A Deep Learning Approach to Infer Cellular Features from Pathology Imaging Data

Zhining Sui

A Thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington
2023

Reading Committee:
Wei Sun
Ziyi Li

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

A Deep Learning Approach to Infer Cellular Features from Pathology Imaging Data

Zhining Sui

Chair of the Supervisory Committee:

Wei Sun
Department of Biostatistics

Recent developments in single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics (ST) provide unprecedented opportunities for studying individual cells and their organization. While these techniques are underutilized in clinical practice due to cost and logistical challenges, we propose an innovative alternative. We suggest leveraging scRNA-seq and ST data to train a deep learning model for histologically stained images, particularly hematoxylin and eosin (H&E) stained whole slide images, commonly used in disease assessment. In the burgeoning field of digital pathology, where deep learning excels in extracting meaningful imaging features, limited annotations for small image segments pose a challenge. To address this, we introduce STApath, a transfer-learning neural network model that automates patch-level annotation by exploiting paired ST data for model training and is therefore capable of predicting cell type proportions and classifying tumor microenvironments. Despite challenges such as uncertain annotations from spatial transcriptomics and image resolution disparities, our work establishes the viability of this pipeline. STApath's evaluation on breast cancer datasets demonstrates promising results even with limited training data and varying proportions and resolutions. Anticipating an influx of ST data, ongoing STApath updates hold the potential to become an invaluable AI tool for pathologists, streamlining diagnostic tasks.

Acknowledgements

I would like to express my gratitude to my thesis advisor, Dr. Wei Sun, whose unwavering support and expertise have been instrumental in guiding my master's research and crafting this thesis. His belief in my potential provided constant motivation, encouraging me to explore new ideas, challenging existing knowledge, and pushing the boundaries of my research. His insightful feedback enriched every phase of the journey, from idea inception to publication. His broad expertise was a blessing during the formulation, development, and composition of this thesis, and I am truly thankful for his invaluable mentorship.

I would also like to thank Dr. Ziyi Li, the second reader of this thesis, for her meticulous review of the thesis and thoughtful feedback that significantly enriched the quality of my work. Her expertise and dedication have been invaluable in shaping this research into a more comprehensive and impactful contribution to the field.

I wish to express my sincere thanks to the many friends and office mates with whom I have exchanged words, thoughts, ideas, and experiences. Their contributions have played an invaluable role in shaping my personal and intellectual growth, creating a dynamic environment of learning and camaraderie. Their presence has been a constant source of inspiration.

In closing, I want to convey my profound thanks to my parents, whose unwavering support and limitless love have been a guiding light not only throughout my master's research but also across the entirety of my life. It is truly impossible to adequately express the depth of my love and gratitude for all they have selflessly given and the immense role they've played in shaping who I am today.

DEDICATION

To my parents.

I can never thank you enough for your boundless and constant support, encouragement, and love, and I remain eternally grateful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Brief Overview

The recent developments of single cell RNA-sequencing (scRNA-seq) and Spatial Transcriptomics (ST) open unprecedented opportunities to study single cells and their organization. However, these techniques are rarely used in large-scale clinical practice due to cost and logistical issues. Instead of applying these techniques in clinical practice, we can use them to improve our understanding of other data modalities that are commonly used in clinical practice. We propose to use scRNA-seq and ST data to train a deep learning method for histologically stained images, which is the gold standard for the evaluation of many diseases. The most common stain is the hematoxylin and eosin (H&E), which provides cell morphology information. Deep learning is a popular method to analyze H&E whole slide images (WSIs) for many purposes, e.g., prediction of cardiac allograft rejection (Lipkova et al. [2022]) or tumor grade/subtype (Nagpal et al. [2019]; Coudray et al. [2018]). A H&E WSI is a giga-pixel image, which is huge and thus is often divided into small patches for patch-level analysis. Most often a WSI is only labeled on the image-level but not at patch-level. Manually annotating each patch by pathologists is possible, but it is labor-intensive and costly. Here we propose an automatic patch-level annotation approach called STApath (Spatial Transcriptomics Annotated pathology imaging data) using a deep learning approach while exploiting paired ST data for model training.

1.2 Background

1.2.1 Spatially Resolved Transcriptomics

Spatial organization and heterogeneity of gene expression within tissues have significant biological effects on tissue properties and functions (Finn and Misteli [2019]). In particular, tumors exhibit spatial variations not only due to genetic differences that occur during clonal expansion but also because of interactions between cancer cells and the surrounding immune and stromal cells in the tumor microenvironment. This interaction results in distinct characteristics in different parts of the tumor (Janiszewska [2020]). Understanding the spatial heterogeneity of tumors is crucial in a clinical setting, especially when limited physical samples of tumors are used to guide treatment decisions (Andor et al. [2016]).

However, traditional methods of analyzing transcriptomes, such as bulk RNA sequencing or single-cell RNA sequencing, fail to capture high-resolution spatial heterogeneity, resulting in a loss of rich spatial information regarding gene expression patterns and cellular communications. Spatially resolved transcriptomics, on the other hand, enables high-throughput measurement of gene expression while simultaneously preserving spatial information about tissue context and cellular organization. Various techniques have been developed for spatially resolved transcriptomics and they can be categorized into fluorescence in situ hybridization (FISH)-based, in situ sequencing (ISS)-based, and in situ capture-based techniques. FISH-based techniques such as MERFISH (Chen et al. [2015]), osmFISH (Codeluppi et al. [2018]), seqFISH+ (Eng et al. [2019]), and SABER-FISH (Kishi et al. [2019]) enable the concurrent acquisition of spatial and gene expression data through iterative imaging cycles. By hybridizing labeled probes that complement specific targets of interest, these methods enable the precise identification and localization of target genes within the spatial context of the sample. Likewise, ISS-based techniques, such as FISSEQ (Lee et al. [2015]), BaristaSeq (Chen et al. [2017]), and STARmap (Wang et al. [2018]), employ iterative cycles of imaging based on ISS in intact tissue samples. These techniques allow for direct RNA sequencing within the tissue context of a cell, enabling the resolution of transcript locations at subcellular levels. To amplify the signal sufficiently for imaging, micrometer- or nanometer-sized DNA balls are utilized. The resulting images obtained from either FISH- or ISS-based techniques are integrated into a spatially aligned dataset encompassing all cycles, and expression data are coded based on the presence or absence of signals in each pixel for each

cycle. These techniques enable the acquisition of data with submicron resolution, making them particularly suitable for examining processes occurring at subcellular scales. Nevertheless, achieving this high resolution entails increased costs of complexity, time, and equipment. While the previously discussed techniques are based on in situ visualization of RNA molecules by hybridization or sequencing, the in situ capturing-based techniques, on the other hand, capture transcripts in situ and subsequently perform sequencing ex-situ. The workflow typically involves placing thin tissue sections on top of an array of spatially barcoded features. These features can be densely packed beads, as utilized by Slide-seq (Rodriques et al. [2019]) and High Definition Spatial Transcriptomics (HDST) (Vickovic et al. [2019]), or printed spots, as utilized by Spatial Transcriptomics (ST) and 10X Visium (Ståhl et al. [2016]). DNA barcodes are employed to capture spatial data and gene expression information are retrieved through RNA sequencing. Although these capturing-based techniques offer gene expression profiles for hundreds and thousands of spots within each tissue section, they have some limitations. One such limation is related to spatial resolution, as the captured spots may not precisely represent the original spatial context of individual cells. Additionally, read coverage may be uneven across the tissue section, leading to potential variations in data quality. Among all spatially resolved transcriptomics techniques, ST/10X Visium appears better suited to clinical applications due to its relative simplicity, low barrier to adoption, similarity of workflows to clinical pathology techniques, and compatibility with histological imaging. Indeed, core facilities offering 10X Visium as a service have been established at many universities and hospitals.

Over the past decade, the rapid advancement of spatial transcriptomics technology has significantly contributed to biological discoveries across various fields. The spatially resolved transcriptomic data were often integrated with single-cell measurements through computational approaches, providing researchers with a powerful tool to generate comprehensive atlases of cellular architecture in tissues. This integration enables in-depth characterization of cell type composition and local cell states in multiple tissue types. Understanding the origins and progression of complex diseases has been made easier with the spatial context provided by ST. For instance, in a study by Chen et al. [2020b], ST was used to identify transcriptional alterations occurring in tissues surrounding amyloid plaques in Alzheimer's disease model. In addition, ST has been used to characterize the composition and spatial heterogeneity of the tumor microenvironment in various tumor types (Thrane et al. [2018]; Berglund et al. [2018]). In another study, Ji et al. [2020] incorporated

scRNA-seq data to computationally deconvolve spatial expression profiles, thereby determining the spatial locations of various cell types within the tissue of human squamous cell carcinoma. Similarly, Moncada et al. [2020] conducted scRNA-seq-based copy number variation analysis in ST data of pancreatic ductal adenocarcinomas. Besides, Friedrich and Sonnhammer [2020] used ST to infer the presence and absence of fusion transcripts in cancer cell line and clinical tissue data. Moreover, spatial analysis of developing organs can trace the cellular processes that organize morphogenesis. Asp et al. [2019] used ST to identify gene profiles that correspond to distinct anatomical regions in three stages of embryonic heart development and the spatial annotation of embryonic cardiac gene expression was enriched by scRNA-seq data. Finally, computational advances can further empower ST by integrating transcriptomic data with morphological features found in histological images. For example, coupled histological and ST data has been used to train machine learning algorithms to predict histopathological annotations based on gene expression data (Yoosuf et al. [2020]) as well as local transcriptional patterns based on histopathologic features (He et al. [2020]) in breast cancer.

1.2.2 Spatial Transcriptomics and 10X Visium

In this section, we introduce more details of two types of spatially resolved transcriptomics: Spatial Transcriptomics (ST) and 10X Visium. ST combines high-resolution tissue imaging and high-throughput transcriptome sequencing data. A key feature of the ST method is its ability to simultaneously register two types of information: histological imaging and gene expression. Histological imaging employs standard staining techniques to capture visual details of the tissue's structure, while gene expression profiling involves processing and sequencing spatially barcoded cDNA to determine gene activity. The gene expression datasets are then aligned with the images to ensure accurate visualization. The primary advantage of ST is its unbiased ability to capture mRNA abundance and their location information, which opens up possibilities for redefining both known and unknown morphological features solely based on molecular characteristics.

Spatial Transcriptomics

In ST, a tissue is cryosectioned into thin sections and placed on a glass slide. The slide bears spatially barcoded DNA oligonucleotide capture probes that are printed as microarrays of spots in a capture area.

To secure the tissue section on the microarray, it is fixed using methanol and then stained by H&E (hematoxylin for nuclei and eosin for the extracellular matrix and cytoplasm). The stained tissue sections are then subjected to imaging.

Next, gentle enzymatic permeabilization of the same tissue enables the release of mRNA molecules, which diffuse onto the slide's surface. The poly-adenylated mRNAs then bind to the poly(dT) sequence present in the capture probes, which in turn serves as a primer for reverse transcription. The reverse transcription generates cDNA sequences that correspond to the mRNAs and contain the spatial barcodes and unique molecular identifiers (UMI) from the capture probes. The resulting cDNA library is collected and subjected to sequencing using standard Illumina workflows. Since the sequences contain spatial barcodes as well as RNA sequences, the RNAseq counts can be mapped to specific tissue regions. This allows for the visualization of gene expression within the original placement of the tissue.

The initial version of ST microarrays consisted of approximately 1,000 spots with 200 μm spacing between spots. Each spot has a diameter of 100 μm. Consequently, they provided an averaged transcriptomic profile obtained from a mixture of several cells (Ståhl et al. [2016]).

10X Visium

To enhance the ST technology, 10X Genomics further developed and commercialized the Visium spatial gene expression technology in 2018, with increased sensitivity and throughput. 10x Genomics Visium Spatial Gene Expression is a next-generation molecular profiling solution for ST. It enables the measurement of the entire transcriptome in a spatially resolved manner by mapping gene expression onto high-resolution microscope images of intact fresh frozen sections or Formalin-Fixed Paraffin-Embedded (FFPE) tissue sections. The Visium Spatial Gene Expression solution consists of slides, reagents, and software tools that facilitate whole transcriptome analysis of tissue samples.

During the process, investigators section the tissue onto a Visium Gene Expression (GEX) slide. The Visium slide will undergo H&E staining, imaging, library prep, sequencing, and SpaceRanger data processing. These GEX slides have either two tissue capture areas measuring 11 mm x 11 mm or four tissue capture areas measuring 6.5 mm x 6.5 mm. Each capture area is defined by a fiducial frame. The standard capture area is 6.5 mm x 6.5 mm and contains approximately 5,000 barcoded spots, while the XL capture area is

11 mm x 11 mm and contains approximately 14,000 barcoded spots. The spot size was reduced from 100 µm to 55 µm in diameter, and the capture area density was enhanced by utilizing hexagonal packing with a distance of 100 µm between the centers of adjacent spots. Depending on the tissue type and thickness, an average of 1-10 cells would cover each spot. This advancement brought the technology closer to achieving single-cell resolution.

Although ST and 10X Visium lack the spatial resolution necessary to distinguish single cells in most samples, the inclusion of a histological image within the workflow allows for computational inference of spatial expression at resolutions finer than the physical capture areas. The average gene expression measurements obtained per spot are supplemented by histological images, which provide connections between molecular characteristics and the morphological attributes of larger tissue structures or even individual cells. Besides, achieving cellular resolution would involve reducing the spot dimensions to a few micrometers in diameter. However, this would require more precise printing technology capable of depositing spatial barcodes into denser array patterns.

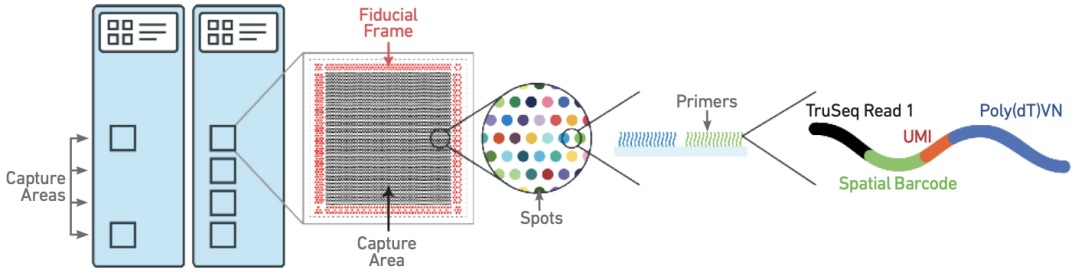


Figure 1.1: Illustration of Visium Gene Expression slide, provided by 10x Genomics.

The spatial barcode assigned to each spot is incorporated during cDNA synthesis, enabling the mapping of gene expression data back to its precise location within the tissue. As shown in Figure 1.1, the spots on the slide contain oligos that consist of the Illumina TruSeq Read 1 sequencing primer (partial), a 16 nt Spatial Barcode (shared by all primers within a specific spot), a 12 nt UMI, and a 30 nt poly(dT) sequence (used for capturing poly-adenylated mRNA for cDNA synthesis) or a probe capture sequence for FFPE samples.

To ensure accurate alignment of RNA-Seq data with stained tissue images, it is recommended that tissue sections do not exceed the size of the capture area (6.5 mm or 11 mm) to avoid covering the fiducial frame. Additionally, placing tissue outside the capture area will not generate any additional gene expression data

and may potentially complicate the interpretation of the generated gene expression data.

Following the staining, imaging, and sequencing steps, the output data is processed using the 10X SpaceRanger analysis software and can be visualized using the Loupe Browser software. SpaceRanger comprises a collection of analysis pipelines designed to handle Visium Spatial Gene Expression data along with brightfield and fluorescence microscope images. This software empowers users to map the entire transcriptome in both fresh frozen and FFPE tissues, facilitating the discovery of novel insights into normal development, disease pathology, and clinical translational research.

To begin the analysis, SpaceRanger takes an input slide image that serves as an anatomical map onto which gene expression measurements are projected. The input for SpaceRanger can be in two formats: a brightfield image stained with H&E, featuring dark tissue on a light background, or a fluorescence image with bright signals on a dark background. While brightfield input consists of a single image, fluorescence input incorporates one or more channels of information generated from separate excitations of the sample.

SpaceRanger seamlessly combines algorithms for processing Visium slide images with the robust gene expression analysis employed within the software. An analysis pipeline called "spaceranger count" is utilized, which takes a microscope slide image and FASTQ files as input. This pipeline carries out alignment, tissue detection, fiducial detection, and barcode/UMI counting. By leveraging the Visium spatial barcodes, this pipeline generates feature-barcode matrices, identifies clusters, and conducts gene expression analysis. The spaceranger count is executed on each individual capture area present in the Visium slide.

1.2.3 Digital Pathology

Digital pathology has revolutionized the traditional field of histopathology by introducing whole slide imaging (WSI) and integrating artificial intelligence (AI) for advanced image analysis. The emergence of this new field has transformed the traditional practice of pathologists, enabling them to view and analyze digitized images in real-time, share data remotely for collaboration and second opinions, and utilize digital pathology for improved diagnostic accuracy and patient care.

WSI is at the core of digital pathology's transformative capabilities. WSI involves scanning multiple images of entire tissue sections on the glass slides at high resolution to create digital slides (Pantanowitz et al. [2013], Zarella et al. [2018]). These digital slides can be archived, reviewed, and analyzed on computer

monitors, eliminating the need for physical slide handling and storage. The WSI scanner's image resolution (i.e., sharpness) is determined by the microscope objective used for scanning, the numerical aperture of the objective, and the quality of the camera's photosensors (Pantanowitz et al. [2013]). WSI provides a rich source of information, capturing not only pixel-level cellular patterns but also color information from stains such as H&E and immunohistochemistry. This vast amount of data presents challenges for human pathologists to extract all relevant information manually since it requires significant time investment and can be prone to human error.

The utilization of digital images enables the use of computational approaches for quantitative analysis of WSIs. However, traditional statistical methods are no longer capable of integrating and assimilating the overwhelming volume of data of WSIs. Classical statistical models commonly exploit relatively simple models that involve limited factors and multiple assumptions, making them not suitable for dealing with high-dimensional and highly correlated data. Fortunately, Artificial Intelligence (AI) algorithms, particularly machine learning (ML) models, have emerged and proven to be effective in addressing the limitations of traditional analysis methods. ML is broadly defined as a set of mathematical models and computational algorithms designed to automatically learn complex patterns from training data, without being explicitly programmed to do so (Bera et al. [2019]). Over time, the performance of the model continuously enhances as it learns from newly accessible data. ML techniques can process and analyze the vast amount of data present in WSIs, extracting morphological patterns, histological features, and biological markers of interest. ML techniques can be categorized into supervised learning and unsupervised learning. Supervised learning refers to learning from labeled data to predict the labels of unlabeled input data, while unsupervised learning refers to learning from unlabelled data to differentiate data into groups or to find patterns in a dataset.

Deep learning (DL), a specialized subset of ML, involves training artificial neural networks with multiple layers of interconnected neurons to automatically learn hierarchical representations of data from input data (Srinidhi et al. [2021]). The application of DL in digital pathology enables molecular classification and predicting response or identifying patients most likely to respond to treatment. A convolutional neural network (CNN) trained with WSIs of H&E stained lung tissue from The Cancer Genome Atlas was used to classify between adenocarcinoma, squamous cell carcinoma, or normal lung tissue (Coudray et al. [2018]). The performance of the trained network was comparable to that of pathologists, with an average area under

the curve (AUC) of 0.97. Similarly, Nagpal et al. [2019] trained a CNN with WSIs of H&E-stained formalin-fixed paraffin-embedded (FFPE) prostatectomy specimens to perform Gleason Grade Group classification. Other than the diagnosis and classification of diseases, DL can also be used to understand the tumor-immune microenvironment (TME). A lymphocyte-infiltrated classification CNN and a necrosis segmentation CNN were trained with H&E-stained WSIs of 13 TCGA tumor types to develop tumor-infiltrating lymphocyte (TIL) maps, with TIL densities and spatial structure differentially enriched among tumor types, immune subtypes, and tumor molecular subtypes (Saltz et al. [2018]). In addition, DL was used to predict gene mutations from histopathology slides. Using a CNN trained with WSIs of H&E-stained hepatocellular carcinoma (HCC) tissue, Chen et al. [2020a] predicted the ten most common prognostic and mutated genes in HCC, with four of them correctly identified by the model.

1.3 Approach

We have created a deep-learning-based pipeline named STApath, which stands for Spatial Transcriptomics Annotated Pathology Imaging Data. This pipeline is designed to provide automatic annotations for WSIs using either quantitative or qualitative features. The process involves segmenting WSIs into smaller image patches, with each patch being input into a CNN. For the CNN, we utilized transfer learning with the pre-trained ResNet50 model as the base model. Our pipeline can handle both regression and classification tasks. The regression model's objective is to quantitatively predict the cell type proportions within each image patch. On the other hand, the classification model aims to classify each image patch into one of multiple cellular composition patterns. For the training data, the response variables, either quantitative or qualitative, were determined based on the cell type deconvolution of ST data.

1.4 Thesis Outline

In Chapter 2, we comprehensively describe the STApath pipeline, including the type of input and output data, imaging preprocessing steps, and the architecture of the convolutional neural networks employed. In Chapter 3, we delve into the quantification of cell type proportions derived through cell type deconvolution of ST data. By quantifying these proportions, we gain valuable insights into the cellular composition of

the observed gene expression patterns. In Chapter 4, we evaluate the convolutional neural networks' ability to extract meaningful features from WSIs. By assessing the extracted features, we ascertain the network's efficacy in capturing relevant information for downstream analyses. This evaluation plays a crucial role in determining the overall performance and reliability of the results obtained through our pipeline. Chapter 5 and Chapter 6 present the results of regression tasks: the predicted cell type proportions. Similarly, in Chapter 7, we present the results obtained from the classification tasks. Chapter 8 discusses the implications of the results obtained throughout our analysis. We highlight significant findings, elucidate their potential impact on the field, and explore the broader implications of our research. Additionally, we critically examine the limitations of the STApath pipeline and its underlying technologies, acknowledging challenges and suggesting avenues for improvement. Lastly, we outline future directions for research, identifying promising areas for further exploration and advancement in the field of digital pathology.

Chapter 2

Methods

In this chapter, we focus on developing deep learning methods that predict cellular/molecular features based on small patches of H&E stained WSIs, named STApath (Spatial Transcriptomics Annotated pathology imaging data). STApath consists of several steps (Figure 2.1). First, we create patches from the WSIs and preprocess these patch images to serve as input data for our neural networks. Each generated patch is then labeled either with quantitative features such as the proportions of relevant cell types or qualitative features such as cellular composition patterns (e.g., a mixture of tumor and lymphocyte cells vs. a mixture of tumor and fibroblast).

STApath serves two main purposes: regression for quantitative features and classification for qualitative features. In the current work, for regression, we predict the proportion of each relevant cell type based on image patches. When training the regression model, the response variables are cell type proportion estimates derived from corresponding 10X Visium gene expression data. For classification, the response variable is cellular composition patterns that are derived from the clustering of spatial transcriptomic data or pathologist annotations.

To enable the automated learning of complex image patterns, we feed the imaging input along with the corresponding responses into a neural network. Specifically, our neural networks are constructed using transfer learning based on ResNet50. This approach allows the neural network to leverage the knowledge gained from pre-trained models and adapt it to our specific task of cell-type composition prediction and annotation pattern recognition.

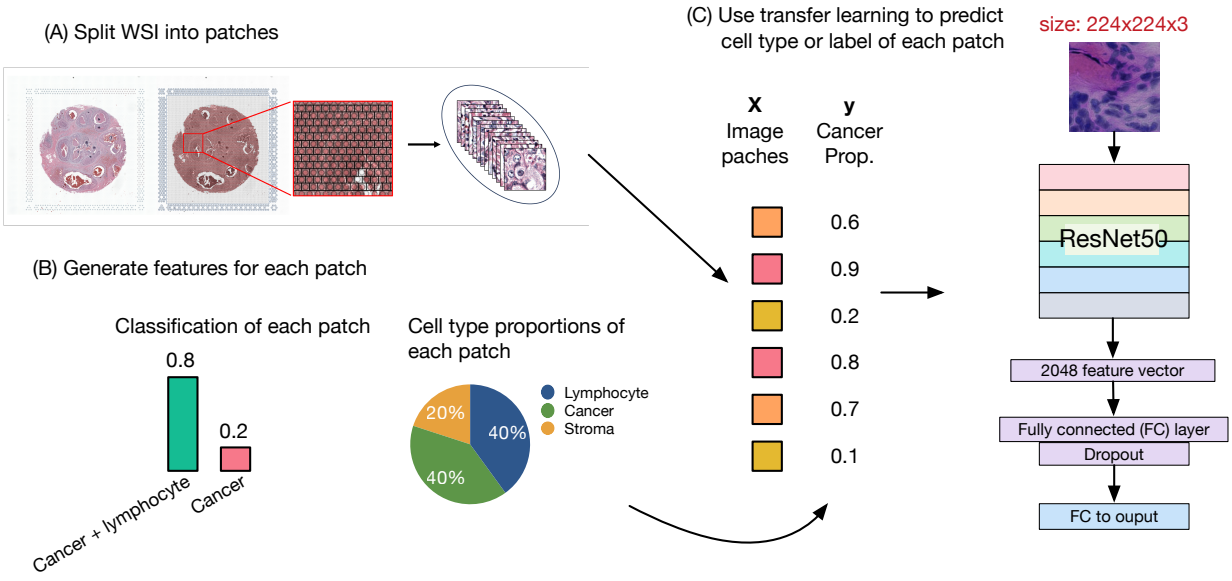
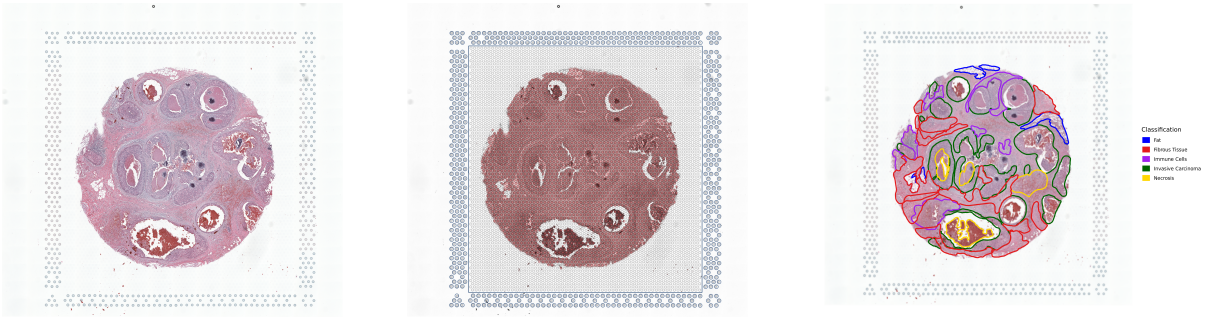


Figure 2.1: An overview of STApath. (A) An illustration to split a WSI into individual patches. The left panel shows the WSI of the breast cancer FFPE sample downloaded from the 10X genomics website. The other panels illustrate the process to split the WSI into patches. (B) Generate features to be learned for each patch, using paired ST data, pathologist annotations, or other resources. (C) Given image patches $\mathbf{X}$ and features associated with each patch $\mathbf{y}$, use transfer learning to build a neural network to learn features from image patches.

2.1 Spatial Transcriptomics Datasets

2.1.1 10X Visium: Two Breast Cancer Tissues

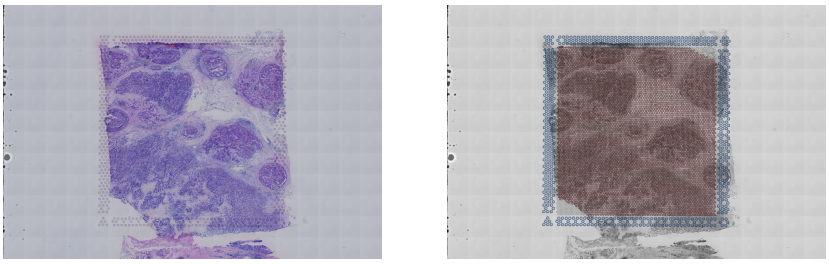
In our study, we utilized spatial gene expression data obtained from 10x Genomics' SpaceRanger 1.3.0. The data consisted of two breast cancer tissue samples. The first sample was derived from FFPE human breast tissue obtained from BioIVT Asterand Human Tissue Specimens. This particular sample was annotated as Ductal Carcinoma In Situ and Invasive Carcinoma. Tissue sections with a thickness of 5 µm were placed on Visium Gene Expression slides. Detailed information regarding the staining, imaging, and sequencing settings, as well as the input and output of SpaceRanger, can be found at 10X genomics website: `https://www.10xgenomics.com/resources/datasets/human-breast-cancer-ductal-carcinoma-in-situ-invasive-carcinoma-ffpe-1-standard-1-3-0`. The corresponding slide image of this sample, along with an overview-level annotation performed by a pathologist, is displayed in Figure 2.2.



(a) High-resolution image **(b)** Detected spots **(c)** Pathologist annotations

Figure 2.2: Images of 10X Genomics FFPE Ductal Carcinoma In Situ, Invasive Carcinoma breast tissue.

The second sample was derived from a fresh frozen Invasive Ductal Carcinoma breast tissue obtained from BioIVT. Tissue sections with a thickness of 10 µm were placed on Visium Gene Expression slides. Unfortunately, we did not have annotations from pathologists for this particular sample. The corresponding slide image of this fresh frozen sample is displayed in Figure 2.3. More details regarding the processing and sequencing settings, as well as the input and output of SpaceRanger, can be found at: `https://www.10xgenomics.com/resources/datasets/human-breast-cancer-visium-fresh-frozen-whole-transcriptome-1-standard`.



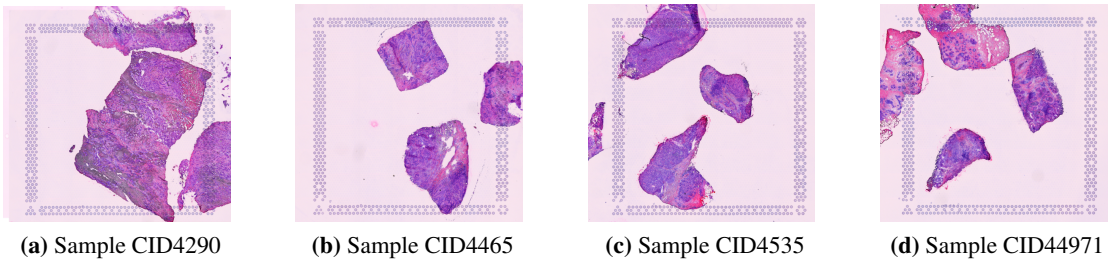
(a) High-resolution image **(b)** Detected spots

Figure 2.3: Images of 10X Genomics fresh frozen Invasive Ductal Carcinoma breast tissue.

In our analysis, we made use of the full-resolution WSIs provided by 10X Genomics. These images were employed to generate patches that served as input for the neural network. Both WSIs were scanned using a 20× magnification microscope objective. Access to these WSIs is available on the websites provided by 10X Genomics as mentioned earlier.

2.1.2 Wu et al. [2021]: Six Breast Cancer Tissues

The ST dataset provided by Wu et al. [2021] consisted of six human breast samples obtained from various breast cancer patients. These samples comprised two samples classified as Estrogen Receptor-positive (ER+) (CID4535 and CID4290) and two samples classified as Triple-negative breast cancers (TNBCs) (CID44971 and CID4465) from their scRNA-seq cohort (as mentioned in section 3.1.1), and two additional TNBCs (1142243F and 1160920F) processed in an independent laboratory (Figure 2.4). The authors generously provided full-resolution WSIs of all six tissues upon request. Notably, the WSIs of the four samples from the scRNA-seq cohort of Wu et al. [2021] had distinct resolutions compared to the WSIs of the two samples processed in an independent laboratory. The WSIs processed by Wu et al. [2021] were scanned using a 20× magnification microscope objective, while the WSIs processed in the independent laboratory were scanned using a 40× objective.



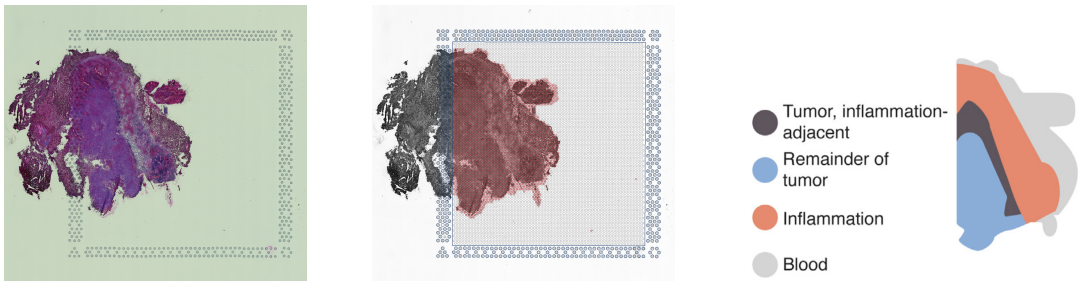
(a) Sample CID4290 **(b)** Sample CID4465 **(c)** Sample CID4535 **(d)** Sample CID44971

(e) Sample 1142243F **(f)** Sample 1160920F

Figure 2.4: Whole slide images of six samples in Wu et al. [2021].

2.1.3 Sudemeier et al. [2022]

The ST data of a melanoma brain metastasis (patient 16) was obtained from Sudmeier et al. [2022], specifically from a H&E-stained fresh frozen section of the tumor parenchyma. The WSI was scanned at 10× magnification. Figure 2.5 displays the H&E-stained image and the capture spot of this sample.



(a) High-resolution image **(b)** Detected spots **(c)** Clustering of spots

Figure 2.5: Hematoxylin-and-eosin-stained section of a melanoma brain metastasis (patient 16) in Sudmeier et al. [2022].

2.2 Image Preprocessing

2.2.1 Patch Extraction

In our analysis, we worked with ST data that featured a standardized capture area on GEX slides, which consisted of approximately 5000 barcoded spots. Some of the spots do not overlap with the tissue sample. When extracting patches from WSIs, we only utilized spots within the tissue that contained gene expression data. Specifically, only patches containing a minimum of 70% tissue were selected for further analysis. To distinguish tissue from the background, we adopt a three-step approach similar to the one utilized by Barker et al. [2016]. Firstly, we converted the patch images to grayscale to merge the three color channels into a single channel, where each pixel was assigned an integer value ranging from 0 to 255. Given that the slide backgrounds were illuminated with white light, the pixel values in the grayscale image's background were mostly close to or equal to 255. Next, the complement of the grayscale image was obtained at an 8-bit depth, so that the background values are close to or equal to 0. Finally, hysteresis thresholding, an automatic edge detection technique, was performed using experimentally-determined high and low thresholds for each sample. The images exhibit varying degrees of edge brightness, where some edges appear stronger, while others are fainter and might be attributed to noise. Consequently, a dual-threshold approach was adopted such that edges with intensities above the high threshold are identified as true edges, those falling below the low threshold are classified as non-edges, and edges with intensities lying between the high and low thresholds are identified as edges only if they are connected to a previously identified true edge; otherwise,

they are discarded.

Standard Patches

The output of the SpaceRanger analysis for the ST data includes the position information and diameter of each spot. To generate standardized patches, we ensured that each patch contained only one spot. We accomplished this by setting the center of each spot as the center of the corresponding square patch. The diagonal of the square was determined in such a way that there was no overlap between different patches. The resolutions of the extracted patches for each sample tissue used in this study are shown in Table 2.1.

	Image capture magnification	Dimensions of WSI (pixel× pixel)	Dimensions of patches (pixel× pixel)
10X Genomics FFPE Breast Tissue	20x objective	27452× 25233	Standard: 208× 208
10X Genomics Fresh Frozen Breast Tissue	20x objective	41572× 27755	Standard: 194× 194
Sample CID4290 Processed by Wu et al.	20x objective	9907×9404	Standard: 90× 90 Large: 290× 290
Sample CID4465 Processed by Wu et al.	20x objective	9648× 9232	Standard: 90× 90 Large: 290× 290
Sample CID44971 Processed by Wu et al.	20x objective	9648× 9232	Standard: 90× 90 Large: 290× 290
Sample CID4535 Processed by Wu et al.	20x objective	9648× 9232	Standard: 90× 90 Large: 290× 290
Sample 1142243F Processed in An Independent Laboratory	40x objective	41572× 27755	Standard: 224× 224 Large: 720× 720
Sample 1160920F Processed in An Independent Laboratory	40x objective	41572× 27755	Standard: 226× 226 Large: 724× 24
Melanoma Brain Metastasis Processed by Sudemeier et al.	10x objective	15796× 14355	Standard: 50× 50

Table 2.1: Summary of resolution and size in WSIs and patches. The samples were scanned using various commercial WSI systems from different manufacturers. It should be noted that different WSI scanning models may have distinct image capture resolutions (microns/pixel) at the same image capture magnification, resulting in varying dimensions of WSIs under identical image capture magnification.

Large Patches

We also generated larger patches that encompassed eight spots within each patch (Figure 2.6). Since the Visium spots are hexagonally spaced, the arrangement of these eight spots followed a specific pattern, forming an hourglass shape. In the first and third rows of the large patch, there were three spots each, while the second row contained two spots. To select the spots for the first and third rows of the large patch, we consistently chose them from the double rows present in the entire tissue slide. For a spot located at column

j and a double row i, we created a large patch that included spots at the following positions: (i, j), $(i, j+2)$, $(i, j+4)$, $(i+1, j+1)$, $(i+1, j+3)$, $(i+2, j)$, $(i+2, j+2)$, $(i+2, j+4)$. The process began with the spot located at row zero and column zero, and each subsequent large patch was created by shifting one spot to the right with each new spot in the double row serving as the top left spot in the next large patch.

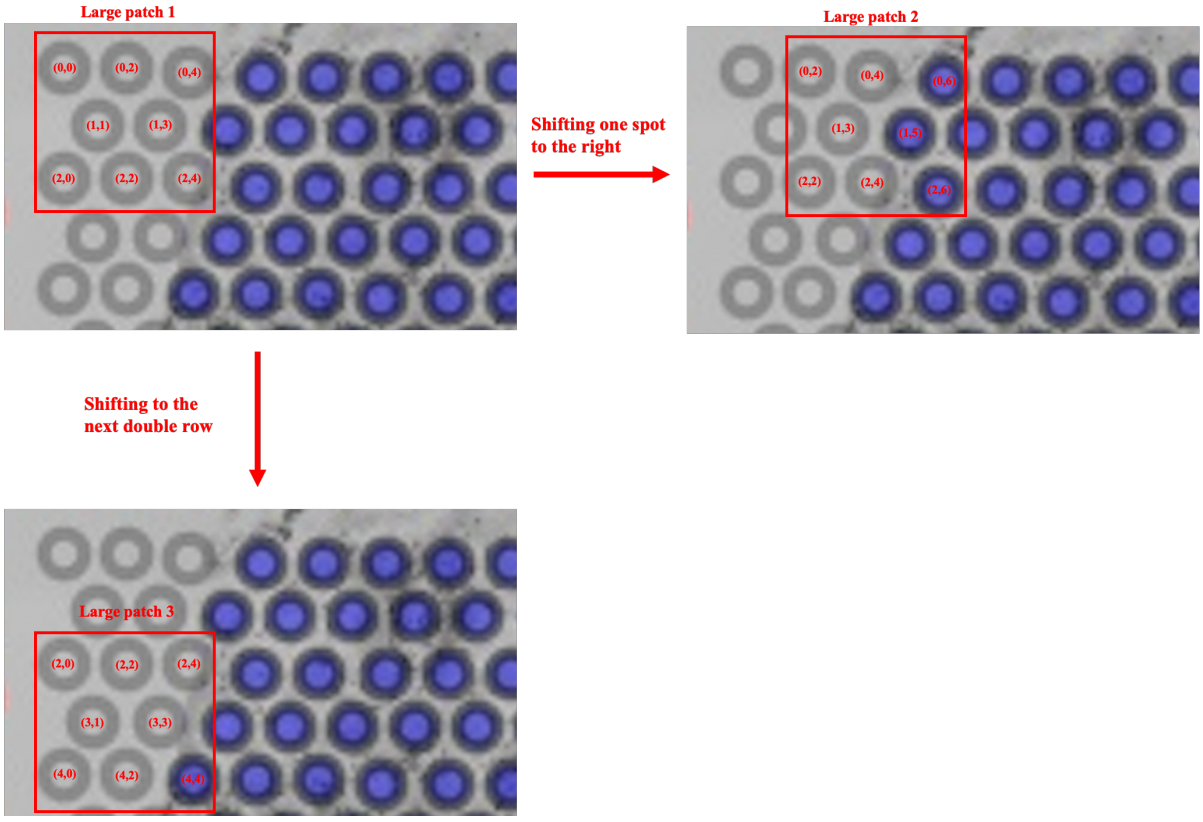


Figure 2.6: Process of generating large patches.

2.2.2 Stain Normalization

For the data from Wu et al. [2021], we reduced the color and intensity biases present in stained images from different laboratories by stain normalization. Two methods were applied, namely the Macenko method (Macenko et al. [2009]) and the Vahadane method (Vahadane et al. [2016]). These algorithms transfer the color style of the source image to that of a carefully picked target image while preserving the other information in the processed image. Figure 2.7 provides two examples of stain normalization, one from sample 1142243F and another from sample CID4465.

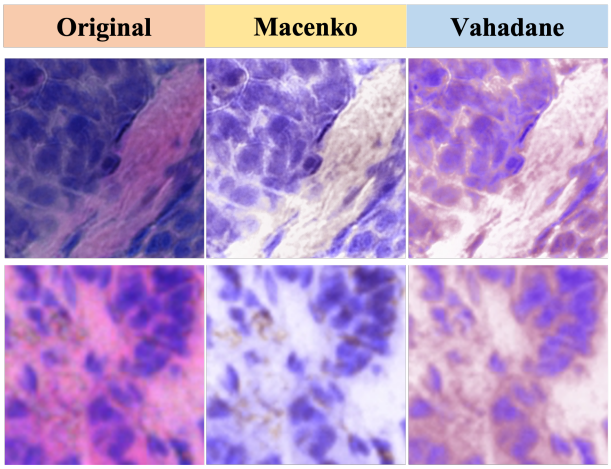


Figure 2.7: Examples of stain normalization.

2.3 Ground Truth for Prediction Tasks

2.3.1 Regression Task: Cell Type Deconvolution

We employed a cell type deconvolution algorithm called CARD (Conditional AutoRegressive model-based Deconvolution) to estimate the composition of cell types in spatial transcriptomics data. CARD takes advantage of the spatial information available in the data and performs reference-based deconvolution, as described in the work by Ma and Zhou [2022]. The relevant cell types of a tissue need to be determined for each tissue type separately, and then scRNA-seq data of these cell types are needed as input for CARD. CARD provides estimates of the proportions of relevant cell types within each spot.

We ran CARD on the ST data for breast cancer obtained from both 10X Genomics and Wu et al. [2021]. We did not run CARD on the ST data retrieved from Sudmeier et al. [2022] due to a lack of scRNA-seq data with annotations for relevant cell types. We used a breast cancer scRNA-seq dataset (GSE176078) provided by Wu et al. [2021] as the reference for cell type deconvolution. The dataset consists of scRNA-seq (Chromium, 10X Genomics) on 26 primary tumors from three major clinical subtypes of breast cancer, including 11 ER+, 5 HER2+, and 10 TNBC tumors. The scRNA-seq identified nine major cell types and 29 minor cell types (Figure 2.8).

In our regression tasks, we focused on predicting the proportions of nine major cell types, as well as four collapsed cell types that were defined as follows: Invasive Cancer (Cancer Epithelial), Stroma (Endothelial,

PVLs, or CAFs), Lymphocyte (T cells, B cells, or Plasmablasts), and Others (Normal Epithelial or Myeloid). We consider these collapsed cell types because some of the nine major cell types have low proportions across patches, making them harder to predict, and the estimates of cell type proportions may be more accurate for collapsed cell types.

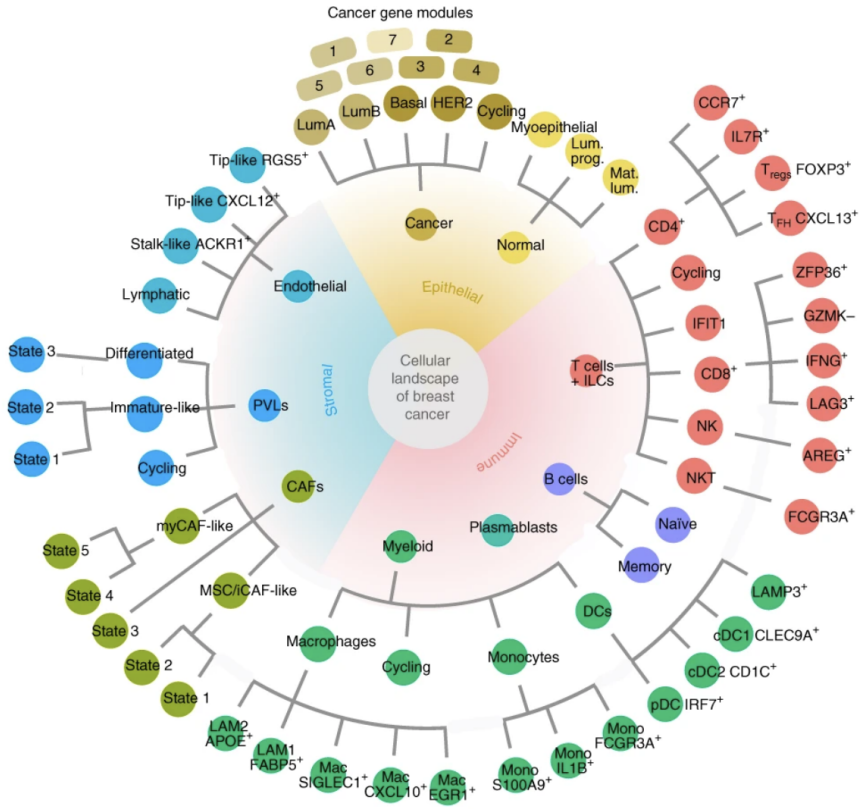


Figure 2.8: Summary of the epithelial, immune, and stromal cell types identified by scRNA-seq grouped by their major (inner), minor, and subset (outer) level classification tiers. Obtained from Figure 8(g) of Wu et al. [2021].

2.3.2 Classification Task: Annotation

We ran STApath for classification tasks on two datasets. For the breast cancer dataset obtained from Wu et al. [2021], the authors provided classifications of ST spots for all samples, as listed in Table 2.2.

	1142243F	1160920F	CID4290	CID4465	CID44971	CID4535	Total
Adipose tissue	0	83	0	0	0	8	91
Artefact	119	48	7	4	1	23	202
Cancer trapped in lymphocyte aggregation	0	9	0	0	0	0	9
Ductal carcinoma in situ (DCIS)	0	12	0	0	273	0	285
Invasive cancer	0	0	0	0	0	418	418
Invasive cancer + adipose tissue + lymphocytes	0	0	0	0	0	3	3
Invasive cancer + lymphocytes	0	0	0	0	317	361	678
Invasive cancer + stroma	0	0	2082	0	0	0	2082
Invasive cancer + stroma + lymphocytes	3627	3146	215	1131	0	0	8119
Lymphocytes	15	186	0	0	81	69	351
Necrosis	568	0	0	0	0	0	568
Normal + stroma + lymphocytes	0	0	0	0	240	0	240
Normal duct	0	0	0	3	0	0	3
Normal glands + lymphocytes	0	278	0	0	0	0	278
Stroma	445	1132	122	73	134	169	2075
Stroma + adipose tissue	0	0	0	0	114	0	114
Tertiary lymphoid structures (TLS)	10	0	0	0	0	0	10
Uncertain	0	0	0	0	0	73	73
NA	0	1	6	0	2	3	12

Table 2.2: Number of spots with each pathology annotation in each patient.

For the brain metastasis data obtained from Sudmeier et al. [2022], the capture spots were clustered based on their gene expression by the authors, and each cluster was annotated by their gene expression and their appearance in H&E staining, as listed in Table 2.3 and illustrated in Figure 2.5(c).

Cluster	Number of Spots
1: Inflammation	451
2: Blood	289
3: Tumor, necrotic	231
4: Tumor	218
5: Tumor, inflammation-adjacent	204
6: Blood	176
7: Inflammation and blood vessels	147

Table 2.3: Number of spots within each cluster defined by Sudmeier et al. [2022].

2.4 Neural Network Architecture for Two Tasks in Whole-Slide Images

ResNets have become the established standard for supervised transfer learning due to their superior performance and efficiency in comparison to other CNN models (He et al. [2016]). We adopted neural networks built upon a ResNet-50 architecture that had been pre-trained on the ImageNet dataset. To design specifically for each of our benchmark tasks, we further fine-tuned the pre-trained ResNet-50 model using transfer

learning techniques.

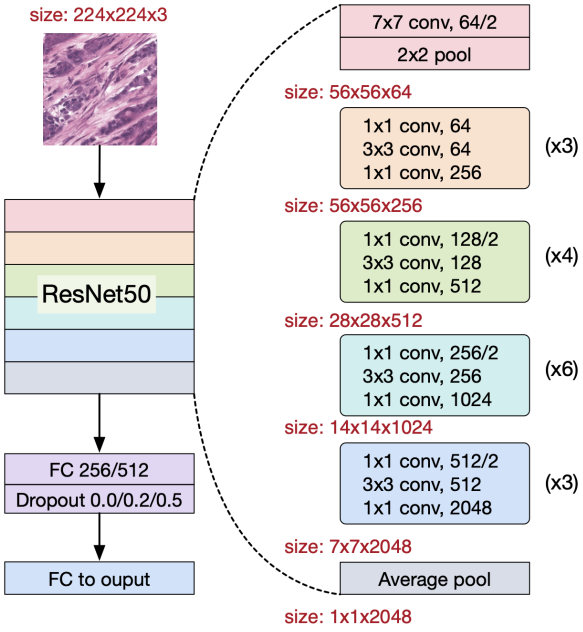


Figure 2.9: CNN architecture overview.

2.4.1 CNN Architecture

STApath utilizes the ResNet-50 model as the base model, which is a pre-trained model consisting of 50 layers. However, we excluded the last output layer with 1000 neurons of the ResNet-50 model, as it focuses on learning task-specific features that were not suitable for our benchmark tasks. To preserve the already learned weights in the base model, we froze the layers, preventing any further weight initialization.

To adapt the base model to our specific benchmark tasks, we introduced new trainable layers on top of the base model. These additional layers were responsible for converting the extracted features from the base model into predictions that aligned with our regression or classification tasks. The final output layer, which is a Fully Connected (FC) layer, played a crucial role in this transformation. Its number of neurons was adjusted based on the nature of each task. For the regression task, the number of neurons in the final output layer matched the number of different cell types we aimed to predict. On the other hand, for the classification task, the number of neurons in the final output layer corresponded to the number of distinct classes within the dataset. The output layer used the softmax activation function for all tasks. This design

allowed our model to generate predictions compatible to the output expectations of each task.

We introduced an additional hidden FC layer using the ReLU activation function between the newly added final output layer and the base model. The purpose of this hidden FC layer was to increase the number of trainable parameters that could be fine-tuned from scratch. By having more parameters to optimize, we aimed to enhance the model's ability to capture task-specific patterns and nuances. To mitigate the risk of overfitting, we incorporated a dropout layer immediately after the intermediate FC layer. Dropout is a regularization technique that randomly deactivates a fraction of the neurons during training, preventing them from relying too heavily on specific features or co-adapting. We tuned the dropout proportion as well as the number of neurons in the hidden FC layer.

2.4.2 Training and Evaluation

While we trained different networks for different datasets and tasks, we always divided the available images into three subsets: 70% for training, 15% for validation, and another 15% for testing. This division allowed us to train our CNN models while also assessing their performance on unseen data. We fine-tuned the hyperparameters of the models, observing any trends in network performance as we made adjustments. It is important to note that the division of the training, validation, and testing datasets remained consistent for networks working on the same dataset and performing the same task. This ensured fairness and comparability between the different network models, allowing us to accurately assess and compare their respective performances.

To train the models, we employed different optimizers, including Adam, SGD, and RMSprop, with the goal of minimizing the categorical entropy loss for classification tasks or the mean squared error loss for regression tasks. To enhance the model's learning, we experimented with varying learning rates and batch sizes.

We saved the weights of the model after each iteration if it demonstrated an improvement in loss on the validation dataset. We set the maximum number of training epochs to 1000 for all CNN models. Additionally, we implemented early stopping based on the validation loss, allowing for a patience of 30 epochs. If the validation loss did not decrease within this patience period, we terminated the training process.

To evaluate the model's performance on the testing dataset, we employed specific metrics depending on

the task. For classification tasks, we utilized the F1 score and categorical entropy loss, while for regression tasks, we measured the mean absolute error and mean squared error. These evaluation metrics provided insights into the accuracy and precision of the model's predictions, giving us a comprehensive understanding of its performance on the test data.

Table 2.4 summarizes the regression tasks on different datasets, along with the corresponding varying parameters of network architecture and training parameters.

	Data	# of Cell Types	Patch	Image Type	Optimizer	Learning Rate	Batch Size	Dropout Layer Proportion	Hidden Dense Layer Units
Section 5.1.1	10X Genomics FFPE breast tissue	Nine	Standard	Original	Adam, SGD	0.01, 0.001, 0.0001	16, 32, 64, 128	0.0, 0.2, 0.5	256, 512
Section 5.1.2	10X Genomics FFPE breast tissue	Four	Standard	Original	Adam	0.01, 0.001, 0.0001	16, 32, 64, 128	0.0, 0.2, 0.5	256, 512
Section 5.2.1	10X Genomics fresh frozen breast tissue	Nine	Standard	Original	Adam, SGD	0.01, 0.001, 0.0001	16, 32, 64, 128	0.0, 0.2, 0.5	256, 512
Section 5.2.2	10X Genomics fresh frozen breast tissue	Four	Standard	Original	Adam	0.01, 0.001, 0.0001	16, 32, 64, 128	0.0, 0.2, 0.5	256, 512
Section 5.3.1	Both 10X Genomics breast tissues	Nine	Standard	Original	Adam	0.001, 0.0001	16, 32, 64, 128	0.0, 0.2, 0.5	256, 512
Section 5.3.2	Both 10X Genomics breast tissues	Four	Standard	Original	Adam	0.001, 0.0001	16, 32, 64, 128	0.0, 0.2, 0.5	256, 512
Section 6.1.1	Six breast tissue samples from Wu et al.	Four	Standard	Original, Macenko, Vahadane	Adam	0.01, 0.001, 0.0001	32, 64, 128, 256, 512	0.0, 0.2, 0.5	0, 256, 512
Section 6.1.2	Six breast tissue samples from Wu et al.	Nine	Large	Original	Adam, RMSprop, SGD	0.01, 0.001, 0.0001	32, 64, 128, 256, 512	0.0, 0.2, 0.5	0, 256, 512
Section 6.1.3	Six breast tissue samples from Wu et al.	Four	Large	Original	Adam, SGD	0.01, 0.001, 0.0001	32, 64, 128, 256, 512	0.0, 0.2, 0.5	256, 512
Section 6.2.1	Two breast tissue samples processed by an independent lab	Nine	Standard	Original	Adam	0.001, 0.0001	32, 64, 128, 256	0.0, 0.2, 0.5	256, 512
Section 6.2.2	Two breast tissue samples processed by an independent lab	Four	Standard	Original	Adam	0.001, 0.0001	32, 64, 128, 256	0.0, 0.2, 0.5	256, 512
Section 6.3.1	Four breast tissue samples processed by Wu et al.	Nine	Standard	Original	Adam	0.001, 0.0001	32, 64, 128, 256	0.0, 0.2, 0.5	256, 512
Section 6.3.2	Four breast tissue samples processed by Wu et al.	Four	Standard	Original	Adam	0.001, 0.0001	32, 64, 128, 256	0.0, 0.2, 0.5	256, 512

Table 2.4: Summary of regression tasks using different datasets together with the corresponding varying parameters of the network architecture.

Chapter 3

Cell Type Proportion Estimates

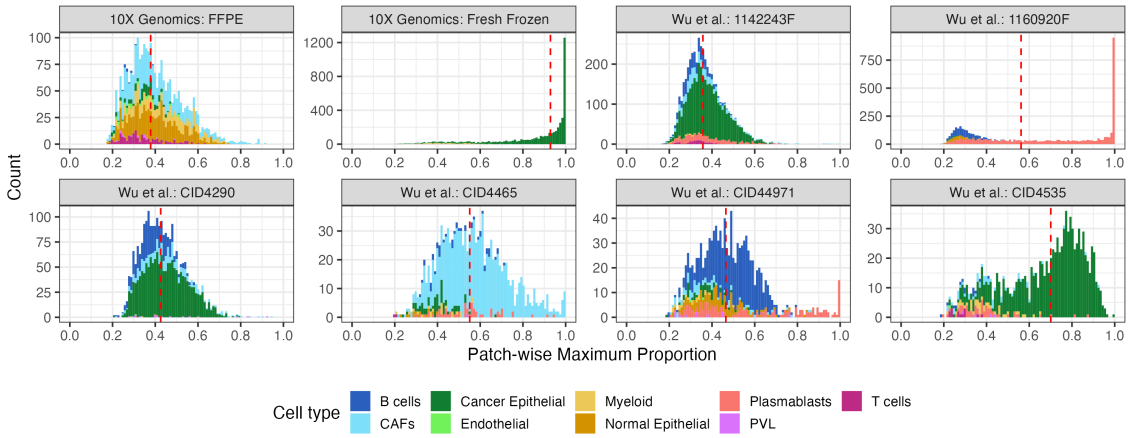
This chapter covers the results of cell type deconvolution using CARD on the breast tissue sections from 10X Genomics and Wu et al. [2021]. Based on the scRNA-seq reference we used, there were nine major cell types and we further calculated the proportions of four collapsed cell types, as described in Chapter 2.

	Number of Patches							
	10X Genomics		Wu et al. [2021]					
	FFPE	Fresh Frozen	1142243F	1160920F	CID4290	CID4465	CID44971	CID4535
B cells	13	18	547	878	531	46	694	14
CAFs	980	49	545	0	324	1008	78	51
Cancer Epithelial	127	4516	3144	61	1556	65	46	946
Endothelial	49	43	0	0	2	0	2	1
Myeloid	431	225	22	1	1	21	19	43
Normal Epithelial	736	0	19	273	0	12	137	0
Plasmablasts	8	33	406	3681	2	56	176	54
PVL	17	13	0	1	16	1	10	4
T cells	157	1	98	0	0	2	0	14

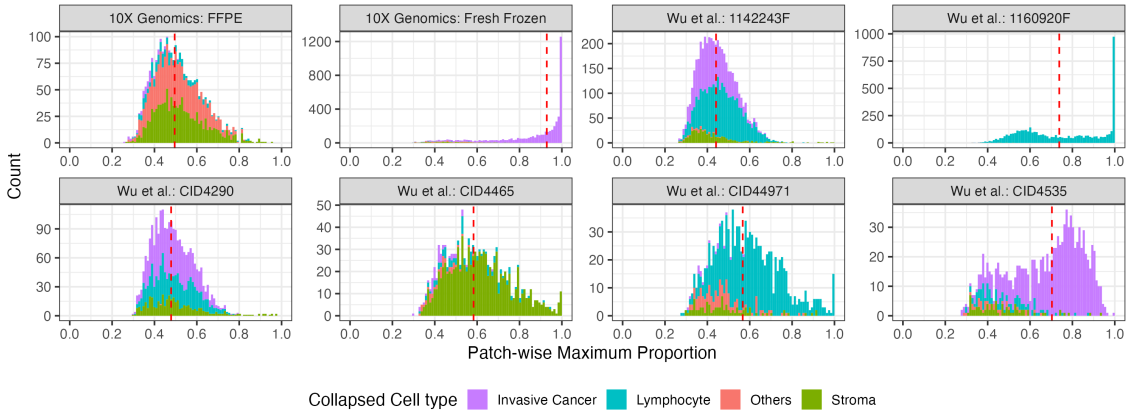
Table 3.1: Number of standard patches in each tissue section with each of the nine major cell types having the highest proportion. PVL: perivascular-like stroma cells.

	Number of Patches							
	10X Genomics		Wu et al. [2021]					
	FFPE	Fresh Frozen	1142243F	1160920F	CID4290	CID4465	CID44971	CID4535
Invasive Cancer	73	4427	1853	2	1197	49	18	922
Lymphocyte	222	71	2447	4882	811	110	926	96
Stroma	1059	181	425	4	423	1007	89	75
Others	1164	219	56	7	1	45	129	34

Table 3.2: Number of standard patches in each tissue section with each of the four collapsed pathological cell types having the highest proportion.

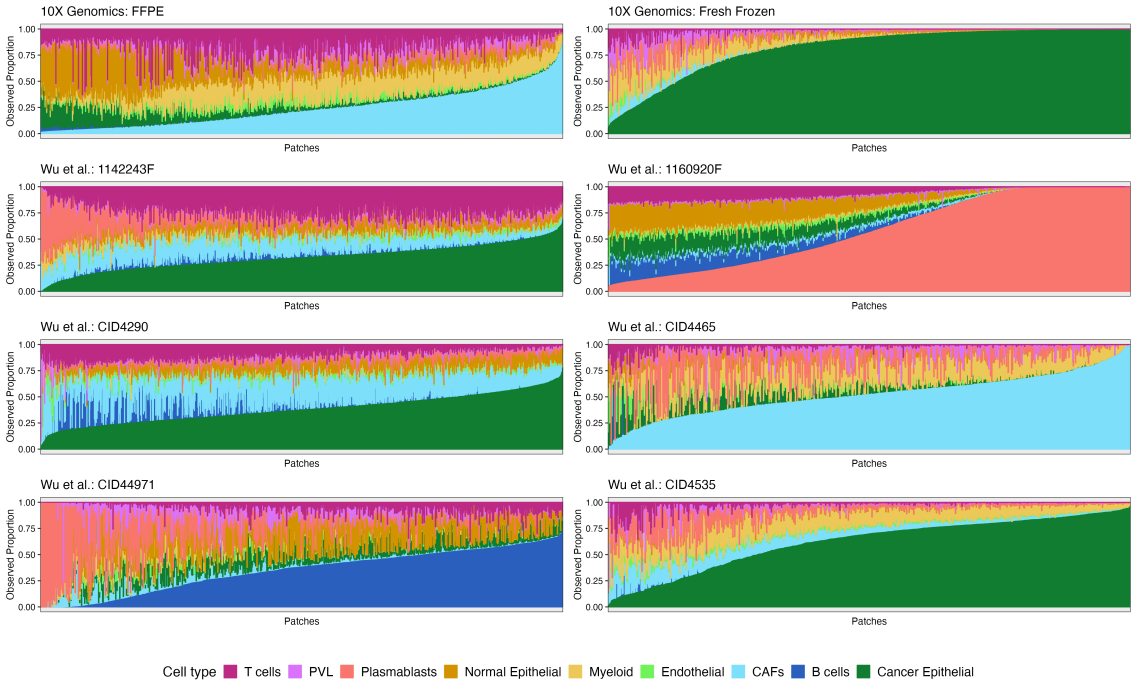


(a) Nine major cell types.

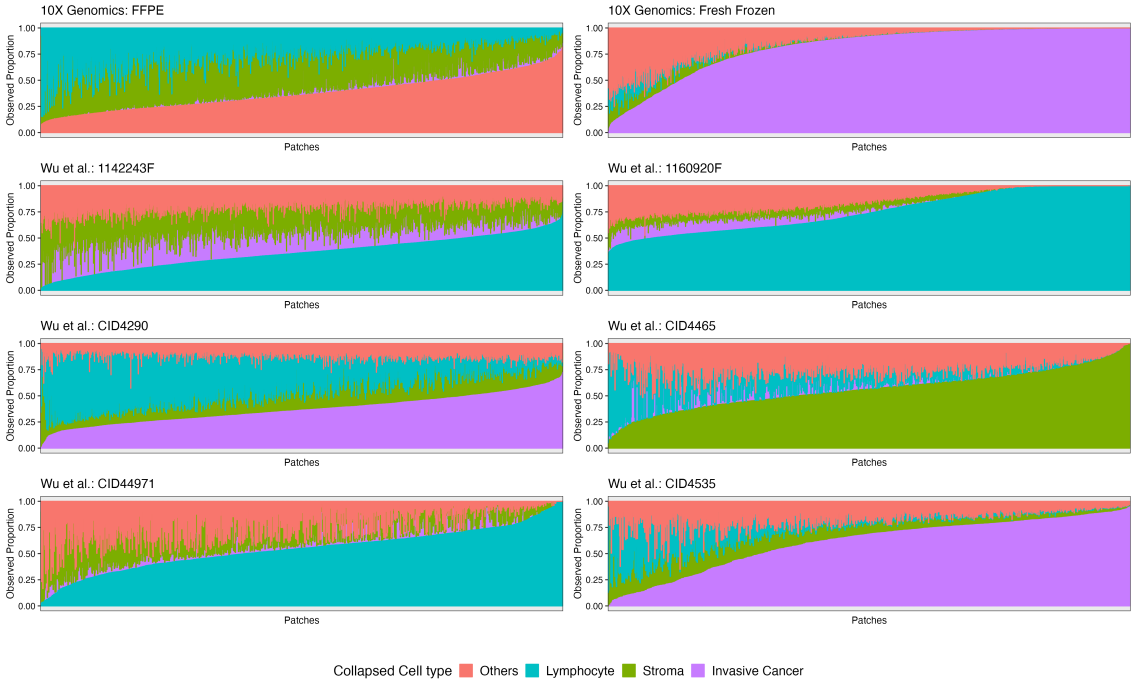


(b) Four collapsed cell types.

Figure 3.1: Histogram of the maximum cell type proportion for each patch, stratified by samples, with the median of the maximum proportions per sample indicated by red dashed lines.

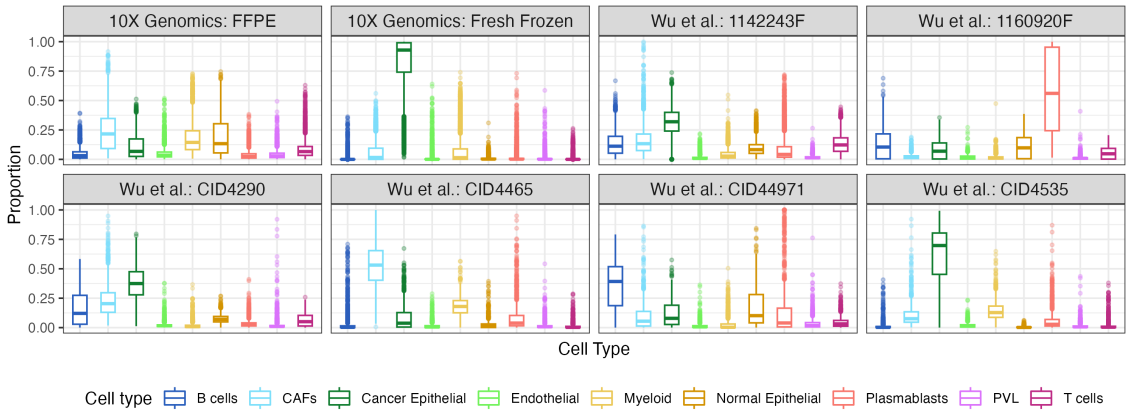


(a) Nine major cell types.

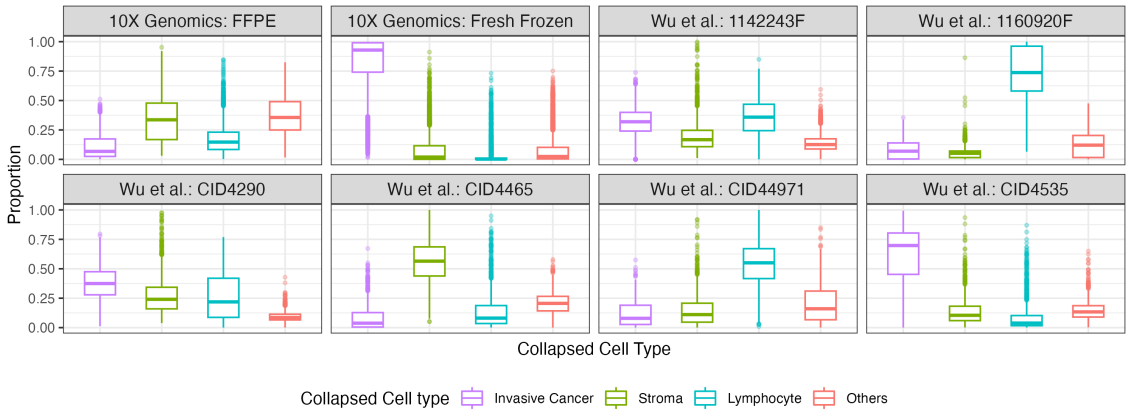


(b) Four collapsed cell types.

Figure 3.2: Cell type compositions of all patches in each sample, with the patches arranged in order based on the proportion of the cell type that most frequently exhibited higher proportions compared to other cell types.



(a) Nine major cell types.



(b) Four collapsed cell types.

Figure 3.3: Boxplot showing the distribution of cell type proportions in all the patches, stratified by sample.

3.1 10X Genomics

A total of 4849 standard patches were obtained from the 10X Genomics fresh frozen breast tissue and subjected to cell type deconvolution. Within this tissue, there was a predominance of cancer epithelial cells, with 4516 patches exhibiting cancer epithelial cells as the most abundant cell type (Table 3.1). The proportion of cancer epithelial cells in the majority of the patches was significantly higher than that of other cell types, as depicted in Figure 3.1(a). This observation suggests a lack of diversity in cell type proportions within the fresh frozen tissue sample. To simplify the analysis, we combined the nine major cell types into four collapsed pathological cell types. Despite this consolidation, invasive cancer cells remained the most abundant cell type (Table 3.2 and Figure 3.1(b)), with a substantial number of patches being predominantly

composed of invasive cancer cells (Figure 3.2, 3.3). For the patches with a relatively lower proportion of invasive cancer, the most prevalent cell type was classified as "Others," encompassing normal epithelial cells and myeloid cells.

A total of 2518 standard patches were extracted from the 10X Genomics FFPE breast tissue and underwent cell type deconvolution. In contrast to the fresh frozen tissue, this FFPE tissue exhibited greater diversity in terms of cell type abundance. CAF was most frequently identified as the cell type with the highest proportion, followed by normal epithelial cells (Figure 3.1). As these major cell types belong to the Stroma and Others categories, respectively, the most abundant collapsed cell type in the majority of patches was either stroma or others (Table 3.2). Additionally, unlike the fresh frozen tissue, most patches in the FFPE tissue had a mixture of more than one cell type. Overall, invasive cancer was not the predominant cell type in the majority of patches (Figure 3.2) and on average it had the smallest proportion over all patches (Figure 3.3).

3.2 Wu et al. [2021]

3.2.1 Cell Type Proportions Obtained by CARD

The number of patches subjected to cell type deconvolution varied across the samples: 4781, 4895, 2432, 1211, 1162, and 1127 for samples 1142243F, 1160920F, CID4290, CID4465, CID44971, and CID4535, respectively.

In sample 1142243F, the predominant cell type in most patches was cancer epithelial (Table 3.1). However, after collapsing the major cell types, lymphocytes became the predominant cell type in most patches, followed by invasive cancer, which had the highest proportion in the second most patches (Table 3.2). Among all nine cell types per patch, the highest proportion had a median below 0.4, ranging from approximately 0.2 to 0.7, indicating that most patches were not pure in one specific cell type (Figure 3.1(a)). This is true also for the collapsed cell types (Figure 3.1(b)). Besides, in general, an increase in cancer epithelial proportion was associated with a decrease in plasmablasts proportion (Figure 3.2). Cancer epithelial, CAFs, B cells, and T cells had higher average proportions than other cell types (Figure 3.3).

In sample 1160920F, plasmablasts were the predominant cell type in most patches, aligning with the

observation that nearly all patches had lymphocytes as the most abundant pathological cell type after collapsing the major cell types (Table 3.1, 3.2). Furthermore, the patches displayed a relatively high level of purity, as evidenced by the median proportion of lymphocytes being approximately 0.7 and around 25% of patches exhibiting a lymphocyte proportion of 1 (Figure 3.1(b), 3.2). Other than the plasmablasts, B cells, cancer epithelial, normal epithelial, and T cells had on average higher proportions than CAFs, endothelial, myeloid, and PVL (Figure 3.3).

In sample CID4290, the predominant cell type in the majority of patches was cancer epithelial (Table 3.1), which is consistent with the finding that approximately half of the patches had invasive cancer as the most prevalent collapsed cell type (Table 3.2). However, B cells and CAFs were also highly abundant in a few patches, resulting in the other half of the patches being characterized by lymphocytes or stroma as the most abundant pathological cell type, with lymphocytes being more prevalent (Table 3.2). The proportion of predominant cell type in each patch was centered around 0.4 to 0.5, indicating intermediate purity in a substantial number of patches (Figure 3.1). Overall, a higher presence of invasive cancer was associated with a lower proportion of lymphocytes, while the proportions of stroma and other cell types remained relatively stable (Figure 3.2(b)). Furthermore, endothelial cells, myeloid cells, plasmablasts, and PVLs were infrequent in most patches, except for a few patches with exceptionally high proportions of PVL (Figure 3.3(a)). This finding can partially explain patches displaying unusually high proportions of stroma, as depicted in Figure 3.3(b).

In sample CID4465, CAF was the most prevalent cell type in the majority of patches, exhibiting a wide range of proportions, ranging from 0.2 to 1.0. (Table 3.1, Figure 3.1(a)). A similar pattern was observed for the collapsed cell types, with stroma being the dominant cell type in most patches (Table 3.2, Figure 3.1(b)). This suggests that there is diversity in the purity of the patches, ranging from patches that are purely composed of CAFs to those that contain a mixture of all cell types with each proportion being less than 0.2 or relatively pure with other cell types. For example, although on average, invasive cancer and lymphocytes had lower proportions compared to stroma and other cell types in most patches, there were patches where more than half of the composition consisted of invasive cancer or lymphocytes, as depicted in Figure 3.2(b) and Figure 3.3(b)).

In sample CID44971, B cells were the most prevalent cell type in over half of the patches, as indicated

in Table 3.1. Some other patches had plasmablasts and normal epithelial cells as the most prevalent cell types, resulting in lymphocytes being the most abundant pathological cell type in the majority of patches and "others" emerged as the second most prevalent cell type in terms of frequency (Table 3.2). The patch-wise maximum proportion of B cells ranged from 0.2 to 0.7, while the patch-wise maximum proportion of plasmablasts ranged from 0.2 to 1.0, making the proportion of the collapsed cell type, lymphocyte, had a wide range of proportion when it was the most abundant cell type in the patch (Figure 3.1). Considering the collapsed proportions, the average proportion of invasive cancer was lower compared to other cell types, and its range of proportion was also more limited than that of other pathological cell types (Figure 3.3).

In sample CID4534, cancer epithelial cells dominated the majority of patches (Table 3.1), aligning with invasive cancer being the predominant pathological cell type across most patches (Table 3.2). Among patches with cancer epithelial as the predominant cell type, the proportion of cancer epithelial had a wide range from 0.3 to 1.0, indicating the diversity of purity with respect to cancer epithelial (Figure 3.1). Note that more than half of these patches had more than 80% cancer epithelial per patch. On the other hand, even in patches where other cell types were the most prevalent, the corresponding dominant cell types exhibited relatively lower proportions (mostly from 0.3 to 0.6) compared to when cancer epithelial cells were dominant (Figure 3.1). This implies that, even within the few patches where cancer epithelial cells did not have the highest proportion, those patches were not pure in other cell types. Instead, they consisted of a mixture of multiple cell types, as illustrated in Figure 3.2.

3.2.2 Comparison Between Cell Type Proportions and Pathology Annotations

Pathology annotations for in-tissue capture spots for all samples have been provided by Wu et al. [2021] as documented in Chapter 2. In this section, we sought to compare cell type compositions obtained through cell type deconvolution versus pathology annotations. We found the two types of information were consistent but represented different information. Patches labeled as NA, Uncertain, or Artefact were excluded from our analysis. For simplicity, we only assessed the agreement between the four collapsed cell type proportions and the annotations. The examination of consistency between the nine major cell type proportions and the annotations was carried out in the Appendix A (Figure A.2, A.1). Figure 3.4 and Figure 3.5 illustrates the cell type compositions of patches categorized under different annotations within each sample.

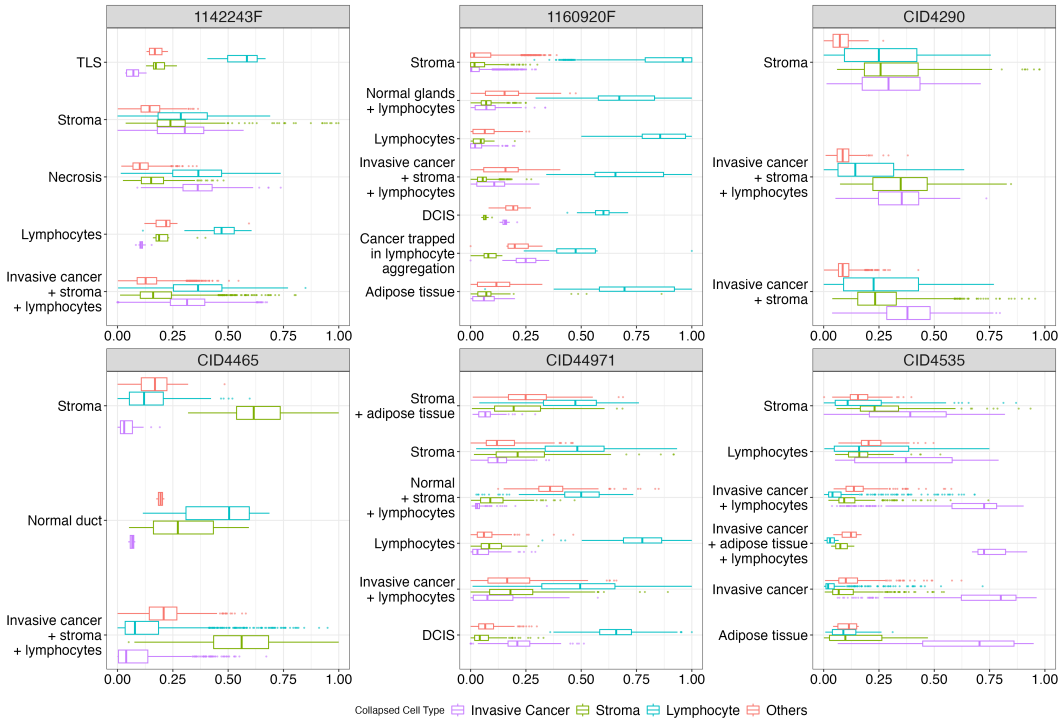


Figure 3.4: Composition of each of the four pathological cell types among patches with each pathology annotation, stratified by sample.

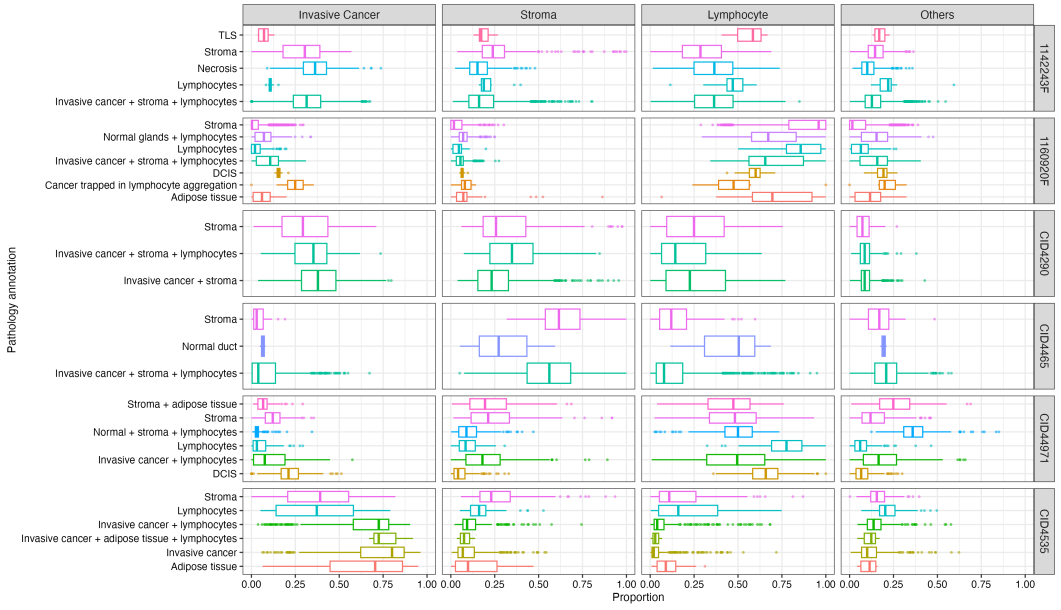


Figure 3.5: Distribution of each of the four pathological cell types among patches with each pathology annotation, stratified by cell types.

In the case of sample 1142243F, patches identified as TLS and Lymphocytes exhibited a higher proportion of lymphocytes compared to the other three cell types, with invasive cancer having the lowest proportion among all four cell types (Figure 3.4). Furthermore, when compared to patches labeled differently, patches labeled as TLS and Lymphocytes showed a lower proportion of invasive cancer, slightly higher proportions of lymphocytes and others, and a relatively similar proportion of stroma (Figure 3.5). These differences in proportions suggest a certain degree of consistency between the cell type deconvolution results and these two pathology annotations. However, it is important to note that out of all the patches in sample 1142243F, only 25 were labeled as TLS or Lymphocyte. Unfortunately, patches labeled as Stroma, Invasive cancer + stroma + lymphocytes, and Necrosis exhibited similar cell compositions (Figure 3.6). In comparison to patches labeled as Invasive cancer + stroma + lymphocytes, patches labeled as Stroma showed comparable proportions of invasive cancer and others (i.e., normal epithelial and myeloid), a slightly higher proportion of stroma, and a slightly lower proportion of lymphocytes. However, these slight differences were not statistically significant (Figure 3.5). Furthermore, except for a few patches with exceptionally high proportions of Stroma, most patches labeled as Stroma did not demonstrate a significantly higher proportion of stroma compared to the other three cell types (Figure 3.4). These findings suggest that the cell type proportions obtained through cell type deconvolution were neither consistent with the expected proportions for certain cell types nor able to differentiate between patches with different labels. Consequently, we can conclude that there is poor consistency between the cell type deconvolution results and the pathology annotations for the majority of patches in sample 1142242F.

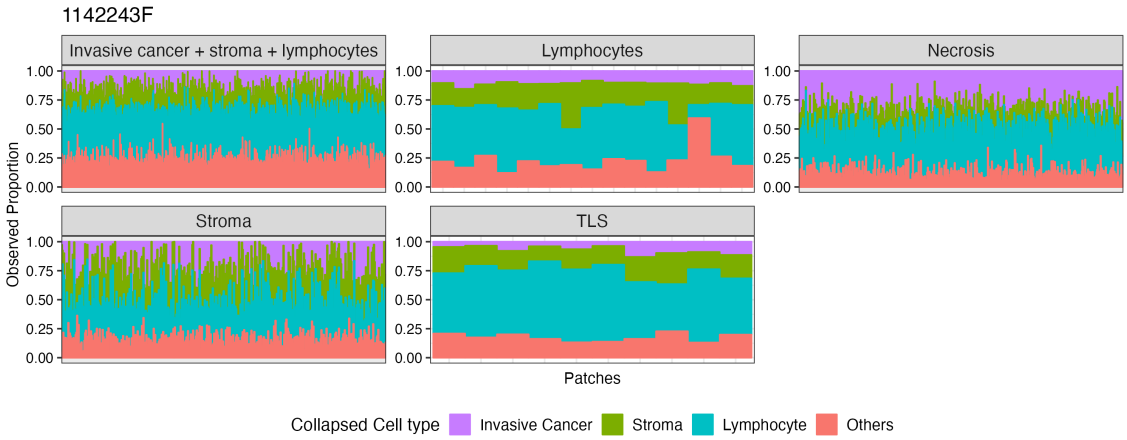


Figure 3.6: Cell type composition of patches from sample 1142243F, stratified by pathology annotations.

For the sample 1160920F, it was observed that lymphocytes had significantly higher proportions compared to other cell types within almost all patches, regardless of their labels (Figure 3.4, 3.5). This finding suggests that the cell type deconvolution results were not able to accurately represent certain pathology annotations that lack lymphocytes, including Adipose tissue, Stroma, and Ductal carcinoma in situ (DCIS). Only a few patches were labeled as Adipose tissue or DCIS, and more than 1000 patches were labeled as Stroma. However, the majority of patches labeled as Stroma exhibited negligible proportions of stroma. Furthermore, for patches labeled as Normal glands + lymphocytes, the proportion of others (i.e., normal epithelial + myeloid) was not as high as expected, and many patches displayed comparable proportions of all cell types except for lymphocytes. Finally, considering the proportion of lymphocytes across different labels (Figure 3.5), patches labeled as pure Lymphocytes had slightly higher proportions of lymphocyte cells compared to patches labeled as Normal glands + lymphocytes or Invasive cancer + stroma + lymphocytes.

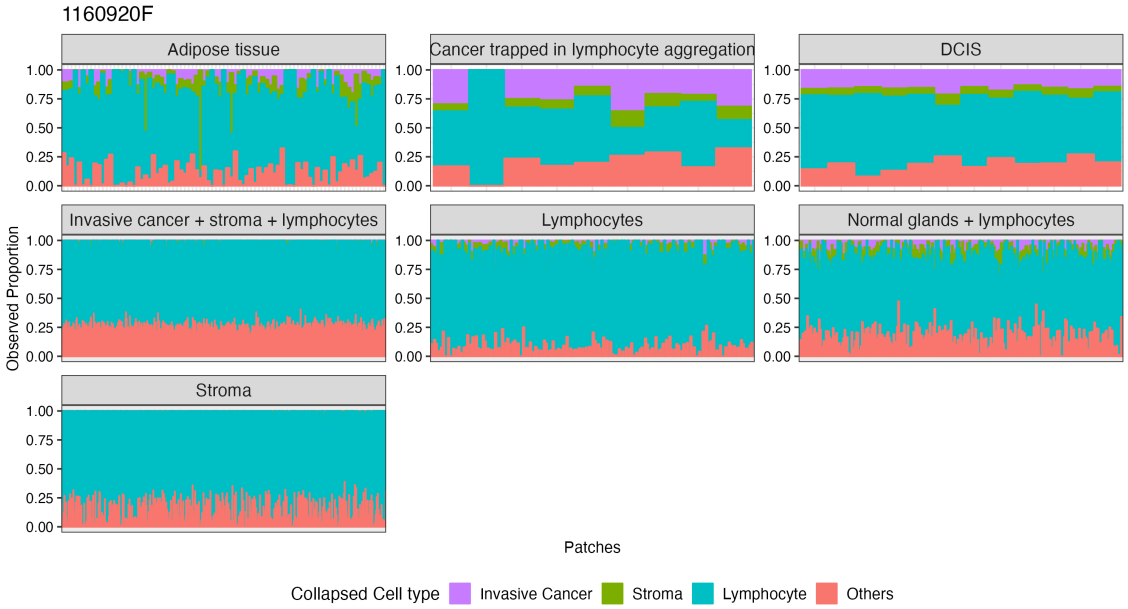


Figure 3.7: Cell type composition of patches from sample 1160920F, stratified by pathology annotations.

In sample CID4290, it was observed that all patches were labeled with annotations that lacked normal cells. This is consistent with the deconvoluted proportion of others (i.e., normal epithelial + myeloid) being lower than other cell types in the majority of patches (Figure 3.4). Similar to the previous samples, patches labeled as Stroma did not exhibit significantly higher proportions of stroma compared to other cell types (Figure 3.4). Overall, the cell type compositions of patches with different annotations were found to be

similar (Figure 3.8). In comparison to patches labeled as Invasive cancer + stroma, patches labeled as Invasive cancer + stroma + lymphocytes did not demonstrate higher proportions of lymphocytes, yet they had slightly higher proportions of stroma (Figure 3.5).

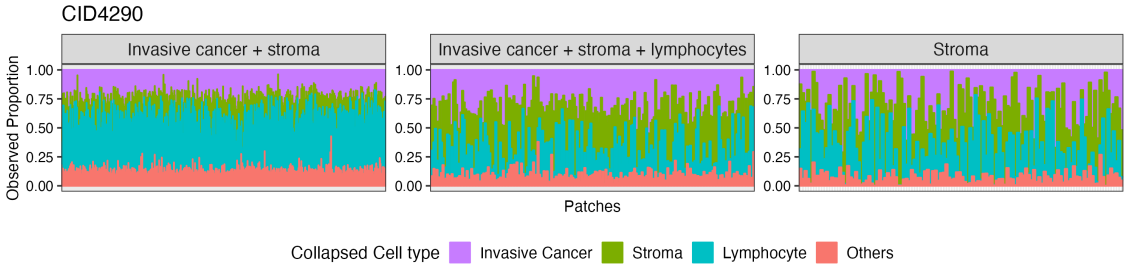


Figure 3.8: Cell type composition of patches from sample CID4290, stratified by pathology annotations.

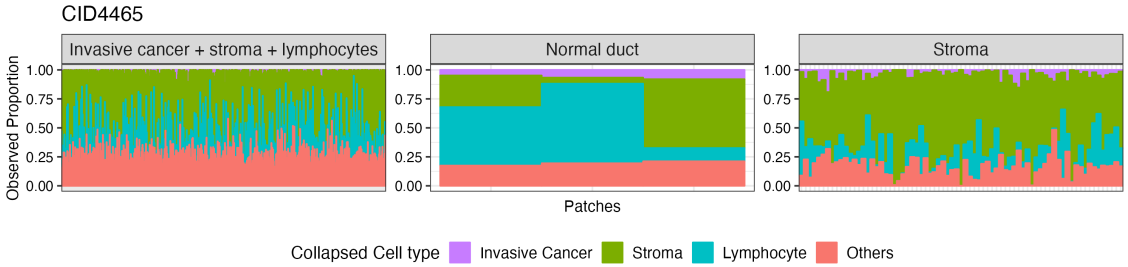


Figure 3.9: Cell type composition of patches from sample CID4465, stratified by pathology annotations.

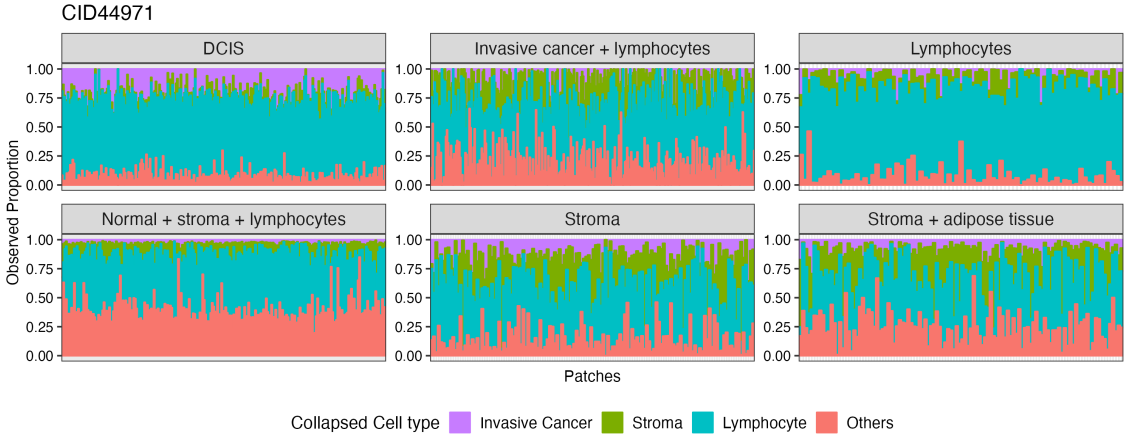


Figure 3.10: Cell type composition of patches from sample CID44971, stratified by pathology annotations.

In sample CID4465, a significant portion of the patches (over 90%) were labeled as Invasive cancer + stroma + lymphocytes, and the majority of the remaining patches were labeled as Stroma. However,

despite patches labeled as Stroma showing significantly higher proportions of stroma compared to other cell types (Figure 3.4), patches labeled as Invasive cancer + stroma + lymphocytes exhibited similar cell type compositions. Therefore, based on the cell type deconvolution results, we were unable to distinguish between these two annotations.

In sample CID44971, the proportion of lymphocytes was found to be on average higher than the proportions of the other three cell types (Figure 3.4). Notably, patches labeled as pure Lymphocytes exhibited relatively higher proportions of lymphocytes compared to patches with other labels (Figure 3.5). The cancer-associated annotations in sample CID44971 were Invasive cancer + lymphocytes and DCIS. Patches with these labels had relatively higher proportions of invasive cancer. Again, patches labeled as Stroma alone did not display high proportions of stroma. Additionally, patches labeled as Normal + stroma + lymphocytes showed higher proportions of others (i.e., normal epithelial + myeloid) compared to other patches that did not contain normal cells in their annotations.

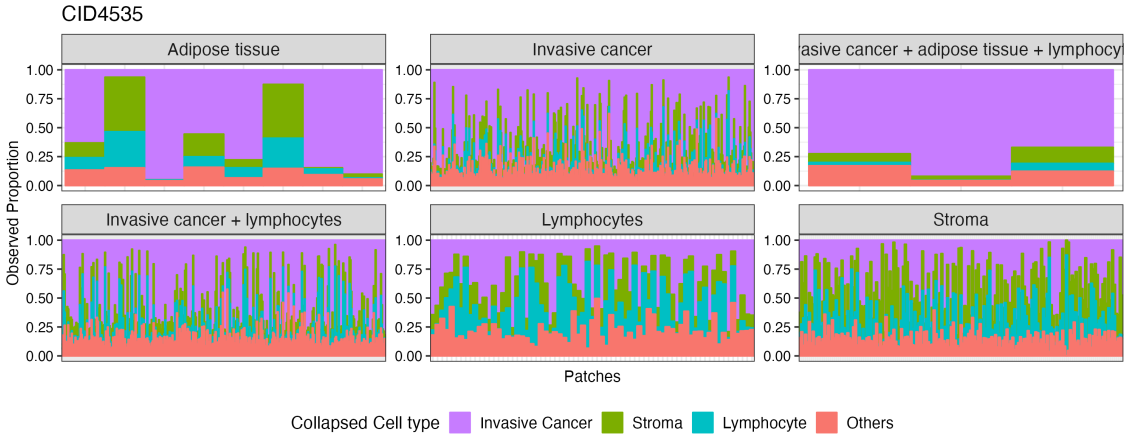


Figure 3.11: Cell type composition of patches from sample CID4535, stratified by pathology annotations.

Many patches from sample CID4535 were annotated as involving Invasive cancer. Patches with these invasive cancer annotations exhibited higher proportions of invasive cancer compared to patches without Invasive cancer in their annotations (Figure 3.5). Interestingly, most of these patches displayed very low lymphocyte proportions, despite some of them having lymphocytes mentioned in their annotations (Figure 3.5). This inconsistency in the proportion of lymphocytes is also evident when comparing patches labeled as Invasive cancer + lymphocytes with patches labeled as Lymphocytes (Figure 3.4). Patches with these two labels demonstrated similar cell type compositions. Furthermore, patches labeled as pure Stroma or pure

Lymphocytes unexpectedly exhibited higher proportions of invasive cancer than the proportions of stroma or lymphocytes, respectively (Figure 3.4), which was also evidence of inconsistency between the cell type deconvolution results and the annotation provided by Wu et al. [2021].

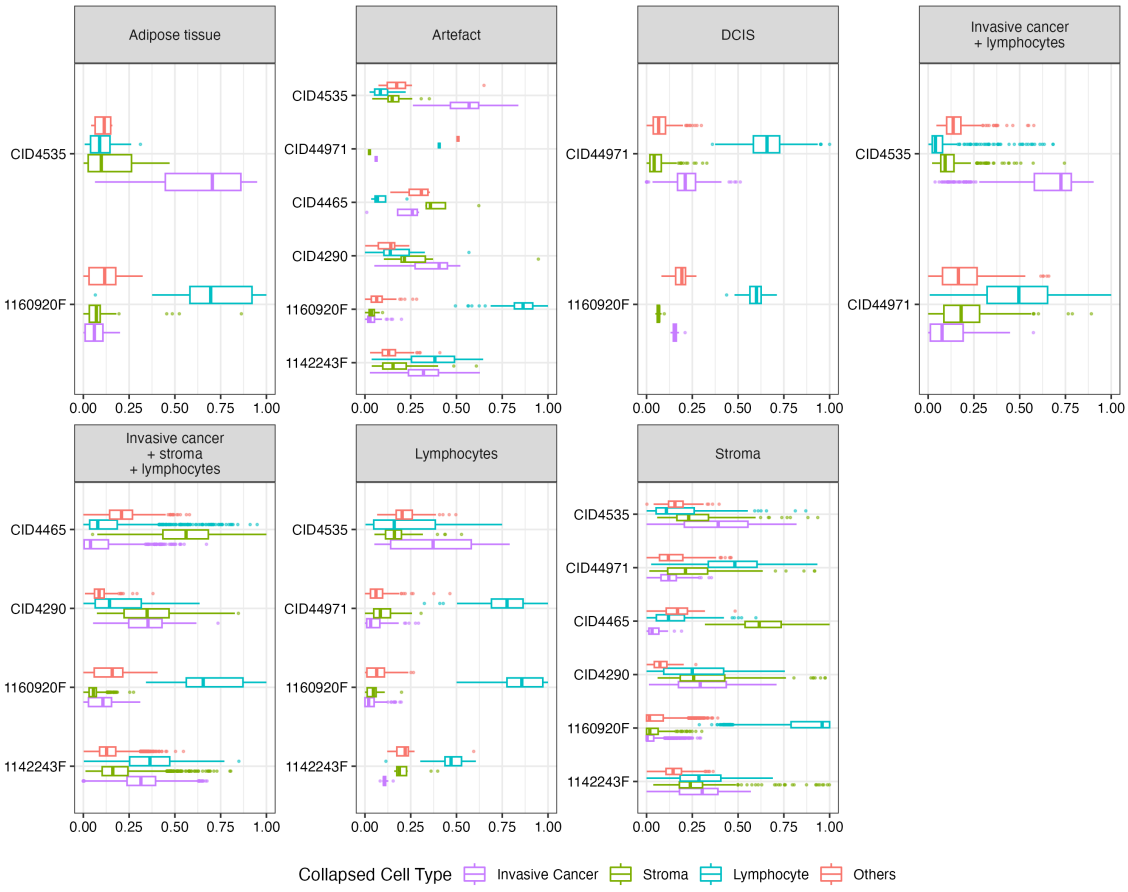


Figure 3.12: Cell type proportions of patches with the same annotation but from different samples are shown. Only annotations that include patches from more than one sample are included.

We also examined whether the cell type deconvolution results varied for patches with the same annotation across different samples (Figure 3.12). For patches labeled as Adipose tissue, the cell type compositions differed significantly among patches from different samples. Specifically, patches from 1160920F displayed significantly higher proportions of lymphocytes compared to other cell types, while those from CID4535 were predominantly composed of invasive cancer. Similar inconsistencies were observed for patches labeled as Invasive cancer + lymphocytes, Invasive cancer + stroma + lymphocytes, and Stroma. However, the cell type compositions of patches labeled as DCIS and Lymphocytes showed relatively consistent pat-

terns across samples CID44971 and 1160920F.

In conclusion, our analysis revealed inconsistencies between the cell type proportions obtained through deconvolution and the pathology annotations provided by Wu et al. [2021]. These results suggest that the cell type deconvolution approach did not align well with the pathology annotations, particularly for annotations related to Stroma. Furthermore, the cell type proportions derived from deconvolution were not effective in differentiating between patches with different labels. Additionally, patches with a single pathology annotation did not exhibit clear purity in the corresponding cell types.

Our original intention was to utilize the deconvoluted cell type proportions to annotate patches and train a neural network for classification tasks. However, these findings underscore the limitations of the cell type deconvolution method in accurately representing specific cell compositions within different pathology annotations. Moreover, the variations in cell type compositions across different samples under the same pathology annotation further complicate the establishment of generalized criteria for creating patch classes and determining cutoff values for each cell type in each annotation. It is important to note that the accuracy of the pathology annotations provided by Wu et al. [2021] is unknown, and thus, the inconsistencies observed may not necessarily be attributed to the cell type deconvolution approach alone. In future research, we should consider conducting gold standard evaluations for deconvolution and pathologist annotation. Realistic simulation tools like scDesign3 (Song et al. [2023]) could potentially be helpful in this regard.

Chapter 4

Examination of Features Extracted by Pre-trained ResNet

This chapter focuses on the analysis of features extracted by the pre-trained ResNets. To extract relevant features from each patch of the histological image, we employed a pre-trained ResNet50 as a feature extractor. Our primary objective was to assess whether this neural network was capable of extracting discernible features that differentiated between images with varying cell type compositions.

To accomplish this, we constructed a base model comprising solely the pre-trained ResNet50 that provided 2048 features per image. It is important to note that we did not train the base model with our specific dataset. Instead, we directly utilized this pre-trained base model to extract features from the datasets. Consequently, we obtained 2048 features for each standard patch by passing them through the pre-trained ResNet base model.

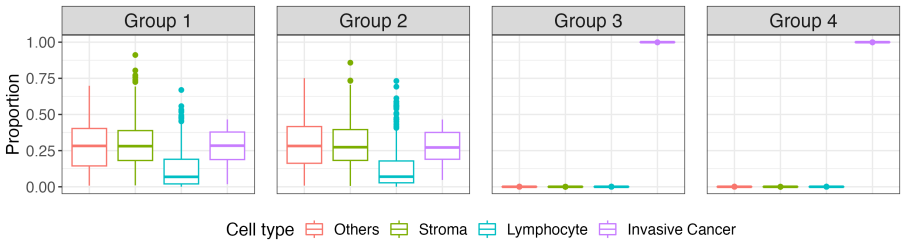
4.1 Comparing Imaging Features Between Patches With Similar or Different Cell Type Proportions.

To evaluate whether the pre-trained ResNet50 could extract informative features for distinguishing patches based on their cell type compositions, we conducted the following analysis. Firstly, we generated four patch groups for each dataset: Groups 1 and 2 with comparable proportions of a specific cell type, and Groups 3

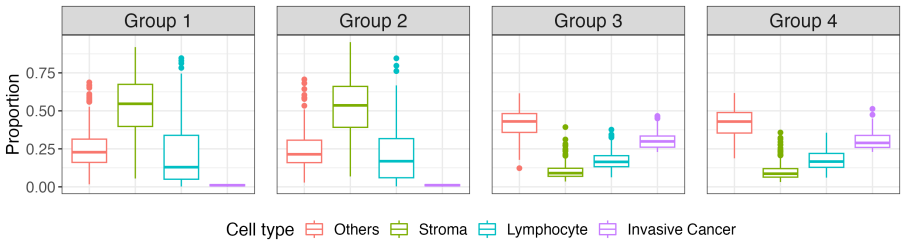
and 4 with similar proportions of that cell type, but with differing proportions compared to Groups 1 and 2. Next, we compared the value of each feature extracted by ResNet50 between any two groups to identify significant differences between the two groups. In each pairwise comparison, for each of the 2048 features, we performed hypothesis testing by T-test or Mann-Whitney U (M-W U) test.

4.1.1 10X Genomics Breast Cancer Tissues

Difference in Invasive Cancer Proportions



(a) 10X Genomics fresh frozen breast tissue



(b) 10X Genomics FFPE breast tissue

Figure 4.1: Cell type composition in four groups of patches from 10X Genomics tissues, categorized by invasive cancer proportion, assessing ResNet50 as feature extractors.

For each of the two 10X Genomics samples, we selected four groups of patches based on the invasive cancer proportions. Specifically, for the 10X Genomics fresh frozen breast tissue section, Groups 3 and 4 exclusively consisted of patches with pure invasive cancer, while Groups 1 and 2 included patches with lower cancer proportions. Each group comprised 300 patches. In the case of the FFPE breast tissue, Groups 1 and 2 contained patches with negligible invasive cancer, while Groups 3 and 4 had patches with some degree of invasive cancer. Each group consisted of 200 patches. Figure 4.1 depicts the cell type composition within the selected patches for the four groups in each tissue section. Notably, for the FFPE tissue sample, Groups 1 and 2 predominantly exhibited stroma, whereas Groups 3 and 4 had limited stroma.

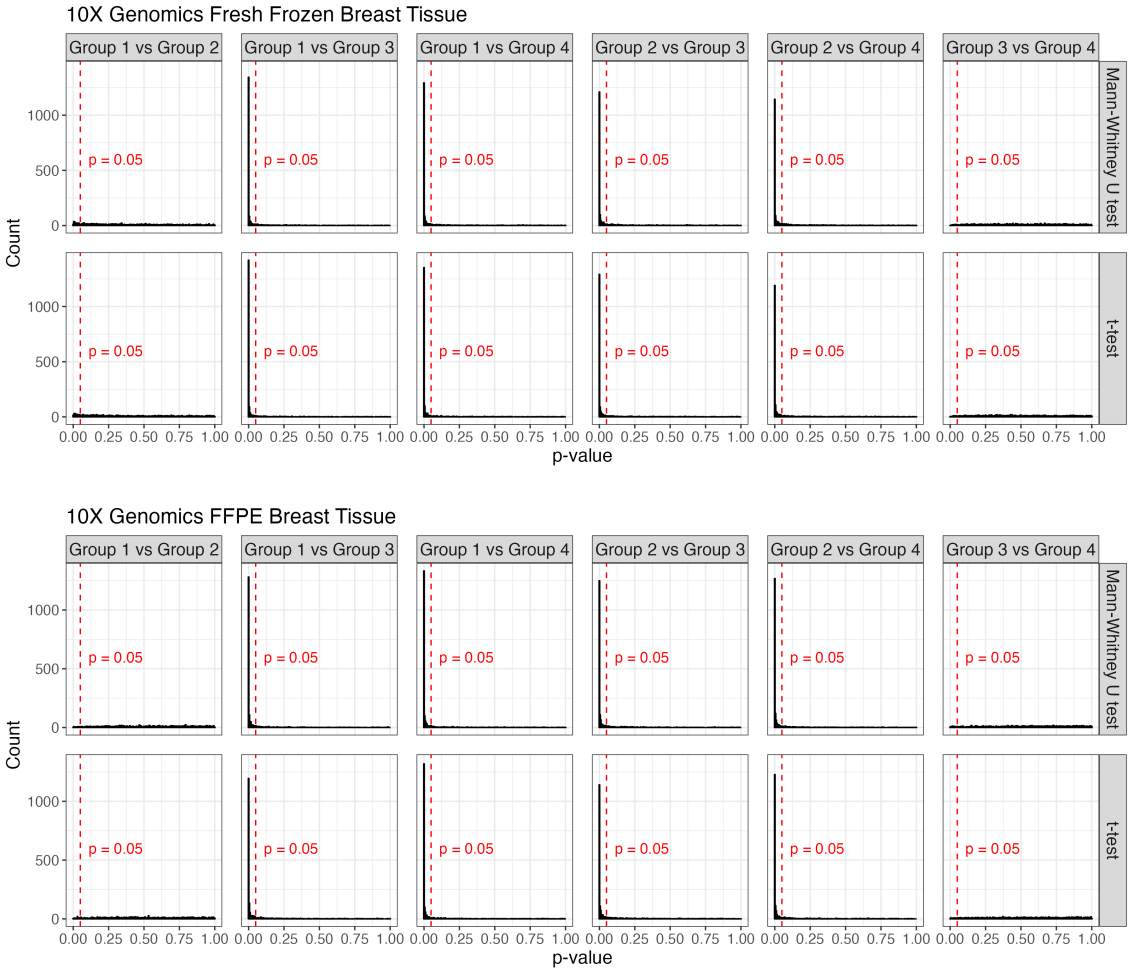


Figure 4.2: Histogram of p-values obtained from pairwise hypothesis testings between four groups of patches from 10X Genomics tissues categorized by invasive cancer proportion.

Figure 4.2 displays the histogram of 2048 p-values obtained from each pairwise hypothesis testing for both tissue samples. Notably, within the same proportion category for both tissues (Group 1 vs. 2 or Group 3 vs. 4), the histogram reveals that less than 5% of features had p-values below 0.05 (Table 4.1). This indicates that there were no significant differences in the overall features extracted by ResNet50 when the patches had similar proportions of invasive cancer. However, when comparing pairs of groups from different proportion categories, over 70% of the features had p-values below 0.05 (Table 4.1). This suggests that if the true proportion of invasive cancer in the patches differed significantly, the features extracted by ResNet50 also exhibited significant differences. These observations highlight the ResNet50's capability to extract informative morphological features that characterize the proportion of invasive cancer, regardless of

the proportion of other cell types.

		10X Fresh Frozen Tissue		10X FFPE Tissue	
		t-test	M-W U test	t-test	M-W U test
Similar tumor prop.	Group 1 vs. 2	218 (10.6%)	234 (11.4%)	78 (3.8%)	65 (3.2%)
	Group 3 vs. 4	68 (3.3%)	65 (3.2%)	70 (3.4%)	81 (4.0%)
Different tumor prop.	Group 1 vs. 3	1650 (80.6%)	1594 (77.8%)	1560 (76.2%)	1606 (78.4%)
	Group 1 vs. 4	1614 (78.8%)	1563 (76.3%)	1612 (78.7%)	1643 (80.2%)
	Group 2 vs. 3	1567 (76.5%)	1503 (73.4%)	1490 (72.8%)	1571 (76.7%)
	Group 2 vs. 4	1519 (74.2%)	1455 (71.0%)	1558 (76.1%)	1603 (78.3%)

Table 4.1: Number and percentage of features with p-value < 0.05 among 2048 extracted features between each pair of the four groups from 10X Genomics tissues categorized by invasive cancer proportion.

In summary, ResNet50 demonstrated a strong performance in extracting similar features from patches with no invasive cancer, some lymphocytes, others, and stroma (in ascending proportions), as well as similar features from patches with minimal stroma, some lymphocytes, invasive cancer, and others (in ascending proportions). Additionally, ResNet50 extracted different features from patches with and without invasive cancer.

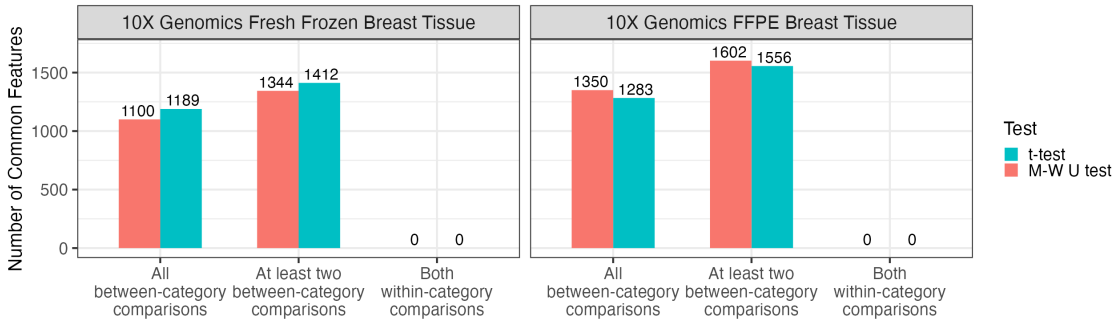
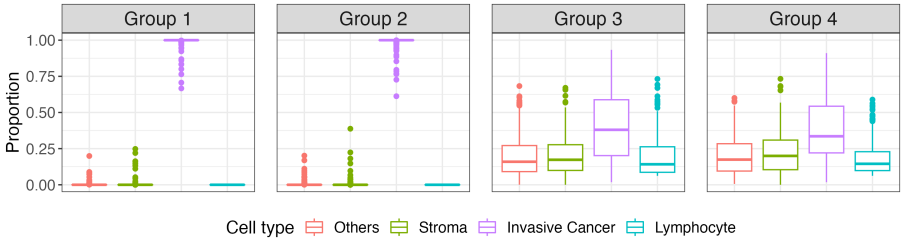


Figure 4.3: Number of significant features (p-value < 0.05) that were commonly identified by different pairwise hypothesis testing between four groups from 10X Genomics tissues categorized by invasive cancer proportion.

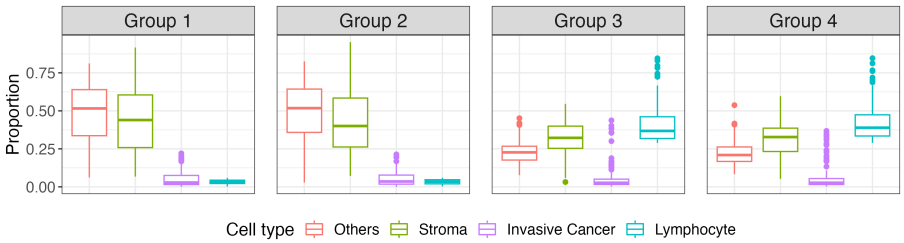
We further explored whether certain features with p-values below 0.05 were consistently identified in pairwise comparisons between groups (Table B.1). We categorized the common features into three groups. The first category includes features that were commonly identified in all between-category hypothesis tests, namely Group 1 vs. 3, Group 1 vs. 4, Group 2 vs. 3, and Group 2 vs. 4. The second category comprises features that were commonly identified in at least two between-category hypothesis tests. The third category consists of features that were commonly identified in both within-category hypothesis tests, namely Group

1 vs. 2 and Group 3 vs. 4. It is important to note that all features in the first category are also present in the second category, and the second and third categories are mutually exclusive. The pre-trained ResNet50 successfully extracted more than 1000 features from both breast tissues that were significantly different in patches differing in the proportion of invasive cancer (Figure 4.3).

Difference in Lymphocyte Proportions



(a) 10X Genomics fresh frozen breast tissue



(b) 10X Genomics FFPE breast tissue

Figure 4.4: Cell type composition in four groups of patches from 10X Genomics tissues, categorized by lymphocyte proportion, assessing ResNet50 as feature extractors.

We selected the standard patches into four groups based on the lymphocyte proportion for each of the two 10X Genomics tissues. Specifically, for the fresh frozen breast tissue section, Groups 1 and 2 exclusively contained patches without lymphocytes, while Groups 3 and 4 included patches with varying degrees of lymphocyte presence. Each group consisted of 250 patches. For the FFPE breast tissue, Groups 1 and 2 comprised patches with very low proportions of lymphocytes, whereas Groups 3 and 4 consisted of patches with higher lymphocyte proportions. Each group consisted of 200 patches. Notably, in the fresh frozen tissue, the patches in Groups 1 and 2 were essentially pure invasive cancer patches, which were a subset of Groups 3 and 4 from the previous section (Figure 4.4). Also, the majority of patches in Groups 3 and 4 had invasive cancer as the dominant cell type. With regards to the FFPE tissue sample, all four groups

predominantly exhibited very low proportions of invasive cancer, with only a few exceptions.

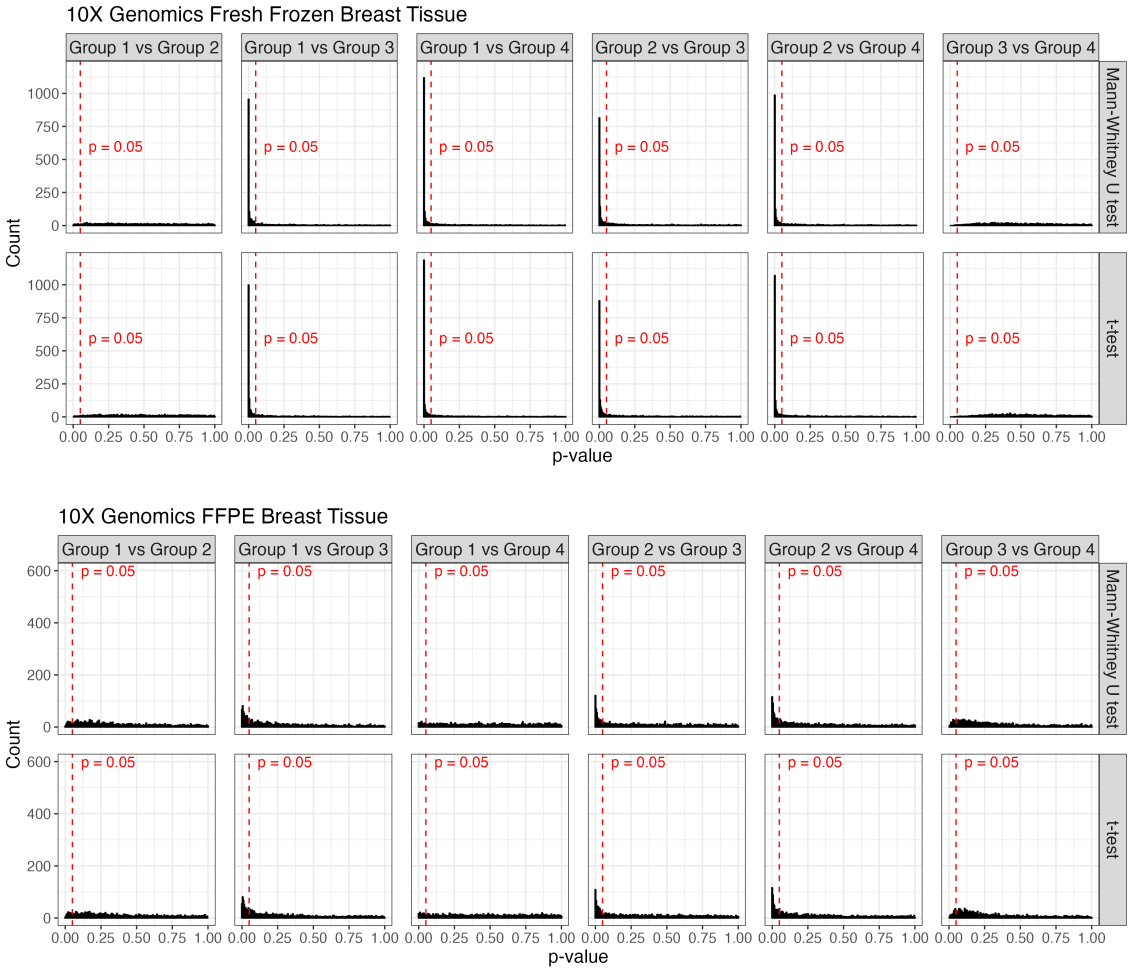


Figure 4.5: Histogram of p-values obtained from pairwise hypothesis testings between four groups of patches from 10X Genomics tissues categorized by lymphocyte proportion.

		10X Fresh Frozen Tissue		10X FFPE Tissue	
		t-test	M-W U test	t-test	M-W U test
Similar lymphocyte prop.	Group 1 vs. 2	49 (2.4%)	87 (4.2%)	155 (7.6%)	149 (7.3%)
	Group 3 vs. 4	9 (0.4%)	18 (0.9%)	181 (8.8%)	155 (7.6%)
Different lymphocyte prop.	Group 1 vs. 3	1375 (67.1%)	1333 (65.1%)	462 (22.6%)	497 (24.3%)
	Group 1 vs. 4	1498 (73.1%)	1443 (70.5%)	164 (8%)	157 (7.7%)
	Group 2 vs. 3	1289 (62.9%)	1216 (59.4%)	433 (21.1%)	440 (21.5%)
	Group 2 vs. 4	1417 (69.2%)	1360 (66.4%)	501 (24.5%)	495 (24.2%)

Table 4.2: Number and percentage of features with p-value < 0.05 among 2048 extracted features between each pair of the four groups from 10X Genomics tissues categorized by lymphocyte proportion.

Table 4.2 and Figure 4.5 provide insights into the significant differences observed between groups for the extracted features. The number of features with a p-value < 0.05 between each pair of groups varied substantially between the two tissue samples. In the case of fresh frozen breast tissue, more than 50% of the features exhibited p-values below 0.05 when comparing patches from different proportion categories. Conversely, over 95% of the features showed no significant differences when comparing patches from Groups 1 and 2. These percentages were similar to the observations made when comparing patches from Groups 3 and 4 in the previous section (Table 4.1), as both sets of groups primarily consisted of pure invasive cancer patches. This indicates that ResNet50 successfully extracted morphological features specific to invasive cancer cells. When comparing Group 3 and Group 4, less than 1% of the features exhibited p-values below 0.05. Since both groups had similar proportions of all four cell types and invasive cancer generally had higher proportions than other cell types (Figure 4.4), this lack of significant differences does not provide evidence regarding whether the extracted features were specific to the morphological characteristics of invasive cancer or lymphocytes.

For the FFPE breast tissue, there were more features with a p-value < 0.05 identified by between-category comparisons than by within-category comparisons, though the contrast is less dramatic as compared to the fresh frozen tissue. The percentages of features with p-values below 0.05 were 7.3% - 8.8% and 7.7% - 24.5% for within-category comparisons and between-category comparisons, respectively. This suggests that the pre-trained ResNet50 did not perform well in extracting distinct morphological characteristics to differentiate between patches in different proportion categories for the FFPE tissue. The disparity in feature extraction by ResNet50 is noteworthy when compared to the findings in the previous section, where the FFPE patches were grouped based on invasive cancer proportions. Given that all patches from which the features were extracted in this section had minimal invasive cancer cells (Figure 4.4), which is different from the patches in the previous section (Figure 4.1), this implies that the pre-trained ResNet50 had a stronger ability to extract distinctive morphological features when the patches had some invasive cancer cells.

We proceeded to investigate whether certain features, characterized by p-values below 0.05, were consistently identified in pairwise comparisons between groups (Table B.2, Figure 4.6). In the case of the fresh frozen breast tissue, the pre-trained ResNet50 effectively extracted over 1000 features that distinguished patches without lymphocytes (with pure invasive cancer) from patches with lymphocytes (with impure in-

vasive cancer). However, for the FFPE tissues, the pre-trained ResNet50 only extracted approximately 50 key features that differentiated patches with minimal lymphocytes from patches with higher lymphocyte proportions.

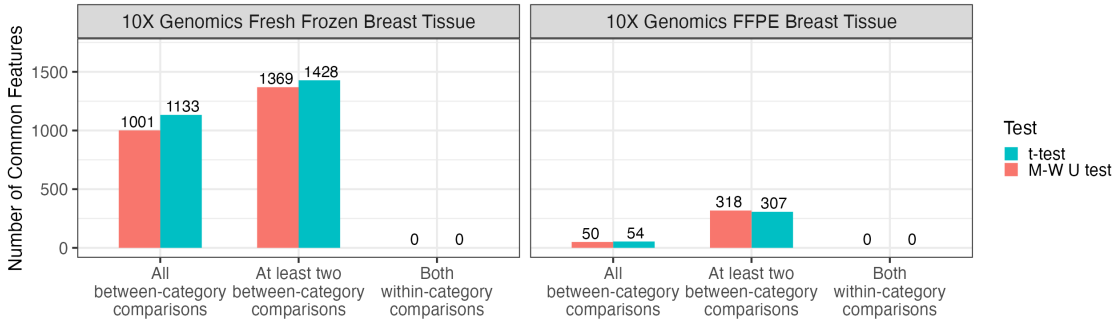


Figure 4.6: Number of significant features (p-value < 0.05) that were commonly identified by different pairwise hypothesis testing between four groups from 10X Genomics tissues categorized by lymphocyte proportion.

Difference in Stroma Proportions

We selected four groups of standard patches based on the stroma proportion. The prevalence of invasive cancer in the majority of the patches obtained from 10X Genomics fresh frozen breast tissue posed a challenge to create four distinct groups of patches with substantial variations in stroma proportion while maintaining an adequate sample size per group. Consequently, our analysis regarding the difference in stroma proportions focused solely on the features derived from patches sourced from 10X Genomics FFPE breast tissue.

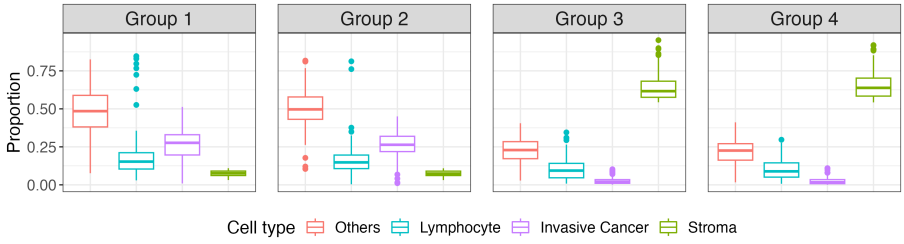


Figure 4.7: Cell type composition in four groups of patches from 10X Genomics FFPE tissue, categorized by stroma proportion, assessing ResNet50 as feature extractors.

Specifically, for the FFPE breast tissue section, each group had 200 patches, with Groups 1 and 2 primarily consisting of patches with relatively lower stroma proportions and Groups 3 and 4 consisting of patches with relatively higher stroma proportions. Notably, Groups 1 and 2 predominantly exhibited

higher proportions of others (i.e., normal epithelial and myeloid) with only a few exceptions of lymphocytes occurring with very high proportions (Figure 4.7). Invasive cancer was the second most prevalent cell type in many patches within Groups 1 and 2. Conversely, Groups 3 and 4 displayed relatively abundant others but had considerably lower proportions of invasive cancer.

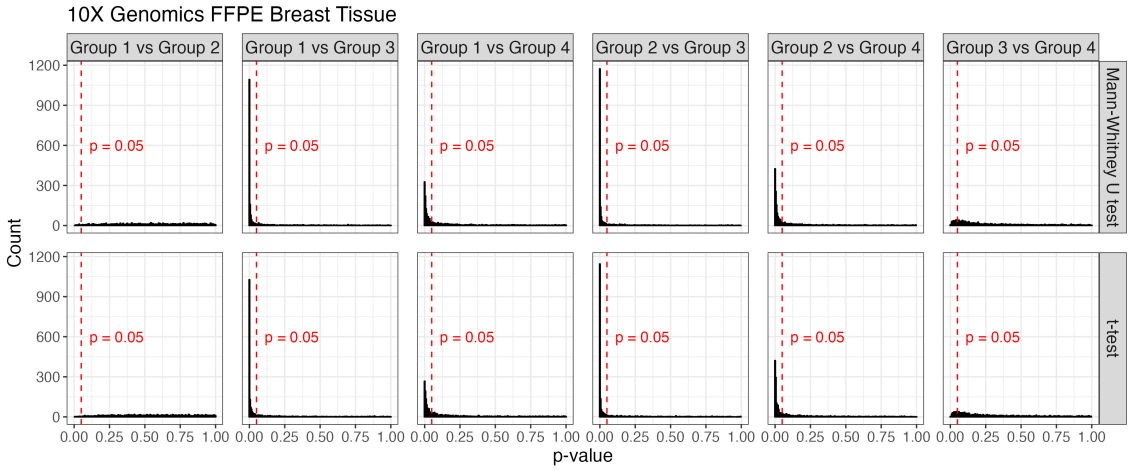


Figure 4.8: Histogram of p-values obtained from pairwise hypothesis testings between four groups of patches from 10X Genomics FFPE tissue categorized by stroma proportion.

		10X FFPE Tissue	
		t-test	M-W U test
Similar stroma prop.	Group 1 vs. 2	29 (1.4%)	44 (2.1%)
	Group 3 vs. 4	275 (13.4%)	308 (15%)
Different stroma prop.	Group 1 vs. 3	1439 (70.3%)	1499 (73.2%)
	Group 1 vs. 4	980 (47.9%)	1052 (51.4%)
	Group 2 vs. 3	1527 (74.6%)	1546 (75.5%)
	Group 2 vs. 4	1179 (57.6%)	1206 (58.9%)

Table 4.3: Number and percentage of features with p-value < 0.05 among 2048 extracted features between each pair of the four groups from 10X Genomics FFPE tissue categorized by stroma proportion.

Table 4.3 and Figure 4.8 provide insights into the significant differences observed between groups for these features. In the comparison between Group 1 and Group 2, less than 5% of the features exhibited p-values below 0.05. However, when comparing patches from Group 3 and Group 4, more than 10% of the features showed p-values below 0.05. This suggests that ResNet50 performed better in extracting features that characterized patches in Groups 1 and 2 compared to those in Groups 3 and 4. The discrepancy in the number of significant features identified in these two within-category comparisons could be attributed to the

absence of invasive cancer proportions in Groups 3 and 4, while Groups 1 and 2 contained a mixture of cell types other than stroma. In other words, ResNet50 demonstrated superior performance in capturing features associated with invasive cancer rather than stroma, given that patches in Groups 1 and 2 exhibited higher proportions of invasive cancer. When comparing patches from different proportion categories, over 50% of the features displayed p-values below 0.05. Interestingly, the between-category comparisons that included Group 4 revealed a smaller number of identified significant features compared to the comparisons involving Group 3.

We conducted an investigation to determine if specific features, identified with p-values below 0.05, consistently emerged in pairwise comparisons between groups. The findings are presented in Table B.3 and Figure 4.9. Notably, the pre-trained ResNet50 model successfully extracted more than 700 features that effectively distinguished patches exhibiting varying proportions of stroma. However, it is crucial to consider that the observed differences in these features were likely primarily driven by variations in the proportion of invasive cancer rather than stroma.

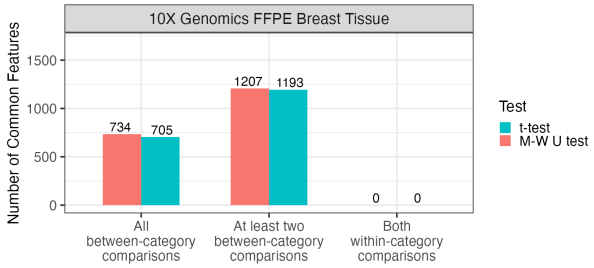


Figure 4.9: Number of significant features (p-value < 0.05) that were commonly identified by different pairwise hypothesis testing between four groups from 10X Genomics FFPE tissue categorized by stroma proportion.

Difference in Others Proportions

We selected four groups of standard patches based on the proportion of others (i.e., normal epithelial and myeloid). Since most patches from 10X Genomics fresh frozen breast tissue had low proportions of others, our analysis regarding the different proportions of others focused solely on the features derived from patches sourced from 10X Genomics FFPE breast tissue. For the FFPE breast tissue, each of the four groups consists of 200 patches. Groups 1 and 2 primarily comprised patches with relatively lower proportions of other cell types, while Groups 3 and 4 consisted of patches with relatively higher proportions of others.

Remarkably, Groups 1 and 2 exhibited predominantly negligible proportions of invasive cancer, with only a few exceptions (Figure 4.10). Within these groups, stroma emerged as the second most prevalent cell type across many patches. In contrast, Groups 3 and 4 displayed slightly higher proportions of invasive cancer compared to Groups 1 and 2, but lower proportions of stroma and lymphocytes. Invasive cancer was found to be the least abundant cell type in many patches within Groups 3 and 4.

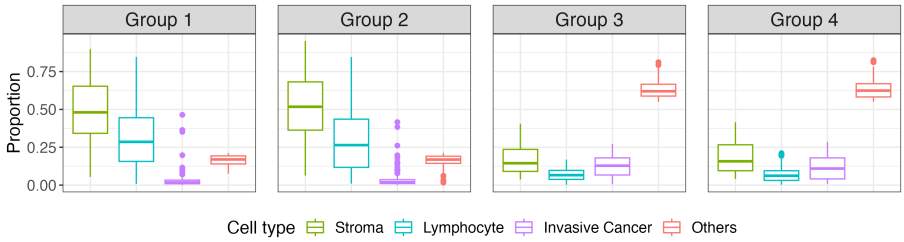


Figure 4.10: Cell type composition in four groups of patches from 10X Genomics FFPE tissue, categorized by others proportion, assessing ResNet50 as feature extractors.

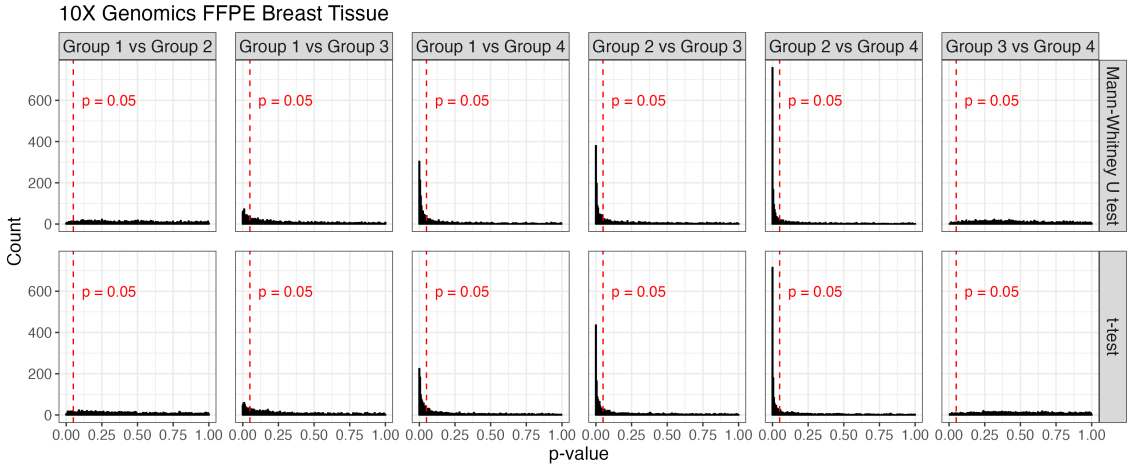


Figure 4.11: Histogram of p-values obtained from pairwise hypothesis testings between four groups of patches from 10X Genomics FFPE tissue categorized by others proportion.

Figure 4.11 and Table 4.4 provide the number of significant differences observed between groups for these features. When comparing patches from the same proportion category, approximately 5% of the features exhibited p-values below 0.05, indicating that there were no significant differences in the features extracted by ResNet50 from patches with similar cell type compositions. On the other hand, in the comparisons between patches from different proportion categories, over 20% of the features displayed p-values

below 0.05. Notably, when Group 1 was included in the between-category comparisons, a smaller number of significant features were identified compared to comparisons involving Group 2. Similarly, the between-category comparisons involving Group 3 revealed a smaller number of significant features compared to comparisons involving Group 4. These findings suggest that the features extracted by ResNet50 from patches with different cell type compositions were significantly different, albeit to a limited extent.

		10X FFPE Tissue	
		t-test	M-W U test
Similar others prop.	Group 1 vs. 2	144 (7%)	87 (4.2%)
	Group 3 vs. 4	67 (3.3%)	59 (2.9%)
Different others prop.	Group 1 vs. 3	435 (21.2%)	487 (23.8%)
	Group 1 vs. 4	886 (43.3%)	1032 (50.4%)
	Group 2 vs. 3	1031 (50.3%)	998 (48.7%)
	Group 2 vs. 4	1254 (61.2%)	1301 (63.5%)

Table 4.4: Number and percentage of features with p-value < 0.05 among 2048 extracted features between each pair of the four groups from 10X Genomics FFPE tissue categorized by others proportion.

Remarkably, the pre-trained ResNet50 model successfully extracted over 300 features that effectively distinguished patches with varying proportions of normal epithelial and myeloid cells (Table B.4, Figure 4.12). However, a larger number of significant features were identified in at least two between-category comparisons. This suggests that the features extracted by ResNet50 characterize different distinguishing patterns for different between-category comparisons. In other words, the patches analyzed in this study lacked consistent signals that could be consistently extracted by ResNet50 to differentiate the patches.

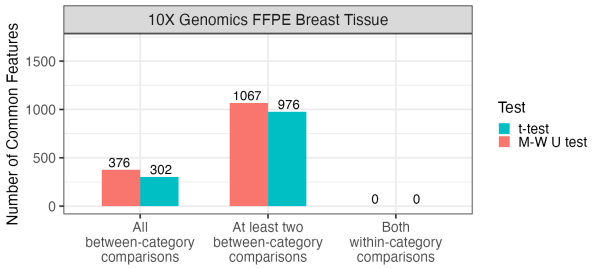


Figure 4.12: Number of significant features (p-value < 0.05) that were commonly identified by different pairwise hypothesis testing between four groups from 10X Genomics FFPE tissue categorized by others proportion.

4.1.2 Breast Cancer Data from Wu et al. [2021]

In this section, we present an analysis of six samples obtained from Wu et al. [2021]. We selected four groups of standard patches within each sample based on the proportions of invasive cancer, lymphocyte, stroma, and others. Notably, Groups 1 and 2 consistently exhibited relatively lower proportions of specific cell types, while Groups 3 and 4 consistently displayed relatively higher proportions of specific cell types.

To provide an overview of the groups and the results of hypothesis testing, we present figures and tables in a concise manner, focusing on the summary of findings without detailed interpretation and explanation. Also, only results obtained by Mann-Whitney U tests are included in this section. This allows for a clear understanding of the distribution of patches across the different groups and the significance of the observed differences without overwhelming the reader with excessive details.

Cell Type Compositions in Each Group

We selected and categorized the standard patches into four groups according to the proportion of each cell type present in each sample. The cell type composition of patches within each of the four groups categorized by the proportions of invasive cancer, lymphocytes, stroma, and others for each sample are illustrated in Figure 4.13, Figure 4.14, Figure 4.15, and Figure 4.16, respectively.

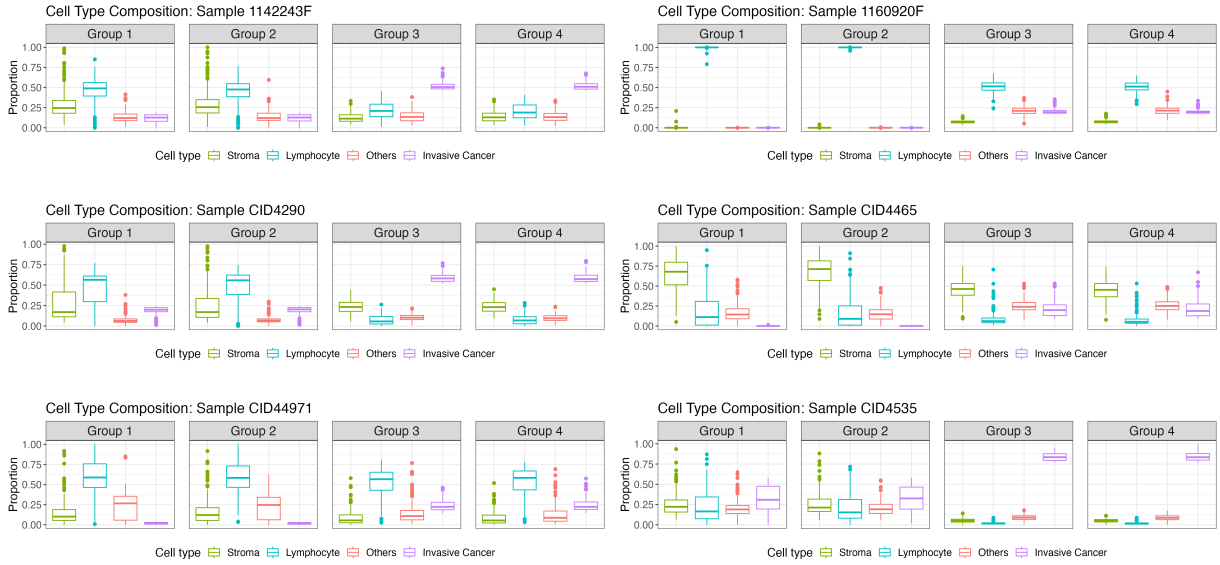


Figure 4.13: Cell type composition in four groups of patches from each tissue sample from Wu et al. [2021], categorized by invasive cancer proportion, assessing ResNet50 as feature extractors.

When we created four groups of patches based on the proportion of invasive cancer (Figure 4.13), we found that the difference in invasive cancer proportion between different proportion categories (i.e., Groups 1 and 2 versus Groups 3 and 4) was not large for all samples. For samples 1142243F, 1160920F, CID4465, and CID44971, most patches in the high-proportion groups had invasive cancer proportions around or lower than 0.5, while those in the low-proportion groups had negligible invasive cancer cells. Additionally, in sample 1160920F, the patches in the low-proportion groups were nearly pure of lymphocytes, whereas in sample CID4535, the patches in the high-proportion groups were almost entirely made up of invasive cancer cells.

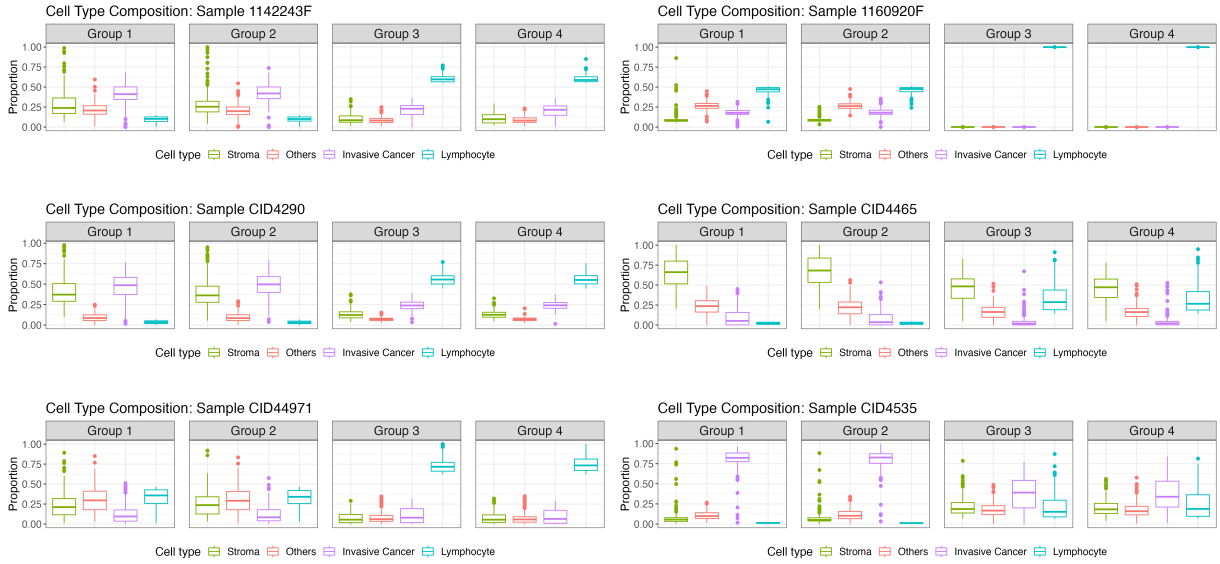


Figure 4.14: Cell type composition in four groups of patches from each tissue sample from Wu et al. [2021], categorized by lymphocyte proportion, assessing ResNet50 as feature extractors.

When we created four groups of patches based on the proportion of lymphocyte (Figure 4.14), the difference in lymphocyte proportion between different proportion categories was not large for most samples. For samples CID4290, CID4465, and CID4535, most patches in the low-proportion groups had negligible lymphocytes. Additionally, in sample 1160920F, the patches in the high-proportion groups were nearly pure of lymphocytes.

When we created four groups of patches based on the proportion of stroma (Figure 4.15), the difference in stroma proportion between different proportion categories was not large for all samples. For samples 1142243F, 1160920F, CID44971, and CID4535, most patches in the low-proportion groups had very few

stromata. Additionally, in sample 1160920F, the patches in the low-proportion groups were nearly pure of lymphocytes.

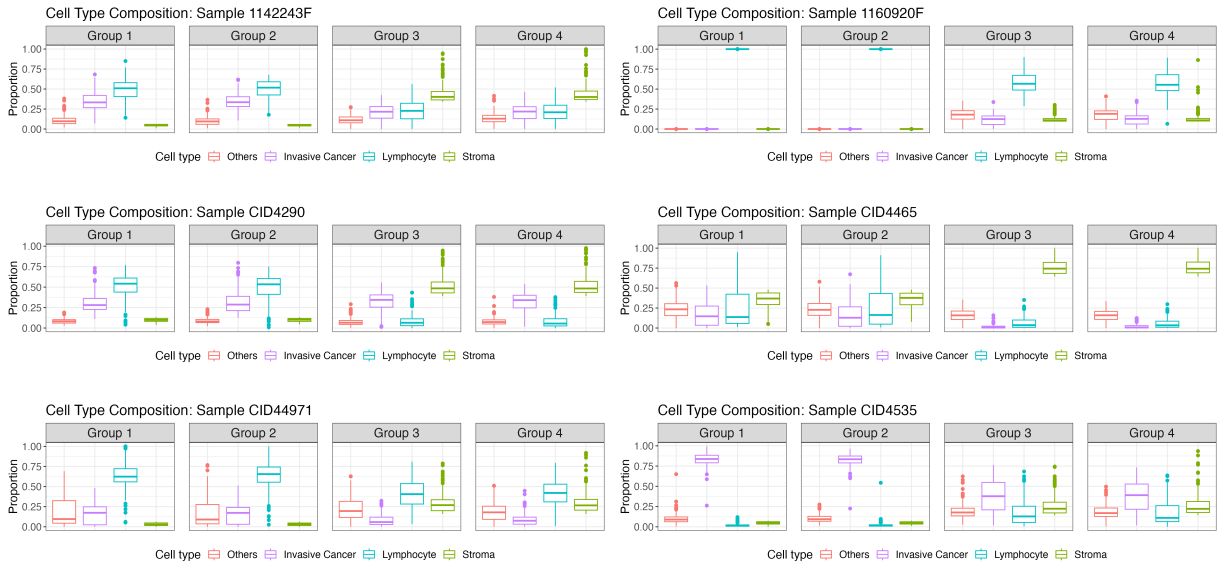


Figure 4.15: Cell type composition in four groups of patches from each tissue sample from Wu et al. [2021], categorized by stroma proportion, assessing ResNet50 as feature extractors.

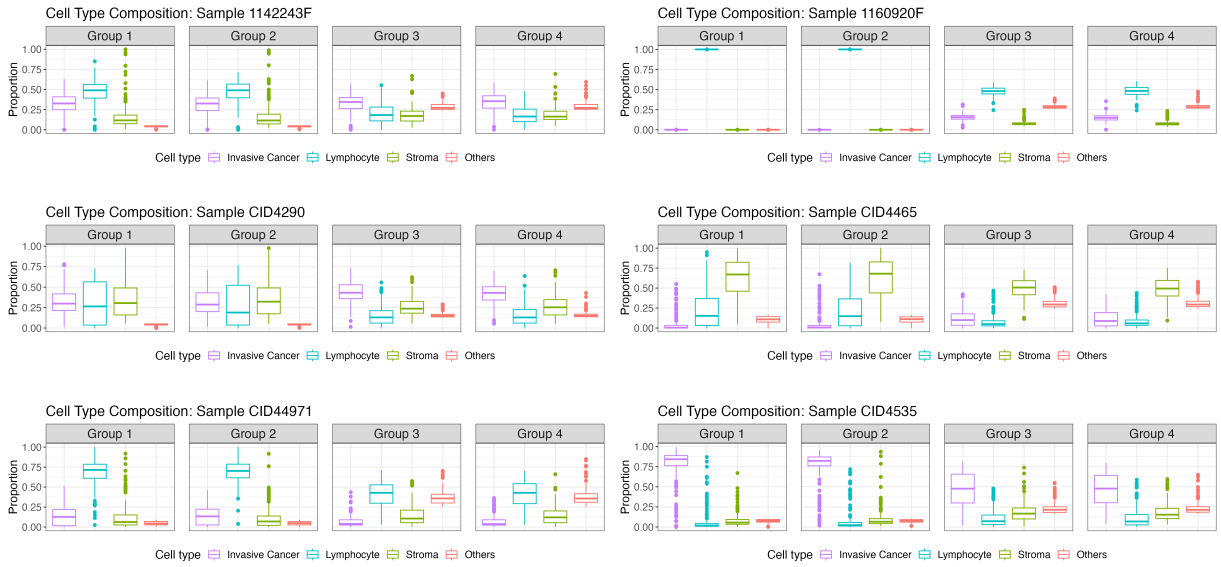


Figure 4.16: Cell type composition in four groups of patches from each tissue sample from Wu et al. [2021], categorized by others proportion, assessing ResNet50 as feature extractors.

When we created four groups of patches based on the proportion of others (Figure 4.16), no sample had

patches with very high proportions of others in the high-proportion groups. For all samples, most patches in the high-proportion groups had proportions of others lower than 0.5. Additionally, in sample 1160920F, the patches in the low-proportion groups were nearly pure of lymphocytes.

P-values Obtained by Pairwise Mann-Whitney U Tests

Table 4.5 provides the number of features with significant differences (rank-sum test p-value < 0.05) between each pair of the four groups categorized by invasive cancer, lymphocyte, stroma, and others, respectively.

		1142243F	1160920F	CID4290	CID4465	CID44971	CID4535
Similar tumor prop.	Group 1 vs. 2	16 (0.8%)	76 (3.7%)	23 (1.1%)	30 (1.5%)	43 (2.1%)	16 (0.8%)
	Group 3 vs. 4	15 (0.7%)	62 (3.0%)	20 (1.0%)	314 (15.3%)	43 (2.1%)	53 (2.6%)
Different tumor prop.	Group 1 vs. 3	1803 (88.0%)	1827 (89.2%)	1543 (75.3%)	830 (40.5%)	757 (37%)	1573 (76.8%)
	Group 1 vs. 4	1787 (87.3%)	1830 (89.4%)	1614 (78.8%)	1157 (56.5%)	712 (34.8%)	1612 (78.7%)
	Group 2 vs. 3	1804 (88.1%)	1812 (88.5%)	1595 (77.9%)	704 (34.4%)	1028 (50.2%)	1641 (80.1%)
	Group 2 vs. 4	1776 (86.7%)	1817 (88.7%)	1632 (79.7%)	997 (48.7%)	962 (47%)	1679 (82%)

(a) Categorized by the invasive cancer proportion.

		1142243F	1160920F	CID4290	CID4465	CID44971	CID4535
Similar lymphocyte prop.	Group 1 vs. 2	114 (5.6%)	95 (4.6%)	15 (0.7%)	17 (0.8%)	101 (4.9%)	79 (3.9%)
	Group 3 vs. 4	56 (2.7%)	64 (3.1%)	43 (2.1%)	216 (10.5%)	160 (7.8%)	71 (3.5%)
Different lymphocyte prop.	Group 1 vs. 3	1494 (72.9%)	1736 (84.8%)	1806 (88.2%)	1340 (65.4%)	737 (36%)	1611 (78.7%)
	Group 1 vs. 4	1513 (73.9%)	1771 (86.5%)	1769 (86.4%)	985 (48.1%)	768 (37.5%)	1657 (80.9%)
	Group 2 vs. 3	1414 (69%)	1766 (86.2%)	1799 (87.8%)	1245 (60.8%)	675 (33%)	1524 (74.4%)
	Group 2 vs. 4	1386 (67.7%)	1776 (86.7%)	1759 (85.9%)	792 (38.7%)	920 (44.9%)	1583 (77.3%)

(b) Categorized by the lymphocyte proportion.

		1142243F	1160920F	CID4290	CID4465	CID44971	CID4535
Similar stroma prop.	Group 1 vs. 2	451 (22.0%)	3 (0.1%)	46 (2.2%)	68 (3.3%)	73 (3.6%)	12 (0.6%)
	Group 3 vs. 4	22 (1.1%)	78 (3.8%)	64 (3.1%)	61 (3.0%)	48 (2.3%)	18 (0.9%)
Different stroma prop.	Group 1 vs. 3	1500 (73.2%)	1700 (83.0%)	1750 (85.4%)	1107 (54.1%)	1524 (74.4%)	1464 (71.5%)
	Group 1 vs. 4	1505 (73.5%)	1719 (83.9%)	1644 (80.3%)	1025 (50.0%)	1597 (78.0%)	1531 (74.8%)
	Group 2 vs. 3	1412 (68.9%)	1699 (83.0%)	1727 (84.3%)	1073 (52.4%)	1285 (62.7%)	1496 (73.0%)
	Group 2 vs. 4	1415 (69.1%)	1709 (83.4%)	1590 (77.6%)	1042 (50.9%)	1424 (69.5%)	1561 (76.2%)

(c) Categorized by the stroma proportion.

		1142243F	1160920F	CID4290	CID4465	CID44971	CID4535
Similar others prop.	Group 1 vs. 2	73 (3.6%)	82 (4.0%)	56 (2.7%)	8 (0.4%)	6 (0.3%)	22 (1.1%)
	Group 3 vs. 4	51 (2.5%)	28 (1.4%)	159 (7.8%)	36 (1.8%)	49 (2.4%)	64 (3.1%)
Different others prop.	Group 1 vs. 3	665 (32.5%)	1771 (86.5%)	1078 (52.6%)	58 (2.8%)	953 (46.5%)	1334 (65.1%)
	Group 1 vs. 4	525 (25.6%)	1757 (85.8%)	1470 (71.8%)	368 (18.0%)	886 (43.3%)	1479 (72.2%)
	Group 2 vs. 3	668 (32.6%)	1831 (89.4%)	525 (25.6%)	39 (1.9%)	919 (44.9%)	1456 (71.1%)
	Group 2 vs. 4	568 (27.7%)	1827 (89.2%)	1240 (60.5%)	474 (23.1%)	871 (42.5%)	1551 (75.7%)

(d) Categorized by the others proportion.

Table 4.5: Number and percentage of features with p-value < 0.05 identified by Mann-Whitney U test between each pair of the four groups from each Wu et al. sample categorized by the proportion of different cell types. The histograms of p-values are provided in Appendix B (Figure B.9, B.10, B.11, B.12).

Figure 4.17 provides a summary and comparison of the percentage of features with a p-value < 0.05. These features were identified through either between-category pairwise comparisons or within-category pairwise comparisons, stratified by either sample or cell type.

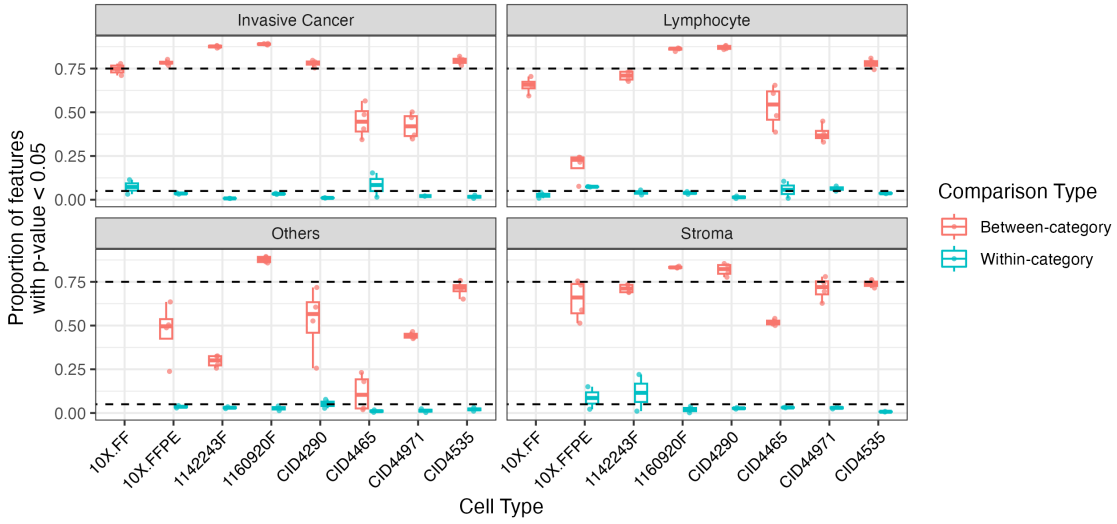


Figure 4.17: The proportion of rank-sum test p-values that are smaller than 0.05 when comparing the 2048 ResNet features between groups defined by cell type proportions and samples. The two horizontal lines indicate 0.05 and 0.75, respectively.

Common Significant Features

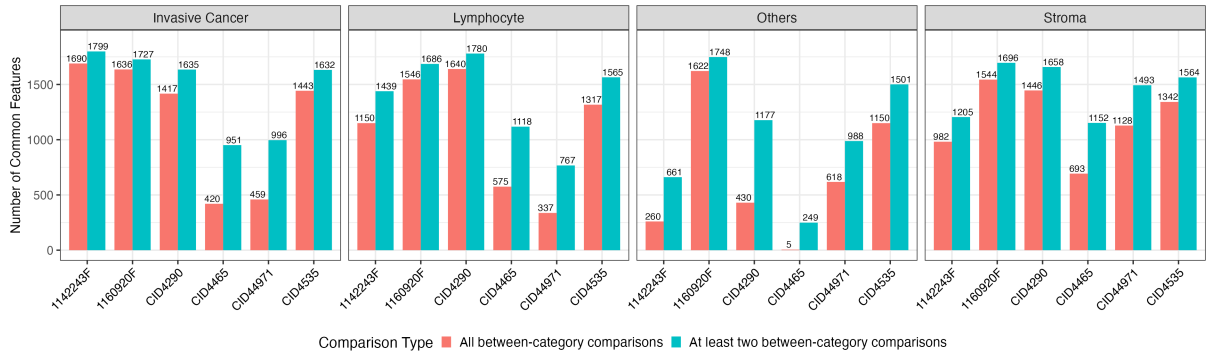


Figure 4.18: Number of significant features (p-value < 0.05) that were commonly identified by different pairwise hypothesis testing between four groups of patches categorized by proportions of invasive cancer, lymphocyte, stroma, or others, within each Wu et al. sample.

Table B.5, Table B.6, Table B.7, and Table B.8 in Appendix B depict the number of features that were consistently identified as having p-value below 0.05 in more than one pairwise comparison among four

groups categorized by the proportions of invasive cancer, lymphocyte, stroma, and others, respectively, within each sample. It is worth noting that no significant features were commonly identified through both within-category comparisons (Group 1 vs. 2 and Group 3 vs. 4) across all samples and groups generated based on similar proportions of cell types. Therefore, in the results presented, we only display the number of significant features that were commonly identified in more than one comparison between two groups from different proportion categories (Figure 4.18).

4.2 Prediction of Cell Type Proportions by Penalized Regressions

In order to further explore the ability of the pre-trained ResNet to differentiate patches with different cell type compositions, we conducted Lasso regressions using the extracted features as predictors and the proportions of the nine major cell types as the response variables. The feature extraction was performed using ResNet50, ResNet101, and ResNet152. For simplicity, we will only present the regression results for the two 10X Genomics breast tissues.

To begin, we divided the patches from each sample into training, validation, and testing datasets, following a split ratio of 70% for training, 15% for validation, and 15% for testing. The same dataset split was later used for training the neural network. The λ values for the Lasso regression were determined through cross-validation using the training dataset. The specific values of λ_{min} and λ_{1se} for each regression can be found in Table 4.6.

Sample	ResNet	λ_{min}	λ_{1se}
10X Genomics fresh frozen	ResNet50	0.0006310	0.0019953
10X Genomics FFPE	ResNet50	0.0012589	0.0050119
10X Genomics fresh frozen	ResNet101	0.0010000	0.0063096
10X Genomics FFPE	ResNet101	0.0015849	0.0050119
10X Genomics fresh frozen	ResNet152	0.0012589	0.0050119
10X Genomics FFPE	ResNet152	0.0015849	0.0039811

Table 4.6: The value of λ that gave the minimum cross-validated error and the largest value of λ such that error is within one standard error of the minimum.

We fitted each Lasso regression with the two chosen λ values, λ_{min} and λ_{1se}, using the training dataset, and subsequently analyzed the predicted values for the testing dataset. It is important to note that the predicted values can be negative, and the nine predicted values for each patch did not sum up to one. We

evaluated the relationship between the predicted values and the observed proportions of each cell type and assessed the prediction accuracy of Lasso regressions using the adjusted R-squared value, the mean squared error (MSE), and the slope of best-fit line obtained by the ordinal least squares (OLS) method (Figure 4.19).

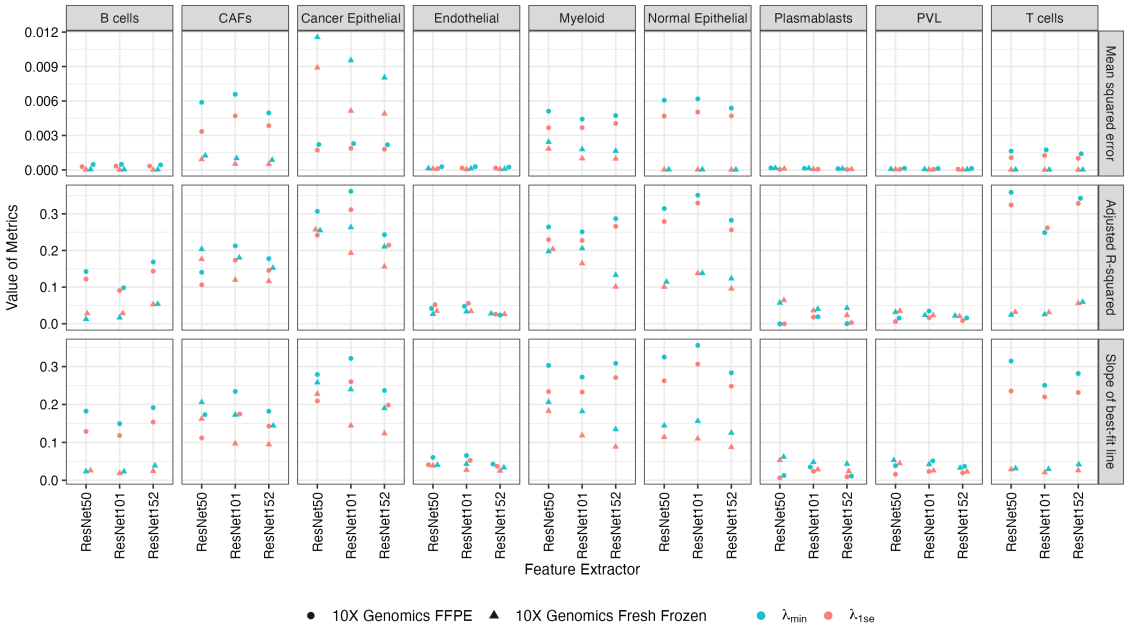


Figure 4.19: The mean squared error, adjusted R-squared, and slope of the best-fit line that represents the relationship between the observed proportion of cell types obtained through cell type deconvolution and the predicted values obtained from the Lasso regression.

Overall, the alignment between the predicted values and observed proportions was generally poor for all cell types. However, when considering the adjusted R-squared and slope, the cell types of cancer epithelial CAFs, myeloid, and normal epithelial in the fresh frozen tissue showed relatively better correlation compared to other cell types. In the FFPE tissue, the proportions of endothelial, plasmablasts, and PVL were poorly predicted by the Lasso regression. On the other hand, when considering the MSEs, the Lasso regressions using λ_{min} consistently had higher MSEs than the corresponding Lasso regressions using λ_{1se} across all cell types in both samples. It is worth noting that although certain cell types exhibited better alignment between the predicted and observed proportions, the mean squared errors varied across cell types. This discrepancy arises from the lack of comparability in the scales of different cell type proportions. Consequently, higher mean squared errors were observed for certain cell types, despite their predictions being relatively well-aligned with the corresponding observed proportions. Besides, there was no consistent trend

in the metrics observed across all cell types with respect to the feature extractors. However, there was some consistency in the values of λ. In the FFPE breast tissue, the adjusted R-squared values indicated that the predictions made by the Lasso regression using λ_{min} were generally better aligned with the observed proportions for most cell types, except for endothelial, plasmablasts, and PVL. The slope of the best-fit line did not show any exceptions in cell types. In the fresh frozen tissue, for the cell types that exhibited some alignment between the observed and predicted values, using λ_{min} generally resulted in better predictions.

As the Lasso regressions with λ_{min} exhibited superior predictive performance compared to those with $\lambda_{1\text{se}}$, we present scatterplots showing the relationship between observed proportions and predicted values obtained from the Lasso regression with λ_{min}, utilizing features extracted by ResNet50 (Figure 4.20). Similar scatterplots for Lasso regressions on features extracted by ResNet101 and ResNet152 are provided in Appendix B (Figure B.13). Notably, within each cell type, the scatterplots exhibit similar patterns regardless of the ResNet feature extractor used.

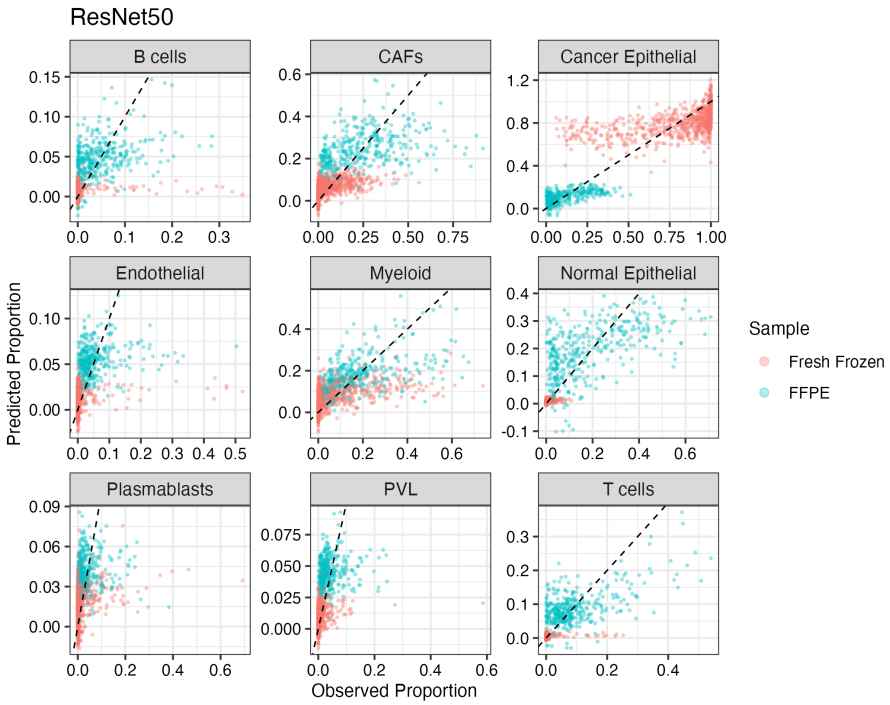


Figure 4.20: The scatterplot of the predicted values obtained from the Lasso regression with λ_{min} on features extracted by pre-trained ResNet50 plotted against the observed values derived from cell type deconvolution. The black dashed line represents the diagonal line.

To address the violation of compositional data restrictions, whereby the predicted values were not non-

negative and did not sum up to one, we implemented a normalization procedure. For each patch, we replaced all negative predicted values with zero and divided each positive predicted value by the sum of all positive predicted values. The metrics evaluating the Lasso regression are presented in Figure 4.21. The patterns observed in the metrics for the normalized predicted proportions are consistent with those for the original predicted values.

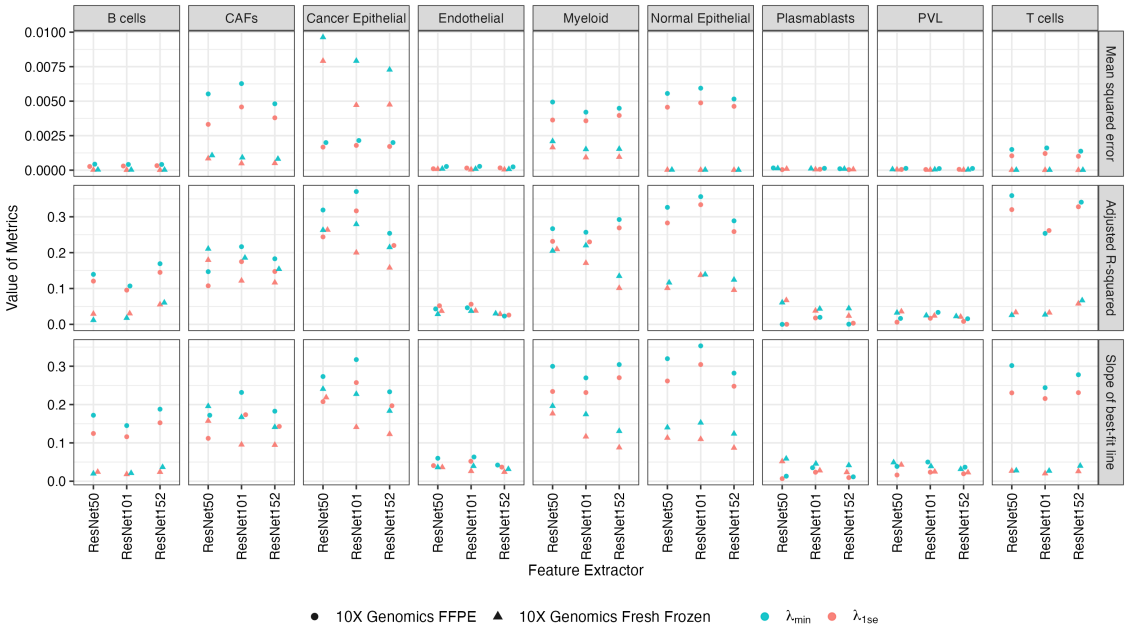


Figure 4.21: The mean squared error, adjusted R-squared, and slope of the best-fit line that represents the relationship between the observed proportion of cell types obtained through cell type deconvolution and the normalized predicted proportion obtained from the Lasso regression.

The scatterplots depicted in Figure 4.22 demonstrate that the predictions generated by the Lasso regression exhibited a narrower range compared to the observed proportions. Notably, while the mean values were very similar between the predicted and observed proportions, the median values varied to varying degrees (Figure 4.23). Similar scatterplots and histograms for Lasso regressions on features extracted by ResNet101 and ResNet152 are provided in Appendix B (Figure B.14, B.15). These findings indicate that the Lasso regression models effectively predicted the mean values of the testing data, but were less accurate in predicting values that deviated from the mean.

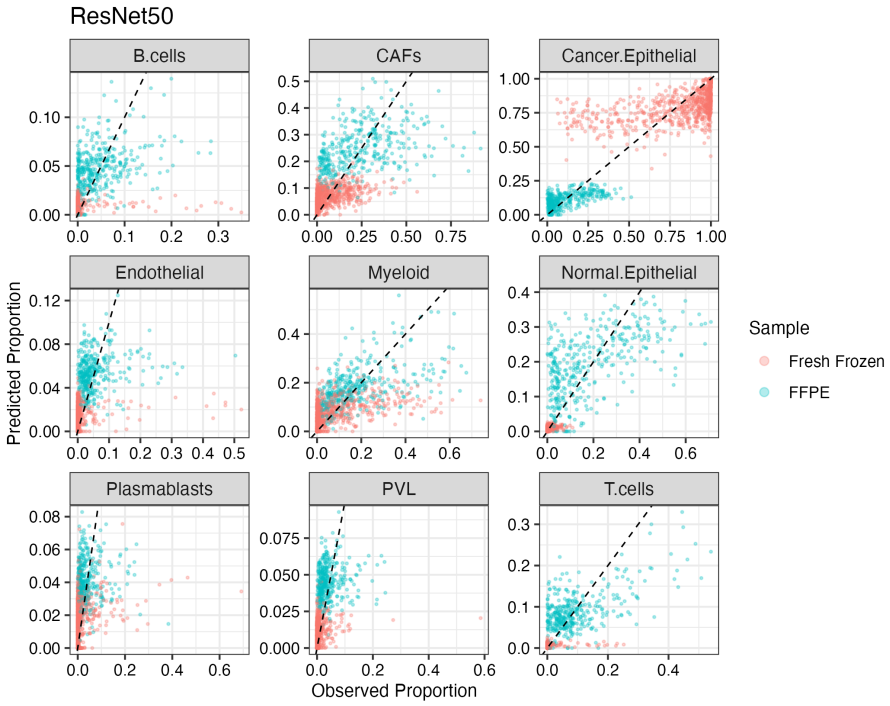


Figure 4.22: The scatterplot of the normalized predicted proportions obtained from the Lasso regression with λ_{min} on the features extracted by pre-trained ResNet50 plotted against the observed values derived from cell type deconvolution. The black dashed line represents the diagonal line.

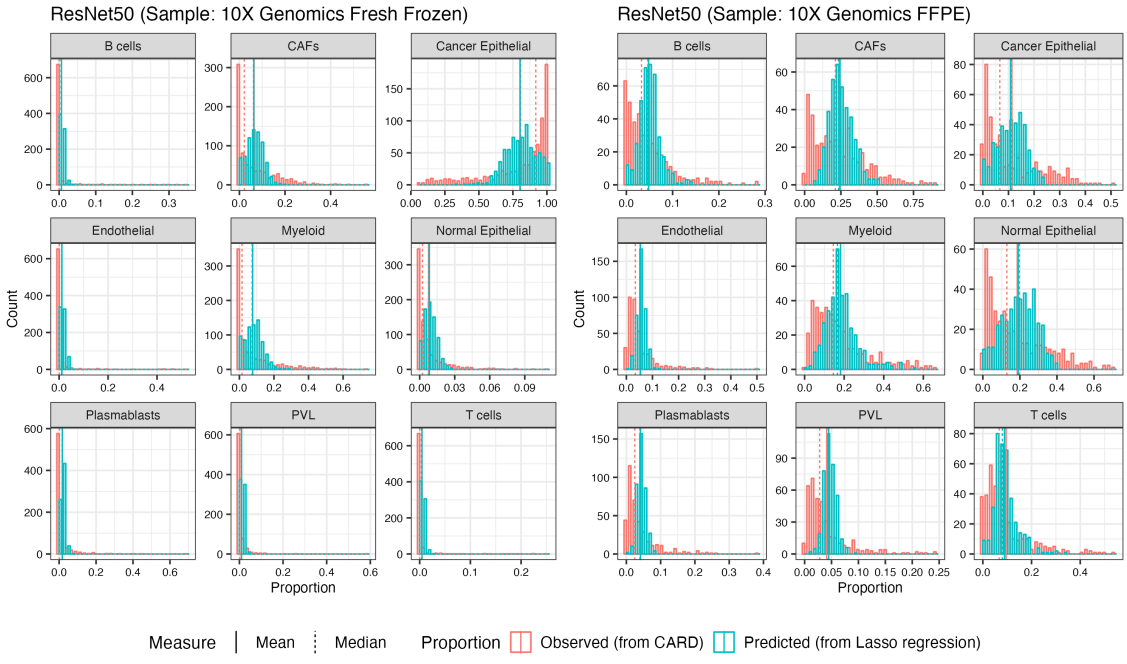


Figure 4.23: The histograms of the normalized predicted proportions obtained from the Lasso regression with λ_{min} on the features extracted by pre-trained ResNet50 and the observed values derived from cell type deconvolution. The solid and dashed lines represent the mean and median, respectively.

Chapter 5

Regression Results for 10X Genomics Samples

This chapter focuses on the analysis of ResNet50 transfer learning for the regression task, using samples from 10X Genomics. The input for the neural networks consisted of patches from the samples, while the output was the cell type proportions. Various configurations were explored, including different batch sizes for training, optimizers, learning rates, dropout layer proportions, and numbers of neurons in the hidden dense layer. Throughout the analysis, the training, validation, and testing datasets remained consistent across all networks with varying parameters for the same task and samples. The loss function employed was mean squared error (MSE), with mean absolute error (MAE) used as the metrics function. The networks were trained for a maximum of 1000 epochs, although some networks terminated early with a patience of 30 epochs. For further information on the dataset and network architecture, please refer to Chapter 2.

5.1 FFPE Human Breast Tissue

5.1.1 Predicting Nine Cell Type Proportions

This section aims to assess the performance of transfer learning using ResNet50 with various hyper-parameter combinations. Neural networks were designed to predict the proportions of nine cell types in each standard patch derived from the 10X Genomics FFPE human breast tissue. The evaluation involves examining differ-

ent optimizers (Adam, SGD), batch sizes for training (16, 32, 64, 128), learning rates (0.01, 0.001, 0.0001), dropout layer proportions (0.0, 0.2, 0.5), and the number of neurons of the hidden dense layer (256, 512). The datasets used for training, validation, and testing remain consistent across all network configurations. The overall dataset comprises a total of 2518 patches, with 1763 patches allocated for training, 378 for validation, and 377 for testing. Notably, the training, validation, and testing sets reveal diversity in the cellular composition (Figure 5.1). The proportions of all cell types in the training set are shown in Figure 5.2.

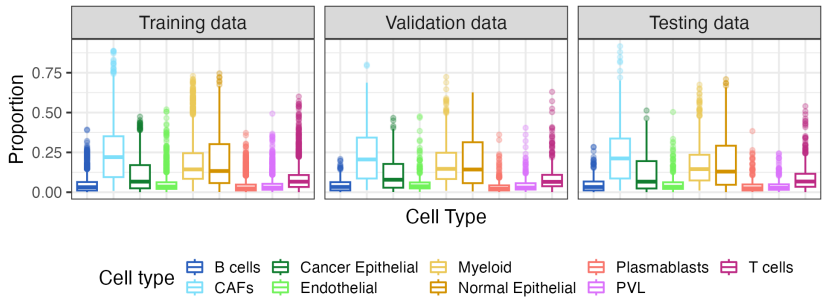


Figure 5.1: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of nine cell types using standard patches from 10X Genomics FFPE breast tissue.

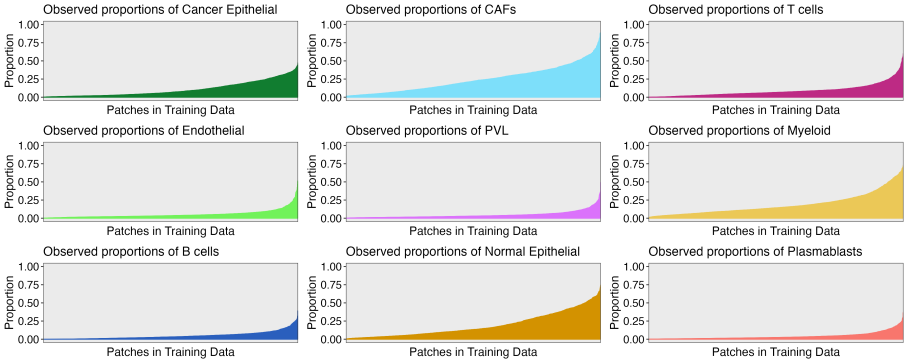


Figure 5.2: Observed proportions of nine cell types in standard patches from the training data of the 10X Genomics FFPE breast tissue.

Learning Curves

Learning curves illustrate loss and metrics over the course of the training epochs. By examining the learning curves in training and validation data for a set of hyper-parameters, we can tell whether the model is doing meaningful learning (i.e., whether the loss has decreased) and whether there is overfitting (i.e., the loss of training data decreases while the loss of validation data increases). Examples of six networks are shown in

Figure 5.3 as representatives of other networks.

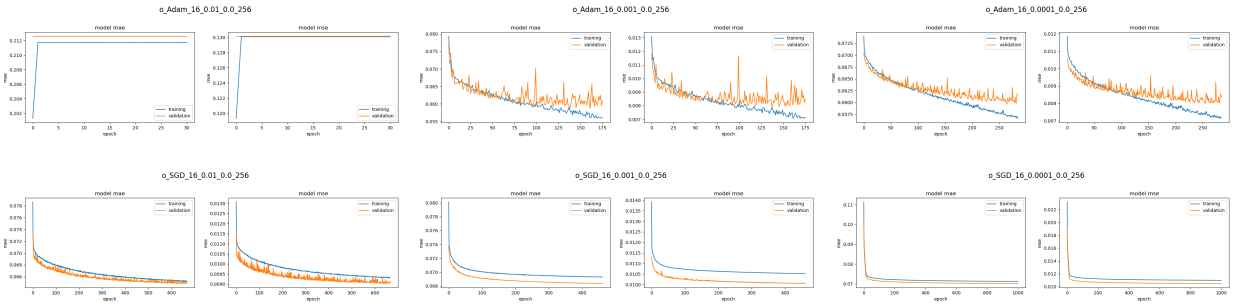


Figure 5.3: Learning curves of neural networks with a batch size of 16, a hidden dense layer of 256 neurons, and no dropout layer aiming to predict the proportions of nine cell types using standard patches from 10X Genomics FFPE breast tissue. Each figure title specifies the image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

The learning curves indicate that when the learning rate was set to 0.01, all networks employing the Adam optimizer failed to learn from the training data. This failure was evident by the training process ending abruptly after the initial 30 epochs. This suggests that a learning rate of 0.01 is not suitable for effective learning with the Adam optimizer, resulting in suboptimal performance and limited training progress. When the learning rate was set to 0.001 or 0.0001, all networks with the Adam optimizer successfully learned from the training data. While some networks consistently had smaller validation loss and metrics than the training loss and metrics, others initially had smaller validation loss and metrics but decreased slower than the training loss and metrics, resulting in larger values.

On the other hand, the SGD optimizer performed reasonably well with all learning rates. However, the reduction in training loss was relatively slow, and the validation loss and metrics consistently remained lower than the training loss and metrics. When the learning rate was set to 0.0001, the networks generally took more epochs to reach a stable validation loss, as compared to networks with learning rates of 0.01 and 0.001.

Evaluation Using Validation Data

This section focuses on analyzing and comparing the validation loss (MSE) and validation metrics (MAE) at the optimal epoch across different hyperparameter combinations. The consistency observed between the validation loss and metrics indicates their agreement in evaluating the model's performance (Figure 5.4).

The neural network that achieved the lowest validation loss had the following specifications: optimizer = Adam, batch size = 32, learning rate = 0.0001, dropout proportion = 0.5, and dense layer units = 512.

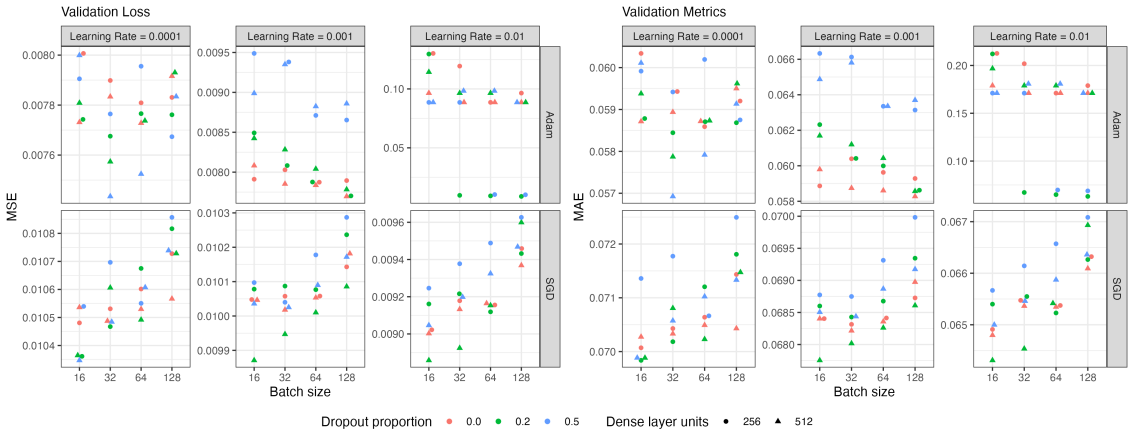


Figure 5.4: Validation loss and metrics from neural networks predicting proportions of nine cell types using standard patches from 10X Genomics FFPE breast tissue.

Figure 5.4 clearly demonstrates that among networks utilizing the Adam optimizer, the validation loss at the optimal epoch was consistently higher when the learning rate was set to 0.01 compared to cases where it was set to 0.001 or 0.0001. In contrast, for networks using the SGD optimizer, the validation loss at the optimal epoch was slightly lower when the learning rate was set to 0.01 compared to cases where it was set to 0.001 or 0.0001. It is important to note that some networks using the Adam optimizer with a learning rate of 0.01 failed to learn from the training data, as evident from the learning curves.

Regarding the learning rate, networks using the Adam optimizer consistently achieved lower validation loss at the optimal epoch compared to networks using SGD when the learning rate was set to 0.0001 and 0.001, after accounting for other parameters. However, when the learning rate was set to 0.01, networks utilizing the Adam optimizer consistently yielded significantly higher validation loss compared to those using the SGD optimizer.

Furthermore, for the same combinations of optimizer, learning rate, training batch size, and dense layer units, the validation loss at the optimal epoch did not exhibit a generalizable trend with increasing dropout proportion. When networks utilized the Adam optimizer with a learning rate of 0.0001, networks without a dropout layer tended to produce higher validation loss compared to those with a dropout layer. On the other hand, when the networks used the Adam optimizer with a learning rate of 0.001, the validation loss

increased with increasing dropout proportion. For the SGD optimizer, the dropout proportion did not have a significant effect on the validation loss, with only small variations observed after controlling for other parameters.

Predication Using Testing Data

To evaluate the relationship between the predicted and observed proportions of each cell type, we employed the OLS method to fit a line of best fit. The slope of this line represents the degree of alignment between the predicted and observed proportions. The slopes for each cell type were consistently worse when using the SGD optimizer as compared to the Adam optimizer, regardless of the combinations of dense layer and dropout layer (Figure C.1). Additionally, networks employing the Adam optimizer with a learning rate of 0.01 also exhibited poorer slopes (Figure C.1). As a result, we excluded these combinations of optimizer and learning rate from our analysis and focused on networks utilizing the Adam optimizer with learning rates of 0.001 and 0.0001 (Figure 5.5).

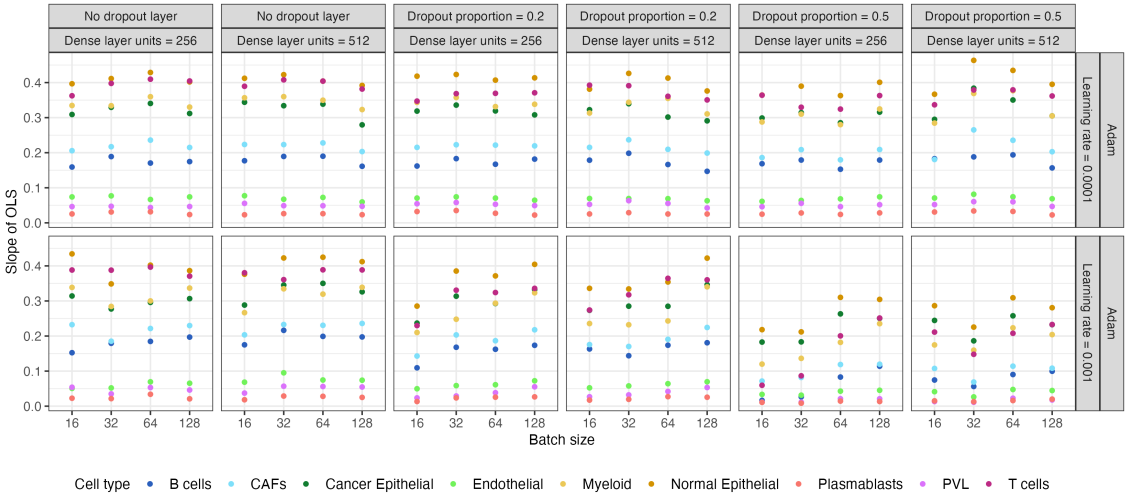


Figure 5.5: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using standard patches from 10X Genomics FFPE breast tissue, excluding networks with limited learning.

Our analysis reveals that normal epithelial cells, T cells, myeloid cells, and cancer epithelial cells consistently demonstrated higher slopes compared to the remaining cell types (Figure 5.5). Conversely, endothelial cells, PVLs, and plasmablasts consistently exhibited smaller slopes when compared to the other cell types.

Since no single network configuration was able to attain the highest slope for all cell types (Table C.1), our focus shifts to the combination of parameters that yielded the lowest validation loss (optimizer = Adam, batch size = 32, learning rate = 0.0001, dropout proportion = 0.5, and dense layer units = 512). Remarkably, this specific parameter combination coincided with achieving the highest slope of the best-fit line for many cell types.

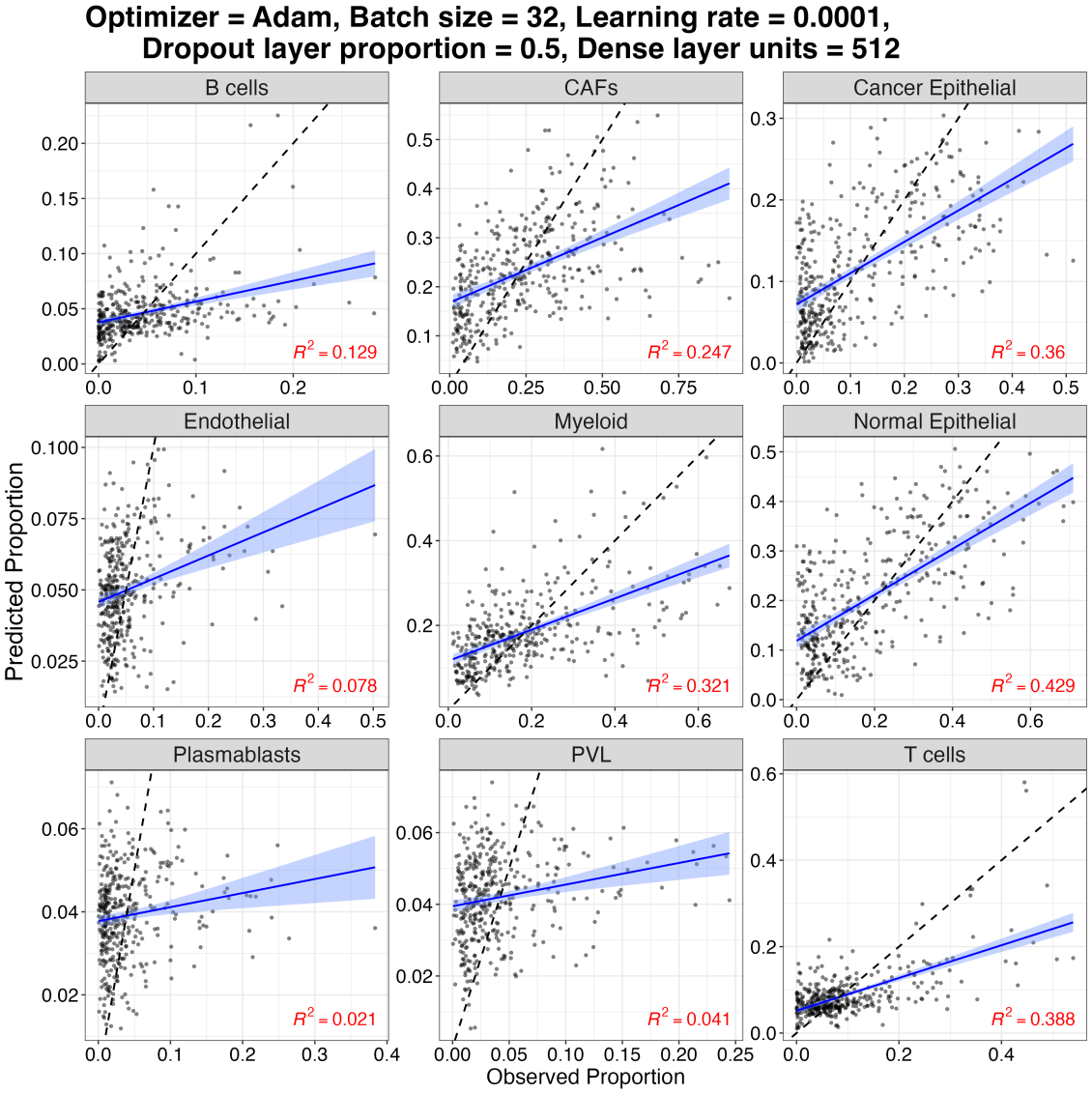


Figure 5.6: Scatterplot comparing predicted and deconvoluted proportions of nine cell types in standard patches from 10X Genomics FFPE breast tissue. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. The R-squared value is indicated in red.

Upon examining the scatterplots in Figure 5.6, we notice that when the true proportions of these cell

types are very small, the predicted proportions tend to vary in a wide range around the true proportions. However, as the true proportions increase, the range of predicted proportions becomes more concentrated around the mean value of the observed proportions.

Additionally, the bar plots in Figure 5.7 illustrate that cell types exhibiting higher abundance within the testing patches, such as T cells, cancer epithelial cells, and normal epithelial cells, demonstrate better correlations between the predicted and observed proportions. Conversely, for cell types that are relatively rare in the testing patches, such as plasmablasts and PVL, a different pattern emerges. This implies that cell types with higher abundance in the training (Figure 5.2) and testing data demonstrated superior predictive performance. Conversely, cell types with extremely small proportions in the sampled tissue were associated with diminished predictive performance.

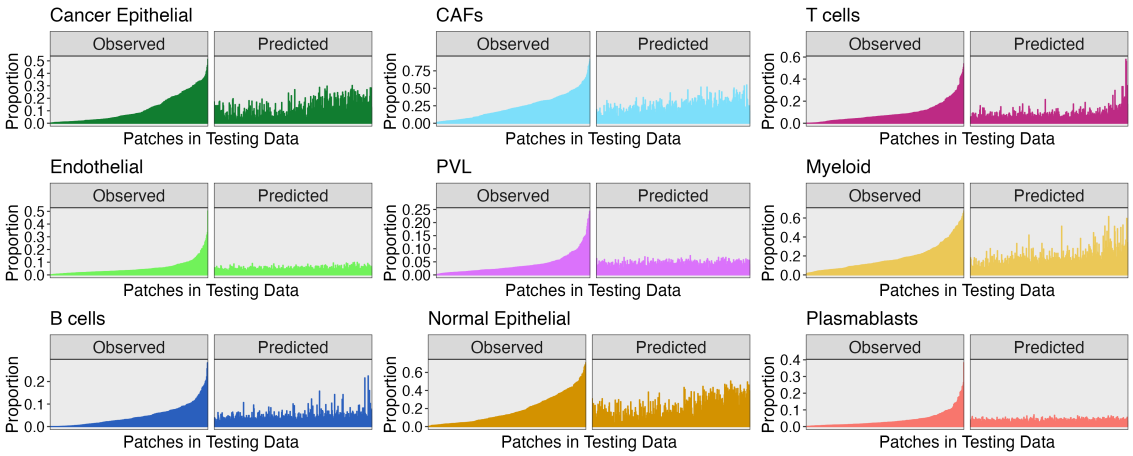


Figure 5.7: Comparison of predicted and observed nine cell type compositions in each of the standard patches from the 10X Genomics FFPE tissue in the testing data.

To explicitly examine the disparities between the distributions of proportions obtained from CARD and those predicted by the neural network, we constructed histograms (Figure 5.8). The predicted proportions tend to be more concentrated than the observed proportions. This observation suggests that the trained networks encountered difficulties in accurately predicting proportions associated with extreme values. Furthermore, our findings indicate that for the majority of cell types, the predicted proportions exhibit different median values compared to the observed proportions, while maintaining similar mean values. This implies that instead of precisely predicting the proportions for individual patches, the networks tended to estimate the mean proportion for each cell type within the testing set.

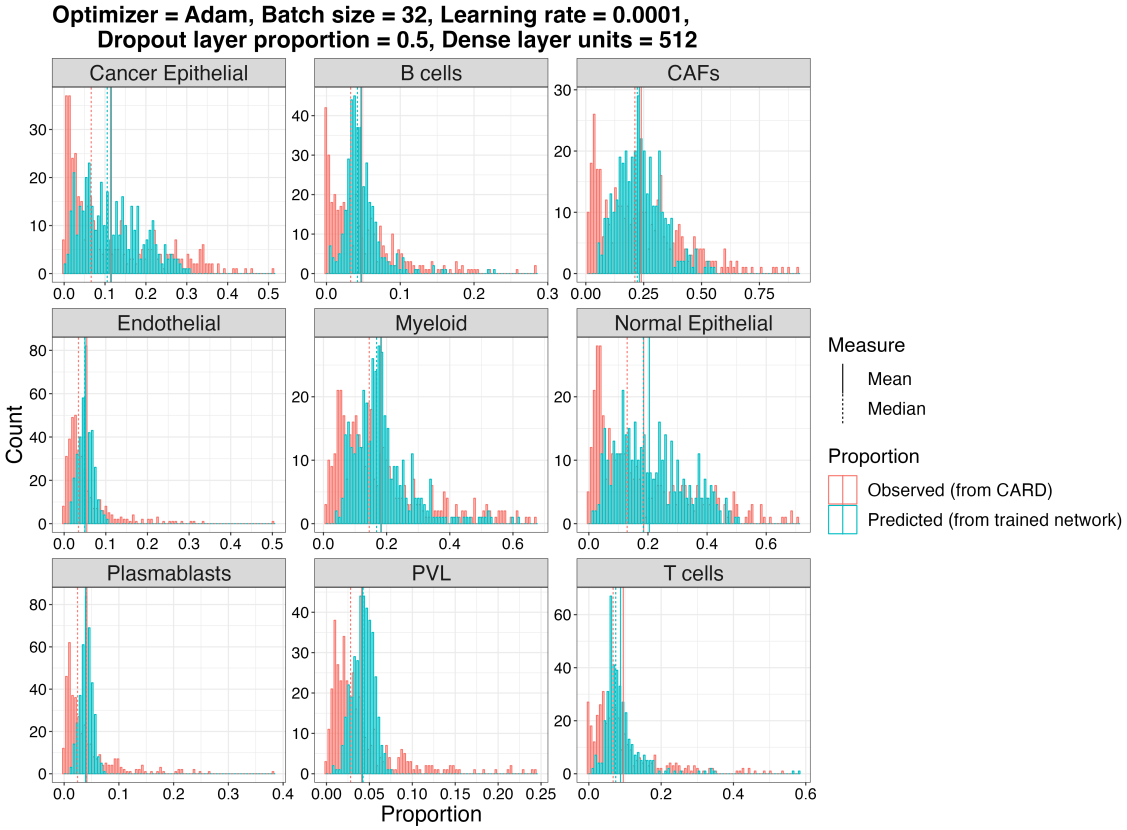


Figure 5.8: Histogram comparing predicted and observed proportions for each of the nine cell types, where the predictions were made by a network trained using standard patches from 10X Genomics FFPE breast tissue.

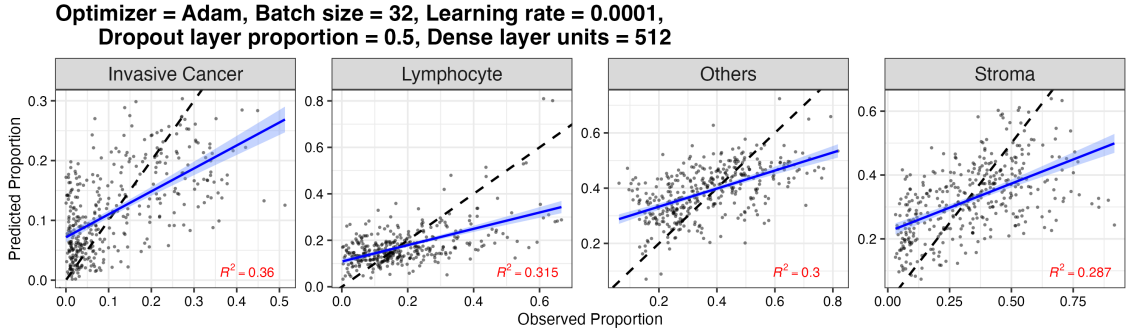


Figure 5.9: Scatterplot comparing four collapsed predicted proportions to four collapsed deconvoluted proportions in the standard patches from 10X Genomics FFPE breast tissue.

Considering the potential presence of cell types with weak signals and morphologically indistinguishable characteristics, accurately predicting proportions for all nine cell types becomes challenging. The limited

proportions of certain cell types also posed challenges. To dig deeper into this matter, we sought to predict the proportions of four collapsed pathological cell type: invasive cancer, stroma, lymphocyte, and others. To accomplish this, we generated scatterplots, illustrating the predicted proportions versus the deconvoluted proportions of the testing dataset for each of the four pathological types (Figure 5.9).

5.1.2 Predicting Four Cell Type Proportions

When we train the neural networks to predict nine cell types and then collapse them into four cell types, there was no noticeable improvement in prediction accuracy (Figure 5.9). Thus, instead of training the neural networks to predict the nine individual cell types, our task shifted towards constructing neural networks designed to directly predict the four pathological cell types. Regarding the results presented in section 5.1.1, we did not include neural networks with SGD as the optimizer in this regression task. However, we kept the remaining parameters in the network architecture and the composition of the training, validation, and testing datasets unchanged from the regression task detailed in section 5.1.1. The cellular compositions of the training, validation, and testing datasets are depicted in Figure 5.10.

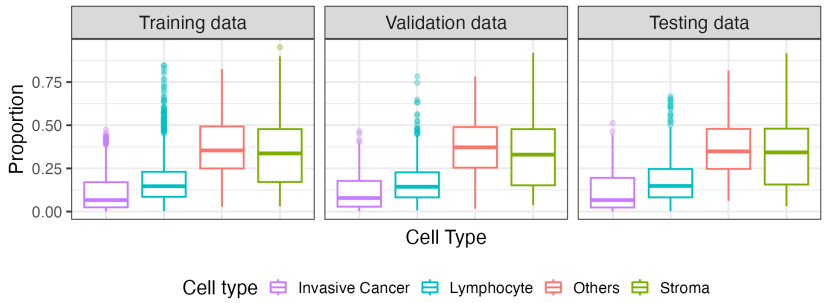


Figure 5.10: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of four cell types using standard patches from 10X Genomics FFPE breast tissue.

Learning Curves

The patterns in the learning curves regarding the trends of loss and metrics remained similar to those in the learning curve of the network designed for predicting nine cell type proportions in 10X Genomics FFPE breast tissue. Examples of learning curves are shown in Figure 5.11.

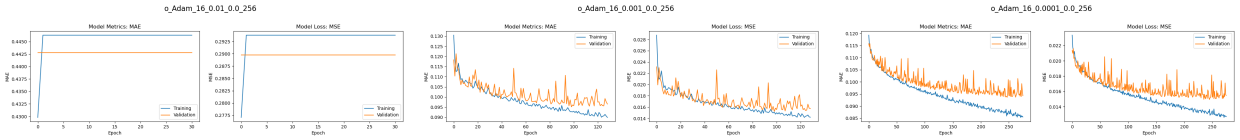


Figure 5.11: Learning curves of neural networks with a batch size of 16, a hidden dense layer of 256 neurons, and no dropout layer aiming to predict the proportions of four cell types using standard patches from 10X Genomics FFPE breast tissue. Each figure title specifies the image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

Evaluation Using Validation Data

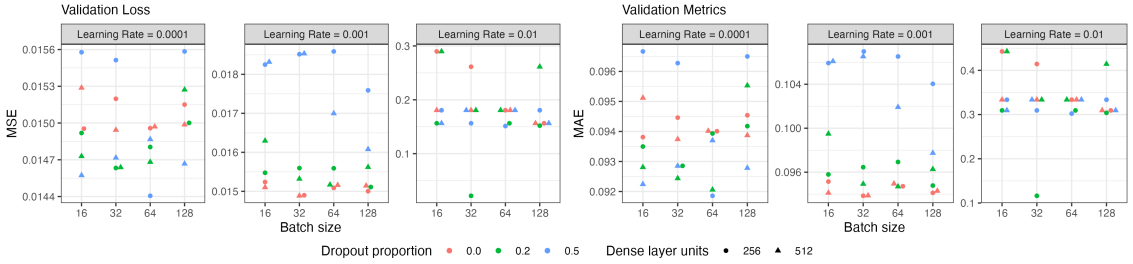


Figure 5.12: Validation loss and metrics from neural networks predicting proportions of four cell types using standard patches from 10X Genomics FFPE breast tissue.

The validation loss and the validation metrics were always consistent with each other (Figure 5.12). The neural network with the lowest validation loss had the following specifications: optimizer = Adam, batch size = 64, learning rate = 0.0001, dropout proportion = 0.5, and dense layer units = 256, which were different from the specifications giving the lowest validation loss in the regression task of predicting nine cell types of the 10X Genomics FFPE breast tissue, as mentioned in section 5.1.1.

The validation loss at the optimal epoch was always higher when the learning rate was set to 0.01 compared to the cases where it was set to 0.001 or 0.0001. Some networks with the Adam optimizer and a learning rate of 0.01 failed to learn from the training data, as indicated in the learning curves. Furthermore, for the same combinations of optimizer, learning rate, batch size of training, and dense layer units, the validation loss at the optimal epoch did not have a generalized trend with increasing dropout proportion. When the networks had learning rate of 0.001, the validation loss increased with increasing dropout proportion.

Predication Using Testing Data

In this section, we analyze the predicted proportions of the testing dataset. We focused only on the networks employing Adam optimizer with learning rates of 0.001 and 0.0001. Our findings indicate that among all cell types, lymphocyte consistently exhibited the highest slopes of best-fit line compared to other cell types, whereas stroma consistently exhibited the lowest slopes (Figure 5.13, C.2). It is noteworthy that there is no single network that can reach the highest slope of all cell types (Table C.2).

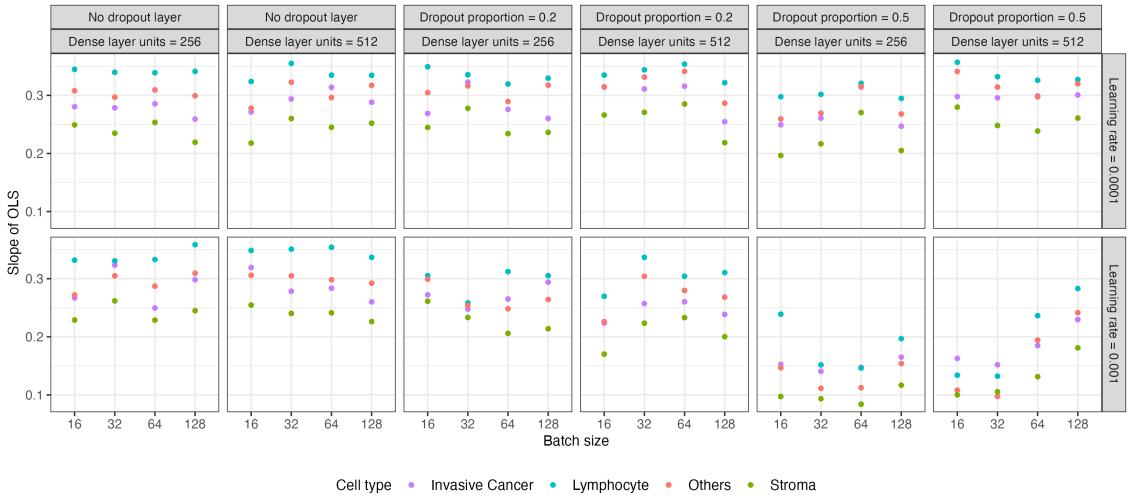


Figure 5.13: Slope of the best-fit line between predicted and observed proportions of four cell types in testing data, predicted by networks trained using standard patches from 10X Genomics FFPE breast tissue, excluding networks with limited learning.

Now we focus on the combination of parameters that gave the lowest validation loss (i.e., optimizer = Adam, batch size = 64, learning rate = 0.0001, dropout proportion = 0.5, dense layer units = 256), which did not give any cell type the highest slope. When examining the scatterplots in Figure 5.14, it can be observed that invasive cancer displayed a slightly weaker correlation between the predicted and observed proportions compared to other cell types. Specifically, when the observed proportions of invasive cancer were low, the predicted proportions exhibited a wider range around the observed values. Conversely, the other cell types demonstrated a relatively consistent range of variation in the predicted proportions around the observed values. The histograms in Figure C.3 suggest that the predicted proportions are slightly more concentrated than the observed proportions around the mean of observed proportions. For invasive cancer and others (i.e., normal epithelial and myeloid), the predicted proportions display different median values compared to the

observed proportions, whereas the predicted proportions of lymphocyte and stroma showed similar median values compared to the corresponding observed proportions.

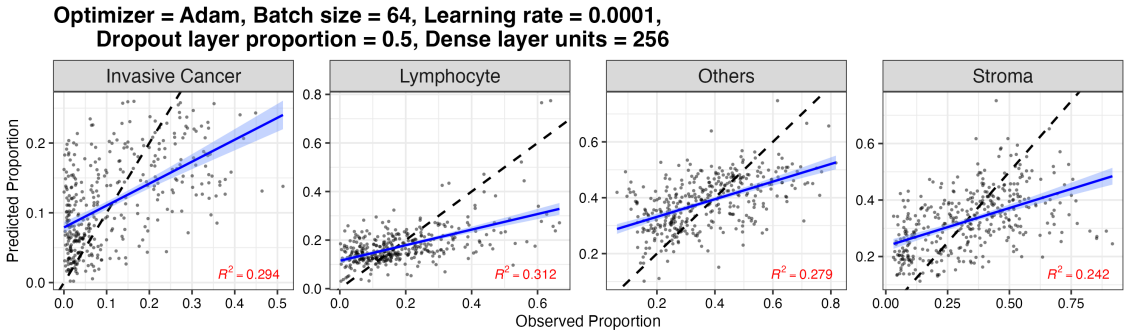


Figure 5.14: Scatterplot comparing predicted and deconvoluted proportions of four cell types in standard patches from 10X Genomics FFPE breast tissue. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. The R-squared value is indicated in red.

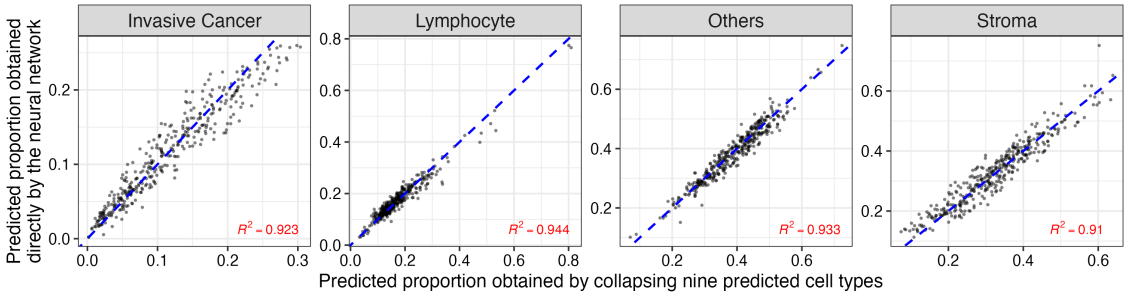


Figure 5.15: Scatterplot comparing predicted proportions achieved by two approaches for standard patches of 10X Genomics FFPE breast tissue derived from identical testing data. The proportions were estimated through two methods: direct prediction by a network with four response variables, and collapsing the proportions of nine major cell types predicted by a network with nine response variables.

The similarity between the two scatterplots, depicted in Figure 5.9 and Figure 5.14, is apparent for all cell types. In Figure 5.9, the predicted proportions were derived by collapsing the proportions of nine major cell types which were predicted by a trained network with nine cell type proportions as the response variable. Conversely, in Figure 5.14, the predicted proportions were obtained directly from another trained network with the four corresponding cell type proportions as the response variable. Both approaches utilized the same patches in the testing dataset, allowing for a comparison between the predicted proportions obtained by the two approaches for each patch (Figure 5.15). Given the concentration of scatter points around the diagonal line, we can conclude that the predicted proportions derived from these two approaches exhibited

similarity, especially for lymphocyte and others (i.e. normal epithelial and myeloid). Nonetheless, based on the adjusted R-squared values, when collapsing the nine predicted proportions, the resulting proportions for all four cell types were better aligned with the observed proportions as compared to predictions obtained directly from the network. This difference in correlation between the predicted and the observed proportions was particularly evident for invasive cancer and stroma.

5.2 Fresh Frozen Human Breast Tissue

5.2.1 Predicting Nine Cell Type Proportions

We evaluated the transfer learning of ResNet50 with different combinations of parameters. The purpose of the neural networks is to predict the nine cell type proportions within individual patches from a WSI of a 10X Genomics fresh frozen Invasive Ductal Carcinoma breast tissue. The varying parameters investigated include the optimizer (Adam, SGD), batch size of training (16, 32, 64, 128), learning rate (0.01, 0.001, 0.0001), dropout layer proportion (0.0, 0.2, 0.5), and the numbers of neurons in the hidden dense layer (256, 512). The training, validation, and testing datasets remained consistent across all network evaluations. The dataset used consisted of a total of 4898 patches, with 3429 patches allocated for training, 734 for validation, and 735 for testing. Notably, the cellular composition of the training, validation, and testing data reveals a predominant presence of cancer epithelial cells (Figure 5.16).

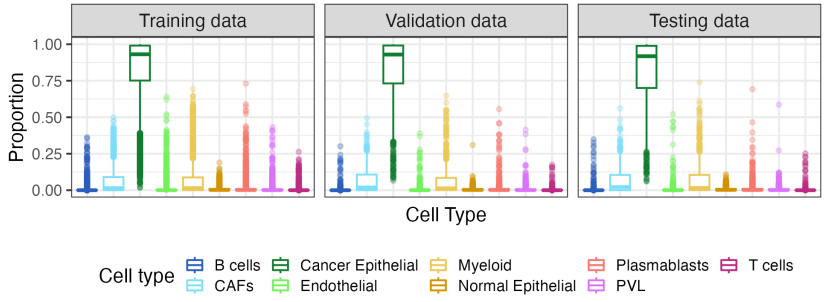


Figure 5.16: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of nine cell types using standard patches from 10X Genomics fresh frozen breast tissue.

Learning Curves

We investigated the learning curves for both the training and validation datasets over the course of the training epochs. Examples of learning curves are shown in Figure 5.17.

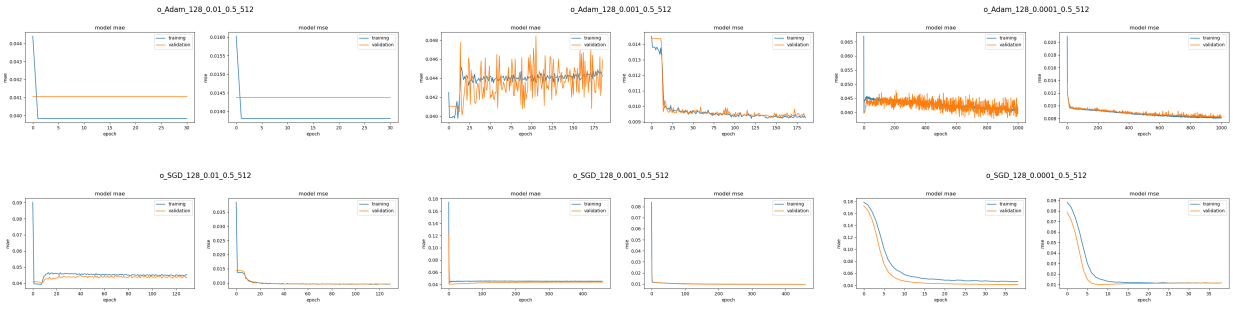


Figure 5.17: Learning curves of neural networks with a batch size of 128, a hidden dense layer of 512 neurons, and a dropout layer proportion of 0.5 aiming to predict the proportions of nine cell types using standard patches from 10X Genomics fresh frozen breast tissue. Each figure title specifies the image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

Similar to the FFPE breast tissue, a learning rate of 0.01 was inadequate for enabling effective learning in neural networks with the Adam optimizer, leading to suboptimal performance and limited progress in training. A majority of the networks employing the Adam optimizer with a learning rate of 0.001 successfully learned from the training data, with only a few networks stopping early at the 31st or 32nd epoch. While the validation loss closely followed the training loss, it exhibited higher levels of fluctuation compared to the training loss. Similarly, the validation metrics displayed much more pronounced fluctuations than the training metrics. These findings suggest that while the Adam optimizer can effectively learn with a learning rate of 0.001, it may struggle to achieve substantial improvements in the training loss and exhibit instability in the validation metrics, potentially indicating challenges in generalizing well to unseen data. Finally, none of the networks employing the Adam optimizer with a learning rate of 0.0001 terminated the training process early right after the initial 30 epochs. These networks exhibited similar patterns in terms of loss and metrics as in those observed in networks using the Adam optimizer with a learning rate of 0.001. However, they generally required more epochs to reach a stage where the validation loss remained unchanged for 30 consecutive epochs.

In contrast, the SGD optimizer demonstrated satisfactory performance with a learning rate of 0.01.

Nonetheless, despite the training loss decreased throughout the training process, the rate of reduction was relatively constrained. The validation loss closely tracked the training loss and did not exhibit signs of overfitting. When the learning rate was set to 0.001, all networks utilizing the SGD optimizer consistently executed the training process for a minimum of 80 epochs. Unlike those with a learning rate of 0.01, these networks demonstrated close alignment not only between the validation loss and the training loss but also between the validation metrics and the training metrics. Notably, certain networks exhibited validation loss and metrics that were lower than the corresponding training loss and metrics. Finally, some networks utilizing the SGD optimizer with a learning rate of 0.0001 terminated the training process early, typically between the 30th and 40th epochs. Additionally, some of these networks displayed validation loss and metrics being smaller than the corresponding training loss and metrics. This observation suggests that the SGD optimizer with a learning rate of 0.0001 may not be optimally suited for achieving significant improvements in performance.

Evaluation Using Validation Data

This section focuses on analyzing and comparing the validation loss and validation metrics for the optimal epoch across different parameter combinations. The validation loss, MSE, and the validation metrics, MAE, were not always consistent with each other. This is likely because larger errors have a larger impact on MSE than MAE. The neural network with the lowest validation loss had the following specifications: optimizer = Adam, batch size = 64, learning rate = 0.0001, dropout proportion = 0.0, and dense layer units = 256.

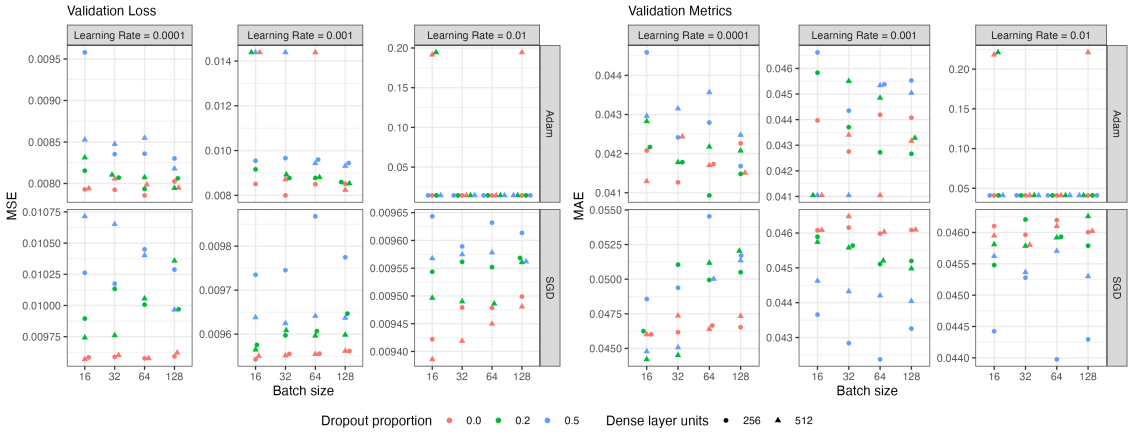


Figure 5.18: Validation loss and metrics from neural networks predicting proportions of nine cell types using standard patches from 10X Genomics fresh frozen breast tissue.

Among the networks utilizing the Adam optimizer, a higher validation loss was generally observed for a learning rate of 0.01, in contrast to a learning rate of 0.001 or 0.0001 (Figure 5.18). It is worth noting that all networks with the Adam optimizer and a learning rate of 0.01 consistently displayed high validation losses. This suggests that these networks failed to learn from the training data, as also indicated in the learning curves. Similarly, certain networks with dense layer units equal to 512 did not effectively learn from the training data when the learning rate was 0.001 for the Adam optimizer. Conversely, the networks employing the SGD optimizer demonstrated more resilient patterns of validation loss in response to variations in the learning rate, except for a slight increase in validation loss when the learning rate was set to 0.0001.

In terms of the learning rate, when it was set to 0.0001, the networks using the Adam optimizer consistently achieved lower validation loss at the optimal epoch compared to networks using SGD, after controlling for other parameters. Similarly, with a learning rate of 0.001, most networks utilizing the Adam optimizer exhibited lower validation loss than those employing the SGD optimizer, except for some networks with dense layer units equal to 512 and not learning from the training data. However, when the learning rate was set to 0.01, networks using the Adam optimizer consistently yielded significantly higher validation loss than those using the SGD optimizer.

Furthermore, the validation loss tends to be higher for higher dropout proportions, suggesting that regulation by dropout is not helpful in this problem. This may be due to the limited sample size.

Predication Using Testing Data

Across all combinations of dense layer and dropout layer, the slopes of the best-fit line between the predicted and observed proportions of each cell type were predominantly close to zero when the learning rate was set to 0.01 for the Adam optimizer, as well as when the SGD optimizer was utilized (Figure C.4). Moreover, networks using the Adam optimizer generally exhibited smaller slopes when the learning rate was set to 0.01 compared to rates of 0.001 and 0.0001 (Figure C.4). As a result, we excluded these combinations of optimizer and learning rate from our analysis, focusing solely on networks utilizing the Adam optimizer with learning rates of 0.001 and 0.0001 (Figure 5.19). It was observed that among all cell types, cancer epithelial cells consistently displayed the highest slope, followed by myeloid cells and CAFs. However, it is important to note that all slopes are relatively small, measuring less than 0.2 (Table C.3).

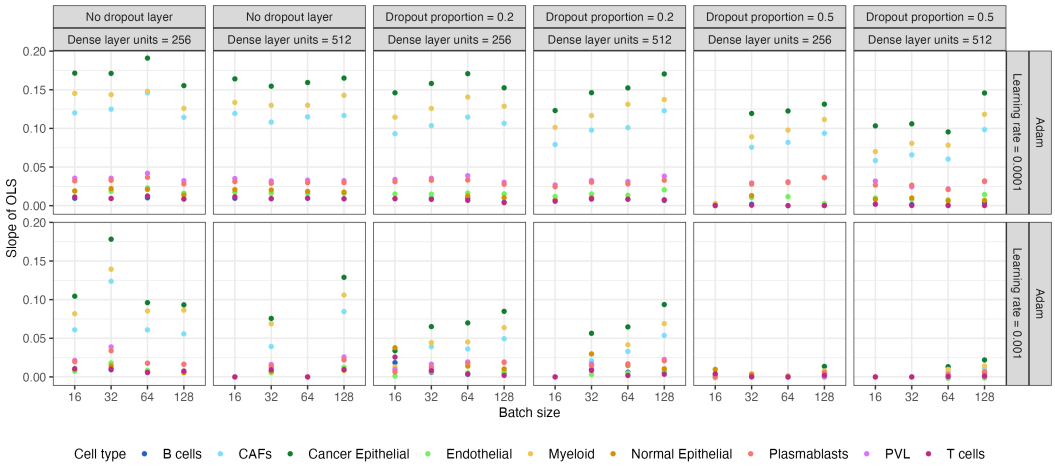


Figure 5.19: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using standard patches from 10X Genomics fresh frozen breast tissue, excluding networks with limited learning.

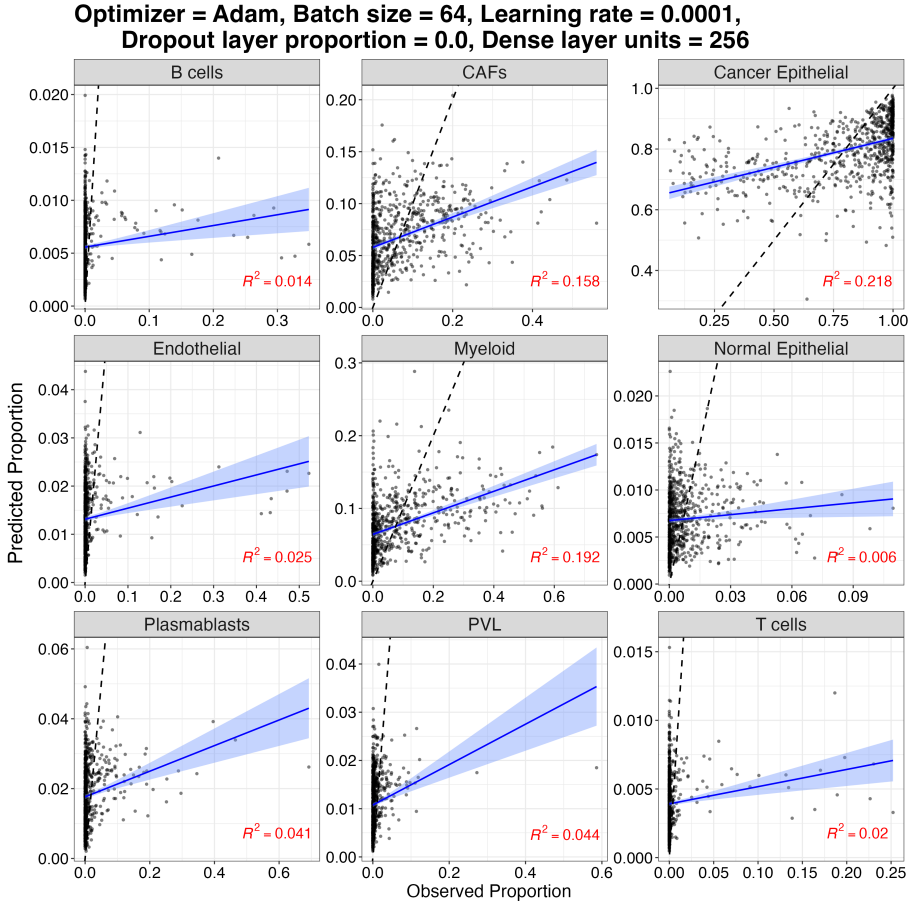


Figure 5.20: Scatterplot comparing predicted and deconvoluted proportions of nine cell types in standard patches from 10X Genomics fresh frozen breast tissue. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. The R-squared value is indicated in red.

We now focus on the parameter combination that yielded the lowest validation loss (optimizer = Adam, batch size = 64, learning rate = 0.0001, dropout proportion = 0.0, dense layer units = 256). Interestingly, this combination also resulted in the highest slope for the best-fit line across most cell types (Table C.3). Similar to the FFPE sample regression tasks, the predicted proportions showed wider ranges around true proportions for very small true proportions (Figure 5.20) and became narrower and more tightly clustered as true proportions increased. Additionally, they exhibited higher concentration compared to observed proportions (Figure C.6), and for most cell types, the predicted proportions had distinct median values while retaining the same mean values as the observed proportions.

Finally, we examined the prediction after collapsing the nine cell types into four cell types. We created scatterplots contrasting the predicted proportions with the deconvoluted proportions from the testing dataset for each of the four cell types (Figure 5.21). There was no noticeable improvement in prediction accuracy after collapsing the nine predicted cell type proportions into four cell types.

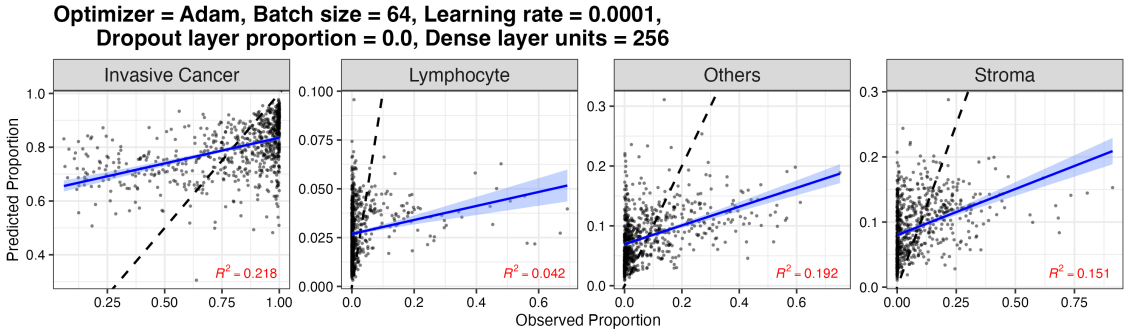


Figure 5.21: Scatterplot comparing four collapsed predicted proportions to four collapsed deconvoluted proportions in the standard patches from 10X Genomics fresh frozen breast tissue.

5.2.2 Predicting Four Cell Type Proportions

Instead of training the neural networks to predict the nine individual cell types, we modified our neural networks to directly predict the four cell types. We did not use SGD as an optimizer due to its poor performance compared to Adam. We kept the remaining parameters in the network architecture and the patches in the training, validation, and testing datasets unchanged from the previous regression task detailed in section 5.2.1. Notably, the cellular composition of the training, validation, and testing sets continued to exhibit a significant predominance of invasive cancer cells (Figure 5.22).

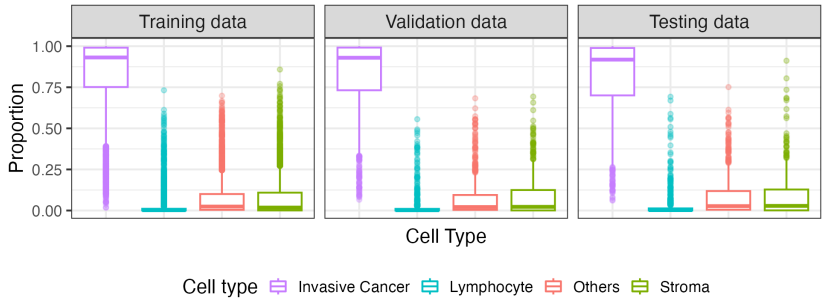


Figure 5.22: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of four cell types using standard patches from 10X Genomics fresh frozen breast tissue.

Learning Curves

Similar to the previous regression task on predicting the nine cell type proportions, all networks with a learning rate of 0.01 and some networks with a learning rate of 0.001 did not learn from the training data (Figure 5.23).

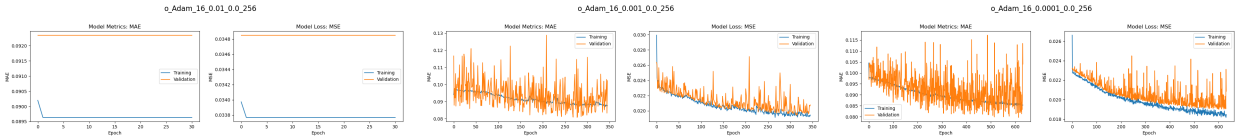


Figure 5.23: Learning curves of neural networks with a batch size of 16, a hidden dense layer of 256 neurons, and no dropout layer aiming to predict the proportions of four cell types using standard patches from 10X Genomics fresh frozen breast tissue. Each figure title specifies the image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

Evaluation Using Validation Data

This section focuses on analyzing and comparing the validation loss (MSE) and validation metrics (MAE) for the optimal epoch across various parameter combinations. The neural network with the lowest validation loss exhibited the following specifications: optimizer = Adam, batch size = 64, learning rate = 0.0001, dropout proportion = 0.0, and dense layer units = 256, which were the same as the specifications giving the lowest validation loss in the regression task of predicting nine cell types, as detailed in section 5.2.1.

Figure 5.24 clearly demonstrates a decreasing trend in the validation loss at the optimal epoch as the

learning rate decreases from 0.01 to 0.0001, while keeping all other parameters constant. It is important to note that all networks with the Adam optimizer and a learning rate of 0.01 failed to learn from the training data. Additionally, some networks with a learning rate of 0.001 also performed poorly, as evident from the learning curves. Consequently, networks with a learning rate of 0.01, which shared the same validation loss and metrics at the optimal epoch, are excluded from the subsequent analysis. For the same combinations of the optimizer, learning rate, batch size of training, and dense layer units, the validation loss at the optimal epoch did not exhibit a consistent trend with increasing dropout proportion. Specifically, when the learning rate was set to 0.001 and the dense layer had 256 units, the validation loss increased as the dropout proportion increased.

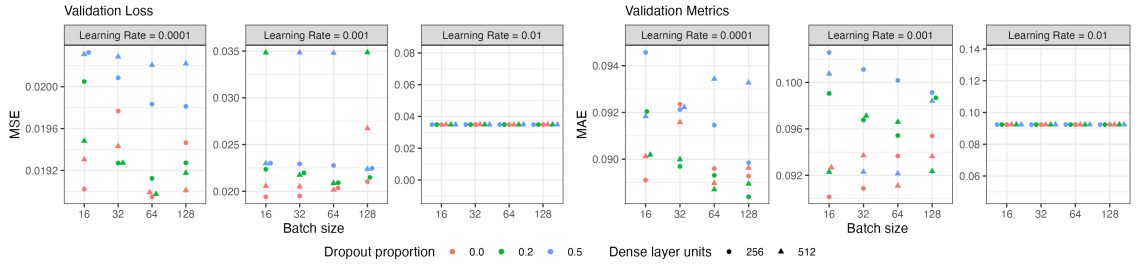


Figure 5.24: Validation loss and metrics from neural networks predicting proportions of four cell types using standard patches from 10X Genomics fresh frozen breast tissue.

Predication Using Testing Data

We analyzed the predicted proportions of the testing dataset and examined the alignment between the predicted and observed proportions using the slope of the best-fit line. Our focus was solely on networks that utilized the Adam optimizer with learning rates of 0.001 and 0.0001 (Figure 5.25). When the learning rate was set to 0.001, we observed that an increase in the dropout layer proportion was associated with a lower slope of the best-fit line for all cell types except for lymphocyte. Similarly, when the learning rate was set to 0.0001, we found that, for all cell types except for lymphocyte, the networks with a dropout layer proportion of 0.5 exhibited lower slopes of the best-fit line compared to networks without a dropout layer or with a dropout proportion of 0.2. Furthermore, it is worth noting that networks trained with a learning rate of 0.0001 consistently displayed an increasing slope across different cell types. Specifically, the order of increasing slope was as follows: lymphocyte, stroma, others, and invasive cancer. Besides, there is no single network that can reach the highest slope of all cell types (Table C.4). The network that gave most

cell types the highest slope of best-fit line was not the one with the lowest validation loss at the optimal epoch. Here we focus on the combination of parameters that gave the lowest validation loss (i.e., optimizer = Adam, batch size = 64, learning rate = 0.0001, dropout proportion = 0.0, dense layer units = 256), which did not give any cell type the highest slope.

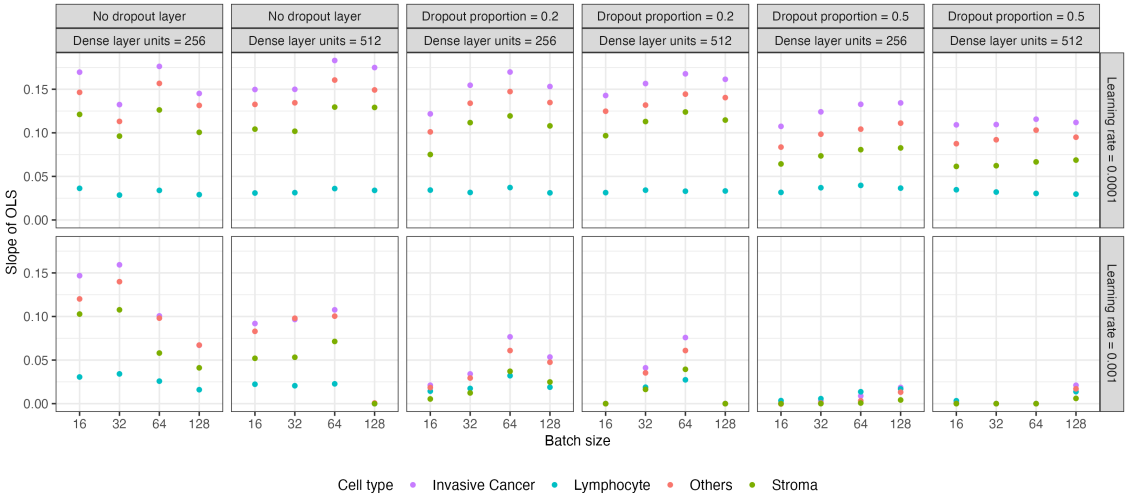


Figure 5.25: Slope of the best-fit line between predicted and observed proportions of four cell types in testing data, predicted by networks trained using standard patches from 10X Genomics fresh frozen breast tissue.

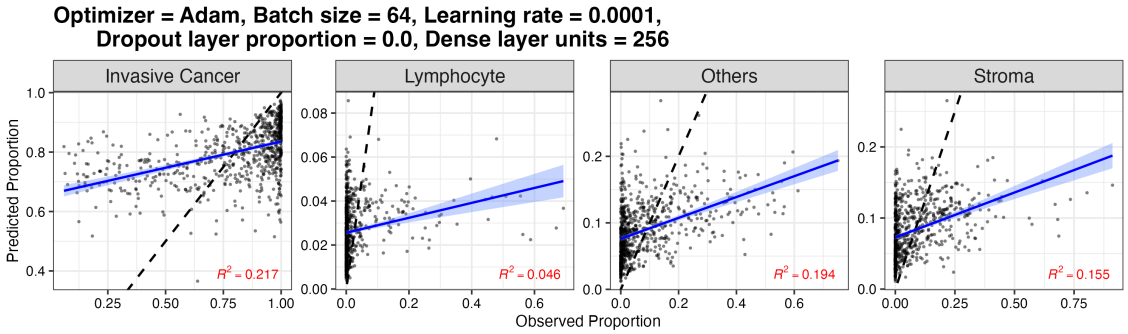


Figure 5.26: Scatterplot comparing predicted and deconvoluted proportions of four cell types in standard patches from 10X Genomics fresh frozen breast tissue. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. The R-squared value is indicated in red.

Based on the scatterplots in Figure 5.26, it is apparent that there is a notable lack of correlation between the predicted and observed proportions across all cell types. Specifically, when focusing on the predominant cell type, invasive cancer, the network tended to overestimate the proportion in patches where the observed

proportion was low. Conversely, for other cell types, the network consistently underestimated the proportion when the observed proportion was high. The predicted proportions are more concentrated than the observed proportions, and for the majority of cell types, the predicted proportions display different median values compared to the observed proportions, but always have the same mean values (Figure C.7).

The predicted proportions of the four cell types can be obtained by either collapsing the nine predicted cell types or directly from the neural network. The predicted proportions derived from the two approaches exhibited remarkable similarity given the close concentration of scatter points around the diagonal line in Figure 5.27. For stroma, the predictions obtained by collapsing nine predicted cell types were generally smaller than the predictions obtained directly by the trained neural network. The opposite pattern was observed for invasive cancer and others (normal epithelial + myeloid). Furthermore, the correlations between the predicted and observed proportions remained nearly unchanged for all cell types when comparing the R-squared values in Figure 5.21 and Figure 5.26 for predictions obtained using both approaches.

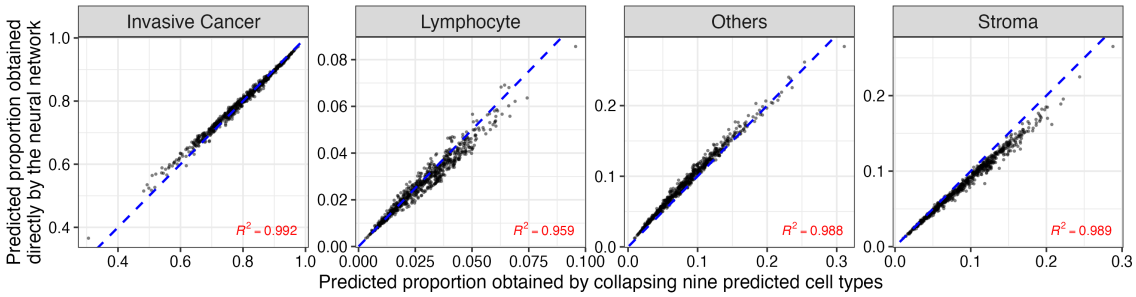


Figure 5.27: Scatterplot comparing predicted proportions achieved by two approaches for standard patches of 10X Genomics fresh frozen breast tissue derived from identical testing data. The proportions were estimated through two methods: direct prediction by a network with four response variables, and collapsing the proportions of nine major cell types predicted by a network with nine response variables.

5.3 Both Fresh Frozen and FFPE Tissues

In general, the neural networks exhibited better predictive performance when applied to FFPE tissue compared to fresh frozen tissue. This could be attributed to the lack of diversity in cell type compositions within the training data for the fresh frozen tissue. As a result, the networks failed to acquire enough information, leading to inaccurate predictions when encountering unseen data with different cell type compositions than the majority in the training data. To address this issue and enhance the predictive performance of the

networks, we merged patches from both tissue types and split the patches into training, validation, and testing datasets. This approach served two purposes: increasing the diversity of cell type compositions within the training data and expanding the sample size for more effective and comprehensive network training. Based on the findings in section 5.1.1 and section 5.1.2, we only constructed networks utilizing Adam as an optimizer with learning rates of 0.001 and 0.0001.

5.3.1 Predicting Nine Cell Type Proportions

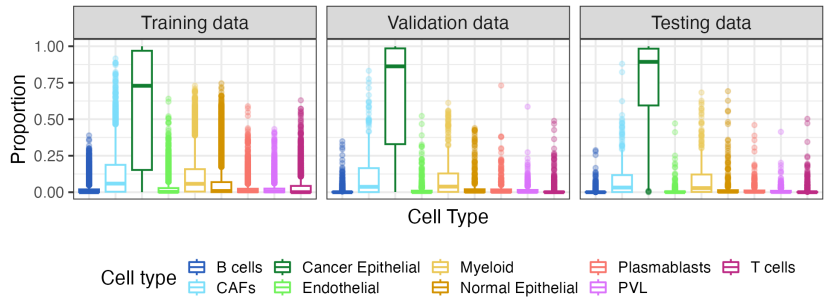


Figure 5.28: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of nine cell types using standard patches from both 10X Genomics tissues.

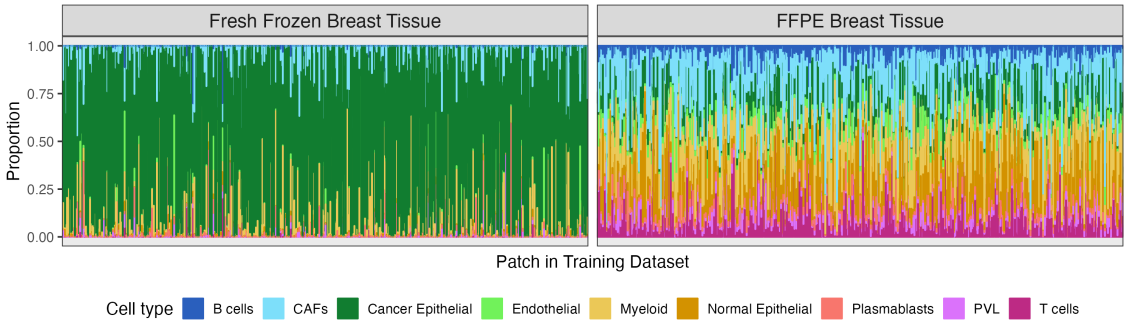


Figure 5.29: Cell type compositions in training dataset for regression task predicting proportions of nine cell types using standard patches from both 10X Genomics breast tissues.

Neural networks were designed to predict the proportions of nine cell types in each patch derived from both fresh frozen and FFPE human breast tissues. The evaluation involves examining different batch sizes for training (16, 32, 64, 128), learning rates (0.001, 0.0001), dropout layer proportions (0.0, 0.2, 0.5), and the number of neurons of the hidden dense layer (256, 512). The datasets used for training, validation, and testing remain consistent across all network configurations. The overall dataset comprises a total of 6250

patches, with 5191 patches allocated for training, 578 patches for validation, and 481 patches for testing (Table 5.1). Notably, the training, validation, and testing data reveal a predominance of cancer epithelial (Figure 5.28) and this is primarily due to the enrichment of cancer epithelial cells from the fresh frozen tissue sample (Figure 5.29).

	Number of patches		
	Training	Validation	Testing
Fresh Frozen	3431	470	416
FFPE	1760	108	65

Table 5.1: Number of standard patches from each of the two 10X Genomics samples in the training, validation, and testing datasets.

Learning Curves

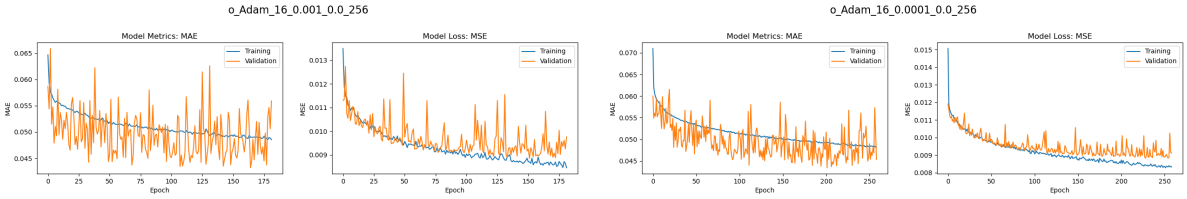


Figure 5.30: Learning curves of neural networks with a batch size of 16, a hidden dense layer of 256 neurons, and no dropout layer aiming to predict the proportions of nine cell types using standard patches from both 10X Genomics breast tissues. Each figure title specifies the image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

The learning curves demonstrate that all networks effectively learned from the training data. Networks with a learning rate of 0.0001 required approximately the same or more epochs to reach a stable state in terms of validation loss compared to networks with a learning rate of 0.001. In general, the validation loss and metrics exhibited more fluctuations than the training loss and metrics. Besides, while the validation loss consistently exceeded the training loss for certain networks, the validation metrics were generally lower than the training metrics across all networks.

Evaluation Using Validation Data

The validation loss and metrics did not always consistent with each other (Figure 5.31). The neural network that achieved the lowest validation loss had the following specifications: optimizer = Adam, batch size = 32,

learning rate = 0.001, dropout proportion = 0.0, and dense layer units = 512.

Among networks with a dropout layer, those with a learning rate of 0.001 consistently had higher validation loss than those with a learning rate of 0.0001, while maintaining all other parameters constant. Regarding the impact of batch size on the validation loss, no consistent trend was observed when the learning rate was set to 0.0001. However, with a learning rate of 0.001 and a dropout layer proportion of 0.5, we observed an increasing trend of validation loss as the batch size increased. Furthermore, a higher dropout layer proportion was found to be associated with increased validation loss at the optimal epoch when the learning rate was set to 0.001. We did not observe this consistent association among networks with a learning rate of 0.0001. In terms of the hidden dense layer neurons, networks with 512 neurons had lower validation loss compared to those with 256 neurons when the learning rate was set to 0.001. However, an exception was observed when the batch size was 128, where a higher number of neurons led to higher validation loss. No consistent association between the hidden dense layer and validation loss was found when the learning rate was set to 0.0001.

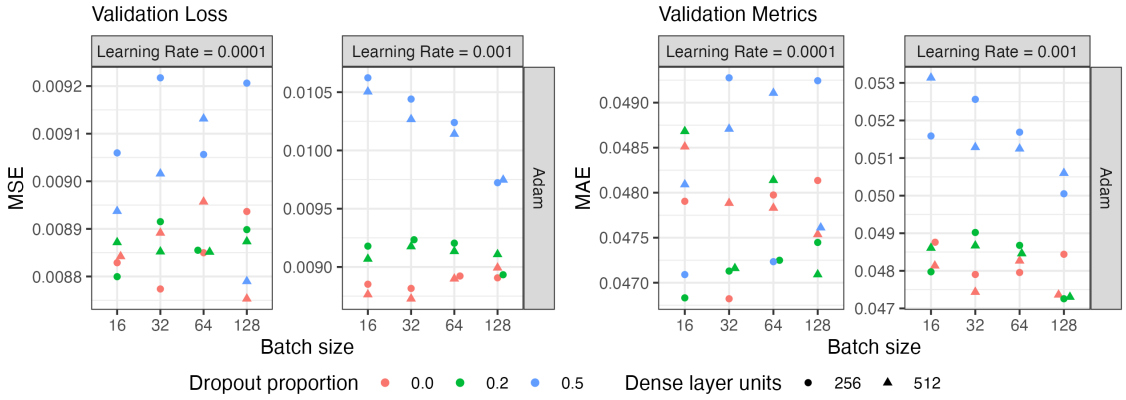


Figure 5.31: Validation loss and metrics from neural networks predicting proportions of nine cell types using standard patches from both 10X Genomics breast tissues.

Predication Using Testing Data

To evaluate the relationship between the predicted and observed proportions of each cell type, we examined the slope of best-fit between the predicted and observed proportions. The slopes of networks with varying learning rates were found to be relatively similar when other parameters were held constant (Figure C.8). T cells consistently displayed higher slopes compared to all other cell types, except when the network included

a dropout layer proportion of 0.5 and had a learning rate of 0.001 (Figure 5.32). Both cancer epithelial and normal epithelial cells also exhibited higher slopes compared to the remaining cell types. CAFs ranked fourth highest in terms of slope among all cell types. On the other hand, plasmablasts consistently showed the smallest slope when compared to the other cell types.

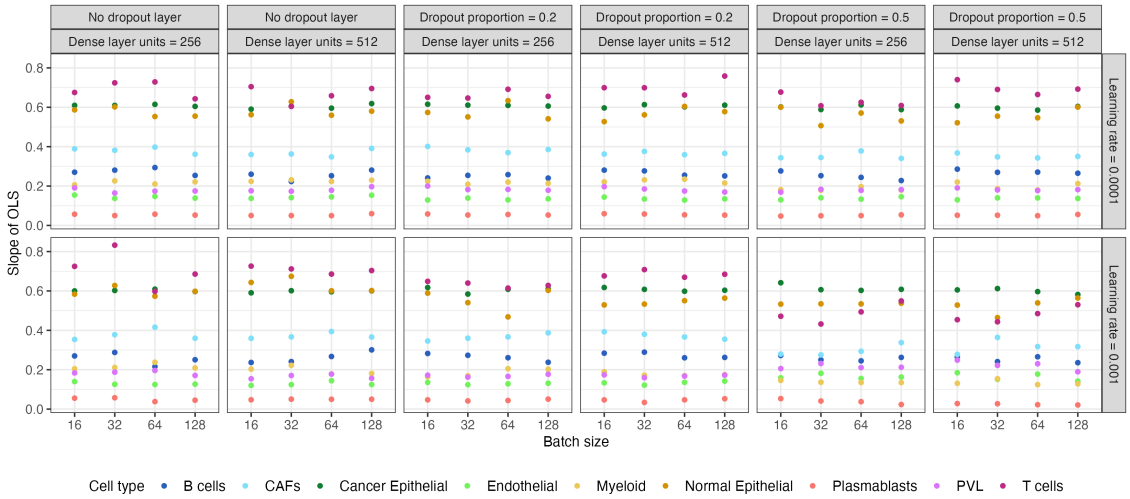


Figure 5.32: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using standard patches from both 10X Genomics breast tissues.

Furthermore, no single network configuration was capable of achieving the highest slope for all cell types (Table C.5). However, with the exception of plasmablasts, which exhibited the lowest slope among all cell types, the remaining cell types attained their highest slopes when utilizing networks with a learning rate of 0.001. These networks either lacked a dropout layer or employed a high dropout proportion of 0.5. We now focus on the combination of parameters that yielded the lowest validation loss, with the following parameters: optimizer = Adam, batch size = 32, learning rate = 0.001, dropout proportion = 0.0, and dense layer units = 512.

The scatterplots presented in Figure 5.33 demonstrate that although certain cell types displayed high values of the slope of the best-fit line, the correlations between the predicted and observed proportions were not as strong as initially suggested. For cell types with high overall R-squared values, the sample-specific R-squared values were considerably smaller compared to the overall value. This implies that while the networks were capable of distinguishing patches with substantial differences in observed proportions and providing predictions within the correct relative range of magnitudes, the accuracy of predictions within each

sample was lacking. For instance, in the case of cancer epithelial cells, the majority of patches from fresh frozen tissue exhibited higher observed proportions compared to patches from FFPE samples. While the network correctly predicted that almost all patches from fresh frozen tissue would have a higher proportion of cancer epithelial cells than the patches from FFPE tissue, it was unable to distinguish the differences in observed proportions within each sample.

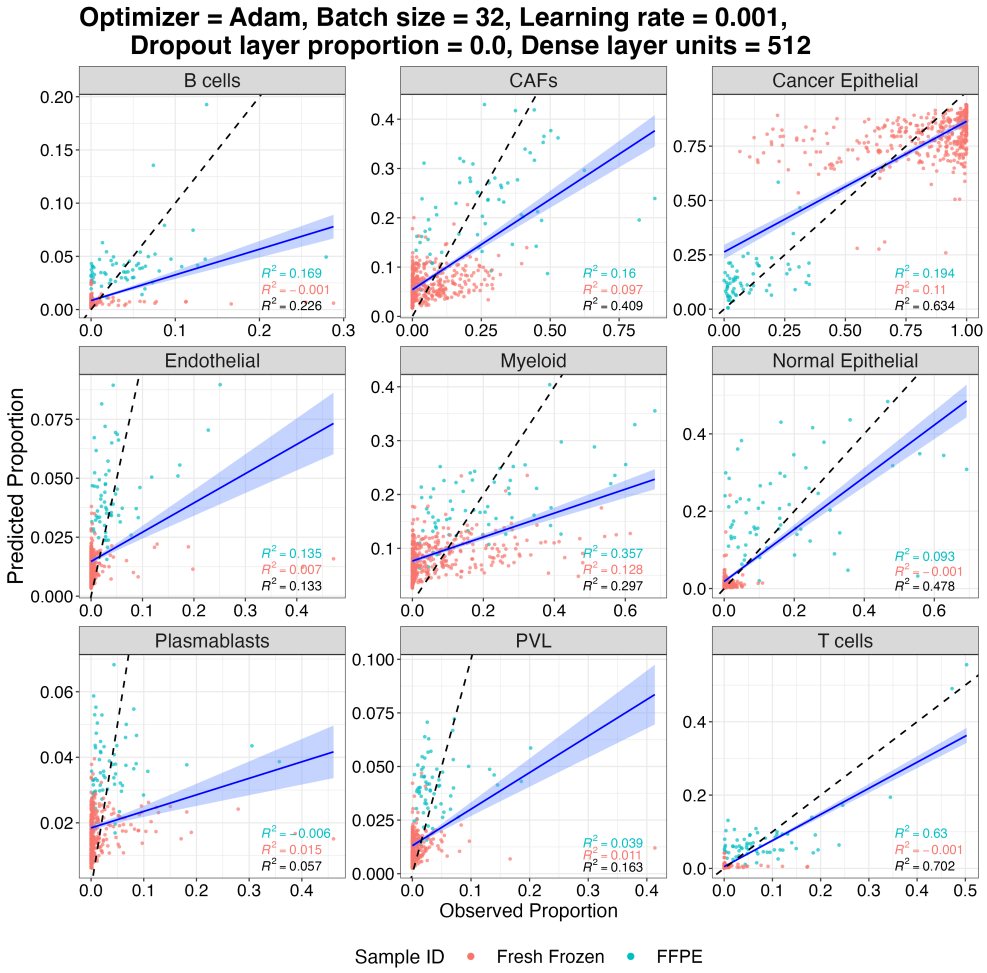


Figure 5.33: Scatterplot comparing predicted and deconvoluted proportions of nine cell types in standard patches from both 10X Genomics breast tissues. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. R-squared values are indicated in black for the overall fit and in corresponding colors for each sample.

For all cell types, the FFPE tissue consistently exhibited higher adjusted R-squared values compared to the fresh frozen tissue. This indicates that the network achieved better predictive performance on the FFPE tissue samples. Figure 5.34 and Figure C.9 also illustrated this difference in predictive performance.

Essentially, conducting the regression task separately on the two tissue types yielded similar results. Despite merging patches from both tissue types, the network still struggled to accurately predict cell type proportions in patches from the fresh frozen tissue.

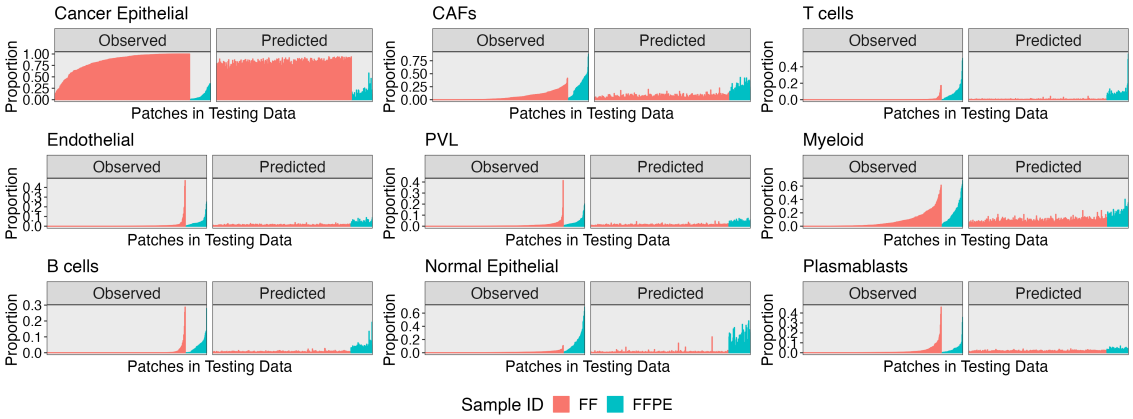


Figure 5.34: Comparison of predicted and observed nine cell type compositions in each of the standard patches from both 10X Genomics tissues in the testing data.

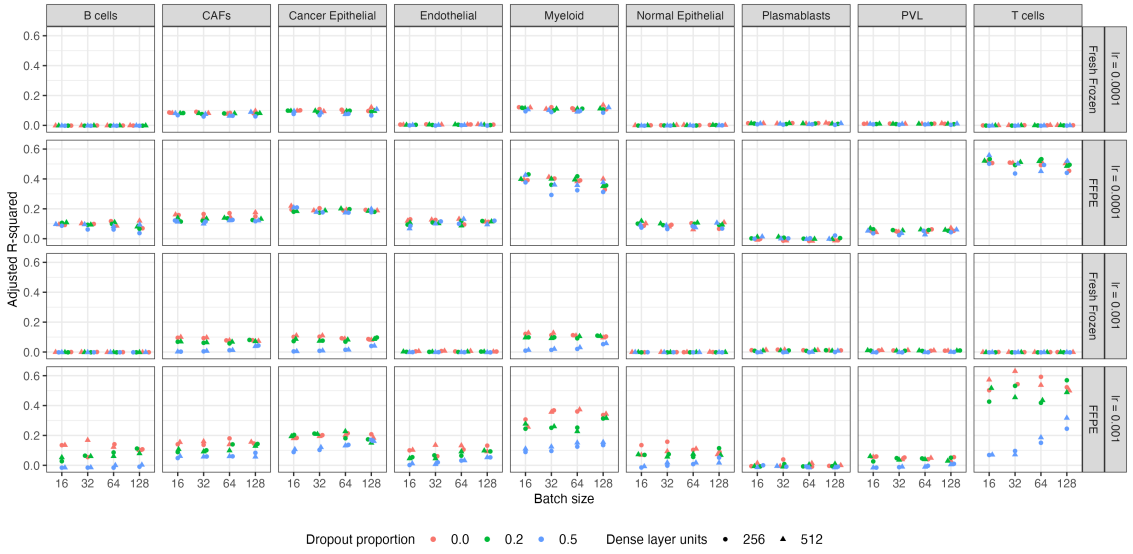


Figure 5.35: Sample-specific adjusted R-squared values of the testing dataset for networks predicting proportions of nine cell types using standard patches from both 10X Genomics breast tissues.

Figure 5.35 illustrates the correlation between the predicted and observed proportions for each cell type in each of the tissue samples. For the fresh frozen tissue, among the cell types, only CAFs, cancer epithelial, and myeloid displayed some level of correlation between the predicted and observed proportions across all

networks. Conversely, for all other cell types, the correlations were negligible across all networks. In the case of the FFPE tissue, only plasmablasts exhibited negligible correlation across all networks. Moreover, for each cell type, the FFPE patches generally demonstrated similar or stronger correlation between the predicted and observed proportions compared to the fresh frozen tissue.

We examined the correlation between the four collapsed pathological cell types by scatterplots, illustrating the predicted proportions versus the deconvoluted proportions of the testing dataset for each of the four pathological types (Figure 5.36).

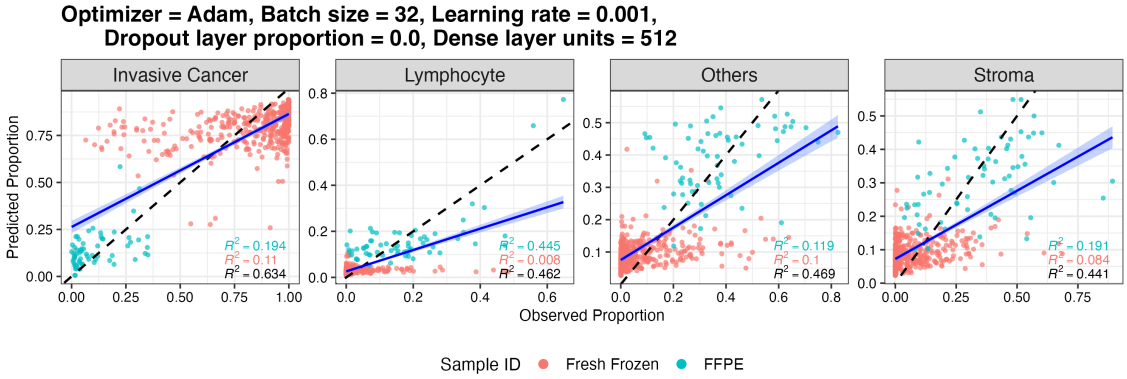


Figure 5.36: Scatterplot comparing four collapsed predicted proportions to four collapsed deconvoluted proportions in the standard patches from both 10X Genomics breast tissues.

5.3.2 Predicting Four Cell Type Proportions

Instead of training the neural networks to predict the nine individual cell types, our task shifted towards constructing neural networks designed to directly predict the four pathological cell types. Regarding the results presented in section 5.3.1, We kept the parameters in the network architecture and the standard patches in the training, validation, and testing datasets unchanged. The cellular compositions of the training, validation, and testing datasets are depicted in Figure 5.37. The predominance of invasive cancer in the fresh frozen tissue caused the predominance of invasive cancer in the overall training dataset (Figure 5.38).

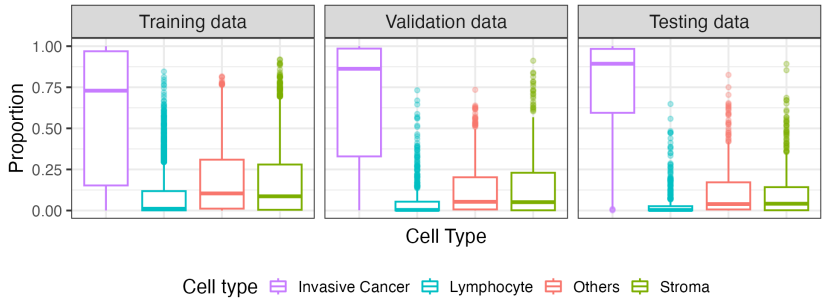


Figure 5.37: Cell type compositions of patches in the training, validation, and testing datasets used for the regression task, which aimed to predict the proportions of four pathological cell types using standard patches from both breast tissues provided by 10X Genomics.

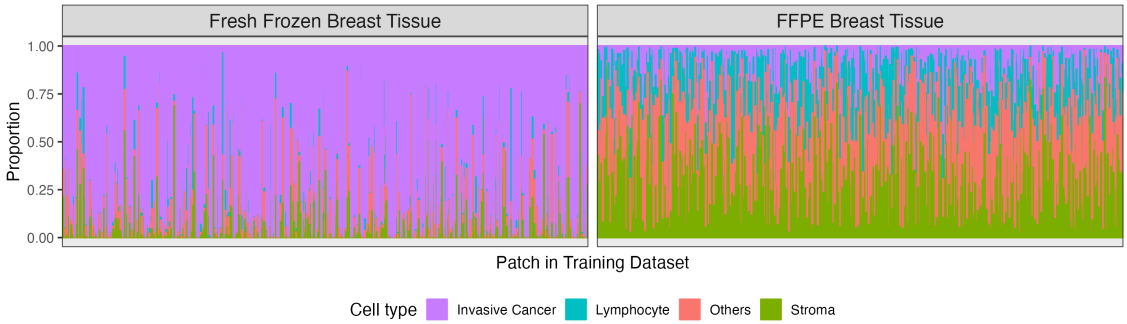


Figure 5.38: Cell type compositions in training dataset for regression task predicting proportions of four cell types using standard patches from both 10X Genomics breast tissues.

Learning Curves

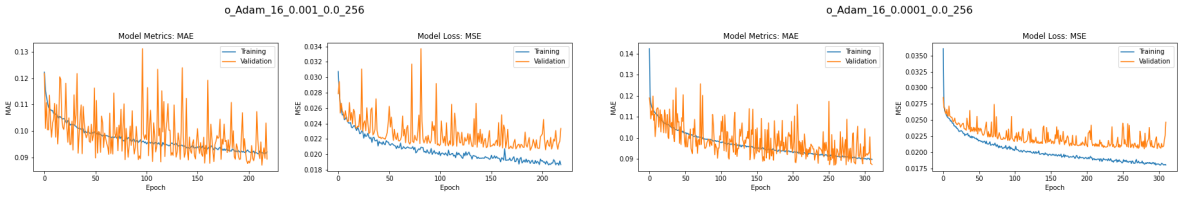


Figure 5.39: Learning curves of neural networks with a batch size of 16, a hidden dense layer of 256 neurons, and no dropout layer aiming to predict the proportions of four cell types using standard patches from both 10X Genomics breast tissues. Each figure title specifies the image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

The learning curves indicate that all networks successfully learned from the training data. Networks with a learning rate of 0.0001 generally required more epochs to stabilize the validation loss compared to net-

works with a learning rate of 0.001. The validation metrics displayed much greater fluctuations than the training metrics. However, unlike the networks designed for the regression task of predicting nine cell type proportions, the validation loss consistently exceeded the training loss for all networks.

Evaluation Using Validation Data

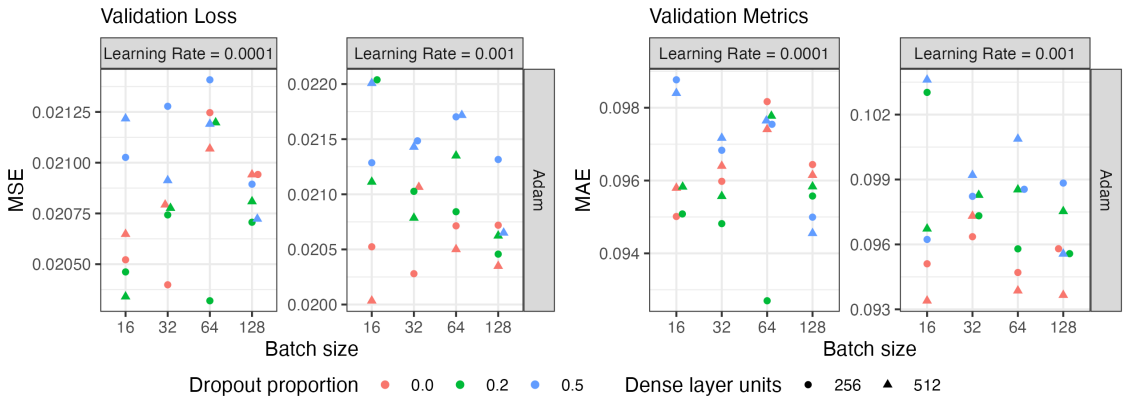


Figure 5.40: Validation loss and metrics from neural networks predicting proportions of four cell types using standard patches from both 10X Genomics breast tissues.

The validation loss and metrics exhibited inconsistencies with each other (Figure 5.40). The neural network with the lowest validation loss had the following specifications: optimizer = Adam, batch size = 16, learning rate = 0.001, dropout proportion = 0.0, and dense layer units = 512. It is worth noting that there was no consistent association observed between the validation loss at the optimal epoch and any of the parameters across all networks.

Prediction Using Testing Data

The slopes of best-fit line between the predicted and the observed proportions in testing data were depicted in Figure 5.41. Invasive cancer cells consistently showed the largest slope when compared to the other cell types. We focus on the combination of parameters that yielded the lowest validation loss, with the following parameters: optimizer = Adam, batch size = 16, learning rate = 0.001, dropout proportion = 0.0, and dense layer units = 512 (Figure 5.42). Similar to all other regression tasks, the predicted proportions had lower variance, different median, and equal mean as compared to the observed proportions (Figure C.11). Besides, this network achieved the highest slopes for both lymphocyte and invasive cancer (Table C.6). Additionally,

another network, which had a different batch size compared to the network with the lowest validation loss, simultaneously achieved the highest slopes for stroma and others.

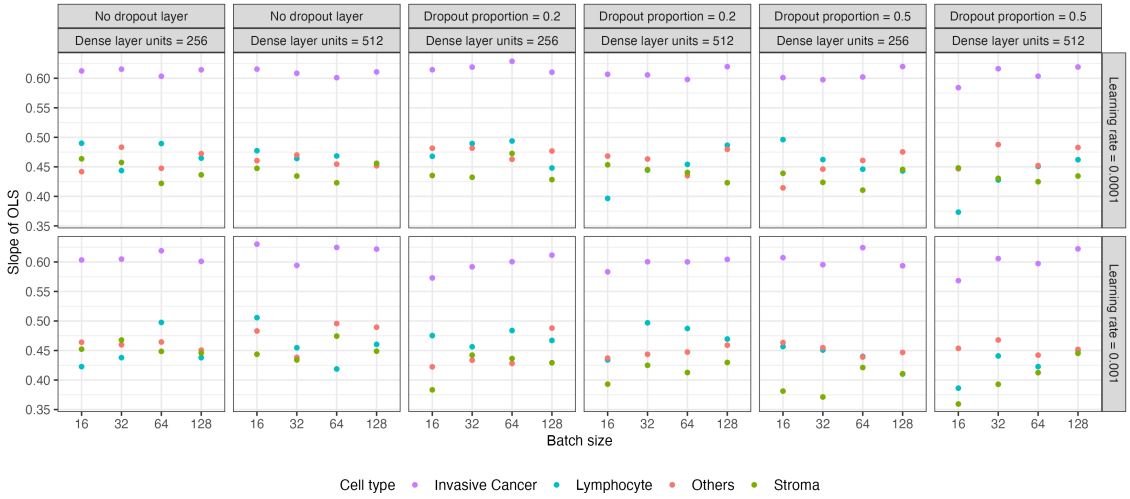


Figure 5.41: Slope of the best-fit line between predicted and observed proportions of four cell types in testing data, predicted by networks trained using standard patches from both 10X Genomics breast tissues.

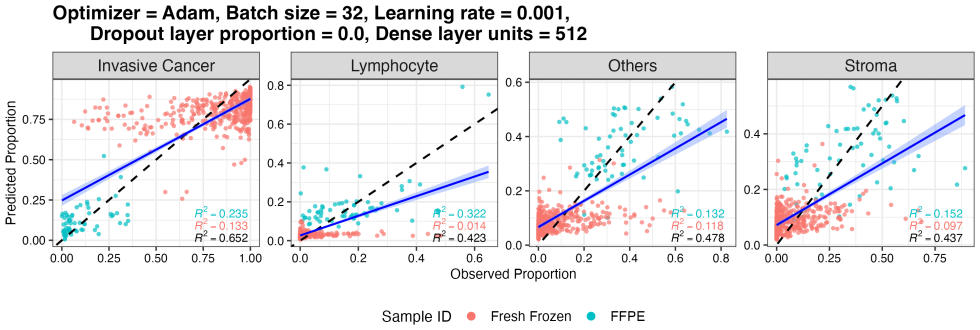


Figure 5.42: Scatterplot comparing predicted and deconvoluted proportions of four cell types in standard patches from both 10X Genomics breast tissues. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. R-squared values are indicated in black for the overall fit and in corresponding colors for each sample.

The two scatterplots shown in Figure 5.36 and Figure 5.42 exhibit similarities for all cell types. Both overall and sample-specific correlations between the predicted and observed proportions were higher for invasive cancer and others (normal epithelial and myeloid) when the predictions were directly obtained from the neural network. Although the overall correlations for lymphocyte and stroma were lower when predictions were obtained directly from the neural network, their sample-specific correlations in fresh frozen tissue

were higher. Moreover, based on Figure 5.43, the overall predictions from the two approaches were similar to each other. However, there were substantial differences in the predictions made by the two approaches for FFPE tissue.

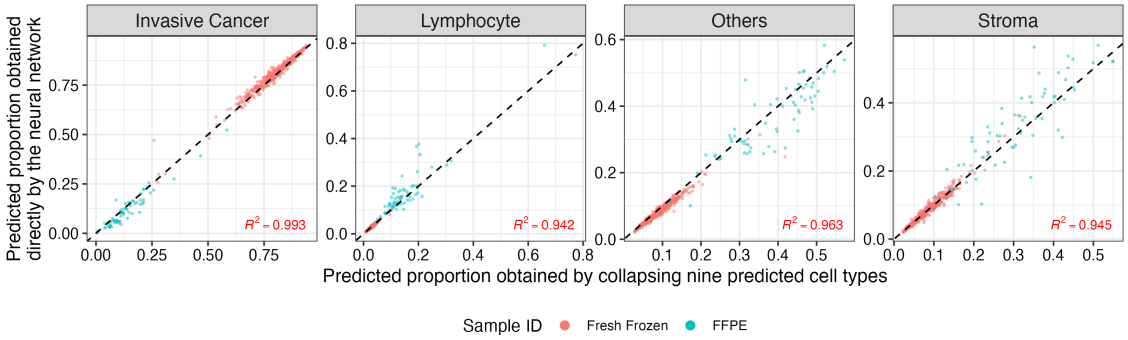


Figure 5.43: Scatterplot comparing predicted proportions achieved by two approaches for standard patches of both 10X Genomics breast tissues derived from identical testing data. The proportions were estimated through two methods: direct prediction by a network with four response variables, and collapsing the proportions of nine major cell types predicted by a network with nine response variables.

5.4 Comparison between Lasso Regression and Transfer Learning

Once the pre-trained ResNet50 model extracted 2048 features, we employed two different approaches to predict nine cell type proportions. The first approach involved using these features as predictors and the cell type proportions as responses, fitting Lasso regressions as described in Chapter 4. The second approach utilized transfer learning from pre-trained ResNet50. Our objective was to compare the predictive performance of these two methods.

In Chapter 4, we obtained three metrics for the Lasso regressions that were fitted using the features extracted by the pre-trained ResNet50: mean squared error, adjusted R-squared, and the slope of the best-fit line. Additionally, we calculated these three metrics for the neural network that yielded the lowest validation loss in each of the two samples. For the networks trained with fresh frozen tissue, the best neural network had the following configurations: optimizer = Adam, batch size = 64, learning rate = 0.0001, dropout proportion = 0.0, dense layer units = 256. For the networks trained with FFPE tissue, the best neural network had the following configurations: optimizer = Adam, batch size = 32, learning rate = 0.0001, dropout proportion = 0.5, and dense layer units = 512.

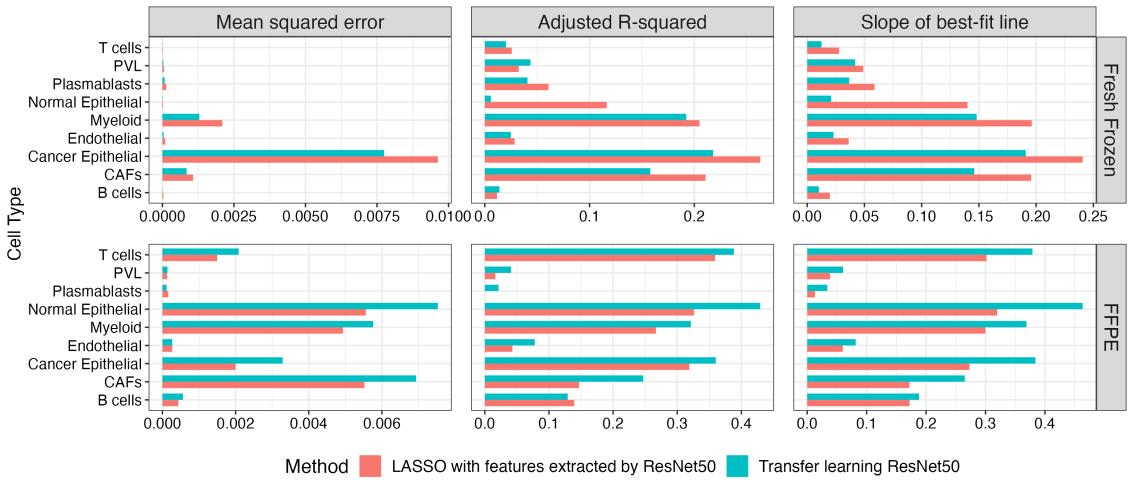


Figure 5.44: Comparison between the predictive performance of Lasso regression and transfer learning from ResNet50 by assessing the mean squared error, adjusted R-squared, and slope of best-fit line on the testing dataset.

Since both the Lasso regressions and the neural networks were trained with the same training dataset and tested with the same testing dataset, we can compare the testing data metrics obtained from the Lasso regressions and the transfer learning of ResNet50. Observing Figure 5.44, it can be concluded that transfer learning exhibited superior predictive performance compared to Lasso regression when applied to FFPE breast tissue. In contrast, Lasso regression demonstrated better performance in predicting the nine major cell types of fresh frozen breast tissue compared to transfer learning. This discrepancy in performance could be attributed to the prevalence of cancer epithelial cells in the fresh frozen tissue. The lack of diversity in the fresh frozen patches hindered the neural networks from acquiring sufficient information to learn about the distinct morphological differences among various cell types. Consequently, the networks updated their weights based on features that represented common characteristics among the patches, which were not relevant to the variations in cell type proportions. As a result, the predictive capability of the networks was further diminished throughout the training process. These findings imply that while transfer learning can be advantageous for certain tissues, where the neural networks can effectively leverage the learned knowledge, it may not perform as well in tissues with a high abundance of certain cell types. In such cases, Lasso regression, with its ability to capture the major cell types, can yield better predictive outcomes.

Chapter 6

Regression Results for Wu et al. Samples

This chapter focuses on the analysis of ResNet50 transfer learning for the regression task, using samples from Wu et al. [2021]. The input for the neural networks consisted of patches from the samples, while the output was the cell type proportions. Various configurations were explored, including different batch sizes for training, optimizers, learning rates, dropout layer proportions, and numbers of neurons in the hidden dense layer. Throughout the analysis, the training, validation, and testing datasets remained consistent across all networks with varying parameters for the same task and samples. The loss function employed was mean squared error (MSE), with mean absolute error (MAE) used as the metrics function. The networks were trained for a maximum of 1000 epochs, although some networks terminated early with a patience of 30 epochs. For further information on the dataset and network architecture, please refer to Chapter 2.

6.1 All samples

6.1.1 Predicting Four Cell Type Proportions Using Standard Patches

In this section, we analyzed the performance of the neural networks that predicted the four cell type proportions for each standard patch in all six samples retrieved from Wu et al. [2021]. The parameters that were varied included the type of image (original, stained normalized by Macenko, stain normalized by Vahadane), batch size of training (32, 64, 128, 256, 512), learning rate (0.01, 0.001, 0.0001), dropout proportion (0.0, 0.2, 0.5), and dense layer units (256, 512). We employed the Adam optimizer. The number of patches from

each tissue sample in training, validation, and testing datasets is listed in Table 6.1. The cellular compositions of the training, validation, and testing data are displayed in Figure 6.1). Cell type compositions of all patches in the training dataset stratified by samples are shown in Figure 6.2.

	Number of patches		
	Training	Validation	Testing
1142243F	3369	710	702
1160920F	3421	746	728
CID4290	1726	348	358
CID4465	823	193	195
CID44971	801	186	175
CID4535	786	158	183

Table 6.1: Number of standard patches from each of the six samples in the training, validation, and testing datasets.

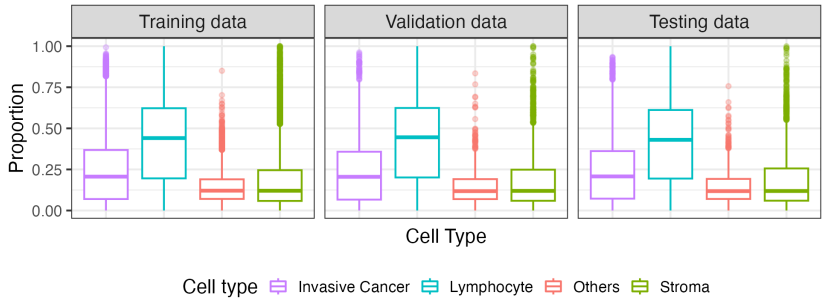


Figure 6.1: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of four cell types using standard patches from all six samples.

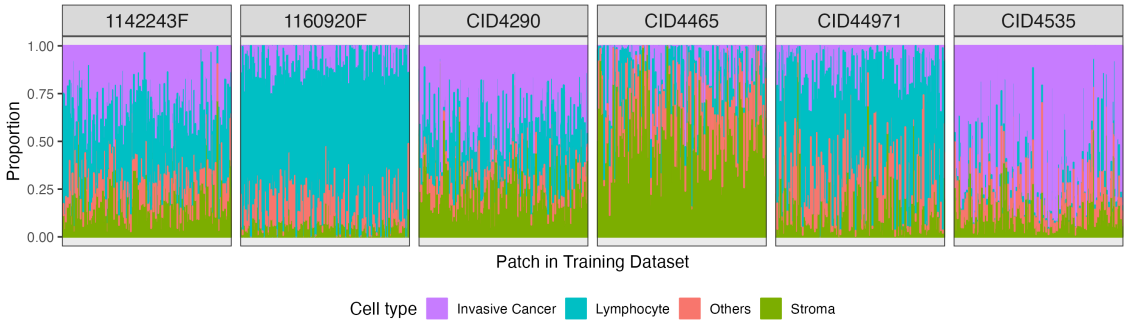


Figure 6.2: Cell type compositions in training dataset for regression task predicting proportions of four cell types using standard patches from all six samples.

Learning Curve

The Adam optimizer did not work well with a learning rate of 0.01 in the sense that all networks with a learning rate of 0.01 terminated early after the initial 30 epochs. For many networks, the validation loss and metrics were generally lower than the training loss and metrics with greater fluctuations. Parameters other than the learning rate did not have a consistent effect on the training process of the networks.

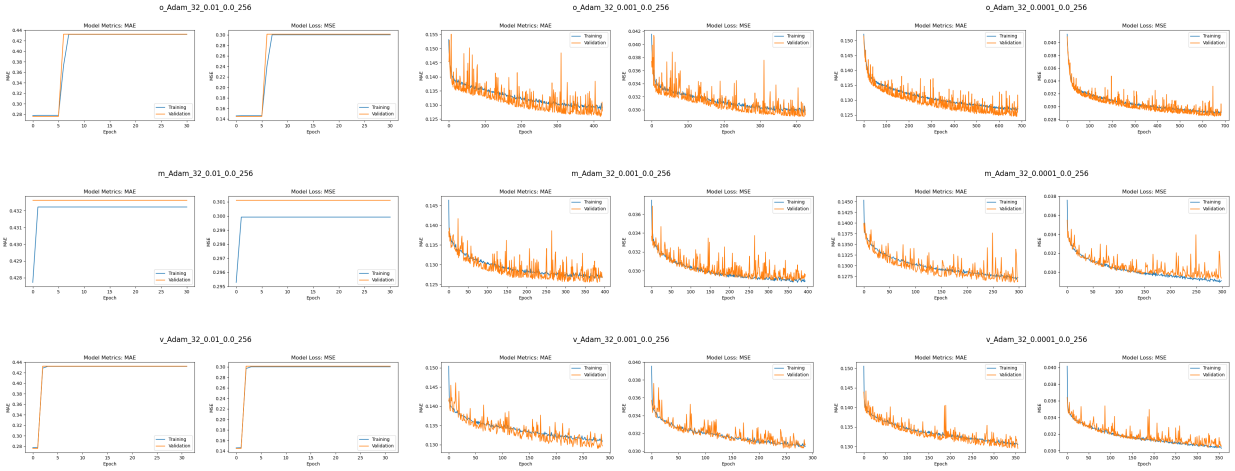


Figure 6.3: Learning curves of neural networks with a batch size of 32, a hidden dense layer of 256 neurons, and no dropout layer aiming to predict the proportions of four cell types using standard patches from all six samples. Each figure title specifies the image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

Evaluation Using Validation Data

Among all combinations of hyperparameters, the neural network that exhibited the lowest validation loss had the following specifications: image type = original, optimizer = Adam, batch size = 32, learning rate = 0.0001, dropout proportion = 0.0, and dense layer units = 512.

First, we compared the validation loss across different learning rates. It was evident that neural networks with a learning rate of 0.01 consistently exhibited significantly higher validation loss at the optimal epoch compared to networks with lower learning rates (Figure 6.4), with only two exceptions. Due to the limited training process for networks with a learning rate of 0.01, we excluded these networks from further analysis. When comparing networks with the same architecture hyperparameters other than the learning rate, it was consistently observed that the network with a learning rate of 0.0001 had lower validation loss than the

corresponding network with a learning rate of 0.001. This trend held true in most cases, with the exception occurring when no dropout layer was incorporated.

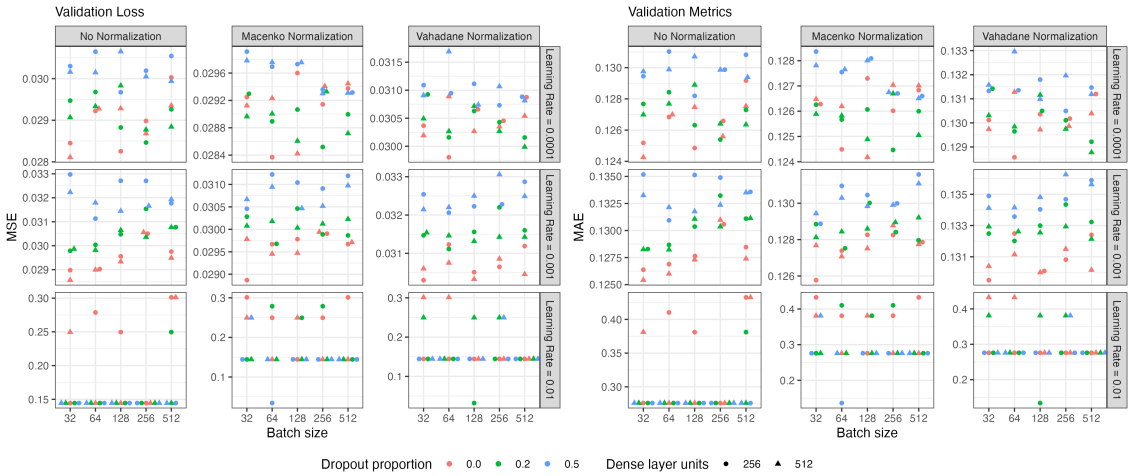


Figure 6.4: Validation loss and metrics from neural networks predicting proportions of four cell types using standard patches from all six samples.

We then examined the impact of stain normalization on the validation loss. Surprisingly, the stain normalization did not result in the expected improvement in validation loss. Generally, when the images were stain normalized using the Vahadane method, the networks' validation loss at the optimal epoch was higher compared to networks with the same architecture trained without stain normalization or with stain normalization using the Macenko method. However, a few exceptions were observed when the learning rate was set to 0.001 and the dropout proportion was set to 0.5. In terms of stain normalization using the Macenko method, we did not observe a consistent increase or decrease in the validation loss when the same network was trained with images before and after the stain normalization.

When comparing networks with different dropout layers but the same other architecture hyperparameters, it was generally observed that most networks with a dropout layer proportion of 0.5 had higher validation loss at the optimal epoch compared to the corresponding networks without a dropout layer or with a dropout layer proportion of 0.2. This trend indicates that incorporating a dropout layer with a higher dropout proportion may have a negative impact on the validation loss of the network.

During our analysis, we did not observe any consistent association between the validation loss at the optimal epoch and either the batch size of training or the number of neurons in the hidden dense layer. How-

ever, it is worth noting that individual cases or specific scenarios still show varying relationships between these factors and the validation loss.

Prediction Using Testing Data

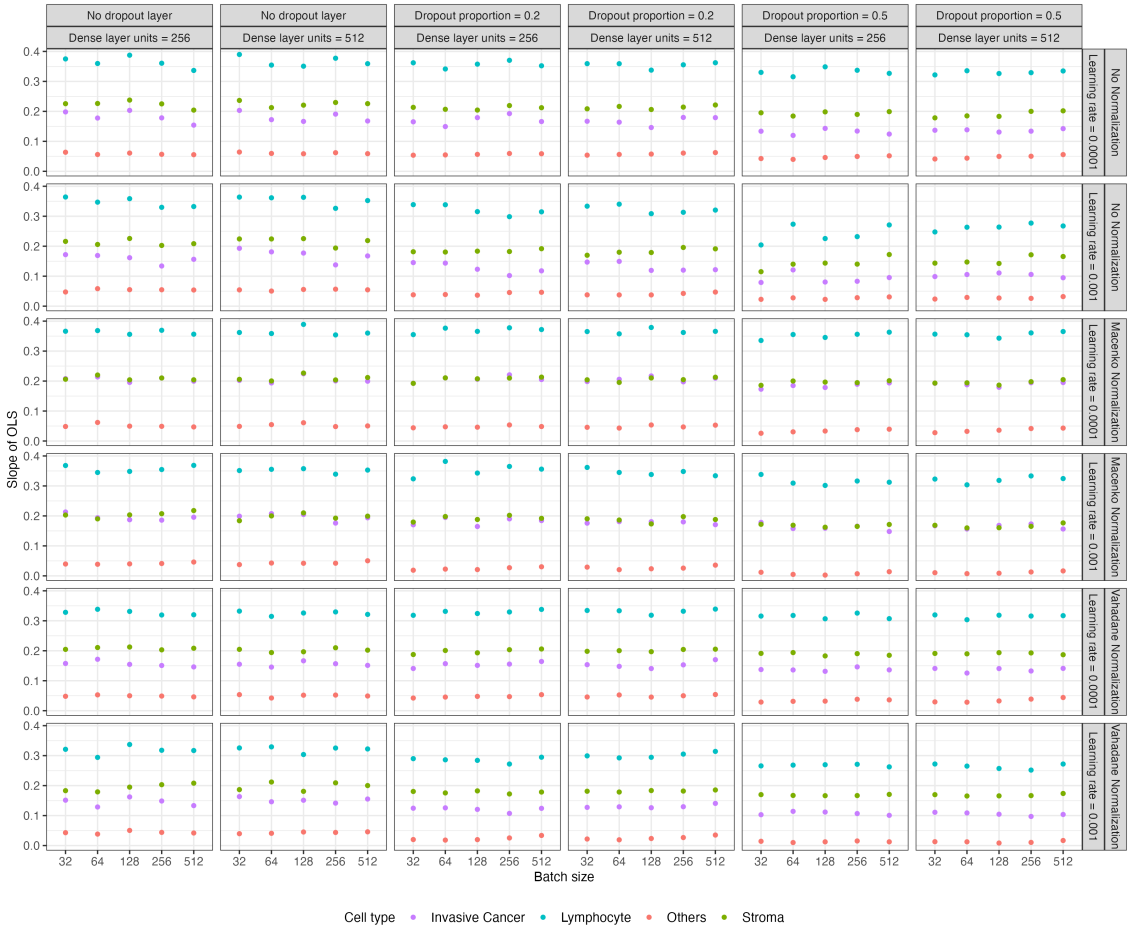


Figure 6.5: Slope of the best-fit line between predicted and observed proportions of four cell types in testing data, predicted by networks trained using standard patches from all six samples, excluding networks with limited learning.

Figure 6.5 depicts the slope of best-fit line between the predicted and observed proportions. The alignment between the predicted and the observed proportions was generally consistent with network's validation loss at the optimal epoch, with lower validation loss associated with higher slope of certain cell types. Among all cell types, lymphocyte consistently exhibited the highest slope, whereas others (normal epithelial and myeloid) consistently exhibited the lowest slope. Besides, there was no single combination of image type

and network architecture that gave the highest slopes of best-fit line of all cell types (Table D.1). Now, our focus shifts to the specific combination of parameters that resulted in the lowest validation loss. Specifically, we examine the network configuration with the following parameters: image type = original, optimizer = Adam, batch size = 32, learning rate = 0.0001, dropout proportion = 0.0, and dense layer units = 512, which gave the highest slopes for lymphocyte and others (normal epithelial and myeloid).

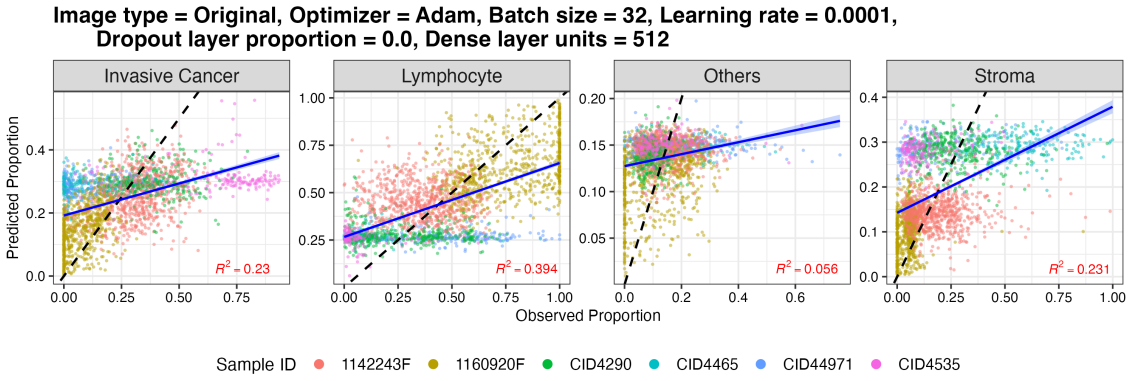


Figure 6.6: Scatterplot comparing predicted and deconvoluted proportions of four cell types in standard patches from all six samples. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. The overall R-squared value is indicated in red.

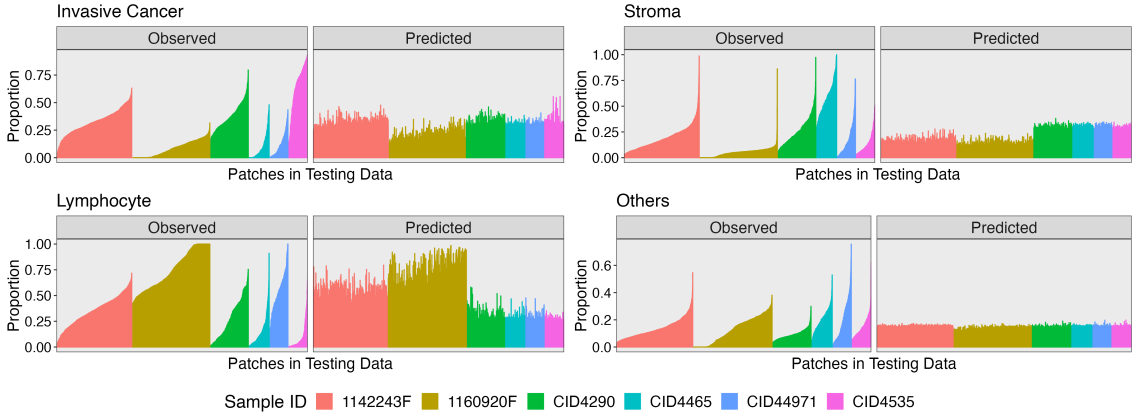


Figure 6.7: Comparison of predicted and observed four cell type compositions in each of the standard patches from all six samples in the testing data.

We observed relatively strong correlations between the predicted and observed proportions for lymphocyte (Figure 6.6). In general, the patches of the two independent lab samples had better correlations between the predicted and the observed proportions than the patches of the other four samples processed by Wu et al (Figure 6.7). The adjusted R-squared values also demonstrate the presence of batch effects resulting from

variations among different samples (Figure D.1). These batch effects are evident in the varying correlations between the predicted and observed proportions of cell types for different samples within the same networks.

6.1.2 Predicting Nine Cell Type Proportions Using Large Patches

The poor performance of networks in sections 6.1.1 and 6.1.2 may be attributed to using small patches for analysis. Small patches might not provide enough context and information for the neural network to accurately differentiate between different cell types. Larger patches are likely to capture more relevant features and context, potentially leading to improved performance. Specifically, one possible morphological difference between cell types could be cell density, which might be difficult to observe and distinguish with small patches, hence negatively affecting the network's ability to distinguish cell types. Moreover, the two samples from the independent lab have better resolution, offering more detailed information that aids in feature extraction. This may contribute to the network's better predictive performance on these samples. Increasing the patch size can help the network capture more features and information, potentially compensating for the resolution variation between the samples since larger patches encompass more context.

Thus, in this section, we analyzed the neural networks that predicted the nine cell type proportions for each large patch in all six samples retrieved from Wu et al. [2021]. The parameters that were varied included the optimizer (Adam, RMSprop, SGD), batch size of the training dataset (32, 64, 128, 256, 512), learning rate (0.01, 0.001, 0.0001), dropout proportion (0.0, 0.2, 0.5), and dense layer units (0, 256, 512). It should be noted that the training, validation, and testing datasets remained consistent across all networks, with 2041, 510, and 465 patches in training, validation, and testing datasets, respectively (Table 6.2). Additionally, all models were trained using original images.

	Number of patches		
	Training	Validation	Testing
1142243F	581	144	169
1160920F	637	172	154
CID4290	292	75	72
CID4465	178	47	27
CID44971	203	39	20
CID4535	150	33	23

Table 6.2: Number of large patches from each of the six samples in the training, validation, and testing datasets.

Each large patch encompassed eight Visium capture spots that had available observed proportions. The observed proportions for each large patch were calculated by averaging the proportions from the eight spots. Notably, there were overlaps in the spots utilized for the large patches in the training and validation datasets, while there were no overlaps between the training and testing datasets. The cellular compositions of the training, validation, and testing data are displayed in Figure 6.8. Cell type compositions of all patches in the training dataset stratified by samples are shown in Figure 6.9.

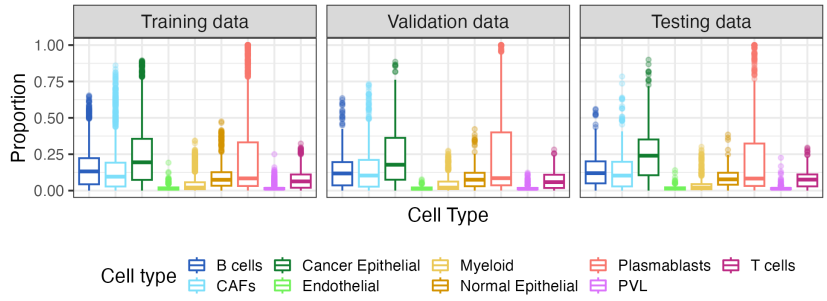


Figure 6.8: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of nine cell types using large patches from all six samples.

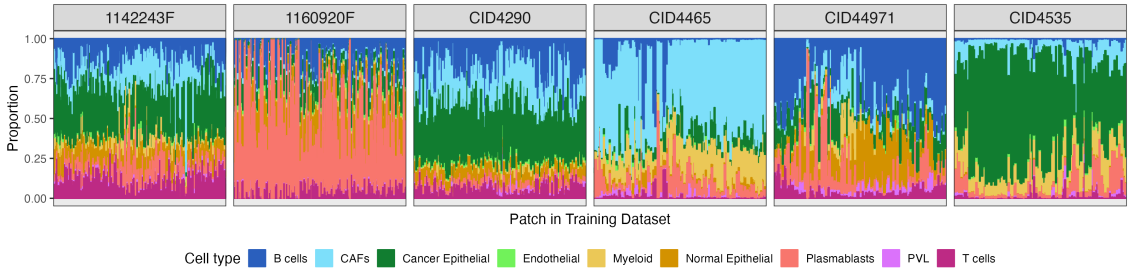


Figure 6.9: Cell type compositions in training dataset for regression task predicting proportions of nine cell types using large patches from all six samples.

Learning Curve

All the networks that utilized the Adam optimizer effectively learned from the training data, except for a few cases with a learning rate of 0.01 where the training process terminated early after 30 epochs. In general, a lower learning rate was found to take a greater number of epochs to converge to a stable validation loss. Some networks consistently demonstrated lower validation loss and metrics compared to their corresponding training loss and metrics. However, there were also instances where certain networks ended up with higher

validation loss and metrics.

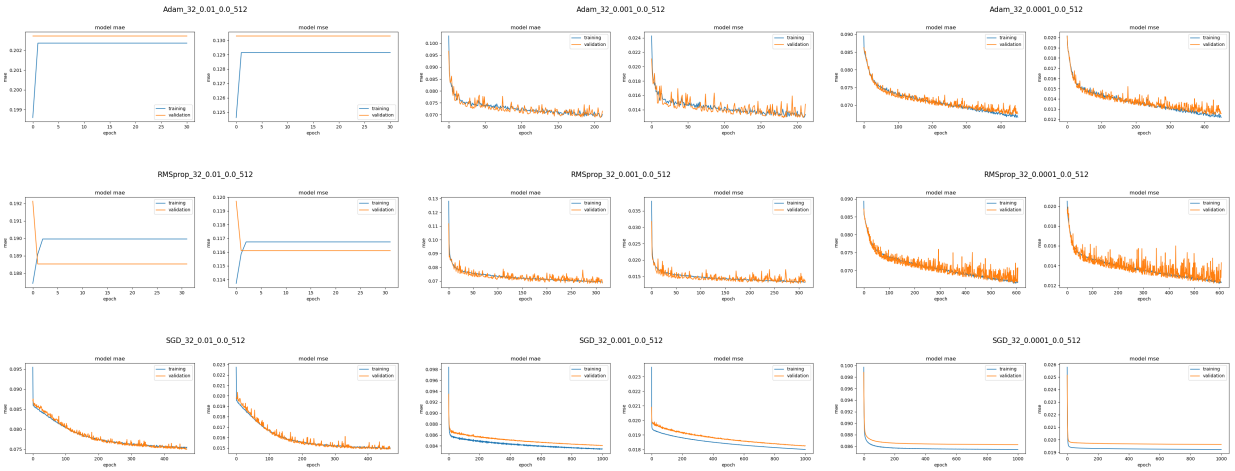


Figure 6.10: Learning curves of neural networks with a batch size of 32, a hidden dense layer of 512 neurons, and no dropout layer aiming to predict the proportions of nine cell types using large patches from all six samples. Each figure title specifies the image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

When the learning rate was set to 0.01, the networks utilizing the RMSprop optimizer faced difficulties in effectively learning from the training data. In fact, many of these networks terminated their training process after the initial 30 to 50 epochs. However, when the learning rates were set to 0.001 and 0.0001, the RMSprop optimizer performed well as the networks were able to decrease and converge to a stable validation loss. Notably, the convergence was slower with a learning rate of 0.0001. Additionally, for the majority of networks, the validation loss and metrics were found to be lower than the corresponding training loss and metrics.

All networks that employed the SGD optimizer exhibited a consistent pattern of the validation loss reducing and eventually converging after a minimum of 200 epochs. However, there were variations among the networks in terms of the relationship between the validation loss and metrics compared to the training loss and metrics. Some networks experienced higher validation loss and metrics in comparison to their corresponding training values, while others showed the opposite pattern with lower validation loss and metrics than the training ones.

Evaluation Using Validation Data

We analyzed and compared the validation loss and metrics at the optimal epoch across various parameter combinations. Among these combinations, the neural network that exhibited the lowest validation loss had the following specifications: optimizer = Adam, batch size = 64, learning rate = 0.0001, dropout proportion = 0.2, and dense layer units = 512.

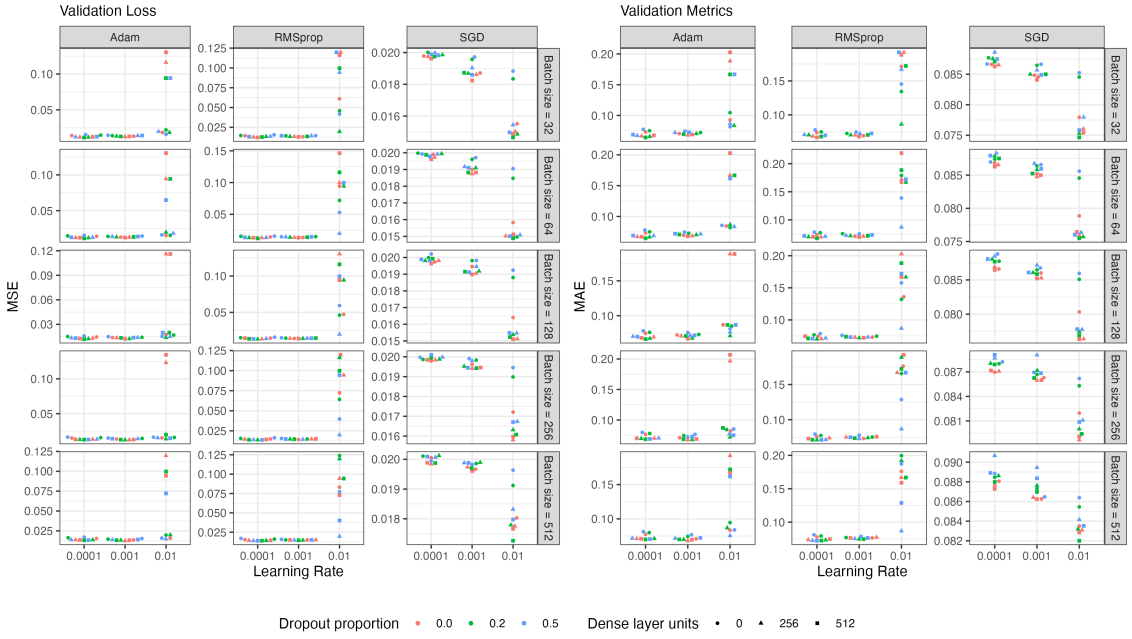


Figure 6.11: Validation loss and metrics from neural networks predicting proportions of nine cell types using large patches from all six samples.

First, we compared the validation loss across different learning rates while considering different optimizers (Figure 6.11). We observed that neural networks utilizing RMSprop as the optimizer consistently exhibited much higher validation loss at the optimal epoch when the learning rate was set to 0.01. This trend was consistent across most cases, except when the dense layer consisted of 256 units and the dropout proportion was set to 0.5. For neural networks with Adam as the optimizer, we discovered that some of these networks demonstrated considerably higher validation loss at the optimal epoch when the learning rate was set to 0.01, especially when a dense layer was present without a dropout layer. Conversely, when we examined neural networks with SGD as the optimizer, we found that the lowest validation loss was either similar to or slightly lower when the learning rate was set to 0.01 compared to learning rates of 0.001 and 0.0001. In

summary, the comparison of validation loss across different learning rates, stratified by optimizers, revealed distinct patterns. Neural networks utilizing RMSprop or Adam as the optimizer generally showed higher validation loss at the optimal epoch with a learning rate of 0.01, whereas networks employing SGD as the optimizer displayed similar lowest validation loss across all three learning rates.

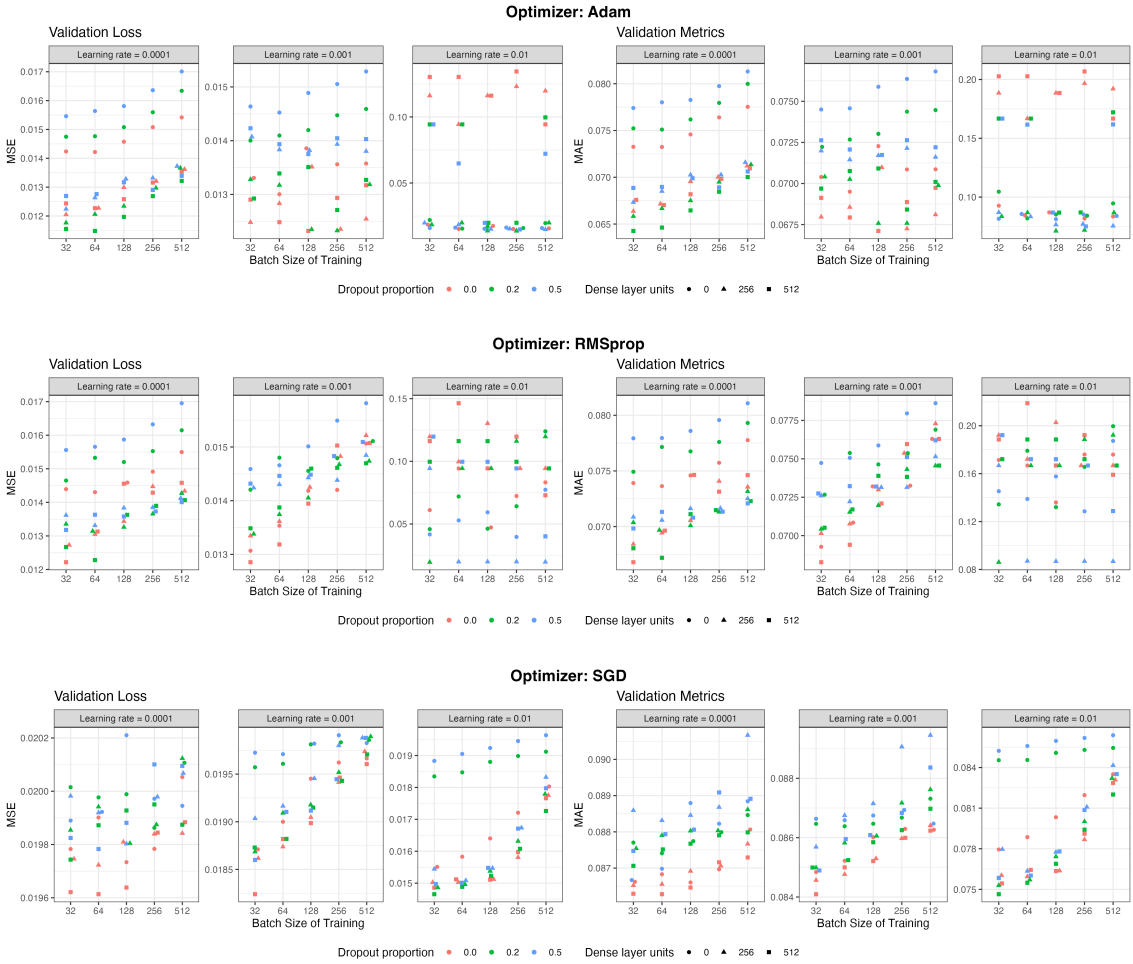


Figure 6.12: Validation loss and metrics from neural networks, stratified by optimizer, for predicting proportions of nine cell types using large patches from all six samples.

We then investigated the impact of batch size on the lowest validation loss for each optimizer individually (Figure 6.12). For networks using the Adam optimizer with a learning rate of 0.0001, we observed a consistent and significant increasing trend in validation loss at the optimal epoch as the batch size increased. This trend was particularly prominent in neural networks without a hidden dense layer. In the other parameter combinations, a general increasing trend in validation loss was observed, with occasional reductions at

certain batch sizes. When the learning rate was set to 0.001, the strict increasing trend in validation loss remained only for networks with a dropout layer of 0.2 proportion but no dense layer. For the remaining parameter combinations, no clear trend in the lowest validation loss with respect to the batch size was identified. The lack of a discernible trend was also observed when the learning rate was set to 0.01. In the case of the RMSprop optimizer, we observed an overall increasing trend in validation loss at the optimal epoch with increasing batch size, not only when the learning rate was 0.0001 but also when it was 0.001. On the contrary, when the optimizer was SGD and the learning rate was set to 0.0001, we did not observe a consistent trend across all combinations of dense layers and dropout layers. However, when the learning rate was set to 0.001 or 0.01, we observed an overall increasing trend in the lowest validation loss for neural networks across all combinations of dense layers and dropout layers.

Furthermore, adding hidden dense layers resulted in a reduction in validation loss at the best epoch for networks that used the Adam optimizer with learning rates of 0.001 and 0.0001, the RMSprop optimizer with a learning rate of 0.0001, and the SGD optimizer with a learning rate of 0.01, while keeping other parameters the same. Within these networks, when no hidden dense layer was present, a greater proportion of dropout layers was associated with higher validation loss at the best epoch.

Predication Using Testing Data

Across all combinations of dense layer and dropout layer, the slope of best-fit line between the predicted and observed proportions of each cell type was predominantly close to zero when the learning rate was set to 0.0001 and 0.001 for the SGD optimizer, as well as 0.01 for the RMSprop optimizer (Figure D.2). Also, the networks employing Adam as the optimizer generally gave worse slopes when the learning rate was set to 0.01 compared to when the learning rate was set to 0.001 and 0.0001 (Figure D.2). So we excluded these four combinations of optimizer and learning rate from our analysis.

We then investigated the relationship between the magnitude of slopes for different cell types (Figure 6.13). Among all cell types, plasmablasts consistently exhibited the highest slope. For networks trained with Adam and RMSprop as the optimizer, the cell types with the second and the third highest slopes of the best-fit lines consisted of cancer epithelial cells, T cells, CAFs, and myeloid cells. However, for networks trained with SGD and a learning rate of 0.01, cancer epithelial cells and CAFs consistently displayed the second

and the third highest slopes, respectively. On the other hand, endothelial cells consistently demonstrated the lowest slope among all cell types, with their slopes approaching nearly zero.

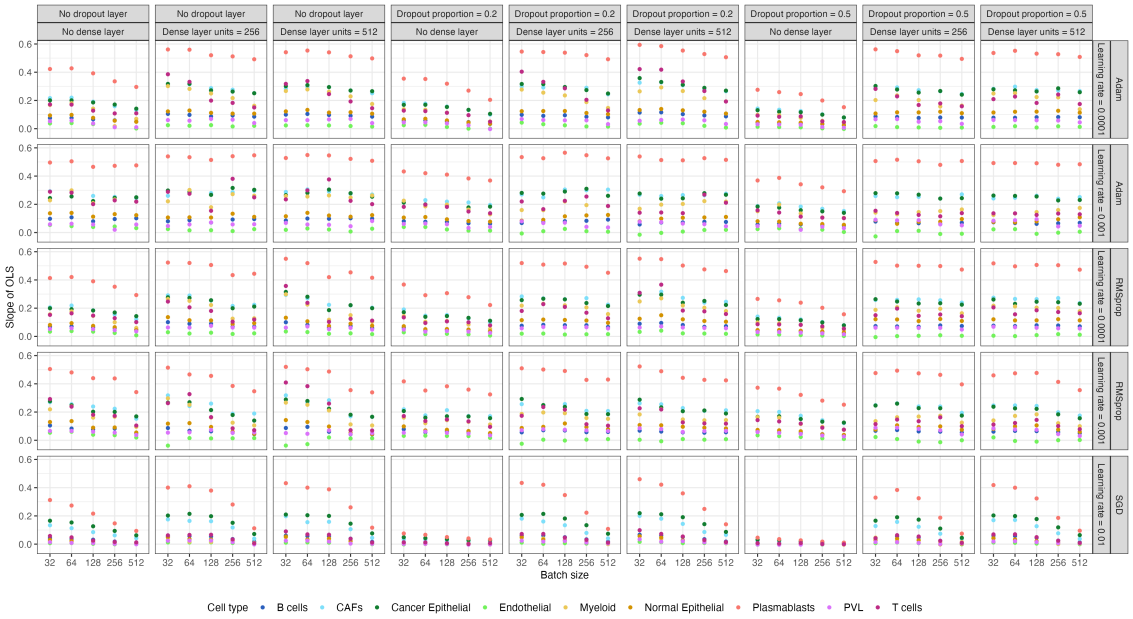


Figure 6.13: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using large patches from all six samples, excluding networks with limited learning.

The network giving the highest slopes of the best-fit lines for the majority of cell types had the following parameters: optimizer = Adam, batch size = 32, learning rate = 0.0001, dropout proportion = 0.2, and dense layer units = 512 (Table D.2). This particular network achieved the second lowest validation loss at the optimal epoch. Comparatively, the network with the lowest validation loss employed the same set of parameters except for a different batch size of 64.

Here, we examined the network that resulted in the lowest validation loss with the following hyperparameters: optimizer = Adam, batch size = 64, learning rate = 0.0001, dropout proportion = 0.2, and dense layer units = 512. We observed relatively strong correlations between the predicted and observed proportions for several cell types, in particular T cells and plasmablasts (Figure 6.14). It is noteworthy that for most cell types, the curves generated by both the OLS and LOESS methods closely aligned with each other. However, certain cell types, for example, endothelial cells and PVLs, exhibited distinct behavior in terms of their LOESS curves. These cell types demonstrate a distinct pattern where the LOESS curves initially

exhibited a more steep slope, nearly diagonal, followed by a transition to a less steep trajectory compared to the OLS line.

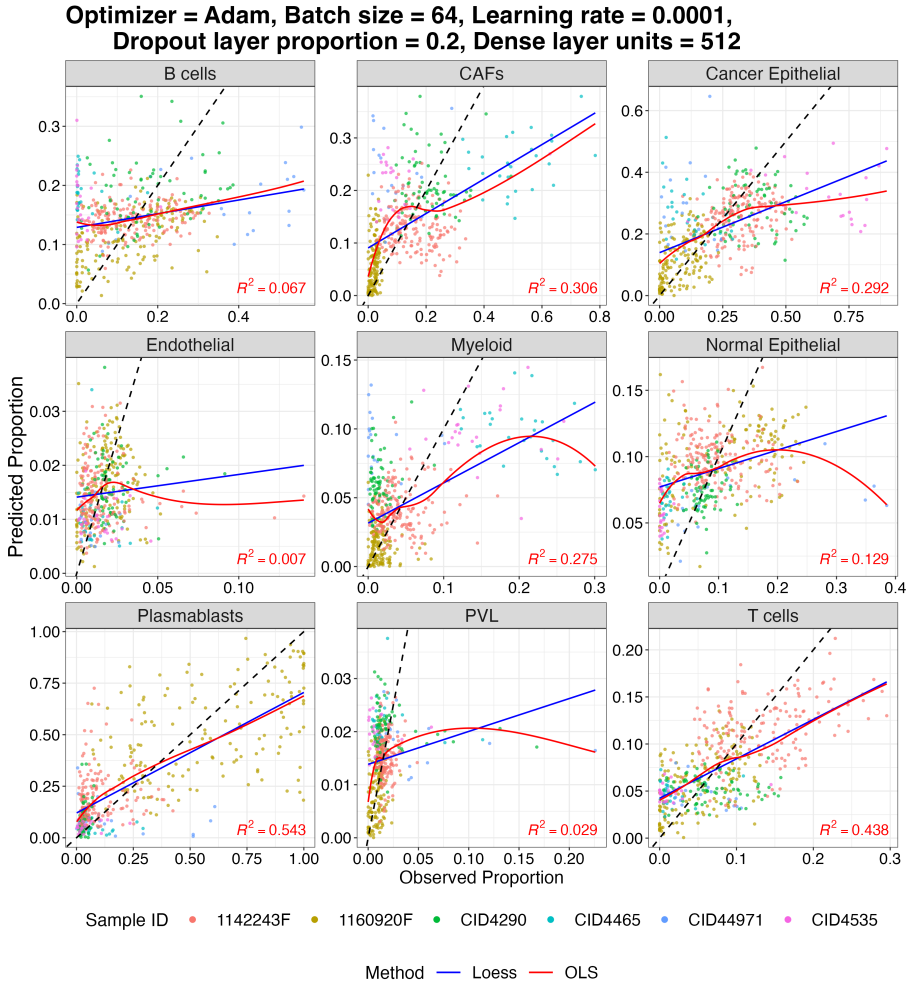


Figure 6.14: Scatterplot comparing predicted and deconvoluted proportions of nine cell types in large patches from all six samples. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. The overall R-squared value is indicated in black.

This distinctive characteristic emphasizes that while the slope of the OLS line appeared to be low for these cell types, the predicted proportions exhibited a strong correlation with the observed proportions when the observed proportions were extremely low (< 0.025). However, as the observed proportions increased to uncommonly high values, the predicted proportions deviated significantly from accuracy. In fact, the trained networks encountered challenges when attempting to accurately predict extreme proportions (Figure D.3). Moreover, for the majority of cell types, the predicted proportions displayed different median values com-

pared to the observed proportions. However, when considering the mean values of predicted and observed proportions, we observed a similarity across all cell types except for plasmablasts.

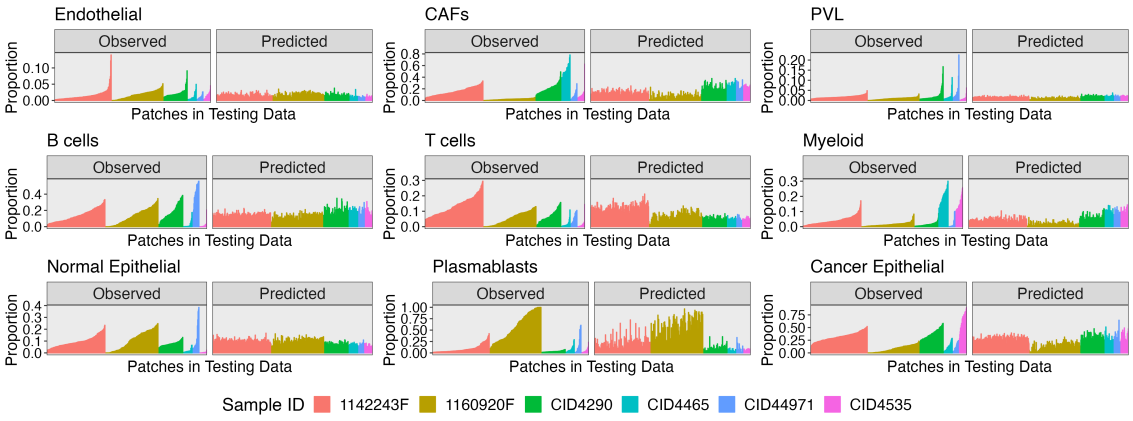


Figure 6.15: Comparison of predicted and observed nine cell type compositions in each of the large patches from all six samples in the testing data.

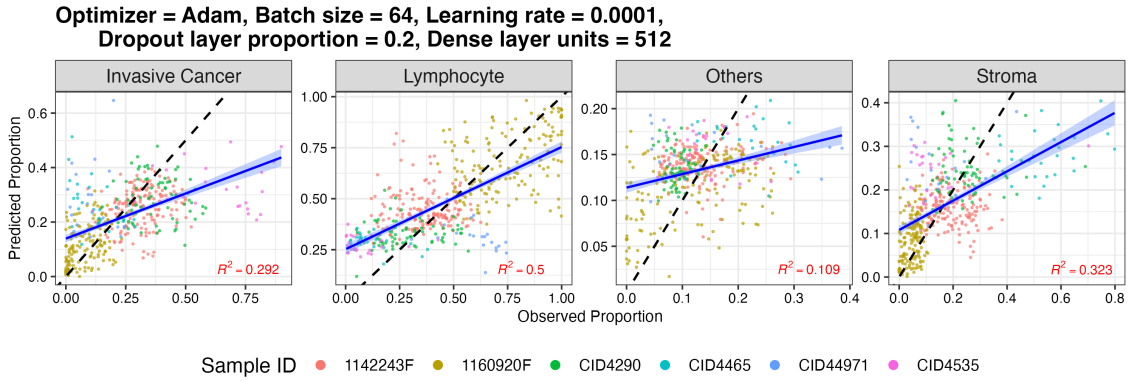


Figure 6.16: Scatterplot comparing four collapsed predicted proportions to four collapsed deconvoluted proportions in the large patches from all six samples in the testing data.

In order to investigate the potential presence of batch effects among different samples, we compared the predicted and observed cell type compositions for each patch within each sample (Figure 6.15). Overall, the neural network's performance varied across samples, with patches from samples 1142243F and 1160920F showing better accuracy compared to the other samples. Particularly, the batch effects are evident in the varying correlations between the predicted and observed proportions of cell types for different samples within the same networks, as indicated by the varying sample-specific R-squared values in the testing data (Figure D.4). This observation suggests that the use of large patches in training the neural network might not

have mitigated the batch effects. However, this observation could also potentially be attributed to the smaller sample sizes of the other four samples processed by Wu et al. in the training dataset. These smaller sample sizes have limited the network's ability to capture the full range of intrinsic differences in these specific samples, thus leading to comparatively lower prediction accuracy. We also investigated the correlation between the predicted and the observed proportions of the four collapsed pathological cell types (Figure 6.16).

6.1.3 Predicting Four Cell Type Proportions Using Large Patches

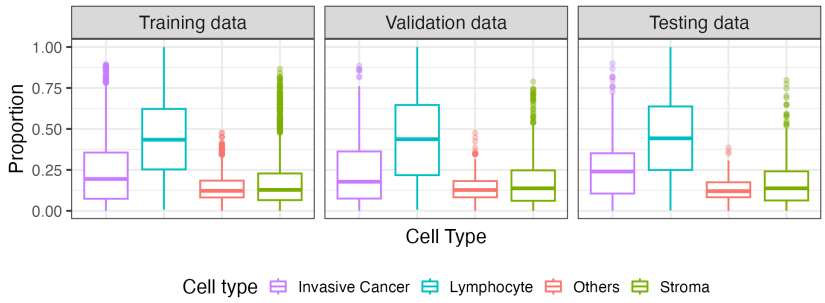


Figure 6.17: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of four cell types using large patches from all six samples.

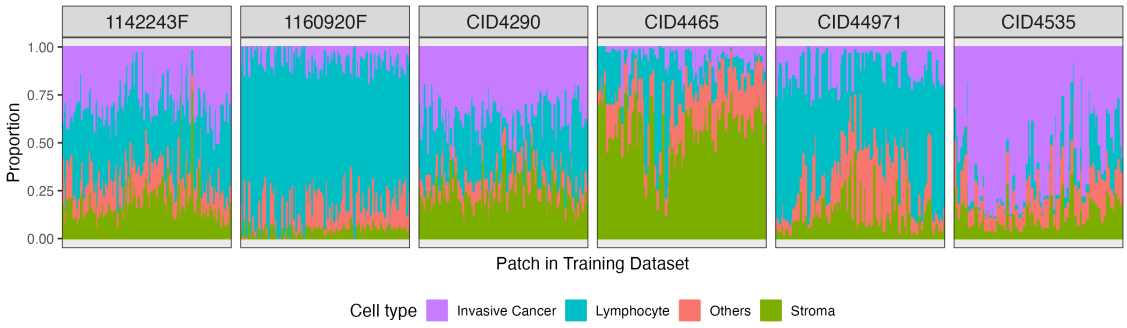


Figure 6.18: Cell type compositions in training dataset for regression task predicting proportions of four cell types using large patches from all six samples.

Instead of utilizing nine cell types as response variables, we opted to use four collapsed cell types for constructing the networks. Following the analysis in section 6.1.2, we maintained the same varying parameters in the network architecture, except for employing the RMSprop optimizer and excluding hidden dense layers. The patches in the training, validation, and testing datasets remained unchanged from the previous

regression task described in section 6.1.2. The training, validation, and testing datasets showed a predominance of lymphocytes (Figure 6.17). Specifically, within the training dataset, patches from 1160920F had a predominance of lymphocytes, patches from CID4465 had a predominance of stroma, and patches from CID4535 had a predominance of invasive cancer cells (Figure 6.18).

Learning Curve

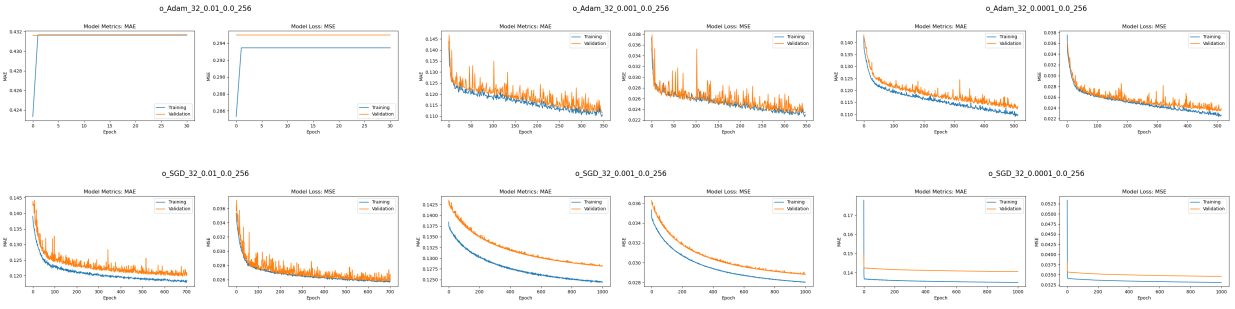


Figure 6.19: Learning curves of neural networks with a batch size of 32, a hidden dense layer of 256 neurons, and no dropout layer aiming to predict the proportions of four cell types using large patches from all six samples.

Evaluation Using Validation data

Among all combinations of parameters, the neural network that exhibited the lowest validation loss had the following specifications: optimizer = Adam, batch size = 128, learning rate = 0.0001, dropout proportion = 0.2, and dense layer units = 512. Since the networks utilizing the Adam optimizer with a learning rate of 0.01 had very high validation losses at the optimal epochs than all the other networks (Figure 6.20), we excluded these networks from our analysis and discussion.

The Adam optimizer consistently resulted in smaller validation loss compared to the SGD optimizer at the optimal epochs for networks with learning rates of 0.001 and 0.0001, while keeping other parameters constant. Additionally, among networks using the SGD optimizer, a higher learning rate was associated with lower validation loss at the optimal epochs, after controlling for other parameters. On the other hand, among networks using the Adam optimizer, a learning rate of 0.001 yielded higher validation loss at the optimal epochs than a learning rate of 0.0001, except when there was no dropout layer present.

Moreover, we examined the effect of batch size on the validation loss after controlling for other pa-

rameters. For networks utilizing the Adam optimizer, we did not observe a strict and consistent trend of validation loss in relation to the batch size. However, for networks using the SGD optimizer, an increasing batch size was consistently linked to higher validation loss at the optimal epochs when the learning rate was set to 0.001, for all combinations of dropout layer and hidden dense layer. A similar trend was observed when the learning rate was set to 0.0001, with some exceptions for specific batch sizes across the majority of dropout and hidden dense layer combinations. We did not find this trend for networks using the SGD optimizer with a learning rate of 0.01.

Finally, there was no consistent association between the lowest validation loss and the dropout layer proportion or the number of neurons in the hidden dense layer.

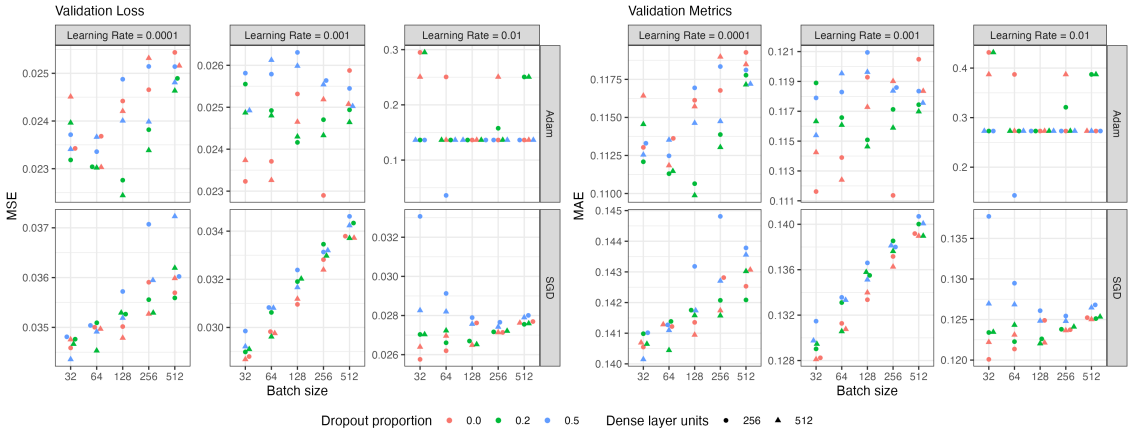


Figure 6.20: Validation loss and metrics from neural networks predicting proportions of four cell types using large patches from all six samples.

Prediction Using Testing data

The slope of best-fit line between the observed proportion and the predicted proportion, obtained from networks using the Adam optimizer with a learning rate of 0.01 and networks using the SGD optimizer with a learning rate of 0.0001, was almost zero for all cell types (Figure D.5). Among all cell types, lymphocytes consistently exhibited the highest slope of the best-fit line, whereas the combination of normal epithelial and myeloid cells consistently displayed the lowest slope of the best-fit line (Figure 6.21). Besides, while all the cell types achieved the highest slope when the networks utilized the Adam optimizer with a learning rate of 0.0001, there was no specific batch size or combination of dropout and hidden dense layers that produced the highest slope for more than one cell type simultaneously (Table D.3).

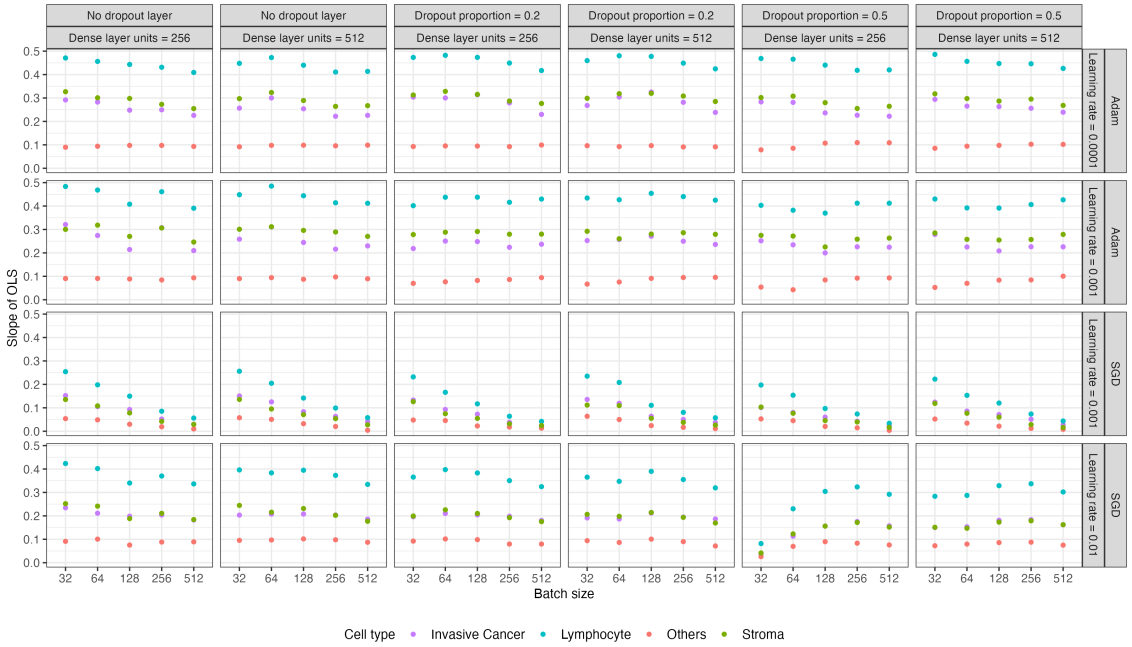


Figure 6.21: Slope of the best-fit line between predicted and observed proportions of four cell types in testing data, predicted by networks trained using large patches from all six samples, excluding networks with limited learning.

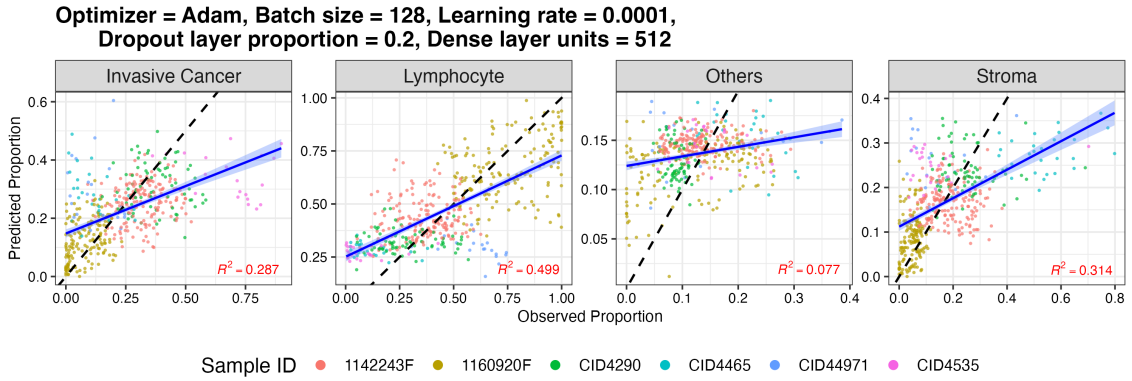


Figure 6.22: Scatterplot comparing predicted and deconvoluted proportions of four cell types in large patches from all six samples. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. The overall R-squared value is indicated in black.

Here, we focused on the predicted proportions obtained from the networks that yielded the lowest validation loss at the optimal epoch, with the following parameter settings: optimizer = Adam, batch size = 128, learning rate = 0.0001, dropout proportion = 0.2, and dense layer units = 512. Specifically, for invasive cancer, the proportions in samples 1160920F, CID4465, and CID44971 were inclined to be overestimated,

while samples 1142243F, CID4290, and CID4535 tended to be underestimated (Figure 6.22). In terms of lymphocyte proportions, all samples except for 1160920F and CID44971 tended to be overestimated. For the stroma proportions, samples 1160920F, CID4535, and CID44971 tended to be overestimated, whereas sample CID4465 tended to be underestimated.

Lastly, we conducted a comparison between the four predicted proportions obtained by collapsing the nine predicted cell type proportions from section 6.1.2, and the four predicted proportions directly obtained by the network using these four cell types as the response variable (Figure 6.23). We can see that the predicted proportions obtained by the two approaches were highly correlated for invasive cancer, lymphocyte, and stroma. However, the combination of normal epithelial and myeloid displayed a relatively weaker correlation, with the correlation between the observed and predicted proportions being smaller in comparison to when the predicted proportions were obtained by aggregating the nine predicted cell type proportions.

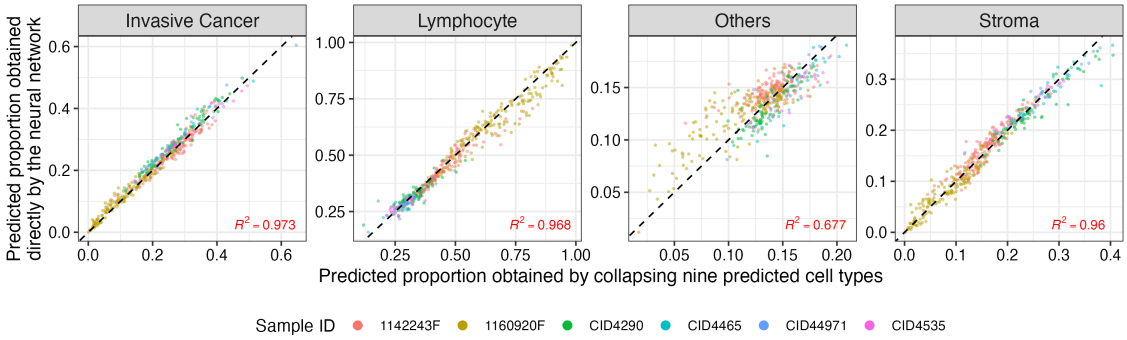


Figure 6.23: Scatterplot comparing predicted proportions achieved by two approaches for large patches of all six samples derived from identical testing data. The proportions were estimated through two methods: direct prediction by a network with four response variables, and collapsing the proportions of nine major cell types predicted by a network with nine response variables.

6.2 Two Samples Provided by An Independent Lab

6.2.1 Predicting Nine Cell Type Proportions

The neural networks predicted the nine cell type proportions for each standard patch in two samples from an independent laboratory. The varying parameters are the batch size of training (32, 64, 128, 256), learning rate (0.001, 0.0001), dropout layer proportion (0.0, 0.2, 0.5), and number of neurons of the dense layer (256, 512). We used Adam as the optimizer. The training, validation, and testing datasets remained consistent

across all network evaluations. There are in total 9676 patches, with 6773 for training, 1452 for validation, and 1451 for testing. The number of patches from each sample is shown in Table 6.3. The cellular compositions of the training, validation, and testing data are displayed in Figure 6.24. Specifically, sample 1160920F exhibits a predominance of plasmablasts in the training dataset (Figure 6.25).

	Number of patches		
	Training	Validation	Testing
1142243F	3358	730	693
1160920F	3415	722	758

Table 6.3: Number of patches from each of the two independent lab samples in the training, validation, and testing datasets.

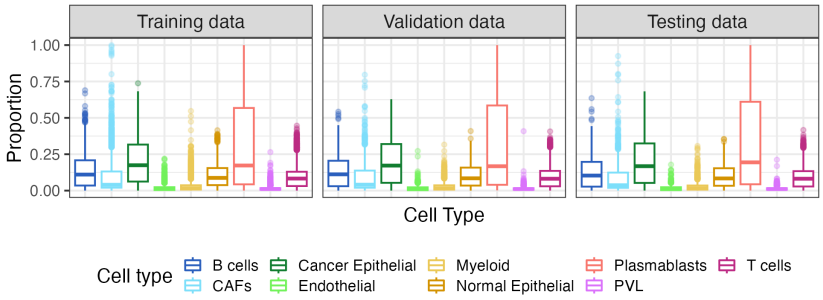


Figure 6.24: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of nine cell types using standard patches from two independent lab samples.

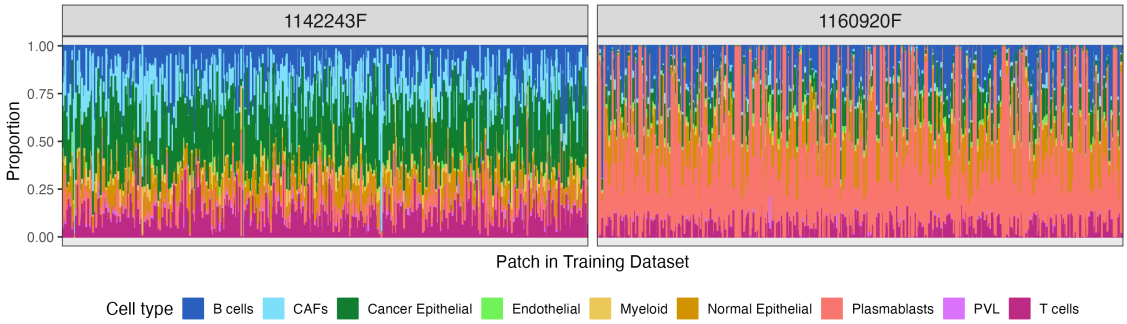


Figure 6.25: Cell type compositions in training dataset for regression task predicting proportions of nine cell types using standard patches from two independent lab samples.

Learning Curves

It can be observed from the learning curves that certain networks underwent training for 1000 epochs, while others terminated early, lasting at least 200 epochs. Typically, networks utilizing a learning rate of 0.0001 required a greater number of epochs compared to those with a learning rate of 0.001 to attain a state where the validation loss did not reach a lower value for 30 consecutive epochs. Moreover, many networks exhibited validation loss and metrics that were lower than their corresponding training loss and metrics. The validation loss and metrics displayed a higher degree of variability and fluctuation compared to the training loss and metrics. The higher fluctuation in validation loss and metrics implies that the networks might not be able to generalize well to unseen data.

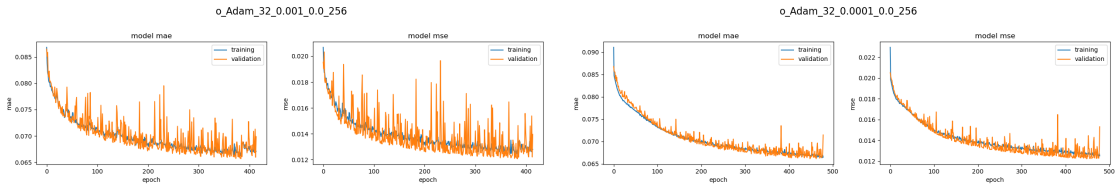


Figure 6.26: Learning curves of neural networks with a batch size of 32, a hidden dense layer with 256 neurons, and no dropout layer aiming to predict the proportions of nine cell types using standard patches from two independent lab samples.

Evaluation Using Validation Data

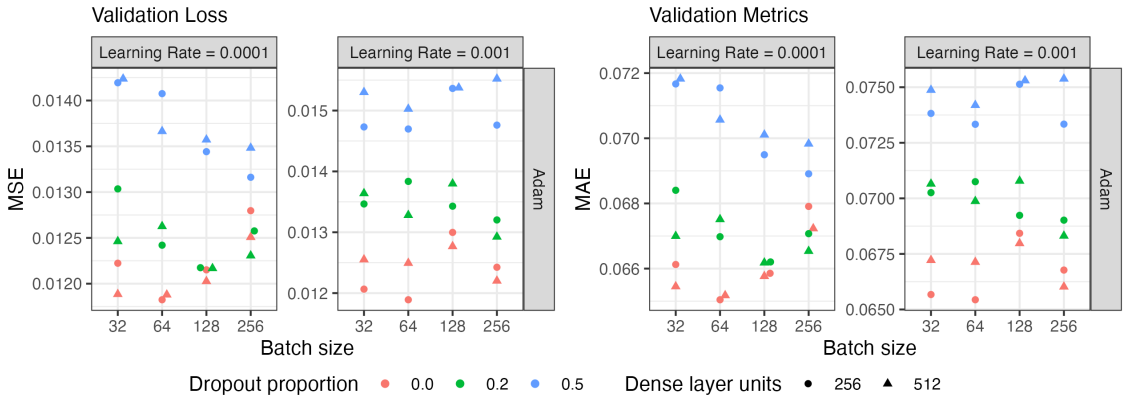


Figure 6.27: Validation loss and metrics from neural networks predicting proportions of nine cell types using standard patches from two independent lab samples.

Across all neural networks, the validation metrics and loss were consistent. Among all combinations of hyperparameters, the neural network that had the lowest validation loss had the following specifications: batch size = 64, learning rate = 0.0001, dropout proportion = 0.0, and dense layer units = 256.

Networks with a learning rate of 0.0001 consistently achieved lower validation loss at the optimal epoch compared to networks with a learning rate of 0.001 (Figure 6.27). However, there were exceptions where a learning rate of 0.001 achieved lower validation loss. These exceptions occurred when there was no dropout layer, the dense layer had 256 units, and the batch size was either 32 or 256.

Additionally, there was no consistent trend observed across all networks in the validation loss with respect to the number of units in the dense layer when controlling for other parameters. However, for specific networks with a learning rate of 0.0001, those with a dropout proportion of 0.5 consistently showed an opposite trend in validation loss compared to networks with dropout proportions of 0.0 and 0.2. For instance, among networks with a learning rate of 0.0001 and a batch size of 32, networks without a dropout layer or with a dropout proportion of 0.2 had lower validation losses at the optimal epoch when the dense layer had 512 units. However, the network with a dropout proportion of 0.5 had a lower validation loss when the dense layer had 256 units. Besides, among networks with a learning rate of 0.001 and a dropout proportion of 0.5, networks with dense layer units of 256 consistently achieved lower validation loss at the optimal epoch. Finally, increasing the dropout layer proportion generally resulted in higher validation loss at the optimal epoch, except when the batch size was 256 and the learning rate was 0.0001.

Predication Using Testing Data

The slopes of best-fit line between the predicted and observed proportions are depicted in Figure 6.28. Among all cell types, plasmablasts consistently exhibited the highest slope, followed by cancer epithelial cell and T cell. Endothelial cells, on the other hand, consistently exhibited the lowest slope. The highest slopes of plasmablasts, cancer epithelial cell, and T cell were all achieved by batch size = 64, learning rate = 0.0001, dropout proportion = 0.0, dense layer units = 256 (Table D.5), which is also the combination of parameters that gave the lowest validation loss. We now focus on this combination of parameters. We noticed that the predicted proportions for some cell types have clear upper bounds, for example, B cells, normal epithelial cells, and CAFs (Figure 6.29).

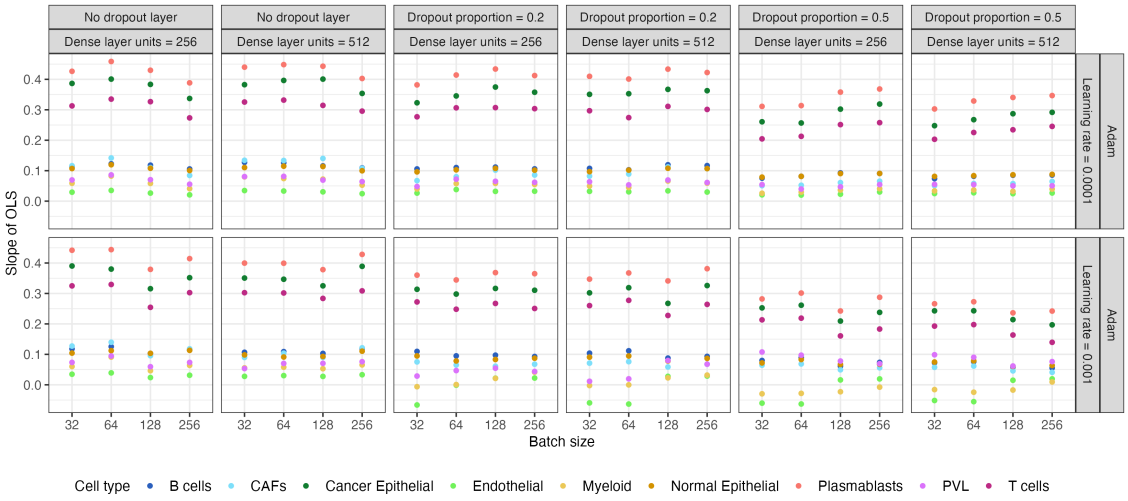


Figure 6.28: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using large patches from two independent lab samples.

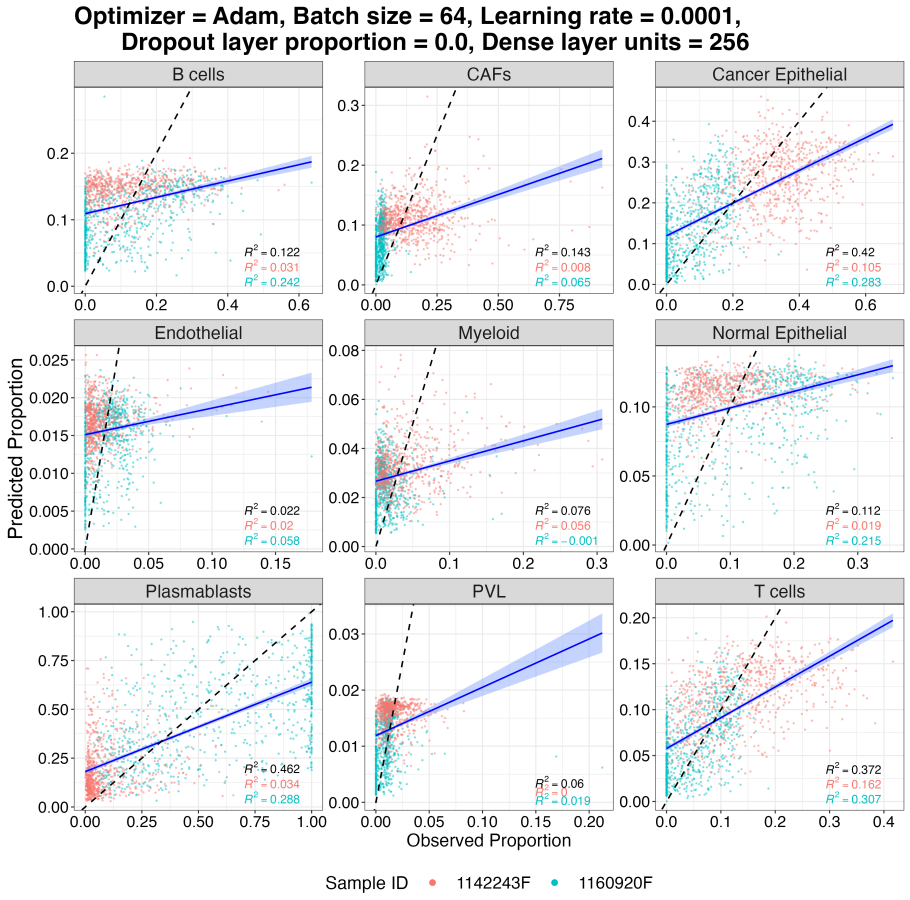


Figure 6.29: Scatterplot comparing predicted and deconvoluted proportions of nine cell types in standard patches from two independent lab samples. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. R-squared values are indicated in black for the overall fit and in corresponding colors for each sample.

Cancer epithelial cells, plasmablasts, and T cells exhibit stronger correlations between the predicted and observed proportions when compared to other cell types. Besides, the predicted proportions exhibited a higher level of concentration compared to the observed proportions. For the majority of cell types, the predicted proportions showed distinct median values in comparison to the observed proportions, while maintaining the same mean values (Figure D.7).

Additionally, when assessing the sample-specific adjusted R-squared values (Figure D.8), it becomes evident that sample 1160920F demonstrates superior correlations between the predicted and observed proportions across all cell types, excluding myeloid cells, in comparison to sample 1142242F. In fact, sample 1160920F consistently displayed stronger correlations between the predicted and observed proportions for all cell types except myeloid, CAFs, and PVL, across all networks. This observation suggests the presence of batch effects introduced by the different samples. Interestingly, no single neural network was capable of achieving the highest correlation for the predicted and observed proportions of any cell type simultaneously in both samples (Table D.4).

Furthermore, it was observed that cell types with higher abundance in the training and testing data demonstrated better predictive performance, while the predictive performance for other cell types was hindered by their extremely small proportions in the sampled tissue (Figure 6.30).

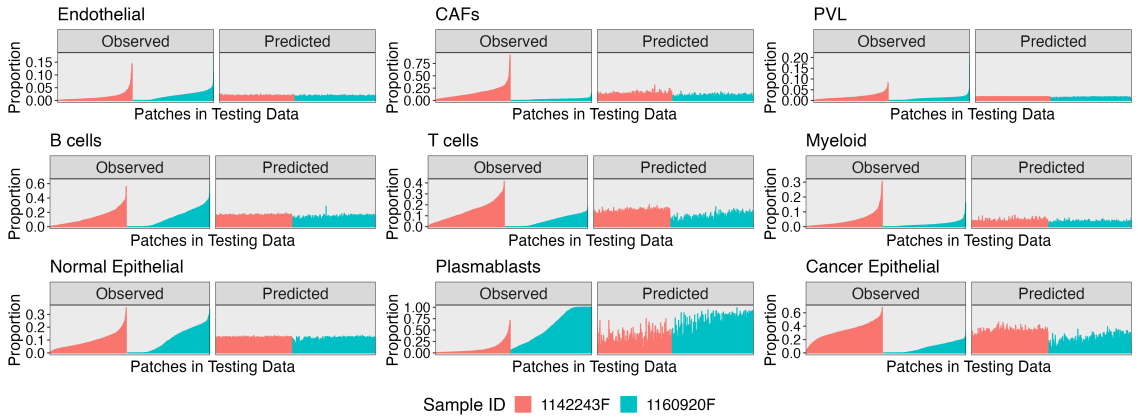


Figure 6.30: Comparison of predicted and observed nine cell type compositions in each of the standard patches from two independent lab samples.

Finally, we investigated whether there are stronger correlations between the predicted and the observed proportions of the four collapsed pathological cell types than those of the nine major cell types. It is evident

that there was no noticeable improvement in the correlation between the predicted proportions and the observed proportions when collapsing the nine major cell types into four pathological cell types (Figure 6.31).

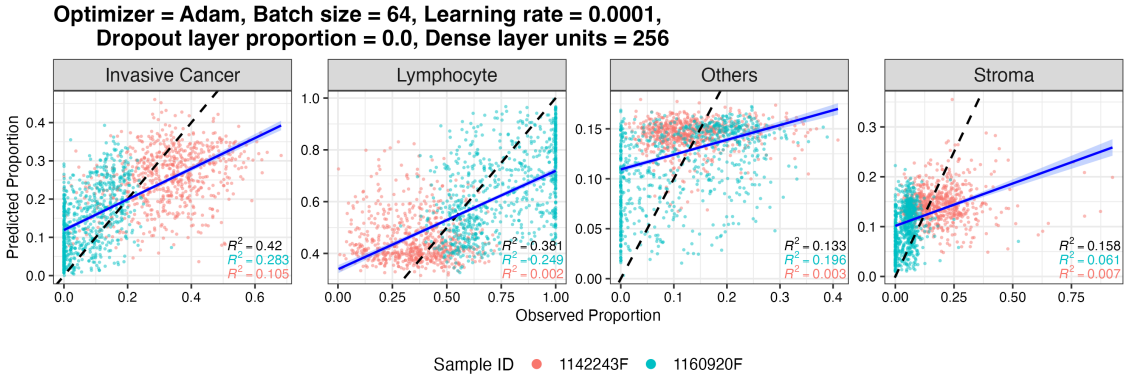


Figure 6.31: Scatterplot comparing four collapsed predicted proportions to four collapsed deconvoluted proportions in the standard patches from two independent lab samples in the testing data.

6.2.2 Predicting Four Cell Type Proportions

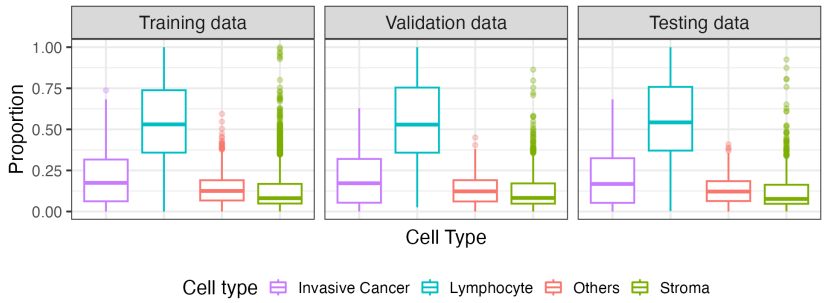


Figure 6.32: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of four cell types using standard patches from two independent lab samples.

Instead of training the neural networks to predict the nine individual cell types, our task shifted towards constructing neural networks designed to directly predict the four pathological cell types. We kept the parameters in the network architecture and the composition of the training, validation, and testing datasets unchanged from the previous regression task detailed in section 6.2.1. Notably, the cellular composition of the training, validation, and testing datasets exhibited a predominance of lymphocytes (Figure 6.32). The predominance of lymphocytes in the training dataset was primarily caused by its predominance in sample

1160920F (Figure 6.33).

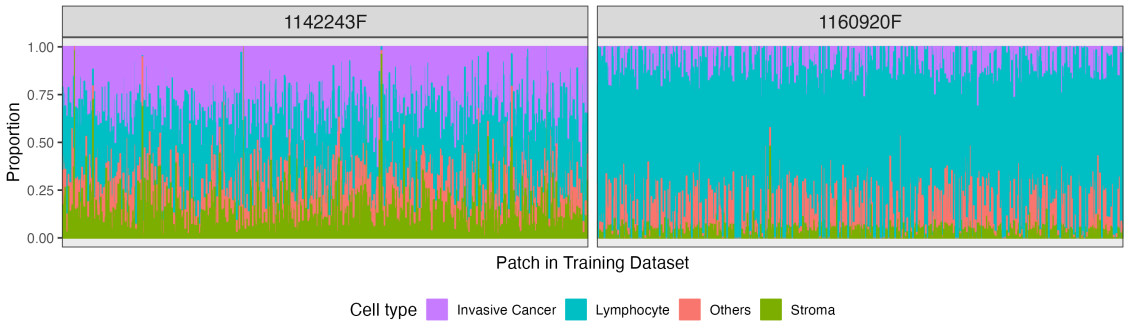


Figure 6.33: Cell type compositions in training dataset for regression task predicting proportions of four cell types using standard patches from two independent lab samples.

Learning Curve

The learning curves demonstrate the successful learning of all networks from the provided training data. Generally, networks with a learning rate of 0.0001 required a greater number of epochs to attain a stable validation loss compared to networks with a learning rate of 0.001 and identical other parameters. Nonetheless, for all networks, the validation loss and metrics were consistently lower than the corresponding training loss and metrics.

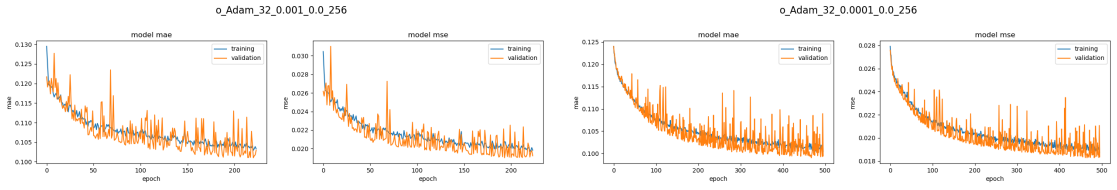


Figure 6.34: Learning curves of neural networks with a batch size of 32, a hidden dense layer with 256 neurons, and no dropout layer aiming to predict the proportions of four cell types using standard patches from two independent lab samples.

Evaluation Using Validation Data

Across all neural networks, the validation metrics and loss were consistent (Figure 6.35). Among these combinations, the neural network that had the lowest validation loss had the following specifications: batch size = 128, learning rate = 0.0001, dropout proportion = 0.0, and dense layer units = 512.

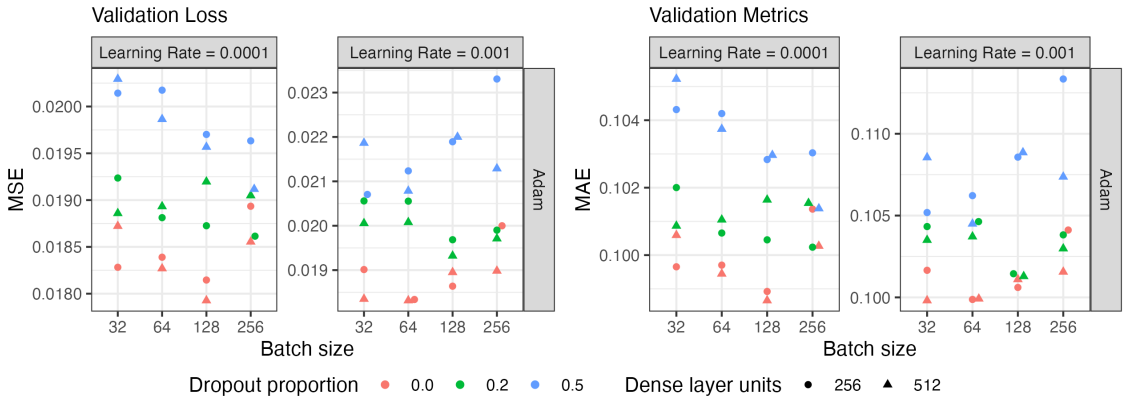


Figure 6.35: Validation loss and metrics from neural networks predicting proportions of four cell types using standard patches from two independent lab samples.

Except for two specific cases, where the networks had a batch size of 32 and 512 neurons in the hidden dense layer, and where the networks had a batch size of 64 and 256 neurons in the hidden dense layer (and in both cases, no dropout layer was present), networks with a learning rate of 0.001 generally exhibited higher validation loss at the optimal epoch compared to networks with the same parameters but a different learning rate of 0.0001.

Additionally, we investigated the impact of the dropout layer proportion on the validation loss while keeping other parameters constant. We observed a consistent relationship between increasing dropout layer proportion and higher validation loss at the optimal epochs, except when the batch size was 256 and the hidden dense layer consisted of 256 neurons.

Lastly, we found no consistent correlation between the number of neurons in the hidden dense layer and the lowest validation loss. Similarly, we did not observe a consistent association between the batch size and the validation loss when considering all combinations of the learning rate, dropout layer proportion, and units in the hidden dense layer.

Prediction Using Testing Data

In general, the predicted proportions of all cell types tended to align more closely with the observed proportions when the predictions were generated by networks that did not include a dropout layer, as opposed to networks that included a dropout layer (Figure 6.36). Besides, based on the slope of the best-fit line, the predicted proportions of invasive cancer cells and lymphocytes exhibited a significantly better alignment with

their observed proportions than the predicted proportions of stroma and other cell types. It is worth mentioning that the specific network, which produced the highest slope of the best-fit line for invasive cancer, lymphocyte, and stroma, also yielded the lowest validation loss among all the networks (Table D.6). Consequently, we concentrated on examining this particular network in more detail. The histogram comparing predicted and observed proportions for each of the four cell types is shown in Figure D.9.

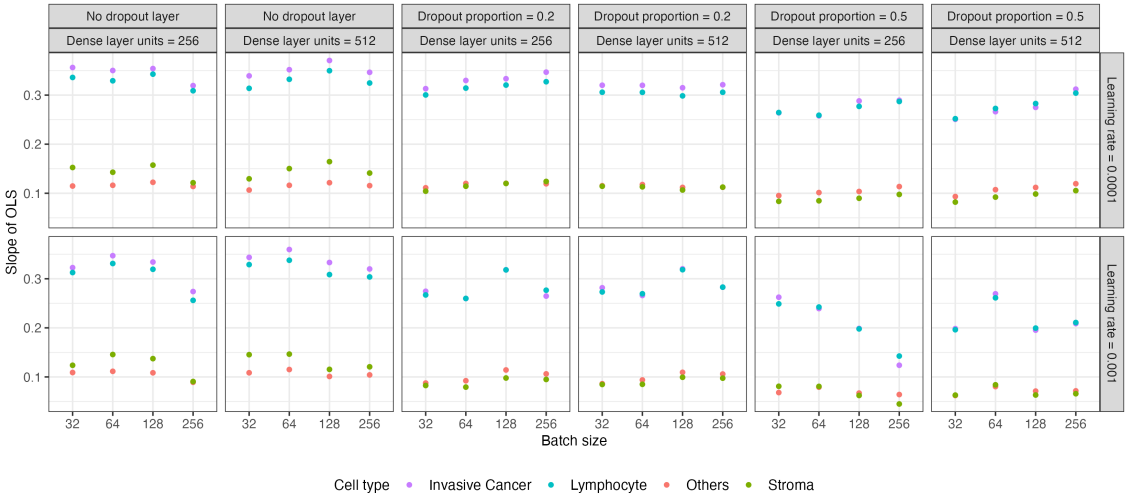


Figure 6.36: Slope of the best-fit line between predicted and observed proportions of four cell types in testing data, predicted by networks trained using standard patches from two independent lab samples.

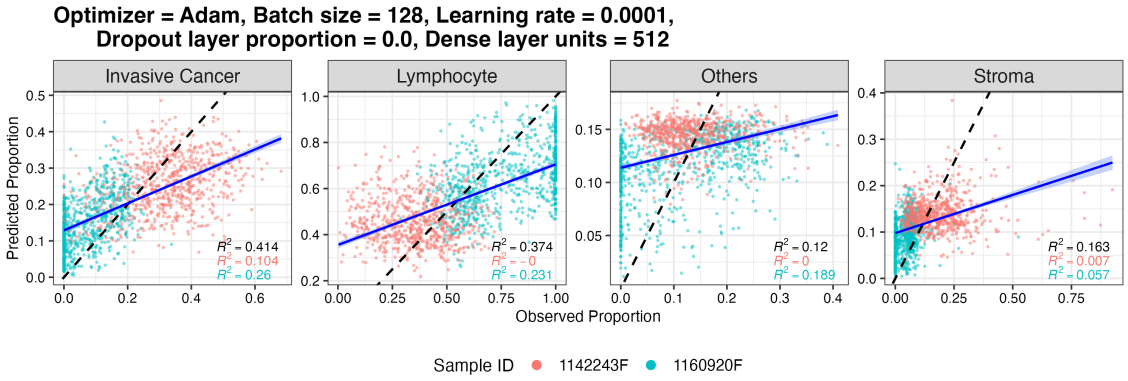


Figure 6.37: Scatterplot comparing predicted and deconvoluted proportions of four cell types in standard patches from two independent lab samples. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. R-squared values are indicated in black for the overall fit and in corresponding colors for each sample.

We observed that the correlations between the predicted and observed proportions varied across in-

dividual samples and were generally lower than the overall correlation (Figure 6.37). Sample 1160920F exhibited a stronger correlation than sample 1142243F. This batch effect was less pronounced in the correlations between the predicted and observed proportions of stroma (Figure 6.38). In terms of the specific proportions, sample 1160920F tended to overestimate the proportion of invasive cancer cells and stroma, while underestimating the proportion of lymphocytes.

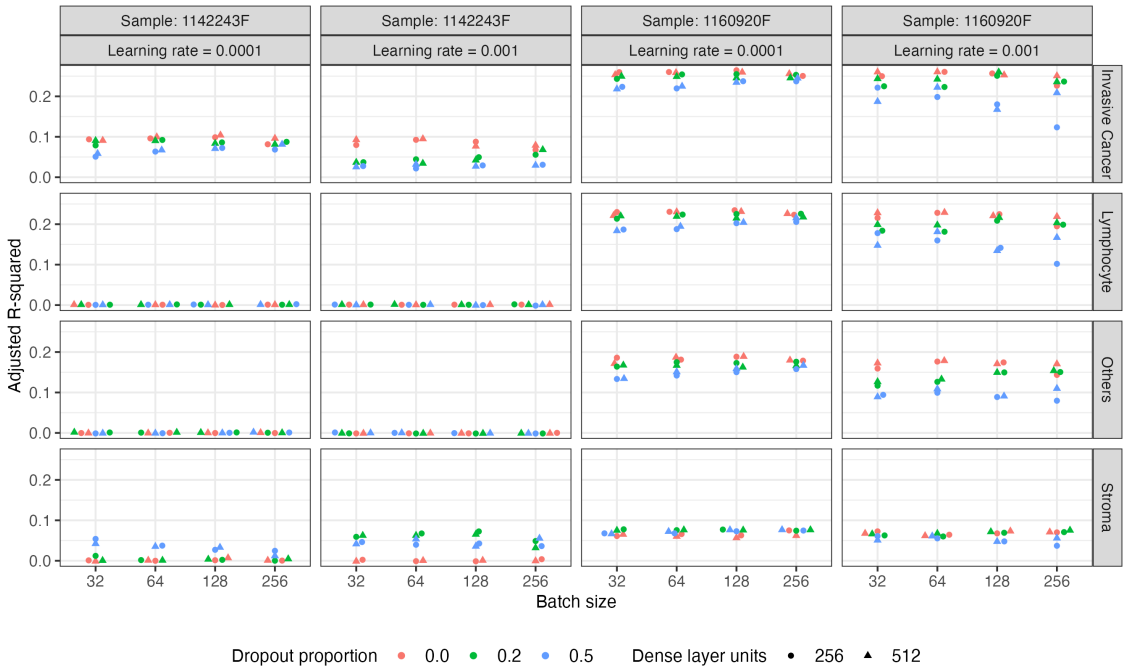


Figure 6.38: Sample-specific adjusted R-squared values of the testing dataset for networks predicting proportions of four cell types using standard patches from two independent lab samples.

Lastly, we conducted a comparison between the four predicted proportions obtained by collapsing the nine predicted cell type proportions from section 6.2.1 and the four predicted proportions directly obtained by the network using these four cell types as the response variable. By comparing Figure 6.31 and Figure 6.37, it becomes evident that the correlations between the observed and predicted proportions for all cell types other than stroma were smaller when the predictions were obtained directly from the neural network, compared to when the predictions were derived by aggregating the nine predicted cell type proportions. Additionally, Figure 6.39 demonstrates a high correlation between the predicted proportions obtained through the two approaches, although there was a relatively weaker correlation for the combined proportions of normal epithelial and myeloid. Notably, for all cell types in sample 1160920F, when the predicted proportions

were relatively small, the predicted proportion obtained directly from the neural network tended to be larger than the predicted proportion obtained by collapsing the proportions of the nine cell types. Conversely, when the predicted proportions were relatively large, the opposite trend was observed.

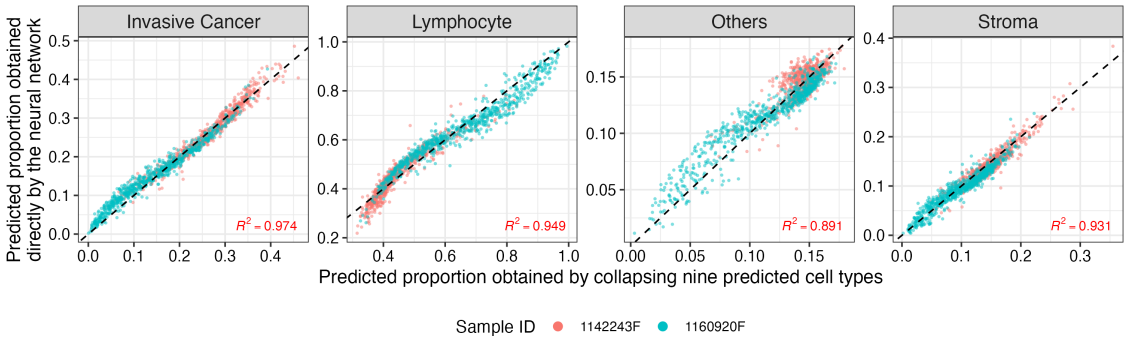


Figure 6.39: Scatterplot comparing predicted proportions achieved by two approaches for standard patches of two independent lab samples derived from identical testing data. The proportions were estimated through two methods: direct prediction by a network with four response variables, and collapsing the proportions of nine major cell types predicted by a network with nine response variables.

6.3 Four Samples Provided by Wu et al.

6.3.1 Predicting Nine Cell Type Proportions

	Number of patches		
Sample	Training	Validation	Testing
CID4290	1710	372	350
CID4465	851	178	182
CID44971	813	166	183
CID4535	778	174	175

Table 6.4: Number of standard patches from each of the four samples processed by Wu et al. in the training, validation, and testing datasets.

In this section, we ananlyzed the performance of the neural networks that predicted the proportions of the nine cell types for each patch in the four samples processed by Wu et al. The network architecture's varying hyperparameters align with those utilized in the regression task networks for two independent lab samples, as elaborated in section 6.2.1. The training, validation, and testing datasets remained consistent across all network evaluations. There are in total 5932 patches, with 4152 for training, 890 for validation, and 890 for

testing. Table 6.4 reveals an imbalance in the number of patches from each sample, with sample CID4290 contributing a higher number of patches to each dataset.

Figure 6.40 presents the cellular compositions of the training, validation, and testing data. Notably, in the training dataset, sample CID4465 exhibits a predominant presence of CAFs, while sample CID4535 demonstrates a higher proportion of cancer epithelial cells (Figure 6.41).

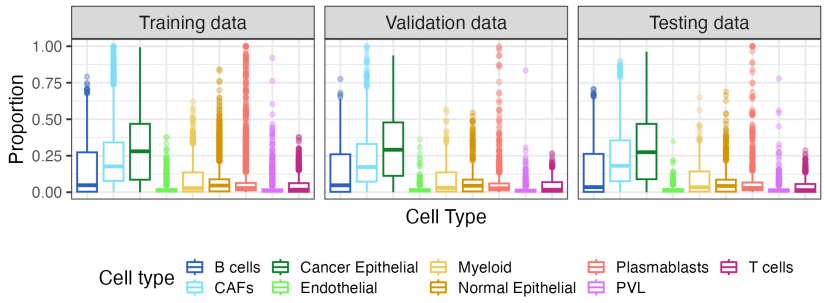


Figure 6.40: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of nine cell types using standard patches from four samples processed by Wu et al.

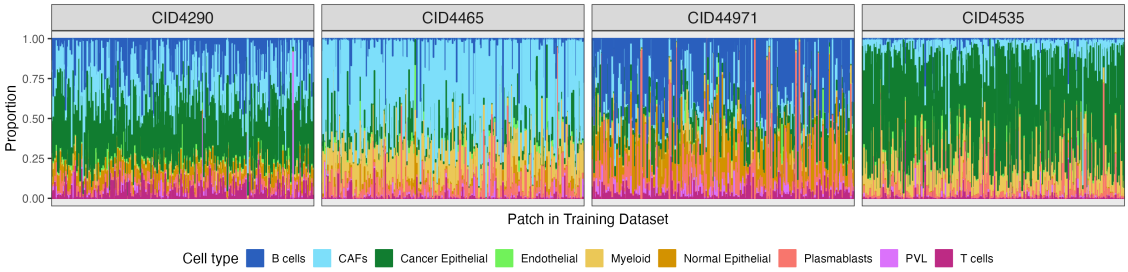


Figure 6.41: Cell type compositions in training dataset for regression task predicting proportions of nine cell types using standard patches from four samples processed by Wu et al.

Learning Curves

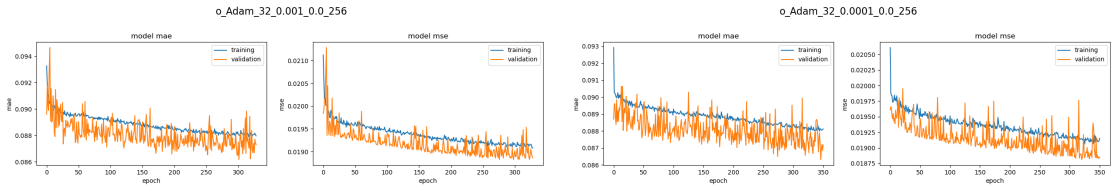


Figure 6.42: Learning curves of neural networks with a batch size of 32, a hidden dense layer with 256 neurons, and no dropout layer aiming to predict the proportions of nine cell types using standard patches from four samples processed by Wu et al.

It is evident from the learning curves that all networks effectively learned from the training data. Notably, the validation loss and metrics consistently outperformed the corresponding training loss and metrics across all hyperparameter combinations.

Evaluation Using Validation Data

Across all hyperparameter combinations, the validation metrics and loss were consistent (Figure 6.43). Among these combinations, the neural network that had the lowest validation loss had the following specifications: batch size = 64, learning rate = 0.001, dropout proportion = 0.0, and dense layer units = 256.

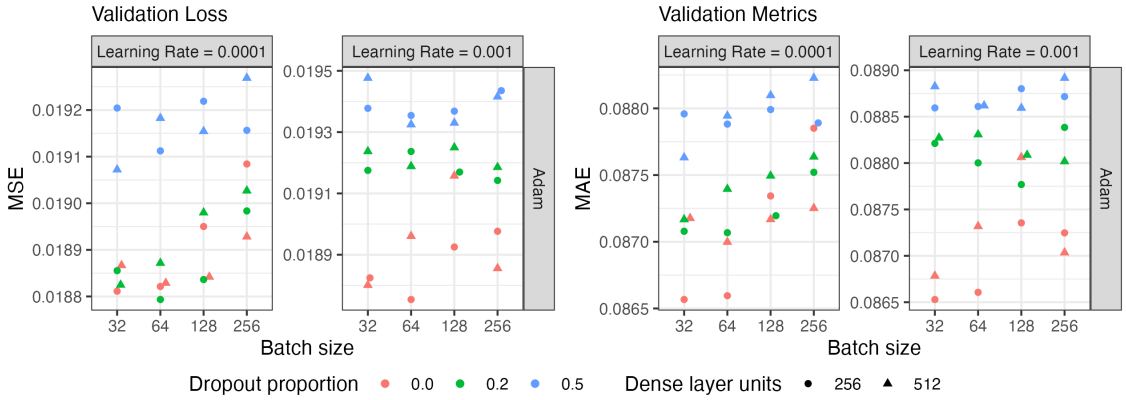


Figure 6.43: Validation loss and metrics from neural networks predicting proportions of nine cell types using standard patches from four samples processed by Wu et al.

Networks with a learning rate of 0.0001 consistently achieved lower validation loss at the optimal epoch compared to networks with a learning rate of 0.001. However, there were exceptions where a learning rate of 0.001 achieved lower validation loss when no dropout layer was included in the model.

An increase in the dropout proportion generally led to higher validation loss at the optimal epoch, particularly when the learning rate was set to 0.001. However, when the learning rate was set to 0.0001, networks with a dropout proportion of 0.5 consistently exhibited higher validation loss compared to other networks. Nevertheless, the presence of a dropout layer with a proportion of 0.2 did not consistently result in a higher validation loss compared to networks without a dropout layer. When the dense layer had 512 units, networks with a dropout proportion of 0.2 showed higher validation loss than those without a dropout layer, while networks with a dropout proportion of 0.2 achieved lower validation loss than those without a dropout layer when the dense layer had 256 units. This pattern was observed for batch sizes 64, 128, and 256.

Lastly, no consistent trend in validation loss was observed in relation to the number of units in the dense layer when controlling for other parameters.

Predication Using Testing Data

In general, all cell types had very small slopes of the best-fit line between the predicted and the observed proportions in the testing dataset (Figure 6.44, Table D.7). Among all cell types, cancer epithelial and myeloid cells consistently exhibited higher slopes than others when predicted with the same network.

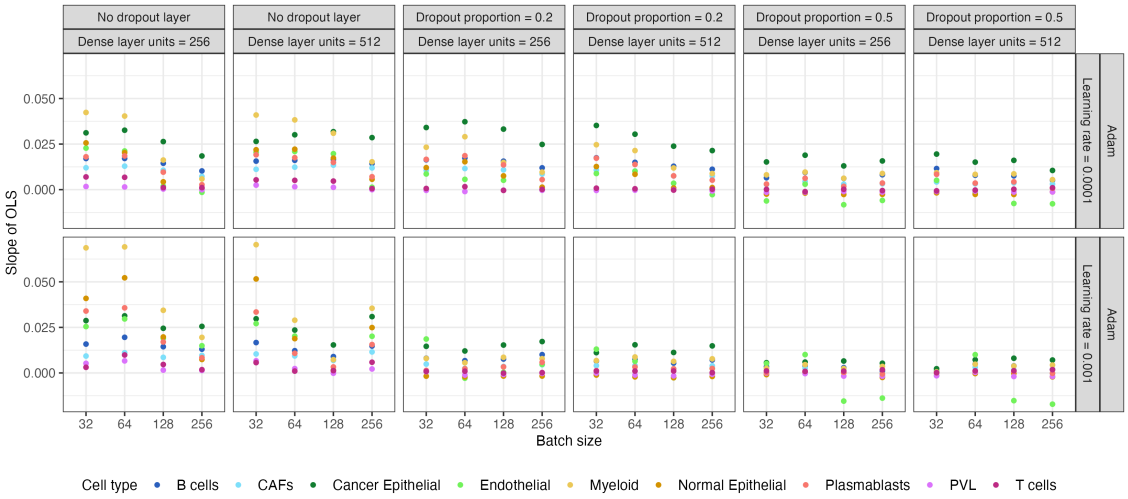


Figure 6.44: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using standard patches from four samples processed by Wu et al.

Here, we specifically examined the neural network with batch size of 64, learning rate of 0.001, dropout proportion of 0.0, and dense layer units of 256, as this parameter combination yielded the lowest validation loss. The scatterplots of predicted proportions against observed proportions demonstrate nearly horizontal best-fit lines (Figure 6.45), indicating weak or negligible correlations between the predicted and observed proportions. Notably, the predicted proportions exhibited limited variation around certain value for each cell type. The predicted proportions tended to cluster around the mean of the observed proportions from the testing data (Figure D.11). It implies that the network struggled to capture the full range of variation present in the observed data.

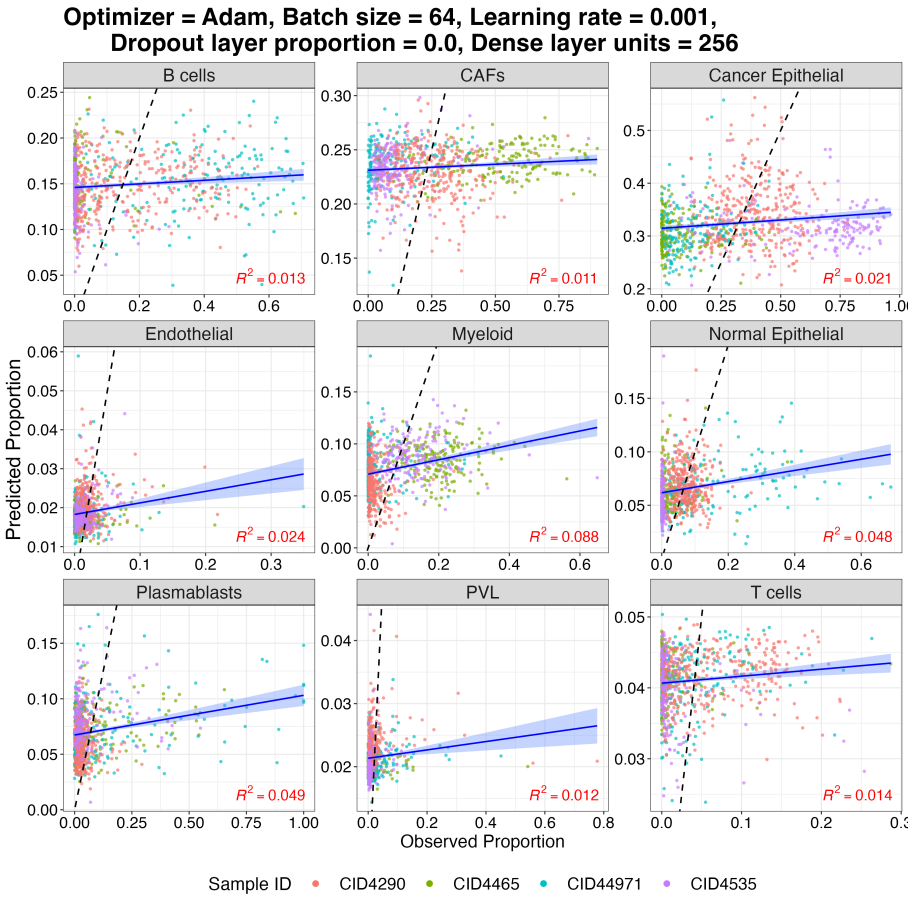


Figure 6.45: Scatterplot comparing predicted and deconvoluted proportions of nine cell types in standard patches from four samples processed by Wu et al. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. The overall R-squared value is indicated in red.

In addition, we can observe that there was no discernible batch effect concerning the correlations between the predicted and observed proportions (Figure 6.46). All samples displayed similarly weak correlations between the predicted and observed proportions. In order to explicitly assess the disparity in correlations between the predicted and observed proportions for each sample, we computed the sample-specific adjusted R-squared values across all networks (Figure D.12). Specifically, sample CID4535 exhibited generally superior correlations between the predicted and observed proportions for the majority of cell types. Sample CID44971 showed better correlations for normal epithelial cells and plasmablasts. In contrast, samples CID4290 and CID4465 displayed relatively weaker correlations for several cell types.

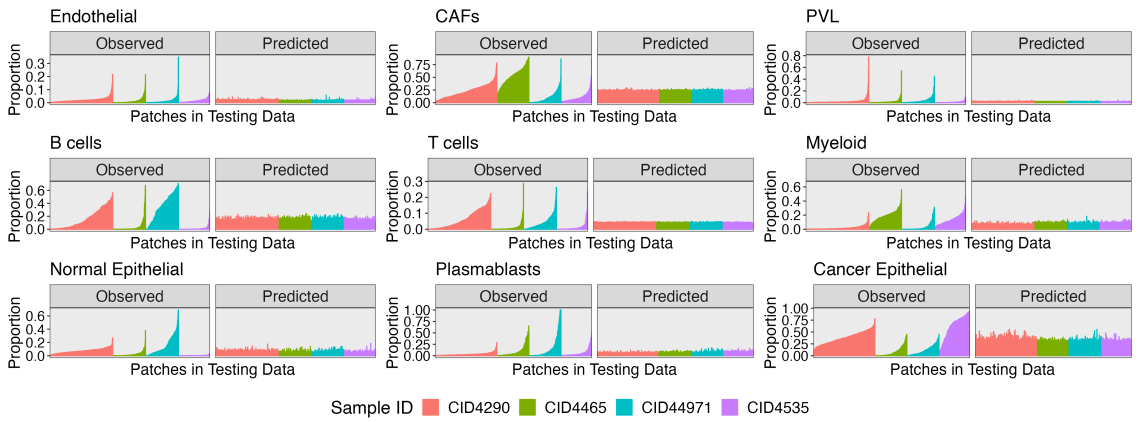


Figure 6.46: Comparison of predicted and observed nine cell type compositions in each of the standard patches from four samples processed by Wu et al.

Finally, we observed that there was no noticeable improvement in the correlation between the predicted proportions and the observed proportions when collapsing the nine major cell types into four pathological cell types (Figure 6.47).

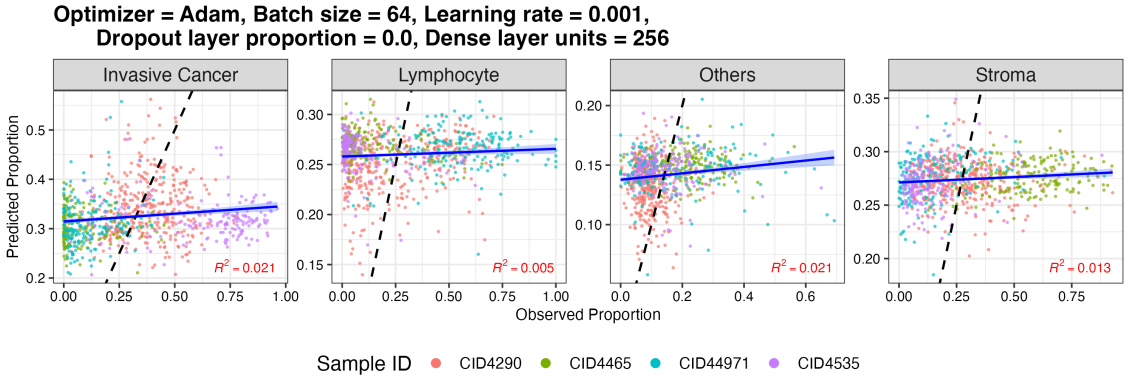


Figure 6.47: Scatterplot comparing four collapsed predicted proportions to four collapsed deconvoluted proportions in the standard patches from four samples processed by Wu et al.

6.3.2 Predicting Four Cell Type Proportions

Our task shifted towards constructing neural networks designed to directly predict the four pathological cell types. We kept the parameters in the network architecture and the composition of the training, validation, and testing datasets unchanged from the previous regression task detailed in section 6.3.1. Notably, the cellular composition of the training, validation, and testing datasets exhibited no predominance of any cell

type (Figure 6.48).

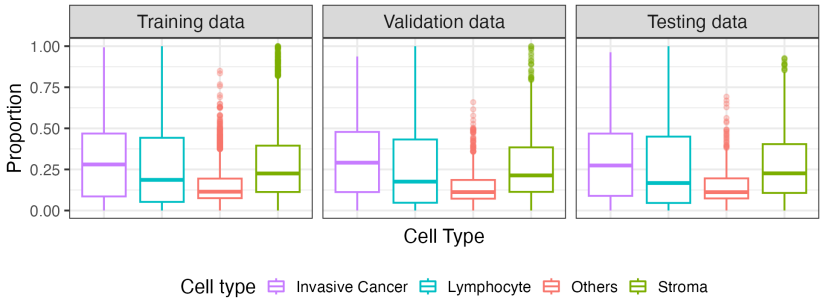


Figure 6.48: Cell type compositions in training, validation, and testing datasets for the regression task predicting proportions of four cell types using standard patches from four samples processed by Wu et al.

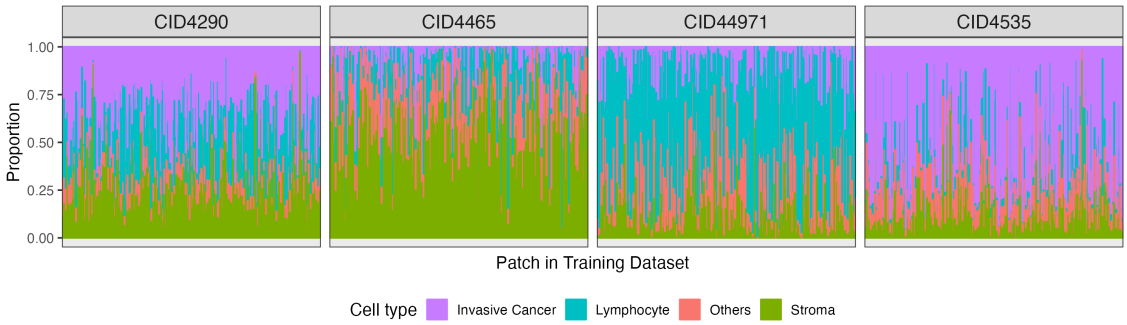


Figure 6.49: Cell type compositions in training dataset for regression task predicting proportions of four cell types using standard patches from four samples processed by Wu et al.

Learning Curve

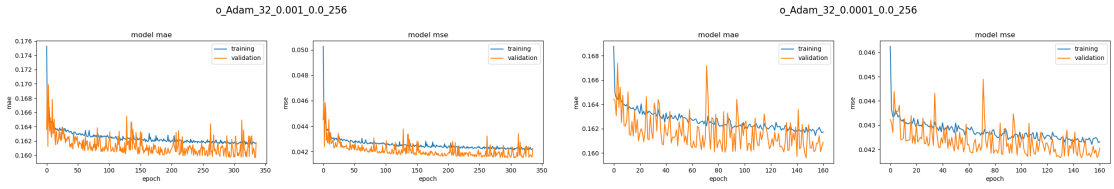


Figure 6.50: Learning curves of neural networks with a batch size of 32, a hidden dense layer with 256 neurons, and no dropout layer aiming to predict the proportions of four cell types using standard patches from four samples processed by Wu et al.

The learning curves reveal that not all networks effectively learned from the given training data, particularly for certain networks with a learning rate of 0.001 that learned for fewer than 100 epochs. However, it is

worth noting that, across all networks, the validation loss and metrics consistently remained lower than the corresponding training loss and metrics.

Evaluation Using Validation Data

Across all networks, the validation metrics and loss were consistent (Figure 6.51). Among all combinations of hyperparameters, the neural network that had the lowest validation loss had the following specifications: batch size = 128, learning rate = 0.0001, dropout proportion = 0.2, and dense layer units = 512.

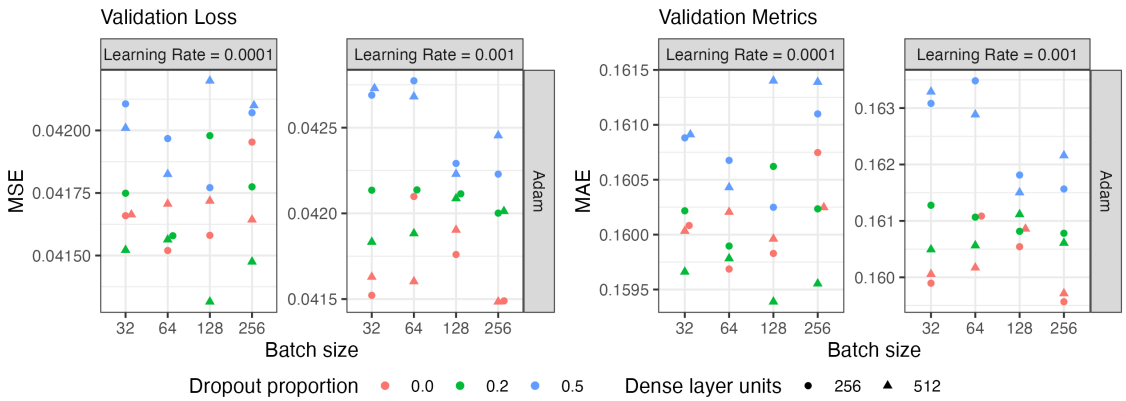


Figure 6.51: Validation loss and metrics from neural networks predicting proportions of four cell types using standard patches from four samples processed by Wu et al.

For the networks that incorporated a dropout layer, those with a learning rate of 0.001 generally exhibited higher validation loss at the optimal epoch compared to networks with the same parameters but a different learning rate of 0.0001. Among networks with a learning rate of 0.001, we observed a consistent association between increasing dropout layer proportion and higher validation loss at the optimal epochs, while keeping other parameters the same. Lastly, we found no consistent correlation between the number of neurons in the hidden dense layer and the lowest validation loss. Similarly, we did not observe a consistent association between the batch size and the validation loss when considering all combinations of the learning rate, dropout layer proportion, and units in the hidden dense layer.

Prediction Using Testing Data

In general, the predicted proportions of all cell types aligned poorly with the observed proportions across all networks (Figure 6.52). Besides, based on the slope of the best-fit line, the predicted proportions of

invasive cancer cells and others (normal epithelial and myeloid) exhibited a better alignment with their observed proportions than the predicted proportions of stroma and lymphocyte. It is worth mentioning that the specific network, which produced the highest slope of the best-fit line for invasive cancer, lymphocyte, and stroma, also yielded the lowest validation loss among all the networks (Table D.9). Consequently, we concentrated on examining this particular network in more detail.

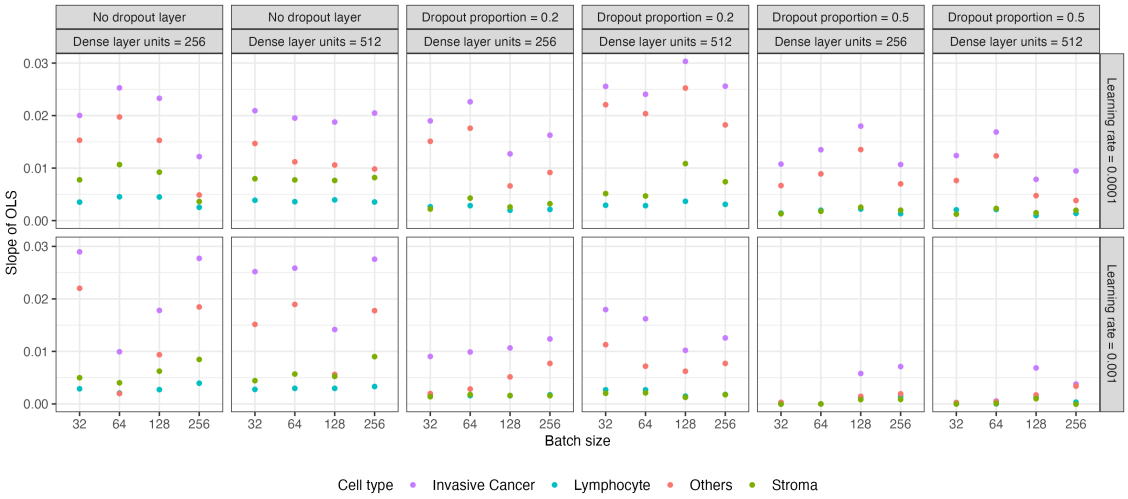


Figure 6.52: Slope of the best-fit line between predicted and observed proportions of four cell types in testing data, predicted by networks trained using standard patches from four samples processed by Wu et al.

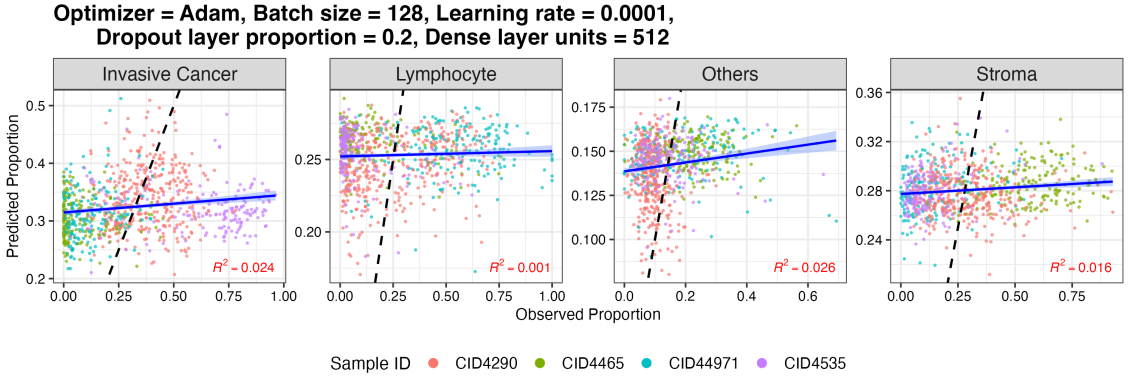


Figure 6.53: Scatterplot comparing predicted and deconvoluted proportions of four cell types in standard patches from four samples processed by Wu et al. The black dashed line represents the diagonal line, while the blue solid line represents the best-fit line. The overall R-squared value is indicated in red.

After analyzing the scatterplots in Figure 6.53 and the histograms in Figure D.13, it became apparent that the predicted proportions for all cell types were tightly clustered around the average observed proportions.

This suggests that the network struggled to accurately predict the proportions of individual cell types within each patch. Instead, the predicted proportions can only be reliably used to estimate the average proportion across a set of patches. Besides, the presence of batch effects was evident in the R-squared values of each cell type across different samples within the same network (Figure D.14).

Lastly, by comparing Figure 6.47 and Figure 6.53, the correlations between the observed and predicted proportions for all cell types except lymphocyte were larger when the predictions were obtained directly from the neural network, compared to when the predictions were derived by aggregating the nine predicted cell type proportions. Additionally, we observed a high correlation between the predicted proportions obtained through the two approaches, although there were relatively weaker correlations for the combined proportions of normal epithelial and myeloid and stroma (Figure 6.54). Notably, for the lymphocyte, when the predicted proportions were greater than 0.225, the predicted proportions obtained directly from the neural network were smaller than the predicted proportions obtained by collapsing the proportions of the nine cell types. Conversely, for the stroma, the predicted proportions obtained directly from the neural network tended to be greater than the predicted proportions obtained by collapsing the proportions of the nine cell types.

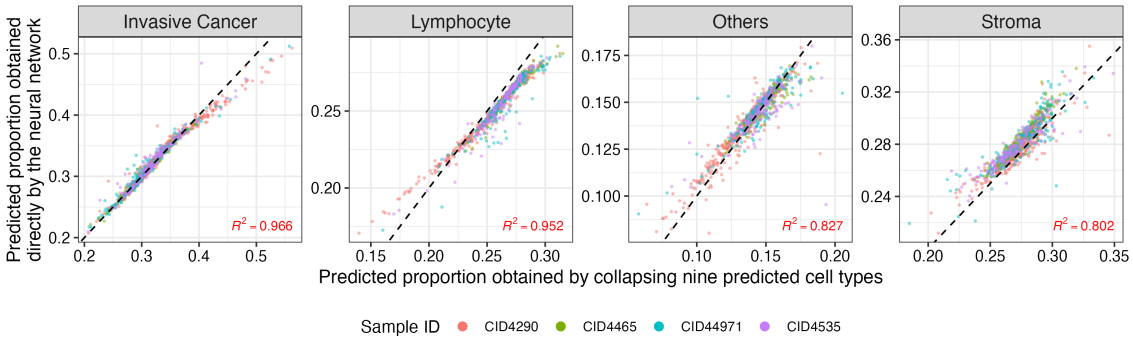


Figure 6.54: Scatterplot comparing predicted proportions achieved by two approaches for standard patches of all four samples processed by Wu et al. derived from identical testing data. The proportions were estimated through two methods: direct prediction by a network with four response variables, and collapsing the proportions of nine major cell types predicted by a network with nine response variables.

Chapter 7

Classification Results

This chapter focuses on the analysis of ResNet50 transfer learning for classification tasks, using samples from either Wu et al. [2021] or Sudmeier et al. [2022]. The input for the neural networks consisted of patches from the samples, while the output was the qualitative annotation of cell composition patterns. Various configurations were explored, including different batch sizes for training, optimizers, learning rates, dropout layer proportions, and numbers of neurons in the hidden dense layer. Throughout the analysis, the training, validation, and testing datasets remained consistent across all networks with varying parameters for the same task and samples. The loss function employed was categorical entropy, with accuracy used as the metrics function. The networks were trained for a maximum of 1000 epochs, although some networks terminated early with a patience of 30 epochs. For further information on the dataset and network architecture, please refer to Chapter 2.

7.1 Six Breast Tissue Samples from Wu et al. [2021]

In the study conducted by Wu et al. [2021], the capture spots in each of the six samples were annotated according to their pathology. Table 2.2 in Chapter 2 provides the number of patches with each pathology annotation within each sample. The classification task of the neural networks in this section involved annotating the standard patches for each individual sample separately. Patches annotated as NA, Artefact, or Uncertain were excluded from the analysis. Additionally, for each sample, rare labels that were assigned to less than 50 patches in the corresponding sample were also excluded. The final annotation and the cor-

responding number of patches for each sample can be found in Table 7.1. In this classification task, the varying network architecture hyperparameters included the batch size of training and the number of neurons in the hidden dense layer units. Based on the previous analysis for the regression tasks in Chapter 5 and Chapter 6, the Adam optimizer with a learning rate of 0.0001 was employed to ensure effective learning. Additionally, since previous regression tasks did not exhibit significant signs of overfitting, a dropout layer after the hidden dense layer was not included in the classification networks.

	Training	Validation	Testing
1142243F			
Invasive cancer + stroma + lymphocytes	2531	537	559
Necrosis	396	97	75
Stroma	321	62	62
1160920F			
Adipose tissue	57	15	11
Invasive cancer + stroma + lymphocytes	2218	463	465
Lymphocytes	123	32	31
Normal glands + lymphocytes	195	44	39
Stroma	785	170	177
CID4290			
Invasive cancer + stroma	1454	319	309
Invasive cancer + stroma + lymphocytes	158	23	34
Stroma	81	21	20
CID4465			
Invasive cancer + stroma + lymphocytes	789	170	172
Stroma	54	10	9
CID44971			
DCIS	194	42	37
Invasive cancer + lymphocytes	214	49	54
Lymphocytes	60	12	9
Normal + stroma + lymphocytes	162	39	39
Stroma	101	18	15
Stroma + adipose tissue	80	14	20
CID4535			
Invasive cancer	296	67	55
Invasive cancer + lymphocytes	249	49	63
Lymphocytes	45	11	13
Stroma	122	25	22

Table 7.1: Patch counts for each of the six samples, categorized by the annotations involved in the classification task.

7.1.1 Learning Curve

Among the networks trained using patches from samples 1142243F and 1160920F, the majority effectively learned from the training data. These networks demonstrated a consistent decrease in validation loss for at least 100 epochs, indicating successful learning. However, the networks trained using the patches from four other samples displayed a different pattern. In most cases, these networks terminated early after the initial 30 epochs. Also, it was observed that the validation loss exhibited much greater fluctuations compared to the corresponding training loss. The validation loss showed larger variations throughout the training process. Moreover, the fluctuation in validation accuracy was consistently high across all networks. In some cases, the training accuracy also displayed fluctuations as significant as the validation accuracy, indicating that the training process was not as stable for those networks.

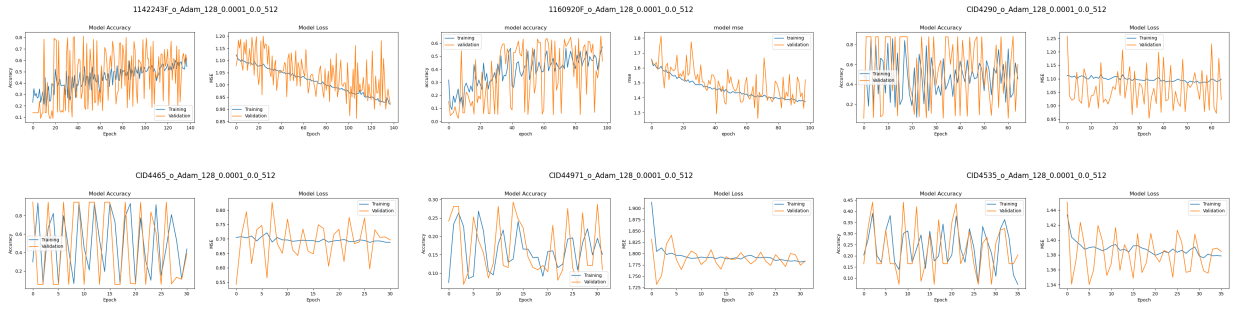


Figure 7.1: Learning curves of neural networks with a batch size of 128 and a hidden dense layer of 512 neurons aiming to classify standard patches from each of the six samples. Each figure title specifies the sample, image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

7.1.2 Evaluation Using Validation Data

We examined and compared the validation loss and metrics at the optimal epoch for each combination of batch size and dense layer units (Figure 7.2). We observed that higher batch sizes were generally associated with higher validation loss, with a few exceptions in all samples except for 1142243F and CID4290. Additionally, concerning the impact of the number of neurons in the hidden dense layer, for the two samples processed by an independent lab, networks with a hidden dense layer of 512 neurons had lower validation loss at the optimal epoch compared to networks with a hidden dense layer of 256 neurons but the same batch size. The remaining four samples processed by Wu et al. displayed a similar association between the dense

layer units and validation loss, with a few exceptions.

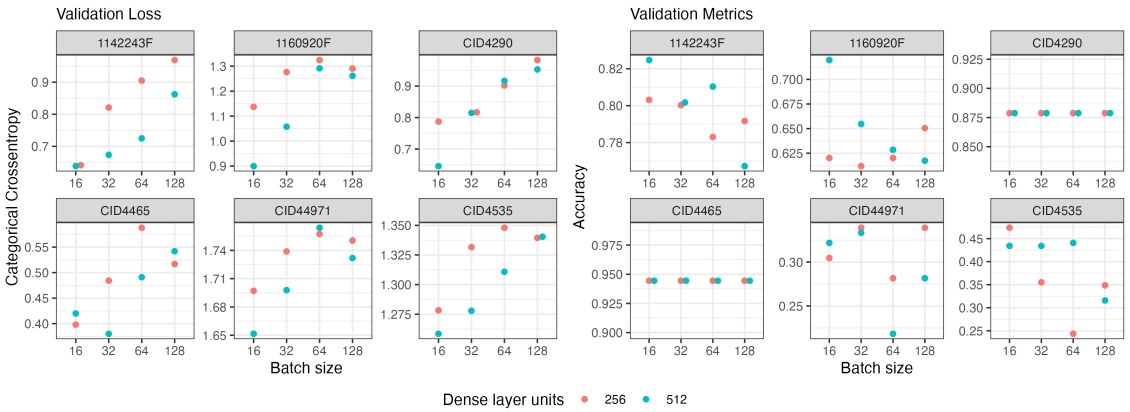


Figure 7.2: Validation loss and accuracy from neural networks classifying standard patches from each of the six samples in Wu et al.

The batch size and the number of neurons in the hidden dense layer of the network that exhibited the lowest validation loss at the optimal epoch for each of the samples are listed in Table 7.2. Specifically, the networks with the lowest validation loss for all samples had a hidden dense layer with 512 neurons and a batch size of 16, except for sample CID4465, where the batch size was 32. The Receiver Operating Characteristic (ROC) curves of the validation data generated by these networks can be found in Appendix E (Figure E.1).

	Validation Loss	Validation Accuracy	Batch size	Dense layer Units
1142243F	0.638	0.825	16	512
1160920F	0.900	0.720	16	512
CID4290	0.646	0.879	16	512
CID4465	0.380	0.944	32	512
CID44971	1.652	0.322	16	512
CID4535	1.258	0.434	16	512

Table 7.2: Batch size and hidden dense layer units of the neural network that gave the lowest validation loss at the optimal epoch of each sample.

Furthermore, we found that the validation loss and validation metrics did not consistently align with each other (Figure 7.2). Networks with lower validation loss did not necessarily exhibit higher validation metrics. Notably, samples CID4290 and CID4465 displayed higher validation accuracy compared to other samples, with 87.9% and 94.4% of patches in the respective validation datasets being correctly classified

regardless of the network used. However, it is important to note that accuracy as a metric has limitations, particularly in imbalanced datasets. Constantly high accuracy can be attributed to misclassifying all patches with less prevalent labels while correctly classifying all patches with the most prevalent label. Therefore, it is essential to consider other class-specific metrics as well (Table E.1).

7.1.3 Classification of Testing Data

We analyzed the classification results of the testing data produced by the network that achieved the lowest validation loss at the optimal epoch for each sample. The ROC curves of the testing data generated by these networks are examined and depicted in Figure 7.3.

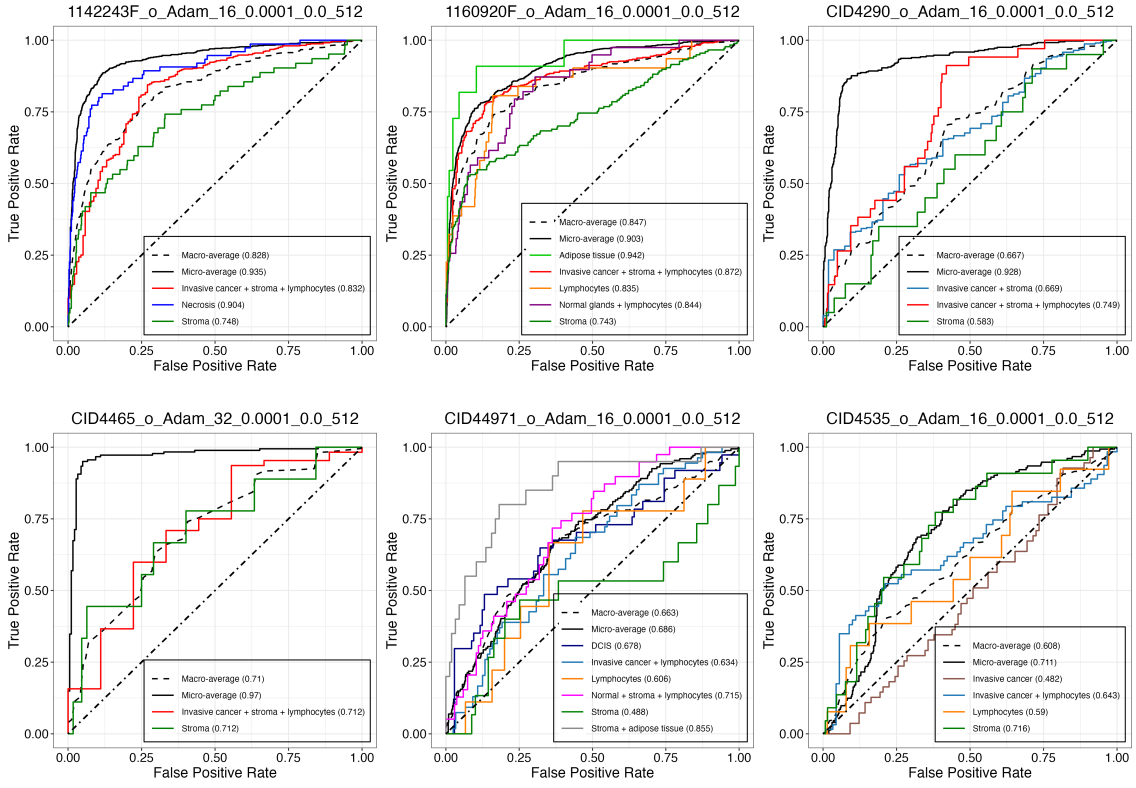


Figure 7.3: ROC curves of the testing data of all six samples obtained by the networks that exhibited the lowest validation loss.

For each class, we computed the Area under the ROC Curve (AUC), as well as the micro-averaged and macro-averaged AUCs. The macro-averaged metrics were obtained by calculating the metrics for each class separately and then taking the average, treating all classes equally. Conversely, the micro-averaged metrics

were obtained by aggregating the contributions of all classes to derive average metrics. In our multi-class classification setup, the micro-average is preferred due to class imbalance. Furthermore, we calculated the precision, recall, and F1-score for each individual class, as well as their averaged values (Table E.2). The performance of the network varies across different samples and across different classes within each sample.

For sample 1142243F, when considering an equal weight for each class, the network's annotations showed a moderate average performance across all classes, with macro-averaged precision, recall, and f1-score being 0.56, 0.52, and 0.51, respectively. When considering an equal weight for each individual prediction (i.e., micro-averages), all three overall metrics improved to 0.78, 0.82, and 0.79, respectively. Besides, the class-specified f1-scores demonstrate that the network performed poorly in annotating a standard patch from sample 1142243F with the true label as Stroma (f1-score = 0.06). Specifically, out of all the patches annotated as Stroma by the network, only 25% were correct. Also, the network detects only 3.2% of the true Stroma patches. On the other hand, the Invasive cancer + stroma + lymphocytes class exhibited a high f1-score of 0.90, indicating excellent overall performance for this specific class. When the network predicted this class, it was correct 86.4% of the time and the network successfully identified 94.5% of the patches belonging to this class.

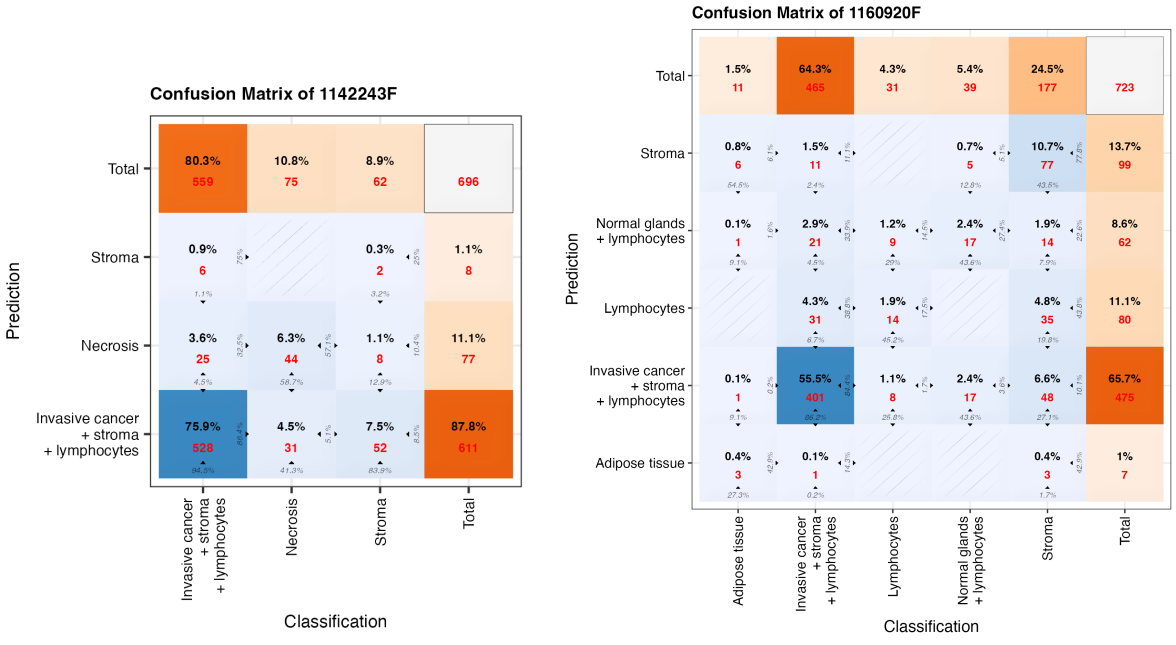


Figure 7.4: Confusion matrices of samples 1142243F and 1160920F.

The network's overall performance for all classes in sample 1160920F was moderate based on the macro-averages (precision = 0.50, recall = 0.49, and f1-score = 0.47). However, the higher micro-averages (precision = 0.76, recall = 0.71, and f1-score = 0.72) suggest that the network performed relatively well when considering an equal weight for each individual prediction. The network demonstrated strong performance in accurately identifying the Invasive cancer + stroma + lymphocytes class, with a high precision of 0.84, recall of 0.86, and f1-score of 0.85. However, it struggled to accurately identify Lymphocytes, with low precision and recall, resulting in a low f1-score of 0.25. The performance for Stroma, Normal glands + lymphocytes, and Adipose tissue classes fell in between, with f1-scores being 0.56, 0.34, and 0.33, respectively.

For CID4290, the macro-averages were low (precision = 0.28, recall = 0.33, and f1-score = 0.31). However, the micro-average metrics showed higher overall accuracy, with precision at 0.72, recall at 0.85, and an f1-score of 0.78. The network performed well in identifying patches of class Invasive cancer + stroma, achieving a high precision of 0.85, a perfect recall of 1.00, and an f1-score of 0.92. However, it failed to correctly annotate all other patches, resulting in precision, recall, and f1-scores of 0.00 for both the other two classes.

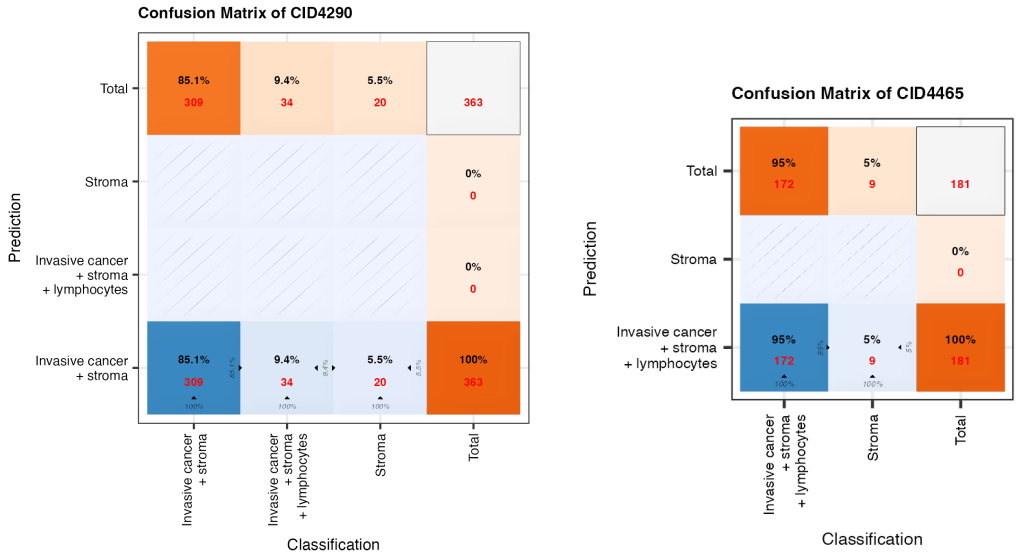


Figure 7.5: Confusion matrices of samples CID4290 and CID4465.

Similar to sample CID4290, for sample CID4465, though the micro-average metrics showed high accuracy, with precision at 0.90, recall at 0.95, and an f1-score of 0.93, the network only perform well on accurately identifying patches of Invasive cancer + stroma + lymphocytes, achieving a precision of 0.95, a

perfect recall of 1.00, and a f1-score of 0.97. However, it failed to correctly label all patches of the Stroma class, with all three metrics being 0.00.

For sample CID44971, the macro-average metrics indicate a relatively low overall performance (precision = 0.35, recall = 0.24, and f1-score = 0.21), while the micro-average metrics show a slightly higher but still limited overall performance (precision = 0.40, recall = 0.32, and f1-score = 0.27). There was a complete lack of performance in labeling patches of Lymphocytes and Stroma classes since the precision, recall, and f1-scores for these classes were all 0.00. Additionally, while the network demonstrates moderate accuracy in identifying patches of DCIS and Normal + stroma + lymphocytes, it struggled in correctly labeling patches of Invasive cancer + lymphocytes and Stroma + adipose tissue.

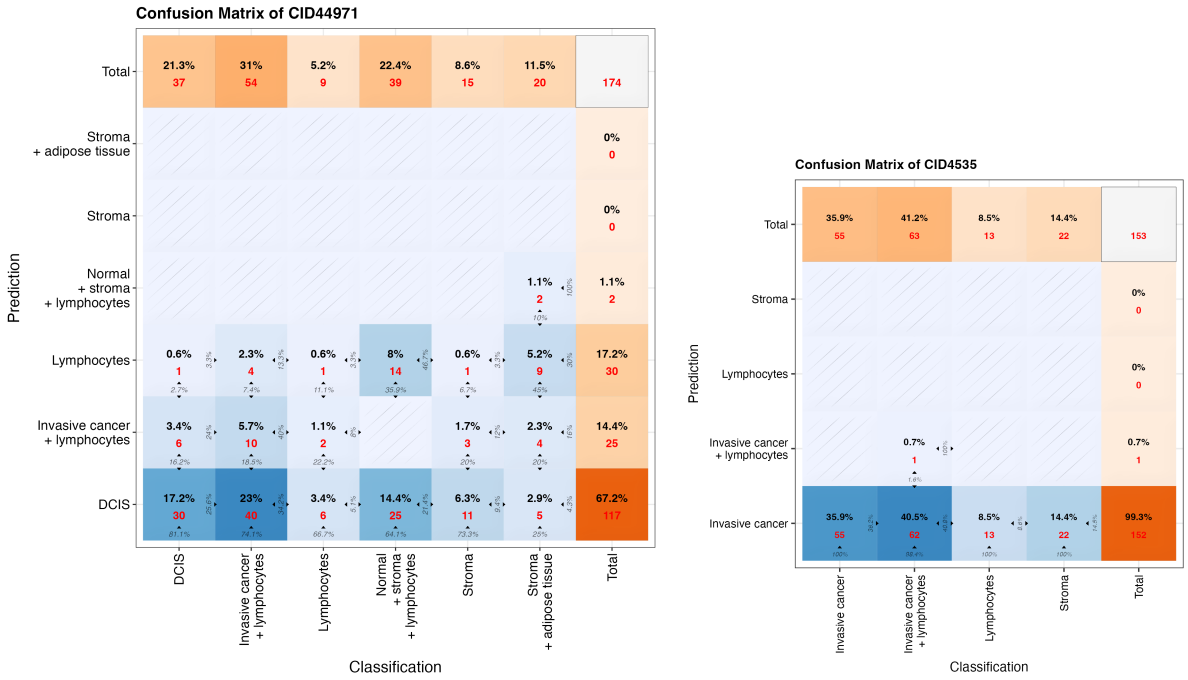


Figure 7.6: Confusion matrices of samples CID44971 and CID4535.

For sample CID4535, the macro-average metrics were very low (precision = 0.09, recall = 0.25, and f1-score = 0.13), and the micro-average metrics also indicated low overall accuracy (precision = 0.13, recall = 0.36, and f1-score = 0.19). The network achieved a precision of 0.36 in identifying patches of Invasive cancer. However, it had a perfect recall of 1.00, suggesting a successful detection of all patches whose true labels were Invasive cancer. Unfortunately, the network failed to correctly label any patches of Invasive cancer + lymphocytes, Lymphocytes, and Stroma classes. The precision, recall, and f1-scores for these

classes were all 0.00, indicating a complete lack of performance in identifying patches belonging to these classes.

7.2 A Melanoma Brain Metastasis Tissue Sample for Sudmeier et al. [2022]

In the study by Sudmeier et al. [2022], the capture spots in a melanoma brain metastasis from patient 16 were annotated based on their transcriptional phenotype. Among the annotated clusters, clusters 1 and 7 were identified as peritumoral inflammation, with cluster 7 also including blood vessels. Clusters 3, 4, and 5 were identified as tumor regions, with cluster 5 specifically being inflammation-adjacent tumor.

The purpose of the neural networks in this section was to classify between tumor (cluster 4,5) and inflammation (cluster 1,7) for each standard patch of patient 16 in Sudmeier et al. [2022]. Patches in cluster 3 were excluded from our dataset due to its annotation as necrotic tumor. The varied parameters were the batch size of training, learning rate, dropout layer proportion, and the size of the dense layer. The optimizer used was Adam. We also trained the same networks with different image types, either before or after stain normalization. The training, validation, and testing datasets were the same across all networks. The number of patches with each label in the training, validation, and testing datasets is listed in Table 7.3. The datasets were imbalanced with more patches annotated as inflammation.

	Training	Validation	Testing
Inflammation	427	85	86
Tumor	287	68	67
Total	714	153	153

Table 7.3: Number of patches with each label in the training, validation, and testing datasets from Sudmeier et al.

7.2.1 Learning Curve

Throughout the training process, all the networks consistently showed a decrease in validation loss for a minimum of 150 epochs, indicating successful learning regardless of whether the patches underwent stain normalization. However, for certain networks, there were greater fluctuations in validation loss and accuracy compared to their corresponding training loss and accuracy. Some networks exhibited higher validation loss and lower validation accuracy than their training loss and accuracy, while others showed the opposite pattern

with lower validation loss and higher validation accuracy.

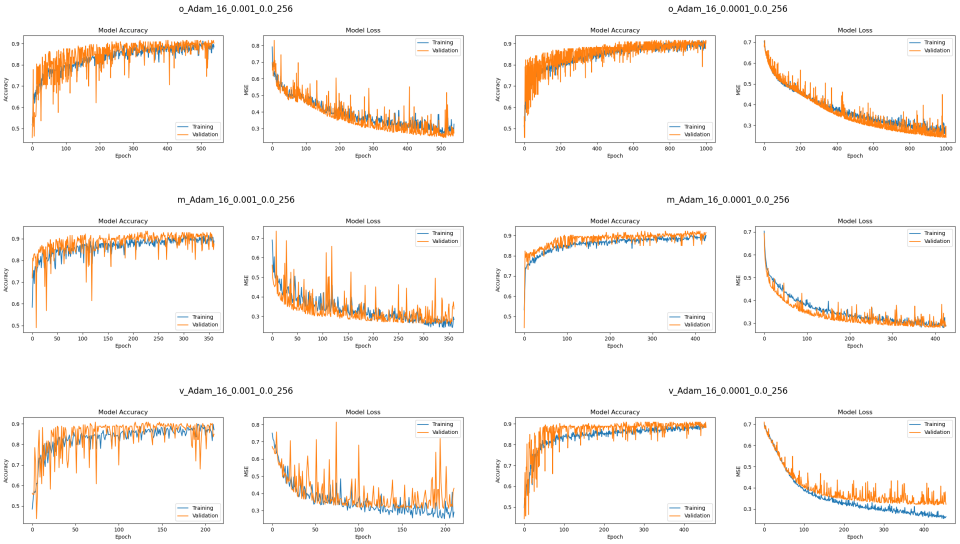


Figure 7.7: Learning curves of neural networks with a batch size of 16, a hidden dense layer of 256 neurons, and no dropout layer aiming to classify standard patches from Sudmeier et al. Each figure title specifies the sample, image type, optimizer, training batch size, learning rate, dropout rate, and size of the dense layer, with these parameters separated by underscores.

7.2.2 Evaluation Using Validation Data

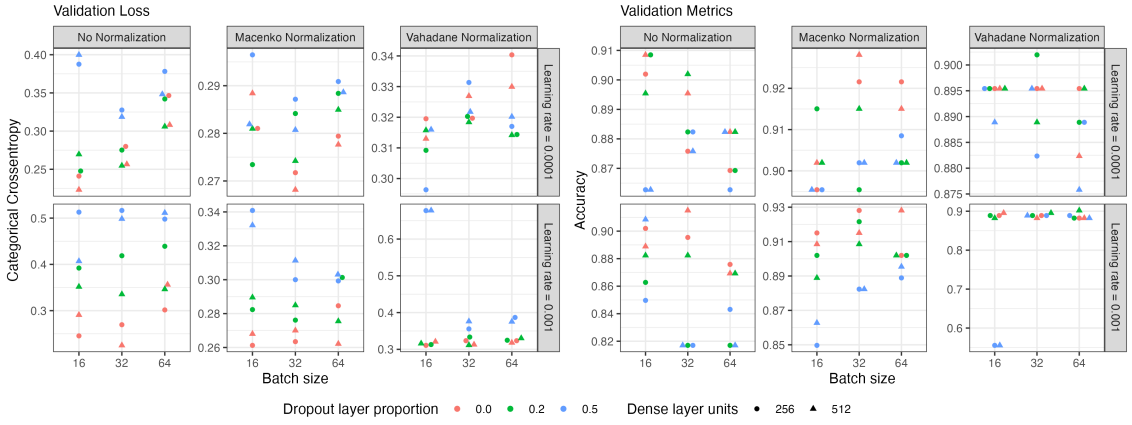


Figure 7.8: Validation loss and accuracy from neural networks classifying standard patches from the sample from Sudmeier et al.

We analyzed and compared the validation accuracy and loss of the optimal epoch for each parameter combination. The network that gave the lowest validation loss had the following configurations: image type =

original, optimizer = Adam, batch size = 16, learning rate = 0.0001, dropout layer proportion = 0.0, dense layer units = 512. Across different image types, training batch sizes, learning rates, dropout layer proportions, and hidden dense layer units, no strict and consistent trend was observed in the validation loss and accuracy over all networks (Figure 7.8). However, some general trends can be observed among certain subsets of networks. For example, among the networks with a dropout rate of 0.5, those with a learning rate of 0.001 had higher validation losses compared to those with a learning rate of 0.0001, given the same image type and other hyperparameters of network architecture. In addition, among networks with a learning rate of 0.001, a higher dropout rate was associated with higher validation loss after controlling for image type and other hyperparameters, with a few exceptions where networks without a dropout layer had slightly higher validation losses than corresponding networks with a dropout rate of 0.2.

7.2.3 Classification of Testing Data

In this section, we analyze the classification results of the testing data generated by the networks. Figure 7.9 shows the ROC curve of the classification performed by the network with the lowest validation loss. The high AUC value of 0.980 indicates the high accuracy of the network in distinguishing between Tumor and Inflammation classes.

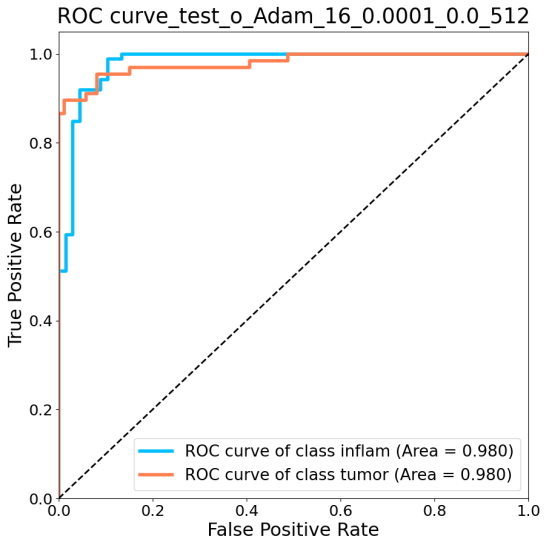


Figure 7.9: ROC curve for the classification task on the testing data from Sudmeier et al. conducted by the network that gave the lowest validation loss at the optimal epoch

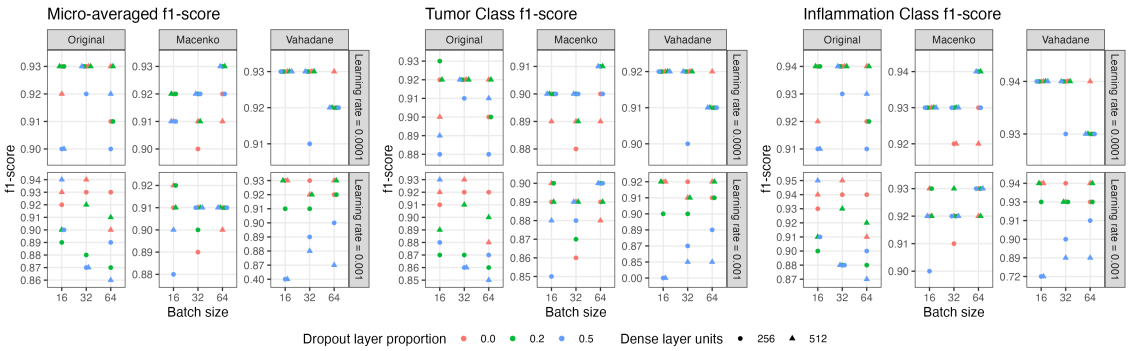


Figure 7.10: Micro-averaged and class-specific f1-scores of the classification task on testing data of patches from Sudmeier et al.

All networks achieved high micro-averaged f1-scores, ranging from 0.86 to 0.94 (Figure 7.10), indicating high precision and recall overall, except for two networks trained using Vahadane normalized images. Furthermore, the class-specific f1-scores for these networks with high micro-averaged f1-scores were also notably high. For the Inflammation class, the f1-scores ranged from 0.87 to 0.95, indicating a strong ability to correctly classify patches of Inflammation. Similarly, for the Tumor class, the f1-scores ranged from 0.85 to 0.93, suggesting a high level of accuracy in identifying Tumor patches. These f1-scores imply the potential of these networks in aiding medical professionals in diagnosing and understanding pathological conditions related to inflammation and tumors.

Finally, we examine the misclassified patches. The testing dataset consists of a total of 153 patches, out of which 67 have been labeled as "Tumor" and 86 have been labeled as "Inflammation" by Sudmeier et al. Further analysis reveals that cluster 4 and 5 account for 24 and 43 Tumor patches, respectively, while cluster 1 and 7 account for 65 and 21 Inflammation patches, respectively.

The classifiers used in this study were trained with different combinations of parameters, resulting in a total of 36 classifiers for each image type. Table E.3 and Table E.4 in Appendix E provide detailed information about the number of times a patch was misclassified in the testing dataset. When considering the classifiers trained with the original images, it was found that out of the 67 patches with an actual label of Tumor in the testing dataset, 17 of them were misclassified by at least one classifier. Out of the 86 patches with an actual label of Inflammation in the testing dataset, only four of them were misclassified by at least one classifier. Notably, Tumor patches labeled with barcodes CCCGTAAGTCTAGGCC-1 and TTGGACATGTGGCTTA-1, both of which were from cluster 5 (i.e., inflammation-adjacent tumor), were

misclassified by all classifiers, regardless of whether they were trained on original or stain-normalized images. This implies that the Tumor patches from cluster 5, which are located near inflammation regions, posed a challenge for the classifiers. On the other hand, it can be seen that none of the Inflammation patches were misclassified by all classifiers trained on either of the three types of images. However, some Inflammation patches were still misclassified by a large number of classifiers trained on each type of image.

	Number of spots misclassified				Number of misclassifications			
	Original	Macenko	Vahadane	Total	Original	Macenko	Vahadane	Total
Inflammation								
Cluster 1	1	9	7	16	9	79	123	211
Cluster 7	3	11	1	14	29	172	34	235
Tumor								
Cluster 4	5	5	24	24	116	99	59	274
Cluster 5	12	5	43	43	322	132	332	786

Table 7.4: Summary of misclassification for patches in each cluster across original, Macenko normalized, and Vahadane normalized images. Each row represents a cluster, and columns show the number of patches misclassified by at least one classifier and total number of misclassifications for patches in that cluster for each image type.

	Number of spots misclassified				Number of misclassifications			
	Original	Macenko	Vahadane	Total	Original	Macenko	Vahadane	Total
Inflammation								
Cluster 1	1	2	4	7	9	65	114	188
Cluster 7	1	8	1	9	17	161	34	212
Tumor								
Cluster 4	3	4	1	6	108	91	11	210
Cluster 5	9	4	8	14	307	129	256	692

Table 7.5: Summary of misclassification for patches in each cluster across original, Macenko normalized, and Vahadane normalized images. Each row represents a cluster, and columns show misclassified count and total number of misclassifications, only for patches misclassified by more than half of the classifiers trained with each image type.

Referring to Table 7.4, the application of the Macenko stain normalization method resulted in an increase in the number of misclassified Inflammation patches in both clusters 1 and 7. After normalization with the Vahadane method, fewer Inflammation patches in cluster 7 were misclassified compared to the original images, while more Inflammation patches in cluster 1 were misclassified. For Tumor patches, the Macenko method increased the number of misclassified patches from cluster 5 while the Vahadane method

was not effective in reducing misclassification for either cluster. It is noteworthy that all Tumor patches were misclassified by at least one classifier after stain normalization by Vahadane method. These results indicate that the effectiveness of stain normalization may vary depending on the type and location of the tissue being analyzed.

By comparing Table 7.4 and Table 7.5, it becomes apparent that a substantial proportion of misclassified Inflammation patches were misclassified by fewer than half of the classifiers trained on either image type. Similarly, among all Tumor patches misclassified by at least one of the classifiers trained on Vahadane normalized images, only nine patches were misclassified by more than half of the classifiers.

We conducted Fisher's exact test to assess the potential association between patch misclassification and the two clusters, i.e., clusters 1 and 7 for Inflammation patches, and clusters 4 and 5 for Tumor patches. Table 7.6 summarizes the results of the test for Inflammation patches across different image types. Specifically, without stain normalization, the odds ratio (OR) of misclassification by at least one classifier was 0.097 with a p-value of 0.044, but the association was no longer significant when examining the odds of misclassification by more than half of the classifiers. In contrast, after stain normalization by Macenko, the association between misclassification of Inflammation patches and the two clusters was significant regardless of the criteria for misclassification. Macenko normalized Inflammation patches from cluster 7 were more likely to be misclassified than those from cluster 1. When Inflammation patches were stain normalized by the Vahadane method, the association between the misclassification and the two clusters was not significant regardless of the criteria for misclassification.

	Misclassified by at least one classifier			Misclassified by more than half of classifiers		
	Odds Ratio	95% CI	p-value	Odds Ratio	95% CI	p-value
Original	0.097	(0.002,1.291)	0.044	0.318	(0.004,25.747)	0.431
Macenko	0.151	(0.042,0.51)	0.001	0.054	(0.005,0.313)	0.000
Vahadane	2.394	(0.278,114.092)	0.673	1.308	(0.12,67.753)	1.000

Table 7.6: Results of the Fisher's exact test for association between patch misclassification and cluster for Inflammation patches (Cluster 1 vs Cluster 7), for classifiers trained on each image type. Misclassification is defined in two ways, one is the patch being misclassified by at least one classifier, and another is the patch being misclassified by more than half of the classifiers.

Similarly, Table 7.7 summarized the results of the test for Tumor patches across different image types. Regardless of the image type and the criteria for misclassification, the association between the misclassifi-

cation of Tumor patches and the two clusters was always not significant.

	Misclassified by at least one classifier			Misclassified by more than half of classifiers		
	Odds Ratio	95% CI	p-value	Odds Ratio	95% CI	p-value
Original	0.684	(0.162,2.511)	0.574	0.544	(0.085,2.517)	0.515
Macenko	1.978	(0.402,9.781)	0.476	1.930	(0.323,11.546)	0.443
Vahadane	0.000	(0,Inf)	1.000	0.194	(0.004,1.613)	0.142

Table 7.7: Results of the Fisher's exact test for association between patch misclassification and cluster for Tumor patches (Cluster 4 vs Cluster 5), for classifiers trained on each image type. Misclassification is defined in two ways, one is the patch being misclassified by at least one classifier, and another is the patch being misclassified by more than half of the classifiers.

Chapter 8

Discussion

This chapter centers around the implications arising from our analysis results. We place particular emphasis on noteworthy findings, elucidating their potential influence on the field. Additionally, we conduct a thorough evaluation of the limitations of STApath and its underlying technologies, recognizing the challenges and proposing possible avenues for improvement. Lastly, we outline future research directions, identifying promising areas for further exploration and advancement in the realm of digital pathology.

8.1 Learning Curve

8.1.1 Validation Loss Lower Than Training Loss

For some trained networks, rather than experiencing overfitting, we observed that the validation loss and metrics were consistently lower than their corresponding training loss and metrics. Over time, the gap between the training loss and validation loss either decreased or remained constant. It is possibly caused by the networks being over-regularized. Regularization methods often sacrifice training accuracy to improve validation accuracy. Dropout is a regularization technique that helps reduce overfitting by randomly dropping out (freezing) neurons during training. As dropout is only applied during training, it affects only the training loss. During validation, all neurons are activated, leading to a generally lower loss on this dataset compared to the training dataset since the network has not experienced dropout during evaluation. However, it is worth noting that higher training loss does not necessarily indicate the network is less accurate on the

training set than on the validation set.

In some cases, the validation loss of a neural network is initially lower than the training loss during the early stages of training. However, as the training progresses, the validation loss becomes similar to or higher than the training loss. This phenomenon can be explained by how we calculate the training and validation losses. An epoch is a complete pass of the entire training dataset through the neural network during training. The training data are split into smaller batches, and each batch is used to update the network's parameters through backpropagation. Thus, an epoch consists of multiple steps of backpropagation. In the first few epochs, the network's parameters are relatively untrained with the training data. Thus, during each backpropagation step, the network's generalization ability can improve significantly. The training loss is the average of the losses for all batches of training data over the current epoch. Since the model is still learning, the loss over the first few batches of an epoch is generally higher than over the last few batches. On the other hand, the validation loss for an epoch is computed after the entire epoch is finished when the network has undergone more training. Under this framework, the validation error has the benefit of being fully updated, while the training error averages over error calculations with fewer updates. As a result, the validation loss tends to be lower because the network has learned from more data and updated its parameters accordingly. Therefore, the network gives a lower validation loss in the first few epochs due to the significant improvements made during each backpropagation step. However, as the training continues, the magnitude of parameter updates becomes smaller, and the validation loss may stabilize or even increase due to overfitting or convergence. This explains why the validation loss can be lower initially but catch up or surpass the training loss later on in some cases.

8.1.2 Fluctuation in the Validation Loss and Metrics

For some networks, the validation loss and metrics have much larger fluctuation than the training loss and metrics. The most likely reason is that the validation set is small and thus average of loss and metrics has relatively large variation. In such circumstances, a change in network parameters after an epoch has a more visible impact on the validation loss and metrics.

8.2 Input Imaging Data

8.2.1 Standard Patches Versus Large Patches

We tested and compared the networks' performance when the image patches used contained either one capture spot or eight capture spots using samples from Wu et al. [2021]. The training dataset for large patches and the training dataset for standard patches had similar cell-type compositions. This is also true for validation and testing sets. When the networks predicted the proportions of different cell types, the correlation between the predicted proportions and the observed proportions was in a similar order for each cell type, regardless of whether the data came from standard patches or large patches. When considering only the proportions predicted by the network that achieved the lowest validation loss for each task, the correlations between the predicted proportions and the observed proportions were slightly higher for all cell types when the network was trained with large patches. This suggests that using large patches during training might have a slight advantage in predicting cell type proportions.

8.2.2 Stain Normalization

In general, the stain normalization techniques did not improve the neural networks' predictive performance in both regression and classification tasks. In fact, compared to no normalization or the Macenko normalization, Vahadane normalization had a negative impact on the networks' performance in regression tasks as networks trained with Vahadane normalized images generally had higher validation loss than those trained without stain normalization or with the Macenko method. Both stain normalization methods used, Macenko and Vahadane, rely on a reference image to estimate stain parameters, but it is hard for one reference image to cover all staining phenomena or represent all input images, which usually causes misestimation of stain parameters and thus delivers inaccurate normalization results (Zheng et al. [2021]; Zhou et al. [2019]).

8.3 Network Architecture

8.3.1 Batch Size

The batch size determines the number of patches in the training data processed collectively before updating the model's parameters during each epoch of training. Larger batch sizes tend to provide more stable updates as they average gradients over more patches. Smaller batch sizes, on the other hand, offer more frequent updates and can potentially lead to better generalization as they introduce more randomness into the training process. It is worth noting that the impact of batch size on the neural network's performance can vary based on its interaction with multiple factors, such as the size of training data, other network architecture hyperparameters, and learning rate. In most of the regression tasks examined in our study, the impact of batch size on the validation loss at the optimal epoch did not exhibit consistent patterns when accounting for other architecture hyperparameters. However, for specific regression tasks and specific combinations of architecture hyperparameters, we observed increasing or decreasing trends in the validation loss as the batch size increased. These trends were more frequently observed when the regression tasks involved predicting nine cell type proportions compared to predicting four cell type proportions.

For instance, when the networks aimed to predict the nine cell type proportions using patches from the 10X Genomics FFPE breast tissue, an increasing batch size was associated with a decreasing validation loss only when the networks were optimized using the Adam algorithm with a learning rate of 0.001. Besides, when the networks aimed to predict the nine cell type proportions using patches from the 10X Genomics fresh frozen breast tissue, an increasing batch size was associated with an increasing validation loss only when the networks were optimized using the SGD algorithm with a learning rate of 0.01. More pronounced trends were observed when the networks utilized large patches from WSIs of samples from Wu et al. [2021] to predict cell type proportions. As the batch size increased, the validation loss also increased for the networks optimized using Adam with a learning rate of 0.0001 or optimized using SGD with learning rates of 0.01 and 0.001.

8.3.2 Hidden Dense Layer Units

The number of neurons in the dense layer of the hidden layer determines the network's ability to learn complex patterns from the data. A larger number of neurons may capture more complex patterns in the data, but it may also increase the risk of overfitting, particularly when the dataset is limited. In our study, the number of neurons in the hidden dense layer, either 256 or 512, did not show a consistent association with the validation loss. However, in the regression task of predicting nine cell types using large patches from Wu et al. [2021], we observed that networks without a dense layer exhibited higher validation loss at the optimal epochs compared to networks that included a dense layer with the same other hyperparameters.

8.3.3 Dropout Layer Proportion

The dropout layer is a regularization technique to prevent overfitting by randomly setting a proportion of neurons to zero during each training epoch. The dropout rate represents the probability of a neuron being dropped out. It introduces stochasticity and makes the network avoid relying too heavily on any particular set of neurons.

For many tasks in our study, we observed an association between the dropout layer proportion and the validation loss at the optimal epoch while keeping other hyperparameters the same. The associations were more frequently observed when the networks were optimized using the Adam algorithm with a learning rate of 0.001. In general, networks with a dropout layer proportion of 0.5 had higher validation loss than those with a dropout layer proportion of 0.2 or without a dropout layer while keeping other hyperparameters the same. Including a dropout layer with a proportion of 0.2 also increased the validation loss compared to networks without a dropout layer, but with occasional exceptions.

The dropout layer proportion played a crucial role in determining the model's performance in various tasks. Strong regularization may not always be appropriate for certain jobs. In order to successfully balance regularization and avoid overfitting in the neural network during training, experimenting with different dropout rates and other hyperparameters can help identify the most suitable configuration for each task.

8.3.4 Optimizer and Learning Rate

During the training process, an optimization algorithm is used to minimize the loss function of the neural network. The optimization process updates the network's parameters (weights and biases) at each epoch to find the optimal set of parameters that minimize the loss function, thereby making the network's predictions as accurate as possible. The learning rate plays a crucial role as it determines the step size at which the parameters are adjusted during optimization. A higher learning rate causes larger steps during training. In some cases, this can enable faster convergence and possibly achieve lower validation loss at the optimal epochs. However, it can also cause overshooting, leading to unstable behavior and potentially higher validation loss. Overall, the behavior of the learning rates in the optimizers influences how quickly the model converges and how close it gets to the optimal solution during training. The choice of an appropriate optimizer with an appropriate learning rate for a specific task and architecture is essential to achieve the best performance and stability during the training process.

The SGD algorithm updates the network's parameters based on the gradient of the loss function computed on a single or a small subset of training samples at each epoch and the learning rate. Different from the SGD algorithm that uses a fixed learning rate for all parameters throughout training, Adam optimizer employed an adaptive learning rate algorithm, which adapts the learning rates of each parameter based on past gradients, allowing it to converge quickly and efficiently, especially in the case of noisy data and sparse gradients.

Based on the learning curves obtained from all tasks conducted on the samples in our study, it is clear that using the Adam optimizer with a learning rate of 0.01 frequently resulted in early stopping of the training process, usually immediately after the first 30 epochs. We also observed similar early stopping for certain tasks when the network was trained with the Adam optimizer and a learning rate of 0.001. This is because, with a high learning rate, the optimization process takes larger steps in the parameter space, potentially overshooting the optimal solution and hence hindering further progress. As a result, the network is not able to reach its true optimal solution, leading to suboptimal performance. In contrast, setting the learning rate to 0.0001 allowed for a successful training process, indicating that a lower learning rate contributed to more stable and effective optimization when the Adam optimizer was used. On the other hand, when employing the SGD optimizer, early stopping was not commonly observed at a very early stage for most

tasks, regardless of the learning rate used. In fact, most of the networks underwent at least 200 epochs of training. The SGD optimizer updates the model's parameters based on the average gradient computed from a randomly selected mini-batch of data. Thus, the SGD optimizer navigates through different parts of the loss landscape and explores various paths toward the optimal solution. This stochastic nature of the SGD optimizer, along with the fixed learning rate, allows the networks to explore the parameter space more extensively and continue training for a longer period without prematurely converging to a suboptimal state.

Besides, when we used the same architecture for the neural networks with the Adam optimizer, we observed that a smaller learning rate of 0.0001 generally resulted in lower validation loss at the optimal epoch compared to using a learning rate of 0.001. This phenomenon can be attributed to the fact that the Adam optimizer takes smaller steps during training with a lower learning rate, allowing for more careful fine-tuning of the model's parameters. Consequently, this finer adjustment often resulted in achieving a lower validation loss at the optimal epoch, as the optimization process was less prone to overshooting the optimal solution. On the contrary, networks with a higher learning rate tended to have lower validation loss at the optimal epochs when the SGD optimizer was applied to the same network architecture. This observation can be explained by the larger steps taken in the parameter space due to the increased learning rate. These larger steps enable the SGD optimizer to avoid getting stuck in local minima and instead explore different regions of the loss landscape, potentially finding a more favorable optimal point that results in lower validation loss.

8.4 Predicting Nine Cell Types Versus Predicting Four Cell Types

We took into account that predicting nine cell type proportions might be challenging for the network to achieve accurate results. This difficulty arises because some cell types may have similar morphology, making it harder for the network to distinguish between them effectively. As a solution, we designed specific networks to predict only four cell types: invasive cancer, lymphocyte, stroma, and others. These four cell types are more distinguishable morphologically, and they provide stronger signals for the network to work with. By focusing on these four distinct cell types, we aimed to improve the network's ability to make more accurate predictions.

However, after conducting a comparison between the four proportions obtained by collapsing the predic-

tions of the nine cell types made by the network and the four proportions directly predicted by the network, we did not find a substantial improvement in the correlation between the predicted and observed proportions. This implies that even though we simplified the task by focusing on predicting only four cell types instead of nine, this modification did not lead to a significant enhancement in the accuracy of the network's predictions. Consequently, if the objective is to predict the proportions of the four pathological cell types, it is inconsequential whether the network is trained with nine cell type proportions or four cell type proportions.

8.5 Limitations and Future Direction

8.5.1 Unable to Predict Extreme Values

In all regression tasks, we noticed that the predicted proportions tended to be more tightly clustered than the observed proportions. While the mean values of the predicted and observed proportions were quite similar, the median values showed distinct differences. This suggests that the networks tended to estimate the mean proportion for each cell type within the testing dataset rather than precisely predicting the proportions for individual patches. This tendency was particularly evident for cell types with poor R-squared values between the predicted and observed proportions.

This observation can potentially be rationalized by examining the characteristics of the training dataset. We found that in many cases a significant majority of patches in the training dataset had similar proportions for certain cell types, leading to a potential limitation in capturing the intricate features associated with other proportions of these cell types. Due to insufficient exposure to information relevant to uncommon proportions of certain cell types during the training progress, the trained neural network encountered a lack of experience when trying to perform prediction on a new patch with uncommon proportions of these cell types, leading to inaccurate predictions. As a result, the trained network found itself unable to make predictions better than random guesses or only a little bit better than random guesses. Consequently, instead of making random guesses, the network took a conservative approach by predicting mean values in the absence of specific guidance. Thus, the lack of diversity in the cell type proportions within the training data can be a potential explanation for this problem. Using more complete training data that cover the full range of inputs that the network is expected to handle can be helpful. In scenarios where samples exhibit high

purity of specific cell types, employing penalized regressions on features extracted by pre-trained ResNets has the potential to yield more accurate predictions of cell type proportions compared to making predictions by the trained neural networks. In addition to the lack of diversity in cell type proportions in the training data, the predictive performance of networks on different cell types can also be affected by the relative abundance of those cell types in the training dataset. When certain cell types had very low proportions in most patches of the training data, their unique characteristics or morphologies had limited representation, causing neural networks to prioritize features and patterns that are more commonly and obviously found in the data.

Another potential explanation is the choice of the loss function. Here, we let Y be the observed outcome and $\hat{Y}$ be the predicted outcome given by the neural network. The network's loss function is the MSE averaged over all input patches in the training dataset $\mathcal{D}$, i.e., MSE $= \mathbb{E}_{\mathcal{D}}[(Y - \hat{Y})^2]$. The network adjusts the probability distribution of $\hat{Y}$ so that it minimizes the MSE. Since MSE is non-negative, the theoretical optimal $\hat{Y}$ is when the probability distribution of $\hat{Y}$ is identical with that of Y, and hence giving MSE $= 0$. We know that, for any continuous variable Y, $\mathbb{E}[(Y-a)^2] \geq \mathrm{Var}(Y)$ with equality holding when $a = \mathbb{E}[Y]$.

If all observed proportions were predicted to be the mean values, i.e., $\hat{Y} = \mathbb{E}_{\mathcal{D}}[Y]$, the MSE is

$$\mathbb{E}[(Y - \hat{Y})^2] = \mathbb{E}[(Y - \mathbb{E}_{\mathcal{D}}[Y])^2] = \mathrm{Var}(Y) \geq 0,$$

the squared deviation of $\hat{Y}$ is

$$(Y - \hat{Y})^2 = (Y - \mathbb{E}_{\mathcal{D}}[Y])^2 = \text{Bias}^2(Y),$$

and the variance of $\hat{Y}$ is

$$\mathbb{E}_{\mathcal{D}}\left[(\hat{Y} - \mathbb{E}_{\mathcal{D}}[\hat{Y}])^2\right] = \mathbb{E}_{\mathcal{D}}\left[(\mathbb{E}_{\mathcal{D}}[Y] - \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\mathcal{D}}[Y]])^2\right] = 0.$$

The predicted proportions have too small a variance and too large a deviation from the observed proportions.

It was observed that our networks tended to predict the mean of Y, meaning that the MSE values were close to the variance of observed proportions, $\mathrm{Var}(Y)$. Although the MSE values were small with $\hat{Y} =$

$\mathbb{E}_{\mathcal{D}}[Y]$, they were not small enough. We want the network to further decrease the MSE such that $0 \leq$ MSE $< \mathrm{Var}(Y)$. This can be done by either adding more parameters to the network or adjusting the complexity of the network. We can also adjust the loss function. Our current loss function is

$$\mathcal{L}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2,$$

where N is the size of the dataset. We can modify it to penalize for the difference between y_i and $\hat{y}_i$ being larger than the difference between $\hat{y}_i$ and $\mathbb{E}_{\mathcal{D}}[Y]$.

8.5.2 Batch Effect

The batch effect is a significant and widespread factor that leads to variations in high-throughput experiments. It refers to sub-groups of measurements that demonstrate different behaviors across conditions, and importantly, these differences are unrelated to the biological or scientific variables being studied (Leek et al. [2010]). The batch effect caused by intrinsic differences between different samples was observed in our study. When attempting regression tasks that merged data from multiple samples to train a network, substantial variations in the network's performance are observed between different samples. Moreover, an interesting observation is that the performance of sample-specific networks, where each network is trained with data from a single sample separately, showed similarity to the performance of the networks trained using the merged data on that specific sample. These observations highlighted the importance of addressing and accounting for the batch effect to ensure reliable and accurate performance of the trained networks.

The observed batch effect in our study is likely a result of the batch effect present in the scRNA-seq data used by cell type deconvolution. The CARD algorithm relies heavily on reference data from scRNA-seq data with cell type annotations, from which it learns cell type-specific gene expression to estimate the proportions of different cell types at every capture spot in ST data. Specifically, batch effects in scRNA-seq experiments occur when cells from one biological group or condition are processed, captured, and sequenced separately from cells in another condition (Hicks et al. [2017]). We utilized the scRNA-seq data provided by Wu et al. [2021] as the reference for cell type deconvolution. They performed scRNA-Seq on 26 primary tumors, representing three major clinical subtypes of breast cancer: 11 ER+, 5 HER2+, and 10 TNBC samples.

In future studies, it is imperative to address the batch effect issue when performing cell type decon-

volution using scRNA-seq data. We can employ reference-free or batch-effect normalized deconvolution methods. Possible methods include RCTD (Cable et al. [2021]) and cell2location (Kleshchevnikov et al. [2022]), which have integrated and validated batch effect normalization procedures within their Bayesian models. Furthermore, BayesSpace (Zhao et al. [2021]), a fully Bayesian statistical method of resolution enhancement, offers cell type deconvolution at a subspot level without relying on independent scRNA-seq reference data. Moreover, this method can be applied to create clusters based on the ST data for classification tasks, enhancing the utility of ST data in studying cellular compositions and interactions.

Another approach is to remove the potential batch effect from the scRNA-seq data prior to using reference-based deconvolution algorithms like CARD. Several batch effect removal techniques have been proposed. Harmony (Korsunsky et al. [2019]) is an algorithm that projects cells into a shared embedding, grouping them by cell type rather than dataset-specific conditions. It employs PCA for dimensionality reduction and iteratively eliminates batch effects from the PCA space. Another method, Seurat Integration (Seurat 3) (Stuart et al. [2019]), utilizes a combination of canonical correlation analysis (CCA) and mutual nearest neighbors (MNN) approach to project the data into a subspace and capture the most correlated data features for batch alignment. Besides, DESC (Li et al. [2020]), an unsupervised deep embedding algorithm, gradually removes the batch effects in the scRNA-seq data through iterative self-learning. These techniques offer effective approaches to mitigate batch effects.

The poor performance of the network may also be caused by the inaccuracy in the cell type deconvolution results. Therefore, we can employ cell type deconvolution techniques other than CARD. For the classification tasks, instead of using the qualitative labels provided by the authors, we can create them by ourselves based on the cell type deconvolution results we obtained.

8.5.3 Incorporate Methods for Compositional Data Analysis

The response variables of interest in our regression tasks are cell type proportions, which can be considered as compositional data. For compositional data the appropriate sample space is the positive simplex:

$$\mathbb{S}^{D-1} = \left\{ (y_1, ..., y_D)^T, y_i \geq 0, \sum_{i=1}^{D} y_i = 1 \right\},$$

where D is the number of variables, better known as components. Here, our regression tasks had either nine or four components. We denote $\mathbf{y}$ as a matrix of size $N \times D$ representing the cell type composition of N patches, where each row of $\mathbf{y}$ corresponds to the proportions of the D cell types within each patch. In other words, the response variables of STApath regression tasks are denoted as $\mathbf{y} = (\mathbf{y}_1, , ..., \mathbf{y}_N)$, with $\mathbf{y}_i = (y_{i,1}, , ..., y_{i,D})^T \in \mathbb{S}^{D-1}$ being the vector of cell type proportions for patch i where each element $y_{i,d}$ represents the proportion of cell type d in patch i.

We utilized the softmax activation function in the final output layer of our neural networks to enable the transformation of the scores predicted by the layers preceding the output layer, denoted by vector $\mathbf{s}_i$, into a probability distribution, denoted by vector $\mathbf{y}_i$, such that each component of the output falls within the interval (0, 1) and that the sum of all components equals 1. The softmax function is mathematically defined as

$$\mathbf{y}_i = \sigma(\mathbf{s}_i) = \frac{e^{s_i}}{\sum_{j=1}^{D} e^{s_j}} \text{ for } i = 1, ..., D \text{ and } \mathbf{s}_i = (s_1, .., s_D).$$

The use of the softmax activation function to normalize predicted scores in the output layer may not be the optimal approach when dealing with compositional data. Since the the cell type proportions of the training data are not symmetrically distributed and had the compositional data constraints, using the MSE as a loss function may not be suitable for optimizing neural networks. To address this issue, two potential approaches can be considered. Firstly, we can use specific loss functions that consider the inherent structure of compositional data and can provide more accurate guidance to the optimization algorithm during training. Alternatively, we can preprocess the data appropriately to handle the compositional nature. The neural network can be trained on the transformed response variables, obtained by applying methods like additive log-ratio (ALR) or centered log-ratio (CLR) transformation to the cell type proportions. During training, the network utilizes MSE loss function and a linear activation function in the output layer.

The centered log ratio (CLR) transformation, proposed by AITCHISON [1983], allows for the escape of unit sum constraint of compositional data by mapping data onto the Euclidean space, $\mathbb{R}^{D-1}$. The CLR-transformed compositional data for patch i is denoted as $\mathbf{c}_i = \text{CLR}(\mathbf{y}_i) = (c_{i,1}, , ..., c_{i,D})^T$, with d-th element being

$$c_{i,d} = \log\left(\frac{y_{i,d}}{g(\mathbf{y}_i)}\right) = \log\left(\frac{y_{i,d}}{(\prod_d y_{i,d})^{1/D}}\right) \text{ for } d = 1, ..., D,$$

where $g(\mathbf{y}_i)$ is the geometric mean of $\mathbf{y}_i$. The isometric log-ratio (ILR) transformation, introduced by Egozcue [2003], offers an alternative method that can be used. This transformation involves orthogonalizing the basis of the Euclidean space, which is constructed from the centered log-ratio (CLR)-transformed compositional data. Mathematically, the ILR transformation is defined as $\mathbf{z}_i = \text{ILR}(\mathbf{y}_i) = \mathbf{c}_i \Psi^T$, where Ψ is a $D \times D$ orthonormal matrix with its first row deleted (Egozcue and Pawlowsky-Glahn [2005]). However, log-ratio based transformations, including CLR and ILR, have a significant limitation associated with their reliance on log-ratio quantities. If any of the observed components in the compositional data are zero, both CLR and ILR transformations cannot be executed, leading to potential issues in the analysis. Thus, these transformation techniques cannot be applied to tissues with relatively high purity of certain cell types, such as the 10X Genomics fresh frozen tissue.

8.5.4 Mitigating the Adverse Effects of Noisy Labels

The good performance of neural networks is dependent on the reliability of labels. Unreliable labels are called noisy labels because they may be corrupted from ground-truth labels. However, in our study, the quality of data labels is a concern. During the training of a network, a mini-batch of patches is randomly sampled from the noisy training dataset at a specific time t. The network's parameters at time t are then updated based on the empirical risk computed on this mini-batch. Due to the presence of noisy labels in the mini-batch, the optimization process lacks tolerance for noise labels. Consequently, the network can easily memorize these noisy labels, leading to a degradation in its ability to generalize on unseen data.

Considering the inevitability of some label errors, in addition to exploring label improvement methods as discussed earlier, it is essential to address the negative impact of noisy labels by making the supervised learning process more robust to label noise. There are various methods that can be employed to enhance the robustness of deep learning networks.

Firstly, to improve the model's handling of label noise, we can either introduce a noise adaptation layer on top of the softmax layer (Goldberger and Ben-Reuven [2016]; Bekker and Goldberger [2016]), enabling it to learn the label transition process and adjust the network's output using estimated label transition probabilities, or design a dedicated architecture that can effectively accommodate various types of label noise (Yao et al. [2019]). Additionally, advanced regularization techniques have been proposed (Wei et al. [2021];

Xia et al. [2021]; Menon et al. [2020]), leading to further improvements in the model's robustness to label noise when combined with conventional methods like data augmentation, dropout, and batch normalization. The key advantage of these techniques is that they can easily collaborate with other approaches by making simple modifications. Employing robust loss functions is another option such that they achieve a small risk for unseen clean data even when noisy labels exist in the training data (Ma et al. [2020]; Lyu and Tsang [2020]). Finally, we can perform sample selection that involves updating the network only for the selected clean instances while excluding the rest of the mini-batch instances that are likely to be falsely labeled (Jiang et al. [2018]; Han et al. [2018]; Yu et al. [2019]; Zhou et al. [2021]; Wei et al. [2020]; Li et al. [2019]).

8.5.5 Processing of the Predicted Response Variables

The initial attempt to predict spot-level quantitative cellular patterns was found to be inaccurate. However, there is a promising direction to enhance the accuracy by integrating spatial information from neighboring spots/patches. Intuitively, cell type compositions in two neighboring locations (patches) of a tissue are expected to exhibit higher similarity compared to those in distant locations. Thus, the cell type compositions of nearby patches hold valuable information that can aid in predicting the cell type composition at the target location. Thus, to improve predictions, we can explore methods that model and leverage this spatial correlation among the rows of the predicted cell type composition matrix $\hat{\mathbf{y}}$. We anticipate promising alternatives that incorporate principles and methods of Gaussian process regression. Specifically, we want to express the proportion of cell type d in patch i, $y_{i,d}$, as a weighted summation of the proportion of cell type d in all other patches, $y_{j,d}$ ($j \neq i$), where the weights are based on the Euclidean distance between pairs of patches.

In classification tasks, we can adopt a two-step approach, starting with patch-level classification followed by whole-slide classification. After we feed each patch in the testing data into the trained network that classifies each patch as one of the target classes, we will able to output a heatmap indicating the categorization of each patch in the tissue section.

Bibliography

J. AITCHISON. 1983. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65.

Noemi Andor, Trevor A Graham, Marnix Jansen, Li C Xia, C Athena Aktipis, Claudia Petritsch, Hanlee P Ji, and Carlo C Maley. 2016. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.*, 22(1):105–113.

Michaela Asp, Stefania Giacomello, Ludvig Larsson, Chenglin Wu, Daniel Fürth, Xiaoyan Qian, Eva Wärdell, Joaquin Custodio, Johan Reimegård, Fredrik Salmén, et al. 2019. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7):1647–1660.

Jocelyn Barker, Assaf Hoogi, Adrien Depeursinge, and Daniel L. Rubin. 2016. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. *Medical Image Analysis*, 30:60–71.

Alan Joseph Bekker and Jacob Goldberger. 2016. Training deep neural-networks based on unreliable labels. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2682–2686.

Kaustav Bera, Kurt A. Schalper, David L. Rimm, Vamsidhar Velcheti, and Anant Madabhushi. 2019. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 16(11):703–715.

Emelie Berglund, Jonas Maaskola, Niklas Schultz, Stefanie Friedrich, Maja Marklund, Joseph Bergenstråhle, Firas Tarish, Anna Tanoglidi, Sanja Vickovic, Ludvig Larsson, Fredrik Salmén, Christoph Ogris, Karolina Wallenborg, Jens Lagergren, Patrik Ståhl, Erik Sonnhammer, Thomas Helleday, and Joakim

Lundeberg. 2018. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.*, 9(1):2419.

Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. 2021. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526.

Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. 2015. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233).

Mingyu Chen, Bin Zhang, Win Topatana, Jiasheng Cao, Hepan Zhu, Sarun Juengpanich, Qijiang Mao, Hong Yu, and Xiujun Cai. 2020a. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *npj Precision Oncology*, 4(1).

Wei-Ting Chen, Ashley Lu, Katleen Craessaerts, Benjamin Pavie, Carlo Sala Frigerio, Nikky Corthout, Xiaoyan Qian, Jana Laláková, Malte Kühnemund, Iryna Voytyuk, Leen Wolfs, Renzo Mancuso, Evgenia Salta, Sriram Balusu, An Snellinx, Sebastian Munck, Aleksandra Jurek, Jose Fernandez Navarro, Takaomi C Saido, Inge Huitinga, Joakim Lundeberg, Mark Fiers, and Bart De Strooper. 2020b. Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182(4):976–991.e19.

Xiaoyin Chen, Yu-Chi Sun, George M Church, Je Hyuk Lee, and Anthony M Zador. 2017. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Research*, 46(4):e22–e22.

Simone Codeluppi, Lars E Borm, Amit Zeisel, Gioele La Manno, Josina A van Lunteren, Camilla I Svensson, and Sten Linnarsson. 2018. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods*, 15(11):932–935.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567.

J. J. Egozcue. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

J. J. Egozcue and V. Pawlowsky-Glahn. 2005. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828.

Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. 2019. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239.

Elizabeth H Finn and Tom Misteli. 2019. Molecular basis and biological function of variability in spatial genome organization. *Science*, 365(6457):eaaw9498.

Stefanie Friedrich and Erik L L Sonnhammer. 2020. Fusion transcript detection using spatial transcriptomics. *BMC Med. Genomics*, 13(1):110.

Jacob Goldberger and Ehud Ben-Reuven. 2016. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. 2020. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. 2017. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578.

Michalina Janiszewska. 2020. The microcosmos of intratumor heterogeneity: the space-time of cancer evolution. *Oncogene*, 39(10):2031–2039.

Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G

Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstråhle, et al. 2020. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Fei-Fei Li. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *International Conference on Machine Learning*, pages 2304–2313.

Jocelyn Y Kishi, Sylvain W Lapan, Brian J Beliveau, Emma R West, Allen Zhu, Hiroshi M Sasaki, Sinem K Saka, Yu Wang, Constance L Cepko, and Peng Yin. 2019. SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nat. Methods*, 16(6):533–544.

Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. 2022. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40(5):661–671.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. 2019. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16:1289–1296.

Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M Church. 2015. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.*, 10(3):442–458.

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739.

Junnan Li, Richard Socher, and Steven CH Hoi. 2019. Dividemix: Learning with noisy labels as semi-supervised learning. *International Conference on Learning Representations*.

Xiangjie Li, Kui Wang, Yafei Lyu, Huize Pan, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Muredach P Reilly, Gang Hu, and Mingyao Li. 2020. Deep learning enables accurate clustering with batch effect removal in single-cell rna-seq analysis. *Nature communications*, 11(1):1–14.

Jana Lipkova, Tiffany Y Chen, Ming Y Lu, Richard J Chen, Maha Shady, Mane Williams, Jingwen Wang, Zahra Noor, Richard N Mitchell, Mehmet Turan, et al. 2022. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. *Nature medicine*, 28(3):575–582.

Yueming Lyu and Ivor W. Tsang. 2020. Curriculum loss: Robust learning and generalization against label corruption.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah M. Erfani, and James Bailey. 2020. Normalized loss functions for deep learning with noisy labels. *CoRR*, abs/2006.13554.

Ying Ma and Xiang Zhou. 2022. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*, pages 1–11.

Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. 2009. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. 2020. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*.

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

Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, et al. 2019. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1):1–10.

Liron Pantanowitz, John H. Sinard, Walter H. Henricks, Lisa A. Fatheree, Alexis B. Carter, Lydia Contis, Bruce A. Beckwith, Andrew J. Evans, Avtar Lal, and Anil V. Parwani. 2013. Validating whole slide imaging for diagnostic purposes in pathology: Guideline from the college of american pathologists pathology and laboratory quality center. *Archives of Pathology & Laboratory Medicine*, 137(12):1710–1722.

Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. 2019. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467.

Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. 2018. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193.

Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun, and Jingyi Jessica Li. 2023. scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*.

Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. 2021. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813.

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.

Lisa J Sudmeier, Kimberly B Hoang, Edjah K Nduom, Andreas Wieland, Stewart G Neill, Matthew J Schniederjan, Suresh S Ramalingam, Jeffrey J Olson, Rafi Ahmed, and William H Hudson. 2022. Distinct phenotypic states and spatial distribution of CD8+ T cell clonotypes in human brain metastases. *Cell Rep. Med.*, 3(5):100620.

Kim Thrane, Hanna Eriksson, Jonas Maaskola, Johan Hansson, and Joakim Lundeberg. 2018. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.*, 78(20):5970–5979.

Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. 2016. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971.

Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, Gabriel K Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L Ståhl. 2019. High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods*, 16(10):987–990.

Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P Nolan, Felice-Alessio Bava, and Karl Deisseroth. 2018. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400).

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13723–13732.

Hongxin Wei, Lue Tao, Renchunzi Xie, and Bo An. 2021. Open-set label noise can improve robustness against inherent label noise. *CoRR*, abs/2106.10891.

Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. 2021. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347.

Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. 2021. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*.

Jiangchao Yao, Jiajie Wang, Ivor W. Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang. 2019. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28(4):1909–1922.

Niyaz Yoosuf, José Fernández Navarro, Fredrik Salmén, Patrik L Ståhl, and Carsten O Daub. 2020. Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. *Breast Cancer Res.*, 22(1):6.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? *International Conference on Machine Learning*, pages 7164–7173.

Mark D. Zarella, Douglas Bowman, Famke Aeffner, Navid Farahani, Albert Xthona, Syeda Fatima Absar, Anil Parwani, Marilyn Bui, and Douglas J. Hartman. 2018. A practical guide to whole slide imaging: A white paper from the digital pathology association. *Archives of Pathology & Laboratory Medicine*, 143(2):222–234.

Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam, Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, Jason H. Bielas, and Raphael Gottardo. 2021. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39(11):1375–1384.

Yushan Zheng, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Dingyi Hu, Shujiao Sun, Jun Shi, and Chenghai Xue. 2021. Stain standardization capsule for application-driven histopathological image normalization. *IEEE Journal of Biomedical and Health Informatics*, 25(2):337–347.

Niyun Zhou, De Cai, Xiao Han, and Jianhua Yao. 2019. Enhanced cycle-consistent generative adversarial network for color normalization of H&E stained images. In *Lecture Notes in Computer Science*, pages 694–702. Springer International Publishing.

Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. 2021. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*.

Chapter A

Appendix for Chapter 3

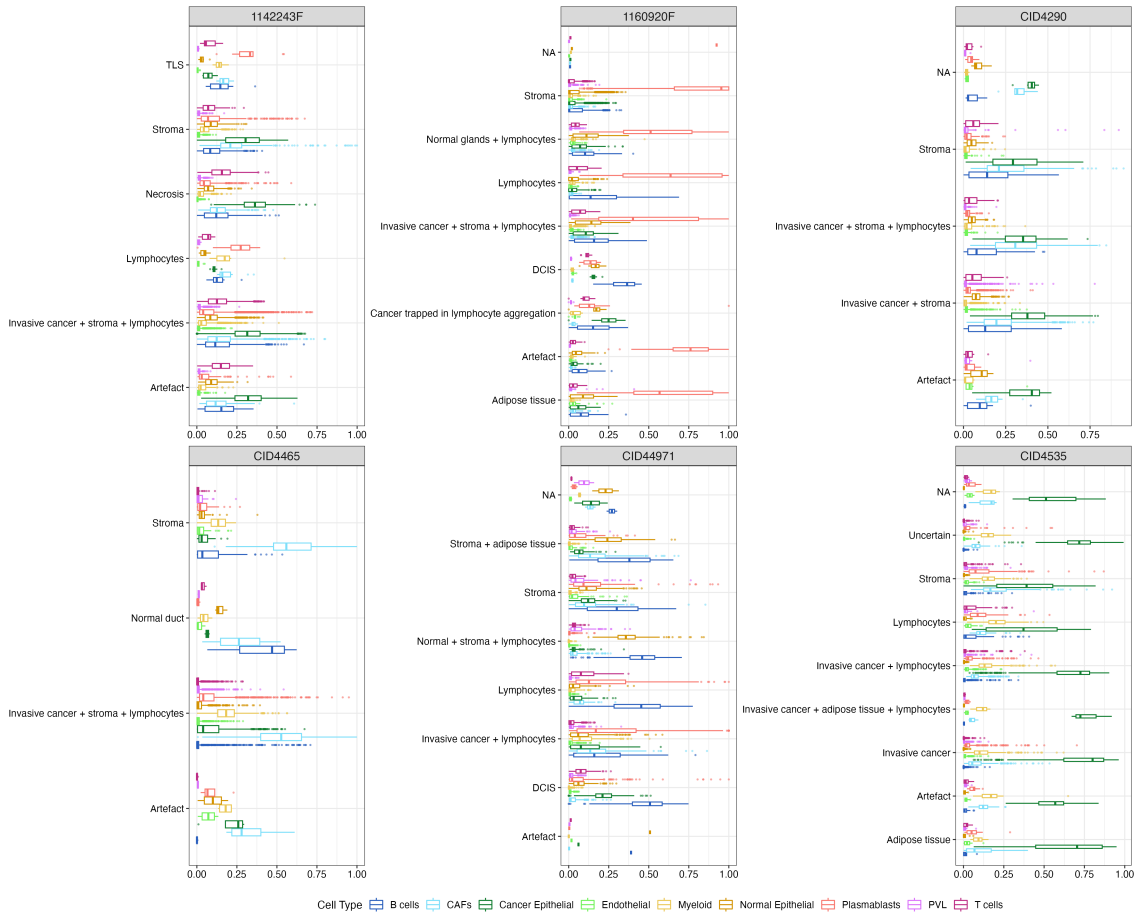


Figure A.1: Composition of each of the nine major cell types among patches with each pathology annotation, stratified by sample.

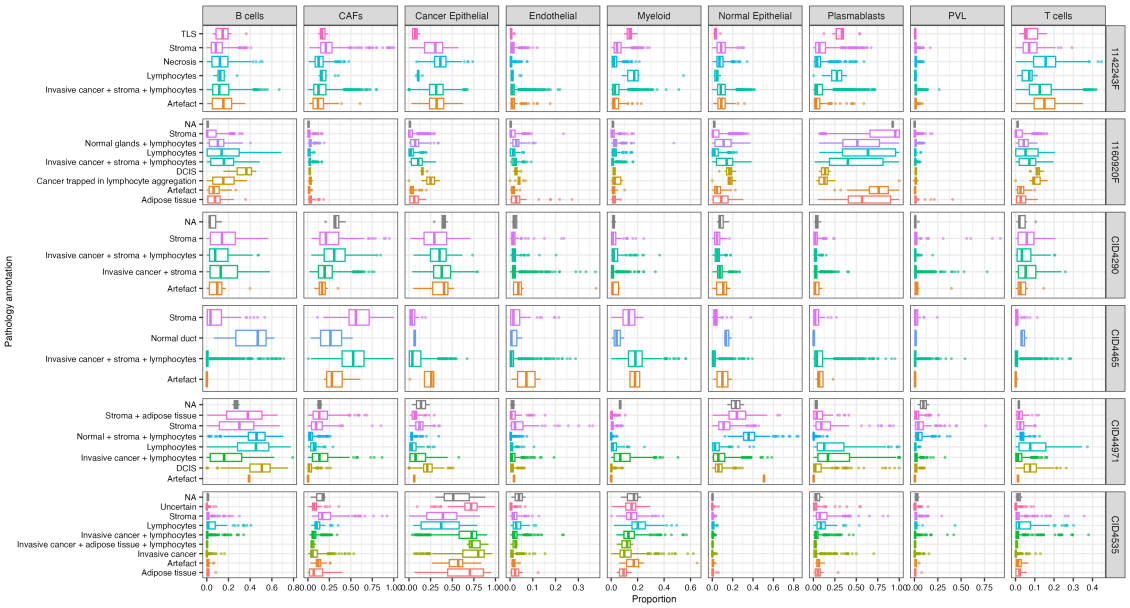


Figure A.2: Distribution of each of the nine major cell types among patches with each pathology annotation, stratified by sample.

Chapter B

Appendix for Chapter 4

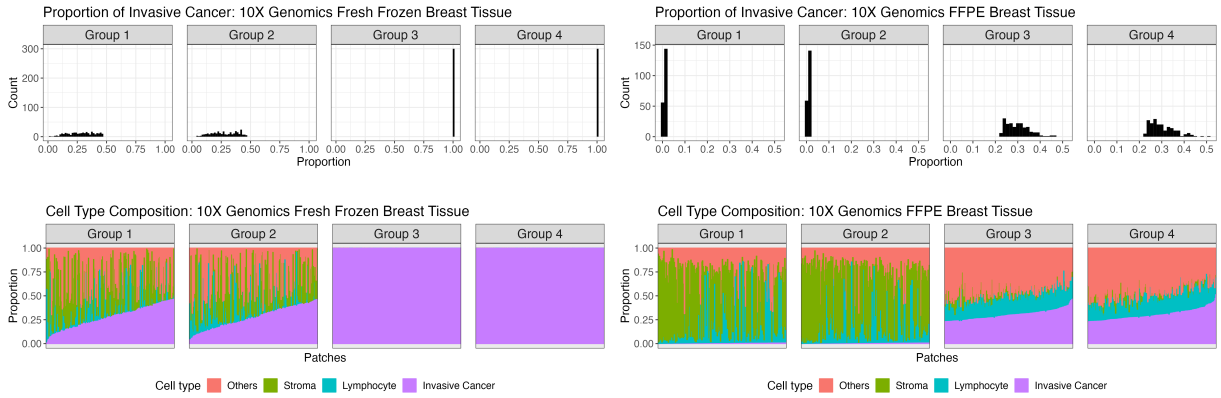


Figure B.1: Histogram of invasive cancer proportion and cell type composition in four groups of selected patches from 10X Genomics tissues used to assess ResNet50 as feature extractors.

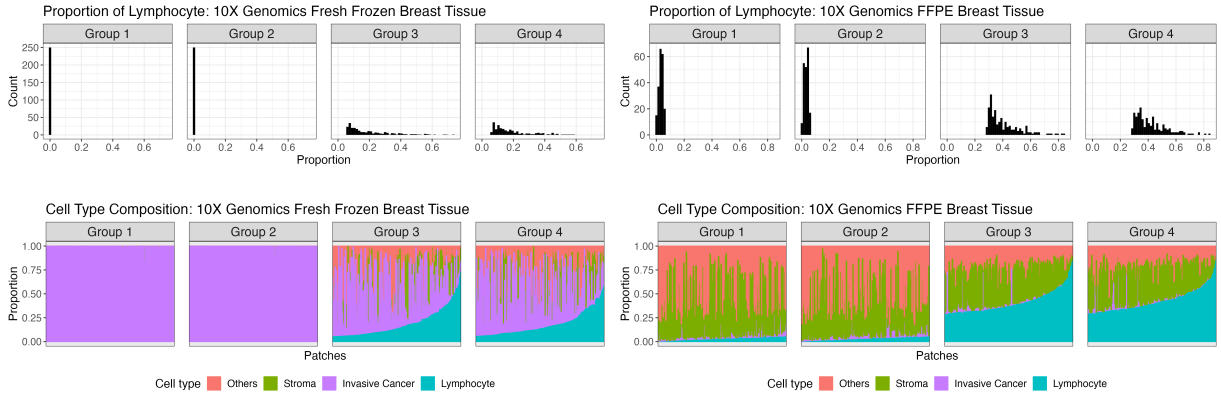


Figure B.2: Histogram of lymphocyte proportion and cell type composition in four groups of selected patches from 10X Genomics tissues used to assess ResNets as feature extractors.

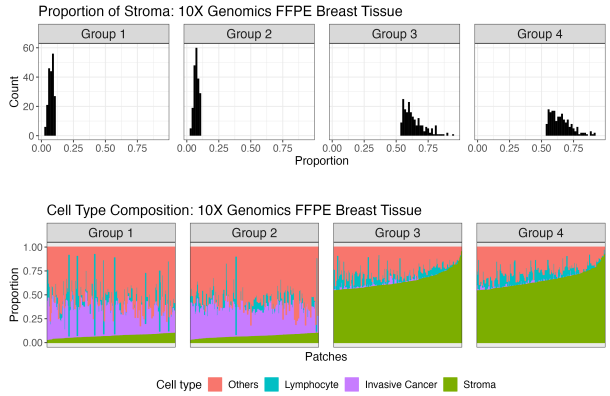


Figure B.3: Histogram of stroma proportion and cell type composition in four groups of standard patches from 10X Genomics FFPE tissue used to assess ResNet50 as feature extractors.

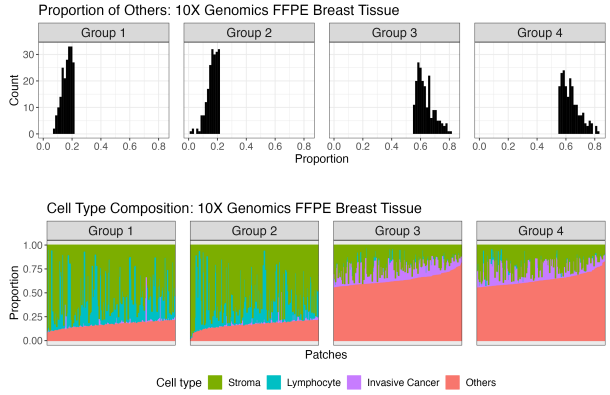


Figure B.4: Histogram of others (normal epithelial and myeloid) proportion and cell type composition in four groups of standard patches from 10X Genomics FFPE tissue used to assess ResNet50 as feature extractors.

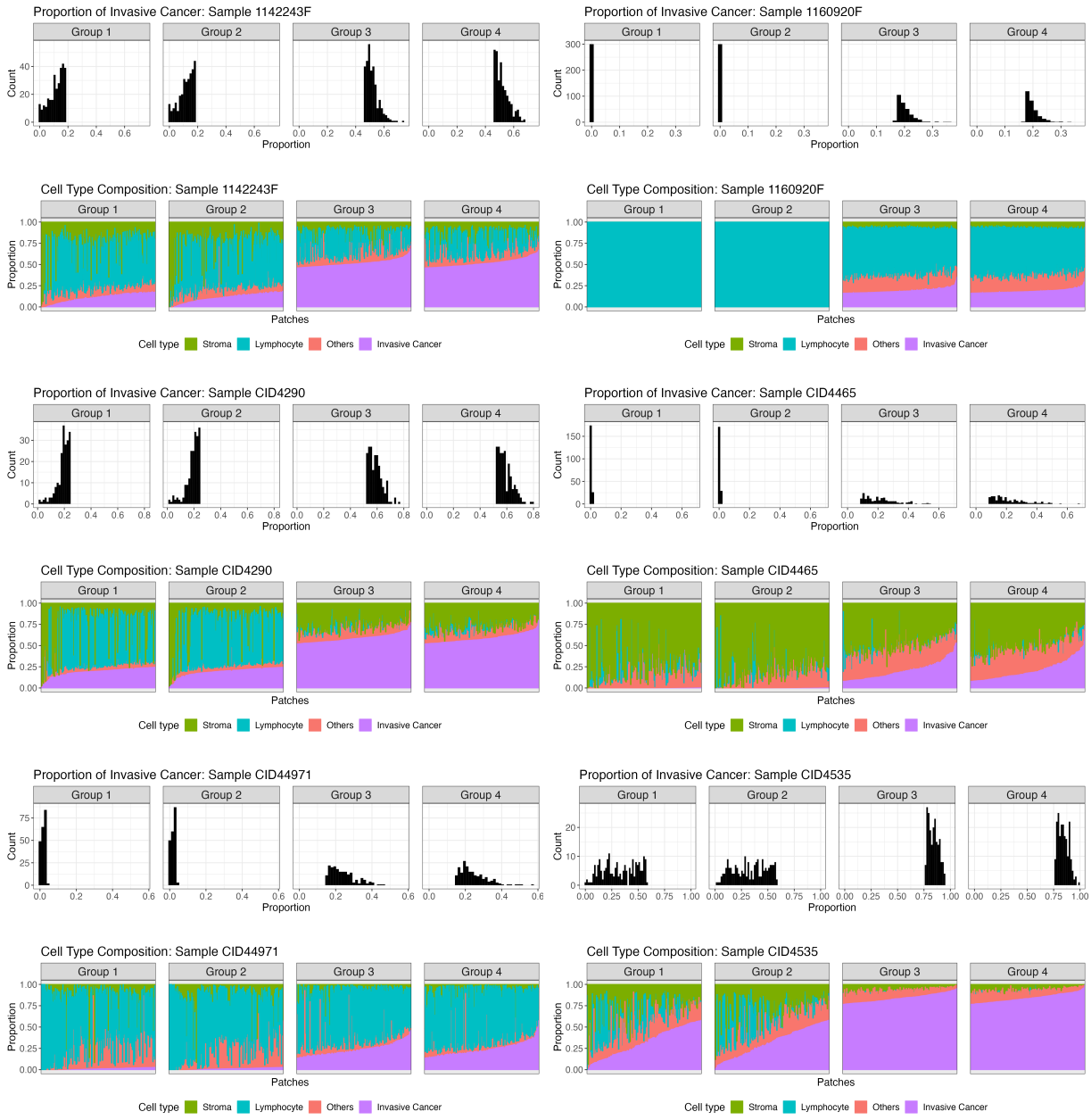


Figure B.5: Histogram of invasive cancer proportion and cell type composition in four groups of standard patches from six tissues from Wu et al. [2021] used to assess ResNet50 as feature extractors.

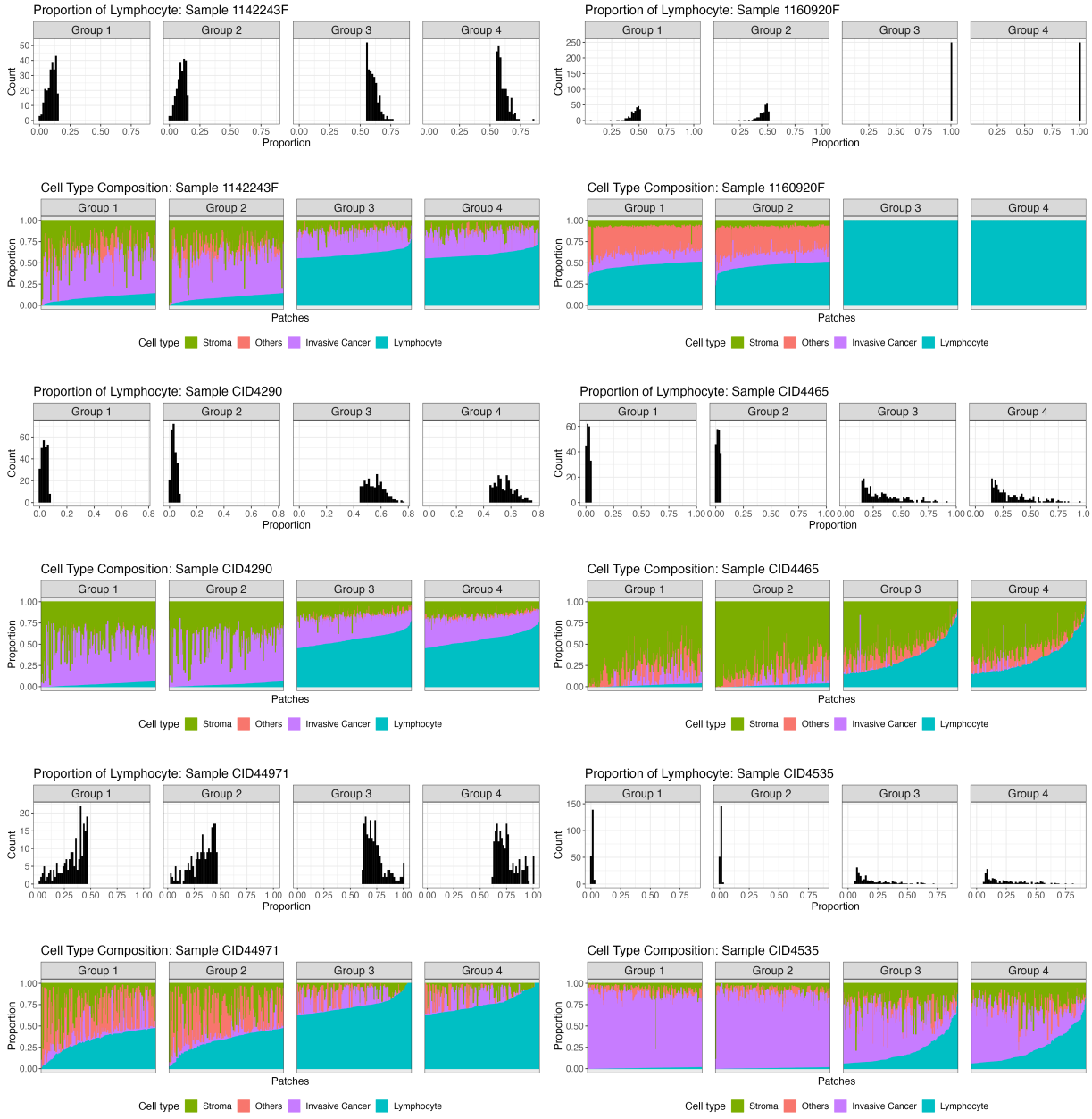


Figure B.6: Histogram of lymphocyte proportion and cell type composition in four groups of standard patches from six tissues from Wu et al. [2021] used to assess ResNet50 as feature extractors.

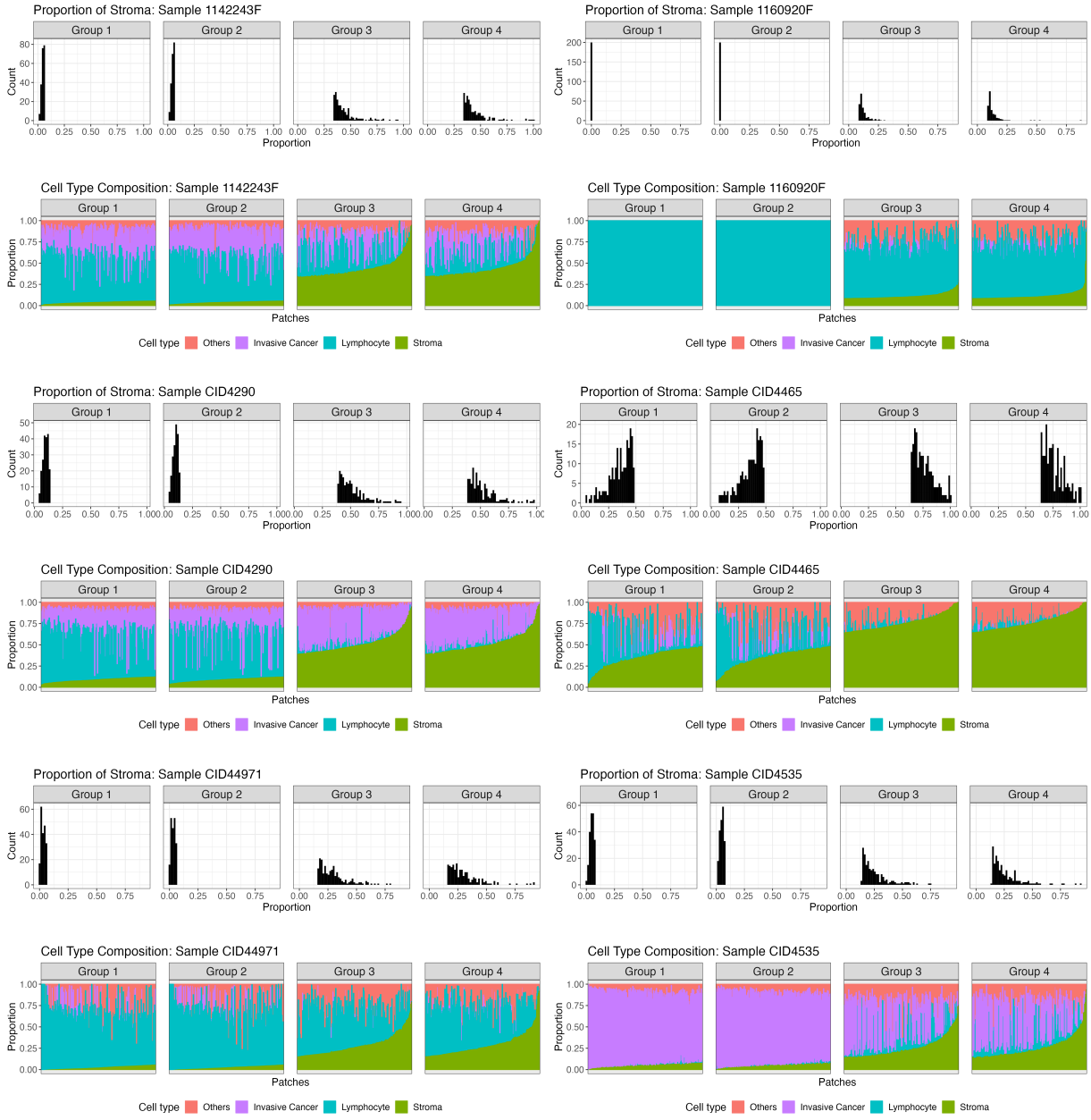


Figure B.7: Histogram of stroma proportion and cell type composition in four groups of standard patches from six tissues from Wu et al. [2021] used to assess ResNet50 as feature extractors.

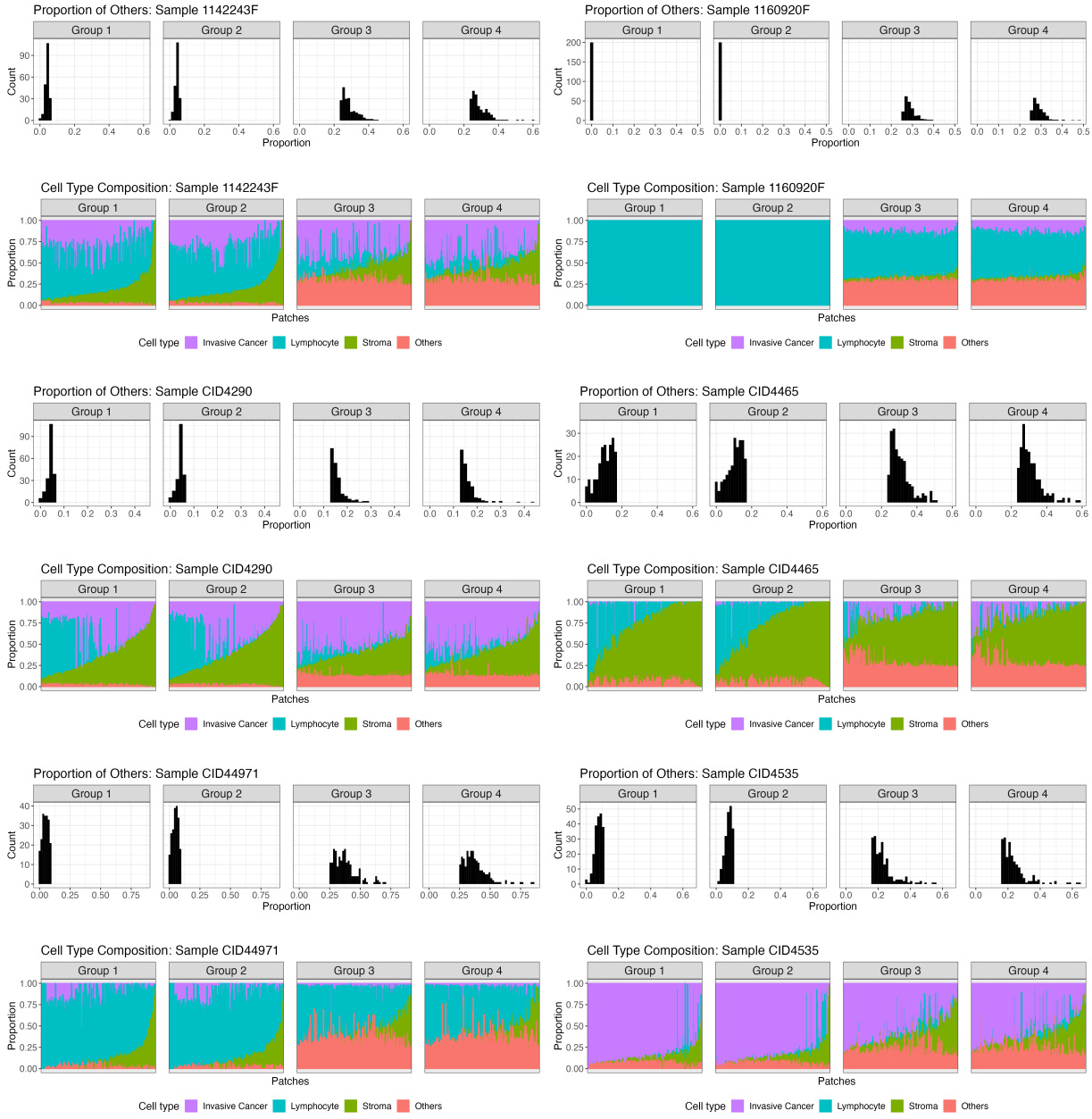


Figure B.8: Histogram of proportions of others and cell type composition in four groups of standard patches from six tissues from Wu et al. [2021] used to assess ResNet50 as feature extractors.

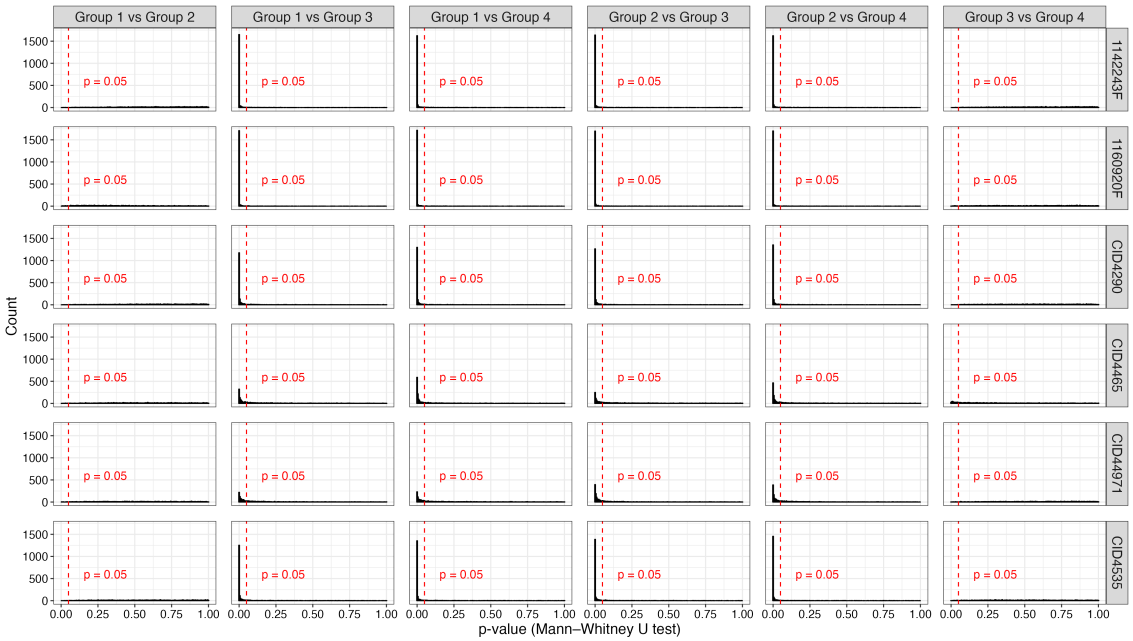


Figure B.9: Histogram of p-values obtained from pairwise Mann-Whitney U tests between four groups of patches from Wu et al. samples categorized by invasive cancer proportion.

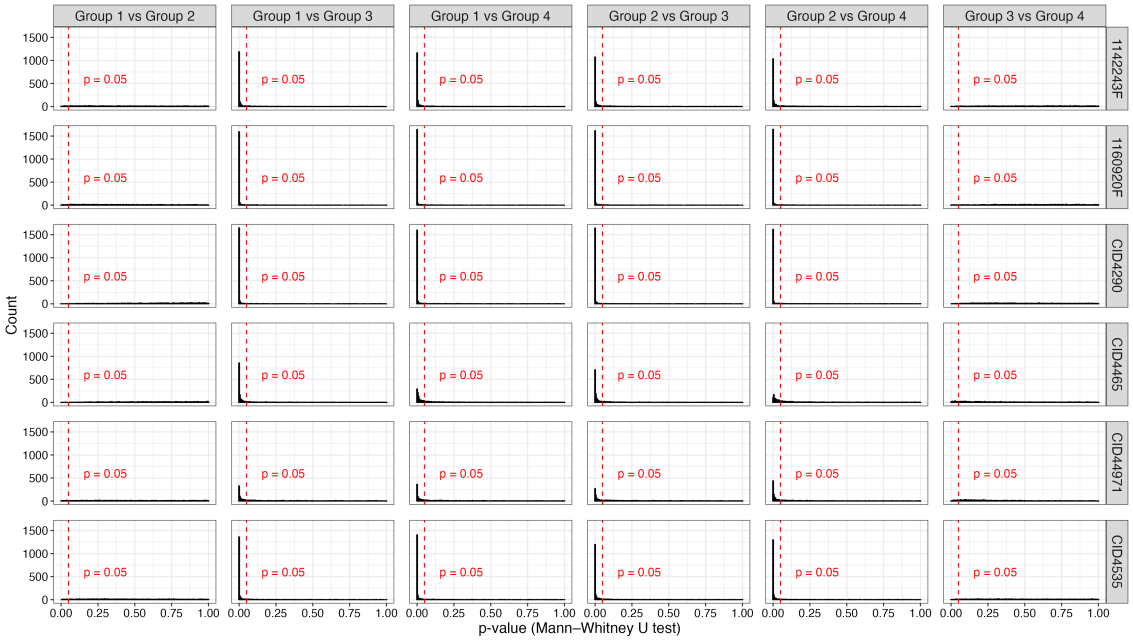


Figure B.10: Histogram of p-values obtained from pairwise Mann-Whitney U tests between four groups of patches from Wu et al. samples categorized by lymphocyte proportion.

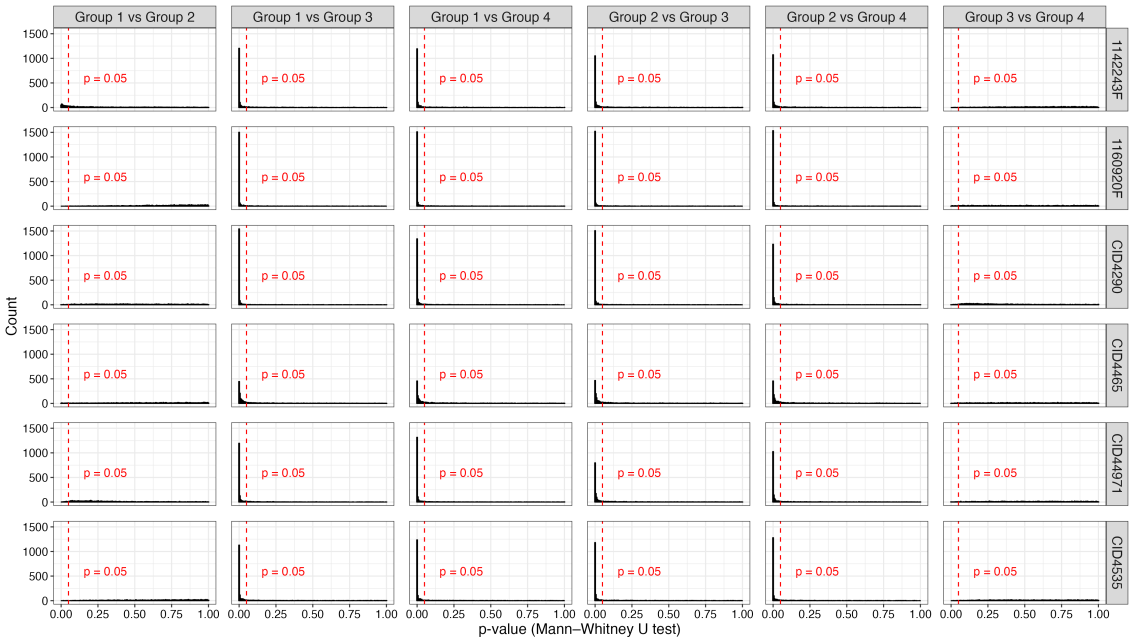


Figure B.11: Histogram of p-values obtained from pairwise Mann-Whitney U tests between four groups of patches from Wu et al. samples categorized by stroma proportion.

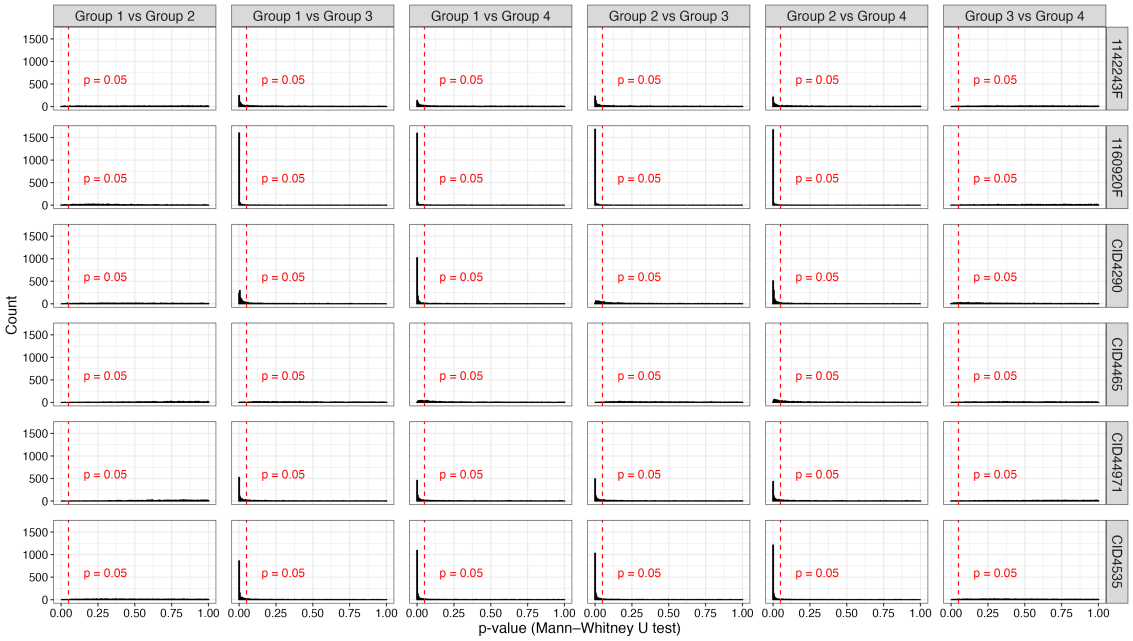


Figure B.12: Histogram of p-values obtained from pairwise Mann-Whitney U tests between four groups of patches from Wu et al. samples categorized by others proportion.

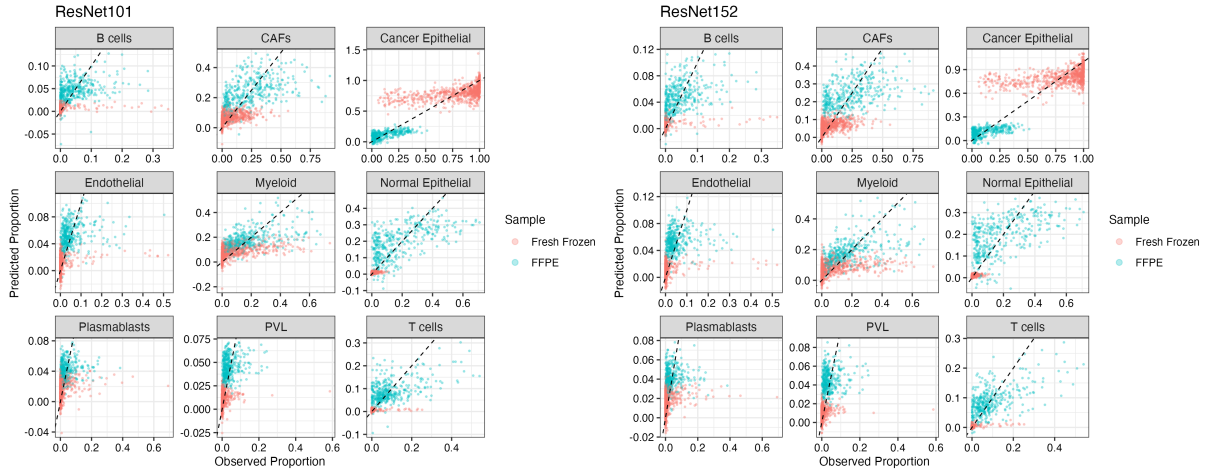


Figure B.13: The scatterplot of the predicted values obtained from the Lasso regression with λ_{min} on features extracted by pre-trained ResNet101 or ResNet512 plotted against the observed values derived from cell type deconvolution. The black dashed line represents the diagonal line.

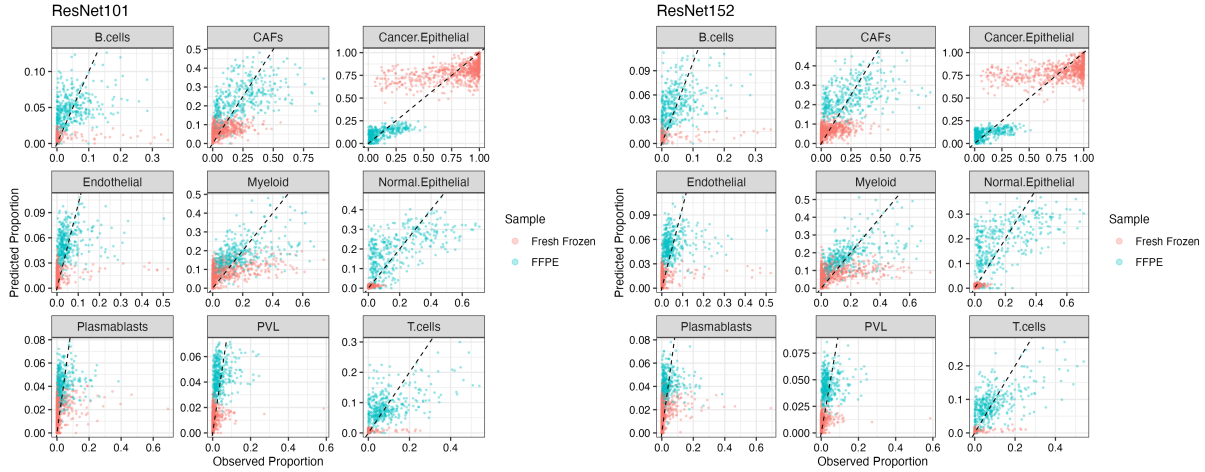


Figure B.14: The scatterplot of the normalized predicted proportions obtained from the Lasso regression with λ_{min} on features extracted by pre-trained ResNet101 or ResNet512 plotted against the observed values derived from cell type deconvolution. The black dashed line represents the diagonal line.

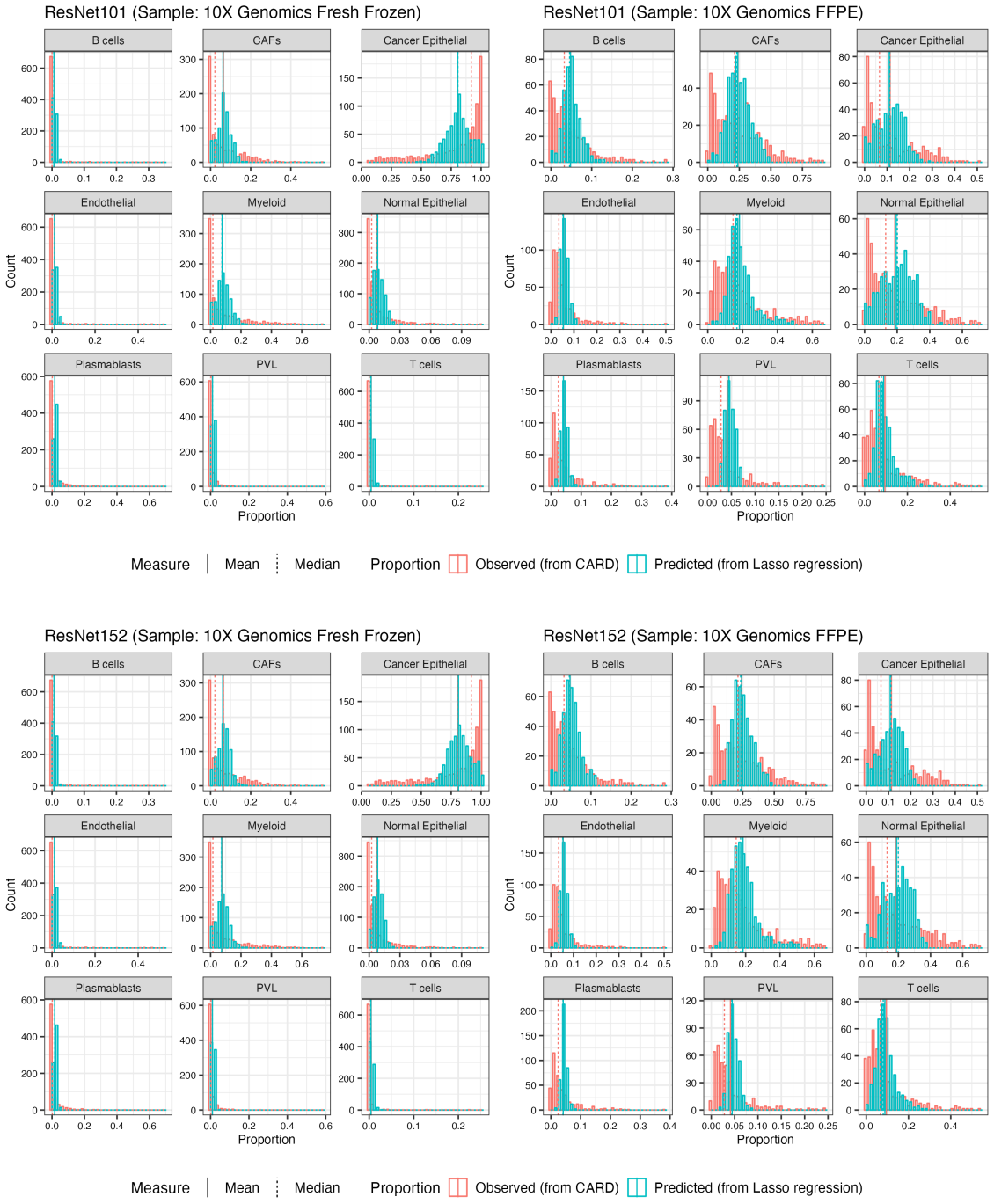


Figure B.15: The histograms of the normalized predicted proportions obtained from the Lasso regression with λ_{min} on features extracted by pre-trained ResNet101 or ResNet512 plotted and the observed values derived from cell type deconvolution. The black dashed line represents the diagonal line.

Comparison where significantly different features were identify						Number of common features			
						Fresh Frozen		FFPE	
Group 1 vs Group 2	Group 1 vs Group 3	Group 1 vs Group 4	Group 2 vs Group 3	Group 2 vs Group 4	Group 3 vs Group 4	t-test	M-W U test	t-test	M-W U test
0	0	0	0	1	1	1	3	1	0
0	0	0	1	0	1	1	1	7	8
0	0	0	1	1	0	23	23	16	19
0	0	0	1	1	1	0	0	0	0
0	0	1	0	0	1	0	2	9	7
0	0	1	0	1	0	15	12	48	46
0	0	1	0	1	1	1	9	6	12
0	0	1	1	0	0	0	0	0	0
0	0	1	1	0	1	0	0	1	0
0	0	1	1	1	0	7	12	19	19
0	0	1	1	1	1	0	0	1	1
0	1	0	0	0	1	3	2	2	1
0	1	0	0	1	0	0	0	0	0
0	1	0	0	1	1	0	0	0	0
0	1	0	1	0	0	32	30	28	32
0	1	0	1	0	1	6	5	9	12
0	1	0	1	1	0	14	18	31	32
0	1	0	1	1	1	3	2	1	2
0	1	1	0	0	0	61	69	36	40
0	1	1	0	0	1	0	0	0	0
0	1	1	0	1	0	25	26	81	49
0	1	1	0	1	1	0	1	5	6
0	1	1	1	0	0	46	54	14	15
0	1	1	1	0	1	3	2	0	1
0	1	1	1	1	0	1189	1100	1283	1350
0	1	1	1	1	1	33	29	8	13
1	0	0	0	0	1	0	0	0	0
1	0	0	0	1	0	0	1	0	0
1	0	0	0	1	1	0	0	0	0
1	0	0	1	0	0	1	2	5	2
1	0	0	1	0	1	1	0	2	1
1	0	0	1	1	0	3	3	6	7
1	0	0	1	1	1	0	0	0	1
1	0	1	0	0	0	1	0	8	7
1	0	1	0	0	1	0	0	2	1
1	0	1	0	1	0	0	0	0	0
1	0	1	0	1	1	0	0	0	2
1	0	1	1	0	0	0	0	1	0
1	0	1	1	0	1	0	0	1	0
1	0	1	1	1	0	0	0	0	0
1	0	1	1	1	1	0	0	0	0
1	1	0	0	0	0	1	0	0	1
1	1	0	0	0	1	0	0	0	0
1	1	0	0	1	0	0	0	0	0
1	1	0	0	1	1	0	0	0	0
1	1	0	1	0	0	0	0	0	0
1	1	0	1	0	1	0	0	0	0
1	1	0	1	1	0	0	0	2	2
1	1	0	1	1	1	0	0	1	0
1	1	1	0	0	0	19	19	13	11
1	1	1	0	0	1	0	0	1	1
1	1	1	0	1	0	3	3	14	7
1	1	1	0	1	1	0	1	2	3
1	1	1	1	0	0	2	2	0	0
1	1	1	1	0	1	0	0	0	0
1	1	1	1	1	0	184	198	16	16
1	1	1	1	1	1	2	1	0	0

Table B.1: Number of common significant features (p-value <0.05) identified by different pairwise comparisons of groups using T-test and Mann-Whitney U test. The groups were created from two 10X Genomics breast tissues, either fresh frozen or FFPE, based on the proportion of invasive cancer.

Comparison where significantly different features were identify						Number of common features			
						Fresh Frozen		FFPE	
Group 1 vs Group 2	Group 1 vs Group 3	Group 1 vs Group 4	Group 2 vs Group 3	Group 2 vs Group 4	Group 3 vs Group 4	t-test	M-W U test	t-test	M-W U test
0	0	0	0	1	1	0	0	23	19
0	0	0	1	0	1	1	1	8	9
0	0	0	1	1	0	22	19	59	59
0	0	0	1	1	1	0	0	0	0
0	0	1	0	0	1	2	2	4	2
0	0	1	0	1	0	68	94	18	20
0	0	1	0	1	1	0	1	5	4
0	0	1	1	0	0	0	0	0	0
0	0	1	1	0	1	0	0	0	0
0	0	1	1	1	0	18	6	25	23
0	0	1	1	1	1	0	0	1	0
0	1	0	0	0	1	0	0	44	46
0	1	0	0	1	0	0	0	0	0
0	1	0	0	1	1	0	0	0	0
0	1	0	1	0	0	25	51	100	101
0	1	0	1	0	1	0	2	30	29
0	1	0	1	1	0	16	22	21	29
0	1	0	1	1	1	0	0	1	2
0	1	1	0	0	0	24	27	14	18
0	1	1	0	0	1	0	0	0	0
0	1	1	0	1	0	110	140	4	3
0	1	1	0	1	1	0	0	0	0
0	1	1	1	0	0	12	9	12	15
0	1	1	1	0	1	0	0	0	0
0	1	1	1	1	0	1133	1001	54	50
0	1	1	1	1	1	4	5	0	0
1	0	0	0	0	1	0	0	0	0
1	0	0	0	1	0	0	0	42	37
1	0	0	0	1	1	0	0	2	1
1	0	0	1	0	0	1	3	16	6
1	0	0	1	0	1	0	1	1	0
1	0	0	1	1	0	1	1	46	64
1	0	0	1	1	1	0	0	0	0
1	0	1	0	0	0	5	10	2	0
1	0	1	0	0	1	0	2	0	0
1	0	1	0	1	0	0	2	0	0
1	0	1	0	1	1	1	0	1	0
1	0	1	1	0	0	0	0	1	0
1	0	1	1	0	1	0	1	0	0
1	0	1	1	1	0	0	0	3	4
1	0	1	1	1	1	0	0	0	1
1	1	0	0	0	0	0	0	13	6
1	1	0	0	0	1	0	0	0	0
1	1	0	0	1	0	0	0	0	1
1	1	0	0	1	1	0	0	1	0
1	1	0	1	0	0	0	0	0	0
1	1	0	1	0	1	1	0	0	0
1	1	0	1	1	0	0	1	3	4
1	1	0	1	1	1	0	0	0	0
1	1	1	0	0	0	6	9	2	2
1	1	1	0	0	1	0	0	0	0
1	1	1	0	1	0	10	8	0	0
1	1	1	0	1	1	0	0	0	0
1	1	1	1	0	0	0	0	0	0
1	1	1	1	0	1	0	0	0	0
1	1	1	1	1	0	22	44	1	1
1	1	1	1	1	1	0	2	0	0

Table B.2: Number of common significant features (p-value <0.05) identified by different pairwise comparisons of groups using the t-test and Mann-Whitney U test. The groups were created from two 10X Genomics breast tissues, either fresh frozen or FFPE, based on the proportion of lymphocytes.

Comparison where significantly different features were identify						Number of common features	
						10X FFPE	
Group 1 vs Group 2	Group 1 vs Group 3	Group 1 vs Group 4	Group 2 vs Group 3	Group 2 vs Group 4	Group 3 vs Group 4	t-test	M-W U test
0	0	0	0	1	1	0	1
0	0	0	1	0	1	7	15
0	0	0	1	1	0	56	34
0	0	0	1	1	1	0	0
0	0	1	0	0	1	2	2
0	0	1	0	1	0	2	4
0	0	1	0	1	1	1	1
0	0	1	1	0	0	0	0
0	0	1	1	0	1	0	0
0	0	1	1	1	0	3	4
0	0	1	1	1	1	0	0
0	1	0	0	0	1	0	1
0	1	0	0	1	0	0	0
0	1	0	0	1	1	0	0
0	1	0	1	0	0	132	113
0	1	0	1	0	1	49	59
0	1	0	1	1	0	176	180
0	1	0	1	1	1	89	82
0	1	1	0	0	0	30	47
0	1	1	0	0	1	0	0
0	1	1	0	1	0	4	4
0	1	1	0	1	1	0	0
0	1	1	1	0	0	85	87
0	1	1	1	0	1	6	5
0	1	1	1	1	0	705	734
0	1	1	1	1	1	115	132
1	0	0	0	0	1	0	0
1	0	0	0	1	0	0	0
1	0	0	0	1	1	0	0
1	0	0	1	0	0	1	6
1	0	0	1	0	1	1	1
1	0	0	1	1	0	9	10
1	0	0	1	1	1	1	0
1	0	1	0	0	0	3	1
1	0	1	0	0	1	0	1
1	0	1	0	1	0	0	0
1	0	1	0	1	1	0	0
1	0	1	1	0	0	0	0
1	0	1	1	0	1	0	0
1	0	1	1	1	0	0	0
1	0	1	1	1	1	0	0
1	1	0	0	0	0	0	2
1	1	0	0	0	1	0	0
1	1	0	0	1	0	0	0
1	1	0	0	1	1	0	0
1	1	0	1	0	0	0	0
1	1	0	1	0	1	0	0
1	1	0	1	1	0	5	8
1	1	0	1	1	1	1	3
1	1	1	0	0	0	4	7
1	1	1	0	0	1	0	0
1	1	1	0	1	0	0	1
1	1	1	0	1	1	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	1	0	0
1	1	1	1	1	0	0	0
1	1	1	1	1	1	0	0

Table B.3: Number of common significant features (p-value <0.05) identified by different pairwise comparisons of groups using t-test and Mann-Whitney U test. The groups were created from the 10X Genomics FFPE breast tissues, based on the proportion of stroma.

Comparison where significantly different features were identify						Number of common features	
						10X FFPE	
Group 1 vs Group 2	Group 1 vs Group 3	Group 1 vs Group 4	Group 2 vs Group 3	Group 2 vs Group 4	Group 3 vs Group 4	t-test	M-W U test
0	0	0	0	1	1	3	3
0	0	0	1	0	1	3	4
0	0	0	1	1	0	171	128
0	0	0	1	1	1	0	0
0	0	1	0	0	1	7	6
0	0	1	0	1	0	131	179
0	0	1	0	1	1	27	25
0	0	1	1	0	0	0	0
0	0	1	1	0	1	0	1
0	0	1	1	1	0	295	313
0	0	1	1	1	1	0	1
0	1	0	0	0	1	0	1
0	1	0	0	1	0	0	0
0	1	0	0	1	1	0	0
0	1	0	1	0	0	21	14
0	1	0	1	0	1	6	1
0	1	0	1	1	0	33	27
0	1	0	1	1	1	0	0
0	1	1	0	0	0	8	9
0	1	1	0	0	1	0	0
0	1	1	0	1	0	13	17
0	1	1	0	1	1	0	1
0	1	1	1	0	0	2	4
0	1	1	1	0	1	1	0
0	1	1	1	1	0	302	376
0	1	1	1	1	1	1	1
1	0	0	0	0	1	0	0
1	0	0	0	1	0	3	2
1	0	0	0	1	1	0	0
1	0	0	1	0	0	6	6
1	0	0	1	0	1	2	1
1	0	0	1	1	0	62	33
1	0	0	1	1	1	0	0
1	0	1	0	0	0	1	1
1	0	1	0	0	1	0	0
1	0	1	0	1	0	0	0
1	0	1	0	1	1	0	0
1	0	1	1	0	0	0	0
1	0	1	1	0	1	1	2
1	0	1	1	1	0	25	17
1	0	1	1	1	1	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	1	0	0
1	1	0	0	1	0	0	0
1	1	0	0	1	1	0	0
1	1	0	1	0	0	0	0
1	1	0	1	0	1	1	0
1	1	0	1	1	0	8	2
1	1	0	1	1	1	0	0
1	1	1	0	0	0	0	2
1	1	1	0	0	1	0	0
1	1	1	0	1	0	0	0
1	1	1	0	1	1	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	1	0	0
1	1	1	1	1	0	30	19
1	1	1	1	1	1	0	0

Table B.4: Number of common significant features (p-value <0.05) identified by different pairwise comparisons of groups using t-test and Mann-Whitney U test. The groups were created from the 10X Genomics FFPE breast tissues, based on the proportion of others.

Comparison where significantly different features were identify						Number of common features											
						1142243F		1160920F		CID4290		CID4465		CID44971		CID4535	
Group 1 vs Group 2	Group 1 vs Group 3	Group 1 vs Group 4	Group 2 vs Group 3	Group 2 vs Group 4	Group 3 vs Group 4	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test
0	0	0	0	1	1	1	0	2	1	1	0	13	9	1	3	2	4
0	0	0	1	0	1	0	0	0	1	0	0	15	13	2	3	0	3
0	0	0	1	1	0	18	12	13	20	73	50	19	24	36	159	49	51
0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	0	0	2	1	2	36	41	2	2	3	0
0	0	1	0	1	0	10	13	5	8	65	38	270	191	45	64	98	36
0	0	1	0	1	1	0	1	2	2	8	6	132	111	6	10	10	9
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	1	0	0	0	0	0	0	1	1	0	0	0	0
0	0	1	1	1	0	12	10	8	5	76	32	32	38	19	67	65	33
0	0	1	1	1	1	0	0	0	1	2	0	2	6	1	0	4	3
0	1	0	0	0	1	0	1	1	1	1	0	10	9	4	4	0	1
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	0	0	0	0	0	1	3	0	0	0	0
0	1	0	1	0	0	25	22	10	14	20	16	46	46	54	85	21	13
0	1	0	1	0	1	0	2	4	2	1	0	35	51	5	5	4	3
0	1	0	1	1	0	13	12	7	5	22	21	14	22	17	92	15	21
0	1	0	1	1	1	0	1	0	0	0	0	0	2	0	0	0	1
0	1	1	0	0	0	20	13	24	19	14	20	49	66	25	26	17	14
0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	0	5	6	7	13	23	32	97	112	18	22	26	12
0	1	1	0	1	1	1	0	0	1	2	2	1	7	0	0	1	2
0	1	1	1	0	0	21	21	2	7	13	9	29	32	21	22	8	9
0	1	1	1	0	1	1	0	2	2	0	0	1	2	0	1	0	0
0	1	1	1	1	0	1664	1690	1658	1636	1252	1417	257	420	113	459	1329	1443
0	1	1	1	1	1	6	10	36	44	4	6	1	5	0	2	16	21
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	2	1	1	1	0	3	2	1	2	4	1	0
1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0
1	0	0	1	0	0	0	1	0	0	1	1	3	2	1	3	1	0
1	0	0	1	0	1	0	0	0	0	0	0	0	2	0	0	0	0
1	0	0	1	1	0	2	1	2	0	1	3	1	5	2	11	4	4
1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	0	1	1	2	4	0	3	2	1	0	0
1	0	1	0	0	1	0	0	0	0	0	0	2	1	0	0	0	0
1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0
1	0	1	1	1	0	0	0	0	0	3	1	0	1	0	1	0	0
1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	1	0	1	0	0	0	0	1	3	2	0	0
1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	1	1	0	0	0	0	1	0	0	0	1	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	0
1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	2	3	7	3	1	3	2	6	3	4	0	0
1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	1	0	0	0	1	2	0	1	0	1	1	1	0	0
1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	1	0	3	1	0	0	0	0	0	0	0	0
1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	0	14	9	88	63	2	5	0	1	0	5	5	10
1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1

Table B.5: Number of common significant features (p-value < 0.05) identified by different pairwise comparisons of groups using t-test and Mann-Whitney U test. The groups were created from six samples from Wu et al. [2021], based on the proportion of invasive cancer.

Comparison where significantly different features were identify						Number of common features											
						1142243F		1160920F		CID4290		CID4465		CID44971		CID4535	
Group 1 vs Group 2	Group 1 vs Group 3	Group 1 vs Group 4	Group 2 vs Group 3	Group 2 vs Group 4	Group 3 vs Group 4	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test
0	0	0	0	1	1	4	3	2	2	0	0	2	0	47	25	0	0
0	0	0	1	0	1	0	1	0	1	0	2	3	8	5	0	2	2
0	0	0	1	1	0	35	18	12	22	20	28	22	25	37	35	5	11
0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	1	0	1	0	0	0	1	18	3	5	2
0	0	1	0	1	0	32	31	23	21	11	10	17	26	278	126	43	58
0	0	1	0	1	1	6	6	9	10	3	2	3	7	212	30	7	10
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	0	28	21	21	31	16	7	9	10	30	73	4	12
0	0	1	1	1	1	3	5	3	7	1	0	0	0	6	6	0	0
0	1	0	0	0	1	4	3	1	2	0	1	20	10	16	16	1	0
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	1	0	0
0	1	0	1	0	0	24	26	13	12	17	21	196	188	39	73	39	44
0	1	0	1	0	1	6	7	1	6	4	3	97	86	19	16	0	7
0	1	0	1	1	0	20	14	5	7	14	16	33	53	16	20	10	15
0	1	0	1	1	1	1	0	0	0	1	1	2	4	0	0	0	0
0	1	1	0	0	0	95	82	15	12	30	27	51	53	35	34	36	33
0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	0	34	30	6	10	11	6	13	28	40	24	72	55
0	1	1	0	1	1	0	1	1	0	1	0	0	0	5	1	3	7
0	1	1	1	0	0	51	67	14	25	22	25	250	160	17	45	19	20
0	1	1	1	0	1	0	2	2	2	1	1	71	48	2	7	0	0
0	1	1	1	1	0	1083	1150	1613	1546	1654	1640	495	575	213	337	1351	1317
0	1	1	1	1	1	11	12	14	21	16	31	26	38	18	12	12	26
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	6	2	9	3	0	1	1	0	3	18	0	0
1	0	0	0	1	1	0	0	1	0	0	0	0	0	1	10	0	0
1	0	0	1	0	0	2	4	0	0	0	0	1	1	1	1	1	0
1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1
1	0	0	1	1	0	19	22	1	4	5	2	0	1	3	12	0	1
1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	1	0	0	0	0	0	4	1	3	4
1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2	1
1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	1	1	0	0	0	2	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	1	0	4	6	5	4	3	1	0	0	2	4	0	0
1	0	1	1	1	1	1	1	1	1	0	0	0	0	1	3	0	0
1	1	0	0	0	0	9	4	6	4	0	0	6	2	3	7	0	0
1	1	0	0	0	1	1	4	0	0	0	0	2	1	0	4	0	0
1	1	0	0	1	0	2	0	0	0	0	0	0	1	1	2	0	0
1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	4	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0
1	1	0	1	1	0	0	0	1	0	1	1	0	0	0	0	0	1
1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	22	22	5	3	0	3	4	7	9	16	4	6
1	1	1	0	0	1	0	2	0	0	0	0	0	0	0	1	0	0
1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	7	8
1	1	1	0	1	1	0	1	0	0	0	0	0	0	1	0	1	2
1	1	1	1	0	0	8	5	2	5	1	2	2	1	1	3	1	0
1	1	1	1	0	1	0	1	0	1	0	0	1	0	0	1	0	0
1	1	1	1	1	0	41	27	68	58	6	5	0	0	1	7	90	49
1	1	1	1	1	1	0	2	2	3	0	0	0	0	0	0	3	3

Table B.6: Number of common significant features (p-value <0.05) identified by different pairwise comparisons of groups using t-test and Mann-Whitney U test. The groups were created from six samples from Wu et al. [2021], based on the proportion of lymphocytes.

Comparison where significantly different features were identify						Number of common features											
						1142243F		1160920F		CID4290		CID4465		CID44971		CID4535	
Group 1 vs Group 2	Group 1 vs Group 3	Group 1 vs Group 4	Group 2 vs Group 3	Group 2 vs Group 4	Group 3 vs Group 4	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test
0	0	0	0	1	1	0	0	1	2	0	1	2	8	0	2	1	0
0	0	0	1	0	1	3	1	1	2	0	1	3	4	1	2	1	0
0	0	0	1	1	0	20	23	7	10	9	12	81	63	11	10	15	26
0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	1	0	0	1	0	0	7	3	0	3	1	0
0	0	1	0	1	0	25	33	38	30	9	12	52	48	42	54	74	64
0	0	1	0	1	1	3	2	9	4	0	0	10	8	13	16	7	7
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
0	0	1	1	1	0	16	10	25	33	12	11	37	40	10	19	43	34
0	0	1	1	1	1	0	0	3	3	0	0	2	1	0	0	3	0
0	1	0	0	0	1	1	2	3	6	1	1	1	2	1	3	0	0
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	1	0	0	28	26	19	20	45	68	53	102	18	27	37	36
0	1	0	1	0	1	5	4	13	8	7	9	1	7	1	5	2	0
0	1	0	1	1	0	24	20	9	7	16	24	41	49	6	8	12	13
0	1	0	1	1	1	2	0	0	1	0	1	0	0	0	0	0	0
0	1	1	0	0	0	67	46	15	18	22	25	64	68	343	108	26	13
0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	0	23	39	13	9	3	3	59	40	244	108	36	24
0	1	1	0	1	1	1	1	1	1	0	0	3	4	1	2	1	0
0	1	1	1	0	0	27	26	11	25	39	57	24	49	39	31	11	12
0	1	1	1	0	1	0	0	7	3	2	9	0	0	0	0	0	0
0	1	1	1	1	0	1099	982	1545	1544	1569	1446	792	693	640	1128	1251	1342
0	1	1	1	1	1	7	8	55	38	10	39	4	5	3	11	6	9
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	4	10	0	1	0	0	5	6	5	3	2	1
1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0
1	0	0	1	0	0	11	21	0	0	0	1	5	4	3	0	0	0
1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	1	0	37	39	1	0	5	6	32	16	5	2	0	1
1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	9	21	0	0	0	0	4	2	11	5	0	0
1	0	1	0	0	1	0	0	0	0	0	0	1	2	0	0	0	0
1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	1	0	0	0	0	1	1	0	0	0	0
1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	1	0	3	4	0	1	0	1	5	1	0	0	1	3
1	0	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0
1	1	0	0	0	0	6	5	0	0	2	3	2	2	6	1	0	0
1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	1	0	11	9	0	0	0	1	4	4	0	0	0	0
1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	56	80	1	0	2	3	12	10	55	15	0	0
1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	1	0	11	21	0	0	0	0	3	2	7	6	0	0
1	1	1	0	1	1	0	1	0	0	0	0	1	0	0	1	0	0
1	1	1	1	0	0	6	3	0	0	4	5	0	0	0	0	0	1
1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0
1	1	1	1	1	0	135	198	1	0	7	22	7	4	16	33	1	3
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Table B.7: Number of common significant features (p-value <0.05) identified by different pairwise comparisons of groups using t-test and Mann-Whitney U test. The groups were created from six samples from Wu et al. [2021], based on the proportion of stroma.

Comparison where significantly different features were identify						Number of common features											
						1142243F		1160920F		CID4290		CID4465		CID44971		CID4535	
Group 1 vs Group 2	Group 1 vs Group 3	Group 1 vs Group 4	Group 2 vs Group 3	Group 2 vs Group 4	Group 3 vs Group 4	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test	t-test	M-W U test
0	0	0	0	1	1	6	4	0	0	2	3	5	6	0	1	0	1
0	0	0	1	0	1	1	2	0	0	0	1	0	0	0	1	0	0
0	0	0	1	1	0	58	57	36	36	9	15	91	15	63	35	32	41
0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	0	0	0	3	4	1	4	5	7	1	0
0	0	1	0	1	0	44	43	9	7	171	168	219	210	41	57	150	96
0	0	1	0	1	1	3	7	1	2	55	85	7	13	8	12	16	20
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0
0	0	1	1	1	0	43	44	15	18	18	30	47	6	15	17	47	100
0	0	1	1	1	1	0	0	1	3	0	2	0	0	0	0	2	6
0	1	0	0	0	1	6	7	0	0	0	0	3	3	6	1	0	4
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	1	0	0	79	109	11	14	10	8	22	1	46	91	24	44
0	1	0	1	0	1	9	15	2	2	5	2	0	0	8	9	1	2
0	1	0	1	1	0	34	42	26	27	3	1	4	0	56	52	10	21
0	1	0	1	1	1	0	1	1	1	0	0	0	0	1	0	0	0
0	1	1	0	0	0	58	44	11	10	158	99	17	10	140	44	61	13
0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	0	8	11	9	8	581	410	1	1	39	47	166	17
0	1	1	0	1	1	0	0	0	1	11	43	0	0	3	5	8	0
0	1	1	1	0	0	62	51	8	6	12	16	3	1	30	27	15	19
0	1	1	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0
0	1	1	1	1	0	275	260	1635	1622	248	430	21	5	390	618	924	1150
0	1	1	1	1	1	1	3	12	15	0	5	0	0	3	6	17	27
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	10	8	1	1	2	2	3	2	0	0	0	0
1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0
1	0	0	1	0	0	2	2	1	1	1	1	0	0	1	0	4	0
1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	1	0	8	16	9	14	0	0	0	0	3	0	2	3
1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	0	0	1	8	7	1	2	2	0	1	0
1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	1	0	4	3	0	1	0	0	0	0	0	0	1	2
1	0	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	10	8	0	1	2	3	0	2	0	1	0	0
1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	1	0	0	0	2	3	0	0	0	0	0	0	0	1
1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	17	12	2	1	35	23	0	0	7	3	2	1
1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	1	0	0	0	0	1	8	7	0	0	1	0	0	0
1	1	1	0	1	1	0	0	0	0	1	1	0	0	0	1	0	0
1	1	1	1	0	0	4	4	0	0	1	0	0	0	2	0	0	0
1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	0	1	5	72	49	0	0	0	0	3	1	1	11
1	1	1	1	1	1	0	0	1	2	0	0	0	0	0	0	0	1

Table B.8: Number of common significant features (p-value <0.05) identified by different pairwise comparisons of groups using t-test and Mann-Whitney U test. The groups were created from six samples from Wu et al. [2021], based on the proportion of others.

Chapter C

Appendix for Chapter 5

	Slope	Optimizer	Batch size	Learning rate	Dropout proportion	Dense layer Units
Plasmablasts	0.0351	Adam	32	0.0001	0.2	256
PVL	0.0632	Adam	32	0.0001	0.2	512
Endothelial	0.0951	Adam	32	0.001	0.0	512
B cells	0.2162	Adam	32	0.001	0.0	512
CAFs	0.2651	Adam	32	0.0001	0.5	512
Myeloid	0.3769	Adam	64	0.0001	0.5	512
Cancer Epithelial	0.3840	Adam	32	0.0001	0.5	512
T cells	0.4096	Adam	64	0.0001	0.0	256
Normal Epithelial	0.4634	Adam	32	0.0001	0.5	512

Table C.1: Parameters of neural networks trained on standard patches from the 10X Genomics FFPE tissue, yielding the highest slope of the best-fit line for each of the nine cell types in the testing dataset.

Cell type	Slope	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
Stroma	0.2852	Adam	64	0.0001	0.2	512
Invasive Cancer	0.3234	Adam	32	0.001	0.0	256
Others	0.3414	Adam	64	0.0001	0.2	512
Lymphocyte	0.3586	Adam	128	0.001	0.0	256

Table C.2: Parameters of neural networks trained on standard patches from the 10X Genomics FFPE tissue, yielding the highest slope of the best-fit line for each of the four cell types in the testing dataset.

	Slope	Optimizer	Batch size	Learning rate	Dropout proportion	Dense layer Units
B cells	0.0185	Adam	16	0.001	0.2	256
Endothelial	0.0230	Adam	64	0.0001	0.0	256
T cells	0.0257	Adam	16	0.001	0.2	256
Plasmablasts	0.0367	Adam	64	0.0001	0.0	256
Normal Epithelial	0.0378	Adam	16	0.001	0.2	256
PVL	0.0419	Adam	64	0.0001	0.0	256
CAFs	0.1460	Adam	64	0.0001	0.0	256
Myeloid	0.1480	Adam	64	0.0001	0.0	256
Cancer Epithelial	0.1910	Adam	64	0.0001	0.0	256

Table C.3: Parameters of neural networks trained on standard patches from the 10X Genomics fresh frozen tissue, yielding the highest slope of the best-fit line for each of the nine cell types in the testing dataset.

	Slope	Optimizer	Batch size	Learning rate	Dropout proportion	Dense layer Units
Lymphocyte	0.0397	Adam	64	0.0001	0.5	256
Stroma	0.1295	Adam	64	0.0001	0.0	512
Others	0.1606	Adam	64	0.0001	0.0	512
Invasive Cancer	0.1830	Adam	64	0.0001	0.0	512

Table C.4: Parameters of neural networks trained on standard patches from the 10X Genomics fresh frozen tissue, yielding the highest slope of the best-fit line for each of the four cell types in the testing dataset.

	Slope	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
Plasmablasts	0.0598	Adam	128	0.0001	0.0	512
Endothelial	0.1848	Adam	16	0.001	0.5	512
Myeloid	0.2378	Adam	64	0.001	0.0	256
PVL	0.2488	Adam	16	0.001	0.5	512
B cells	0.3008	Adam	128	0.001	0.0	512
CAFs	0.4159	Adam	64	0.001	0.0	256
Cancer Epithelial	0.6417	Adam	16	0.001	0.5	256
Normal Epithelial	0.6743	Adam	32	0.001	0.0	512
T cells	0.8326	Adam	32	0.001	0.0	256

Table C.5: Parameters of neural networks trained on standard patches from both 10X Genomics breast tissues, yielding the highest slope of the best-fit line for each of the nine cell types in the testing dataset.

	Slope	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
Stroma	0.4743	Adam	64	0.001	0.0	512
Others	0.4957	Adam	64	0.001	0.0	512
Lymphocyte	0.5056	Adam	16	0.001	0.0	512
Invasive Cancer	0.6301	Adam	16	0.001	0.0	512

Table C.6: Parameters of neural networks trained on standard patches from both 10X Genomics breast tissues, yielding the highest slope of the best-fit line for each of the four cell types in the testing dataset.

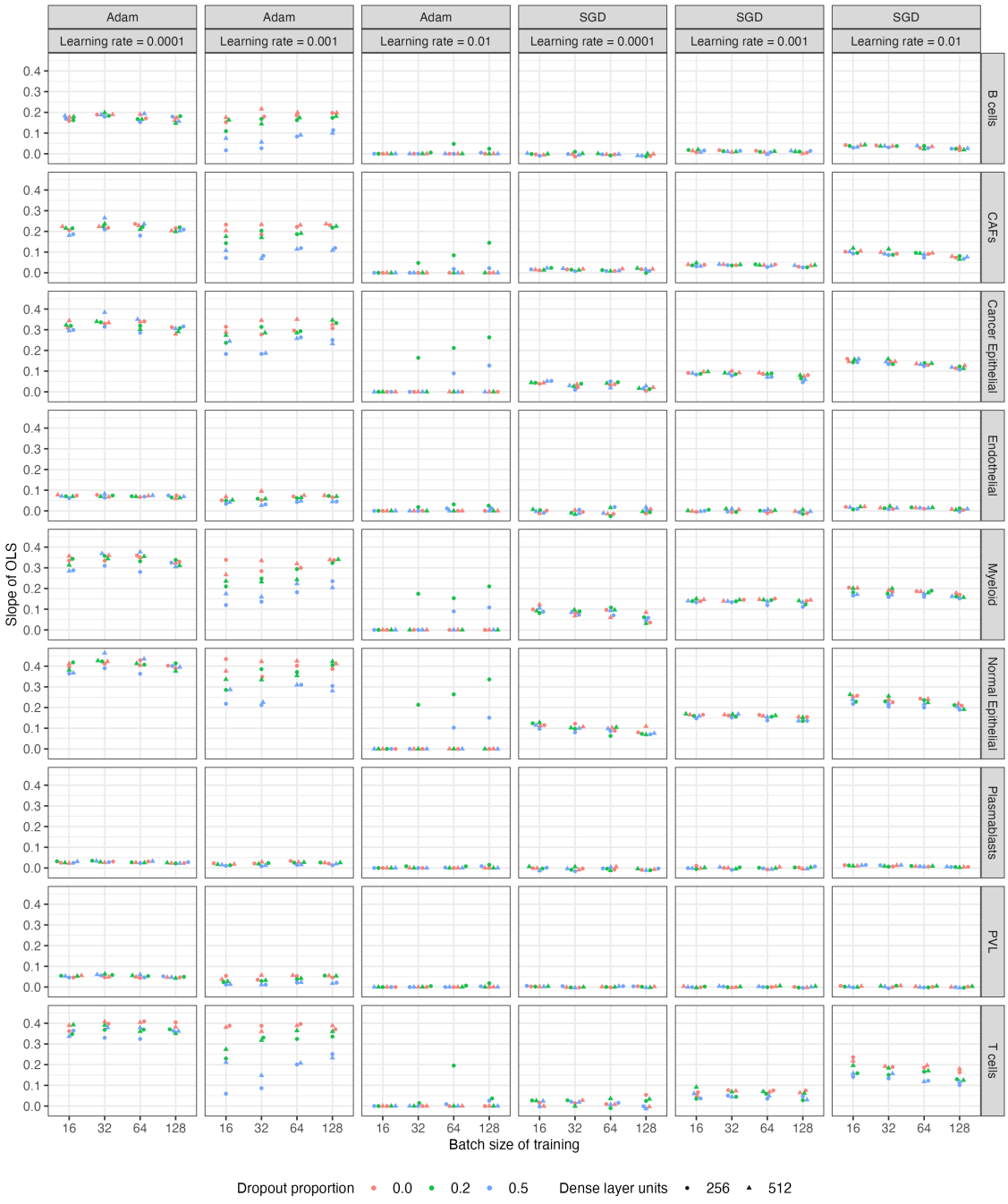


Figure C.1: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using standard patches from 10X Genomics FFPE breast tissue.

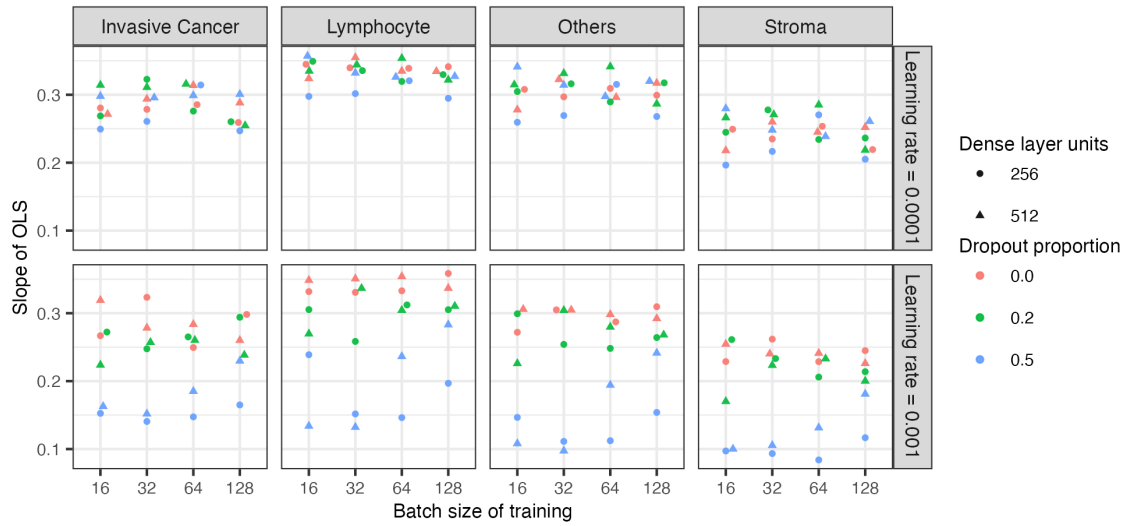


Figure C.2: Slope of the best-fit line between predicted and observed proportions of four cell types in testing data, predicted by networks trained using standard patches from 10X Genomics FFPE breast tissue, excluding networks with limited learning.

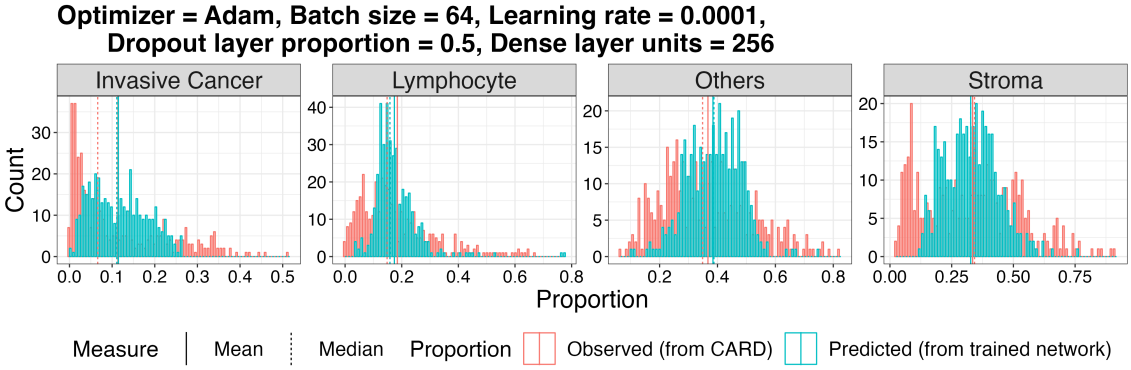


Figure C.3: Histogram comparing predicted and observed proportions for each of the four cell types, where the predictions were made by a network trained using standard patches from 10X Genomics FFPE breast tissue.

Figure C.4: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using standard patches from 10X Genomics fresh frozen breast tissue.

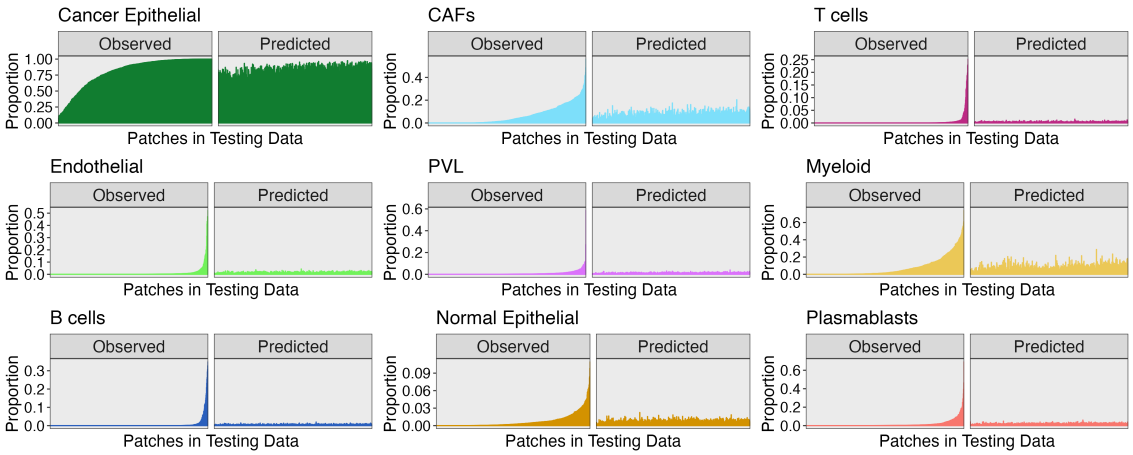


Figure C.5: Comparison of predicted and observed nine cell type compositions in each of the standard patches from the 10X Genomics fresh frozen tissue in the testing data.

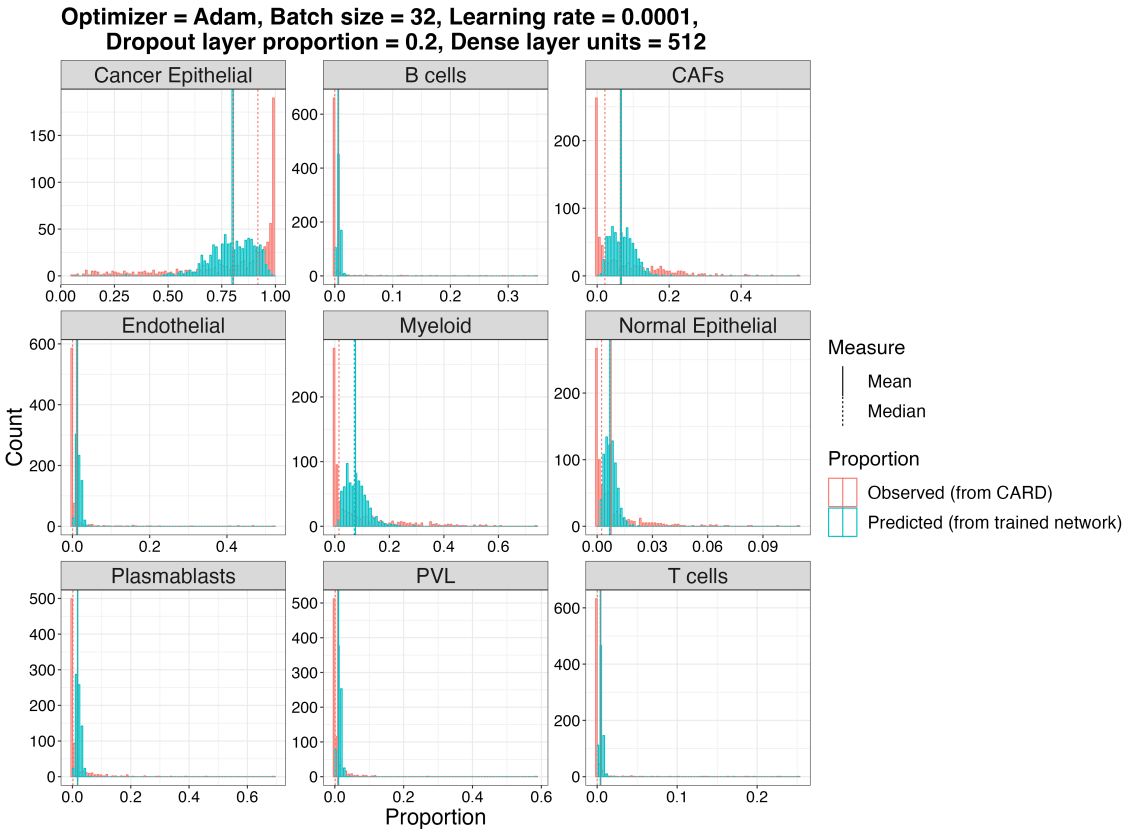


Figure C.6: Histogram comparing predicted and observed proportions for each of the nine cell types, where the predictions were made by a network trained using standard patches from 10X Genomics fresh frozen breast tissue.

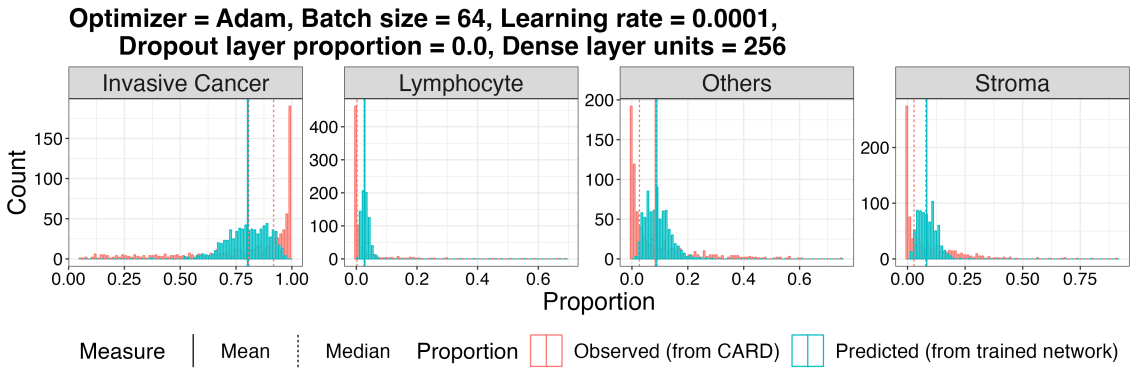


Figure C.7: Histogram comparing predicted and observed proportions for each of the four cell types, where the predictions were made by a network trained using standard patches from 10X Genomics fresh frozen breast tissue.

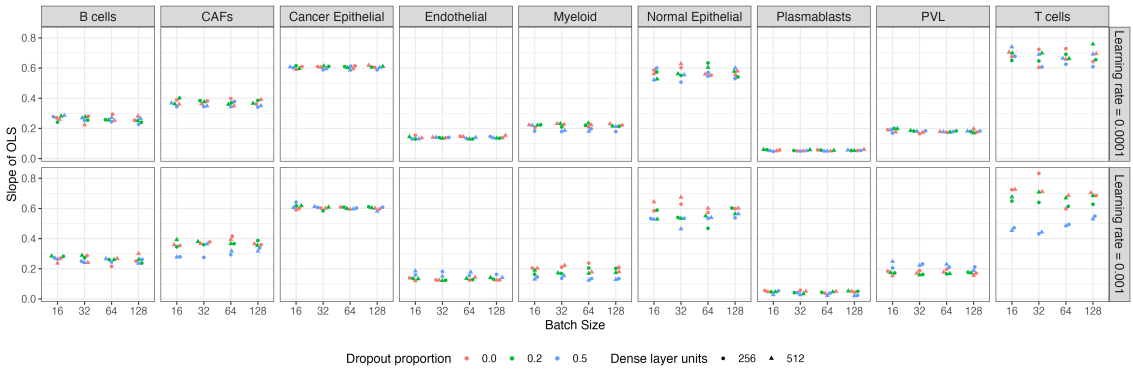


Figure C.8: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using standard patches from both 10X Genomics breast tissues.

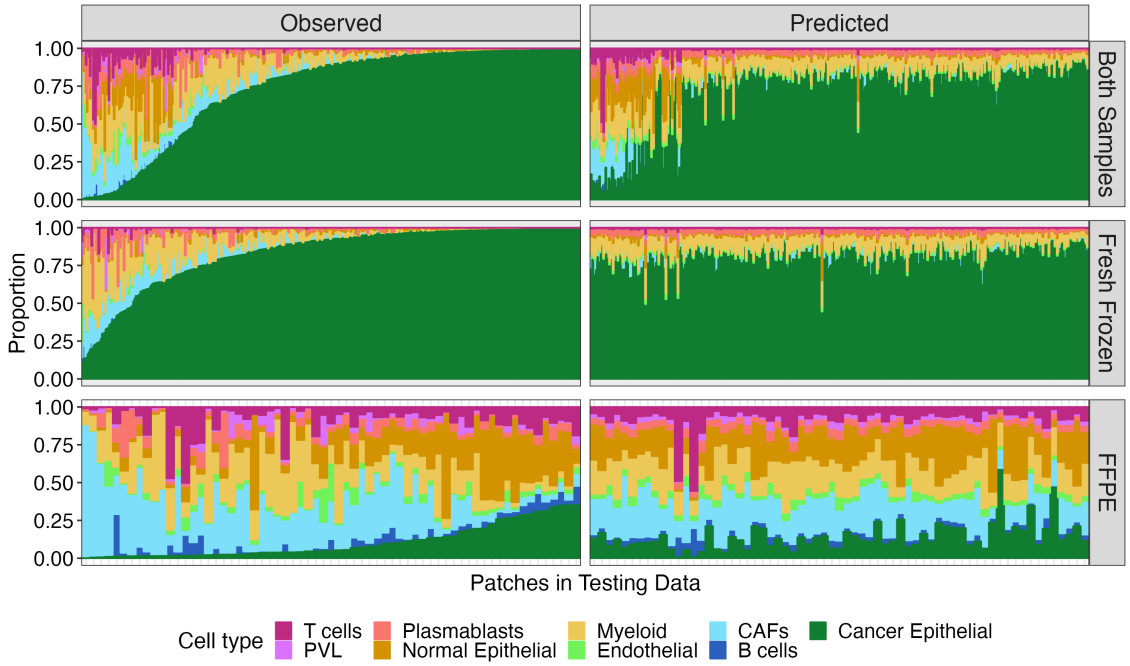


Figure C.9: Comparison of predicted and observed nine cell type compositions in each of the standard patches from both 10X Genomics tissues in the testing data.

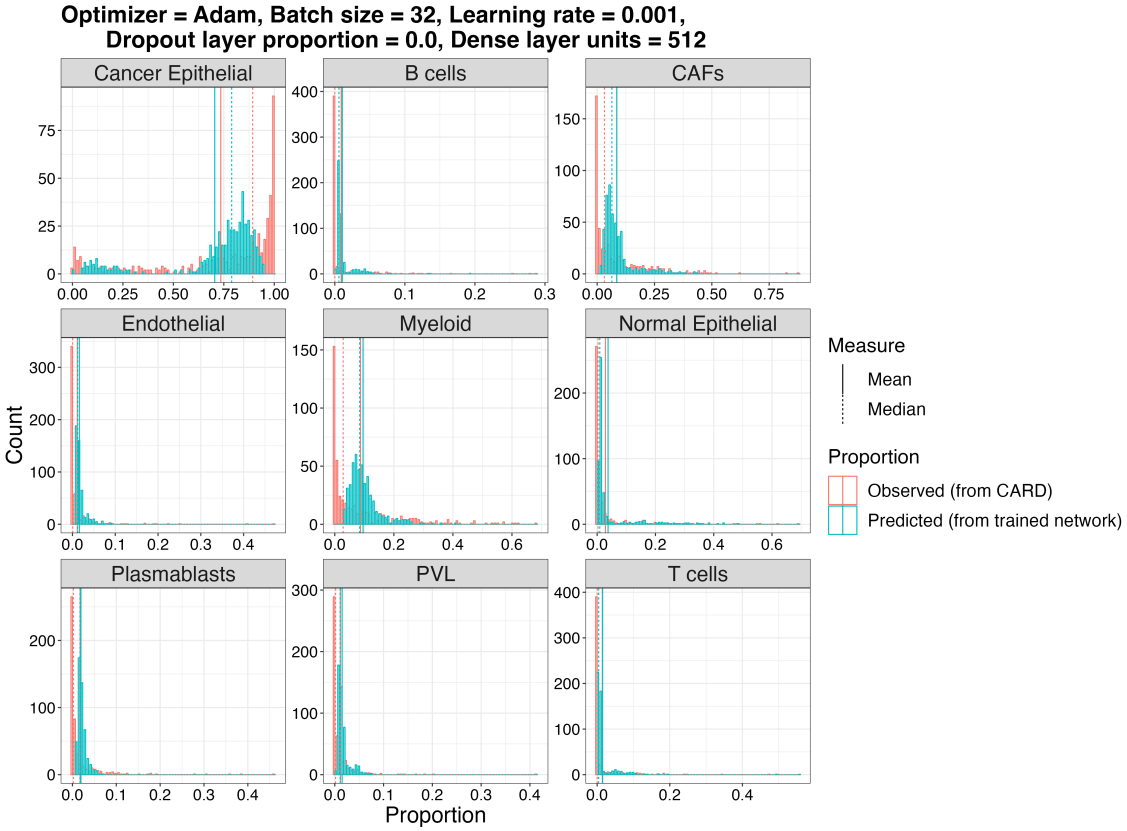


Figure C.10: Histogram comparing predicted and observed proportions for each of the nine cell types, where the predictions were made by a network trained using standard patches from both 10X Genomics breast tissues.

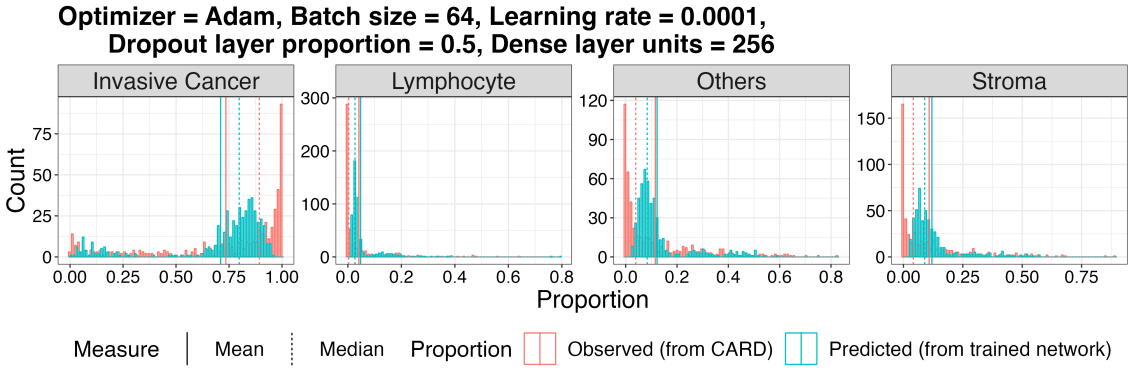


Figure C.11: Histogram comparing predicted and observed proportions for each of the nine cell types, where the predictions were made by a network trained using standard patches from both 10X Genomics breast tissues.

Chapter D

Appendix for Chapter 6

	Slope	Image	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
Others	0.0643	Original	Adam	32	0.0001	0.0	512
Invasive Cancer	0.2244	Macenko	Adam	128	0.0001	0.0	512
Stroma	0.2377	Original	Adam	128	0.0001	0.0	256
Lymphocyte	0.3898	Original	Adam	32	0.0001	0.0	512

Table D.1: Parameters of neural networks trained on standard patches from all six samples, yielding the highest slope of the best-fit line for each of the four cell types in the testing dataset.

	Slope	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
B cells	0.1158	Adam	64	0.0001	0.2	512
Endothelial	0.1182	Adam	128	0.01	0.0	0
Normal Epithelial	0.1500	RMSprop	64	0.0001	0.2	512
PVL	0.2651	Adam	128	0.01	0.0	0
Myeloid	0.3016	Adam	64	0.001	0.0	256
Cancer Epithelial	0.3582	Adam	32	0.0001	0.2	512
CAFs	0.3774	Adam	32	0.01	0.0	0
T cells	0.4220	Adam	32	0.0001	0.2	512
Plasmablasts	0.5938	Adam	32	0.0001	0.2	512

Table D.2: Parameters of neural networks trained on large patches from all six samples, yielding the highest slope of the best-fit line for each of the nine cell types in the testing dataset.

	Slope	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
Others	0.1096	Adam	256	0.0001	0.5	256
Invasive Cancer	0.3254	Adam	128	0.0001	0.2	512
Stroma	0.3283	Adam	64	0.0001	0.2	256
Lymphocyte	0.4861	Adam	32	0.0001	0.5	512

Table D.3: Parameters of neural networks trained on large patches from all samples, yielding the highest slope of the best-fit line for each of the four cell types in the testing dataset.

	Adjusted R-squared	Batch size	Learning rate	Dropout Proportion	Dense layer Units
B cells					
1142243F	0.033	32	0.0001	0.0	512
1160920F	0.249	64	0.001	0.0	256
CAFs					
1142243F	0.093	128	0.001	0.2	512
1160920F	0.075	128	0.0001	0.2	256
Cancer Epithelial					
1142243F	0.111	64	0.001	0.0	256
1160920F	0.289	32	0.001	0.0	256
Endothelial					
1142243F	0.026	32	0.001	0.0	256
1160920F	0.073	32	0.0001	0.2	512
Myeloid					
1142243F	0.062	64	0.001	0.0	256
1160920F	0.008	128	0.001	0.5	512
Normal Epithelial					
1142243F	0.024	128	0.001	0.0	512
1160920F	0.215	64	0.0001	0.0	256
PVL					
1142243F	0.012	32	0.001	0.5	512
1160920F	0.025	64	0.001	0.0	512
Plasmablasts					
1142243F	0.038	32	0.0001	0.2	512
1160920F	0.289	64	0.0001	0.0	512
T cells					
1142243F	0.162	64	0.0001	0.0	256
1160920F	0.312	64	0.001	0.0	256

Table D.4: Parameters of neural networks trained on standard patches from two independent lab samples, yielding the highest adjusted R-squared in the testing dataset for each cell type and each sample.

	Slope	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
Endothelial	0.0392	Adam	64	0.001	0.0	256
Myeloid	0.0911	Adam	64	0.001	0.0	256
PVL	0.1078	Adam	32	0.001	0.5	256
Normal Epithelial	0.1196	Adam	64	0.0001	0.0	256
B cells	0.1275	Adam	32	0.0001	0.0	512
CAFs	0.1418	Adam	64	0.0001	0.0	256
T cells	0.3351	Adam	64	0.0001	0.0	256
Cancer Epithelial	0.4008	Adam	64	0.0001	0.0	256
Plasmablasts	0.4587	Adam	64	0.0001	0.0	256

Table D.5: Parameters of neural networks trained on standard patches from two independent lab samples, yielding the highest slope of the best-fit line for each of the nine cell types in the testing dataset.

	Slope	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
Others	0.1224	Adam	128	0.0001	0.0	256
Stroma	0.1644	Adam	128	0.0001	0.0	512
Lymphocyte	0.3497	Adam	128	0.0001	0.0	512
Invasive Cancer	0.3705	Adam	128	0.0001	0.0	512

Table D.6: Parameters of neural networks trained on standard patches from two independent lab samples, yielding the highest slope of the best-fit line for each of the four cell types in the testing dataset.

	Slope	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
PVL	0.0067	Adam	32	0.001	0.0	512
T cells	0.0099	Adam	64	0.001	0.0	256
CAFs	0.0134	Adam	128	0.0001	0.0	512
B cells	0.0196	Adam	64	0.001	0.0	256
Endothelial	0.0296	Adam	64	0.001	0.0	256
Plasmablasts	0.0357	Adam	64	0.001	0.0	256
Cancer Epithelial	0.0373	Adam	64	0.0001	0.2	256
Normal Epithelial	0.0522	Adam	64	0.001	0.0	256
Myeloid	0.0704	Adam	32	0.001	0.0	512

Table D.7: Parameters of neural networks trained on standard patches from four samples processed by Wu et al., yielding the highest slope of the best-fit line for each of the nine cell types in the testing dataset.

	Adjusted R-squared	Batch size	Learning rate	Dropout Proportion	Dense layer Units
B cells					
CID4290	0.078	64	0.001	0.5	512
CID4465	0.068	32	0.001	0.5	256
CID44971	0.045	256	0.0001	0.2	512
CID4535	0.117	32	0.001	0.5	256
CAFs					
CID4290	0.041	256	0.001	0.2	256
CID4465	0.065	128	0.001	0.0	512
CID44971	0.034	128	0.0001	0.0	512
CID4535	0.075	128	0.0001	0.0	512
Cancer Epithelial					
CID4290	0.003	256	0.001	0.5	256
CID4465	0.017	64	0.001	0.5	256
CID44971	0.049	64	0.001	0.0	256
CID4535	0.064	256	0.001	0.5	256
Endothelial					
CID4290	0.011	256	0.0001	0.0	256
CID4465	0.028	256	0.0001	0.0	256
CID44971	0.013	128	0.0001	0.5	256
CID4535	0.172	64	0.001	0.2	512
Myeloid					
CID4290	0.022	256	0.001	0.5	512
CID4465	-0.001	128	0.0001	0.5	512
CID44971	-0.002	32	0.001	0.0	512
CID4535	0.021	32	0.001	0.2	512
Normal Epithelial					
CID4290	0.011	32	0.001	0.5	256
CID4465	0.010	32	0.001	0.2	512
CID44971	0.111	64	0.001	0.0	256
CID4535	0.045	128	0.001	0.5	512
PVL					
CID4290	0.011	256	0.001	0.0	256
CID4465	-0.002	256	0.0001	0.5	256
CID44971	0.030	64	0.001	0.0	256
CID4535	0.065	128	0.0001	0.0	512
Plasmablasts					
CID4290	0.014	256	0.001	0.5	512
CID4465	0.044	32	0.001	0.2	512
CID44971	0.101	32	0.0001	0.2	512
CID4535	0.043	128	0.001	0.5	256
T cells					
CID4290	0.103	64	0.0001	0.0	512
CID4465	0.005	256	0.001	0.0	512
CID44971	0.075	64	0.0001	0.0	256
CID4535	0.100	32	0.0001	0.0	512

Table D.8: Parameters of neural networks trained on standard patches from four samples processed by Wu et al., yielding the highest adjusted R-squared in the testing dataset for each cell type and each sample.

	Slope	Optimizer	Batch size	Learning rate	Dropout Proportion	Dense layer Units
Lymphocyte	0.0045	Adam	64	0.0001	0.0	256
Stroma	0.0109	Adam	128	0.0001	0.2	512
Others	0.0252	Adam	128	0.0001	0.2	512
Invasive Cancer	0.0303	Adam	128	0.0001	0.2	512

Table D.9: Parameters of neural networks trained on standard patches from four samples processed by Wu et al., yielding the highest slope of the best-fit line for each of the four cell types in the testing dataset.

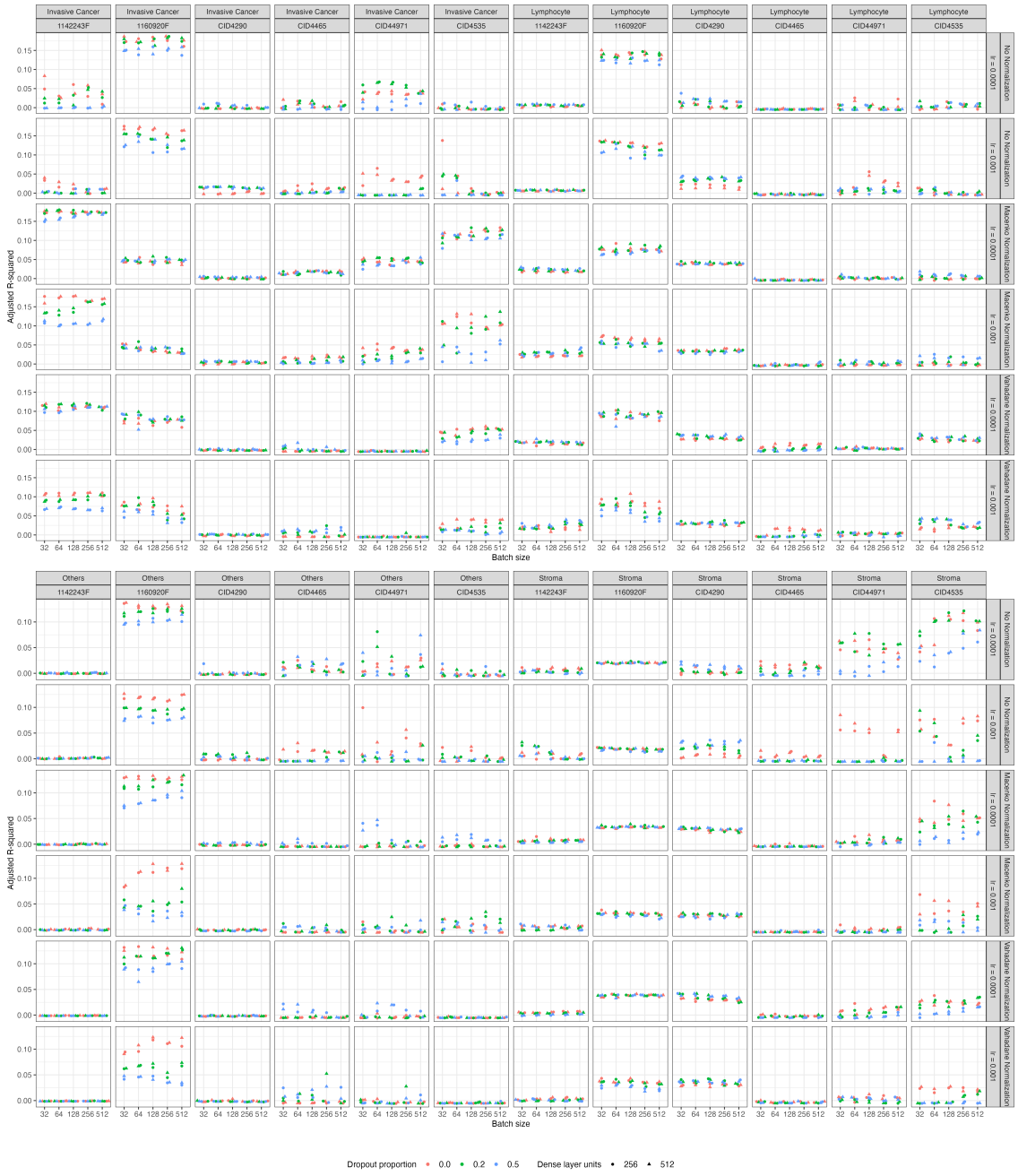


Figure D.1: Sample-stratified adjusted R-squared values of the testing dataset for all networks designed to predict the proportions of four cell types using standard patches from all six samples.

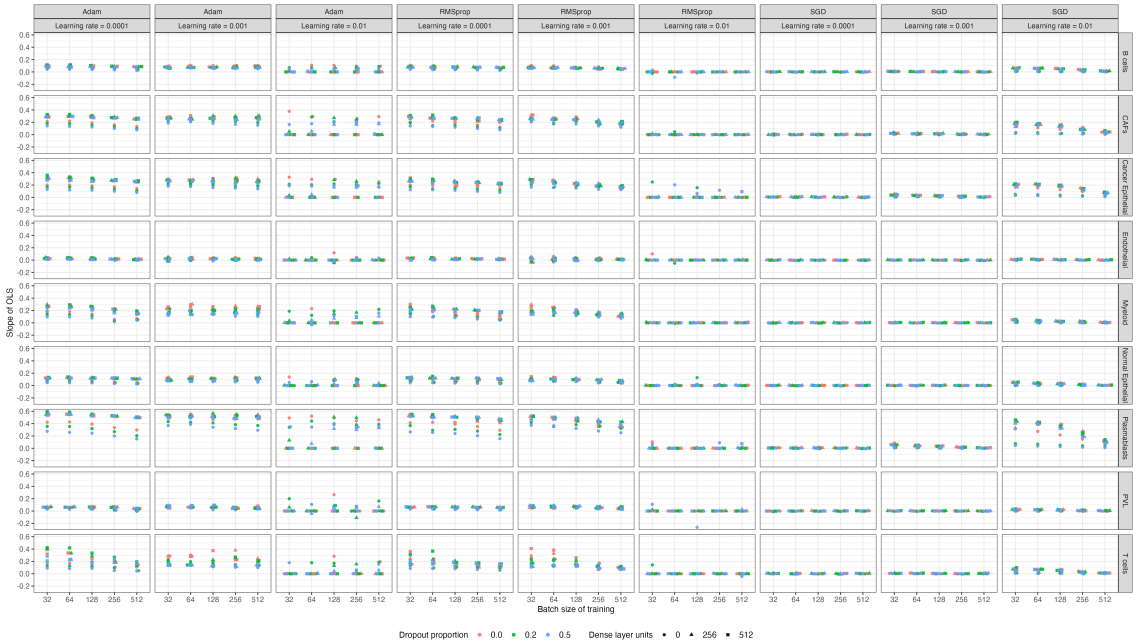


Figure D.2: Slope of the best-fit line between predicted and observed proportions of nine cell types in testing data, predicted by networks trained using large patches from all six samples.

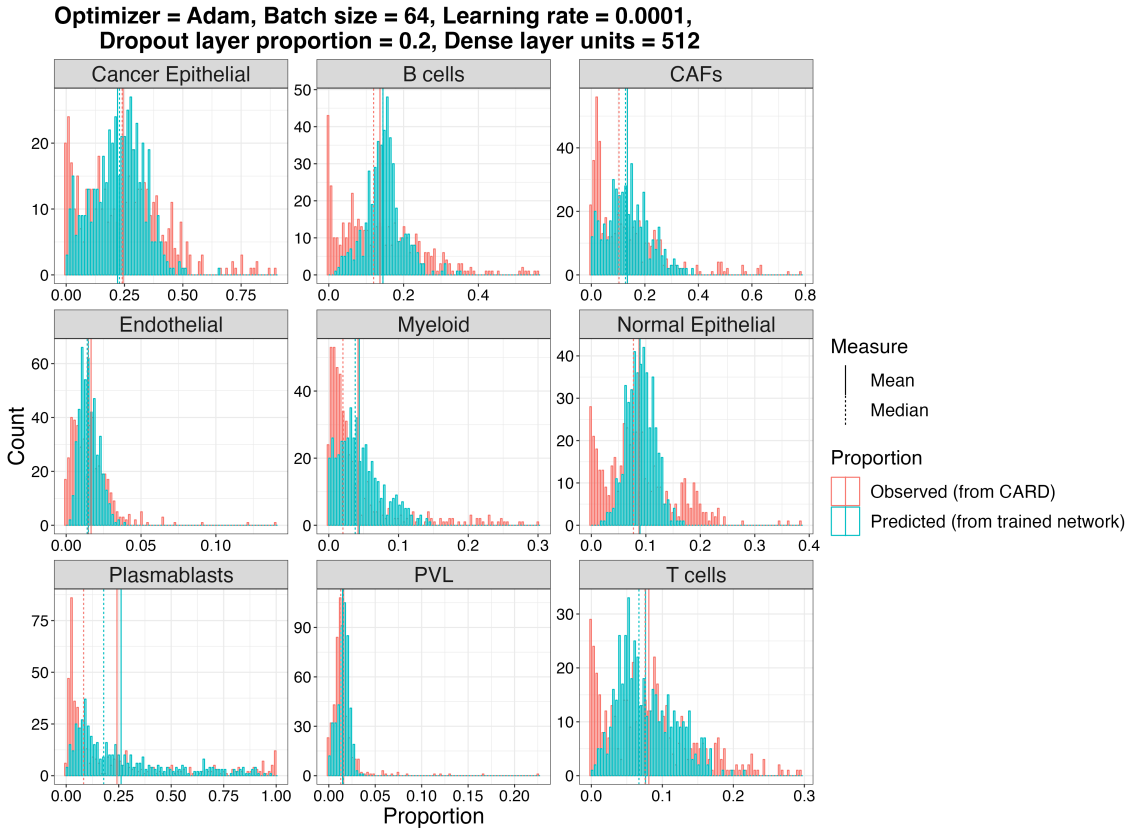


Figure D.3: Histogram comparing predicted and observed proportions for each of the nine cell types, where the predictions were made by a network trained using large patches from all six samples.

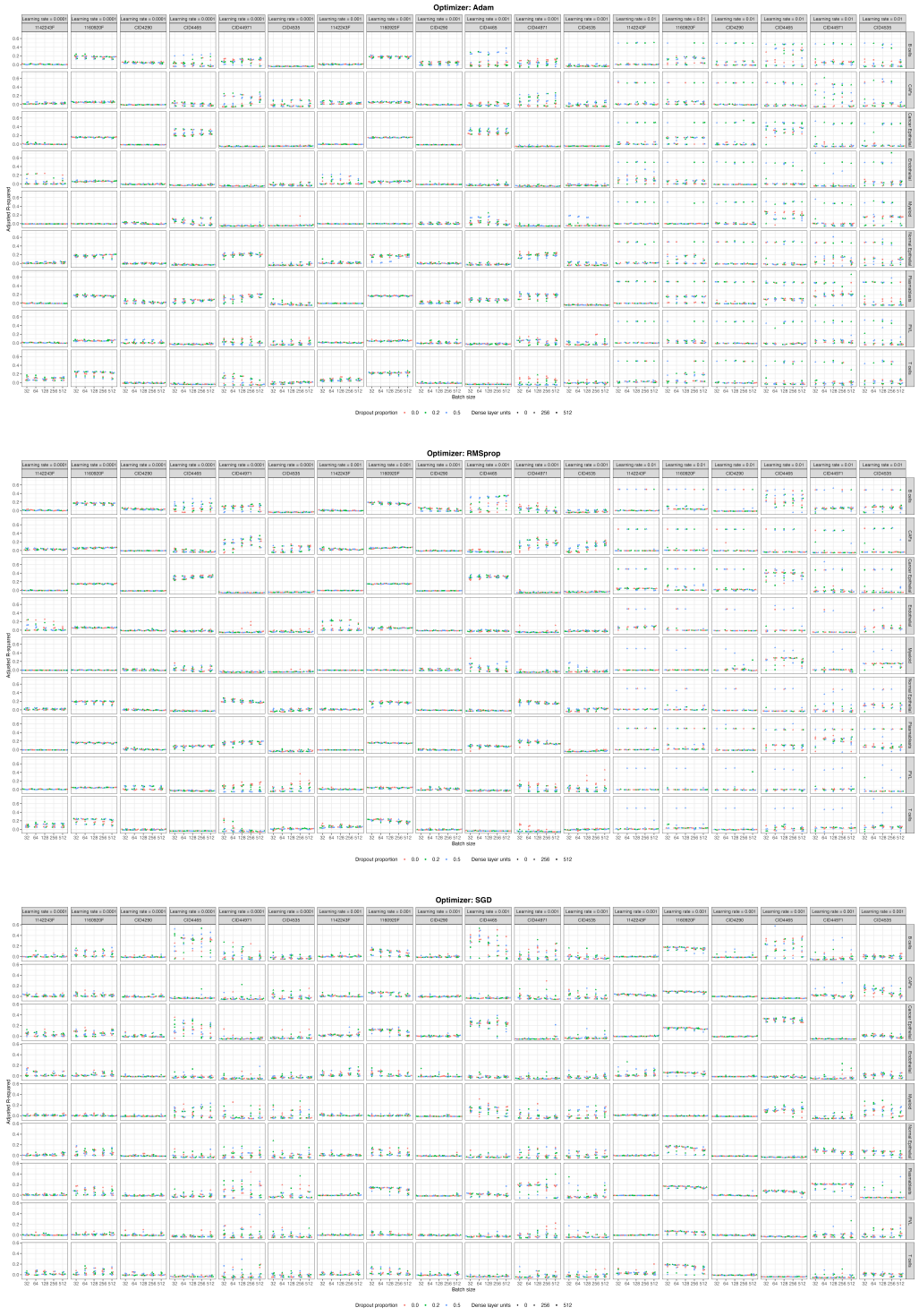


Figure D.4: Sample-stratified adjusted R-squared values of the testing dataset for all networks designed to predict the proportions of nine cell types using large patches from all six samples.

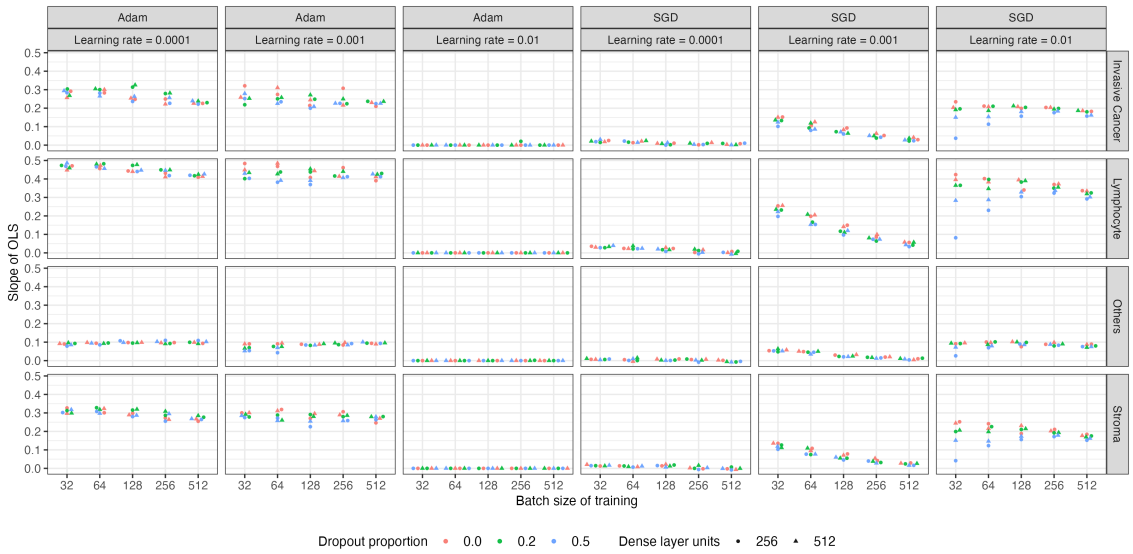


Figure D.5: Slope of the best-fit line between predicted and observed proportions of four cell types in testing data, predicted by networks trained using large patches from all six samples.

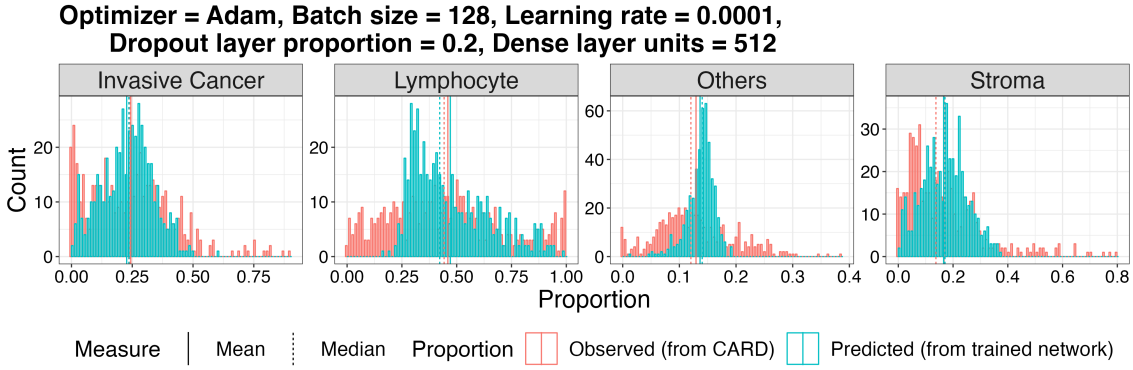


Figure D.6: Histogram comparing predicted and observed proportions for each of the four cell types, where the predictions were made by a network trained using large patches from all six samples.

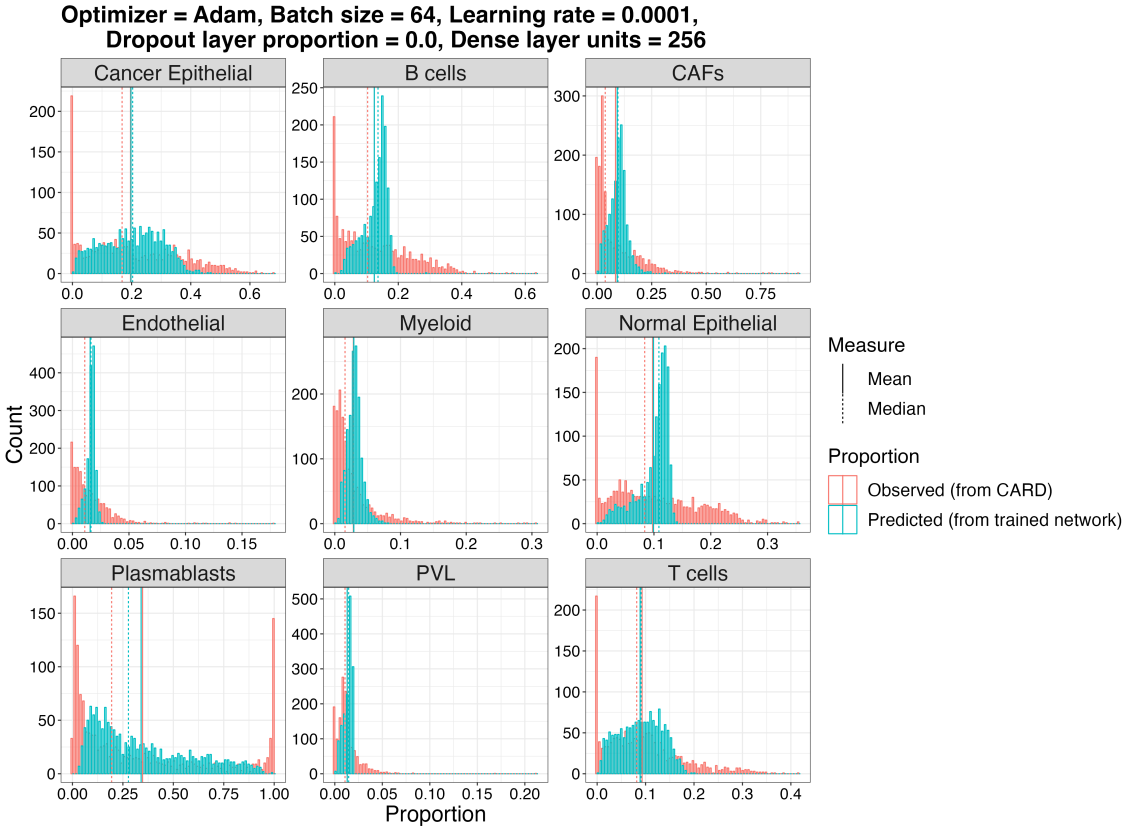


Figure D.7: Histogram comparing predicted and observed proportions for each of the nine cell types, where the predictions were made by a network trained using standard patches from two independent lab samples.

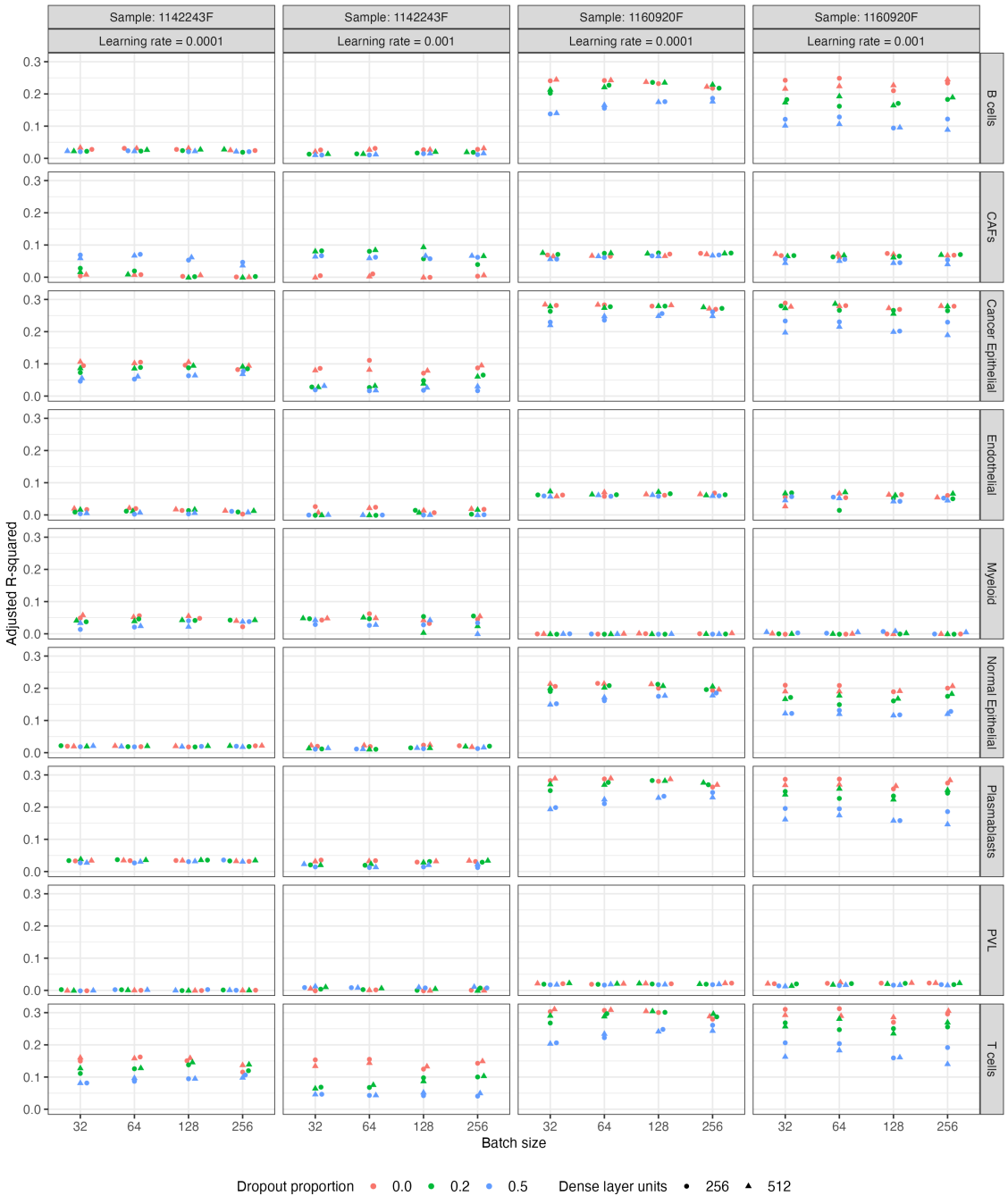


Figure D.8: Sample-specific adjusted R-squared values of the testing dataset for networks predicting proportions of nine cell types using standard patches from two independent lab samples.

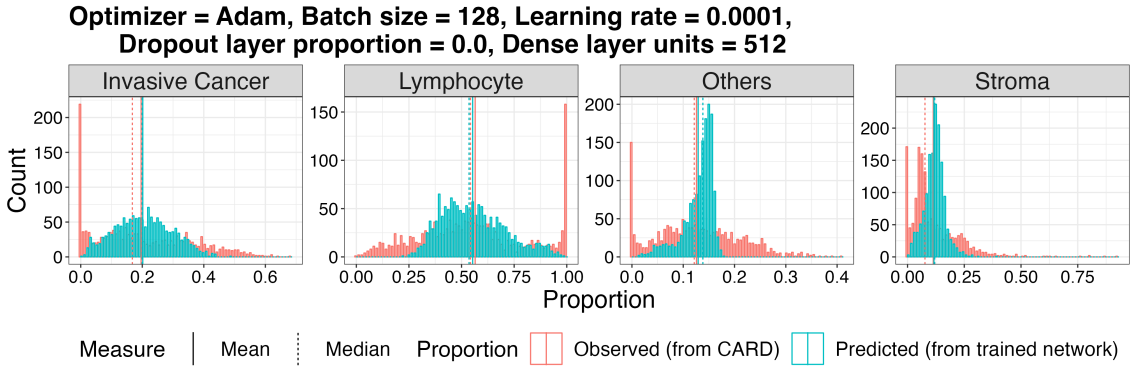


Figure D.9: Histogram comparing predicted and observed proportions for each of the four cell types, where the predictions were made by a network trained using standard patches from two independent lab samples.

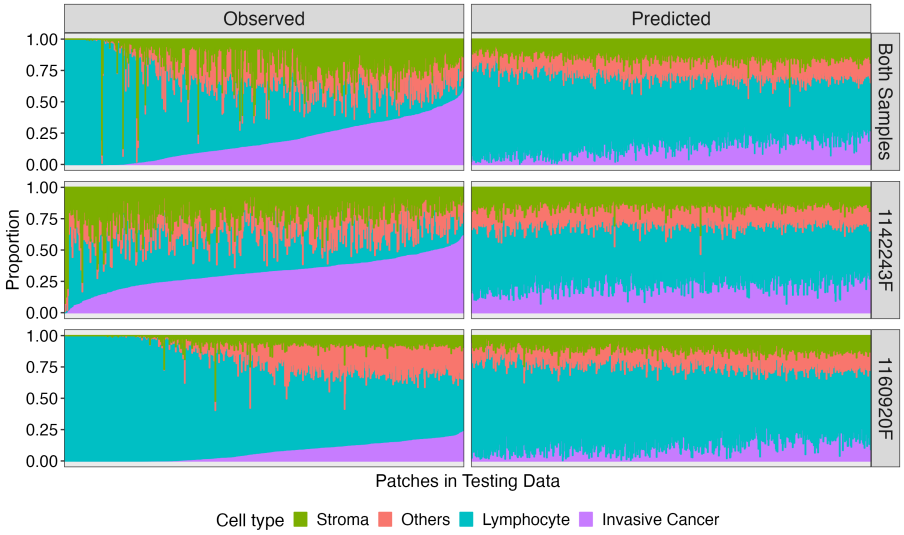


Figure D.10: Comparison of predicted and observed four cell type compositions in each of the standard patches from two independent lab samples.

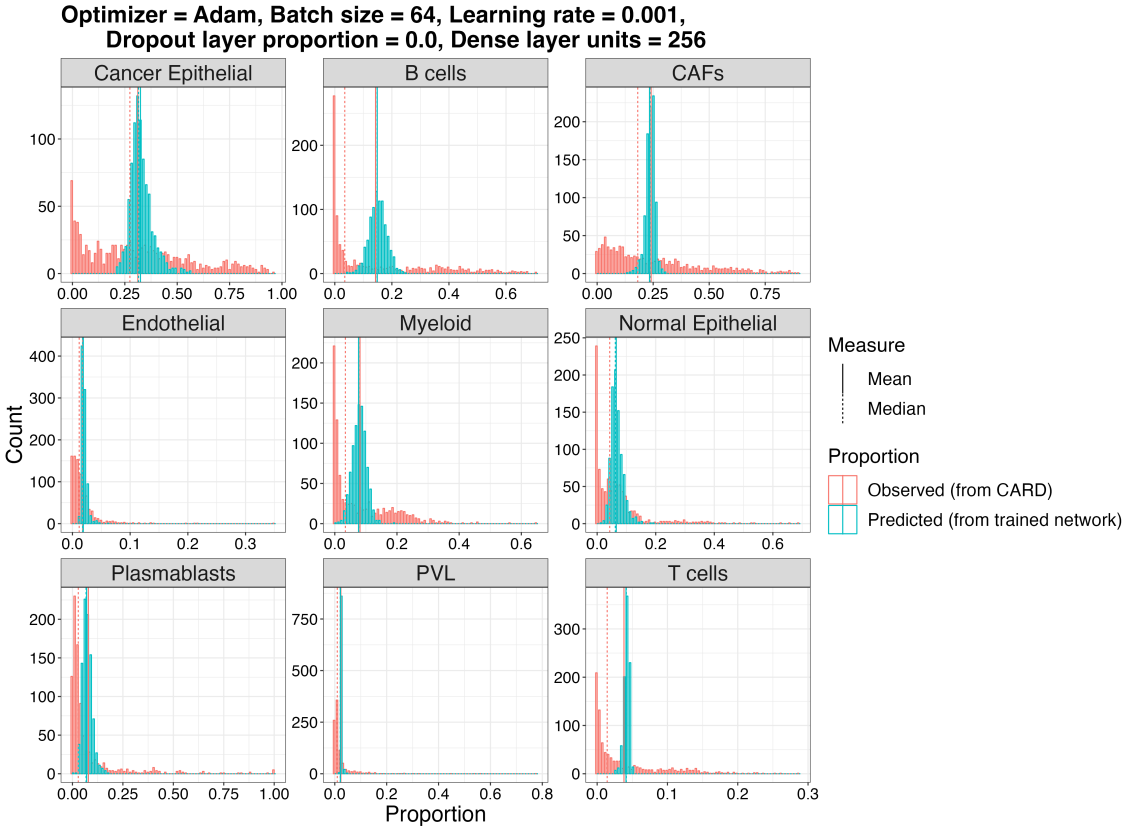


Figure D.11: Histogram comparing predicted and observed proportions for each of the nine cell types, where the predictions were made by a network trained using standard patches from four samples processed by Wu et al.

Figure D.12: Sample-specific adjusted R-squared values of the testing dataset for networks predicting proportions of nine cell types using standard patches from four samples processed by Wu et al.

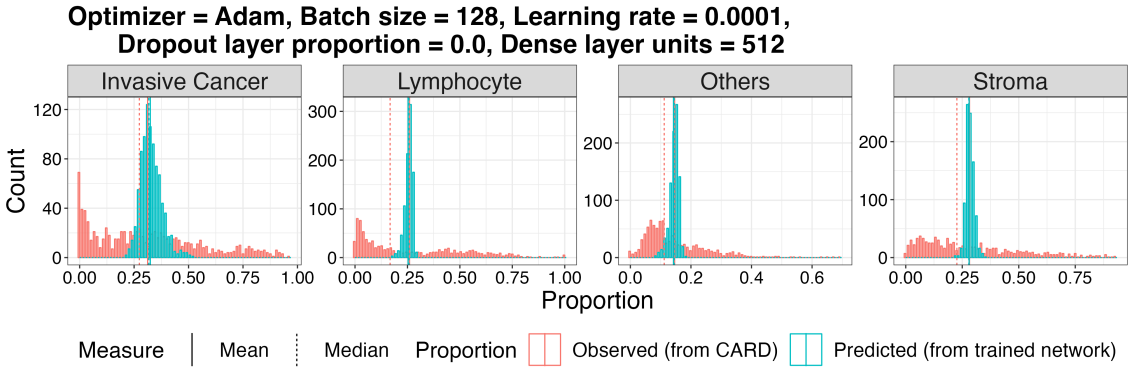


Figure D.13: Histogram comparing predicted and observed proportions for each of the four cell types, where the predictions were made by a network trained using standard patches from four samples processed by We et al.

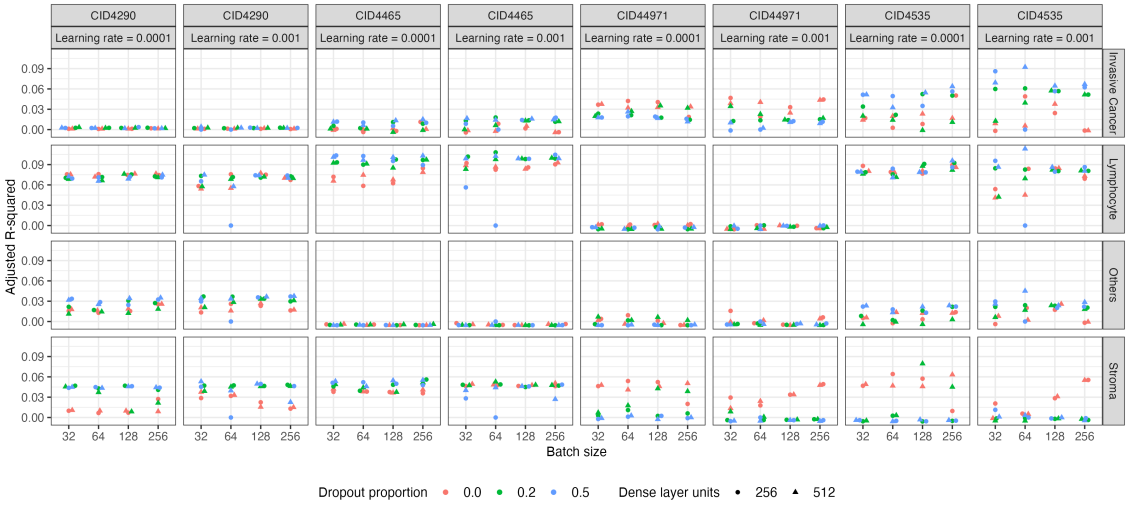


Figure D.14: Sample-specific adjusted R-squared values of the testing dataset for networks predicting proportions of four cell types using standard patches from four samples processed by Wu et al.

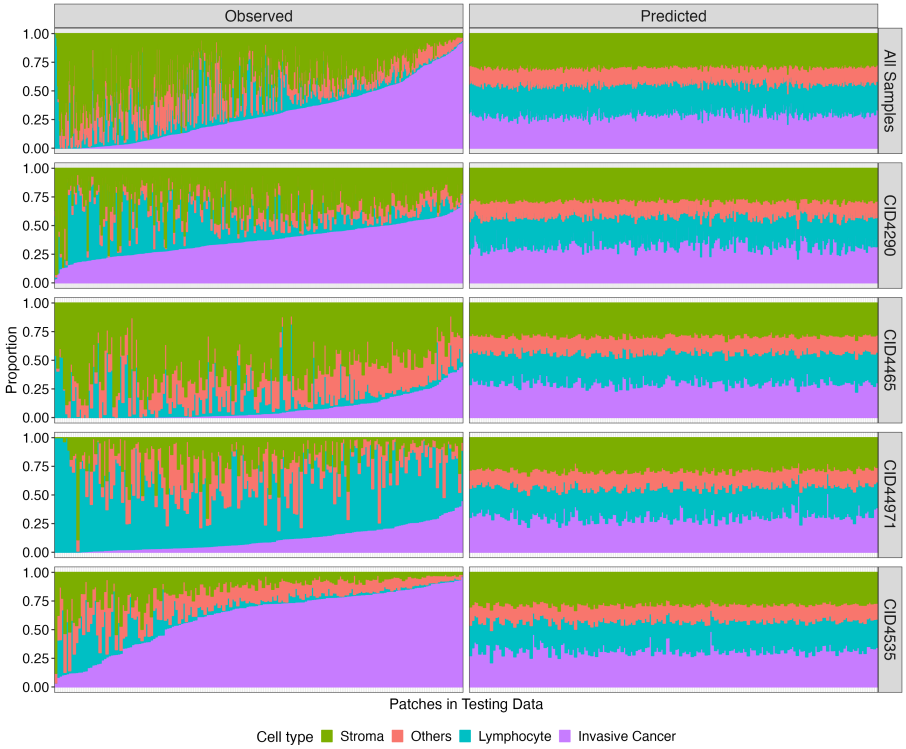


Figure D.15: Comparison of predicted and observed four cell type compositions in each of the standard patches from four samples processed by Wu et al. in the testing data.

Chapter E

Appendix for Chapter 7

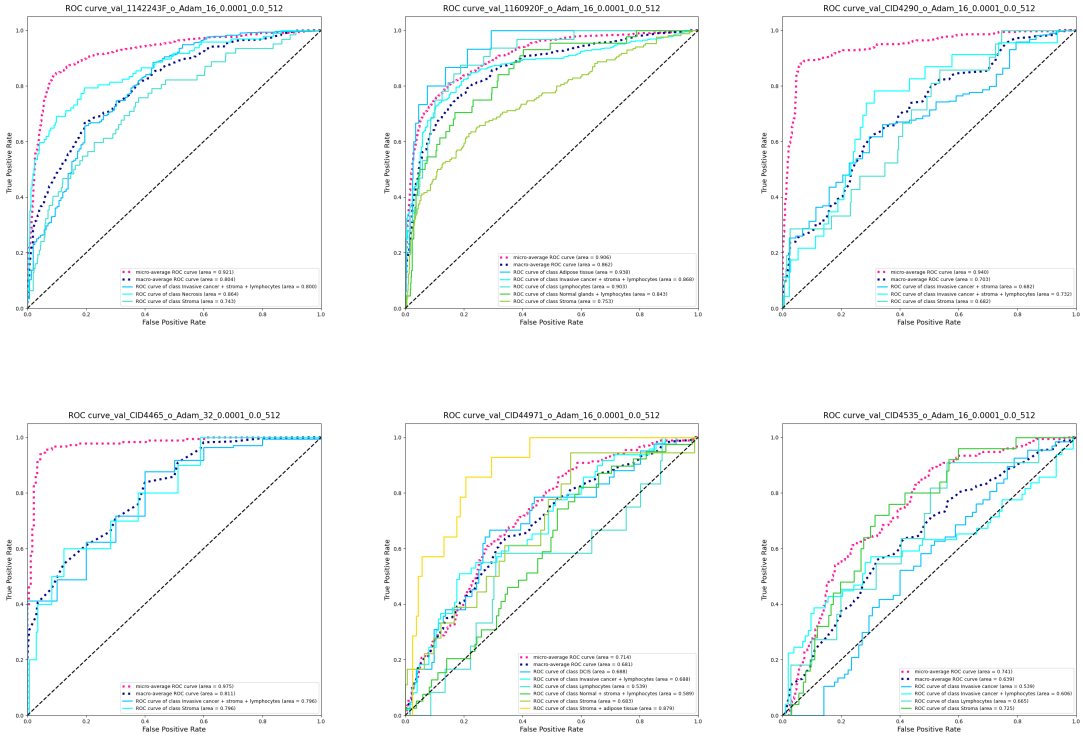


Figure E.1: ROC curves of the validation data of all six samples obtained by the networks that exhibited the lowest validation loss.

	Precision	Recall	f1-score
1142243F			
Macro-average	0.76	0.52	0.53
Micro-average	0.81	0.82	0.79
Invasive cancer + stroma + lymphocytes	0.84	0.96	0.90
Necrosis	0.68	0.56	0.6
Stroma	0.75	0.05	0.09
1160920F			
Macro-average	0.53	0.52	0.49
Micro-average	0.75	0.72	0.72
Adipose tissue	0.50	0.20	0.29
Invasive cancer + stroma + lymphocytes	0.85	0.88	0.86
Lymphocytes	0.26	0.56	0.36
Normal glands + lymphocytes	0.32	0.52	0.40
Stroma	0.72	0.42	0.53
CID4290			
Macro-average	0.29	0.33	0.31
Micro-average	0.77	0.88	0.82
Invasive cancer + stroma	0.88	1.00	0.94
Invasive cancer + stroma + lymphocytes	0.00	0.00	0.00
Stroma	0.00	0.00	0.00
CID4465			
Macro-average	0.47	0.50	0.49
Micro-average	0.89	0.94	0.92
Invasive cancer + stroma + lymphocytes	0.94	1.00	0.97
Stroma	0.00	0.00	0.00
CID44971			
Macro-average	0.18	0.21	0.17
Micro-average	0.26	0.32	0.25
DCIS	0.29	0.81	0.43
Invasive cancer + lymphocytes	0.41	0.33	0.36
Lymphocytes	0.00	0.00	0.00
Normal + stroma + lymphocytes	0.35	0.15	0.21
Stroma	0.00	0.00	0.00
Stroma + adipose tissue	0.00	0.00	0.00
CID4535			
Macro-average	0.11	0.25	0.15
Micro-average	0.19	0.43	0.27
Invasive cancer	0.44	0.99	0.61
Invasive cancer + lymphocytes	0.00	0.00	0.00
Lymphocytes	0.00	0.00	0.00
Stroma	0.00	0.00	0.00

Table E.1: Class-specific and averaged classification metrics in the validation data of each of the six samples.

	Precision	Recall	f1-score
1142243F			
Macro-average	0.56	0.52	0.51
Micro-average	0.78	0.82	0.79
Invasive cancer + stroma + lymphocytes	0.86	0.94	0.90
Necrosis	0.57	0.59	0.58
Stroma	0.25	0.03	0.06
1160920F			
Macro-average	0.50	0.49	0.47
Micro-average	0.76	0.71	0.72
Adipose tissue	0.43	0.27	0.33
Invasive cancer + stroma + lymphocytes	0.84	0.86	0.85
Lymphocytes	0.17	0.45	0.25
Normal glands + lymphocytes	0.27	0.44	0.34
Stroma	0.78	0.44	0.56
CID4290			
Macro-average	0.28	0.33	0.31
Micro-average	0.72	0.85	0.78
Invasive cancer + stroma	0.85	1.00	0.92
Invasive cancer + stroma + lymphocytes	0.00	0.00	0.00
Stroma	0.00	0.00	0.00
CID4465			
Macro-average	0.48	0.50	0.49
Micro-average	0.90	0.95	0.93
Invasive cancer + stroma + lymphocytes	0.95	1.00	0.97
Stroma	0.00	0.00	0.00
CID44971			
Macro-average	0.35	0.24	0.21
Micro-average	0.40	0.32	0.27
DCIS	0.26	0.81	0.39
Invasive cancer + lymphocytes	0.40	0.19	0.25
Lymphocytes	0.00	0.00	0.00
Normal + stroma + lymphocytes	0.47	0.36	0.41
Stroma	0.00	0.00	0.00
Stroma + adipose tissue	1.00	0.10	0.18
CID4535			
Macro-average	0.09	0.25	0.13
Micro-average	0.13	0.36	0.19
Invasive cancer	0.36	1.00	0.53
Invasive cancer + lymphocytes	0.00	0.00	0.00
Lymphocytes	0.00	0.00	0.00
Stroma	0.00	0.00	0.00

Table E.2: Class-specific and averaged classification metrics in the testing data of each of the six samples.

Barcode	Image types			Label by Sudmerier et al.
	Original	Macenko	Vahadane	
AATAGAACAGAGTGGC-1	0	0	34	Inflammation - Cluster 1
ACAGGTGGAGGTGAGG-1	0	0	2	Inflammation - Cluster 1
AGCACTACCGGCCTGT-1	0	36	4	Inflammation - Cluster 1
CCAGAGACAAAGCCGG-1	0	1	0	Inflammation - Cluster 1
CTGGATTTACACTTGA-1	0	0	3	Inflammation - Cluster 1
GACGCTTGCTTCTAAA-1	0	0	34	Inflammation - Cluster 1
GCGATGTCTGTGCTTG-1	0	2	0	Inflammation - Cluster 1
GGGCGTCACCACGTAA-1	0	0	34	Inflammation - Cluster 1
GGTGAAGTACAGGGAT-1	0	1	0	Inflammation - Cluster 1
GTAATCTGATTCTTCG-1	0	1	0	Inflammation - Cluster 1
GTCTATCTGAGTTTCT-1	9	0	0	Inflammation - Cluster 1
GTTTGGTAGGGTCAAC-1	0	0	12	Inflammation - Cluster 1
TCATCACTCGAGCTCG-1	0	5	0	Inflammation - Cluster 1
TGATGGGACTAAGTCA-1	0	3	0	Inflammation - Cluster 1
TTCATAGGGTGTCCAT-1	0	29	0	Inflammation - Cluster 1
TTCCGGCCTTGAGGCT-1	0	1	0	Inflammation - Cluster 1
AAACAGGGTCTATATT-1	17	0	0	Inflammation - Cluster 7
ACATGGCTCAATTTAG-1	0	17	0	Inflammation - Cluster 7
ACCCTATAGGACTGAG-1	0	5	0	Inflammation - Cluster 7
AGACCATGGGATACAA-1	0	10	0	Inflammation - Cluster 7
AGGCCCAGTGACTGGT-1	0	16	0	Inflammation - Cluster 7
AGTTGCTGACTGATAT-1	0	10	0	Inflammation - Cluster 7
ATCTAACGTCCCTATG-1	0	36	34	Inflammation - Cluster 7
ATTGATAGCAACGAGA-1	0	5	0	Inflammation - Cluster 7
CAGTAATCCCTCCCAG-1	5	0	0	Inflammation - Cluster 7
CAGTGGTTGCACATGA-1	0	1	0	Inflammation - Cluster 7
CCAAAGCAGTTGGTTG-1	0	17	0	Inflammation - Cluster 7
CGTCCAGATGGCTCCA-1	0	19	0	Inflammation - Cluster 7
GAGGAGATCCTCATGC-1	7	0	0	Inflammation - Cluster 7
GCCAGGCTTAGTGGTA-1	0	36	0	Inflammation - Cluster 7

Table E.3: Number of times a spot that was identified as inflammation by Sudmeier et al. was misclassified by classifiers being trained by original images, Macenko normalized images, and Vahadane normalized images.

Barcode	Image Type			Label by Sudmerier et al.
	Original	Macenko	Vahadane	
AAATCGCGGAAGGAGT-1	7	0	2	Tumor - Cluster 4
AACCCATCCCATGATC-1	0	0	2	Tumor - Cluster 4

Table E.4 continued from previous page

Barcode	Image Type			Label by Sudmerier et al.
	Original	Macenko	Vahadane	
AACGAAAGTCGTCCCA-1	0	0	2	Tumor - Cluster 4
ACGCGTTTCTTAAGAG-1	36	0	2	Tumor - Cluster 4
ACGCTGTGAGGCGTAG-1	0	8	2	Tumor - Cluster 4
ACTCGTAACCCGTCCT-1	0	0	2	Tumor - Cluster 4
ACTTTGACTGCATCCT-1	0	23	11	Tumor - Cluster 4
AGACCAAACCACACCT-1	0	0	2	Tumor - Cluster 4
AGGATCACGCGATCTG-1	36	36	2	Tumor - Cluster 4
AGGTCAGGTGAGAGTG-1	0	0	2	Tumor - Cluster 4
CATATGTCAGGCTACG-1	0	0	4	Tumor - Cluster 4
CCATAACCTGTGCAGT-1	0	13	2	Tumor - Cluster 4
CCTTTGAATTATGGCT-1	1	19	2	Tumor - Cluster 4
CGCCACAGGTCGCGAT-1	0	0	2	Tumor - Cluster 4
CTCGAGACATACGATA-1	0	0	2	Tumor - Cluster 4
GCTGGTTTAGGCCATA-1	0	0	2	Tumor - Cluster 4
GGACAAGTTGCAGTGA-1	36	0	2	Tumor - Cluster 4
GGCGAGCGAAACGGCA-1	0	0	2	Tumor - Cluster 4
TACCTATCCCTAGAGG-1	0	0	2	Tumor - Cluster 4
TAGAGATCATGCAACT-1	0	0	2	Tumor - Cluster 4
TGACATATATGACGAT-1	0	0	2	Tumor - Cluster 4
TGGCATGAAGTTTGGG-1	0	0	2	Tumor - Cluster 4
TGGTCGTGCAAGGCAA-1	0	0	2	Tumor - Cluster 4
TTGGAAGAATACAGTC-1	0	0	2	Tumor - Cluster 4
AAACCGGGTAGGTACC-1	30	0	2	Tumor - Cluster 5
AACGTTATCAGCACCT-1	36	0	2	Tumor - Cluster 5

Table E.4 continued from previous page

Barcode	Image Type			Label by Sudmerier et al.
	Original	Macenko	Vahadane	
AAGTGACGACCGAATT-1	0	0	2	Tumor - Cluster 5
AATAGTCGCGAGTCGG-1	0	0	2	Tumor - Cluster 5
AGCCCATACATGTAAG-1	0	0	2	Tumor - Cluster 5
AGTCGTGGGCATTACG-1	0	3	2	Tumor - Cluster 5
AGTCTCACAAGACTAC-1	0	0	2	Tumor - Cluster 5
AGTTACCGCACATGGT-1	0	0	2	Tumor - Cluster 5
AGTTTGGCCAGACCTA-1	25	22	2	Tumor - Cluster 5
ATCGTATTCCGAGAAC-1	1	0	2	Tumor - Cluster 5
CCACTGGTGGCTGGTT-1	7	0	2	Tumor - Cluster 5
CCAGCTTCCGCCCGCA-1	36	0	2	Tumor - Cluster 5
CCCGTAAGTCTAGGCC-1	36	36	36	Tumor - Cluster 5
CCTACATTCACAGACG-1	0	35	2	Tumor - Cluster 5
CGCCTGGCCTACGTAA-1	0	0	2	Tumor - Cluster 5
CTCTCGCTGTACTATG-1	36	0	32	Tumor - Cluster 5
CTGTTGGCTCTTCTGA-1	36	0	36	Tumor - Cluster 5
CTTAGCCCGGATAGTG-1	0	0	5	Tumor - Cluster 5
CTTTGGCGCTTTATAC-1	0	0	2	Tumor - Cluster 5
GATATGGATTACGCGG-1	0	0	2	Tumor - Cluster 5
GATCTTCATTGTCCTC-1	0	0	9	Tumor - Cluster 5
GATGCGAATGGTATTA-1	0	0	2	Tumor - Cluster 5
GCACAAGTGGATCATA-1	0	0	36	Tumor - Cluster 5
GCTCAATCCGTTTATT-1	0	0	5	Tumor - Cluster 5
GGGCACGAATTGGCCG-1	0	0	2	Tumor - Cluster 5
GGTACATCTGGGACGA-1	7	0	2	Tumor - Cluster 5

Table E.4 continued from previous page

Barcode	Image Type			Label by Sudmerier et al.
	Original	Macenko	Vahadane	
GTCATTGCATTGACCC-1	36	0	2	Tumor - Cluster 5
GTGGACCAACCCGATT-1	0	0	2	Tumor - Cluster 5
GTTCAAATCAGATGTC-1	0	0	35	Tumor - Cluster 5
TAAACGTCGTCAATGA-1	0	0	2	Tumor - Cluster 5
TAATTTCCGTCCAGTA-1	0	0	2	Tumor - Cluster 5
TACTCGGCACGCCGGG-1	0	0	2	Tumor - Cluster 5
TATAGGGTACTCATGA-1	0	0	2	Tumor - Cluster 5
TATCTAGCCTAAAGGA-1	0	0	2	Tumor - Cluster 5
TCGCCGGATGGGCAAG-1	0	0	2	Tumor - Cluster 5
TCTGCAGATTCGAGTC-1	0	0	2	Tumor - Cluster 5
TGAAGAGCGGTCCTAG-1	0	0	2	Tumor - Cluster 5
TGGTTATGCTTGCGGT-1	0	0	2	Tumor - Cluster 5
TGTGTTCGTATCCAAG-1	0	0	36	Tumor - Cluster 5
TGTTCATAAATGTGCT-1	0	0	2	Tumor - Cluster 5
TTAAACCTGGTTCCTT-1	0	0	2	Tumor - Cluster 5
TTAACTCACGCGTGGA-1	0	0	2	Tumor - Cluster 5
TTGGACATGTGGCTTA-1	36	36	36	Tumor - Cluster 5

Table E.4: Number of times a patch that was identified as tumor by Sudmeier et al. was misclassified by classifiers being trained by original images, Macenko normalized images, and Vahadane normalized images.