For simplicity, we return to the single batch setup to further define the deconvolution problem under two different scenarios - with and without ground truth cell type compositions.

Unsupervised solution:

$$B^*(Y; X) = \underset{B}{argmin}\ \underset{\theta}{min}\ L\big(Y, \widehat{Y}(B, \theta; X)\big), \text{ where}$$

$$\widehat{Y}(B, \theta; X) = h(X; \theta)B$$

Supervised solution:

$$B^*(Y; X) = \widehat{B}(\theta^*; X, Y), \text{ where}$$

$$\widehat{B}(\theta; X, Y) = f(X, Y; \theta) \text{ and } \theta^* = \underset{\theta}{argmin}\ L(B, \widehat{B}(\theta; X, Y))$$

Baseline unsupervised solutions:

(1) multivariate linear regression - independently for each sample mixture

$$\boldsymbol{\beta}^*_{LS} = \underset{\boldsymbol{\beta}}{argmin}\, ||\mathbf{y} - X\boldsymbol{\beta}||_2$$

(2) non-negative constraint of (1)

$$\boldsymbol{\beta}^*_{NNLS} = \underset{\boldsymbol{\beta} \geq 0}{argmin}\, ||\mathbf{y} - X\boldsymbol{\beta}||_2$$

(3) multivariate regression with multiplicative log-normal errors - independently for each sample mixture

$$\boldsymbol{\beta}^*_{lnm} = \underset{\boldsymbol{\beta} \geq 0}{argmin}\, ||log(\mathbf{y}) - log(X\boldsymbol{\beta})||_2$$

Typically, cell type deconvolution models are evaluated on datasets with ground truth cell type proportions using mean squared error (MSE), mean absolute error (MAE), correlation, cross-entropy and Jensen-Shannon divergence (JSD). In most cases, however, non-simulated datasets do not have ground truth cell type proportions. In this unsupervised setting, if profiled marker proteins are also provided with the dataset, one evaluation metric Danaher et al. (2022) is the correlation between predicted cell type proportions and the respective marker proteins.

4.3 Survey of Methods

Most classical methods for cell type deconvolution are based on non-negative matrix factorization (NNMF). The most basic method is non-negative least squares (NNLS), where some reference scRNA-seq gene expression is used to create a cell-profile matrix $\widetilde{X} \in \mathbb{R}^{d \times K}$, which is then regressed onto the bulk gene expression. The resulting (non-negative) coefficients are then used as the cell type composition estimates.

$$\widehat{B} = \underset{B \geq 0}{argmin} ||Y - \widetilde{X}B||_F \tag{4.4}$$

Here, the idea is that the single-cells' expression will aggregate linearly, respective to their proportion in the bulk sample. The cell profile or signature matrix is typically constructed through the median or mean across cells within each cell type of interest. A penalized NNLS approach is taken with DWLS Tsoucas et al. (2019), which applies a dampened weighting

scheme to the standard NNLS framework. Here, each gene's error term is weighted by the squared inverse of the predicted bulk expression level. This is done to reduce bias towards highly expressed genes, or genes that are highly represented across cell types. Most other traditional methods build on these ideas.

NMFreg Rodriques et al. (2019a) applies non-negative matrix factorization (NNMF) on the reference X to construct a basis in a lower dimensional gene space,

$$W, H = \underset{W', H' \geq 0}{argmin} \big|\big| X - W'H' \big|\big|_F \tag{4.5}$$

where the rows of $H \in \mathbb{R}^{K \times N}$ are the cell-topic embeddings, and the columns of $W \in \mathbb{R}^{d \times K}$ the corresponding weightings. The cell-topic profiles are then used for the deconvolution of the bulk data via NNLS

$$\widehat{B} = \underset{B \geq 0}{argmin} ||Y - HB||_F \tag{4.6}$$

Building on NMFReg, SPOTlight Elosua-Bayes et al. (2021b) uses non-negative matrix factorization to produce the cell-topic profile matrix. Taking W, H from the first step of NMFReg, SPOTligt then constructs spot-topic profiles $P \in \mathbb{R}^{K \times n}$ through NNLS of X onto W

$$P = \underset{P' \geq 0}{argmin} ||X - WP'||_F \tag{4.7}$$

Cell-topic profiles $\widetilde{H} \in \mathbb{R}^{K \times K}$ are then constructed from H by taking the median over each cell type. Finally, the estimator of cell type compositions for the bulk data is then given by

$$\widehat{B} = \underset{B \geq 0}{argmin} ||P - \widetilde{H}B||_F \tag{4.8}$$

Altering the classical assumption of an additive error linear model, SpatialDecon Danaher et al. (2022) implements a non-negative linear regression-based method that assumes a log-normal multiplicative error model. The log-normal error model is given by

$$log(_{i\cdot}) = log(\widetilde{X}_{i\cdot}^T B) + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_n) \text{ and } B \in \mathbb{R}^{K \times n} \tag{4.9}$$

The estimator of cell type compositions for the n cell-pools is then given by

$$\widehat{B} = \underset{B \geq 0}{argmin} \big|\big| log(Y) - log(\widetilde{X}^T B) \big|\big|_2 \tag{4.10}$$

One of the first methods to incorporate spatial information in the deconvolution spatial transcriptomics data is Conditional AutoRegressive Model-based Deconvolution (CARD) Ma and Zhou (2022). CARD applies a conditional autoregressive (CAR) assumption on the coefficients of the classical non-negative linear model between the bulk expression Y and a cell-profile matrix $\widetilde{X}$. The linear model is given by

$$Y = \widetilde{X}B + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_e^2 I_n) \tag{4.11}$$

The CAR assumption then incorporates 2D spatial information $S \in \mathbb{R}^{n\times 2}$ through an intrinsic prior on the cell type compositions (the model coefficients) by modeling compositions in each location as a weighted combination of compositions in all other locations. This modeling assumption is given by

$$B_{ki} = b_k + \phi \sum_{j=1, j\neq i}^{n} W_{ij}(B_{kj} - b_k) + \epsilon_{ki}, \epsilon_{ki} \sim \mathcal{N}(0, \sigma_{ki}^2) \tag{4.12}$$

where the weights W_{ij} are given by the Gaussian kernel

$$W_{ij} = K_G(s_i, s_j; \sigma^2) = \exp(-\frac{||s_i - s_j||_2^2}{2\sigma^2}) \tag{4.13}$$

with default scaling parameter $\sigma^2 = 0.1$. CARD then estimates the cell type composition of the spatial transcriptomic data through constrained maximum likelihood estimation.

Some recent developments in cell type deconvolution have applied deep learning-based methods. These approaches typically apply a transfer learning scheme wherein they first simulate bulk data from scRNA reference data, and use a common network to predict the cell type composition of both the simulated and real bulk data. A notable feature of the deep learning-based methods is that they model the cell type compositions directly, i.e. the model objective is on the predicted cell type compositions. This contrasts with most classical methods, where the predicted cell type proportions are the optimal parameters/coefficients from some regression model.

One of the early deep learning approaches to the cell type deconvolution problem is Scaden Menden et al. (2020). First, scRNA reference data is randomly sampled from scRNA

reference data to generate simulated mixed-cell samples. A fully-connected network is then trained to predict the true cell type compositions of the simulated bulk data, with cross-entropy loss function. This trained model is then applied to the real bulk data to get cell type compositions. Building on this approach, DSTG Song and Su (2021b) is a Graphical Neural Network (GNN) based method, modeling similarities in expression between different bulk samples. First, the pseudo bulk expression data is generated taking n_p random samples (with replacement) of 2 to 8 cells from the scRNA-seq reference, and aggregating their UMI counts, downsampling to adjust for realistic bulk UMI counts. The pseudo and real bulk data are then aligned in a lower dimensional ($S < d$) gene-space using Canonical Correlation Analysis (CCA). The projections to the $s = 1, 2, ..., S$ dimensions are given by the canonical variables

$$\begin{aligned} U_s &= \widetilde{X}\mu_s^* \\ V_s &= X\nu_s^* \end{aligned} \tag{4.14}$$

where

$$\mu_s^*, \nu_s^* = \underset{\mu_s, \nu_s \in \mathbb{R}^d}{argmax}\{\nu_s^T \widetilde{X}^T X \nu_s\} \text{ s.t. } U_s^T U_{s'} = V_s^T V_{s'} = \delta_{ss'} \tag{4.15}$$

are the canonical correlation vector pairs. These embeddings are then used to construct a graph by considering Mutual Nearest Neighbors (MNN) as adjacent in the graph. That is, given a pair of sample cell-pools i, j, we let

$$A_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are mutual nearest neighbors} \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

Here, adjacencies can be between simulated-to-real and real-to-real samples. With $X_{in} = [\widetilde{X}X] \in \mathbb{R}^{d \times N}$ ($N = n_p + n$) and the normalized adjacency matrix $\widetilde{A}$ as input, the $L \geq 1$ (default 1) graph convolution (GCN) layers of the DSTG model are given by

$$\begin{aligned} H^{(0)} &= X_{in} \\ H^{(l)} &= \text{ReLU}(\widetilde{A}H^{(l-1)}W^{(l)}) \text{ for } l \in [1, L] \end{aligned} \tag{4.17}$$

where $W^{(l)}$ is the weight matrix for the l_{th} layer. The output of the DSTG model is the predicted composition of K cell-types, given by

$$\begin{bmatrix} \widehat{B}_p \\ \widehat{B} \end{bmatrix} = \text{softmax}(\widetilde{A}H^{(L)}W) \in \mathbb{R}^{n \times K} \tag{4.18}$$

where $\widehat{B}_p$ and $\widehat{B}$ are the predictions for the pseudo and real cell-pools, respectively. The loss function is then defined as the cross-entropy between the predicted and true cell-type compositions of the pseudo cell-pools

$$\mathcal{L} = -\sum_{i=1}^{n_p}\sum_{k=1}^{K} y_{i,k}^{(p)} ln(\hat{y}_{i,k}^{(p)}) \tag{4.19}$$

where $\hat{y}_{i,k}^{(p)}$ and $y_{i,k}^{(p)}$ are the predicted and true composition of cell-type k in the i_{th} pseudo cell-pool.

I have summarized a list of existing cell type deconvolution methods in Table 4.1, and in Table 4.2 I highlight some select benchmark sets in table. Note that this benchmark collection was made prior to our development of large-scale benchmark data.

Table 4.1 A summary of tools for cell type deconvolution.

Tool	Algorithm	Description	Language	Availability
NMFReg	Classical	A non-negative matrix factorization of an annotated scRNA reference matrix	Matlab, Python	NMFReg Rodriques et al. (2019a) NMFReg-Python
SPOTlight	Classical	Extension of NMFReg, with non-negative matrix factorization applied to both the scRNA reference matrix, and the bulk expression matrix	R, Python	SPOTlight Elosua-Bayes et al. (2021b) Dance Ding et al. (2022)
DWLS	Classical	Weighted NNLS; dampened weighting is applied to genes	R	DWLS Tsoucas et al. (2019)
SpatialDWLS	Classical	A subset of cell types chosen via PAGE enrichment analysis	R	SpatialDWLS Dong and Yuan (2021)
SpatialDecon	Classical	A multiplicative log-normal error model	R, Python	SpatialDecon Danaher et al. (2022) Dance Ding et al. (2022)
cell2location	Variational Inference	Bayesian hierarchical model of spatial expression counts with a spatially informed prior on cell type compositions	Python	cell2location Kleshchevnikov et al. (2022b)
CARD	Classical	Conditional autoregressive based model that incorporates spatial correlation of cell type compostion	R, Python	CARD Ma and Zhou (2022) Dance Ding et al. (2022)
RNA-Sieve	Classical	A likelihood based inference model that estimates cell type proportion through a maximum-likelihood method	Python	RNA-Sieve Dan D. Erdmann-Pham and Song (2021)
Scaden	GNN	A fully-connected network that is trained on simulated bulk data, and used to predict cell type compositions of real bulk data	Python	Scaden Menden et al. (2020)
DSTG	GNN	A graph convolutional network whose graph is constructed on Mutual Nearest Neighbors of low-dimensional embeddings of simulated and real bulk data	R, Python	DSTG Song and Su (2021b) Dance Ding et al. (2022)

4.4 Cell Type Deconvolution Benchmark Dataset and Model Development

As mentioned in the background chapter, section Spatial & Bulk Deconvolution, there is a lack for quality benchmark data for cell type deconvolution. Again, such datasets must be

Table 4.2 A summary of datasets for cell-type deconvolution.

Dataset	Species	Tissue	Dataset Dimensions	Protocol	Availability
Mouse Posterior Brain 10x Visium Data	Mouse	Posterior brain	3,353 spots 31,053 genes	10X Visium	MPB10xV lin (d)
Mouse Olfactory Bulb	Mouse	Olfactory bulb	1,185 spots 11,176 genes	10X Visium	MOB10xV lin (c)
HEK293T and CCRF-CEM cell line mixture	Human		56 mixtures 1,414 genes	NanoString GeoMx	CelllineGeoMx lin (a)
Human PDAC	Human	Pancreas	1,819 spots 19,738 genes	Spatial Transcriptomics	HPdacST lin (b)

generated through either synthetic mixture experiments or some form of random sampling from a reference scRNA-seq dataset. In either case, most datasets are limited in size, and/or do not reflect real conditions. Towards this end, we develop a method to generate large yet realistic cell type deconvolution benchmark datasets, from which we have generated a human tumor microenvironment dataset consisting of *1.8 million cells*.

Additionally, we build on ideas from DSTG and develop a spatially informed Graph Neural Network based method, GNNDECONVOLVER. Here, we build the model framework to incorporate spatial information, if it is available. Prior to this, only a small set of classical-shallow methods have incorporated spatial information, such as CARD. With this method, we can leverage reference scRNA-seq data with and without spatial information, to infer cell type compositions of bulk mixtures with and without spatial information.

To validate GNNDECONVOLVER, we carry out a compilation of experiments on the large-scale benchmark dataset we've generated. In this benchmarking, we will see that GNNDECONVOLVER performs strongly against a set of existing state-of-the-art methods. For fairness, we have included methods that incorporate spatial information.

An outcome of this method to generate cell type deconvolution benchmark datasets is an open tool that takes single-cell resolution spatial trancriptomics data and generates synthetic mixtures of varying size. This tool accepts data from many popular spatial transcriptomic platforms, such as 10x Visium, MERFISH, and sci-Space.

Here, we developed cell type deconvolution benchmarking datasets that are larger in scope and quality than current datasets. In terms of quality, most benchmark datasets for

Table 4.3 A summary of datasets for cell-type deconvolution.

Tissue sections	8
Cells	771,236
Genes	980
Fields of view (FOV)	239
FOV size	$\sim 984.96\mu m \times 656.64\mu m$

cell type deconvolution are created through random sampling of scRNA-seq data, wherein the number of sampled cells is randomly chosen within some range that matches typical ranges found in a given spatial transcriptomics platform. This sampling process lacks spatial context, as spatial context is lost with scRNA-seq methods. We used single-cell resolution spatial transcriptomics datasets generated by the CosMx Spatial Molecular Imager (SMI) to create benchmark datasets with preserved spatial context and large sample size. However, while SMI is high-throughput (up to nearly 1 million cells), it has relatively low multi-plex capability (can target around 1,000 genes and 100 proteins per panel) He et al. (2022).

Data from the CosMx Spatial Molecular Imager (SMI) consists of transcriptomic, cell type annotations, spatial, histology images, and some protein data. Cell type annotations are made from a negative binomial likelihood model with the mean given by cell type reference profiles, bias added from expected background, and a large size parameter (default 10) to account for overdispersion due to technical sources of variance. This is the model given in 2.2, with detection scale factor set to 1.

4.4.1 Non-small cell lung cancer tumors

For this experiment, we used the NanoString CosMx open-source non-small-cell lung cancer (NSCLC)dataset. This dataset consists of transcriptomic, spatial, and histology image data for 8 samples of 5 Non-small cell lung cancer (NSCLC) tumors. A data summary is given in Table 4.3. The cell type compositions of each sample can be seen in physical space, and in gene space through gene expression UMAP projections, which can be seen in Figure 4.1.

To generate pseudo-spot data, we divided each FOV spatially into a uniform grid (see

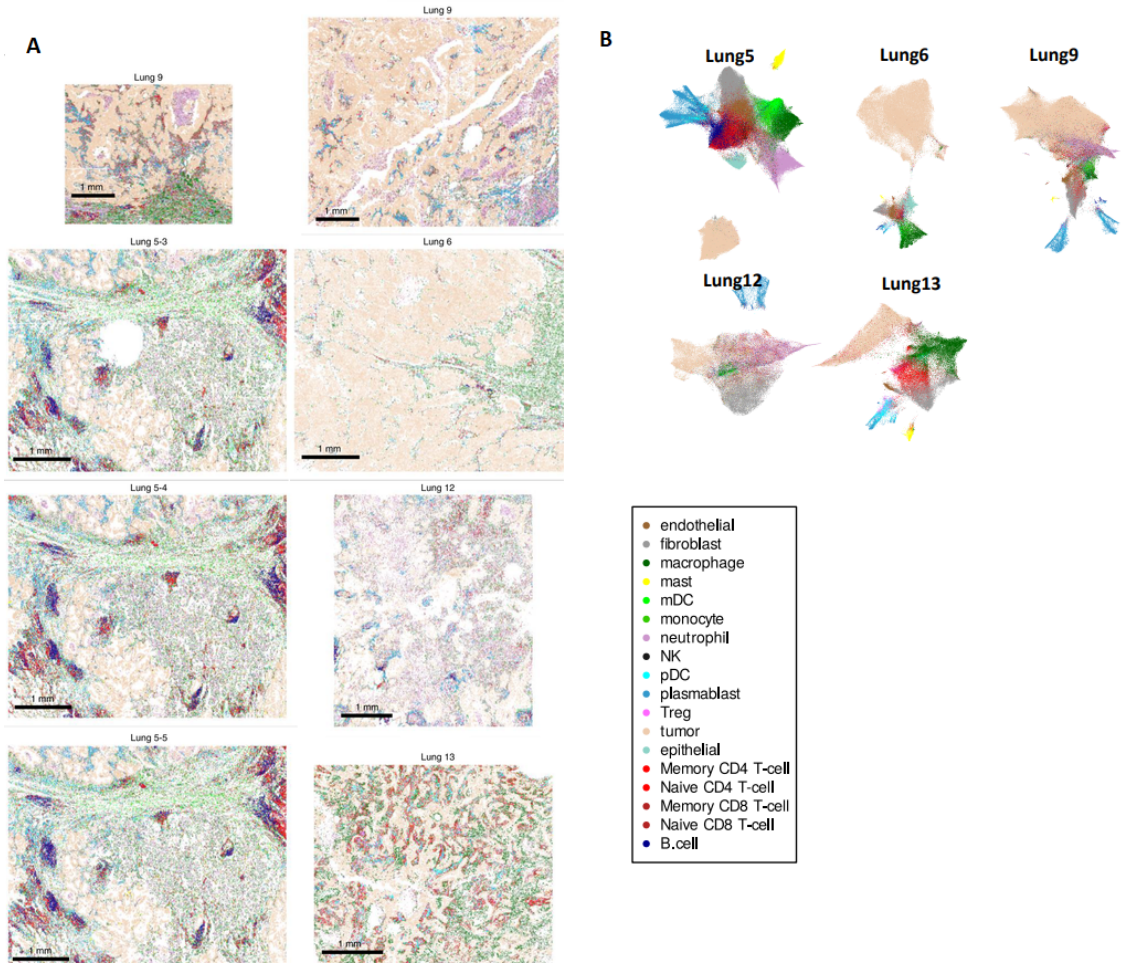


Figure 4.1 Composition of CosMx Lung Samples in A. physical space, and B. gene space through gene expression UMAP projections. Figures from He et al. (2022).

4.3,). The uniform spot (grid rectangle) sizes were chosen to cover an area of $37456.28\mu m^2$, which is the mean area of spots in another NSCLC spatial transcriptomics dataset from Nanostring's GeoMx platform (not single-cell resolution). The ultimate purpose of this is to test these pseudo-spot data as a reference for deconvoltion of the GeoMx generated data. To allow for spatial context in the pseudo spots, we simply use spot centers as the coordinates of each spot. Figures 4.2- 4.3 illustrate this two step process of applying an FOV grid on the tissue, and a second layer grid on each FOV.

In addition to this benchmark dataset, we developed a GNN-based model by modifying and building upon ideas from the GNN-based model DSTG Song and Su (2021b) that is detailed in the Survey of Methods subsection. An important change we made was in the graph

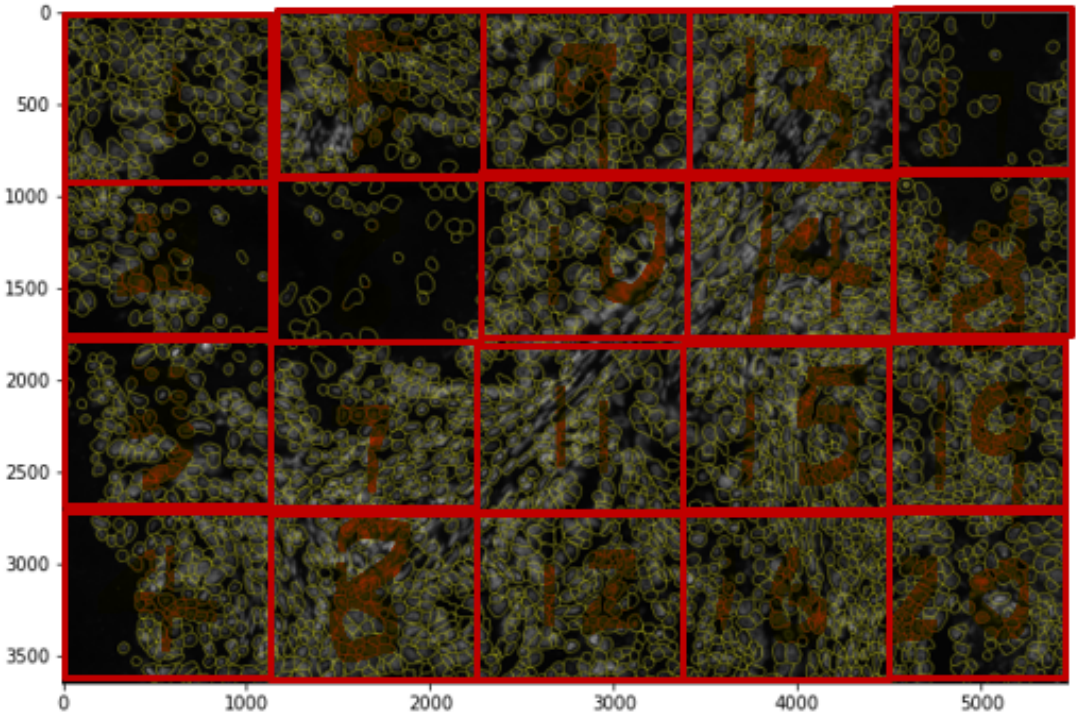


Figure 4.2 FOV grid overlaying tissue sample.

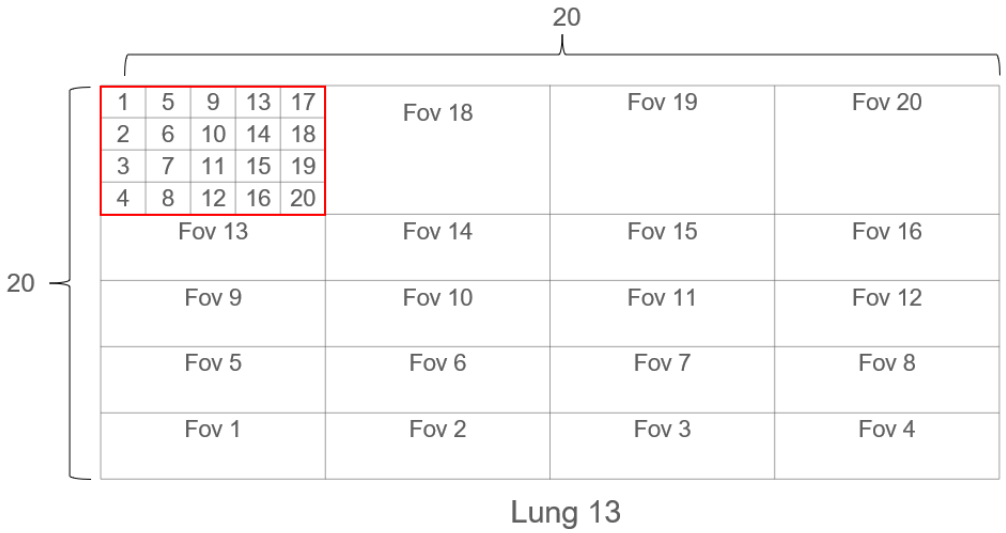


Figure 4.3 Layout of pseudo-spots as a grids over each FOV.

- **Train, valid, test: lung 5-1,2,3**

		Our GNN		SpatialDecon		NNLS		CARD(no spatial info)		CARD(with spatial info)	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
18 cell types	Infer Lung 5-1	0.0027	0.027	0.0021	0.0273	0.005	0.0387	0.0027	0.0308	0.0022	0.027
	Infer Lung 5-2	0.0028	0.027	0.0025	0.0285	0.004	0.0351	0.0026	0.0313	0.0022	0.0284
	Infer Lung 5-3	0.003	0.029	0.0021	0.0279	0.009	0.0516	0.0022	0.0275	0.0018	0.0262
17 cell types	Infer Lung 5-1	0.0028	0.031	0.0072	0.0413	0.0103	0.0512	0.01	0.0477	0.0086	0.0441
	Infer Lung 5-2	0.003	0.031	0.0069	0.0416	0.0075	0.0449	0.0093	0.0477	0.0082	0.0447
	Infer Lung 5-3	0.0028	0.031	0.0069	0.0426	0.0153	0.065	0.0078	0.0442	0.0067	0.0407

- **Train, valid: lung 5-1,2,3, test: lung 12**

		Our GNN		SpatialDecon		NNLS		CARD(no spatial info)		CARD(with spatial info)	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
18 cell types	Infer Lung 12	0.0199	0.068	0.0065	0.0432	0.0138	0.0521	0.0089	0.0545	0.0088	0.0536
17 cell types	Infer Lung 12	0.0077	0.049	0.01145	0.0531	0.0115	0.0487	0.0159	0.0707	0.0158	0.0707

- **Train, valid: lung 5-1,2,3, test: lung 13**

		Our GNN		SpatialDecon		NNLS		CARD(no spatial info)		CARD(with spatial info)	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
18 cell types	Infer Lung 13	0.0102	0.056	0.0112	0.0579	0.0354	0.0922	0.0102	0.0622	0.0116	0.0678
17 cell types	Infer Lung 13	0.0162	0.071	0.0123	0.0631	0.0339	0.0894	0.0129	0.0726	0.0122	0.07

Figure 4.4 Cell type deconvolution benchmark data results - NSCLC tumors.

construction. DSTG only uses CCA embedded expression data to define adjacencies through Mutual-Nearest-Neighbors, and does not allow for the integration of spatial information. Our model incorporates spatial information by defining graph adjacencies from expression data and spatial location.

We used these benchmark pseudo-spot data to validate our model. The common experimental design we use is to one or more samples as references (training and/or validation sets), and one sample as the query (test set). We used both mean squared error (MSE) and mean absolute error as evaluation metrics. Our GNN performed best in the majority of experiments, in both the 18 cell type and 17 cell type settings. See Figure 4.4 for a breakdown of the experimental settings, and the performance results.

4.5 Spot Data and Deconvolution Methodology

Here, we build on the preliminary developments of spot data generation and the graph-based deconvolution scheme proposed.

4.5.1 Dataset

The single-cell resolution spatial transcriptomic data we used is from the CosMx platform (He et al., 2021) by Nanostring, which uses a spatial molecular imaging technique. Through our collaboration with Nanostring, we collected 20 samples from human tissue in lung, kidney, and liver. All the samples contain tumor micro environments. Each dataset was generated from 960-to-1000-plex CosMx RNA panel run on CosMx SMI. Here we describe the data from each tissue in detail.

Human Lung. This dataset consists of 8 samples over 5 NSCLC (non-small cell lung cancer) tissues. The resulting dataset contains measurements from 960 targets over 800,327 cells, of which 766,313 cells are analyzed. In more detail, 259,604,214 transcripts are detected. In these samples, the cells were experimentally labeled (by CosMx) from 18 detected cell types.

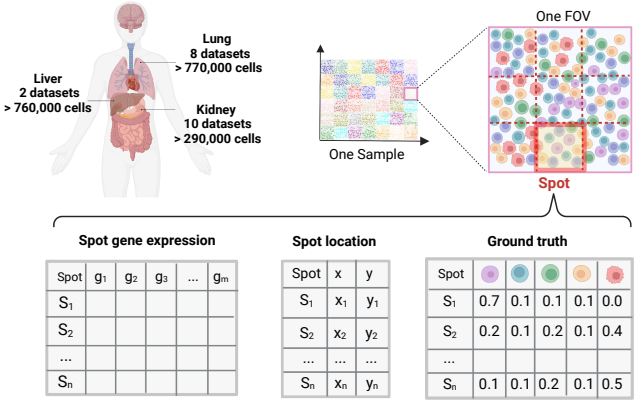
Human Kidney. This dataset consists of 10 samples of tissue taken from lupus nephritis patients, via kidney core biopsy. The resulting dataset contains nearly 300,000 cells.

Human Liver. This dataset had subcellular resolution, and consists of 1,000 genes over 800,000 cells. These samples cover a 180 mm^2 area of liver tissue, from 1 normal liver and 1 hepatocellular carcinoma tissues.

4.5.2 Pseudo spot generation

As in the pseudo spot generation process described in the previous section, we impose a grid on the single-cell resolution spatial transcriptomics data. We choose a spot size to yield multiple cells per spot, with an average size similar to lower resolution spatial transcriptomics methods. We then have the cell type compositions of these pseudo spots, but in a realistic setting since we are generating them within their spatial context of the tissue samples. We assign the cells to pseudo spots based on the centroid coordinates of each cell. That is, the cell gets assigned to the pseudo spot that contains its centroid within its defined boundary. Interestingly, we did not have any cells whose centroid sat on a boundar line, so we did not deal with that assignment problem. Nonetheless, we could randomly assign a cell living on a boundary line to a single pseudo spot forming that boundary. We take the expression

a



Spot	g_1	g_2	g_3	...	g_m
S_1					
S_2					
...					
S_n					

Spot	x	y
S_1	x_1	y_1
S_2	x_2	y_2
...	...	...
S_n	x_n	y_n

Spot					
S_1	0.7	0.1	0.1	0.1	0.0
S_2	0.2	0.1	0.2	0.1	0.4
...					
S_n	0.1	0.1	0.2	0.1	0.5

b

	Sample	# FOV	# Cells	# Spots	# Average Cells /Spot	# Cell Types
Lung Cancer	Lung 5-1	30	98,002	600	163	18
	Lung 5-2	30	105,800	600	176	18
	Lung 5-3	30	97,809	600	163	18
	Lung 6	30	89,975	600	150	18
	Lung 9-1	20	87,606	400	219	18
	Lung 9-2	45	139,504	900	155	18
	Lung 12	28	71,304	560	127	18
	Lung 13	20	81,236	400	203	18
Kidney Cancer	Run1087_SP19_4061	8	16,474	160	102	38
	Run1087_SP18_8471	8	13,455	160	84	37
	Run1087_SP20_1098	17	32,534	340	95	38
	Run1087_SP17_8693	16	19,616	320	61	37
	Run1081_SP17_2566	13	28,880	260	111	38
	Run1081_SP18_3323	12	28,000	240	116	38
	Run1080_SP21_213	13	49,500	260	190	35
	Run1080_SP19_1139	18	61,073	360	169	36
	Run1080_SP20_10838	11	34,902	220	158	38
	Run1080_SP20_642	7	12,404	140	88	35
Liver Cancer	Normal Liver	353	312,691	2,520	124	19
	Hepatocellular Carcinoma	365	447,815	3,276	136	17

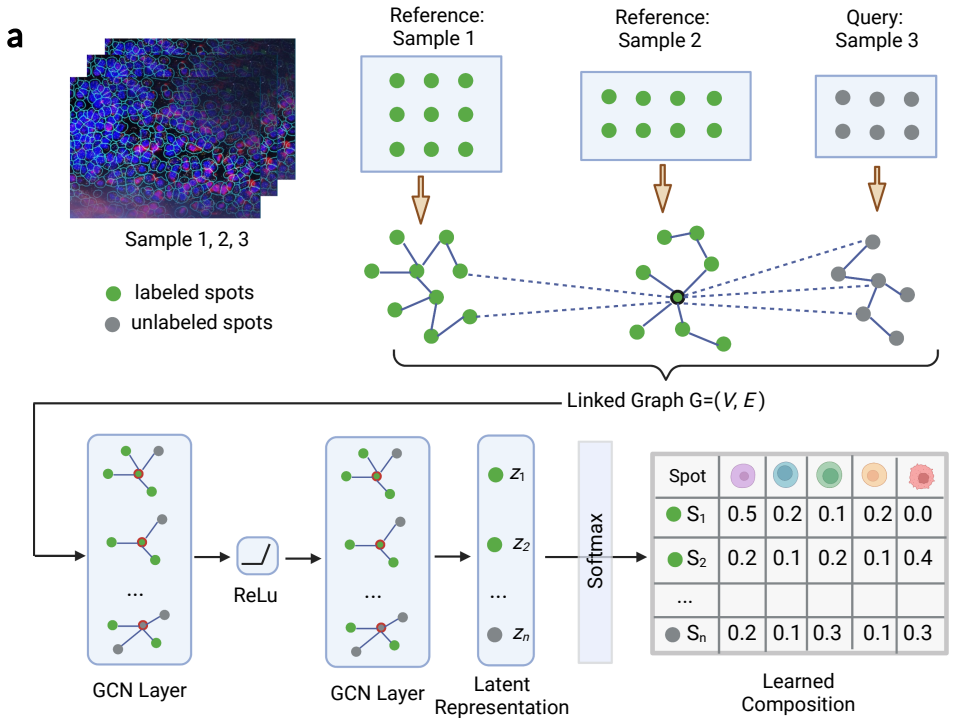
Figure 4.5 An Overview of SPATIALCTD. (**a**). The method for generating the SPATIALCTD dataset. SPATIALCTD comprises three distinct human tissues, namely, the lung, kidney, and liver. For each sample in tissues, SPATIALCTD consists of a spot gene expression file, a spot location file, and a ground truth file. (**b**). A summary of SPATIALCTD.

measure of the pseudo spots by aggregating the expression over the cells within the pseudo spot. One thing to note is we are operating under the simple assumption that expression from bulk mixtures is the sum of the expression of the cells within the mixture. This may not be the case, indeed some studies have shown a log-normal aggregation, but it is a widely used assumption (hence NMF based methods) and a simple starting point. We then compute the cell type composition of the pseudo spots from the cell types of the cells within the pseudo spots, which serves as the ground truth labels of the pseudo spots. We define the spatial location of the spots from their centroid coordinates. This process generates the following data: (1) number of cells in each spot, (2) cell ID, spot ID, and their mapping that defines the cell-to-spot assignment, (3) spatial location of each spot, (4) spot level gene expression, and (5) the ground truth cell type compositions. This procedure has spot size as a parameter, which we set to a realistic size seen in Nanostring's GeoMx platform, which is lower resolution (bulk mixtures). Here, we aimed for spot sizes near the mean and/or median spot size taken by GeoMx, 37456.28 μm^2, and 24168.74 μm^2 respectively.

Human Lung. We set the FOV accordingly: 5,472 pixels * 3,648 pixels, 0.18 μm per pixel. Dividing each FOV into 20 pseudo spots, we get a spot area 32338.2067 μm^2, which is within the mean and median GeoMx spot size. Lastly, we filtered out low quality (spots without cells), resulting in a total of 4,660 spots over the 8 samples, and 771,236 cells. In 4.5 we outline the generation procedure and give a summary of the generated datasets.

Human Kidney. We set the FOV accordingly: 5,472 pixels * 3,648 pixels, 0.18 μm per pixel. Dividing each FOV into 20 pseudo spots, we get a spot area 32338.2067 μm^2. After filtering, we have 2,460 spots over 10 samples, consisting of 296,838 cells.

Human Liver. We set the FOV accordingly: 4,236 pixels * 4,236 pixels, 0.12 μm per pixel. Dividing each FOV into 9 pseudo spots, we get a spot area 28709.9136 μm^2. After filtering, we have 5,796 spots over 2 samples, consisting of 760,506 cells.



b

Setting	Reference		Query
Setting 1	Patient 1 Sample 1	Patient 1 Sample 2	Patient 1 Sample 3
Setting 2	Patient 1 Sample 1	Patient 1 Sample 2	Patient 2 Sample 1
Setting 3	Patient 1 Sample 1	Patient 2 Sample 1	Patient 1 Sample 2
Setting 4	Patient 1 Sample 1	Patient 2 Sample 1	Patient 3 Sample 1

Figure 4.6 An overview of GNNDECONVOLVER and experimental settings. **a**. An overview of GNNDECONVOLVER. Note that reference refers to the training data with known cell type proportion labels, and query indicates test data the model will predict. **b**. Four types of experimental settings.

4.5.3 GNNDECONVOLVER

To begin, let's assume we have t samples which measure expression of d genes over n_1, n_2,..., n_t spots. We treat the spots as graph nodes, constructing the graph from their log-normalized expression values (Lytal et al., 2020). The nodes in each sample are then connected according to both expression level and spatial distance. To do this, let A_{spatial} and A_{gene} be the adjacency matrices of the distances and expression respectively. We define A_{spatial} by nearest-neighbors, with $K = 5$. Meaning, each spot is connected to its 5 nearest nodes (spatially). We also considered defining this by specifying a distance threshold. We apply the same construction setting for A_{gene}, except we define distance here with cosine similarity between expression levels. We then define the final adjacency matrix A_{sample} as a weighted sum of these adjaceny matrices $A_{\text{sample}} = \alpha A_{\text{spatial}} + \beta A_{\text{gene}}$. We experimentally tested and set where α and β to 0.3 and 0.7 respectively. Then, for each sample t the graph is defined by $A_{\text{sample}_t} \in R^{n_t \times n_t}$ and $X_{\text{sample}_t} \in R^{n_t \times d}$ where A_{sample_t} is the final adjacency matrix, and X_{sample_t} is the node representation matrix.

We also connect nodes between different samples, but the spatial context is only within samples so we contruct the between sample graphs via gene expression only. First, we compute expression similarity of nodes between each sample. We define the adjacency matrix using a nearest neighbors scheme, again with $K = 5$. This yields a graph that connects nodes across all t samples, both labeled and unlabeled (cell type compositions). The between sample graph is then given by $A_{\text{all}} \in R^{n \times n}$ and $X_{\text{all}} \in R^{n \times d}$ where $n = n_1 + n_2 + \cdots + n_t$.

This graph construction yields a linked graph $G = (V, E)$. The task of GNNDECONVOLVER is to predict cell type compositions of unlabeled spots, with both spot features and the graph features defined by the graph of node connections between labeled and unlabeled spots. Namely, we have the input as $[AX]$, where $A \in R^{n \times n}$ is the adjacency matrix, and $X \in R^{n \times d}$ is the node representation matrix. Again, we are have n spots with expression measurements over d genes.

GNNDECONVOLVER consists of two graph convolutional layers, where the second layer

is treated as the output layer, i.e. no activation function is applied. These layers are define accordingly:

$$H^{(l+1)} = \sigma\left(\tilde{A}H^{(l)}W^{(l)}\right) = \mathrm{ReLU}\left(\tilde{A}H^{(l)}W^{(l)}\right) \tag{4.20}$$

where $\tilde{\mathbf{A}} = \tilde{\mathbf{D}}^{-1/2}(\mathbf{A}+\mathbf{I})\tilde{\mathbf{D}}^{-1/2}$ with $\tilde{\mathbf{D}}$ the diagonal matrix of $\mathbf{A}+\mathbf{I}$ and $\mathbf{I}$ the identity matrix. $H^{(l)}$ is the input from the previous layer. $W^{(l)}$ is the weight matrix of the l-th layer. $\mathrm{ReLU}(\cdot)$ is the nonlinear activation function. Here, the input for the first layer would be the original node representation $\mathrm{H}^{(0)} = \mathrm{X}$.

We can define GNNDECONVOLVER as the following composition:

$$\hat{Y} = \tilde{A}\,\mathrm{ReLU}\left(\tilde{A}X^{T}W^{(0)}\right)W^{(1)} \tag{4.21}$$

where $W^{(0)}$ and $W^{(1)}$ are learned weight matrices, and $\hat{Y}$ is the predicted cell type compositions with F unique cell types. The loss function is defined as the cross-entropy between ground truth and predicted cell type composition:

$$\mathcal{L} = -\sum_{i=1}^{n_q}\sum_{f=1}^{F} y_{i,f}\ln\left(\hat{y}_{i,f}\right) \tag{4.22}$$

Here, we have n_q labeled nodes, $\hat{y}_{i,f}$ and $y_{i,f}$ represent the predicted and ground truth cell type proportion of cell type f in spot i, respectively. We train the model by minimizing the cross-entropy L on training sets via stochastic gradient descent using backpropogation.

4.5.4 Results

Here we test GNNDECONVOLVER against 8 other deconvolution methods, and compare our generated dataset with other synthetic bulk mixture data. We see that GNNDECONVOLVER outperforms all other methods in each evaluation metric. We continue to see this trend as we vary the spot size. Refer back to Figure 4.6 for an overview of the model and experimental setup.

These results show GNNDECONVOLVER to be a useful deconvolution method, and the formative ideas may help guide future method developments. Particulary, the to integrate reference scRNA-seq data with spatial transcriptomics data. We also see that this

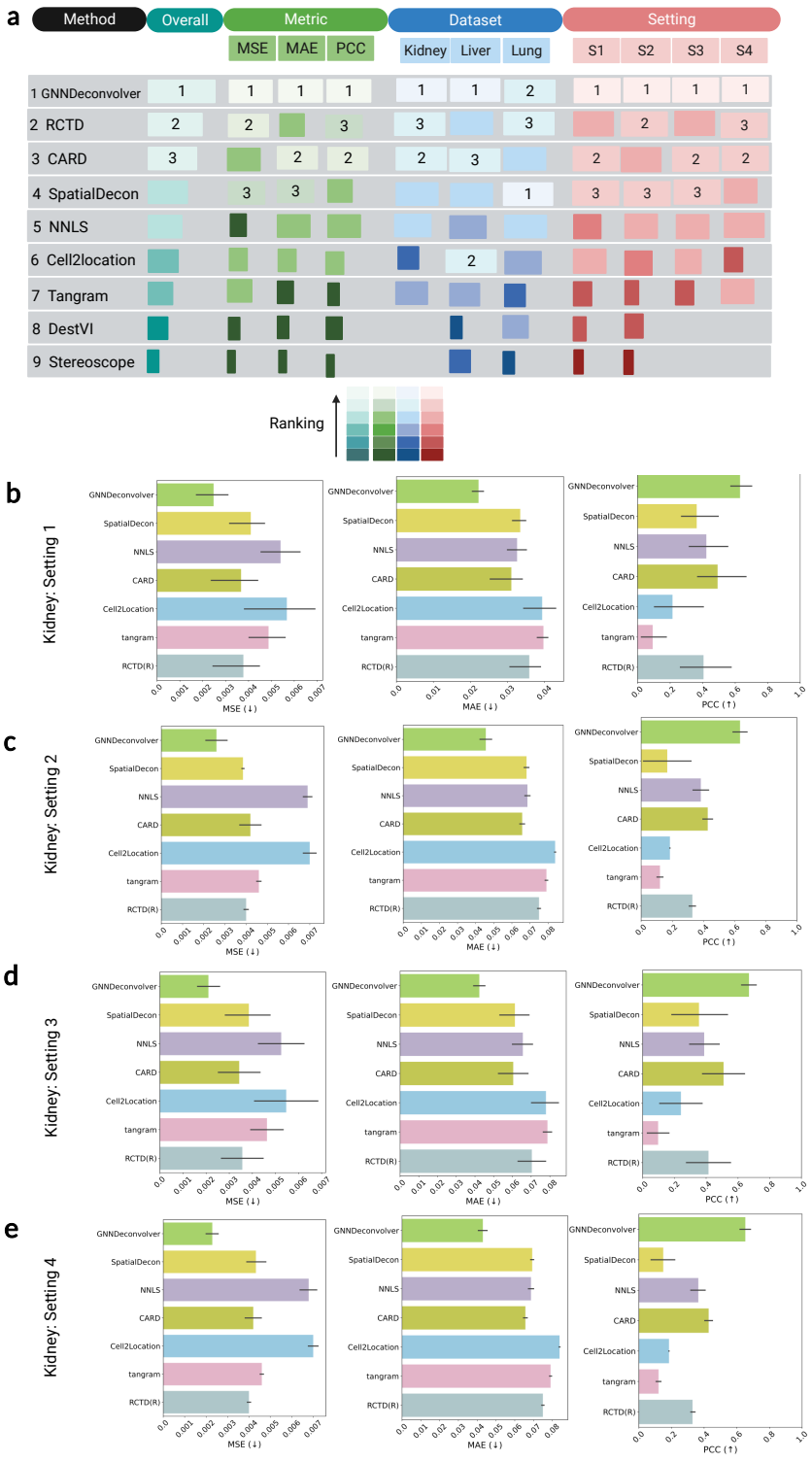


Figure 4.7 Performance of 9 methods in cell type deconvolution. **a**. A summary of results for 9 cell type deconvolution methods. **b-e**. Comparison of the models under different settings on SPATIALCTD kidney tissue in terms of MSE, MAE and PCC.

a

Dataset	# Cells	# Cell Types	Spot Image	Cellular Location	Subcellular Location	Cell Composition	Human	TME
Mouse Embryo	17k	20	✗	✓	✗	✓	✗	✗
MPOA	59k	16	✗	✓	✗	✓	✗	✗
Mouse Brain	4.7k	15	✗	✗	✗	✗	✗	✗
Mouse Cortex	14k	23	✗	✗	✗	✓	✗	✗
Mouse Visual Cortex	14k	15	✗	✗	✗	✓	✗	✗
Simulated ST for Human Lung	60k	6	✗	✗	✗	✗	✓	✗
Simulated ST for Human Liver	29k	9-19	✗	✗	✗	✗	✓	✗
Mouse Posterior Brain	NA	23	✓	✗	✗	✗	✗	✗
Mouse Olfactory Bulb	NA	NA	✓	✗	✗	✗	✗	✗
HEK293T & CCRF-CEM	NA	2	✗	✗	✗	✓	✓	✗
Human PDAC	NA	20	✗	✗	✗	✗	✓	✗
SPOTlight Synthetic	2k-8k	8	✗	✗	✗	✓	✗	✗
Our Human Lung	**771k**	18	✓	✓	✓	✓	✓	✓
Our Human Kidney	**296k**	35-38	✓	✓	✓	✓	✓	✓
Our Human Liver	**760k**	17-19	✓	✓	✓	✓	✓	✓

Figure 4.8 Statistical comparison between SPATIALCTD and existing cell type deconvolution benchmark datasets.

pseudo spot generation procedure can provide us with more realistic cell type deconvolution benchmark data, at a large scale.

CHAPTER 5

FURTHER EXPLORATIONS IN DECONVOLUTION

5.1 Towards a Probabilistic Framework for Deconvolution

Let C be the random variable of observing a single cell of type $[1, K] = \{1, 2, ..., K\}$, with distribution

$$C \sim Categorical_K(\mathbf{w}),\ \ p(C = k; \mathbf{w}) = w_k,\ \ \mathbf{w} \sim \pi_w \tag{5.1}$$

Let $\mathbf{X}$ be the random variable of a single-cell's gene expression for D genes, and suppose

$$p_k(\mathbf{x}; \theta_k) = p(\mathbf{X} = \mathbf{x}|C = k),\ \ \boldsymbol{\theta} = (\theta_1, ..., \theta_K) \sim \pi_\theta \tag{5.2}$$

Then $\mathbf{X} \sim f$ where f is the mixture density

$$f(\mathbf{x}; \boldsymbol{\theta}, \mathbf{w}) = \sum_{k=1}^{K} w_k p_k(\mathbf{x}; \theta_k) \tag{5.3}$$

To simplify notation, for $k \in [1, K]$ we let

$$\begin{gathered} \mathbf{X}_k = \mathbf{X}|C = k \\ \boldsymbol{\mu}_k = \mathbb{E}(\mathbf{X}_k),\ \ \boldsymbol{\Sigma}_k = cov(\mathbf{X}_k) \end{gathered} \tag{5.4}$$

Suppose now that we randomly sample n cells $C^{(1)}, ..., C^{(n)} \overset{iid}{\sim} Categorical_K(\mathbf{w})$, then the total number for each of the K cell-types are given by the random variable

$$\begin{gathered} \mathbf{N} = (N_1, ..., N_k) \sim Multinomial_K(n, \mathbf{w}) \\ p(n_1, ..., n_K; n, \mathbf{w}) = \frac{n!}{n_1! \cdots n_K!} \prod_{k=1}^{K} w_k^{n_k},\ \text{for } (n_1, ..., n_K)/n \in \Delta^{K-1} \end{gathered} \tag{5.5}$$

Suppose we measure the expression $\mathbf{X}_k^{(1)}, ..., \mathbf{X}_k^{(N_k)} \overset{iid}{\sim} p_k, k \in [1, K]$ and aggregate over each cell-type

$$\mathbf{Y} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \mathbf{X}_k^{(i)} = \sum_{k=1}^{K} N_k \mathbf{H}_k,\ \text{where } \mathbf{H}_k = \overline{\mathbf{X}}_k \tag{5.6}$$

Letting $B_k = \frac{N_k}{n}$ and $B = (B_1, ..., B_K) \in \Delta^{K-1}$, taking the sample mean we get

$$\frac{\mathbf{Y}}{n} = \sum_{k=1}^{K} \frac{N_k}{n} \mathbf{H}_k = \sum_{k=1}^{K} B_k \mathbf{H}_k \tag{5.7}$$

Note that $\mathbb{E}_{p_k}(\mathbf{H}_k) = \boldsymbol{\mu}_k$ and $cov_{p_k}(\mathbf{H}_k) = \frac{1}{N_k}\boldsymbol{\Sigma}_k$, so

$$\mathbb{E}(\mathbf{Y}) = \sum_{k=1}^{K} N_k \boldsymbol{\mu}_k,\ cov(\mathbf{Y}) = \sum_{k=1}^{K} N_k \boldsymbol{\Sigma}_k \tag{5.8}$$

and

$$\mathbb{E}\left(\frac{\mathbf{Y}}{n}\right) = \sum_{k=1}^{K} B_k \boldsymbol{\mu}_k,\ cov\left(\frac{\mathbf{Y}}{n}\right) = \sum_{k=1}^{K} \frac{B_k}{n} \boldsymbol{\Sigma}_k \tag{5.9}$$

If we let $\mathbf{Y} = \sum i = 1^n \mathbf{X}^{(i)}$, where $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(n)} \overset{iid}{\sim} f$, then the distribution of $\mathbf{Y}$ is the $n-$fold convolution of f:

$$f_Y(y) = (f * f * \cdots * f * f)(y) = f^{*n}(y) \tag{5.10}$$

Also, $n_k\mathbf{H}_k \sim p_k^{*n_k}$, and $\mathbf{Y} = \sum_{k=1}^{K} n_k\mathbf{H}_k \sim (p_1^{*n_1} * p_2^{*n_2} * \cdots * p_{K-1}^{*n_{K-1}} * p_K^{*n_K})(y)$

5.2 Learning Cell Profiles

Cell type expression profiles have played a key role in cell type deconvolution, as they are what the most basic methods are built on. In chapter 2, section 2, we have shown a standard way of constructing cell type expression profiles from reference scRNA-seq data. These methods are mostly rule based, where we take some normalized reference scRNA-seq data, account for background, and take the median or mean. A direction I thought would be interesting is to develop a deep learning method that learns the cell type expression profile. As I thought about this task, it made sense to try this through the task of deconvolution, which often relies on the cell type profile.

5.2.1 Architecture

The most basic form of this learning method is to take the full reference scRNA-seq data set and pass it through a locally connected neural network (LocNet), specifically organized to only share parameters within each cell type. Setting the dimension of the penultimate layer to the number of unique cell types and applying a non-negative activation would then yield at least the form of a cell type profile matrix. This is then used in the final output layer as a regression task, where the coefficients are considered the cell type compositions. So, in this setup, we not only learn a cell type profile from reference scRNA-seq data, but

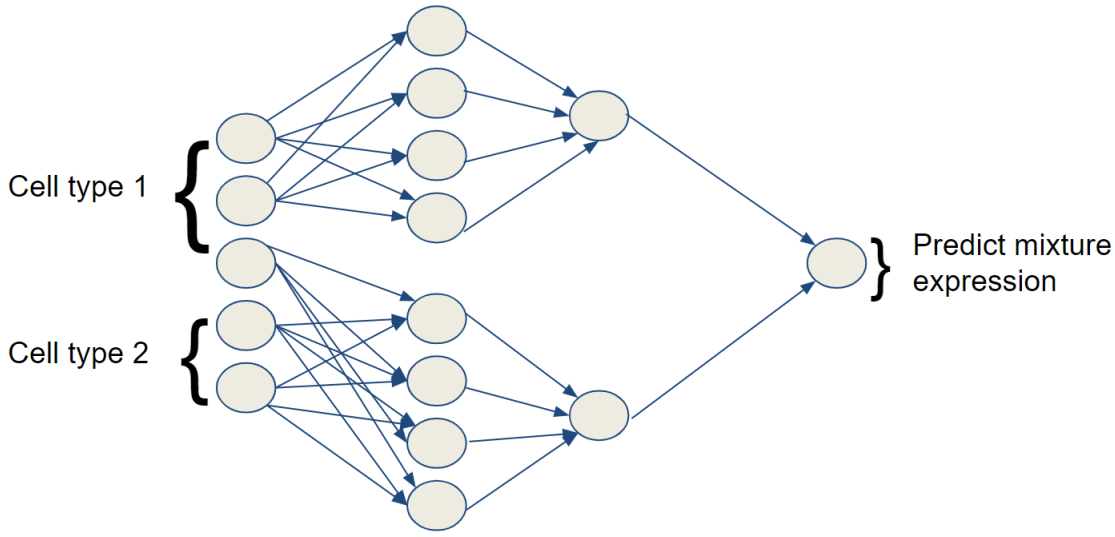


Figure 5.1 Locally-connected Neural Network.

we also perform deconvolution and get an estimate of cell type compositions. Here is a brief setup of the method. Note that we are using the log error model framework. Here we have D genes, N cells from reference scRNA-seq composed of K cell types, and M bulk mixtures.

$$\text{Reference scRNA expression (raw): } \mathbf{X} = [\mathbf{X}_1\, \mathbf{X}_2\, \cdots\, \mathbf{X}_K] \in \mathbb{R}^{D\times N},$$
$$\mathbf{X}_c \in \mathbb{R}^{D\times N_c},\ \sum_{c=1}^{K} N_c = N$$
$$\text{Mixture expression (raw): } \mathbf{Y} \in \mathbb{R}^{D\times M}$$

We then define LocNet and its locally connected layer as follows.

$$\text{Locally connected layer: } \mathbf{H} = \sigma\big(\mathbf{X} diag(\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_K)\big) \in \mathbb{R}^{D\times PK},$$
$$\text{where } \mathbf{W}_c \in \mathbb{R}^{D\times P} \text{ for c=1,...,K}$$
$$= \Big[\sigma\big(\mathbf{X}_1\mathbf{W}_1\big)\, \sigma\big(\mathbf{X}_2\mathbf{W}_2\big)\, \cdots\, \sigma\big(\mathbf{X}_K\mathbf{W}_K\big)\Big]$$

We can extend LocNet to have multiple locally connected layers, with

$$\mathbf{H}^{(0)} = \mathbf{X},\ H^{(i)} = \sigma\big(\mathbf{H}^{(i-1)} diag(\mathbf{W}_1^{(i)}, \mathbf{W}_2^{(i)}, ..., \mathbf{W}_K^{(i)})\big)$$

and we define the cell type expression profile by $F(\mathbf{X}) = \mathbf{H}^{(L)} \in \mathbb{R}^{D\times K}$. The objective is then to minimize the MSLE: $\frac{1}{D}||log(\mathbf{Y}) - log(F(\mathbf{X})\mathbf{B})||_2^2$

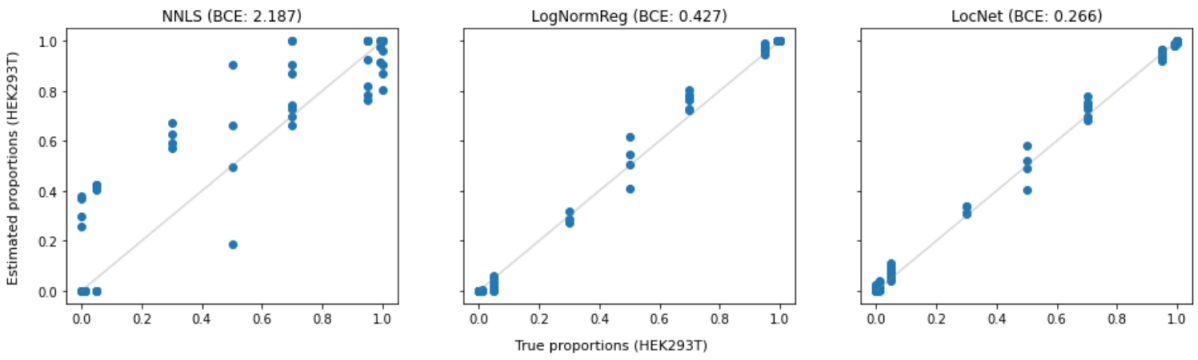


Figure 5.2 True vs estimated cell type proportions.

Proportion Loss		NNLS	LogNormReg	LocalNet
	MSE	0.075	0.009	**0.0009**
	MaxAD	0.405	0.115	**0.096**
	MedAD	0.05	0.011	**0.010**
	Corr	0.903	0.997	**0.998**
	BCE (**min bce - entropy of target 0.2451**)	2.187	0.427	**0.266**
Model Loss				
	MSE	**86614.27**	2289353.54	756179.56
	LogLog MSE	0.409	0.209	**0.064**

Figure 5.3 LocNet preliminary validation results table.

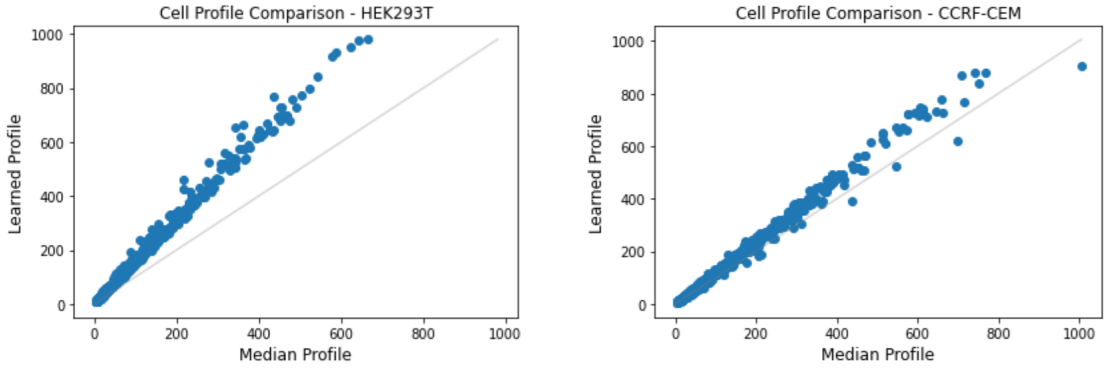


Figure 5.4 LocNet learned profile vs median profile.

Minimizing this objective we learn a cell type expression profile $F(\mathbf{X}) = \mathbf{H}^{(L)}$ and an estimate for the cell type compositions $\mathbf{B}$. A small example given in Figure 5.1 will help illustrate the local nature of this network.

5.2.2 Preliminary Experimental Results

To validate this method I chose a small cell line mixture dataset from Nanostring GeoMx, which came from a cell pellet array study done by Nanostring Danaher et al. (2022). It is a small dataset, but this is just an early validation to test the utility of the model. It consists of expression data from two cell lines (HEK293T, CCRF-CEM) mixed in cell-pellet array at varying proportions (40 mixtures, 16 pure cell-lines). 700 um regions were profiled for 1414 genes with GeoMx platform. Further, I normalized the expression data with 27 housekeeping genes that were selected using geNorm on the 50 highest mean expression genes.

I tested this method against the basic Non-negative Least Squares (NNLS) method, and the Log-Normal Regression (LogNormReg) method. I chose these two methods because they both use cell type profiles directly. For these two methods, I used the standard cell type profile construction, with the median expression within each cell type.

The results show LocNet performs strongly across various metrics in estimating true cell type composition. Since the loss function of LocNet is just the Log-Normal Regression model objective, this may suggest that LocNet is learning a better cell type profile than the standard rules based profile used in NNLS and LogNormReg. Further, looking at the results we may suggest the median is underestimating the true cell type profile.

5.3 Bivariate normal genes for 2 cell-types

Here I wanted to consider deconvolution in a probabilistic framework, which I began with a small digestible example consisting of only two cell types. Here, I consider 2 cell types by taking samples from bivariate normal distributions. The setup is as follows.

$$\boldsymbol{\nu}_c \sim N\Bigg(\boldsymbol{\mu}_c = \begin{bmatrix} \mu_{1,c} \\ \mu_{2,c} \end{bmatrix}, \boldsymbol{\Sigma}_c = \begin{bmatrix} \sigma_{1,c} & \rho_c\sigma_{1,c}\sigma_{2,c} \\ \rho_c\sigma_{1,c}\sigma_{2,c} & \sigma_{2,c} \end{bmatrix}\Bigg), \text{ for cell-types } c = 1, 2, \text{ with } \boldsymbol{\nu}_1 \perp\!\!\!\perp \boldsymbol{\nu}_2$$

Reference scRNA: $\mathbf{x}_c^{(i)} \sim \boldsymbol{\nu}_c$ (i.i.d.), for $i = 1, ..., N$. We put this in matrix form as:

$$\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2] \in \mathbb{R}^{2\times 2N}, \text{ where } \mathbf{X}_c = [\mathbf{x}_c^{(1)}\ \mathbf{x}_c^{(2)}\ \cdots \mathbf{x}_c^{(N)}]$$

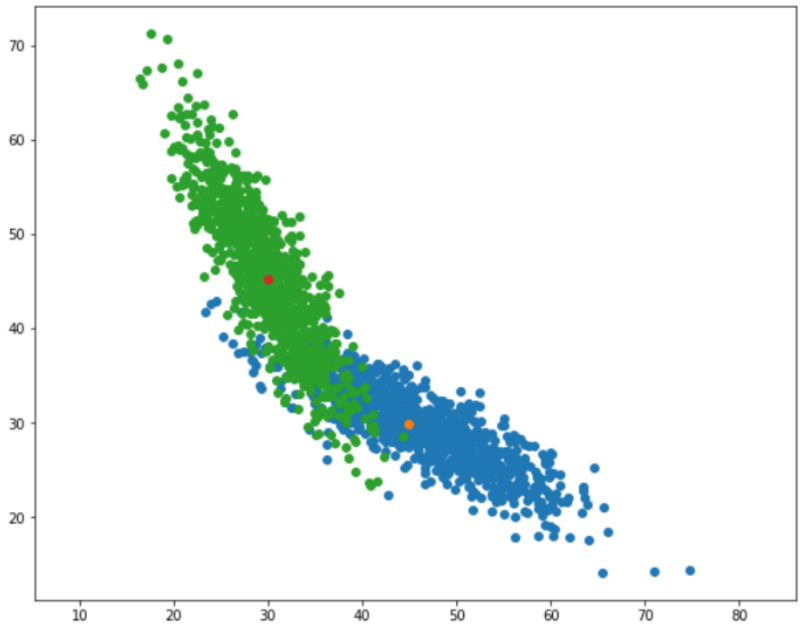


Figure 5.5 toy example - bivariate normal samples.

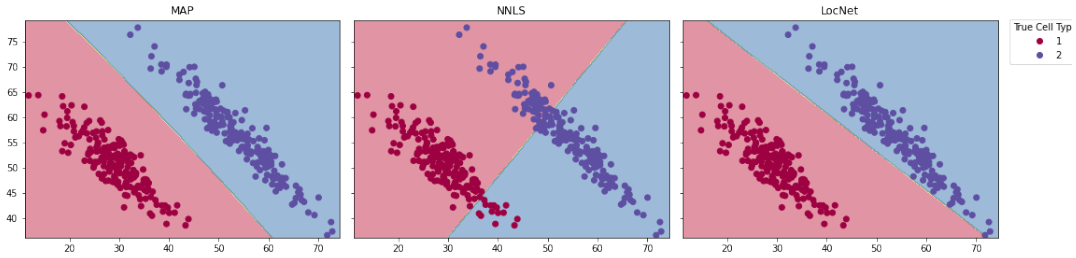


Figure 5.6 Decision boundary and classifications of MAP, LocNet and NNLS on the toy bivariate normal gene expression samples.

Mixed-cell RNA: $\mathbf{y}^{(i)} \sim \beta_1^{(i)}\tilde{\mathbf{x}}_1^{(i)} + \beta_2^{(i)}\tilde{\mathbf{x}}_2^{(i)}$, for $i = 1, ..., M$, where $\tilde{\mathbf{x}}_c^{(i)} \sim \boldsymbol{\nu}_c$ (i.i.d.), and $1 - \beta_1^{(i)} = \beta_2^{(i)} \in \{0, 1\}$. We put this in matrix form as:

$$\mathbf{Y} = [\mathbf{y}^{(1)}\, \mathbf{y}^{(2)}\, \cdots\, \mathbf{y}^{(M)}] \in \mathbb{R}^{2\times M}$$

In this toy example we set the mean and variance parameters as

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 45 \\ 30 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 60 & -30 \\ -30 & 20 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 30 \\ 45 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 20 & -30 \\ -30 & 60 \end{bmatrix}$$

We take 1000 samples from this distribution, and plot their 2D coordinates in Figure 5.5.

With this toy example, I am able to compute the Maximum a posteriori (MAP) estimate for the cell type compositions, against which I compare estimates from LocNet and NNLS for preliminary tests. Interestingly, I found LocNet was able to perform nearly as well as the MAP, and defines very similar decision boundaries.

CHAPTER 6

CONCLUSION

Going back to the Cell Type Deconvolution chapter, recall that scRNA-seq data are used as references for cell type composition estimation of spatial transcriptomic data. There are many variables at play between any pair of these datasets. One major variable is the differences in library preparation of the two sequencing technologies, which can lead to systematic bias that confounds cell type deconvolution results, termed as platform effects. RCTD Cable et al. (2022) and cell2location Kleshchevnikov et al. (2022b) are two methods that statistically account for platform effects and other sources of gene expression variations to model cell type compositions in spatial transcriptomic data. A direction I'd like to take my research is how to synthesize these ideas with deep learning methods, especially a GNN-based method that allows for easier multimodal data integration. Another aim of mine is to continue developing the cell type deconvolution benchmark datasets, and use them to validate my model developments.

A problem highlighted in the Plant Stress Response chapter is that of negative transfer. This is a problem the occurs across many domains, and interests me greatly. Going back to the experimental results in the Plant Stress Response chapter, DeepCAT is shows decent performance relative to well-established shallow and deep learning methods, but accuracy is still low in absolute terms. Thus there are still some challenges to overcome. From our results, we see that the stress grouping is a significant matter, and more sophisticated methods to learn the best groupings (i.e. the most related stresses) could help increase testing accuracy. Additionally, we saw leveraging transfer learning from both the big data human model, and the experimentally verified TFBMs helped increase the predictive accuracy. This is another lever for increased accuracy, and hence is a direction of great interest. However, One issue I found with the transfer learning schemes is there was no control mechanisms on the transferred information. We simply took the source data as parameter initializations and trained all of the target data (Arabadopsis) from there. We didn't account for similarities

or differences, for example, in the DNA sequences between humans and arabadopsis. The negative transfer is also found in the multi-task learning scheme. Grouping the heat and non-heat related stresses separately did improve from sharing across all stresses. However, even within those groups I found that the model performed particularly bad for a hand full of certain stresses. Again, control mechanisms would be helpful to limit the sharing of information between stresses when it leads to negative transfer.

BIBLIOGRAPHY

Hek293t and ccrf-cem cell line mixture data. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174746.

Human pdac data. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672.

Mouse olfactory bulb data. https://www.10xgenomics.com/resources/datasets/adult-mouse-olfactory-bulb-1-standard-1.

Mouse posterior brain 10x visium data. https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Posterior.

Andersson, A., Bergenstråhle, J., Asp, M., Bergenstråhle, L., Jurek, A., Fernández Navarro, J., and Lundeberg, J. (2020). Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications biology*, 3(1):565.

Asp, M., Bergenstråhle, J., and Lundeberg, J. (2020). Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays*, 42(10):1900221.

B. Alipanahi, A. Delong, M. T. W. and Frey, B. J. (2015). Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and Di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular systems biology*, 3(1):78.

Baron, M. and Yanai, I. (2017). New skin for the old rna-seq ceremony: the age of single-cell multi-omics. *Genome Biology*, 18(1):1–3.

Bartosovic, M., Kabbe, M., and Castelo-Branco, G. (2021). Single-cell cut&tag profiles histone modifications and transcription factors in complex tissues. *Nature biotechnology*, 39(7):825–835.

Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C. R., Segerstolpe, Å., Zhang, M., et al. (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature methods*, 18(11):1352–1362.

Brady, G., Barbara, M., and Iscove, N. (1990). Representative in vitro cdna amplification from individual hemopoietic cells and colonies. *Methods in Molecular and Cellular Biology*, 2:17—25.

Brehm-Stecher, B. F. and Johnson, E. A. (1990). Single-cell microbiology: Tools, technologies, and applications. *Microbiology and Molecular Biology Reviews*, 68(2):538—-559.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nature methods*, 10(12):1213.

Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., and Irizarry, R. A. (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526.

Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., Wang, X., Lu, H., Chen, X., Liu, X., Huang, X., Li, Z., Hong, Y., Jiang, Y., Peng, J., Liu, S., Shen, M., Liu, C., Li, Q., Yuan, Y., Wei, X., Zheng, H., Feng, W., Wang, Z., Liu, Y., Wang, Z., Yang, Y., Xiang, H., Han, L., Qin, B., Guo, P., Lai, G., Muñoz-Cánoves, P., Maxwell, P. H., Thiery, J. P., Wu, Q.-F., Zhao, F., Chen, B., Li, M., Dai, X., Wang, S., Kuang, H., Hui, J., Wang, L., Fei, J.-F., Wang, O., Wei, X., Lu, H., Wang, B., Liu, S., Gu, Y., Ni, M., Zhang, W., Mu, F., Yin, Y., Yang, H., Lisby, M., Cornall, R. J., Mulder, J., Uhlén, M., Esteban, M. A., Li, Y., Liu, L., Xu, X., and Wang, J. (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21.

Coons, A. H., Creech, H. J., Jones, R. N., and Berliner, E. (1942). The demonstration of pneumococcal antigen in tissues by the use of fluorescent antibody. *The Journal of Immunology*, 45(3):159–170.

Crosetto, N., Bienko, M., and Van Oudenaarden, A. (2015). Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics*, 16(1):57–66.

Dan D. Erdmann-Pham, Jonathan Fischer, J. H. and Song, Y. S. (2021). A likelihood-based deconvolution of bulk gene expression data using single-cell references. *Genome Research*. Code Link: https://github.com/songlab-cal/rna-sieve.

Danaher, P., Kim, Y., Nelson, B., Griswold, M., Yang, Z., Piazza, E., and Beechem, J. M. (2022). Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nature communications*, 13(1):1–13.

Ding, J., Liu, R., Wen, H., Tang, W., Li, Z., Venegas, J., Su, R., Molho, D., Jin, W., Wang, Y., et al. (2024a). Dance: A deep learning library and benchmark platform for single-cell analysis. *Genome Biology*, 25(1):72.

Ding, J., Venegas, J., Li, L., Lu, Q., Wang, Y., Wu, L., Jin, W., Wen, H., Liu, R., Tang, W., Dai, X., Li, Z., Zuo, W., Chang, Y., Leo, Y., Lulu-Shang, L., Danaher, P., Xie, Y., and Tang, J. (2024b). Spatialctd: A large-scale tumor microenvironment spatial transcriptomic dataset to evaluate cell type deconvolution for immuno-oncology. *Journal of Computational Biology*.

Ding, J., Wen, H., Tang, W., Liu, R., Li, Z., Venegas, J., Su, R., Molho, D., Jin, W., Zuo,

W., et al. (2022). Dance: A deep learning library and benchmark for single-cell analysis. *bioRxiv*. Code Link: https://github.com/OmicsML/dance.

Dong, R. and Yuan, G. (2021). Spatialdwls: accurate deconvolution of spatial transcriptomic data. *Genome Biology*, 22(1). Code Link: https://github.com/rdong08/spatialDWLS_dataset.

Dong, S., Wang, P., and Abbas, K. (2021). A survey on deep learning and its applications. *Computer Science Review*, 40:100379.

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., and Coleman, P. (1992). Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 89:3010–3014.

Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., and Heyn, H. (2021a). Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic acids research*, 49(9):e50–e50.

Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., and Heyn, H. (2021b). Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9):e50–e50.

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239.

Fan, Z., Luo, Y., Lu, H., Wang, T., Feng, Y., Zhao, W., Kim, P., and Zhou, X. (2023). Spascer: spatial transcriptomics annotation at single-cell resolution. *Nucleic Acids Research*, 51(D1):D1138–D1149.

Gerdes, M. J., Sevinsky, C. J., Sood, A., Adak, S., Bello, M. O., Bordwell, A., Can, A., Corwin, A., Dinn, S., Filkins, R. J., Hollman, D., Kamath, V., Kaanumalle, S., Kenny, K., Larsen, M., Lazare, M., Li, Q., Lowes, C., McCulloch, C. C., McDonough, E., Montalto, M. C., Pang, Z., Rittscher, J., Santamaria-Pang, A., Sarachan, B. D., Seel, M. L., Seppo, A., Shaikh, K., Sui, Y., Zhang, J., and Ginty, F. (2013). Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences*, 110(29):11982–11987.

Giesen, C., Wang, H. A., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, 11(4):417–422.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial*

Intelligence and Statistics, volume 9, pages 249–256. PMLR.

Goldman, S. L., MacKay, M., Afshinnekoo, E., Melnick, A. M., Wu, S., and Mason, C. E. (2019). The impact of heterogeneity on single-cell sequencing. *Front. Genet.*, 10:8.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G. P. (2018). Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, 23(12):2126–2135.

He, S., Bhatt, R., Birditt, B., Brown, C., Brown, E., Chantranuvatana, K., Danaher, P., Dunaway, D., Filanoski, B., Garrison, R. G., et al. (2021). High-plex multiomic analysis in ffpe tissue at single-cellular and subcellular resolution by spatial molecular imaging. *bioRxiv*, pages 2021–11.

He, S., Bhatt, R., and Brown, C. e. a. (2022). High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature Biotechnology*.

Hendrickson, D., Soifer, I., Wranik, B., Botstein, D., and McIsaac, S. (2018). Simultaneous profiling of dna accessibility and gene expression dynamics with atac-seq and rna-seq. *Methods in Molecular Biology*, 1819:317–333.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nature Reviews Genetics*, 11(12):855–866.

Jin, H. and Liu, Z. (2021). A benchmark for rna-seq deconvolution analysis under dynamic testing environments. *Genome Biology*.

Karagiannis, T., Monti, S., and Sebastiani, P. (2022). Cell type diversity statistic: An entropy-based metric to compare overall cell type composition across samples. *Frontiers in genetics*.

Keren, L., Bosse, M., Thompson, S., Risom, T., Vijayaragavan, K., McCaffrey, E., Marquez, D., Angoshtari, R., Greenwald, N. F., Fienberg, H., et al. (2019). Mibi-tof: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science advances*, 5(10):eaax5851.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., et al. (2022a). Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology*, 40(5):661–671.

Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., Jain, M. S., Park, J. S., Ramona, L., Tuck, E., Arutyunyan, A., Vento-Tormo, R., Gerstung, M., James, L., Stegle, O., and Bayraktar, O. A. (2022b). Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40(5):661–671. Code Link: https://github.com/BayraktarLab/cell2location.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620.

Kornberg, R. D. (1974). Chromatin structure: a repeating unit of histones and dna. *Science*, 184(4139):868–871.

Kornberg, R. D. and Lorch, Y. (1974). Chromatin structure and transcription. *Annual Review of Cell Biology*, 8:563–587.

Kulkarni, A., Anderson, A. G., Merullo, D. P., and Konopka, G. (2019). Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr. Opin. Biotechnol.*, 58:129–136.

Lähnemann, D. and et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1).

Lewis, Z. R., Phan-Everson, T., Geiss, G., Korukonda, M., Bhatt, R., Brown, C., Dunaway, D., Phan, J., Rosenbloom, A., Filanoski, B., et al. (2022). Subcellular characterization of over 100 proteins in ffpe tumor biopsies with cosmx spatial molecular imager. *Cancer Research*, 82(12_Supplement):3878–3878.

Li, B., Zhang, W., Guo, C., Xu, H., Li, L., Fang, M., Hu, Y., Zhang, X., Yao, X., Tang, M., et al. (2022a). Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature methods*, 19(6):662–670.

Li, H., Ma, T., Hao, M., Wei, L., and Zhang, X. (2022b). Decoding functional cell-cell communication events by multi-view graph learning on spatial transcriptomics. *bioRxiv*.

Li, X. and Wang, C.-Y. (2021). From bulk, single-cell to spatial rna sequencing. *International Journal of Oral Science*, 13(1):1–6.

Li, Z., Kuppe, C., Ziegler, S., Cheng, M., Kabgani, N., Menzel, S., Zenke, M., Kramann, R., and Costa, I. G. (2021). Chromatin-accessibility estimation from single-cell atac-seq data

with scopen. *Nature communications*, 12(1):1–14.

Lin, J.-R., Fallahi-Sichani, M., and Sorger, P. K. (2015). Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nature communications*, 6(1):1–7.

Lin, J.-R., Izar, B., Wang, S., Yapp, C., Mei, S., Shah, P. M., Santagata, S., and Sorger, P. K. (2018). Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-cycif and conventional optical microscopes. *Elife*, 7.

Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods*, 11(4):360–361.

Lytal, N., Ran, D., and An, L. (2020). Normalization methods on single-cell rna-seq data: an empirical survey. *Frontiers in genetics*, 11:501166.

Ma, Y. and Zhou, X. (2022). Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*. Code Link: https://github.com/YingMa0107/CARD.

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

Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., Catallini, J. L., Tran, M. N., Besich, Z., Tippani, M., et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436.

McManus, J., Cheng, Z., and Vogel, C. (2015). Next-generation analysis of gene expression regulation–comparing the roles of synthesis and degradation. *Molecular bioSystems*, 11(10):2680–2689.

Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D. S., Kloiber, K., Heutink, P., and Bonn, S. (2020). Deep learning-based cell composition analysis from tissue expression profiles. *Science Advances*, 6(30):eaba2619. Code Link: https://github.com/KevinMenden/scaden.

Merritt, C. R., Ong, G. T., Church, S. E., Barker, K., Danaher, P., Geiss, G., Hoang, M., Jung, J., Liang, Y., McKay-Fleisch, J., et al. (2020). Multiplex digital spatial profiling of proteins and rna in fixed tissue. *Nature biotechnology*, 38(5):586–599.

Moffitt, J. and Zhuang, X. (2016). Chapter one - RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH). In Filonov, G. S. and Jaffrey, S. R., editors, *Visualizing RNA dynamics in the cell*, volume 572 of *Methods in enzymology*, pages 1–49. Academic Press. ISSN: 0076-6879.

Molho, D., Ding, J., Tang, W., Li, Z., Wen, H., Wang, Y., Venegas, J., Jin, W., Liu, R., Su, R., et al. (2024). Deep learning in single-cell analysis. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–62.

Moor, A. E. and Itzkovitz, S. (2017). Spatial transcriptomics: paving the way for tissue-level systems biology. *Current opinion in biotechnology*, 46:126–133.

Moses, L. and Pachter, L. (2022). Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546.

Muzio, G., O'Bray, L., and Borgwardt, K. (2021). Biological network analysis with deep learning. *Briefings in Bioinformatics*, 22(2):1515–1530.

Nguyen, Q. H., Pervolarakis, N., Nee, K., and Kessenbrock, K. (2018). Experimental Considerations for Single-Cell RNA Sequencing Approaches. *Frontiers in Cell and Developmental Biology*, 6:108.

Okendo, J., Okanda, D., Mwangi, P., and Nyaga, M. (2022). Proteomic deconvolution reveals distinct immune cell fractions in different body sites in sars-cov-2 positive individuals. *medRxiv*.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., and Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36.

Quang, D. and Xie, X. (2016). Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, 44(11):e107–e107.

Rao, A., Barkley, D., França, G. S., and Yanai, I. (2021a). Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220.

Rao, A., Barkley, D., França, G. S., and Yanai, I. (2021b). Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220.

Raredon, M. S. B., Yang, J., Kothapalli, N., Lewis, W., Kaminski, N., Niklason, L. E., and Kluger, Y. (2023). Comprehensive visualization of cell–cell interactions in single-cell and spatial transcriptomics with niches. *Bioinformatics*, 39(1):btac775.

Repsilber, D., Kern, S., and Telaar, Anna, e. a. (2010). Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019a). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*,

363(6434):1463–1467. Code Link: https://github.com/broadchenf/Slideseq.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019b). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357.

Singh, O., Pratt, D., and Aldape, K. (2021). Immune cell deconvolution of bulk dna methylation data reveals an association with methylation class, key somatic alterations, and cell state in glial/glioneuronal tumors. *Acta Neuropathologica Communications*.

Song, Q. and Su, J. (2021a). Dstg: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in bioinformatics*, 22(5):bbaa414.

Song, Q. and Su, J. (2021b). Dstg: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in Bioinformatics*, 22(3):1–13. Code Link: https://github.com/Su-informatics-lab/DSTG.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., and Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, 16(3):133–145.

Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Bella, D. J. D., Arlotta, P., Macosko, E. Z., and Chen, F. (2020). Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqV2. *Nature Biotechnology*, 39(3):313–319.

Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., Arlotta, P., Macosko, E. Z., and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Large-scale simultaneous measurement of epitopes and transcriptomes in single cells. *Nature methods*, 14(9):865.

Sturm, G., Finotello, F., Petitprez, F., Zhang, J. D., Baumbach, J., Fridman, W. H., List, M., and Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type

quantification methods for immuno-oncology. *Bioinformatics*, 35(14):i436–i445.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Åke Borg, Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., and Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A survey on deep transfer learning. *CoRR*, abs/1808.01974.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382.

Teves, J. M. and Won, K. J. (2020). Mapping cellular coordinates through advances in spatial transcriptomics technology. *Molecules and Cells*, 43(7):591.

Thorsen, T., Roberts, R. W., Arnold, F. H., and Quake, S. R. (2001). Dynamic pattern formation in a vesicle-generating microfluidic device. *Physical Review Letters*, 86(18):4163–4166.

Thurman, R. et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Tian, L., Chen, F., and Macosko, E. Z. (2023). The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41(6):773–782.

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G.-C. (2019). Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, 10(1). Code Link: https://github.com/dtsoucas/DWLS.

Uygun, S., Seddon, A. E., Azodi, C. B., and Shiu, S.-H. (2017). Predictive models of spatial transcriptional response to high salinity. *Plant Physiology*, 174(1):450–464.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vistain, L. F. and Tay, S. (2021). Single-cell proteomics. *Trends in biochemical sciences*, 46(8):661–672.

Wang, G., Moffitt, J. R., and Zhuang, X. (2018). Multiplexed imaging of high-density libraries of rnas with merfish and expansion microscopy. *Scientific reports*, 8(1):1–13.

Wang, Y. and Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Molecular cell*, 58(4):598–609.

Waylen, L. N., Nim, H. T., Martelotto, L. G., and Ramialison, M. (2020). From whole-mount to single-cell spatial assessment of gene expression in 3d. *Communications biology*, 3(1):1–11.

Weingarten-Gabbay, S. and Segal, E. (2014). The grammar of transcriptional regulation. *Hum Genet*, 133(6):701–711.

Wen, L. and Tang, F. (2022). Recent advances in single-cell sequencing technologies. *Precision Clinical Medicine*, 5(1):pbac002.

Whitesides, G. M. (2006). The origins and the future of microfluidics. *Nature*, 442(7101):368–373.

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.-M., Pham, G. M., Nicotra, A. B., Gregorio, G. B., Jagadish, S. K., Septiningsih, E. M., Bonneau, R., and Purugganan, M. (2016). Egrins (environmental gene regulatory influence networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *The Plant Cell*, 28(10):2365–2384.

Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., and Haque, A. (2022). An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1):68.

Yan, L. and Sun, X. (2023). Benchmarking and integration of methods for deconvoluting spatial transcriptomic data. *Bioinformatics*, 39(1):btac805.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934.

APPENDIX

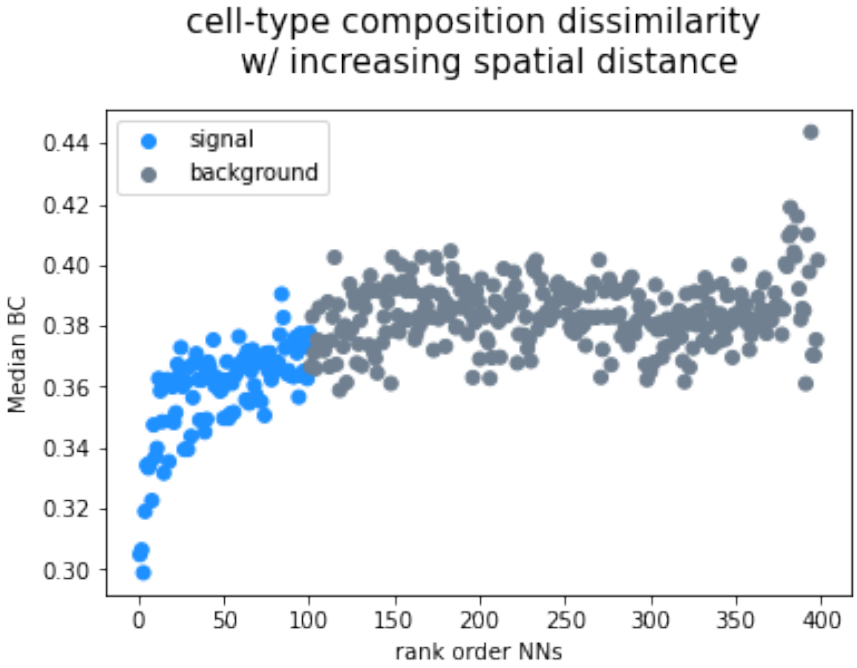


Figure A.1 Median of rank-ordered Bray-curtis dissimilarity taken over all spots.

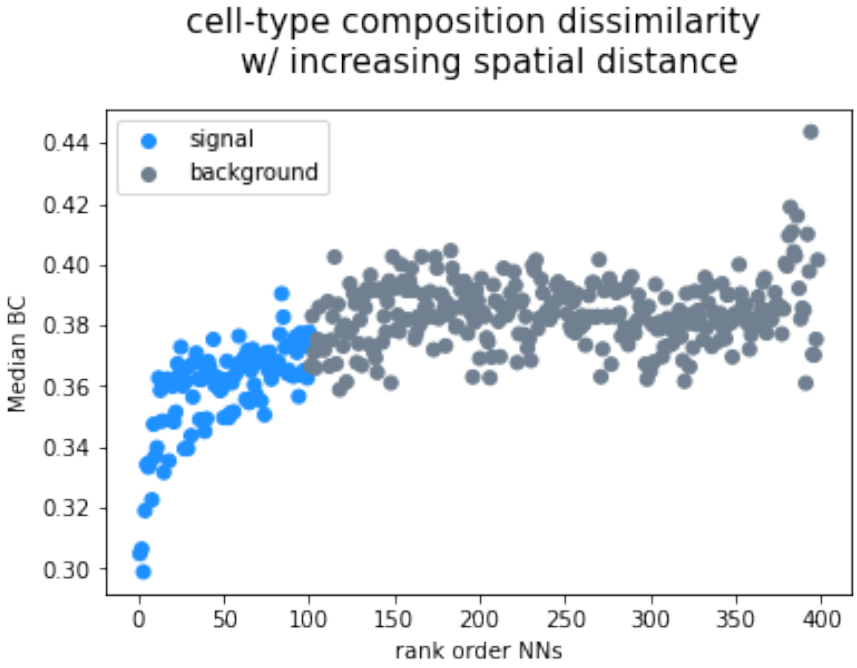


Figure A.2 Median of rank-ordered Bray-curtis dissimilarity taken over all spots.

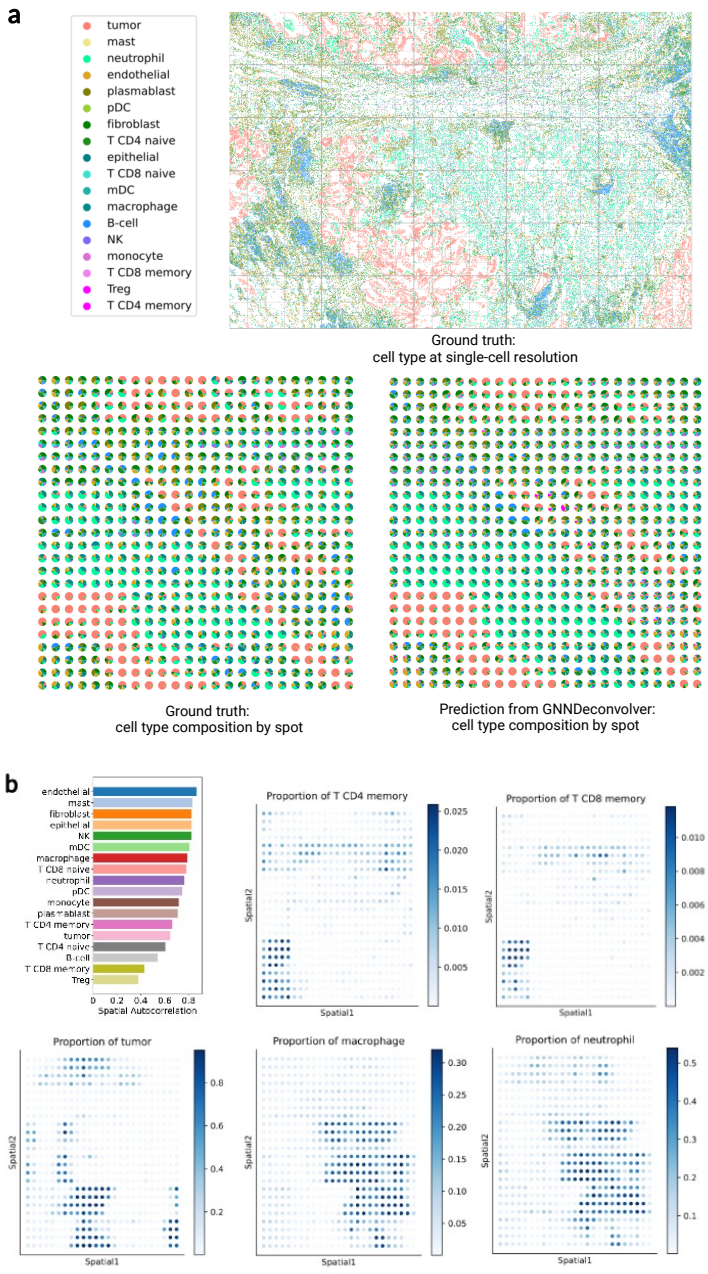


Figure A.3 GNNDECONVOLVER deconvolution on SPATIALCTD lung 5-2 sample. **a**. Ground truth single-cell resolution on SPATIALCTD Lung 5-2 sample. Each dot is a single cell colored by its ground truth cell type label. Proportions of deconvolved cell types from ground truth and GNNDECONVOLVER represented as pie charts for each spot. **b**. Spatial autocorrelation of the cell type proportions computed using Hotspot. Spatial distribution of cell type proportion for T CD4 memory cells, T CD8 memory cells, tumor, macrophage and neutrophil cells, as inferred by GNNDECONVOLVER. Each dot represents a spot. The depth of the point indicates the proportions of the cell type in the spot.

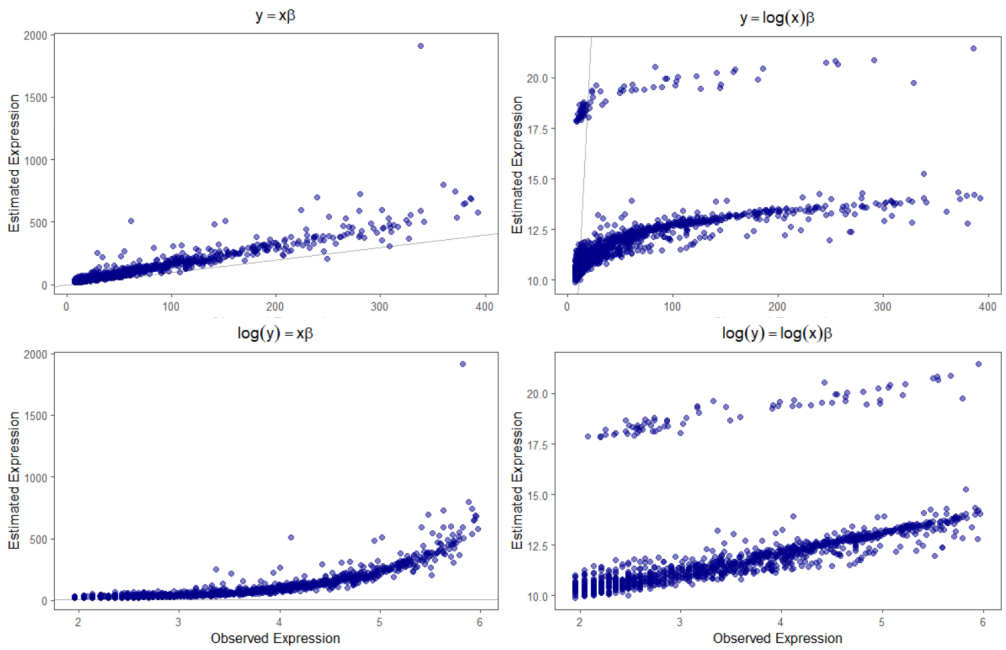


Figure A.4 HEK293T and CCRF-CEM cell line mixture observed expression vs estimated expression from various models.

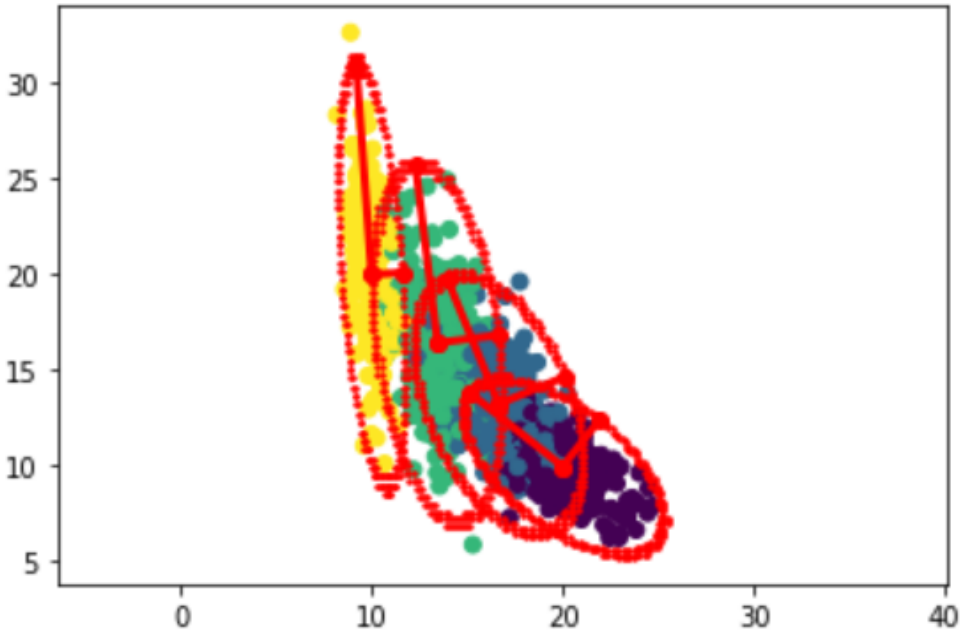


Figure A.5 PC Region Reconstruction of multivariate normal distributions - applied towards deconvolution methods.

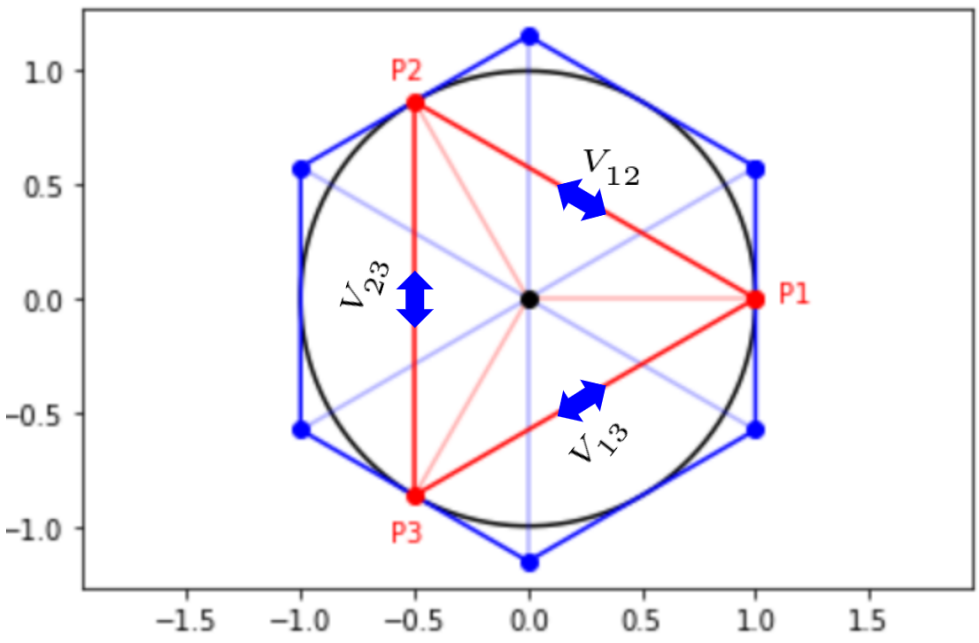


Figure A.6 The essence of the work for which I am most proud, and excited about.

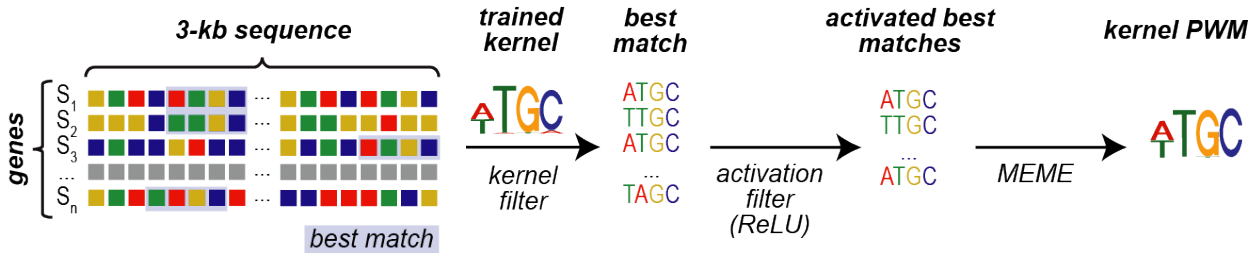


Figure A.7 Pipeline to translate kernel weights to position weight matrices, which can be compared experimentally verified motifs.

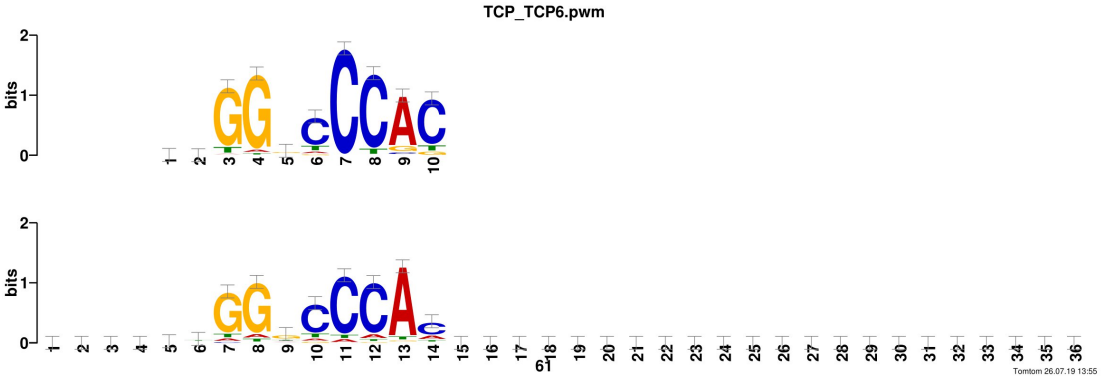


Figure A.8 Motifs learned from DeepCAT aligned with a known TFBM.

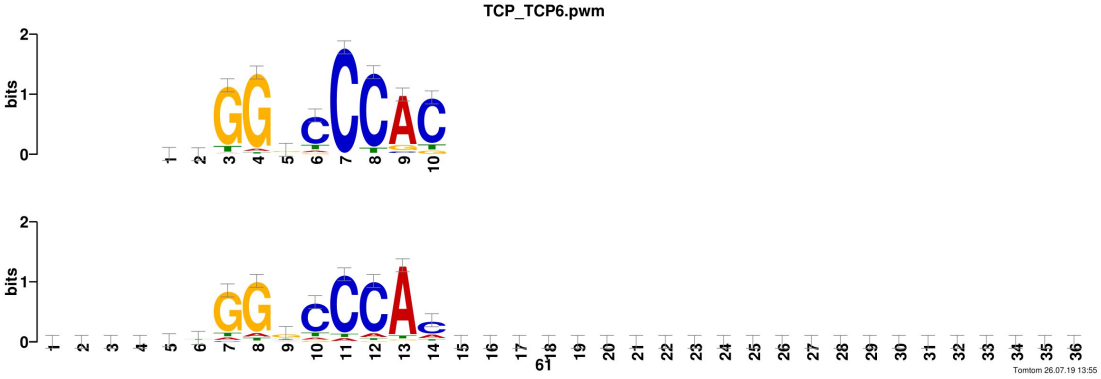


Figure A.9 Motifs learned from DeepCAT aligned with a known TFBM.

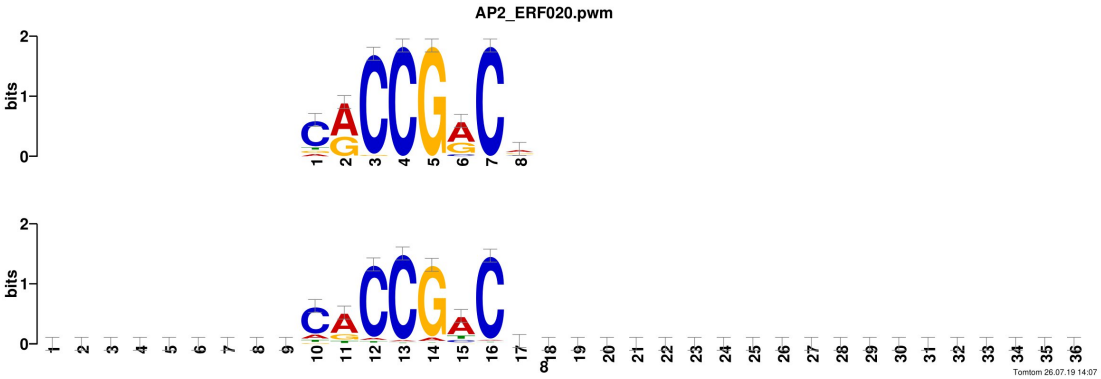


Figure A.10 Motifs learned from DeepCAT aligned with a known TFBM.

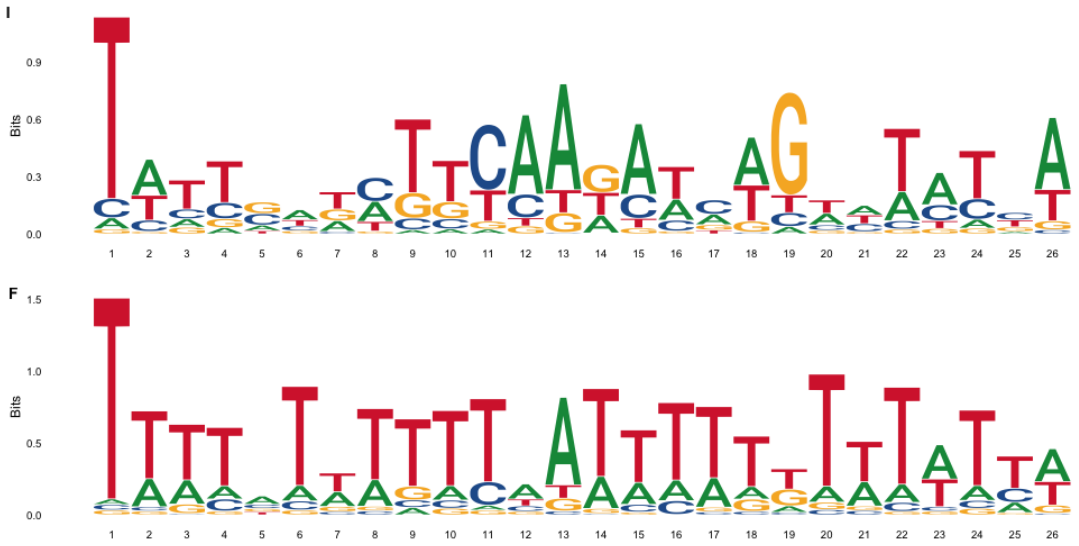


Figure A.11 Potential novel kernel PWM.

	LocalNet	LocalNet + min-max normalization	LocalNet (no scaling)
MSE	0.0121	0.001	**0.0009**
MaxAD	0.2781	**0.0896**	0.096
MedAD	0.0297	**0.01**	**0.01**
BCE	0.2959	0.2672	**0.266**

Figure A.12 This table examines scaling effects when only one cell line's expression is scaled (by 1000 here), a common problem in cell type expression profiling.