Stain by numbers: toward the next generation of biomarkers in cancer systems biology

Author:
Erik Ames BURLINGAME

Supervisor:
Dr. Young Hwan CHANG

A DISSERTATION

Presented to the
Department of Biomedical Engineering
Oregon Health & Science University
School of Medicine
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

September 7, 2021

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of

Erik Ames BURLINGAME

has been approved

Young Hwan Chang	08/25/2021
Mentor	Date
Joe W Gray	08/25/2021
Member	Date
Christopher Corless	08/26/2021
Member	Date
Guillaume Thibault	08/27/2021
Member	Date
Xubo Song	09/02/2021
Member	Date
Gali Jia	09/06/2021
Member	Date

Contents

List of Figures

List of Tables

List of Abbreviations

BC	**B**reast **C**ancer
cGAN	**c**onditional **G**enerative **A**dversarial **N**etwork
CK	**C**yto**K**eratin
CyCIF	**Cy**Clic **I**mmuno**F**luorescence
DL	**D**eep **L**earning
GAN	**G**enerative **A**dversarial **N**etwork
GPU	**G**raphics **P**rocessing **U**nit
H&E	**H**ematoxylin and **E**osin
IF	**I**mmuno**F**luorescence
IHC	**I**mmuno**H**isto**C**hemistry
IMC	**I**maging **M**ass **C**ytometry
JSD	**J**ensen-**S**hannon **D**ivergence
LFD	**L**abel-**F**ree **P**rediction
MCC	**M**atthews **C**orrelation **C**oefficient
ML	**M**achine **L**earning
MSE	**M**ean **S**quared **E**rror
MTI	**M**ultiplexed **T**issue **I**maging
mIHC	**m**ultiplex **I**mmuno**H**isto**C**hemistry
PDAC	**P**ancreatic **D**uctal **A**deno**C**arcinoma
PSNR	**P**eak **S**ignal-to-**N**oise **R**atio
ROI	**R**egion **O**f **I**nterest

SHIFT **S**peedy **H**istological-to-**I**mmuno**F**luorescent **T**ranslation

SMA **S**mooth **M**uscle **A**ctin

SSIM **S**tructural **SIM**ilarity

TMA **T**issue **M**icro**A**rray

VAE **V**ariational **A**uto**E**ncoder

WSI **W**hole **S**lide **I**mage

Abstract

Cancers are complex diseases that operate at multiple biological scales—from atom to organism—and the purview of cancer systems biology is to integrate information between scales to derive insight into their mechanisms and therapeutic vulnerabilities. From this holistic perspective, the field has come to appreciate that the spatial context of the tumor microenvironment in intact tissues not only enables a more granular definition of disease, but also the design of more personalized and effective therapies. In spite of this promise, spatial context-preserving cytometry paradigms like multiplex tissue imaging (MTI) are beset with many challenges related to cost, computational complexity, and study design. In this work, we introduce computational approaches to integrate, analyze, and interpret MTI data which address some of these challenges. First, we present two deep learning methods which (1) leverage morphological features captured in digitized pathology slides to learn realistic and precise virtual stains which can be deployed for a fraction of the cost and in a fraction of the time required by conventional MTI, and (2) define an optimal sample selection strategy which improves the generalizability of virtual staining models. Second, we extend virtual staining to reconstruct the 3D microenvironment of a tumor resection and present a deep learning approach to the integration of histology and MTI data with implications for objective region-of-interest selection in whole slide images. Finally, we present a computationally-efficient machine learning workflow for reproducible, scalable, and robust analysis of single-cell MTI data, and the first cross-validation of breast cancer cell phenotypes derived using two different MTI platforms. The discovery and development of the next generation of biomarkers in cancer systems biology will require computational tools which can cope with the increasing scale and complexity of our measurements, and the work we share within serves as a step toward achieving that requirement.

Acknowledgements

Foremost, I would like to thank my mentor Dr. Young Hwan Chang for his unwavering commitment to my scientific training—first as an intern, then as an out-of-department graduate student, and finally as a graduate student in his own group—and especially for passing on his enthusiasm for outstanding research. Also, thank you to my committee members and examiners Dr. Joe W. Gray, Dr. Christopher Corless, Dr. Guillaume Thibault, Dr. Xubo Song, and Dr. Yali Jia, for their time and feedback throughout the development of this dissertation.

I would further like to thank the members of the Quantitative BioImaging Lab, especially Dr. Geoffrey Schau and Luke Ternes for the synergy in collaboration. I hope you all feel, as I do, that we made each other better scientists through our interactions.

Finally, thank you, Taylor, for making this whole enterprise worth it.

Dedicated to all mothers, grandmothers, great-grandmothers, and so on.

Chapter 1

Introduction

Lud! child, how stupid you are!
There's [turtles] all the way down!

Unwritten Philosophy

1.1 Image-based characterization in cancer systems biology

1.1.1 Brief summary

Physicians depend on histopathology—the visualization and pathological interpretation of thin sections of biopsied tissue—as an essential indicator of disease. For instance, imaging thin sections of formalin-fixed, paraffin-embedded (FFPE), and hematoxylin and eosin (H&E)-stained biopsy tissue provides low-cost, rapid, and direct insight into the cancer tissue morphology which guides diagnosis, grading and staging, and prognosis. Additionally, determining the spatially-resolved molecular profile of a cancer is important for disease subtyping and choosing a patient's course of treatment [29], as is routine in breast cancer (BC) for determining which receptor status subtype a patient's tumor is presenting [131] and whether or not receptor expression can be targeted as a therapeutic vulnerability of the disease.

Many of the fundamental techniques used in histopathology have gone long unchanged. In particular, the H&E stain combination has been in use and largely unchanged since its first description in 1876 [126], while immunohistochemistry (IHC), and to a much lesser extent immunofluorescence (IF), became prominent in the last quarter of the 20th century [91]. Pathologists's enduring preference for light-absorptive IHC over light-emissive IF is largely driven by the compatibility of IHC with hematoxylin counterstaining, which provides morphological context for marker expression that is easily assessed by light microscopy. However, due to the enzymatic basis of signal amplification, limited dynamic range due to the physical properties of chromogens, and propensity for saturation in IHC assays, IHC-based clinical measures of marker expression are typically reported in binary or nominal scores that are inherently non-quantitative. Moreover, it is common for the expression of only a few markers to be assessed due to the colorimetric limitations of single-slide imaging [91]. This coarseness and sparseness of assessment can belie the complexity of the disease. If we aim to understand the organization and interaction of tumor and non-tumor cells that is now accepted to be critical for developing effective treatments [130], then the need for continuous, quantitative, multiplexed, and spatially-resolved measures of marker expression is clear.

To meet this demand, numerous multiplex tissue imaging (MTI) and molecular quantitation platforms have been developed (Table 1.1). The choice of MTI platform depends on the scope and scale of the questions asked, as well as the tolerance or preparedness for dealing with technical confounders. For instance, the fluorescence-based platforms like

Platform	Developer	Multiplexing	Resolution	Runtime	Throughput	Probe	Reference
CyCIF	Harvard	high (>50)	very high	weeks	high	fluor-Ab	[60]
MxIF	GE	high (>50)	high	weeks	high	fluor-Ab	[35]
mIHC	OHSU	low (∼30)	medium	days	medium	enzyme-Ab	[111]
IMC	Fluidigm	high (>50)	low	days	low	metal-Ab	[36]
MIBI	Ionpath	high (>50)	low	days	low	metal-Ab	[4]
CODEX	Akoya	high (>50)	high	days	medium	DNA-fluor/Ab	[37]

TABLE 1.1: Comparisons of multiplex characterization platforms. Adapted from a slide by Jia-Ren Lin, Co-Director of the Tissue Imaging Platform at the Labratory of Systems Pharmacology, Harvard Medical School.

cyclic immunofluorescence (CyCIF) [60], multiplex immunofluorescence (MxIF) [35], and co-dectection by indexing (CODEX) [37] maximize the spatial context of their measurements through their ability to characterize whole tissue slides, but must contend with the natural autofluorescence of FFPE tissues which can confound single-cell measurements of marker expression. By contrast, the mass spectrometry-based platforms like imaging mass cytometry (IMC) [36] and multiplex ion beam imaging (MIBI) [4] have relatively limited resolution and field of view, but also have relatively high signal-to-background ratios by virtue of a detection approach which circumvents the autofluorescence complication.

Among these MTI platforms, CyCIF stands out because it enables visualization of tens or hundreds of markers in whole slides and it uses milder label stripping conditions than multiplex immunohistochemistry (mIHC) [111], making it less prone to tissue and antigen degradation. Together with the keen definition of tissue morphology provided by H&E, CyCIF can provide clear pictures of where specific cell types lie within the tissue, help functionally characterize the tumor microenvironment, and resolve questions of cancer staging and metastatic origin.

Despite the richness of information obtained by CyCIF and other MTI platforms in 2D tissue sections, the data are limited to the information contained within a single tissue section from a single biopsy from a single tumor, which could bias analysis and interpretation of the tumor bulk. Further, this essentially 2D representation of tissue is a relatively poor representation of tissues like prostate, pancreas, breast, and colon which have highly convoluted 3D ductal or glandular structures. Motivated by the undersampling and misrepresentation challenges of 2D pathology—where a standard 5 µm tissue section can represent just a fraction of a percent of a whole specimen and a cross-sectional view of convoluted 3D structures—several recent studies have applied MTI or standard H&E staining to the full set of serial sections of whole specimens to reconstruct 3D atlases of the tumor microenvironment [54, 15, 61], heralding a new era of unprecedented measurable depth and spatial resolution in cancer biology.

1.1.2 Challenges to MTI applications in cancer systems biology

The advance of MTI promises to increase our understanding of heterogeneity and cellular interactions within the tumor microenvironment, both of which play increasingly important roles in the development of effective treatments [130]. Although its clinical potential is immense, CyCIF and other MTI platforms are time- and labor-intensive, technically complicated, and high-cost, so assessment is typically limited to only a small subset of a given biopsy, which is unlikely to be fully representative of a patient's disease. Also, the cost associated with MTI will undoubtedly limit its use to within highly-developed clinical settings for the foreseeable future, further widening the quality-of-care gap between high- and low-income communities. Until MTI matures into an economy of scale, these challenges will only be further amplified in 3D applications. The technological gap between H&E and MTI technologies highlights the broader need for automated tools that leverage information attained by a low-cost technique to infer information typically attained by a high-cost technique.

Aside from the *ex silico* challenges of feasibility and accessibility in MTI, downstream data analysis is fraught with *in silico* computational challenges related to the increasing scale and complexity of MTI data and the increasing need to cross-validate findings within and between MTI platforms. To enable the biggest discoveries, the MTI meta-analyses of the near future will require the integration of billions of single-cell measurements coming from different MTI platforms. To date, very few computational methods for single-cell integration and analysis are capable of operating at this scale [5]. As we move into the megascale and beyond, some of the foremost computational challenges to single-cell analysis are (1) normalization to enable batch compilation of measurements [5]; (2) robust definition of cell phenotypes based on their feature-level representations [64, 124, 65], e.g. marker expression or morphology; and (3) the development of insightful spatial features to characterize the tumor microenvironment, and so enable discrimination between tissues

that vary over important clinical parameters [62].

1.2 Computational cancer systems biology

1.2.1 Brief summary

Parallel to the development of CyCIF and other multiplex platforms, recent advances in machine learning (ML) have made it possible to automatically extract valuable yet human-imperceptible information from images acquired by light and fluorescence microscopy [72]. In particular, deep learning (DL) algorithms—a class of ML algorithms which use multi-layer artificial neural networks to learn abstract feature representations of data—hold the potential to significantly improve the ability of humans to identify and characterize cancer [46].

A subclass of DL algorithms called generative adversarial networks (GANs) has gained considerable traction in medical imaging fields owing to their exceptional capacity to learn to generate realistic data [129], e.g. generating new image instances from noise or some other prior distribution [39], or through conditioning on an input image [48]. GAN architectures are as varied as their applications. When applied to digitized histology slides, GANs have been used for stain normalization [132], semantic segmentation [121], and various supervised and unsupervised image-to-image translation tasks [106, 33, 116].

Aside from computational applications which operate on images directly, parallel strides have been made in the development of ML-based methods for single-cell analysis using "hand-crafted" features which are derived from images [13, 64], e.g. single-cell marker intensity and morphology. On the basis of these feature-level representations, many ML algorithms have been developed to quantitatively define cell phenotypes in lieu of tedious and subjective manual gating methods [124, 65, 64]. Using either prior knowledge or internal data structure, these ML algorithms group similar cells into clusters and facilitate all manner of downstream analyses.

Once cell phenotypes have been defined in high-dimensional feature space, the distributions of phenotypes in tissues in real space can be quantified. Many of the same ML-based clustering methods referenced above can also be used to define local spatial neighborhoods of various cell phenotype combinations, some of which can be associated with clinical parameters [37, 50]. By representing whole tissues as spatial graphs of interconnected cells, graph-based DL models are able to learn both local and global tissue features which improve patient stratification and outcome prediction when integrated with image- and genomic-based DL models [20, 32]. Computational approaches like these which enable multiscale and multimodal integration of cellular features—from molecule to tissue morphology and architecture to clinical outcome—will undoubtedly help to fulfill the promise of precision medicine.

1.2.2 Challenges to computational applications in MTI

Computational approaches to image analysis may be able to help bridge the technological gap highlighted in subsection 1.1.2, but they are not without their own challenges. For instance, DL models typically require large quantities of annotated H&E images to satisfactorily learn tasks. Moreover, training labels defining tumor boundaries or other pathological features must often be generated by domain-expert pathologists through a tedious manual annotation process, which keeps them from their other responsibilities. Even when annotations can be acquired, it remains unclear which specimens should be chosen to best train models which generalize to unseen specimens at deployment time.

As the number of MTI platforms and 2D or 3D—or rather N-dimensional—tissue atlases increases, so too must the capacity of our models to ingest and derive insight from these data. Many of the models used in MTI analysis are cross-overs from other domains. Data integration methods inherited from the bulk sequencing domain assume identical

cellular composition between specimens or batches, and are therefore unfit for discovery-based single-cell studies [5]. Many cell phenotyping algorithms are inherited from single-cell RNA sequencing or flow cytometry domains, where datasets contain measurements either of relatively few cells or without spatial context preserved. Among these cell phenotyping approaches, some make strong assumptions about feature distributions that may not be met or are biased toward known cell types, while others which leverage internal data structure are unstable or prohibitively inefficient at scale and can misidentify rare or unexpected phenotypes [65]. As such, these approaches are unlikely to scale well to atlas-level MTI data and constitute a crucial bottleneck in our search for the next generation of biomarkers in cancer systems biology.

1.3 Dissertation contributions

We attempt to address the challenges above in this dissertation. In chapter 2, we introduce a GAN-based virtual staining paradigm which enables the prediction of biomarker distribution in digitized H&E slides without manual annotation. We demonstrate that virtual staining models can generalize both within and between patient samples, and provide a quantitative framework for selecting specimen cohorts for CyCIF characterization using unsupervised image features from digitized H&E slides.

In chapter 3, we extend the virtual staining paradigm into the third dimension. Using a single, serially-sectioned tumor specimen, we demonstrate that virtual staining models are capable of learning enough information from a single pair of sections stained with either H&E or CyCIF to reconstruct virtual CyCIF images for the entire specimen. Further, we introduce a quantitative region-of-interest (ROI) selection scheme that is enabled through integration of H&E and CyCIF image representations.

In chapter 4, we introduce a graphics processing unit (GPU)-accelerated workflow for normalization, phenotyping, and spatial analysis of single-cell MTI data. By transferring

intensive computations from CPU to GPU, we realize an improvement in analysis efficiency by several orders of magnitude, without using data subsampling strategies which can miss rare cell phenotypes. We deploy this workflow on a BC tissue microarray (TMA) to derive a BC cell type dictionary, which we validate against a published BC cell type dictionary derived using a different MTI platform. Finally, we illustrate spatial features of tissue structure which could be used to distinguish between BC subtypes. We present a brief summary for each of these contributions in the following subsections.

1.3.1 Virtual staining and optimizing histological sample selection

Pathologists rely on the morphological contrast and molecular specificity provided by H&E and immunostains, respectively, when examining tissue slides for indicators of cancer. As a proof of concept, we used a limited but heterogeneous set of human pancreatic cancer samples to demonstrate that a GAN framework we call speedy histological-to-immunofluorescent translation (SHIFT) is able to learn the relationship between H&E and IF images of the same tissue, enabling the near-real-time generation of virtual IF images that are highly similar to the corresponding real IF images. This suggests that information obtained by IF is encoded by features in histological images, and deep learning provides the means to extract this information where such a relationship exists. Importantly, we demonstrate that SHIFT can generalize both within [115] and between [116] samples acquired from multiple patients, even in a data-limited setting. Moreover, our validation of SHIFT is undertaken using complex, associated human tissues, which is in contrast to other virtual IF staining methods which were validated using relatively homogeneous rat or human cell lines or cultures [22, 81].

DL approaches require substantial training data to be robust and generalizable, so we begin by exploring the possibility that the required training samples can be reduced by selecting representative samples. We describe the use of a data-driven method to select samples that optimizes the morphological heterogeneity of the dataset and promotes SHIFT

generalizability. As a proof of concept, we objectively measure the ability of SHIFT to infer the spatial distribution of pan-cytokeratin (panCK), a common cancer biomarker, and show preliminary results from SHIFT inference of α-smooth muscle actin (α-SMA), a common stromal marker. We benchmark SHIFT against Label-Free Determination (LFD) [81], a state-of-the-art supervised DL-based virtual staining method, and demonstrate that either SHIFT or the ensemble of SHIFT and LFD generate more realistic virtual IF images of panCK than LFD alone. This result is consistent with the growing opinion among DL practitioners that adversarial learning methods, like that used in SHIFT, will be required to overcome problems associated with strictly-supervised learning methods [22].

The contents of chapter 2 are adapted from the publications listed below in chronological order:

- **Erik A. Burlingame**, Mary McDonnell, Geoffrey F. Schau, Guillaume Thibault, Christian Lanciault, Terry Morgan, Brett E. Johnson, Christopher Corless, Joe W. Gray, and Young Hwan Chang. "SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning". In: *Scientific Reports* 10.11 (2020), p. 17507. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74500-3

- Young Hwan Chang, **Erik A. Burlingame**, Geoffrey Schau, and Joe W. Gray. *Translation of images of stained biological material*. 2020. URL: https://patents.google.com/patent/WO2020142461A1/en?q=erik+burlingame&inventor=Erik+BURLINGAME

- **Erik A. Burlingame**, Adam A. Margolin, Joe W. Gray, and Young Hwan Chang. "SHIFT: speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks". In: *Proceedings of SPIE–the International Society for Optical Engineering* 10581 (2018). ISSN: 0277-786X. DOI: 10.1117/12.2293249. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6166432/

1.3.2 3D virtual staining and image mode integration for objectively-guided ROI selection

Tumors are not 2D, but many of the imaging characterization platforms in both research and clinical practice make the assumption that tissue microarrays (TMAs) containing small core samples of essentially 2D tissue sections are a reasonable approximation of bulk tumor. However, a recently published 3D tumor atlas charted using H&E- and CyCIF-stained serial sections of colorectal cancer (CRC) specimens strongly challenges this assumption [61]. In spite of the additional insight gathered by measuring the tumor microenvironment in 3D, it can be prohibitively expensive and time consuming to process tens or hundreds of tissue sections with CyCIF. Even when resources or time are not limiting, the criteria for ROI selection in tissues for downstream analysis remain largely qualitative and subjective.

In chapter 3, we extend the virtual staining paradigm to the 3D CRC atlas [61] and demonstrate that GANs can learn from a minimal subset of the atlas to reconstruct the remaining sections of the CyCIF portion of the atlas and recapitulate quantitative endpoints derived using the real CyCIF data. We also implement and evaluate a novel GAN architecture which integrates paired H&E and CyCIF data into a shared representation and demonstrate that the model can be used as a quantitative and objective guide for ROI selection, with the integrated H&E/CyCIF representations being more informative than H&E representations alone.

The contents of chapter 3 are adapted from a manuscript in preparation for submission to *Cell Systems* as a Report:

- **Erik A. Burlingame***, Luke Ternes*, Jia-Ren Lin, and et al. "Histology-based multiplexed 3D reconstruction and channel embedding for optimized region-of-interest selection," manuscript in preparation (2021), *equal contributors sorted alphabetically.

1.3.3 Megascale single-cell phenotyping and spatial analysis of human breast cancer across multiplex tissue imaging platforms

In chapter 4, we analyzed 180 tissues spanning BC subtypes using CyCIF and a marker panel targeting tumor, immune, and stromal cell types. The key contributions of this work are (1) an expanded application and validation of RESTORE [17], our recently published normalization method, which enables compilation and batch processing of such data, (2) a distributed and graphics processing unit (GPU)-accelerated implementation of PhenoGraph, the popular, graph-based algorithm for subpopulation detection in high-dimensional single-cell data, and (3) an integrative analysis using this toolkit which identifies spatial features which discriminate between some of the canonical BC subtypes.

For RESTORE normalization of each core, we leverage the fact that tumor, immune, and stromal cells exhibit mutually exclusive expression of cell type-specific markers, and use a graph-based clustering to define positive and negative cells and normalization factors. Following normalization, shared cell types between TMAs are co-clustered, an indication that the normalization was successful. To define cell types among the $\sim$1.3 million cells in the feature table, we used the CPU-based version of the widely-used algorithm PhenoGraph [58], but found it to be inefficient at this data scale. To break this computational bottleneck, we re-implemented PhenoGraph to be compatible with GPUs and observed multiple orders of magnitude improvement in speed, without sacrificing clustering quality. Our implementation identified diverse tumor, immune, and stromal cell types across tissues and subtypes. We validated our identified cell types through comparison with a recently published survey of BC characterized by IMC, and found highly-correlated clusters for stromal, immune, basal, and proliferating cell types, suggesting that shared cell types could be matched across cohorts and imaging platforms, a necessary step for data integration. We next considered the tumor differentiation states of BC subtypes through their CK expression, and find that while CK+ cells in HER2+, ER+, and HER2+/ER+ tissues are

primarily positive for CKs 19, 7, and 8, triple negative (TN) tissues exhibit a broad heterogeneity of differentiation state, consistent with the genetic and histological heterogeneity of TNBC described in other studies. Finally, we considered the spatial architectures of BC subtypes by building cell type neighborhood graphs, which either quantify the strength and direction of inter-cell type interactions, or the centrality of intra-cell type distributions. When we look across tissues, we observe that ER+ tissues exhibit significantly high higher tumor centrality than other BC subtypes, and forthcoming work will involve validation of this finding in a cohort with more extensive clinical annotation to assess its significance.

The contents of chapter 4 are adapted from the publications listed below in chronological order:

- Young Hwan Chang, Koei Chin, Guillaume Thibault, Jennifer Eng, **Erik A. Burlingame**, and Joe W. Gray. "RESTORE: Robust intEnSiTy nORmalization mEthod for multiplexed imaging". In: *Communications Biology* 3.11 (2020), 1–9. ISSN: 2399-3642. DOI: 10.1038/s42003-020-0828-1

- **Erik A. Burlingame**, Jennifer Eng, Guillaume Thibault, Koei Chin, Joe W. Gray, and Young Hwan Chang. "Toward reproducible, scalable, and robust data analysis across multiplex tissue imaging platforms". In: *Cell Reports Methods* 0.0 (2021). ISSN: 2667-2375. DOI: 10.1016/j.crmeth.2021.100053. URL: https://www.cell.com/cell-reports-methods/abstract/S2667-2375(21)00101-6

1.4 Other contributions

Other contributions, manuscripts, and publications completed during my doctoral studies have been omitted to maintain a clear focus in this dissertation. These works are listed below in chronological order:

- **Erik A. Burlingame**, Jennifer Eng, Guillaume Thibault, Geoffrey F. Schau, Koei Chin, Joe W. Gray and Young Hwan Chang, "Balanced learning of cell state representations," poster presentation at the Learning Meaningful Representations of Life workshop at the Conference on Neural Information Processing Systems (2019).

- Geoffrey F. Schau, **Erik A. Burlingame**, Guillaume Thibault, Tauangtham Anekpuritanang, Ying Wang, Joe W. Gray, Christopher Corless, and Young Hwan Chang. "Predicting primary site of secondary liver cancer with a neural estimator of metastatic origin". In: *Journal of Medical Imaging* 7.1 (2020), p. 012706. ISSN: 2329-4302, 2329-4310. DOI: 10.1117/1.JMI.7.1.012706

- Orit Rozenblatt-Rosen, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E. Rood, Orr Ashenberg, Ethan Cerami, Robert J. Coffey, Emek Demir, and et al. "The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution". In: *Cell* 181.2 (2020), 236–249. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.03.053

- Geoffrey Schau, **Erik A. Burlingame**, and Young Hwan Chang. "DISSECT: DISentangle SharablE ConTent for Multimodal Integration and Crosswise-mapping". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. 2020, 5092–5097. DOI: 10.1109/CDC42340.2020.9304354

- Brett E. Johnson, Allison L. Creason, Jayne M. Stommel, Jamie Keck, Swapnil Parmar, Courtney B. Betts, Aurora Blucher, Christopher Boniface, Elmar Bucher, **Erik A. Burlingame**, and et al. "An Integrated Clinical, Omic, and Image Atlas of an Evolving Metastatic Breast Cancer". In: *bioRxiv* (2020), p. 2020.12.03.408500. DOI: 10.1101/2020.12.03.408500

- Matthew S. Dietz, Thomas L. Sutton, Brett S. Walker, Charles E. Gast, Luai Zarour, Sidharth K. Sengupta, John R. Swain, Jennifer Eng, Michael Parappilly, Kristen Limbach, and et al. "Relevance of Circulating Hybrid Cells as a Non-Invasive Biomarker for Myriad Solid Tumors". In: *bioRxiv* (2021), p. 2021.03.11.434896. DOI: 10.1101/2021.03.11.434896

- Geoffrey F. Schau, Hassan Ghani, **Erik A. Burlingame**, Guillaume Thibault, Joe W. Gray, Christopher Corless, and Young Hwan Chang. "Transfer Learning for Inference of Metastatic Origin from Whole Slide Histology". In: *bioRxiv* (2021). DOI: 10.1101/2021.04.21.440864

- Denis Schapiro, Clarence Yapp, Artem Sokolov, Sandro Santagata, and others including **Erik A. Burlingame**, "MITI Minimum Information guidelines for highly multiplexed tissue images," manuscript under review at *Nature Methods* (2021).

- Avathamsa Athirasala, Paula P. Menezes, Anthony Tahayeri, **Erik A. Burlingame**, Anushka Naiknaware, Ashley Sercia, Christina Hipfinger, Young Hwan Chang, and Luiz E. Bertassoni, "Screening 3D microenvironments reveals the interplay of microgeometry and matrix mechanics in regulation of stem cell differentiation," manuscript in preparation for submission (2021).

- CSBC/PS-ON Image Analysis Working Group, Juan Carlos Vizcarra, **Erik A. Burlingame**, Yury Goltsev, Brian S. White*, Darren Tyson*, Artem Sokolov*, "A community-based approach to image analysis of cells, tissues and tumors," manuscript under review at *Computerized Medical Imaging and Graphics* (2021), *equal contributors.

Chapter 2

Virtual staining foundations

The minute there's a map, there is no art.
Paint by numbers is not art.
Paint by numbers is a mechanical activity.

Seth Godin

2.1 Abstract

Spatially-resolved molecular profiling by immunostaining tissue sections is a key feature in cancer diagnosis, subtyping, and treatment, where it complements routine histopathological evaluation by clarifying tumor phenotypes. In this work, we present a deep learning method called speedy histological-to-immunofluorescent translation (SHIFT, see Figure 2.1) which takes histologic images of hematoxylin and eosin (H&E)-stained tissue as input, then in near-real time returns inferred virtual immunofluorescence (IF) images that estimate the underlying distribution of the tumor cell marker pan-cytokeratin (panCK). To build a dataset suitable for learning this task, we developed a serial staining protocol which allows IF and H&E images from the same tissue to be spatially registered. We show that deep learning-extracted morphological feature representations of histological images can guide representative sample selection, which improved SHIFT generalizability in a

small but heterogenous set of human pancreatic cancer samples. With validation in larger cohorts, SHIFT could serve as an efficient preliminary, auxiliary, or substitute for panCK IF by delivering virtual panCK IF images for a fraction of the cost and in a fraction of the time required by traditional IF.

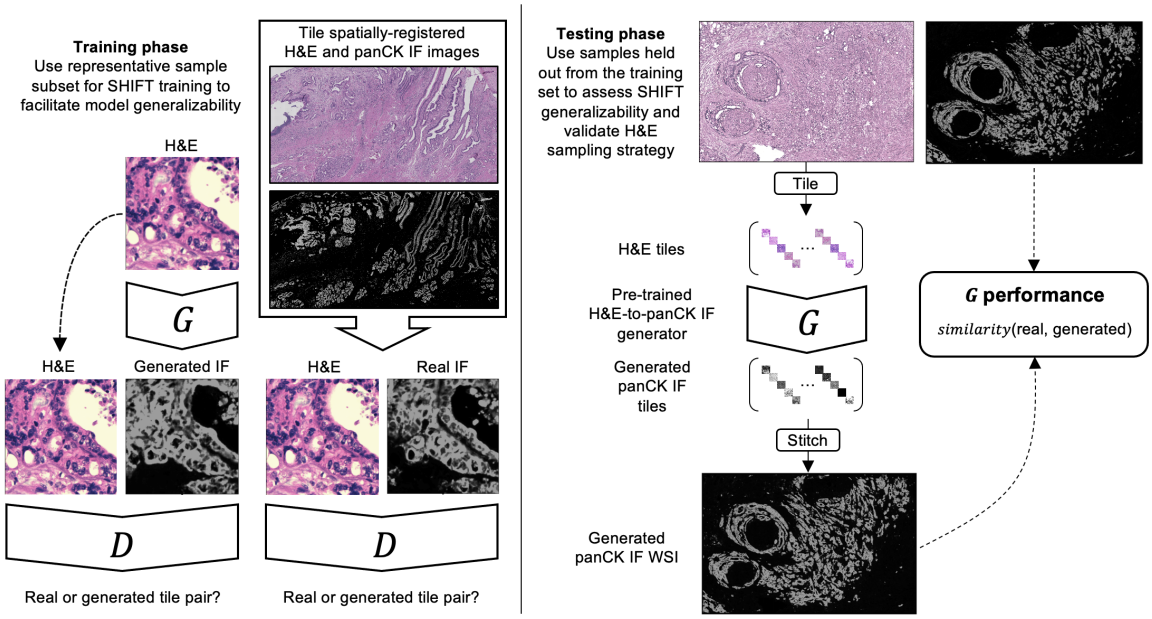


FIGURE 2.1: Schematic of SHIFT modeling for training and testing phases of a model which predicts the distribution of the tumor marker panCK conditioned on an H&E image. The generator network G generates virtual IF tiles conditioned on H&E tiles. The discriminator network D learns to discriminate between real and generated image pairs.

2.2 Introduction

Physicians depend on histopathology—the visualization and pathological interpretation of tissue biopsies—to diagnose cancer. Hematoxylin and eosin (H&E)-stained histologic sections (∼3–5 µm-thick formalin-fixed paraffin-embedded tissue biopsies) are the standard of care routinely employed by pathologists to make diagnoses. On the basis of the visual attributes that make H&E-stained sections useful to pathologists, a broad field of

histopathological image analysis has flourished [40]. Image features derived from H&E-stained sections have been used for tasks ranging from the segmentation of glands in the prostate [77], grading of breast cancer pathology [78], to automated classification of early pancreatic cancer [57].

In spite of the rich information highlighted by the non-specific H&E stain, in challenging cases with indeterminate histology or tumor differentiation, antibody labeling of tumor cells by a molecular imaging technique like immunofluorescence (IF) provides further characterization. It is becoming increasingly apparent that determining the spatially-resolved molecular profile of a cancer is important for disease subtyping and choosing a patient's course of treatment [29]. Despite its clinical value, IF is time- and resource-intensive and requires expensive reagents and hardware, so assessment is typically limited to a small representative section of a tumor, which may not be fully representative of the neoplasm, which can be the case in areas of squamous differentiation in an adenocarcinoma [44]. Also, the cost associated with IF may in some cases limit its use to within highly-developed clinical laboratories, further widening the quality-of-care gap between high- and low-income communities. The gaps between H&E and IF technologies highlight the broader need for automated tools that leverage information attained by a low-cost technology to infer information typically attained by a high-cost technology.

Recent advances in digital pathology and deep learning (DL) have made it possible to automatically extract valuable, human-imperceptible information from H&E-stained histology images [66, 19, 92, 93, 14]. In [49], H&E images and spatially-registered SOX10 immunohistochemistry (IHC) images from the same tissue section were used to train a DL model to infer SOX10 nuclear staining from H&E images alone. Apart from histology and IHC images, other studies have described supervised DL-based methods for inferring fluorescence images from transmitted light images of unlabeled human or rat cell lines or cell cultures [22, 81], but not in complex, associated human tissues. Their methods were also based on supervised pixel-wise learning frameworks, which are known to produce

incoherent or discontinuous patterns in the virtual stains for some markers, though some of the authors suggest that an adversarial learning framework could address the problem [22].

Here were introduce a conditional generative adversarial network (cGAN)-based method called speedy histological-to-IF translation (SHIFT) and begin by demonstrating its ability to infer IF images from images of adjacent H&E-stained tissue from a single patient with pancreatic ductal adenocarcinoma (PDAC) [115]. To better define the virtual staining problem in subsequent studies, we developed a serial staining protocol which enables co-registration of H&E and IF data in the same tissue section [116]. In this setting, we go on to test the generalizability of virtual IF staining by SHIFT through model evaluation on PDAC samples from four patients which were selected by an expert pathologist on the basis of their high inter-sample morphological heterogeneity [116].

DL models require a large amount of heterogeneous training data to generalize well across the population from which the training data was drawn. Since data limitations are common to many biomedical data domains [14, 113, 24, 127], we begin the multi-patient study by exploring the possibility that the choice of training samples can be optimized by selecting the few samples that are most representative of the population of samples at our disposal. Some DL-based applications have been proposed for histological image comparison and retrieval [80, 43], but to the best of our knowledge none have been proposed for the express purpose of image feature-guided training set selection in a data-limited biomedical imaging domain. We describe the use of a data-driven variational autoencoder (VAE)-based method [56] to select samples that optimizes the morphological heterogeneity of the dataset and promotes SHIFT model generalizability.

As a proof of concept for virtual staining in the multi-patient setting, we objectively measure the ability of SHIFT models to infer the spatial distribution of a pan-cytokeratin (panCK) antibody which labels tumor cells, and provide benchmarking comparisons with Label-Free Determination (LFD) [81], a state-of-the-art DL-based virtual staining method.

We also show preliminary results for inference on the stromal marker α-smooth muscle actin (α-SMA). By leveraging a morphological signature of a molecular tumor phenotype and proposing feature-guided sample selection for model generalizability, our approach is a small step toward the development of a generalized platform for multiplexed virtual IF imaging of markers in human tissues for which there exists an association between tissue morphology and an underlying molecular phenotype.

2.3 Methods

2.3.1 Human tissue samples for multi-patient study

Four cases of moderately differentiated pancreatic ductal adenocarcinoma (PDAC) were retrieved from the Oregon Health & Science University (OHSU) Surgical Pathology Department under the Oregon Pancreas Tissue Registry (IRB00003609). Informed written consent was obtained from all subjects. All experimental protocols were approved by the OHSU Institutional Review Board. All methods were carried out in accordance with relevant guidelines and regulations. Sample A was from a male aged 83 at diagnosis; sample B was from a female aged 74 at diagnosis; sample C was from a female aged 57 at diagnosis; and sample D was from a female aged 73 at diagnosis. H&E-stained sections were secondarily reviewed by two board-certified surgical pathologists tasked to identify and classify areas of tumor heterogeneity in representative sections from each case. Discrepancies between pathologists were ameliorated by consensus review. Samples were chosen via pathological review as exemplifying a spectrum of both histological differentiation and heterogeneity.

2.3.2 Pathological evaluation of human tissue samples

Gold standard review of histologic sections by pathologists tasked with identifying heterogeneous differences in PDAC tumor morphology and grade revealed interobserver agreement in the identification of areas of squamous differentiation in one case and various tumor grades within neoplasms in the other three cases. All four cases were predominantly grade 2 adenocarcinoma and there was no disagreement evaluating marked regions of interest. The case with areas of squamous differentiation did not clearly meet the 30% threshold for adenosquamous classification. The other three cases were predominantly grade 2 with foci of grade 1 and others with grade 3.

2.3.3 Preparation of tissue for immunofluorescence staining

Formalin-fixed paraffin-embedded tissue blocks were serially sectioned by the OHSU Histopathology Shared Resource. From each block, three sections were cut in order to generate a standard H&E for pathological review and downstream analysis, a second serial section of tissue for immunofluorescence staining/post-immunofluorescence H&E staining, and a third section for secondary only control. After sectioning, the second serial tissue section was immediately baked at 55 °C for 12 h and subjected to standard deparaffinization; the slides underwent standard antigen retrieval processing, washing, and blocking. Upon completion, primary antibodies were diluted and applied.

2.3.4 Application of antibodies

Alpha-Smooth Muscle Actin (Mouse monoclonal antibody, IgG2a, Clone: 1A4; Pierce/Invitrogen, cat#MA5-11547) was diluted to 1:200 with Ki-67 (D3B5), (Rabbit monoclonal antibody, IgG, Alexa Fluor 647 Conjugate; Cell Signaling Technology, cat#12075S) diluted to 1:400, along with Pan Cytokeratin (AE1/AE3) (Mouse monoclonal antibody, IgG1, Alexa Fluor 488 Conjugate; ThermoFisher, cat#53-9003-82), which was diluted to 1:200 in 10%

Normal Goat Serum in 1% Bovine Serum Albumin in Phosphate Buffered Saline. Primary antibodies were diluted and incubated overnight at 4 °C. After incubation, secondary antibody (Goat anti-mouse monoclonal antibody, IgG2A, Alexa Fluor 555 Conjugate; Life Technologies, cat#A21137), at 1:200 dilution was applied to the slides and incubated at room temperature for one hour. After incubation slides were washed and mounted with Slowfade Gold Antifade Mountant with DAPI (Fisher Scientific, cat#S36936) in preparation for image acquisition.

2.3.5 Post-IF H&E staining of tissue samples

After the IF stained slides were scanned and the immunofluorescence staining verified, the glass coverslips were removed and the slides were processed for post-IF H&E staining. Post-IF H&E staining was performed with the Leica Autostainer XL staining system at the OHSU Histopathology Shared Resource with the staining protocol in Table 2.1.

Solution	**Time**
Hematoxylin	10 min
Wash in water	1 min
Acid alcohol (0.5% HCl in 70% Ethanol)	8 s
Wash in water	25 s
Bluing solution	2 min
Wash in water	20 s
80% Ethanol/water	25 s
Eosin	10 s
80% Ethanol/water	25 s
95% Ethanol/water	20 s
100% Ethanol (two times)	25 s
Xylene (five times)	25 s

TABLE 2.1: Sequential IF and H&E staining protocol for FFPE tissues

2.3.6 Image acquisition and presentation

Slides were scanned with the Zeiss Axio Scan.Z1 slide scanner of the OHSU Advanced Multiscale Microscopy Shared Resource with the 20X objective in both brightfield and immunofluorescence scanning. Carl Zeiss Images (CZI) were acquired using Zeiss Zen software. CZI images from the Zeiss Axioscan Slide Scanner were processed with the Zeiss

Blue Zen Lite microscope software package. All brightfield and immunofluorescence images were exported as TIFF files for downstream image processing.

2.3.7 Image pre-processing

Raw H&E and IF whole slide images (WSIs) must be pre-processed to remove technical noise, account for between-sample intensity variation, and align paired H&E and IF WSIs in a shared coordinate system. To do so, we use the following pipeline:

1. Quality control: formalin-fixed pancreatic tissue is prone to high levels of autofluorescence, which can mask specific IF signal. Regions of WSIs which exhibited low IF signal-to-noise due to autofluorescence as determined by pathologist review were excluded from our analysis. Divisions of samples were based on the geometries of the image regions determined unaffected by autofluorescence. Some acceptable regions were relatively small due to surrounding regions of autofluorescence.

2. Downscaling: 20X WSIs are downscaled by a factor of 2 in x and y dimensions to generate 10X WSIs. We experimented with using either 20X or 10X images and found that models performed best when using 10X images.

3. Registration: H&E and IF WSIs are spatially registered using an affine transformation that is estimated using matched SURF features [16] extracted from hematoxylin and DAPI binary masks of nuclei generated by Otsu's thresholding method, respectively. Concretely, registration of an H&E WSI and a corresponding IF WSI of the same tissue was achieved using MATLAB [73] through the following steps:

   (a) Conversion of H&E images from RGB colorspace to grayscale using the MATLAB function rgb2gray.

(b) Binarization and complementation of the grayscale H&E and DAPI WSIs using the MATLAB functions imbinarize and imcomplement, creating nuclei masks from each of the H&E and DAPI WSIs.

(c) Detection and extraction of SURF features from each of the H&E and DAPI nuclei masks using the MATLAB functions detectSURFFeatures, selectStrongest, and extractFeatures. We constrained the number of features selected to $min(10,000, \text{number of features detected})$ to reduce the computational cost of feature matching in the next step.

(d) Feature matching between the features extracted from the H&E and DAPI nuclei masks using the MATLAB function matchFeatures.

(e) Estimation and application of the affine transformation matrix which correctly registers H&E and DAPI nuclei masks using the MATLAB functions estimateGeometricTransform and imwarp. The same transformation which correctly registers DAPI to the H&E WSI is used to register IF WSIs.

4. Technical noise reduction: IF WSIs are median filtered with a 5-pixel-radius disk structuring element.

5. Intensity normalization: H&E WSI pixel intensities are normalized as previously described [71]. Following [22], IF WSI pixel intensities are normalized to have a fixed mean= 0.25 and standard deviation= 0.125, then clipped to fall within (0,1).

6. Image tiling: WSIs are tiled into non-overlapping 256×256 pixel tiles, such that each H&E tile has a corresponding spatially-registered IF tile. H&E tiles that contained more than 50% background pixels were removed along with the corresponding IF tiles. Background pixels were defined as those with 8-bit RGB intensities all greater than 180. Each 10X WSI is comprised of hundreds or thousands of non-overlapping 256×256 pixel tiles.

2.3.8 SHIFT Model Architectures

Conditional Generative Adversarial Networks (cGANs)

Image-to-image translation—the mapping of pixels from one scene representation to pixels of another representation of the same scene—is a fundamental image processing problem. The cGAN [39, 76] is a compelling DL-based solution to the image-to-image translation problem and has been deployed for many tasks, including detection of skin lesions [113], retinal image synthesis [24], super-resolution fluorescence image reconstruction [82], and virtual H&E staining [93]. To approach the problem of translating H&E images to their IF counterparts, SHIFT adopts the cGAN-driven architecture pix2pix [48], which benefits from its bipartite formulation of generator and discriminator. Like other methods proposed for image-to-image translation, cGANs learn a functional mapping from input images x to ground truth target images y, but, unique to a cGAN architecture, it is the task of a generator network G to generate images $\hat{y}$ conditioned on x, i.e. $G(x) = \hat{y}$, that fool an adversarial discriminator network D, which is in turn trained to tell the difference between real and generated images (Figure 2.1). What ensues from this two-network duel is a G that generates realistic images that are difficult to distinguish from real images, some GAN-generated images being sufficiently realistic to be considered as a proxy for the ground truth when labeled data are scarce or prohibitively expensive. Concretely, the cGAN objective is posed as a binary cross-entropy loss:

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x, y)] + \mathbb{E}_{x\sim p_{data}(x)}[\log(1 - D(x, G(x)))] \tag{2.1}$$

where G seeks to minimize the objective and thus minimize the distinguishability of generated and real images, while D seeks the opposite. In addition to the task of fooling D, G is also encouraged to generate images that resemble real images through incorporation of

an L1 reconstruction loss term:

$$\mathcal{L}_{\text{L1}}(G) = \mathbb{E}_{x,y\sim p_{data(x,y)}}[\|y - G(x)\|_1] \tag{2.2}$$

The full cGAN objective is:

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \mathcal{L}_{\text{L1}}(G) \tag{2.3}$$

where the L1 tuning parameter $\lambda = 100$ is adapted according to the IF stain prevalence in the current batch of IF tiles [115] i.e. if 50% of the pixels in the current batch of IF tiles are positively stained above the mean intensity of the WSI, then $\lambda = 100 \times 0.5 = 50$. Training data consist of spatially registered pairs of H&E image tiles x and IF image tiles y, while the test data consist of H&E and IF image pairs withheld from the training data. Models were trained using the Adam optimizer [55] with a learning rate of 0.002 for 500 epochs. Training batch sizes were set to 64. The first layers of both the generator and discriminator networks were 128 filters deep (see Figure 2.2 for additional architectural details). Models were trained and tested using a single NVIDIA V100 graphics processing unit (GPU). Once trained, models were capable of processing a 10X (0.44 µm/pixel) H&E image tile containing 256×256 pixels into its corresponding virtual IF tile in 10 µs, corresponding to a virtual staining rate of 22 mm^2 tissue per second, or approximately one virtual IF WSI generated per 20 s. Full model details are available at `https://gitlab.com/eburling/shift`.

2.3.9 Model Ensembles

Single Patient Model Ensembles

In the context of machine learning, aggregating several trained models can increase prediction accuracy, especially when the aggregated models capture distinct features of their shared input. Thus, we also combined the output of independently-trained models, i.e.

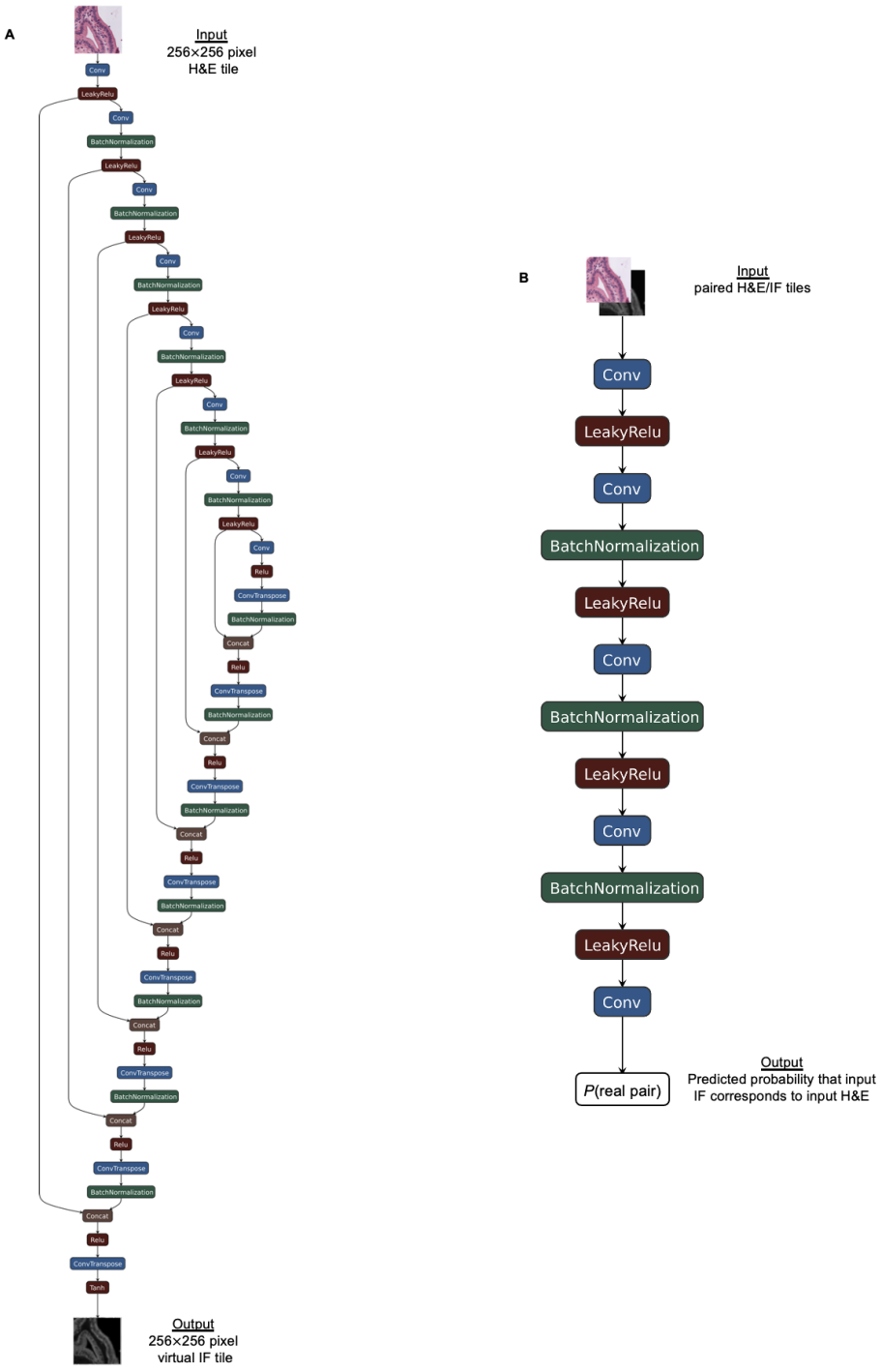


FIGURE 2.2: Schematics of cGAN architecture used by SHIFT. The cGAN architecture used by SHIFT is based on the pix2pix framework [48]. (A) Architecture of generator network G which is based on the U-net architecture [95]. (B) Architecture of discriminator network D. The input to D is a single image of H&E and IF concatenated along the channel axis.

models utilizing Equation 2.3 and Equation 2.6, to form an ensemble distribution, under the assumption that the training strategies put forward in Equation 2.3 and Equation 2.6 are complementary. By doing this, we can smoothen the final output and improve performance by reducing substantial disagreement patterns between models.

Multi-Patient Model Ensembles

In addition to testing the ability of independent SHIFT and LFD models to generate virtual IF images, we also tested model ensembles. Ensemble images were generated by simply averaging the virtual IF image outputs of SHIFT and LFD models trained to generate the same stain using the same training set.

2.3.10 Prevalence-Based Adaptive Regularization

Cancer cells typically remain clustered together (Figure 2.1) and thus it is challenging to balance the reconstruction loss term for positive/negative instances according to the stain prevalence for each training image. For instance, for low-prevalence (sparse) panCK-stained regions, G is more likely to generate an "unstained" pattern rather than generate a sparsely localized stain pattern because the reconstruction loss is relatively small compared to the reconstruction loss for high-prevalence (dense) panCK-stained regions. In order to achieve high sensitivity and specificity, a generative model should be encouraged to be conservative by being maximally penalized when it makes false-positive classifications on low-prevalence ground truth tiles during training. Thus, we propose a prevalence-based adaptive regularization parameter λ' that may be more suitable for the translation of signals from H&E to IF:

$$\lambda' = \lambda \left(\epsilon + \frac{1}{n} \sum_{i=1}^{n} I_{\Omega(p_i)} \right)^{-1} \tag{2.4}$$

where $\epsilon = 0.1$ is chosen to offset in cases where stain prevalence is zero, n is the total number of pixels in the ground truth IF tile, and:

$$I_{\Omega(p_i)} = \begin{cases} 1, & \text{if } p_i \text{ in } \Omega \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

where Ω represents the ground truth mask, and p_i represents the i-th pixel. Our final objective is:

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda' \mathcal{L}_{\text{L1}}(G) \tag{2.6}$$

Utilization of the adaptive regularization parameter λ' maximizes the penalty for generator errors on low-prevalence ground truth tiles and minimizes the penalty for errors on high-prevalence ground truth tiles. By doing this, we can improve localization characteristics and help minimize false classification errors at a distance from true-positive pixels.

2.3.11 Variational Autoencoders

The VAE architecture [56] is designed to elucidate salient features of data in a data-driven and unsupervised manner. A VAE model seeks to train a pair of complementary networks: an encoder network θ that seeks to model an input x_i as a hidden latent representation z_i, and a decoder network ϕ that seeks to reconstitute x_i from its latent representation z_i. The VAE cost function shown below penalizes model training with an additional Kullback–Leibler (KL) divergence term that works to conform the distribution of z with respect to a given prior, which in our case is the standard normal distribution:

$$\mathcal{L}_i(x_i, \theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + \text{KL}(q_\theta(z|x_i)p(z)) \tag{2.7}$$

where $p(z) = \mathcal{N}(0, 1)$.

By specifying a latent dimension z less than the input dimension of x, a VAE model learns a pair of optimal encoding and decoding functions that enable reconstruction of an input sample subject to capacity constraints of the latent feature space within the model. In general, this formulation learns encoding functions that compress the information content in the high-dimensional input into a low-dimensional embedding space that learns dataset features sufficient to reconstitute the original input sample while preserving an expected distribution over the learned features. This interpretation enables a specified selection criteria function designed to sample whole slide images whose constituent tiles maximally cover the entire learned feature space with a minimal number of samples.

2.3.12 Feature-Guided Histological Sample Selection

Although DL approaches like SHIFT and LFD require substantial training data to be robust and generalizable, due to resource constraints we hope that a small number of paired H&E and IF image samples is required for model training. Typically, archival WSIs of H&E-stained tissue sections exist on-hand for each sample, which allows for the screening of samples to identify the minimal number of samples that maximally represent the morphological spectrum of the disease being considered. Recent studies demonstrate that DL systems are well-suited for image retrieval tasks in digital pathology [80, 43], wherein a pathologist submits a query image or region of interest and the DL system returns similar images based on their DL-defined feature representations. We seek to solve the inverse task of heterogeneous training set selection in digital pathology, though our approach could be extended to any data-limited biomedical imaging domain.

Since PDAC is a morphologically heterogeneous disease [44], building a representative training set is crucial to the design of a model that will generalize across heterogeneous biopsy samples after deployment. In order to minimize the required resources for acquiring paired H&E and IF images but still cover a broad spectrum of heterogeneous morphological features in the selected H&E samples, we propose a clustering method to learn a

heterogeneous representation of H&E sample images. To assess the morphological features of each sample, we use a variational autoencoder (VAE) [56] to extract 16-dimensional feature vectors from each H&E tile to establish comparisons between samples. Since texture and morphological features on H&E tiles in each cluster of samples will be comparatively more similar than those of the other cluster, we only select representative H&E samples from each cluster for our training dataset. We also tried using other feature vector sizes for representation learning, e.g. 2, 4, 8, 32, but found that a feature vector size of 16 yielded the lowest reconstruction losses.

For example, if there are four samples being considered for IF staining, but resources limit the number of samples that can be stained to two, a decision must be made about which samples should be selected. For the four samples, we aggregate their archival H&E WSIs, extract features from H&E tiles for each sample using a VAE, and quantitatively determine the samples needed to maximally cover the feature space over which the H&E tile set is distributed. By screening and selecting samples in this data-driven fashion, we exclude homogeneous or redundant samples that would not contribute to model generalizability. This maximizes model performance by ensuring that our training dataset is representative of the disease being modeled, thus minimizing cost through the selection of the fewest samples required to do so. Perhaps more importantly, when we fail to generate reliable virtual IF images for certain tissue samples or IF markers, this framework will be useful to examine whether or not their morphological features are present in the training dataset, which can guide how we select additional samples when updating our dataset.

To identify the sequence of samples that should be selected, we adapt an information-theoretic sample selection algorithm [83] which is more capable of generating representative subsets of data with imbalanced features than other classical algorithms used for sample selection, like maximum coverage [45] or k-medoid clustering [52]. The algorithm is parameterized using the following notation:

TABLE 2.2: Parameters for optimal sampling scheme.

Parameter	Description
X	Complete tile set of all examples, $X = \{x_1, x_2, \ldots, x_n\}$
x_i	Single tile, $x_i \in X$
X_i	Subset of X corresponding to the ith sample, $X_i \subset X$
F	Complete VAE-learned feature set, $F = \{f_1, f_2, \ldots, f_m\}$
f_i	Single feature, $f_i \in F$
A	Random variable defined over F
T	Random variable defined over X

We begin with a tiles $\times$ features table, where we set m=16 for our experiments:

TABLE 2.3: Example of unnormalized VAE feature values.

	f_1	f_2	...	f_m
x_1	-1.64266	1.36952	...	1.23509
x_2	-0.792104	-0.481497	...	1.07938
...	...	...	...	...
x_n	0.00163981	-0.0162441	...	-0.95883

We normalize across rows of the table, such that each tile is now represented as a probability distribution over the feature domain:

TABLE 2.4: Example of normalized VAE feature values.

	f_1	f_2	...	f_m	Sum
x_1	0.00418311	0.0850498	...	0.0743519	1
x_2	0.0148384	0.0202433	...	0.0964193	1
...	...	...	...	...	...
x_n	0.0208721	0.0205021	...	0.00798802	1

We define the random variables T and A over tile domain X and the feature domain F, respectively, such that $P(A = f_1|T = x_1) = 0.418311$, $P(A = f_2|T = x_2) = 0.0202433$, and so on. With this conditional probability table, we can define probability distributions over each subset X_i : $P(A|X_i) = \frac{1}{|X_i|} \sum_{x \in X_i} P(A|x)$. To measure the representativeness

of sample X_i to the full dataset X, we compute the Kullback–Leibler (KL) divergence between $P(A|X_i)$ and $P(A|X)$: $\mathrm{KL}(P(A|X_i)||P(A|X)) = \sum_{f\in F} P(f|X_i)\log\frac{P(f|X_i)}{P(f|X)}$. We then weight this divergence by the proportion of X that X_i comprises, $\frac{|X_i|}{|X|}$, to prioritize subsets that contribute many tiles to X. We define the single most representative sample as $\hat{X}_1 = \min_{X_i\subset X}\left(\frac{|X|}{|X_i|}\mathrm{KL}(P(A|X_i)||P(A|X))\right)$, the most representative duo of samples as $\hat{X}_2 = \hat{X}_1 + \min_{X_i\subset X-\hat{X}_1}\left(\frac{|X|}{|Xi|}\mathrm{KL}(P(A|X_i+\hat{X}_1)||P(A|X))\right)$, the most representative trio of samples as $\hat{X}_3 = \hat{X}_2 + \min_{X_i\subset X-\hat{X}_2}\left(\frac{|X|}{|X_i|}\mathrm{KL}(P(A|X_i+\hat{X}_2)||P(A|X))\right)$, and so on. In this way, we define the sequence of samples that should be chosen to optimally increase the representativeness of the training set.

2.4 Results

2.4.1 Virtual staining in single PDAC patient

Developing the single patient dataset

This study utilizes a dataset [16] containing WSIs of tumorigenic pancreas tissue acquired at 20X-magnification from two adjacent thin sections: one stained with H&E and and the other co-stained with the fluorescent nuclear marker DAPI and fluorescent antibodies against panCK and α-SMA, two markers commonly used in tumor evaluation [7, 107]. The paired 20X images were registered [16] and cropped into four sites, with each site image being ∼12,000×8,000 pixels in size. 10X WSIs were created by half-scaling 20X WSIs. Training data were created by first taking ∼10,000 random 256×256 pixel H&E and IF tile pairs from three sites, then applying single operation manipulations—i.e. jitter, rotation, flipping, Poisson noise—to each tile, yielding ∼20,000 total images in the augmented training data. For a given stain, we trained four leave-one-site-out SHIFT models and generated virtually-stained WSIs for each site, i.e. each of four models were trained on random tiles from three sites and tested on non-overlapping tiles of the left-out site, which

could then be stitched into cohesive WSIs. In this way, we were able to perform a fourfold cross-validation of the SHIFT method for each stain in an intra-patient context. To reduce the deleterious effects of tiling artifacts in the generated panCK WSIs, we utilized three additional test datasets of non-overlapping tiles from each site—one of each test dataset offset by 128 pixels in either x or y or both—and evaluated model performance using the jointly-scaled blend of the four generated WSIs.

Model parameterization

The network architectures and implementations for D and G for all models are as described in [48], except where explicitly specified in Figure 2.2. Training batch size was set to 4 for all experiments and for fair comparison, we tuned the regularization setting for each model by training over a range of λ: 50-5000 and selected the models with optimal λ^* that yielded the best performance. Models were trained for 20 epochs at a fixed learning rate of 0.0002, followed by 10 epochs over which the learning rate linearly decayed to zero. Once trained, each SHIFT model was capable of computing WSI-level translation in approximately one minute.

Model evaluation

For evaluation of SHIFT model performance, we measured the Matthews correlation coefficient (MCC) [74], the Dice similarity coefficient (DSC), as well as other standard classification performance metrics for comparison of the ground truth and generated IF masks produced using a global 10%-luminance threshold on the contrast-adjusted 8-bit ground truth and generated IF WSIs. We also measured the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [122] between raw ground truth and raw generated IF WSIs.

Representative results for the translations from H&E-to-panCK (SHIFT2panCK) for all four sites are shown in Figure 2.3 and translations from H&E-to-DAPI (SHIFT2DAPI),

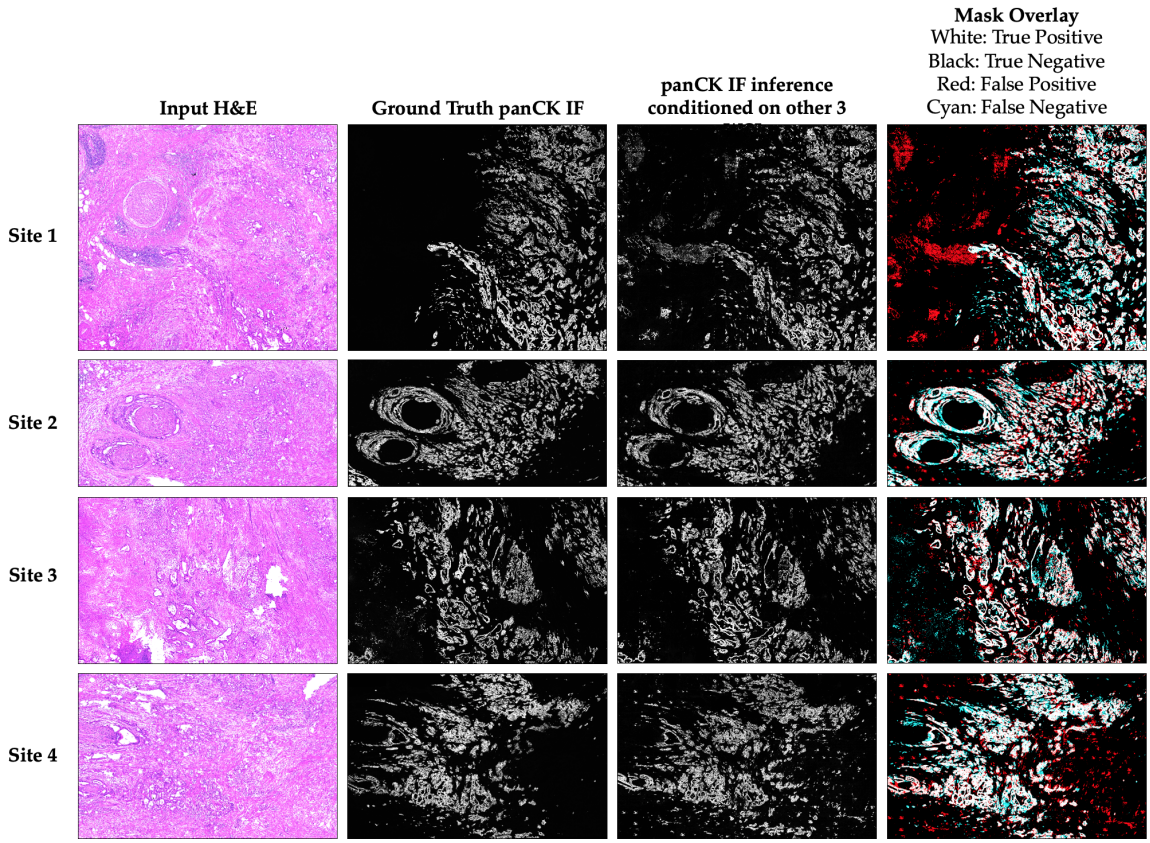


FIGURE 2.3: Single patient SHIFT result for panCK for all four sites.

H&E-to-panCK, and H&E-to-α-SMA (SHIFT2α-SMA) for just site 1 are shown in Figure 2.4. All quantitation of model performance is reported in Table 2.5.

We performed SHIFT2DAPI experiments at both 10X- and 20X-magnification to assess whether or not SHIFT model inference is sensitive to image resolution, and found minor improvements in most metrics when models were trained on 20X tiles (Table 2.5, top), suggesting that localized features of the DAPI stain may be more important for SHIFT2DAPI inference than higher-level architectural features. Since hematoxylin and DAPI are both robust stains for cell nuclei, the task of a SHIFT2DAPI model is theoretically trivial—translate hematoxylin intensity into DAPI intensity—and thus provides insight into the upper limits of SHIFT performance. Note that there exists µm-scale structural differences between

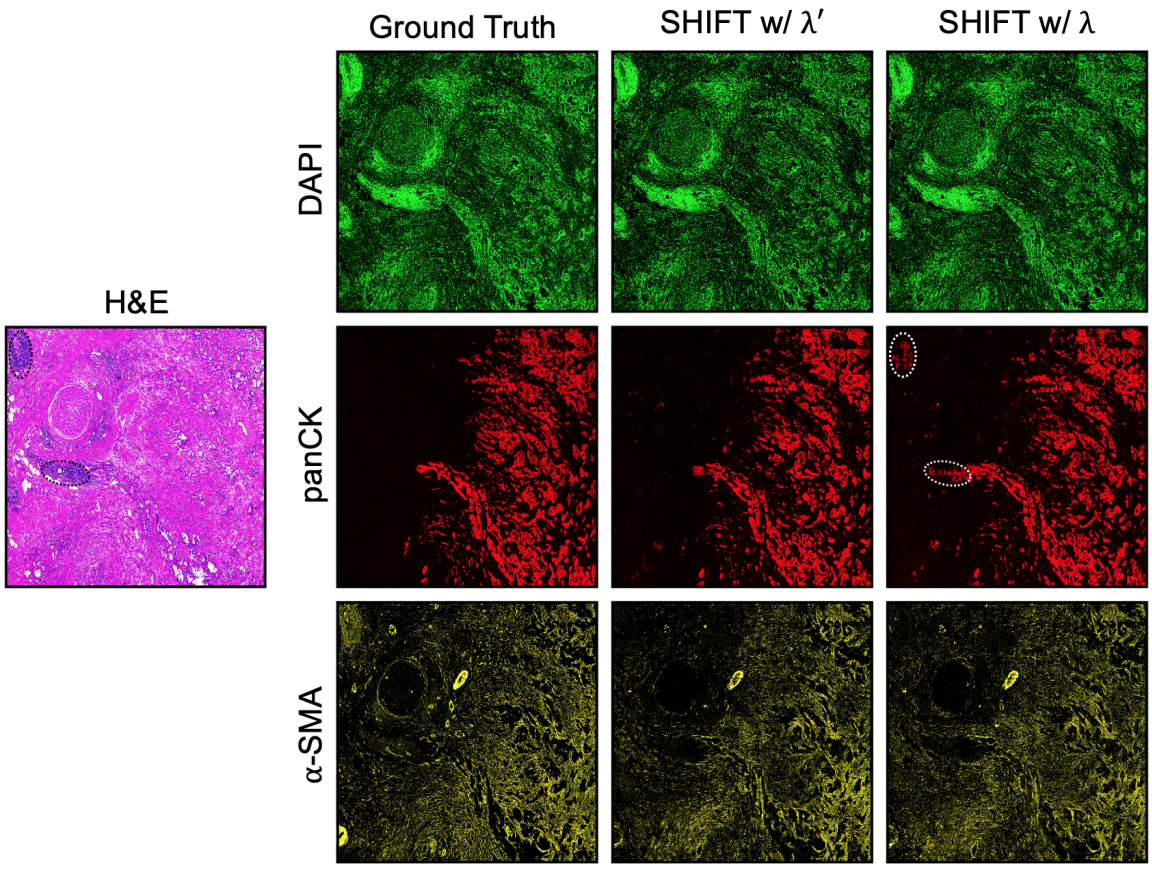


FIGURE 2.4: SHIFT model results for DAPI, panCK, and $\alpha - SMA$ for site 1 (12,656 $\times$ 10,858 pixels at 20X magnification). Each SHIFT image represents the result for the model with optimal λ^* which yielded the best performance (Table 2.5). The circled dark regions in the H&E image are clusters of invading lymphocytes which the SHIFT model with fixed λ (Equation 2.3) misclassified as being panCK-positive (see the corresponding circled regions in the middle-right image). The SHIFT model with adaptive λ' (Equation 2.6) did not commit these errors.

ground truth H&E and IF WSIs due to serial tissue acquisition. Nevertheless, the results for models utilizing Equation 2.6 are consistent with those from a comparison between the DAPI mask and a cell nucleus segmentation mask derived from the H&E image (data not shown), indicating that SHIFT2DAPI achieves good performance up to the fundamental limit.

Given that panCK will stain only the subset of cells which are CK-positive, rather than stain a ubiquitous cytological landmark as do hematoxylin and DAPI, the translation from

Model translation	Mag.	Site generated	G^*	λ^*	MCC	DSC	Accu.	Spec.	Prec.	Sens.	PSNR	SSIM
SHIFT2DAPI	10X	1	Eq (2.3)	5000	0.838	0.885	0.932	0.938	0.857	**0.916**	30.89	0.883
			Eq (2.6)	1000	0.845	0.890	0.936	0.951	0.881	0.898	31.40	0.887
	20X	1	Eq (2.3)	500	0.857	0.897	0.942	0.965	0.910	0.886	**31.53**	0.883
			Eq (2.6)	5000	**0.861**	**0.900**	**0.944**	**0.966**	**0.913**	0.887	31.50	**0.898**
SHIFT2panCK	10X	1	Eq (2.3)	1000	0.704	0.749	0.909	0.918	0.662	0.863	22.99	0.769
			Eq (2.6)	1000	**0.754**	**0.793**	**0.933**	**0.953**	**0.766**	0.822	22.95	**0.791**
			Ensemble	–	0.729	0.769	0.917	0.922	0.679	**0.887**	**23.19**	0.782
		2	Eq (2.3)	1000	0.817	0.855	0.937	0.946	0.812	**0.903**	28.21	0.819
			Eq (2.6)	1000	0.814	0.853	**0.939**	**0.959**	**0.845**	0.861	27.89	0.816
			Ensemble	–	**0.821**	**0.859**	0.938	0.948	0.819	**0.903**	**28.66**	**0.828**
		3	Eq (2.3)	1000	**0.790**	**0.822**	**0.945**	0.965	0.810	0.834	26.36	0.815
			Eq (2.6)	1000	0.777	0.807	**0.945**	**0.978**	**0.860**	0.760	26.16	0.818
			Ensemble	–	**0.790**	**0.822**	0.944	0.958	0.786	**0.862**	**26.69**	**0.828**
		4	Eq (2.3)	1000	0.812	0.849	0.940	0.967	0.865	**0.833**	26.05	0.807
			Eq (2.6)	1000	0.792	0.826	0.936	**0.981**	**0.908**	0.758	25.87	0.810
			Ensemble	–	**0.819**	**0.854**	**0.943**	0.972	0.881	0.828	**26.35**	**0.818**
SHIFT2α-SMA	10X	1	Eq (2.3)	1000	–	–	–	–	–	–	24.70	0.603
			Eq (2.6)	1000	–	–	–	–	–	–	24.84	0.608
			Ensemble	–	–	–	–	–	–	–	**25.09**	**0.611**
		2	Eq (2.3)	1000	–	–	–	–	–	–	25.69	0.634
			Eq (2.6)	1000	–	–	–	–	–	–	25.81	0.642
			Ensemble	–	–	–	–	–	–	–	**26.02**	**0.643**
		3	Eq (2.3)	1000	–	–	–	–	–	–	24.19	0.588
			Eq (2.6)	1000	–	–	–	–	–	–	24.41	0.598
			Ensemble	–	–	–	–	–	–	–	**24.74**	**0.606**
		4	Eq (2.3)	1000	–	–	–	–	–	–	25.21	0.634
			Eq (2.6)	1000	–	–	–	–	–	–	26.34	**0.675**
			Ensemble	–	–	–	–	–	–	–	**26.39**	0.674

TABLE 2.5: SHIFT model parameters and performances. The result for the model with the optimal λ^* that yielded the best performance (MCC for DAPI and panCK, SSIM for α-SMA) is shown for each combination of magnification and G^*. Models were trained until errors stabilized.

H&E to panCK is a more interesting but challenging task. Although SHIFT2panCK models performed less well than SHIFT2DAPI in most categories, it is difficult to visually distinguish the generated from the ground truth panCK IF sites, as shown in Figure 2.4. With one exception (the sensitivity of SHIFT2panCK for site 4), either the models utilizing the proposed method Equation 2.6 alone or the ensemble approach performed as well as or better than models utilizing Equation 2.3 alone, i.e. pix2pix. Notably, models utilizing the proposed method Equation 2.6 showed better localization characteristics (Figure 2.4, circled misclassified regions for model utilizing Equation 2.3).

In contrast to DAPI and panCK stain patterns, the α-SMA stain pattern is sinuous and high-frequency (Figure 2.4, bottom). When these attributes are compounded by spatial

deformity and other complications from the serial acquisition of H&E and IF WSIs, pixel-level evaluation of generated α-SMA WSIs becomes exceedingly challenging. For this reason, we excluded evaluation metrics that were contingent on α-SMA mask generation in favor of metrics which reflect the global configurations of the α-SMA IF WSIs (Table 2.5, bottom). While the ensemble approach performed best in both categories for most sites, all models utilizing the proposed method Equation 2.6 alone outperformed the models utilizing Equation 2.3 alone.

2.4.2 Virtual staining in multiple PDAC patients

Building a dataset of spatially-registered H&E and IF images

SHIFT requires spatially-registered pairs of H&E and IF whole slide images (WSIs) for model training and testing (Figure 2.1). Such data would usually be acquired by processing two adjacent tissue sections, one stained by H&E and another stained by IF, then spatially registering the images into the same coordinate system based on their shared features [16]. Unfortunately, this can lead to inconsistencies between H&E and IF image contents when high-frequency cellular features differ between adjacent sections, even when the sections are as few as 5 µm apart. To alleviate this issue, we developed a protocol that allows for H&E and IF staining in the same section of tissue.

Clinical samples of PDAC from four patients (Samples A, B, C and D) were chosen via pathological review of archival H&E images as exemplifying a spectrum of both histological differentiation and heterogeneity (Figure 2.5A). Chosen samples were sectioned, processed, and stained with DAPI nuclear stain and panCK monoclonal antibody; the staining was confirmed and the slides were scanned. After scanning, the coverslips were removed and the slides were stained with the designed modified H&E protocol (Table 2.1), permanently cover slipped and then scanned again. Nuclear information from the hematoxylin and DAPI stains in pairs of H&E and IF images were used to register images in a common

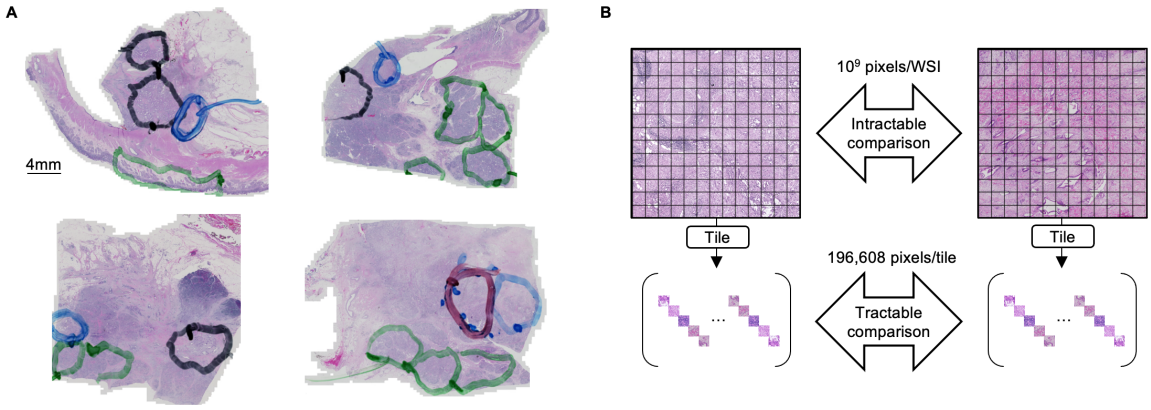


FIGURE 2.5: Overview of PDAC histological samples used for multi-patient SHIFT modeling. (A) Four heterogeneous samples of H&E-stained PDAC biopsy tissue used in the current study. Pathologist annotations indicate regions that are benign (green), grade 1 PDAC (black), grade 2/3 PDAC (blue), and grade 2/3 adenosquamous (red). (B) Making direct comparisons between H&E whole slide images (WSIs) is intractable because each WSI can contain billions of pixels. By decomposing WSIs into sets of non-overlapping 256 $\times$ 256 pixel tiles, we can make tractable comparisons between the feature-wise distribution of tile sets.

coordinate system. Images were then pre-processed to minimize noise and account for technical variability in staining and image acquisition. To exclude regions of autofluorescence that greatly diminished the signal-to-noise ratio of the real IF images, images from samples B and D were subdivided into image subsets B1, B2, B3 and D1, D2, D3, D4, D5.

Feature-guided identification of representative histological samples

For a SHIFT model to generalize well across the population of PDAC samples, it must be trained on a representative subset of the population, which motivated the development of a means to quantitatively compare images. In particular, we wished to learn which sample—or sequence of samples—should be selected to build a training set that is most representative of the population of samples. As a consequence of their large dimensions, direct comparison between gigapixel H&E images is intractable, so we decomposed each image into sets of non-over-lapping 256 $\times$ 256 pixel tiles (Figure 2.5B). Even the small 256 $\times$ 256 pixel H&E tiles contain 196,608 (256 $\times$ 256 $\times$ 3 channels $\times$ 196,608) pixel values each and are difficult to compare directly. To establish a more compact but still expressive

representation of the H&E tiles, we trained a variational autoencoder (VAE) [56]—an unsupervised DL-based method for representation learning and feature extraction—to learn 16-dimensional feature representations of each tile, which makes comparing tiles more tractable (Figure 2.6).

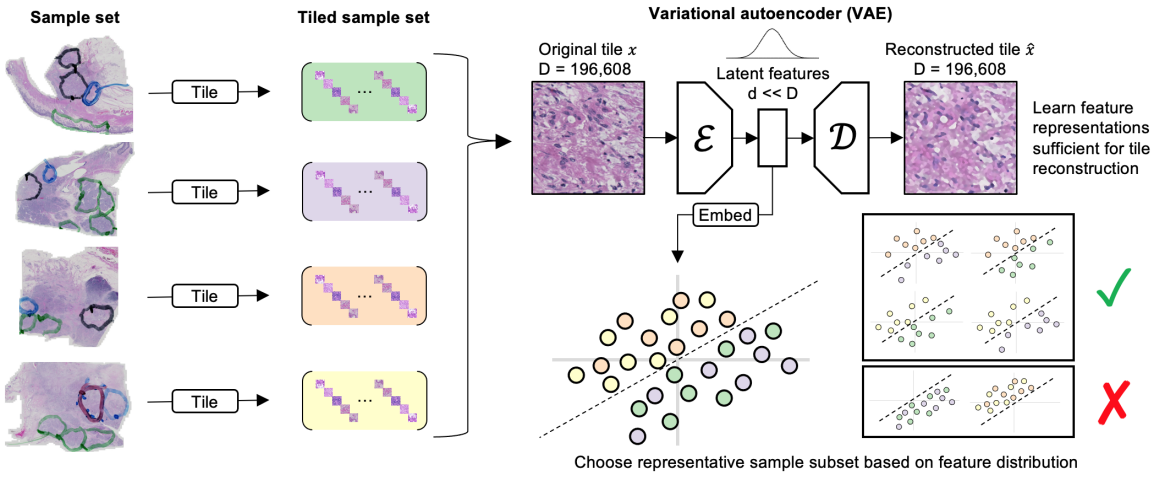


FIGURE 2.6: Schematic of feature-guided H&E sample selection. First, H&E samples are decomposed into 256 × 256 pixel tiles. Second, all H&E tiles are used to train a variational autoencoder (VAE) to learn feature representations for all tiles; for each 196,608-pixel H&E tile in the dataset, the encoder $\mathcal{E}$ learns a compact but expressive feature representation that maximizes the ability of the decoder $\mathcal{D}$ to reconstruct the original tile from its feature representation. Third, the tile feature representations are used to determine which samples are most representative of the whole dataset.

Using the VAE which we pre-trained on all samples, we extracted features from each H&E tile and assessed how each feature was distributed across samples, finding that several of the features discriminated between samples (Figure 2.7A).

In particular, the bimodal distribution of some features suggested two sample clusters, one formed by samples A and B, and another formed by samples C and D. These clusters were corroborated by visualization of the sample tiles embedded in the reduced-dimension feature space generated by *t*-SNE [69] (Figure 2.6 and Figure 2.7B). The *t*-SNE embedding is strictly used as a visual aid and validation of the quantitative selection of the most representative set of samples by the algorithm based on the full 16-dimensional VAE

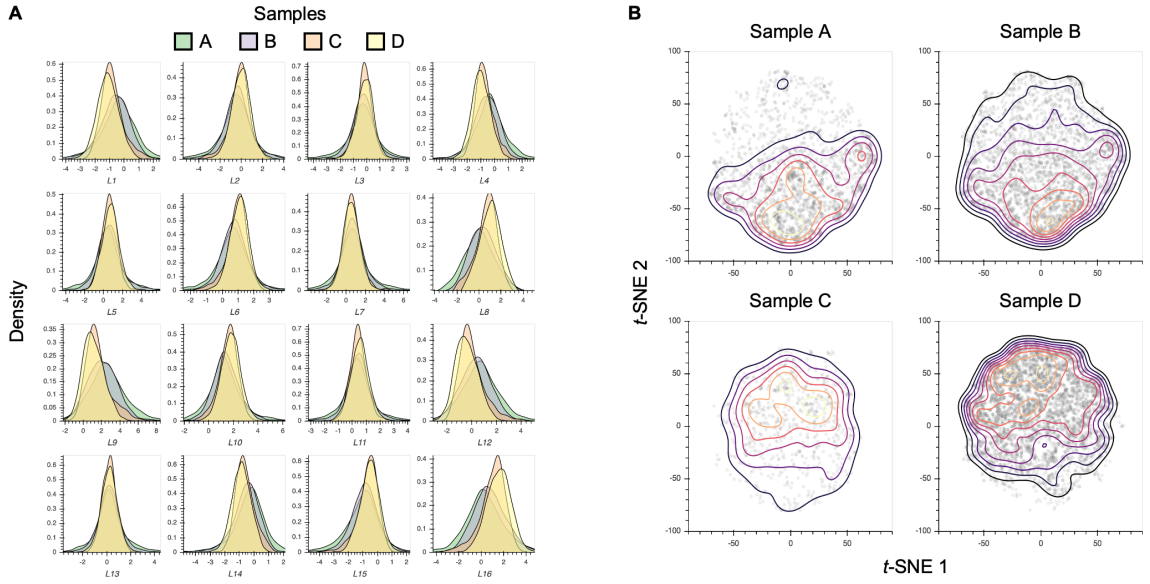


FIGURE 2.7: VAE features derived for feature-guided H&E sample selection. (A) Distribution of the 16 latent features (L1-L16) extracted by VAE from sample H&E tiles. (B) *t*-SNE embedding of latent feature representations of sample H&E tiles, faceted by sample identity. Each point in each plot represents a single H&E tile. Contour lines indicate point density.

features. Since the feature distributions of the H&E tiles highlighted the redundancy between clustered samples, we reasoned that a balanced selection of samples from each cluster would yield a more representative training set and ultimately improve SHIFT model generalizability. Using the full 16-dimensional feature representations of the H&E tiles and an information-theoretic framework for representative sample selection [83] (see subsection 2.3.12), we were able to quantitatively identify sample B and the duo of samples B and D as the single and two most representative samples, respectively, which were then considered for training sets in subsequent experiments. Figure 2.8 illustrates the feature distributions of several sample combinations in comparison to that of the full dataset.

Virtual IF staining in histological samples

SHIFT models are built on an adversarial image-to-image translation framework [48], with a regularization strategy designed to improve inference on sparse IF images (Figure 2.1 and

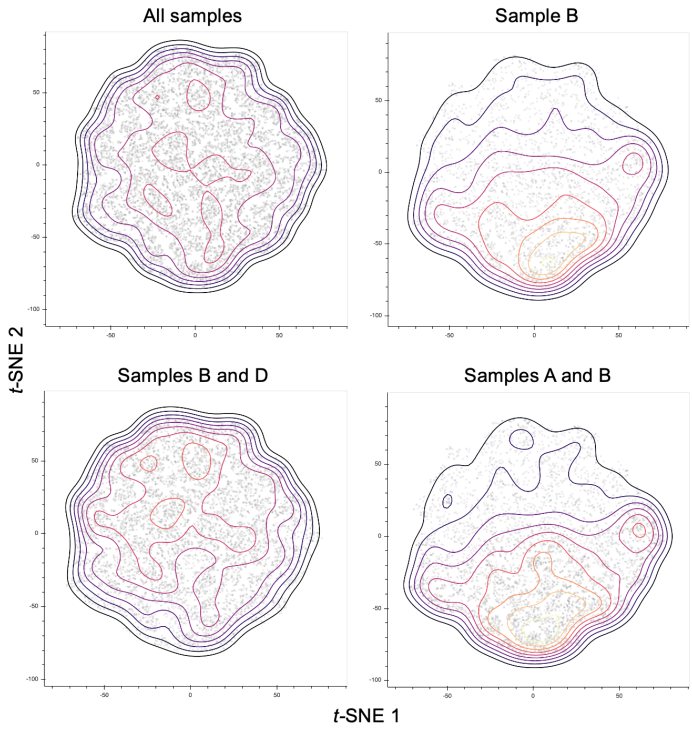


FIGURE 2.8: H&E tile feature distributions of experiment sample combinations. The VAE-learned 16-dimensional feature vector representations for each H&E are embedded into 2 dimensions using *t*-SNE. Each point in each plot represents a single H&E tile. Contour lines indicate point density. Sample B is the single sample that is most representative of the whole dataset. Samples B and D are the duo of samples that are most representative of the whole dataset. Samples A and B are a duo of samples that poorly represent the whole dataset.

Figure 2.2) [115]. Adversarial learning frameworks compute their losses over images, in contrast to strictly supervised learning frameworks where losses are computed over pixels, which has been suggested as a means to improve model inference on virtual staining tasks [22]. Having identified the most representative samples in our dataset, we next tested whether or not a SHIFT model could learn a correspondence between H&E and IF images that generalizes across samples.

We hypothesized that if tissue and cell morphologies observed in H&E-stained tissue are a function of a given marker, then it should be possible to infer the spatial distribution of that marker based on the H&E-stained tissue alone; that is, H&E-to-IF translation should be learnable and generalizable such that a model can be extended to samples from patients

that were not included in the training set. To test this hypothesis, we trained SHIFT models to generate virtual IF images of the cancer marker panCK conditioned on input H&E images alone. To simultaneously assess the utility of our sample selection method, we trained models using different combinations of sample subsets in the training set. Training sets consisted of paired H&E and IF image tiles from either sample subset B1 (from the most representative single sample), sample subsets B1 and D5 (from the most representative duo of samples), or sample subsets A1 and B1 (from a less representative duo of samples as counterexample). Sample subsets B1 and D5 were selected because they contained a similar number of tiles, providing a balance between the sample clusters. Once trained, SHIFT models are capable of translating H&E WSIs into virtual IF WSIs in tens of seconds. Model performance was quantified by measuring the structural similarity (SSIM) [122, 93], a widely used measure of image similarity as perceived by the human visual system, between corresponding virtual and real IF images from samples left out of the model's training set. The SSIM between two images is calculated over pixel neighborhoods in the images and provides a more coherent measure of image similarity than pixel-wise measures like Pearson's r. Considering the SSIM performance of the model trained on B1 as the baseline, we see a significant improvement in model generalizability on held-out sample C (Friedman statistic $= 428.4$, $p = 9.6 \times 10^{-94}$) and held-out subsets D1 (Friedman statistic $= 587.4$, $p = 2.7 \times 10^{-128}$), D2 (Friedman statistic $= 298.0$, $p = 2.0 \times 10^{-65}$), and D4 (Friedman statistic $= 3099.1$, $p < 2.2 \times 10^{-308}$) when the more representative sample subsets B1 and D5 are used for training than when the less representative sample subsets A1 and B1 are used (Figure 2.9). By stitching together the virtual IF tiles from a given sample in the test set, we were able to make large-scale comparisons between real and virtual panCK IF images (Figure 2.10). We also experimented with SHIFT inference of the stromal marker α-SMA (Figure 2.11).

SHIFT is not the only virtual staining method to have been recently proposed. Label-free determination (LFD) [81] is a supervised DL-based virtual staining method which

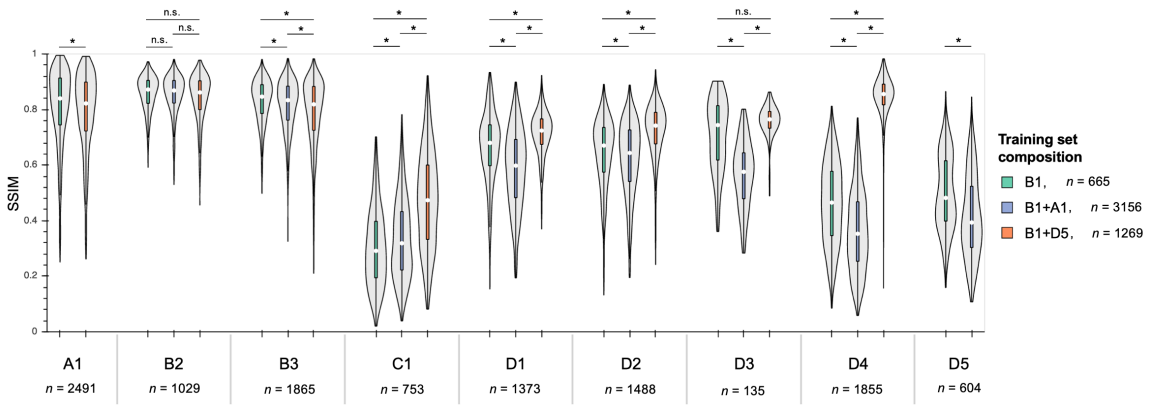


FIGURE 2.9: SHIFT panCK model test performance for optimal (B and D) and non-optimal (A and B) training set sample compositions. The paired H&E and IF images from samples B and D were subdivided into smaller images B=B1,B2 and D=D1,D2,D3,D4,D5 to avoid regions of IF that exhibited substantial autofluorescence. The x-axis labels indicate sample identity, where each letter corresponds to a unique sample and each number corresponds to a subset of that sample. Each n denotes the number of image tiles that were extracted from that sample. Plots for sample subsets are not show if that sample subset was a component of a model's training set. *p<.05; for three group comparisons we used the Friedman test with Nemenyi post-hoc test; for two group comparisons we used the Wilcoxon signed-rank test. White dots in violin plots represent distributional medians.

produces models that were shown to have learned the relationship between images of cell cultures visualized by transmitted light or fluorescence, where sub-cellular structures have been labeled with genetically-encoded fluorescent tags. Because the SHIFT generator G and the LFD are both based on the popular U-Net architecture [95], we compared these models that generate images using a similar architecture, but have differing training formulae and loss functions. To make a fair comparison between the adversarial SHIFT and supervised LFD models, we trained a LFD model using the representative sample subsets B1 and D5, matching the number of optimization steps taken by the SHIFT model that was trained using the same training set (Figure 2.12).

In addition to the performance of independent SHIFT and LFD models, we also considered the ensemble result, taken as the average image of the SHIFT and LFD output images (Figure 2.13A). Across all samples in the test set, either SHIFT alone or the ensemble of SHIFT and LFD tended to perform better than LFD alone (Figure 2.13B). In addition to

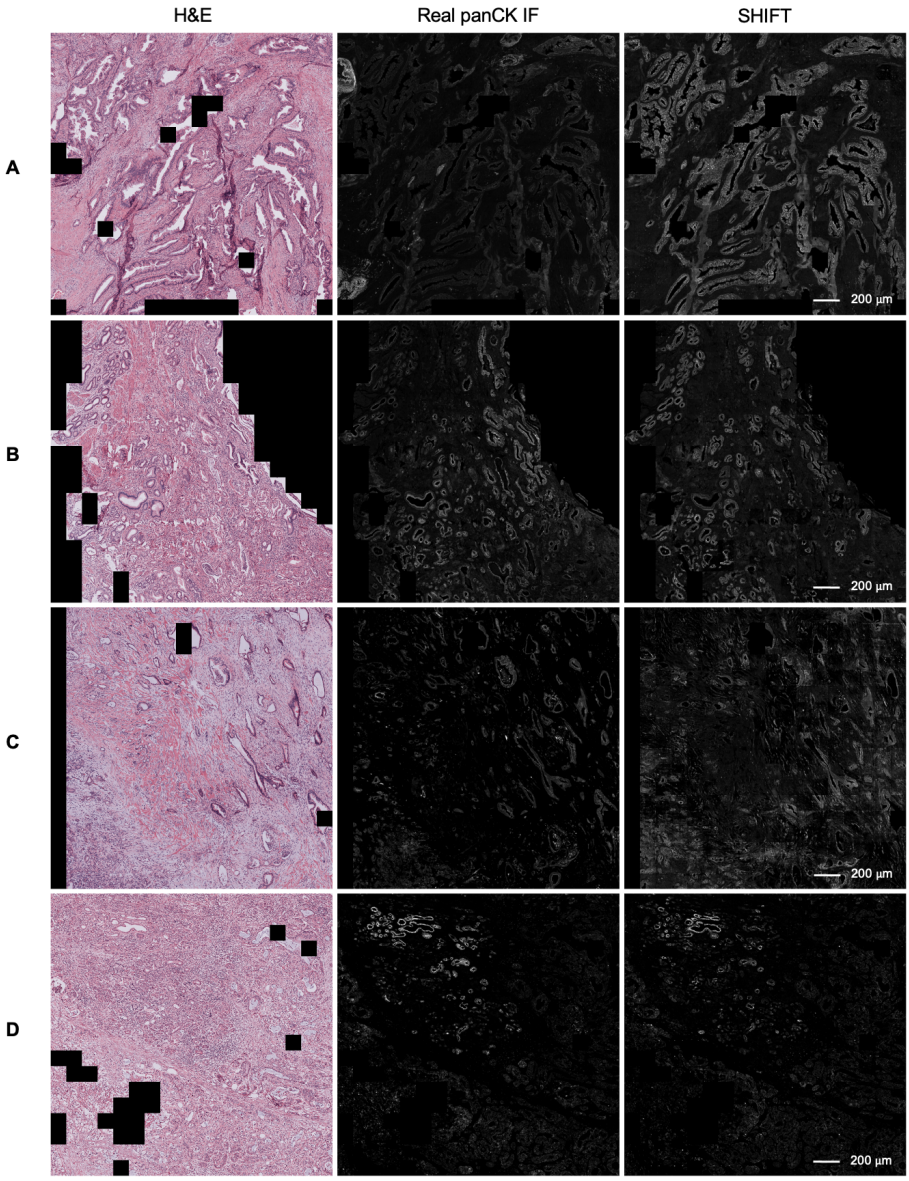


FIGURE 2.10: Large-scale comparison of real and virtual panCK staining generated by SHIFT. SHIFT images were generated by a model trained on sample subsets B1 and D5. Results shown are from the test set. Tiles were excluded if they contained more than 50% background in the H&E representation (black tiles). (A) Representative images taken from sample A. Robust staining of fibrotic vasculature in the upper- and lower-left of the real panCK image is not recapitulated in the SHIFT image because panCK+ fibrotic vasculature was not present in the model's training set of samples B and D. Rather, the desired virtual staining of tumor epithelium is generated. (B) Representative images taken from sample B. (C) Representative images taken from sample C. (D) Representative images taken from sample D.

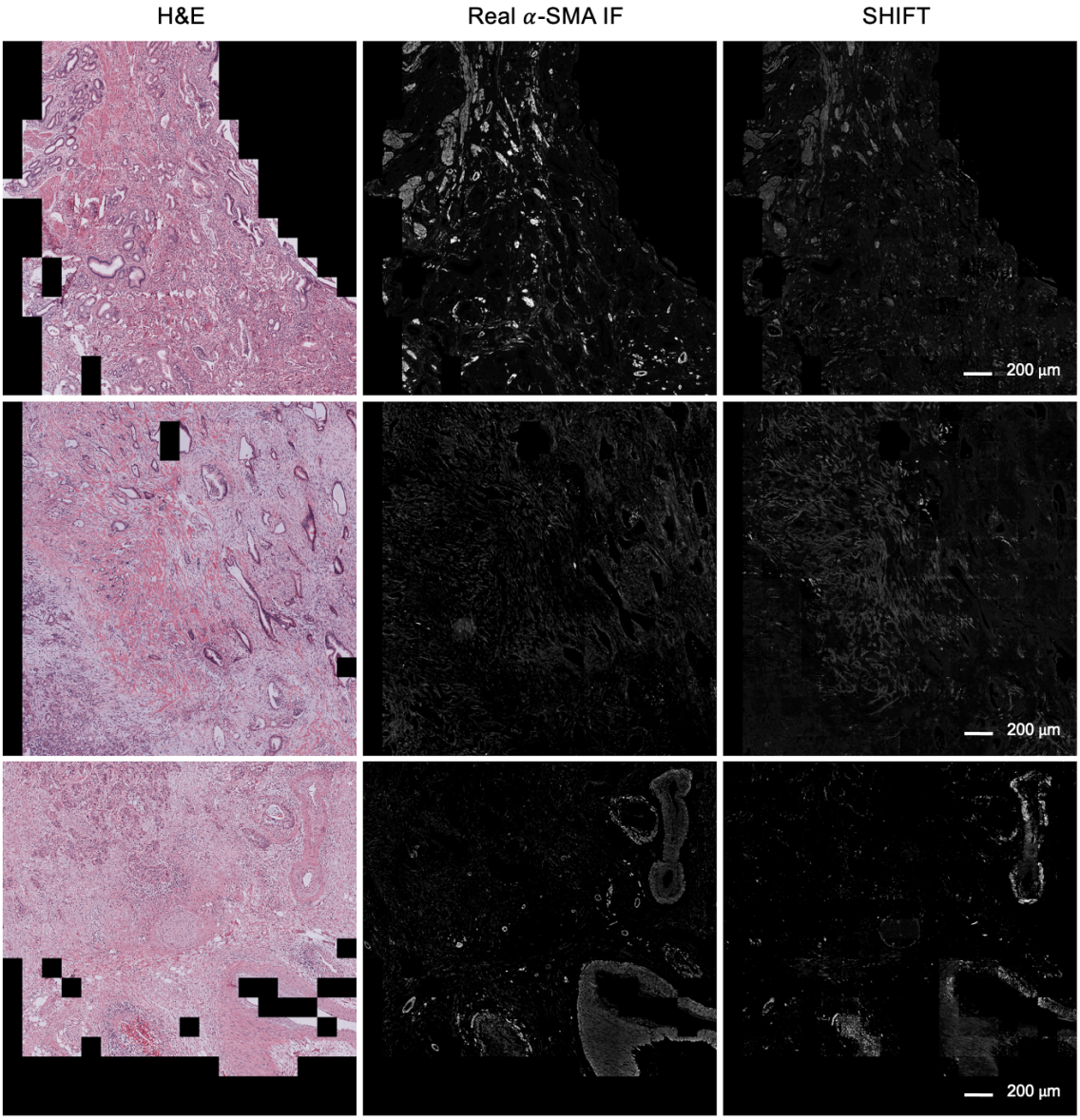


FIGURE 2.11: Large-scale comparison of real and virtual α-SMA staining generated by SHIFT. SHIFT images were generated by a model trained on sample subsets B1 and D5. Results show are from the test set. Tiles were excluded if they contained more than 50% background in the H&E representation (black tiles). Discrepancies between the real and virtual stains suggest that our dataset is of insufficient size to optimally model the inter-sample heterogeneity of α-SMA expression.

comparing SHIFT and LFD models, we also tried removing the discriminator and adversarial loss term from the panCK SHIFT model, leaving just the U-net generator. We find

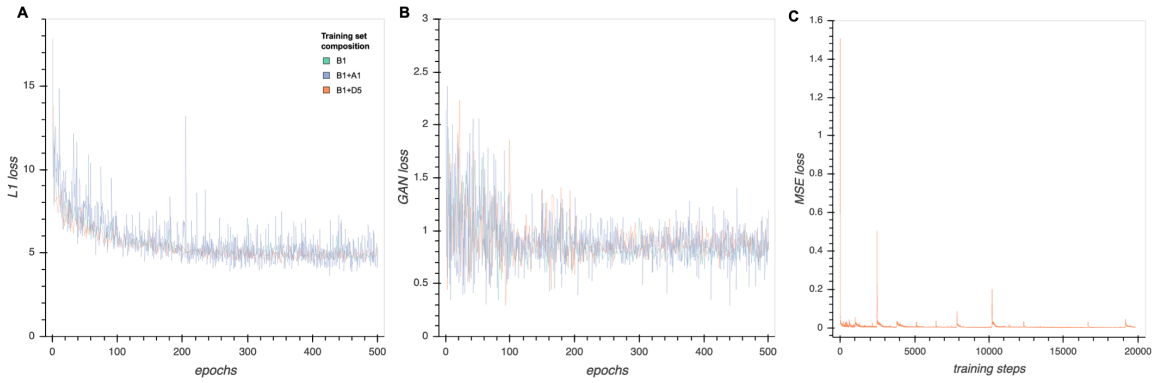


FIGURE 2.12: Training losses for virtual staining models. (A) L1 training loss for SHIFT models for each training set composition. (B) GAN training loss for SHIFT models for each training set composition. (C) Mean squared error (MSE) training loss for Label-Free Determination (LFD) model.

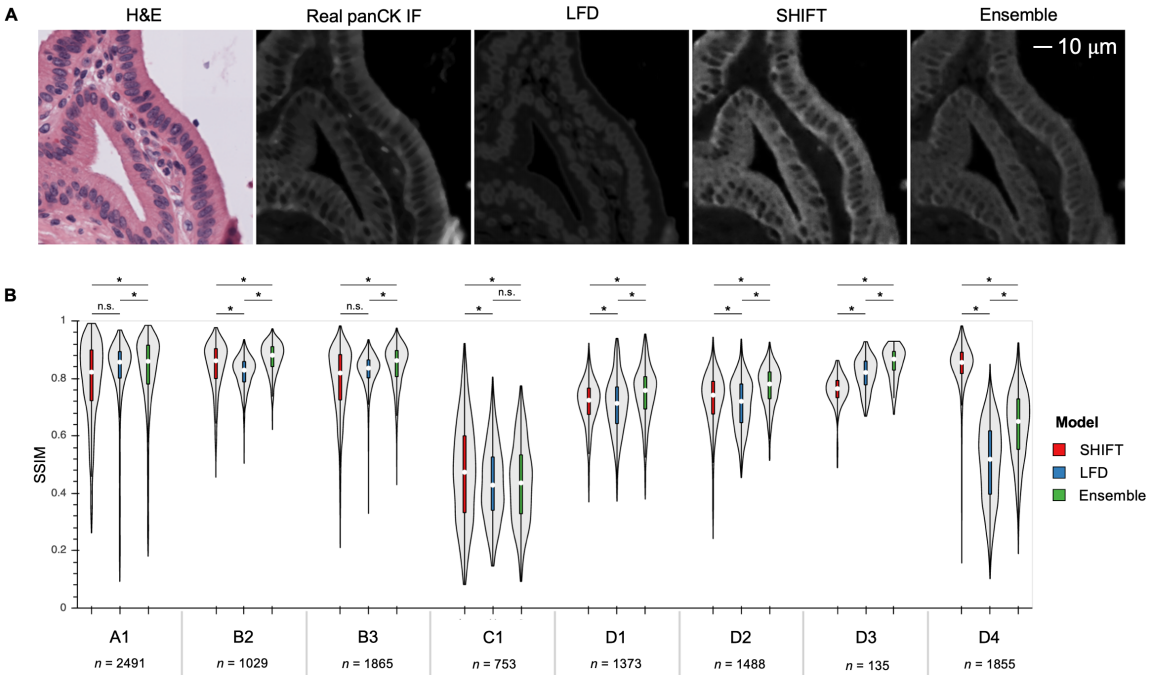


FIGURE 2.13: (A) Visual comparison of virtual staining methods. The ensemble results are attained by averaging the output images of SHIFT and Label-Free Determination (LFD) models. See also Figure 2.2. (B) Test performance comparison of virtual staining methods. The x-axis labels indicate sample identity, where each letter corresponds to a unique sample and each number corresponds to a subset of that sample. Each n denotes the number of image tiles that were extracted from that sample. Plots for sample subsets B1 and D5 are not show because those sample subsets were components of the models' training sets. *p<.05; Friedman test with Nemenyi post-hoc test. White dots in violin plots represent distributional medians.

that these models trained using the pixel-wise L1 loss alone produced virtual panCK staining with good localization, but poor resolution of finer cellular structure, highlighting the importance of the adversarial loss for producing realistic virtual stains (Figure 2.14).

2.5 Discussion

Spatially-resolved molecular profiling of cancer tissues by technologies like IF provides more information than routine H&E histology alone. However, the rich information obtained from IF comes at significant expense in time and resources, restricting IF access and use. Here, we present and extend the validation of SHIFT, a DL-based method which takes standard H&E-stained histology images as input and returns virtual panCK IF images of inferred marker distributions. Using a limited but heterogeneous dataset, we demonstrated that SHIFT models are able to generalize across samples drawn from different PDAC patients, even for training sets that are over an order of magnitude smaller than the test set (train n=665 and test n=11,593 for models trained on sample subset B1 only). Results from our sampling experiments are consistent with the expectation that an automated and quantitative method for representative sample selection will be critical to the effective development and deployment of DL models on large-scale digital pathology datasets. Finally, we compared the adversarial SHIFT method with an alternative, supervised virtual staining method and found that the virtual staining task tends to be best accomplished by the ensemble of both methods. With the incorporation of an adversarial loss term, a SHIFT model computes loss over images, rather than strictly over pixels as do other virtual staining methods [22, 81], which may explain its positive contribution to the model ensemble. Based on the success of DL-based ensemble methods in other biomedical domains [127, 23], we expect ensemble methods to become increasingly relevant to the development of virtual staining applications.

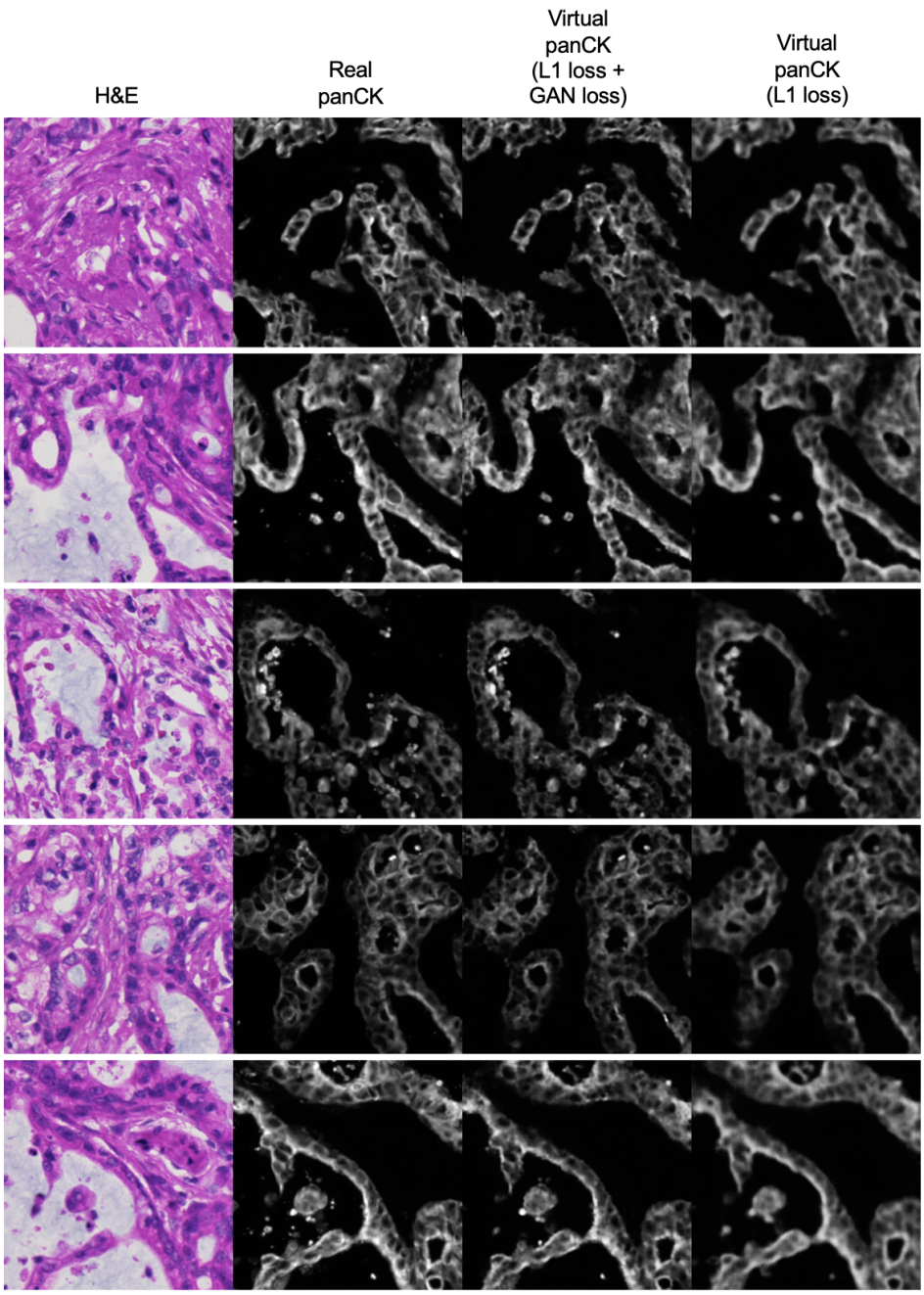


FIGURE 2.14: Comparison of different training losses for models estimating panCK. Training examples for models trained to convergence using either L1 and GAN losses, or L1 loss alone. Stain estimation for the model trained using L1 loss alone lacks high-frequency textural details of the panCK stain when compared to the model using both L1 and GAN losses.

While we have demonstrated the application of SHIFT for the estimation of an IF tumor signature, there are emerging opportunities to test the extensibility of our approach and relate it to the approaches of others. In particular, a comparison between SHIFT and methods which estimate single-stain IHC status conditioned on H&E images [49, 25] would be insightful, since it remains unclear whether there is value added in learning from either IF or chromogenic signals. Following recent advances in cyclic immunofluorescence and multiplex immunohistochemistry (CyCIF/mIHC) technology [59, 112, 38, 90], it is now possible to visualize tens or hundreds of distinct markers in a single tissue section. On their own, these technologies promise a more personalized medicine through a more granular definition of disease subtypes, and will undoubtedly broaden our understanding of cellular heterogeneity and interaction within the tumor microenvironment, both of which play increasingly important roles in the develop-ment and selection of effective treatments [130, 68]. With a paired H&E and CyCIF/mIHC dataset that encompasses the expression of hundreds of markers within the same (or serially-sectioned) tissue, we could begin to quantify the mutual information between histology and expression of any marker of interest. Notwithstanding the prospect of virtual multiplexing, virtual panCK IF alone could be of useful to spatial profiling platforms which use panCK IF to label tumor regions for localized spatial profiling of protein and RNA abundances in formalin-fixed tissues [75].

There are obvious limitations and challenges to both the feature-guided sampling and virtual staining methods we present here. Both methods assume an association between H&E and IF representations of tissue. Since this is unlikely to be the case in general, determining which markers have a histological signature will be essential to the evaluation of clinical utility for virtual staining methods. For markers without a H&E-to-IF association, our methods may fail to maximize representativeness or make incorrect estimates of IF signals. This should be seen as a feature rather than failure of our methods, since they provide a means of quantitatively delineating markers that have a H&E-to-IF association from

those that do not. Notwithstanding, even when there is an association, finding a meaningful way to compare real and virtual images remains a challenge, as we have experienced in experiments modeling markers with fine, high-frequency distributions like α-SMA (Figure 2.11). Like other virtual staining methods that have been deployed on whole human tissues [92, 93], we used SSIM as a measure of image similarity between real and virtual IF images. This classical perceptual measure is used in many imaging domains, but we found that it is sensitive to perturbations commonly associated with image registration and technical or instrumentation noise (Figure 2.15). In light of this, we advocate for the development and use of perceptual measures that are more aware of such perturbations and better correlate with human perception of image similarity or quality [6, 85].

It must be restated that our results are supported by a dataset comprised of samples from just four patients, so some fluctuation in performance between samples should be expected, and indeed was observed (Figure 2.10). Our goal in choosing a relatively small dataset was to demonstrate that, even when limited, SHIFT could learn a general relationship between H&E and IF tissue representations, and we believe that the fluctuation in performance between samples could be addressed by increasing the sample size. With the emergence of digital pathology datasets containing tens of thousands of whole slide images [14], the opportunities to improve virtual staining technologies are only becoming more numerous.

In spite of representing four patients, the samples in our study were selected by a board-certified pathologist to be as heterogenous as possible, encompassing the spectrum of PDAC morphology, albeit in as few samples as possible. While we used these heterogeneous and associated human tissues for our study, other virtual staining methods have only been demonstrated on relatively homogeneous human or rat cell lines [22, 81] and with far less total image area (Table 2.6). Given the precedent set by these prior works, we feel that the work we present here is within scope as a proof-of-concept study of sample selection and histology-based IF prediction in a digital pathology application. Moreover, the

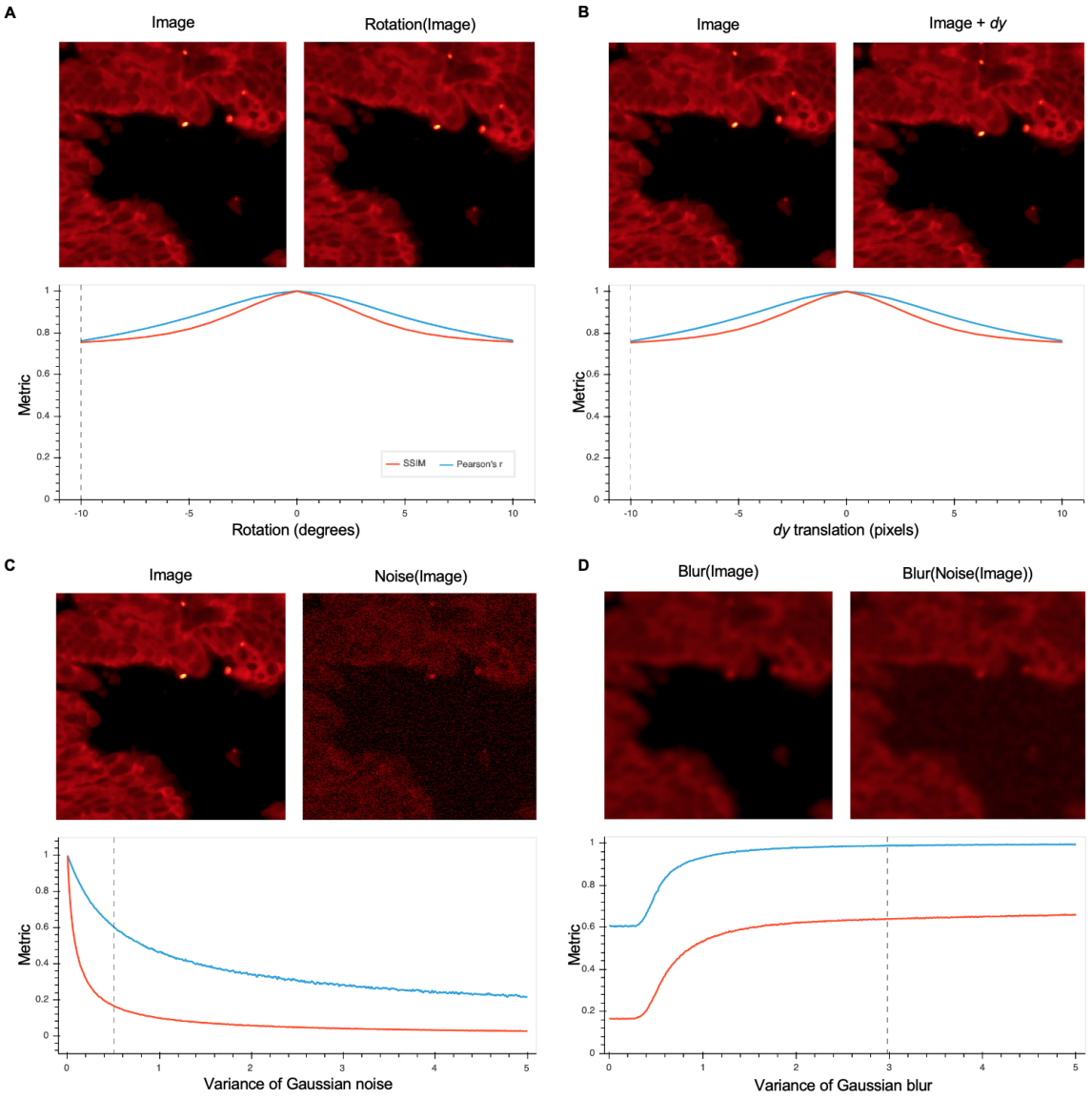


FIGURE 2.15: Model performance metric sensitivity to common technical perturbations. Structural similarity (SSIM) and Pearson's r are two commonly used metrics for image comparison, both having been used to make comparisons between real and generated biological images in recent related work [93, 22, 81]. When comparing raw real IF and raw generated IF images, a low SSIM value may be the result of sensor/technical noise in the IF procedure, which is impossible for the SHIFT model to predict based on the H&E image it is given as input. When considering a stereotypical panCK IF tile, both the SSIM (red traces) and Pearson's r (blue traces) are found to be sensitive to rotation (A) and translation (B), perturbations common to image registration. Grey dotted lines indicate the parameter selected to generate each transformed image. We also observe that both measures are sensitive to simulations of technical noise (C). By applying a Gaussian filter with variance (sigma) set to 3, we recover the SSIM between real and perturbed IF images without sacrificing global image details (D).

protocol that we developed to allow H&E and IF staining in the same tissue sections will be of significant value to the community, since spatially-paired H&E and IF data is difficult to generate from adjacent sections due to tissue deformation and cellular discontinuity between sections.

Reference	Target marker	Image tiles	XY tile resolution (pixels)	Pixel resolution (microns/pixel)	Total image area	Difference in area from current study
[22]	DAPI	2	3500×3500	0.32	78 mm2	45-fold less
[81]	Lamin	40	924×624	0.108	$25\,mm^2$	140-fold less
Current study	panCK	12258	256×256	0.44	$3.5\,cm^2$	-

TABLE 2.6: Comparison of total image area used for training and testing of virtual staining methods.

Since SHIFT can infer virtual panCK IF images as H&E-stained tissue section are imaged, SHIFT could provide pathologists with near-real-time interpretations based on standard H&E-stained tissue in augmented set-tings. Therefore, SHIFT could serve as an efficient preliminary, auxiliary, or substitute technology for traditional panCK IF in both research and clinical settings by delivering comparable virtual panCK IF images for a fraction of the cost and in a fraction of the time required by traditional IF or CyCIF/mIHC imaging. With clinical validation in larger cohorts, the advantages of a SHIFT model over traditional IF would include (1) eliminating the need for expensive imaging hardware, bulk reagents, and technical undertaking of IF protocols; (2) virtual IF images can be generated in near-real time; and (3) the portability of SHIFT allows it to be integrated into existing imaging workflows with minimal effort. As such, we see further validation of SHIFT as an opportunity to simultaneously economize and democratize advanced imaging technologies in histopathology workflows, with implications for multiplexed virtual imaging. Further, we see our methods for the optimal selection of representative histological images, which promote morphological heterogeneity in the training dataset as well as reduce unnecessary effort on IF staining, as a complement to data augmentation, transfer learning, and other means of addressing the problem of limited training data. This will contribute to saving resources and minimizing unnecessary efforts to acquire additional staining or manual annotation for DL applications in biomedical imaging.

Chapter 3

Extending virtual staining into 3D

Spatial patterns in TMAs are pure noise.

An esteemed systems biology professor, 2021

3.1 Abstract

Tumors are not 2-dimensional (2D), but many multiplex tissue imaging platforms (MTIs) make the assumption that tissue microarrays containing small core samples of 2D tissue sections are a good approximation of bulk tumor. However, emerging 3D tumor atlases which employ MTIs like cyclic immunufluorescence (CyCIF) strongly challenge this assumption. In spite of the additional insight gathered by measuring the tumor microenvironment in 3D, it can be prohibitively expensive and time consuming to process tens or hundreds of tissue sections with CyCIF. Even when resources are not limiting, the criteria for region-of-interest (ROI) selection in tissues for downstream analysis remain largely qualitative and subjective. To address these challenges, herein we demonstrate that generative modeling enables a 3D virtual CyCIF reconstruction of a colorectal cancer specimen given a small subset of the imaging data at training time. By co-embedding histology and MTI features, we go on to formulate a generative basis for objective ROI selection.

3.2 Introduction

Cancers are complex diseases that operate at multiple biological scales—from atom to organism—and the purview of cancer systems biology is to integrate information between scales to derive insight into their mechanisms and therapeutic vulnerabilities. From this holistic perspective, the field has come to appreciate that the spatial context of the tumor microenvironment in intact tissues not only enables a more granular definition of disease, but also the design of more personalized and effective therapies. This has motivated the National Cancer Institute's Human Tumor Atlas Network (HTAN) to begin charting 3D tissue atlases which capture the multiscale organizations and interactions of immune, tumor, and stromal cells in their anatomically native states [96].

The HTAN-SARDANA [61] is one such atlas which aimed to deeply characterize the architecture of one whole colorectal cancer (CRC) specimen via histology and a spatial context-preserving multiplex tissue imaging (MTI) platform called cyclic immunofluorescence (CyCIF) [60] (Figure 3.1). Histology is an essential component of the clinical management of cancer. For around 150 years, pathologists have interrogated thin sections of tissue stained with hematoxylin and eosin (H&E) to determine the morphological correlates of cancer grade, stage, and prognosis. However, this essentially 2D representation of tissue is a relatively poor representation of tissues like prostate, pancreas, breast, and colon which have highly convoluted 3D ductal structures [62, 54, 15, 61]. Moreover, histology alone lacks the molecular specificity to unequivocally determine the identity and function of cells in tissue. In contrast, the more recently developed CyCIF enables the co-labelling of tens of markers in tissue and can broadly characterize the tumor, immune, and stromal compartments. By coupling histology and CyCIF in the same specimen, the HTAN-SARDANA atlas integrates both top-down (pathology-driven) and bottom-up (single-cell phenotype-driven) perspectives of CRC and provides a framework for the charting of 3D atlases for other cancers [61].

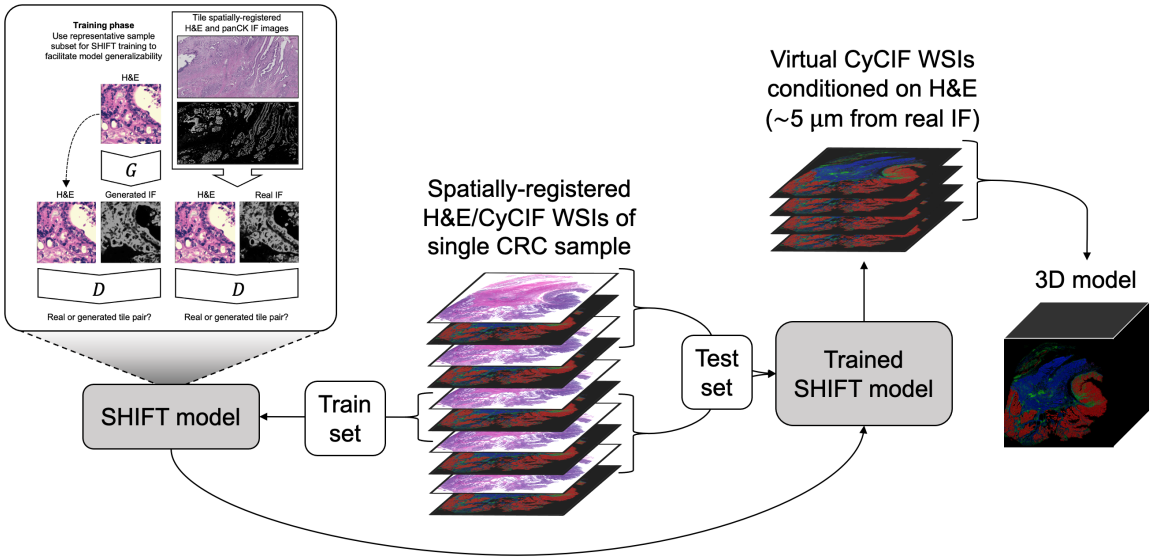


FIGURE 3.1: HTAN-SARDANA dataset and SHIFT modeling overview. Extending SHIFT to 3D using adjacent spatially-registered H&E/CyCIF WSIs from a single CRC sample. We make the assumption that the central pair of H&E/CyCIF sections is a good representation of the collection of sections, i.e. that it is most likely to capture features from either end of the section stack. Using this central section pair, we train individual SHIFT models to predict individual CyCIF images conditioned on H&E images. At test time, we apply the trained SHIFT models to the remaining held-out H&E sections. Using the stack of SHIFT-generated virtual CyCIF WSIs, we can reconstruct a 3D virtual stain volume.

In spite of these advances, 3D atlases require a tremendous amount of resources and effort to build. For the HTAN-SARDANA atlas, a single CRC specimen was serially sectioned and processed yielding 22 H&E slides interleaved with 25 CyCIF slides, with the CyCIF slides taking days to process due to the cycles of antibody incubation. For the breast cancer atlas described in [15], a single breast cancer specimen was serially sectioned and processed into 156 slides which were characterized using imaging mass cytometry, which enables simultaneous labeling of 40 antigens with a single incubation step, but has relatively limited spatial scope compared to CyCIF. For the pancreas cancer atlas described in [54], specimens were serially sectioned and processed into over 1000 H&E slides, some of which had histological regions of interest labeled through a tedious and subjective manual annotation process. These annotations were used as training data for a deep learning

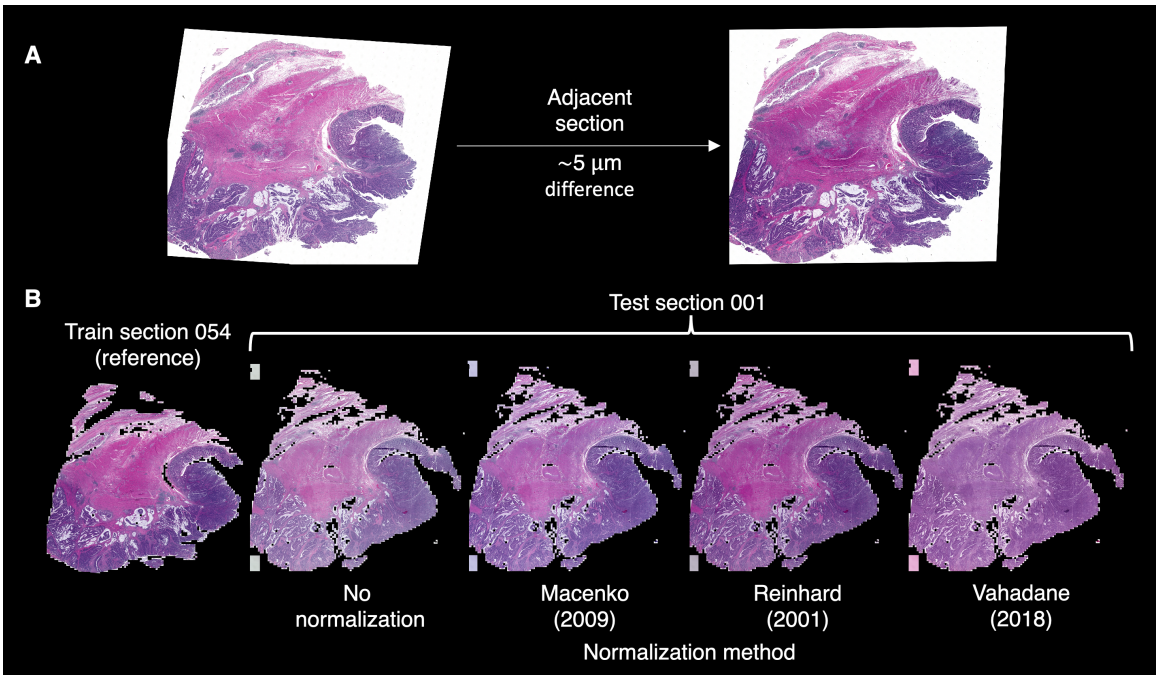


FIGURE 3.2: H&E stain normalization overview. (A) Tissue sections are subject to technical variability in stain intensity, even between adjacent sections that are separated by only ~5 µm. (B) Representative results of H&E stain normalization. The stain intensity distribution of the test section 001 is transformed to match that of the reference section 054 which was used for SHIFT model training. We experimented with the application of three different H&E stain normalization methods [71, 89, 117] and found that the Reinhard method best matched the test stain distribution to the training stain distribution by qualitative comparison. This result was consistent with a quantitative comparison that found the Reinhard method conferred better generalizability to DL models in an analogous digital pathology application [110].

segmentation model which was used to fully reconstruct the semantically-labeled 3D specimen with high accuracy.

We have previously demonstrated methods for learning virtual IF stains [116], wherein we use spatially-registered H&E and IF data and generative deep learning to model the correspondences between these imaging modes. In doing so we learn to compute near-real time virtual IF stains conditioned on H&E-stained tissue alone. From a biological perspective, these data and approach allow us to ask which markers in an IF panel have a quantifiable histological signature, what that signature might be, and a means to estimate the distribution of markers in histological images for which such a signature exists. From

an applications perspective, the approach could be useful for automated compartment labeling and ROI suggestion in histologically-labeled tissues.

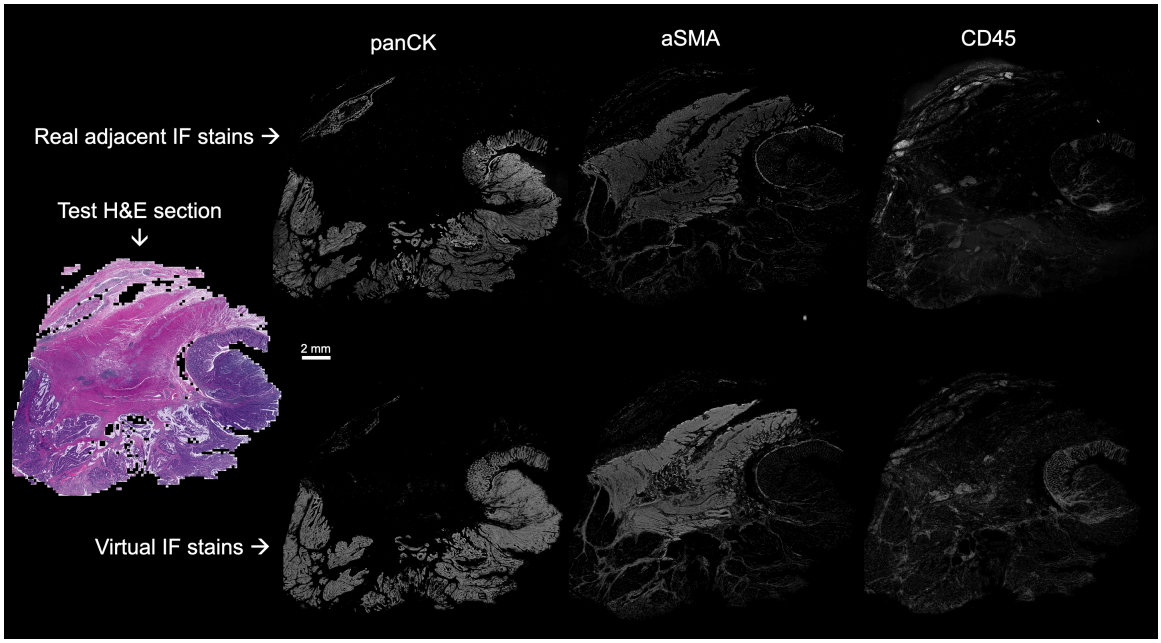


FIGURE 3.3: WSI virtual staining test results for panCK, aSMA, and CD45. Models trained to predict single-channel CyCIF images conditioned on the H&E/CyCIF training sections were applied to H&E test section 096 to generate virtual stain WSIs for the markers panCK, aSMA, and CD45. The input H&E test section is shown at left, and the real and virtual CyCIF WSIs are shown in the rows above and below, respectively, for ease in comparison.

In the present study, we extend the virtual staining paradigm into the third dimension by deploying it on the coupled H&E and CyCIF image data from the HTAN-SARDANA atlas of CRC. We demonstrate that what generative models learn from less than 5% of coupled H&E and CyCIF images is sufficient to generate a virtual 3D CyCIF reconstruction of the whole CRC specimen and that quantitative endpoints derived from real and virtual CyCIF images are highly correlated.

3.3 Results

3.3.1 3D virtual CyCIF reconstruction and evaluation

The conceptual overview of the HTAN-SARDANA dataset and our virtual staining experiments is presented in (Figure 3.1). Spatially registered H&E and IF images are a requirement for SHIFT model training and evaluation. To register the H&E and CyCIF data for this task, we begin with sequential registration of the H&E stack beginning from the middle sections and propagate to outer sections. We then co-register ROIs of adjacent H&E and CyCIF images using their respective nuclear masks for a finer local registration of the adjacent sections.

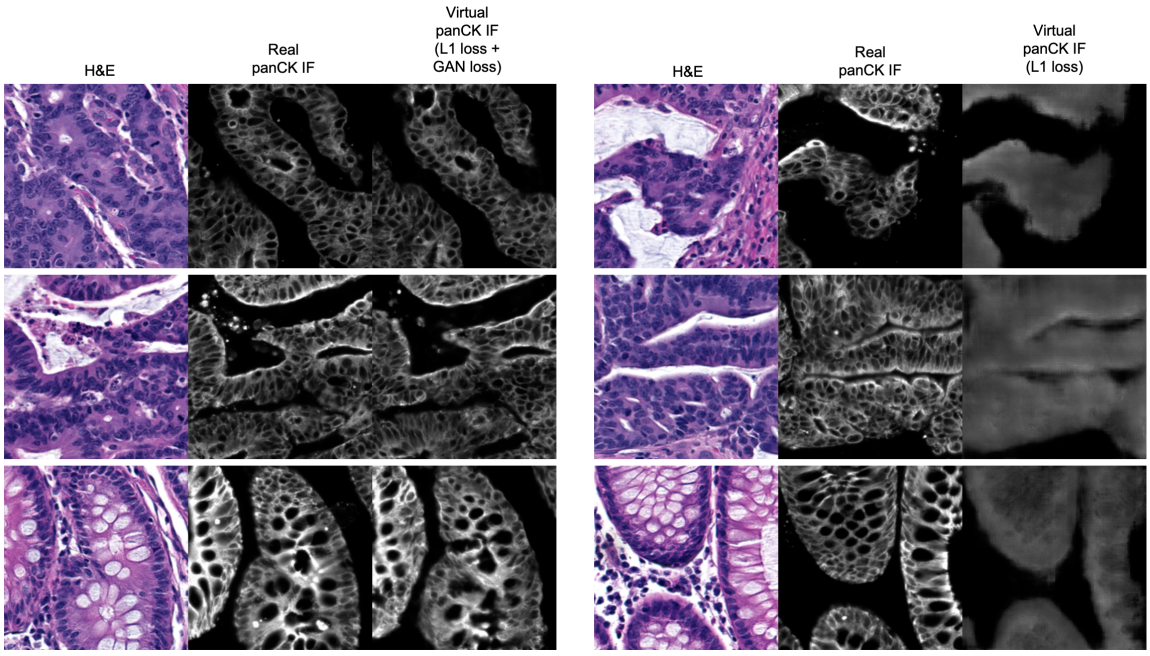


FIGURE 3.4: Virtual staining outcomes with different loss functions.

Before SHIFT model training could begin, we had to account for the section-to-section variability in H&E stain intensity, which helps to ensure a model trained on one H&E section generalizes well to the other sections. Using the training H&E section as reference, we tried several stain normalization methods, and found that the Reinhard method worked

best at normalizing stain intensities to the reference (Figure 3.2). With registered data in hand, we set out to generate a virtual 3D CyCIF reconstruction in an effort to measure how faithfully we can characterize the full SARDANA dataset with virtual IF staining by learning from only one adjacent pair of H&E and real CyCIF sections. We go about this by first selecting the middle pair of H&E and CyCIF sections for training SHIFT models, under the assumption that they are a good representation of the tissue on either side of the sample block. We then decompose the WSIs into thousands of pairs of matching H&E and IF image tiles, and use those to train a conditional generative adversarial network (cGAN) [48, 116] to synthesize virtual IF tiles conditioned on H&E tiles. Briefly, the generator network of the model is responsible for synthesizing virtual IF images conditioned on H&E images, and the discriminator network is responsible for quality assurance of the virtual IF images synthesized by the generator. Once trained on the middle sections, the model can then be tested by feeding it tiles from the held-out H&E sections to generate virtual IF images for comparison with the real CyCIF images. Importantly, a virtual IF image is conditioned on an H&E section, and there is natural variation between it and its adjacent real IF section 5 µm away, which complicates pixelwise evaluation of model accuracy.

We trained individual SHIFT models to predict single CyCIF channels conditioned on H&E inputs from the central H&E/CyCIF training sections 053/054 (Figure 3.1). Representative test results from the application of trained SHIFT models on H&E/CyCIF test sections 096/097 are shown in Figure 3.3. These qualitative results indicated that the SHIFT models were fitting well to the training sections, and the representations learned were useful for extension to held-out test sections, motivating a quantitative evaluation of model performance.

We also assessed the value added by the discriminator network of the GAN by training models without it, leaving the generator network to learn the virtual panCK stain alone (Figure 3.4). We found that while the generator-only virtual panCK stain has good localization, it lacks the naturalistic texture of the real and GAN-generated virtual stains, which

highlights the compromise of a more efficient and portable generator-only model.

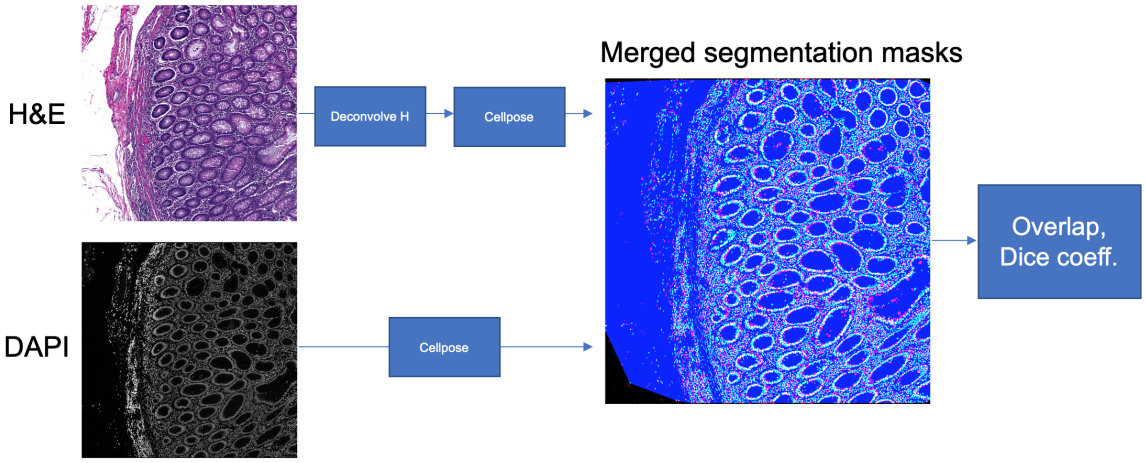


FIGURE 3.5: Difference in image content between adjacent sections estimated using nucleus overlap. Estimating upper bound on SHIFT performance by measuring concordance between nuclei in adjacent sections for locally- registered ROIs from H&E/CyCIF test sections 096/097. For H&E ROIs, we deconvolve the hematoxylin stain to extract nuclear content intensity [99], then segment the intensity to derive binary nuclear masks using Cellpose [109]. For CyCIF ROIs, we use Cellpose to segment DAPI intensity to derive binary nuclear masks. The white regions indicate overlap between nuclear masks from adjacent sections, and magenta and cyan regions indicate non-overlapping nuclear masks from the H&E and CyCIF sections, respectively. The blue region indicates the convex hull of the merged nuclear masks, which was used to measure and compare the spatial extent of nuclear staining over adjacent sections. The Dice coefficients describing the overlap of nuclear masks from ROIs of adjacent sections were used as compensation factors for evaluating virtual stains.

The virtual CyCIF images generated by SHIFT models are conditioned on H&E sections which are 5 µm adjacent to the real CyCIF sections, so the cellular contents are slightly different between sections and images, a difference which can be compounded by slight errors in image registration. Recognizing that these differences would hamper pixelwise comparisons between the real and virtual images [115, 116], we estimated an upper bound on SHIFT performance by measuring the concordance between nuclear content from the adjacent sections of the H&E/CyCIF test sections 096/097 (Figure 3.5).

The test sections were first subdivided into 135 non-overlapping ROIs and each ROI was locally registered to improve the alignment of H&E and CyCIF image content, then we measured the Dice coefficient of nuclear masks derived from the H&E and DAPI images from each ROI (Figure 3.6A). We used the Dice coefficient for each ROI as a compensation

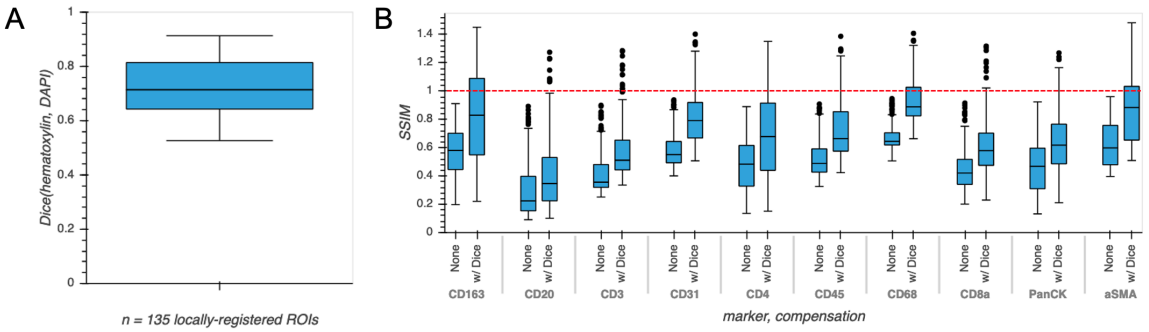


FIGURE 3.6: Nuclear overlap compensation for virtual staining evaluation. (A) Boxplot describing the distribution of Dice coefficients of the 135 locally-registered ROIs from H&E/CyCIF test sections 096/097. (B) Boxplots describing the distributions of structural similarity (SSIM) of real vs. virtual CyCIF ROIs over the 135 locally-registered ROIs from H&E/CyCIF test sections 096/097. The Dice-compensated SSIM values are calculated by taking the SSIM of the virtual CyCIF ROI with respect to the real CyCIF ROI and dividing it by the Dice coefficient of nuclear overlap between the hematoxylin and DAPI nuclear masks from sections 096/097 for that ROI. The red dotted line indicates the unity line where SSIM==Dice compensation factor for the boxplots describing Dice-compensated SSIM.

factor when evaluating the quality of the virtual stains for each ROI by dividing raw quality scores by the Dice coefficients corresponding to each ROI. Virtual CyCIF image quality was evaluated using structural similarity (SSIM), which is established as a metric for assessing virtual stain quality [92, 93, 116]. The median compensated SSIM for virtual stains ranged from 0.36 for CD20 up to 0.89 for aSMA. This result suggested that there was significant room for improvement for some SHIFT models, but we hypothesized that the virtual images might still be useful in the hands of a CyCIF domain expert, since SSIM is sensitive to slight differences in image contrast which may not significantly affect downstream processing [116].

Using a selection of pathologist-annotated regions within H&E test section 096 (Figure 3.7), we quantified the positive cell ratio for multiple markers in each region using either real or virtual CyCIF images to assess how such an endpoint might be impacted when using virtual images which may or may not be of high quality with respect to SSIM (Figure 3.8).

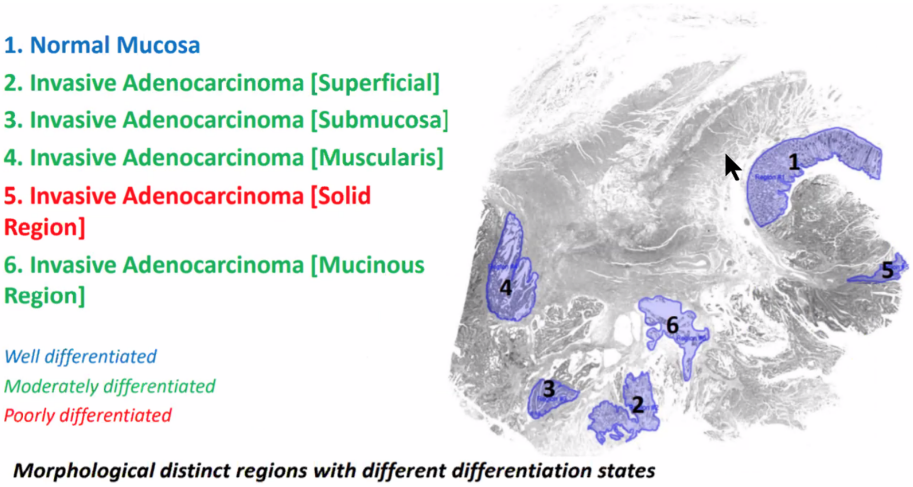


FIGURE 3.7: Pathologist annotation of H&E test section 096 into 6 different ROIs. Adapted from a slide by Jia-Ren Lin, Co-Director of the Tissue Imaging Platform at the Labratory of Systems Pharmacology, Harvard Medical School.

In spite of the adjacency complication explained above, there was substantial correlation between positive cell ratios using real and virtual CyCIF images, suggesting that virtual images could be used in place of real without significantly affecting some downstream endpoints. Having established the utility of the virtual images, we performed a full virtual 3D reconstruction of the CyCIF images be passing all held-out H&E test sections to the SHIFT models trained on H&E/CyCIF training sections 096/097 (Figure 3.9).

3.3.2 Co-embedding H&E and CyCIF image representations

Virtual staining is enabled through the rich latent representations that generative models are capable of learning from paired H&E and CyCIF image data. We hypothesized that these latent representations could be useful for the related and unsolved problem of objective ROI selection. To that end, we built an architecture which learns to co-embed H&E and CyCIF representations of the same tissue into the same latent representation (Figure 3.10). This architecture builds upon previous works in cross-domain data translation [63, 102].

As a proof of concept of the proposed architecture in a minimal working example, we

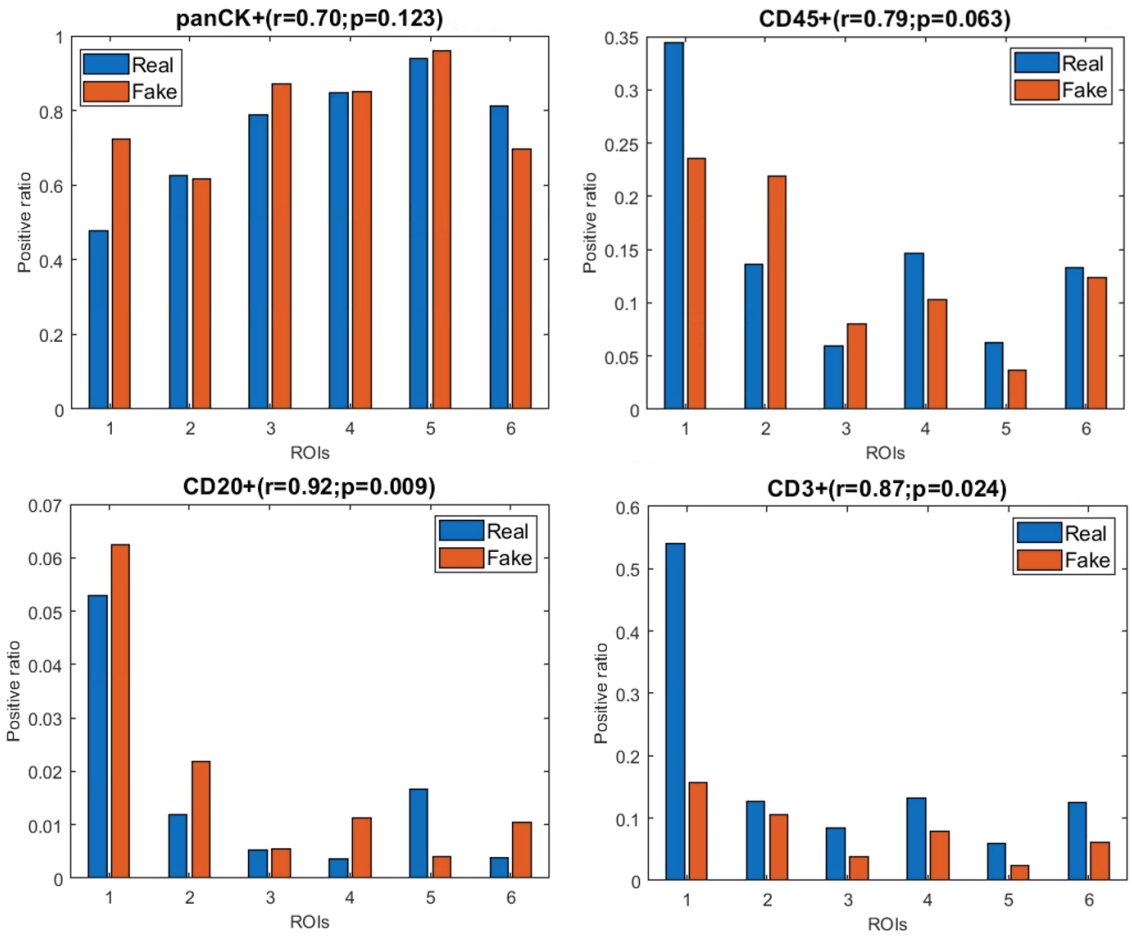


FIGURE 3.8: ROI cell composition correlation between real and virtual CyCIF. For each of the ROIs shown in Figure 3.7, the positive ratio of cells for each of panCK, CD45, CD20, and CD3 are calculated using the same workflow and displayed for either real or virtual CyCIF WSIs. Pearson's correlations and and p-values describing the association between positive ratios derived from real and virtual CyCIF WSIs for each marker are indicated above each bar plot. Adapted from a slide by Jia-Ren Lin, Co-Director of the Tissue Imaging Platform at the Labratory of Systems Pharmacology, Harvard Medical School.

performed a simple ablation experiment with the CyCIF encoder of the model removed (Figure 3.11A). For this experiment, the model was tasked with H&E reconstruction and H&E-to-(DAPI and panCK) translation. To assess goodness of fit, the model was trained to convergence and evaluated on a training batch (Figure 3.11B). Visual inspection of model outputs indicated that the model was functioning as intended (Figure 3.11B).

In our original design, the XAE included skip connections that connected across the

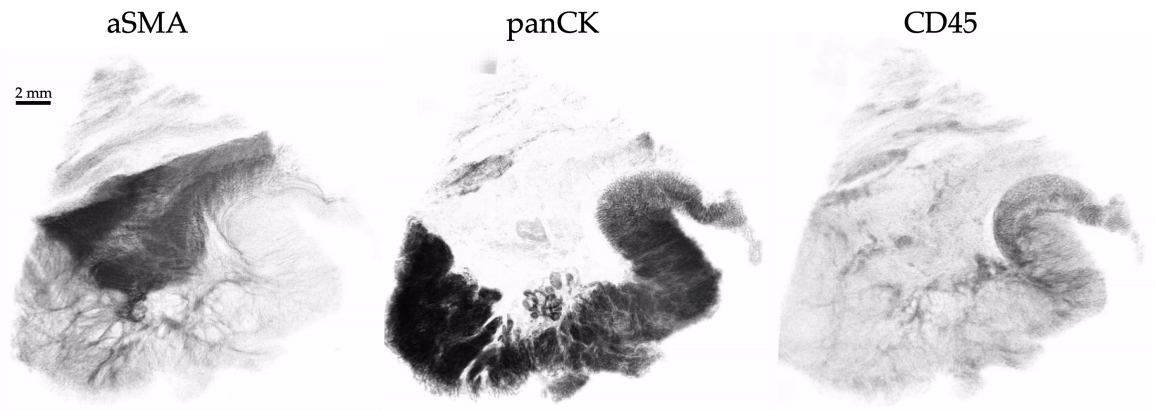


FIGURE 3.9: 3D virtual stain volumes conditioned on held-out H&E test sections.

U-Net generator blocks, but we discovered that the models did not learn useful latent representations of images (Figure 3.12), a direct effect of the absence of loss function gradient flow through the interior layers of the models enable by skip connections (Figure 3.12B). We removed the skip connections in subsequent experiments and found that these models exhibit good convergence properties (Figure 3.13A) and have appreciable loss function gradient flow through the model interior (Figure 3.13B).

Having confirmed that the trained XAE had fit its training distribution (Figure 3.14), we next wanted to assess the representativeness and interpretability of the latent feature space that it learned with respect to pathologically interesting regions of the sample. To do this, we used the H&E encoder of the trained XAE to encode tiles from H&E test section 096 into 512-dimension feature representations and assessed how the features were distributed over tiles drawn from each of several pathologist-defined ROIs in the test section. We found that many of the learned image features were associated with pathologically-distinct regions of the sample (Figure 3.15).

To quantify the utility of these representations in terms of their ability to discriminate between pathological features, we used the representations to train a random forest classification model to predict which ROI tiles were drawn from based on the XAE representations (Figure 3.16A). The median Matthews correlation coefficient (MCC) and

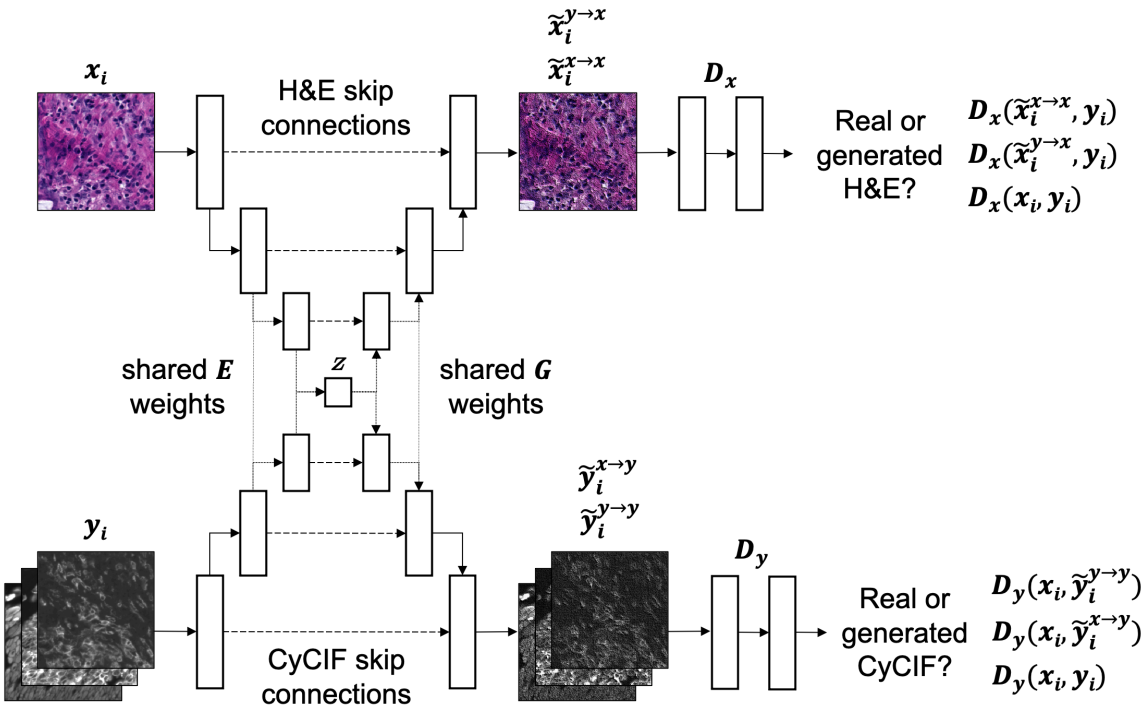


FIGURE 3.10: Overview of XAE architecture for H&E and CyCIF channel co-embedding. The model has two input heads, one for H&E encoder inputs (x_i) and another for CyCIF encoder inputs (y_i), both of which encode into a shared latent space (z). The model also has two output heads, one for H&E decoder outputs ($\tilde{x}^{x\to x}$ if $z_i = E_{\text{H\&E}}(x_i)$ or $\tilde{x}^{y\to x}$ if $z_i = E_{\text{CyCIF}}(y_i)$) and another for CyCIF decoder outputs ($\tilde{y}^{y\to y}$ if $z_i = E_{\text{CyCIF}}(y_i)$ or $\tilde{y}^{x\to y}$ if $z_i = E_{\text{H\&E}}(x_i)$). The weights of the last (first) layer of the encoders E (decoders D) are shared. In this weight-sharing design, we assume for any given pair of spatially-registered pair x_i and y_i, there exists a shared latent code z_i in the shared latent space z such that we can recover either that we can recover either x_i or y_i from z_i, and we can compute z_i from either x_i or y_i [63]. Discriminators D_x and D_y are trained to tell the difference between real and virtual H&E and CyCIF images, respectively. Discriminators are also provided the real image of the opposite domain via a channel-wise concatenation with in in-domain image (e.g. D_x always sees y_i), a form of discriminator conditioning which improves the visual quality of virtual outputs [48]. Optionally, skip connections can be added to confer a U-Net-like design to the XAE [95], but we found that this produced models which were incapable of learning an adequately structured latent space. Full XAE model architecture is described in Table 3.2

weighted area under the receiver operating characteristic curve (AUC) following a 5-fold cross-validation were 0.72 and 0.96, respectively, indicating the utility of XAE features for downstream tasks.

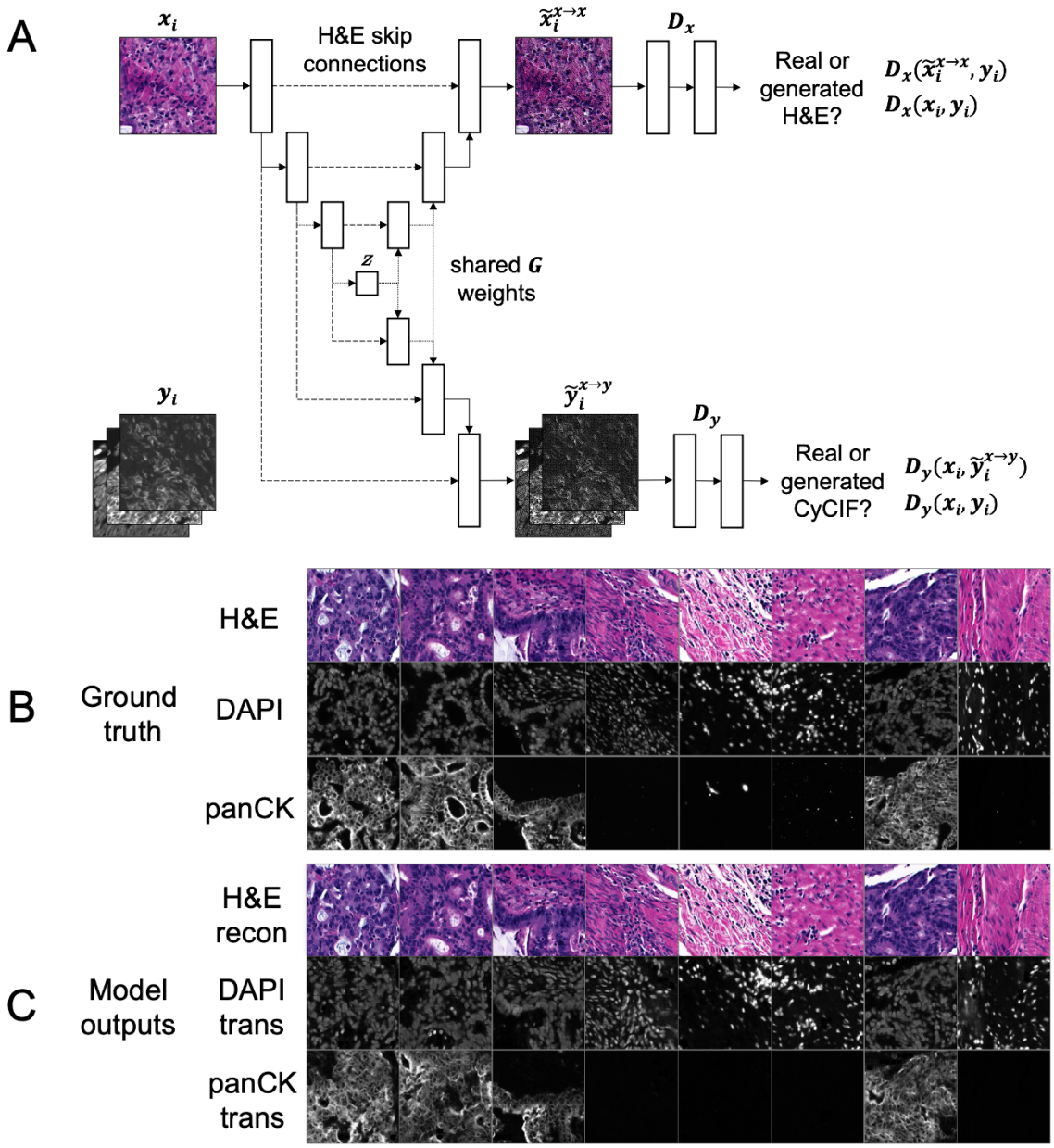


FIGURE 3.11: Model ablation validation to demonstrate a minimal working example and proof of concept for the XAE. (A) XAE architecture with CyCIF encoder removed. The only tasks learned are H&E reconstruction and H&E-to-CyCIF translation. (B) Ground truth tiles representing a single training batch. (C) Trained XAE model results for the tasks of H&E-to-H&E reconstruction (recon) and H&E-to-CyCIF translation (trans) using the ground truth training batch from (B).

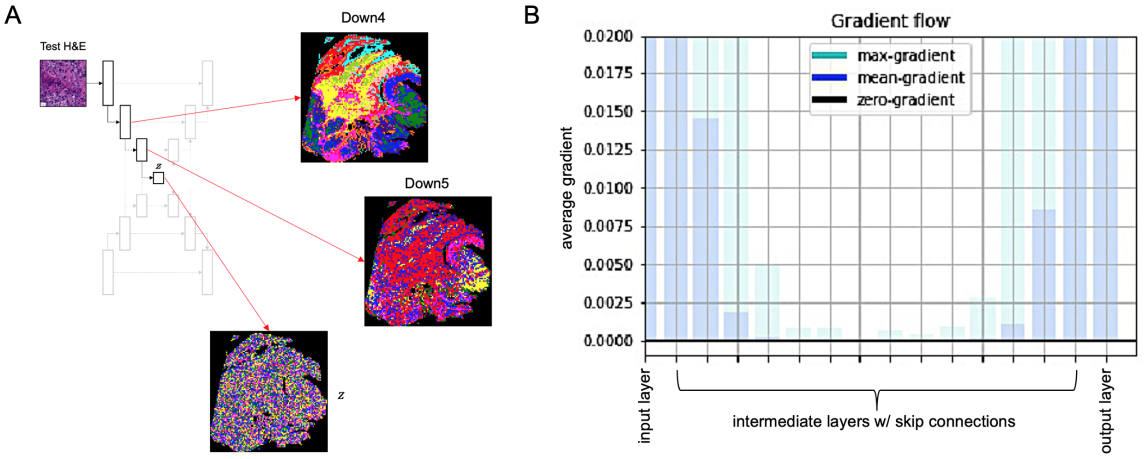


FIGURE 3.12: XAE with skip connections (XAE-SC) fails to learn representative latent space. (A) Assessing the quality of intermediate and latent representations in a XAE-SC model. An XAE-SC model was trained to convergence, then each test H&E tile was passed to the trained model. Feature maps at intermediate and latent layers of the downward pass for each tile were max-pooled along the channel dimension, then these representations were clustered using PhenoGraph [58] for each layer independently. Each tile of the test WSI is colored based on its cluster label from the respective layer in the three inset images. Deeper layers have diminished semantic meaning, with latent space z encoding nothing more than random noise. (B) Interrogation of the average loss function gradient flow over XAE-SC layers showed that layers in the interior of the model were not learning during training.

3.3.3 XAE captures unseen biologically relevant information from H&E images

In order to evaluate how well deep learning can capture and represent unseen complex information using H&E images alone, VAE and XAE features were compared to cell types defined by CyCIF expressions and pathologist tissue annotations. Clustering tiles within the whole slide image based on cell type composition resulted in 7 clusters, and the pathologist annotated 5 key tissue types to be used as ground truth (Figure 3.17A). Ground truth tile labels were compared against one another to create a baseline for evaluation (Figure 3.17B). When annotations were used to predict cell type, there was a baseline performance of 57.1 cluster purity and 0.44 NMI. Conversely when cell type was used to predict annotations, there was a baseline performance of 66.8 cluster purity and 0.44 NMI. In all metrics, XAE outperformed VAE predictions, achieving a 56.1 cluster purity and 0.35 NMI against cell

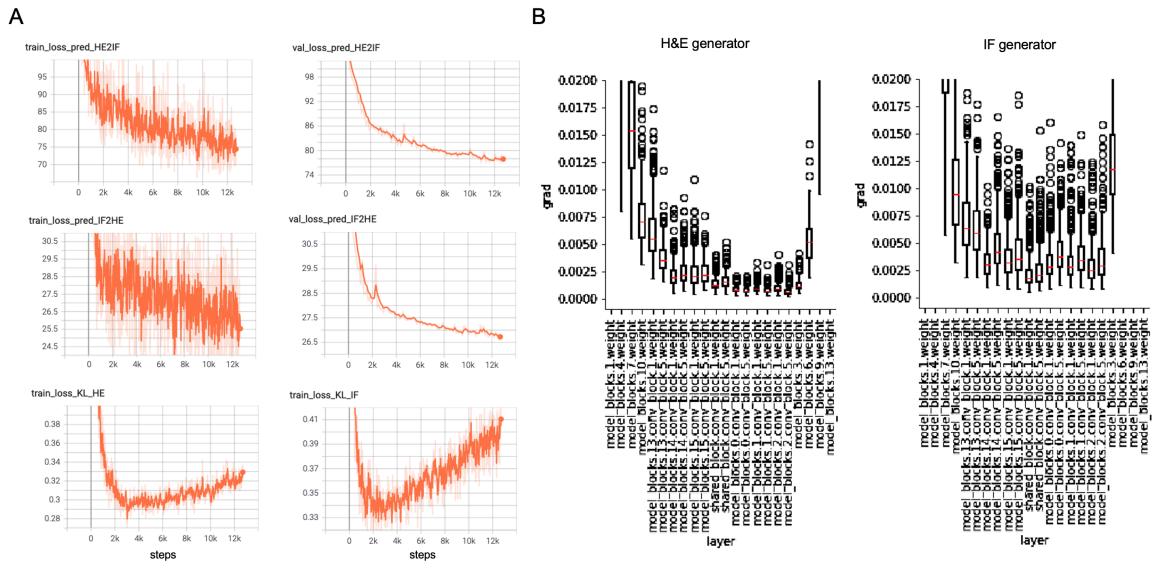


FIGURE 3.13: XAE training dynamics improve without skip connections. The XAE model was parameterized to have a latent space z channel dimension of 512. (A) Training (left) and validation (right) loss dynamics move toward convergence and suggest a trade-off between virtual image quality (pred_HE2IF and pred_IF2HE are the L1 losses for the respective image translation tasks) and the normal prior constraint on the latent space (KL for Kullback-Leibler divergence). (B) Interrogation of the average loss function gradient flow over layers for the H&E encoder and decoder paths (left) and CyCIF encoder and decoder paths (right) of the XAE model indicate that gradient flow through interior layers can be recovered by removing skip connections.

type, and 70.2 cluster purity and 0.38 NMI against pathologist annotation. It is also notable that on the metric of cluster purity against annotations, the XAE outperformed the baseline metric; this indicates that the XAE is better at predicting tissue type than even cell type compositions.

Analysis of complex information, deeper than large scale clustering, was conducted using canonical correlations between the model embedding space and the tile-wise CyCIF expressions. Visually both VAE and XAE show a good overlap between cell type embeddings from CyCIF and model embeddings produced from HE images (Figure 3.17C); the XAE, however, achieves higher canonical correlations (0.93 and 0.92 compared to 0.91 and 0.88 for VAE). Using these embedding spaces we can plot the density of ground truth clusters and their respective correlates for each architecture. In both modalities, we can see that all clusters are adequately covered by the predictions with no populations only captured

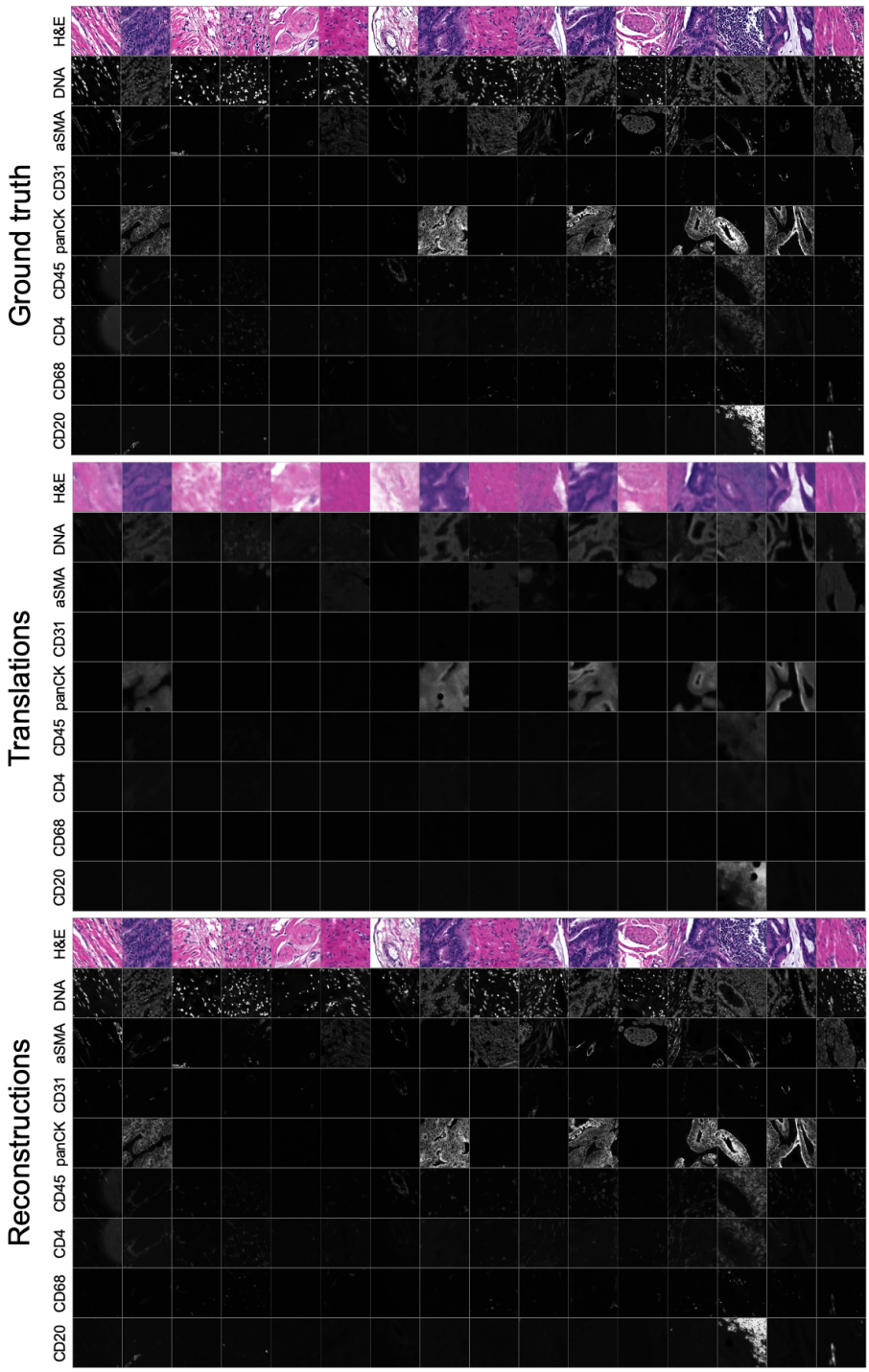


FIGURE 3.14: XAE tile-level training results. An XAE model was trained to convergence (same model as in Figure 3.13) using the indicated CyCIF image channels then evaluated on a training tile batch. Ground truth tiles are shown at top, cross-domain translation tiles are shown in middle, and in-domain reconstruction tiles are shown at bottom.

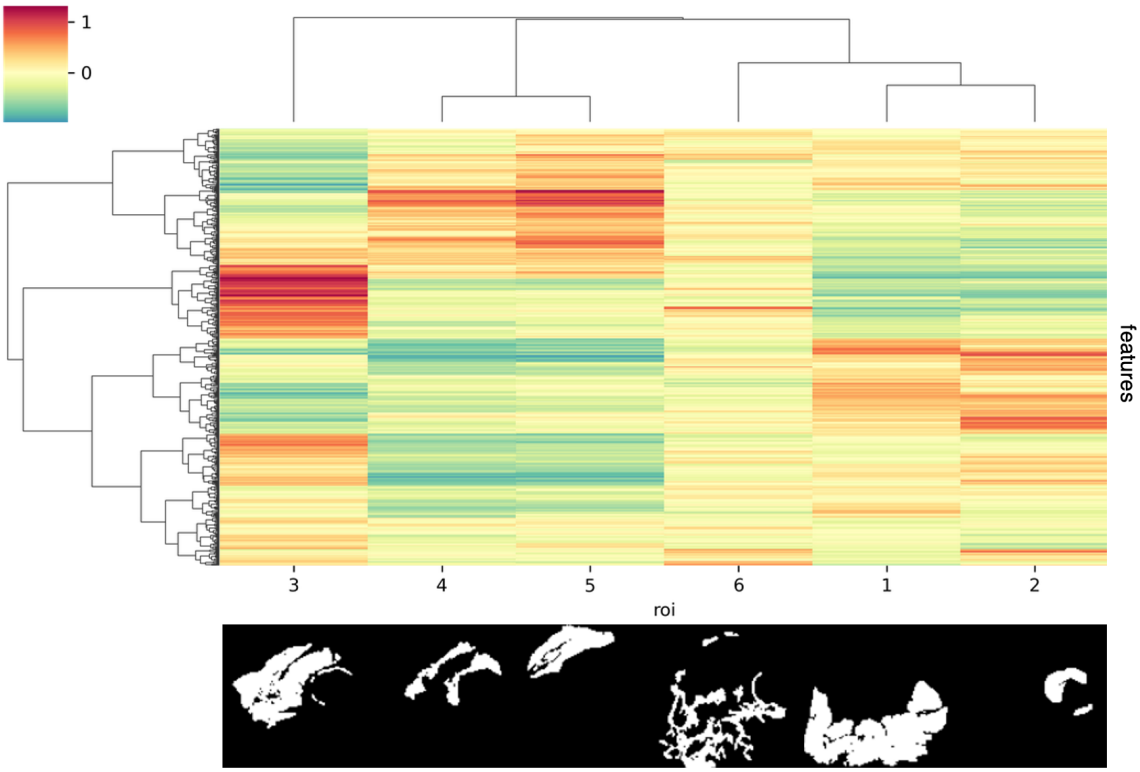


FIGURE 3.15: XAE latent feature clustering. Using the trained XAE described in Figure 3.13 and Figure 3.14, the 6742 non-overlapping tiles from H&E test section 096 which had at least one pixel of pathologist annotation were each encoded into 512-dimension latent feature maps, which were then each max-pooled along the channel dimension such that tiles were encoded by feature vectors of length 512. Features were z-scored, then tiles were mean-aggregated based on their ROI and features were hierarchically clustered. The ROI label keys are 1: tumor adenocarcinoma ($n = 2501$ tiles); 2: normal mucosa ($n = 362$ tiles); 3: proper muscle ($n = 1576$ tiles); 4: submucosa ($n = 473$ tiles); 5: subserosa, loose connective tissue ($n = 782$ tiles); and 6: fibrosis, inflammation, lymphoid aggregate ($n = 1048$ tiles). The color scale corresponds to the mean of z-scored feature values for each ROI. The inset image indicates the binary mask corresponding to each ROI with respect to the layout of the H&E test section 096.

by ground truth.

To confirm that we were extracting relevant and rare cell types with the representation models, we computed the Spearman correlation between every predicted cluster and ground truth cluster (Figure 3.17D). From this we can see that XAE has consistently higher magnitudes of correlation compared to VAE clusters, and that a reasonable correlate exists

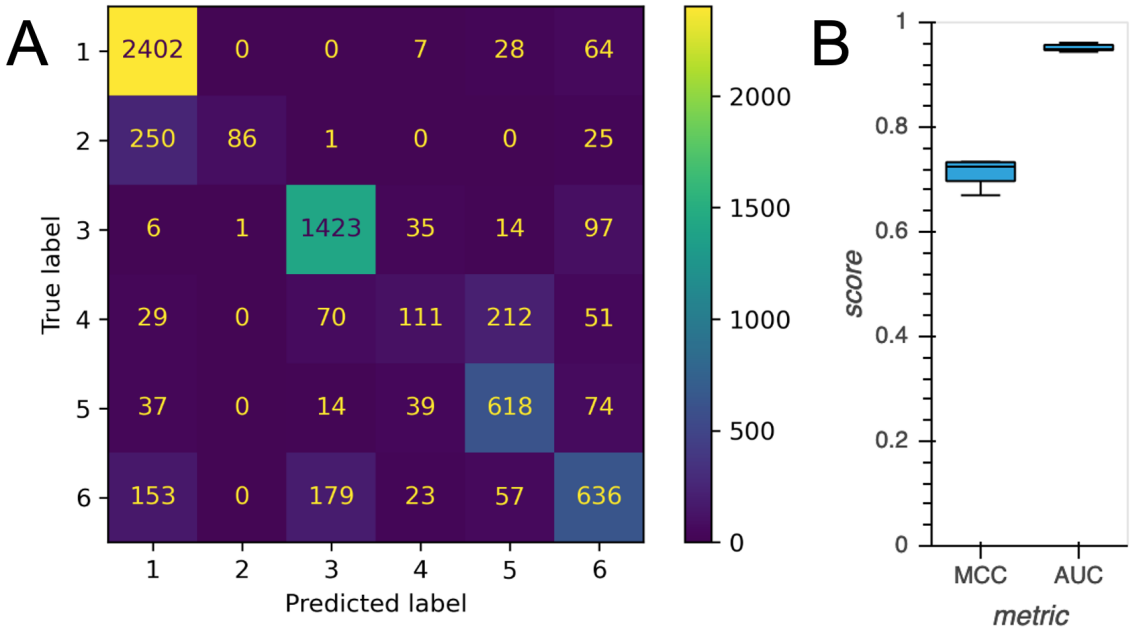
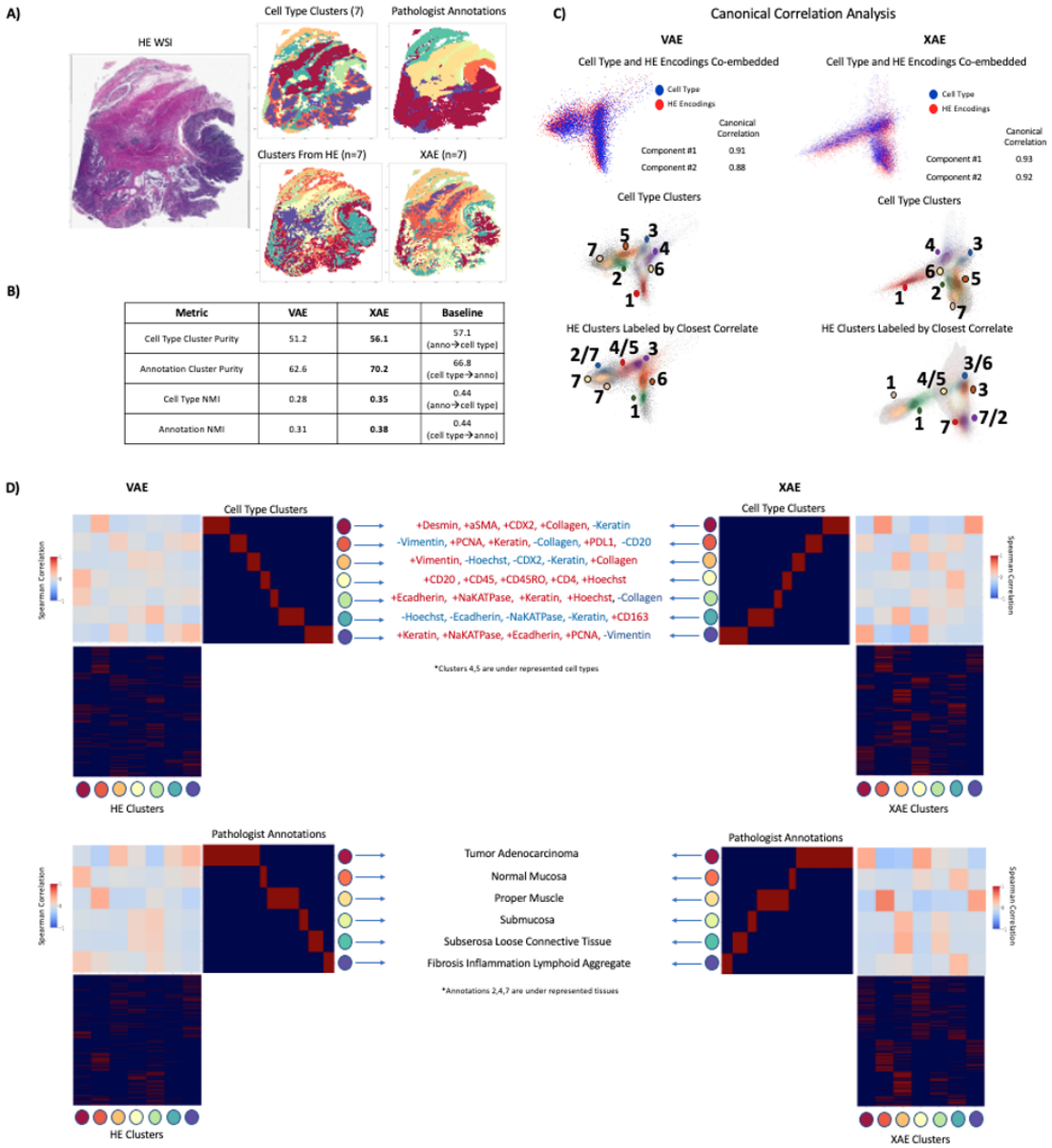


FIGURE 3.16: Random forest classification of histopathological ROIs based on XAE image features. (A) Confusion matrix resulting from a 5-fold cross validation of random forest ROI classification over all tiles from H&E test section 096. Value in cell i, j represents the number of tiles from ROI i which were predicted as being drawn from ROI j. True positive predictions are along the diagonal. (B) Boxplot displaying the distribution of Matthew's correlation coefficients (MCC) and the weighted area under the curve receiver operating characteristic curve (AUC) over the 5 folds of cross validation.

for every ground truth cluster except for cell type clusters 4 and 5 which are underrepresented populations. By comparison VAE on has strong correlates to a few abundant cell types. Furthermore, the cell types that the XAE is able to capture are largely explained by changes in Na-K ATPase, E-Cadherin, and PCNA, which were shown to be important indicators for cell phenotypes in prior research on this tissue [61].

It is shown by numerous metrics that the XAE model outperforms the VAE in capturing detailed information from H&E images alone, which are able to adequately recapitulate unseen information from CyCIF expression data and pathologist annotations. Because the XAE encodings are able to adequately recapitulate the information in CyCIF from H&E, we can use them for proxy analyses such as selecting representative regions of the WSI for further analysis.



Metric	VAE	XAE	Baseline
Cell Type Cluster Purity	51.2	**56.1**	57.1 (anno→cell type)
Annotation Cluster Purity	62.6	**70.2**	66.8 (cell type→anno)
Cell Type NMI	0.28	**0.35**	0.44 (anno→cell type)
Annotation NMI	0.31	**0.38**	0.44 (cell type→anno)

FIGURE 3.17: Deep learning architectures recapitulate unseen complex information using HE. (A) Annotations of ROIs as defined by CyCIF cell type clustering, pathologist annotation, H&E-only VAE feature clustering, and H&E/CyCIF-combined XAE feature clustering. (B) Evaluation metrics of clustering quality for the VAE and XAE clusters against pathologist annotations or cell type clustering. (C) Canonical correlation analysis results for comparison between the model embedding space and the tile-wise CyCIF expressions for both the H&E-only VAE model and the H&E/CyCIF-combined XAE model. (D) Spearman correlations between ground truth clusters and model-derived clusters.

3.3.4 Co-embedding H&E and IF representations improves ROI selection

Currently ROI selection within WSI images is done either randomly, which is inaccurate and is likely to select an area that doesn't represent the WSI, or with manual selection of ROI, which is biased, un-quantitative, and has been shown to miss whole tissue patterns [61]. Using the XAE representations, which capture the complex cell type and annotation information using H&E, we can construct a quantitative methodology to select ROIs that are more representative than random sampling while being repeatable and biologically driven. To measure this, we use three metrics: mean squared error (MSE) between the cell type composition of selected ROIs and the WSI for a discrete interpretation of composition difference; Jensen-Shannon Divergence (JSD) between the cell type composition vectors of selected ROIs and the WSI for a probabilistic interpretation of composition difference; and mean entropy of the selected ROIs' cell type compositions to assess the cellular heterogeneity of a given ROI selection. Four methods for ROI selection were tested: random sampling, linear optimization to match cell composition only versus entropy, convex optimization minimizing MSE and maximizing entropy, and a genetic optimization algorithm.

When regions are randomly sampled, we see that the cell type compositions struggle to converge to the whole slide cell type composition, taking upwards of 20-30 ROIs before reaching a reasonable representation (Figure 3.18). Adding a simple linear optimization to select ROIs drastically decreases the number of ROIs necessary to around 7. This number of ROIs is equivalent to the number of cell type clusters for which we were optimizing, which is indicative that the algorithm was selecting primarily homogeneous regions that reconstruct the whole slide composition. This is validated looking at the mean entropy of ROIs for the base linear optimization method, which shows very low mean ROI entropy in the 1000 pixel size ROIs and middling mean ROI entropy in the 2500 pixel size ROIs. When entropy is considered in the convex optimization, we see convergence much earlier at 3-4 representative ROIs. Unlike the simple linear optimization, however, the ROIs selected are

not homogenous and include much more biologically interesting regions with diverse cell populations.

Although cell type composition and entropy were used as metrics of biological relevance in this setting, it is likely that other experiments would have different priorities. Some examples of this might include: weighting cell type clusters by level of interest; weighting entropy negatively if homogeneous regions are desired; weighting some other extracted scores such as co-localization of two cell types of interest. The method of optimization is versatile and amenable to many different functions. The key takeaway is that this pipeline allows for intelligent representation from H&E images, which enables a plethora of subsequent analyses on this representation space.

3.4 Discussion

The advance of MTI platforms like CyCIF promises to increase our understanding of heterogeneity and cellular interactions within the tumor microenvironment, both of which play increasingly important roles in the development of effective treatments [130]. Although its clinical potential is immense, CyCIF is time- and labor-intensive, technically complicated, and high-cost, so assessment is typically limited to only a small subset of a given biopsy, which is unlikely to be fully representative of a patient's disease. Also, the cost associated with MTI will undoubtedly limit its use to within highly-developed clinical settings for the foreseeable future, further widening the quality-of-care gap between high- and low-income communities. Until MTI matures into an economy of scale, these challenges will only be further amplified in 3D applications. The technological gap between standard histology and MTI technologies highlights the broader need for automated tools that leverage information attained by a low-cost technique to infer information typically attained by a high-cost technique.

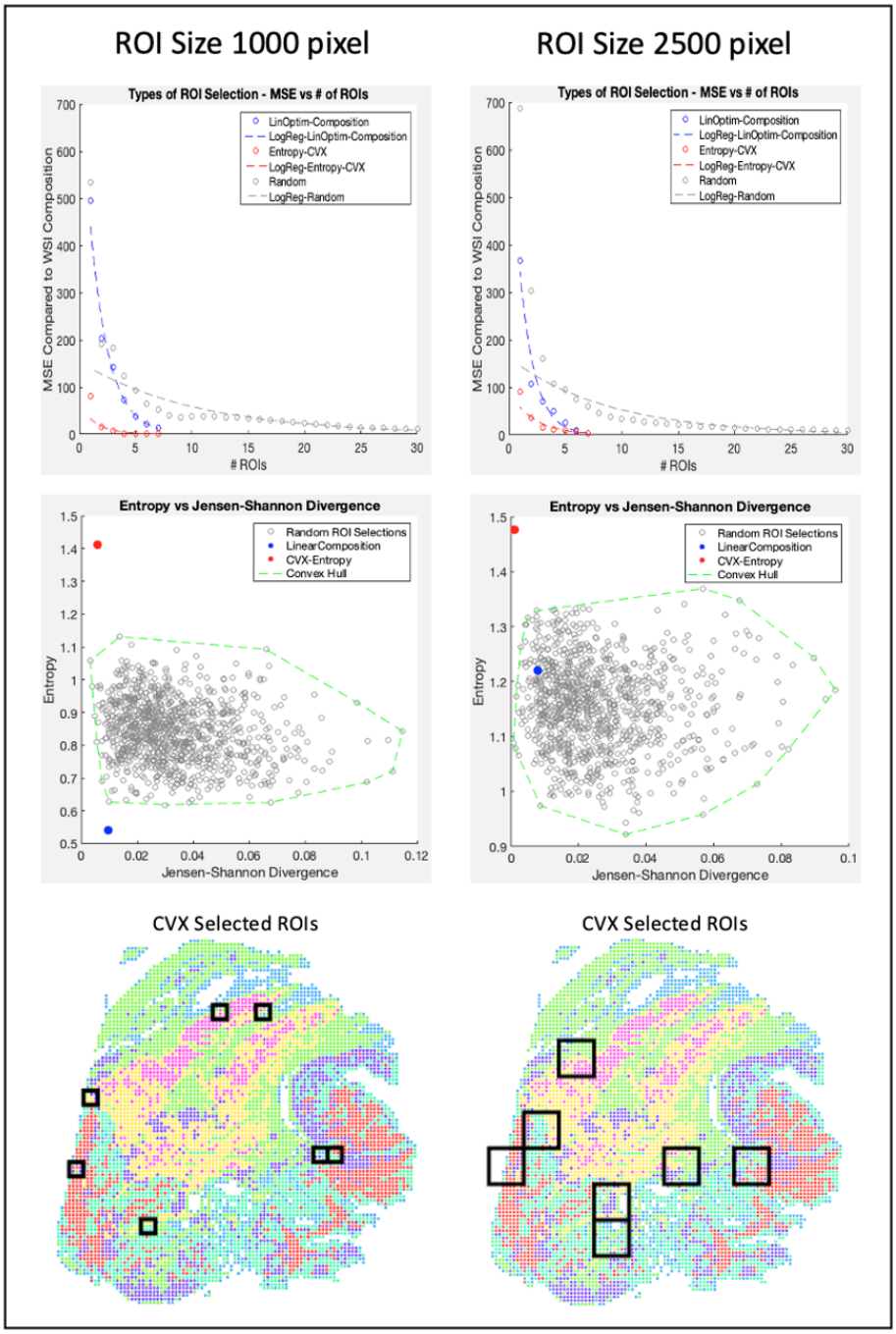


FIGURE 3.18: Optimization of ROI selection. ROI selections are optimized based on either the MSE between the cell type compositions of the ROIs and the WSI (top row), or both the JSD between the cell type compositions of the ROIs and the WSI and the feature-level entropy of the ROI selection (middle row). Two different pixel sizes are used for the ROIs, to highlight the trade-offs for scaling up or down.of different pixel sizes for ROIs. The bottom row shows an example of CVX selected ROIs for the the two ROI sizes.

To help address some of these challenges, in the present study we extend a virtual staining paradigm to a 3D CRC atlas [61] and demonstrate that generative models can learn from a minimal subset of the atlas to reconstruct the remaining sections of the CyCIF portion of the atlas and recapitulate quantitative endpoints derived using the real CyCIF data. We also implement and evaluate a novel deep learning architecture which integrates paired H&E and CyCIF data into a shared representation and demonstrate that the model can be used as a quantitative and objective guide for ROI selection, with the integrated H&E/CyCIF representations being more informative than H&E representations alone.

One of the takeaways from qualitative comparison of real and virtual CyCIF stains was identification of a discrepancy between the real and virtual CD45 stains of an immune aggregate in section 001. In particular, the virtual CD45 model underestimated the real intensity of the aggregated cells. We attribute this to the fact that virtual stains are standardized to their training sets, and because the same immune aggregate in the training section 054 did not match the outlying high intensity CD45 distribution of 001, the model had not learned to match it. This sort of discrepancy can either be a feature of the virtual stain, in that in standardized the section-to-section technical variability of the IF stain, or a bug in our training procedure, in that our training set did not capture an important biological feature present in the held-out test set. However, by looking at the real CD45 image from next nearest section to section 001, about 30 µm away, we see that the immune aggregate does not exhibit outlying high intensity, which suggests that the assumption made by the model may have been a good one. If this were not the case, such a bug could be corrected through trial and error on a larger or more representative training set with guidance from a pathologist, in a process similar to the design and validation of a real stain.

Quantitative comparisons of real and virtual CyCIF stains exposed the challenge of using adjacent sections to train models, where image contents are subtly but appreciably different between sections at single-cell resolution. This challenge could be overcome in future studies by staining each tissue section first with CyCIF then terminally with H&E

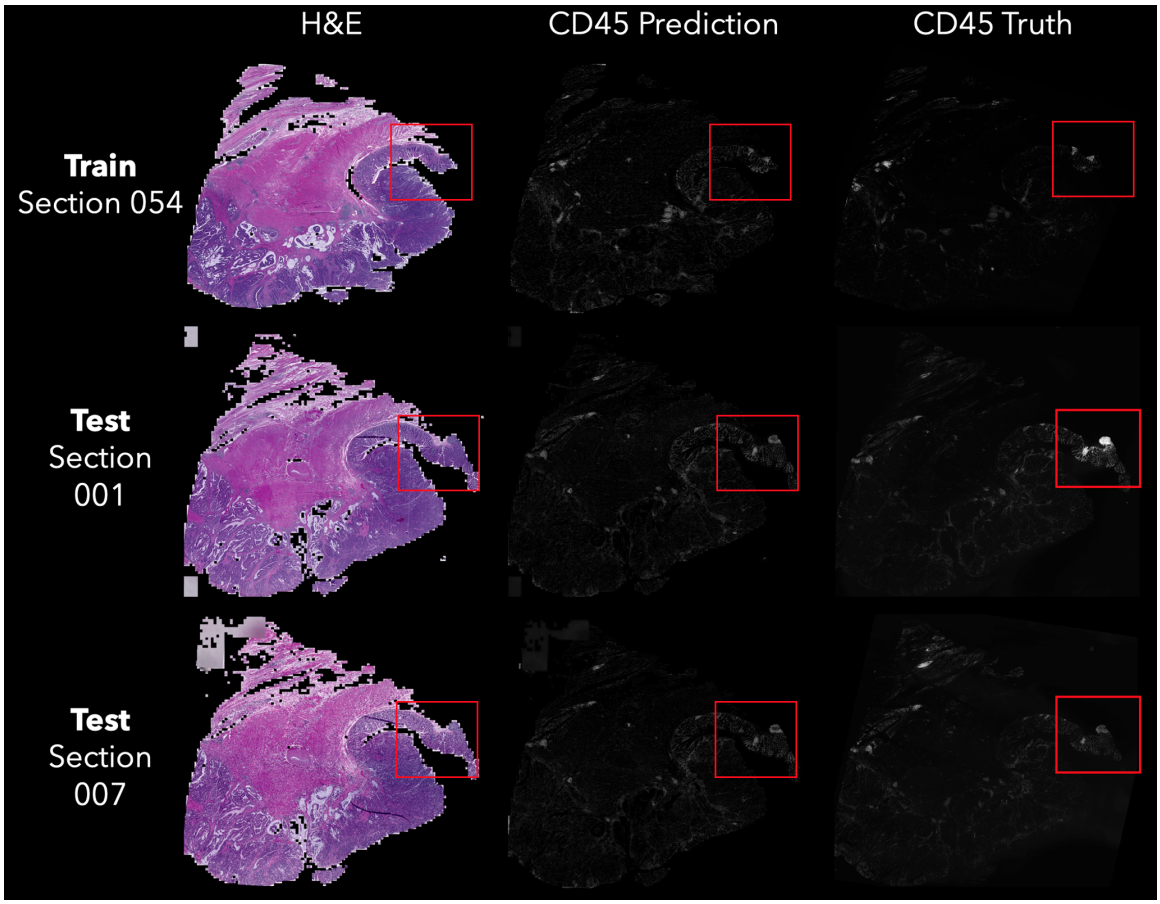


FIGURE 3.19: Inconsistency in ground truth CD45 stain intensity between proximal sections is attenuated by virtual staining.

[116], which would ensure that the image contents are exactly the same between H&E and CyCIF images and minimize the difficulty of image co-registration. This study takes for granted that histology is an inherently destructive procedure. Studies which require serial sectioning and processing of tissue can preclude tissue from being used in other assays. Alternatively, a non-destructive 3D microscopy approach using tissue clearing and light-sheet microscopy could be deployed, which would also preserve tissues for other assays [62]. However, the slow diffusion rate of antibodies in whole tissues limits limit the deep multiplexing potential of the CyCIF platform in this non-destructive approach, but the use of small molecule dyes and affinity agents could help to overcome this challenge and

constitutes one of the most promising avenues forward for 3D virtual staining applications [128].

3.5 Methods

3.5.1 H&E stain normalization

To minimize the influence of technical variability on stain color between H&E sections, we experimented with the application of several stain normalization methods to the H&E WSIs [89, 71, 117] using the Python package staintools (https://github.com/Peter554/StainTools). To identify and mask out background regions of each WSI (white regions of slide without tissue), WSIs were each cropped into non-overlapping 256 × 256-pixel tiles and tiles containing greater than 70% area of pixels with 8-bit intensity greater than (210, 210, 210) were excluded from subsequent normalization steps. To help identify and mask out background pixels in the remaining tiles before model fitting and normalization, the foreground tiles from each H&E WSI were independently standardized such that 5% of all pixels were luminosity saturated. For all normalization methods, we used the H&E WSI from section 054 as the stain reference to which the stain intensity distributions of all other H&E WSIs would be fit. After normalizing the foreground tiles of each non-reference WSI to fit the reference stain distribution, tiles were restitched to form cohesive WSIs. On the basis of visual inspection (Figure 3.2), we opted to use the Reinhard normalization method, which has also been shown to maximize deep learning model performance on digital pathology applications [110].

3.5.2 CyCIF image preprocessing for SHIFT and XAE modeling

To control for variations in raw contrast between CyCIF WSIs, we rescaled the intensities of CyCIF WSIs to have a min-max range fit to the 70th-99.99th intensity percentiles of the

input WSIs. Intensity percentiles were empirically chosen based on their exclusion of low- and high-intensity artifacts.

3.5.3 SHIFT models

SHIFT models were built using PyTorch as previously described [116]. Model architectures are described in Table 3.1. Models were trained to predict single channel images corresponding to one of the CyCIF stains from input H&E tiles from section 054, e.g. H&E $\rightarrow$ CD45 or H&E $\rightarrow$ CD31. Paired H&E and CyCIF image tiles from section 054 were split into 80% training (8134 tiles) and 20% validation (2034 tiles) sets and each model was trained with a batch size of 4 and learning rate of 0.0002 for 100 epochs. Best models were selected based on the lowest validation loss at each epoch end and were then used for downstream application to held-out H&E WSIs.

3.5.4 XAE models

XAE models were built using PyTorch. Model architectures are described in Table 3.2. The XAE architecture used here is an adaptation of the UNIT architecture [63] and the imaging-to-omics XAE architecture [102]. XAE models have two input encoders (FIGURE XXX), one accepting H&E image tiles (batch size $\times 3 \times 256 \times 256$), and the other accepting the corresponding paired CyCIF images (batch size $\times$ N CyCIF channels $\times$ 256 $\times$ 256). Both encoders compress their inputs into a shared latent space z. From z, image representations can be upscaled by either H&E or CyCIF decoders. Hence, there are four forward paths through the model: (1) H&E reconstruction: H&E $\rightarrow z \rightarrow$ H&E; (2) H&E-to-CyCIF translation: H&E $\rightarrow z \rightarrow$ CyCIF; (3) CyCIF reconstruction: CyCIF $\rightarrow z \rightarrow$ CyCIF; and (4) CyCIF-to-H&E translation: CyCIF $\rightarrow z \rightarrow$ H&E. Models were trained with a batch size of 16 and a learning rate of 0.0001 for 100 epochs. Best models were selected based on

the lowest validation loss at each epoch end and were then used for downstream application to held-out H&E WSIs. We also experimented with a U-Net-like architecture with skip connections between encoder and decoders [95], but found that loss gradients did not propagate to the most internal layers of these models such that meaningful latent representations were not learned.

3.5.5 Comparing VAE vs XAE tile-based representations

Tile cluster identity

Ultimately, we want to evaluate whether deep learning architectures can recapitulate the biological information of both cell type and pathologist, but since VAEs and XAEs operate on a tile by tiles basis, it is necessary to cluster tiles based on their cell type composition. For every tile in the WSI, a vector was created that represented the composition of cell types. The ground truth cell type information was made by KMeans clustering these composition vectors (Figure 3.17A). Using the elbow method, we determined that 7 clusters was the optimum for evaluation. A smaller number of clusters within the elbow was chosen to better match the number of pathologist annotations for consistency in evaluation. Pathologist information was created manually by an expert pathologist, resulting in 5 distinct tissue types (Figure 3.17A). Tiles were assigned a ground truth tissue type based on the maximum pixel-wise tissue type within the region. 7 clusters were computed for both the standard VAE and the XAE encoding vectors to evaluate against the cell type ground truth clusters.

Several metrics were used to evaluate different aspects of the ground truth recapitulation. Cluster purity was used to evaluate how well the two methodologies were able to reconstruct the same clusters as ground truth:

$$Purity = \frac{1}{N}\sum_{i=1}^{k}\max(c_i \cap t_j) \tag{3.1}$$

where N is the number of datapoints, k is the number of clusters, c is the set of predicted clusters and t is the set of ground truth clusters. The sklearn [86] implementation of Normalized Mutual Information (NMI) was used as another metric to evaluate the same question:

$$NMI(U, V) = \frac{MI(U, V)}{mean(H(U), H(V))} \tag{3.2}$$

where U and V are the predicted and ground truth cluster labels. The predicted tile-type clusters were paired to ground truth cell-type clusters and annotations using the Spearman correlation.

To evaluate whether the deep learning models capture the same level of feature information as CyCIF staining, we used the pyrcca [9] implementation of canonical correlation on the encoded latent feature space and the paired CyCIF tilewise expressions. The outputs from this process produced two components shared between the two modalities. Quantitatively the correspondence of the two modalities can be measured by the canonical correlation of each component, and qualitatively the correspondence can be observed by the overlap in the scatter plot of the new components.

3.5.6 ROI sampling

Random sampling

Random sampling was conducted by randomly drawing a new non-overlapping ROI repeatedly. For bulk analysis and comparison, 2000 random combinations of k ROIs were selected where k is the number of ROIs found to be optimal for the other sampling methods.

Linear optimization on composition and entropy

If b represents counts of cells across clustered group and a_i represents the cell number belong to the i-th ROI, by solving $\min_x ||x||_1$ s.t. $b = Ax$ we could identify the minimum

number of ROIs to match WSI cellular population. The main issue of this approach is that it often selects ROIs with homogeneous cell populations. Since we do not have cell composition beforehand, we will use cluster results based on latent representation of tiles within ROIs via embedding both H&E and CyCIF. Underlying assumption here is that H&E/CyCIF embedding reflects tile-based cell composition.

For optimization on cluster composition, $b = Ax$ where $b \in \mathbb{R}^N$, b represents the composition vector of clustered groups within the WSI, each column of $A \in \mathbb{R}^{N \times M}$ represents a possible ROI and each row contains the percentage of tiles in that ROI for each cluster; N, M represent the number of clusters and the number of possible ROIs in the WSI respectively. Then, we solve the optimization problem: $\min_x ||x||_1$ s.t. $b = Ax$ and s.t. $0 \leq x \leq 1$. Implementation of this function was conducted using the intlinprog function in MATLAB.

Linear optimization on composition

To optimize both composition and ROI heterogeneity, we take the entropy of the composition vector into account using the convex optimization function: $\min(norm(Ax - b) - \lambda Ex)$ s.t. $b = Ax$, where $E \in \mathbb{R}^M$ represents the vector of entropies and λ is a hyperparameter governing the weight of entropy. We solve the optimization such that $0 \leq x \leq 1$ and $\sum x = 1$. Implementation of this function was conducted using CVX in MATLAB.

Evaluation

The quality of the selected representative ROIs was evaluated based on three metrics: Mean squared error (MSE) compared to WSI composition; Jensen-Shannon Divergence (JSD) of the ROI and WSI compositions; and mean ROI entropy. Mean squared error was calculated using:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (R_i - W_i)^2 \tag{3.3}$$

where n is the number of predicted clusters, R is the percent composition of each cluster within all selected ROIs combined, and W is the percent composition of each cluster within the WSI. JSD was calculated using:

$$JSD = \frac{1}{2}\sum_{i=1}^{n} R_i log_2\left(\frac{R_i}{\frac{1}{2}(R_i + Wi)}\right) + \frac{1}{2}\sum_{i=1}^{n} W_i log_2\left(\frac{W_i}{\frac{1}{2}(R_i + Wi)}\right) \tag{3.4}$$

where n is the number of predicted clusters, R is the percent composition of each cluster within all selected ROIs combined, and W is the percent composition of each cluster within the WSI. The mean entropy was calculated using:

$$\text{mean entropy} = \frac{1}{m}\sum_{i=1}^{m}\sum r_i log(r_i) \tag{3.5}$$

where m is the number of selected ROIs and r is the percent composition within each individual ROI.

Layer	Generator
D1	Conv2d(3, 64, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) LeakyReLU(negative_slope=0.2, inplace=True)
D2	Conv2d(64, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D3	Conv2d(128, 256, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D4	Conv2d(256, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D5	Conv2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D6	Conv2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D7	Conv2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D8	Conv2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) ReLU(inplace=True)
U1	ConvTranspose2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U2	ConvTranspose2d(1024, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U3	ConvTranspose2d(1024, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U4	ConvTranspose2d(1024, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U5	ConvTranspose2d(1024, 256, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U6	ConvTranspose2d(512, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U7	ConvTranspose2d(256, 64, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U8	ConvTranspose2d(128, 1, kernel_size=(4,4), stride=(2,2), padding=(1,1)) Tanh()

Layer	Discriminator
1	Conv2d(4, 64, kernel_size=(4,4), stride=(2,2), padding=(1,1)) LeakyReLU(negative_slope=0.2, inplace=True)
2	Conv2d(64, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
3	Conv2d(128, 256, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
4	Conv2d(256, 512, kernel_size=(4,4), stride=(1,1), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
5	Conv2d(512, 1, kernel_size=(4,4), stride=(1,1), padding=(1,1))

TABLE 3.1: SHIFT model architecture. Layers are represented in PyTorch pseudocode. For the layer column, D and U represent down- and up-sampling layers of the U-Net architecture [95], respectively.

Layer	Encoders	Shared?
1	ReflectionPad2d((3, 3, 3, 3)) Conv2d(3, 64, kernel_size=(7,7), stride=(1,1)) InstanceNorm2d(64, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
2	Conv2d(64, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1)) InstanceNorm2d(128, eps=1e-05, momentum=0.1, affine=False) ReLU(inplace=True)	No
3	Conv2d(128, 256, kernel_size=(4,4), stride=(2,2), padding=(1,1)) InstanceNorm2d(256, eps=1e-05, momentum=0.1, affine=False) ReLU(inplace=True)	No
4	ResBlock(N=256 ,K=3, S=1)	No
5	ResBlock(N=256, K=3, S=1)	No
6	ResBlock(N=256, K=3, S=1)	No
z	ResBlock(N=256, K=3, S=1) Reparameterization()	Yes

Layer	Decoders	Shared?
1	ResBlock(N=256, K=3, S=1)	Yes
2	ResBlock(N=256, K=3, S=1)	No
3	ResBlock(N=256, K=3, S=1)	No
4	ResBlock(N=256, K=3, S=1)	No
5	ConvTranspose2d(256, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1)) InstanceNorm2d(128, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
6	ConvTranspose2d(128, 64, kernel_size=(4,4), stride=(2,2), padding=(1,1)) InstanceNorm2d(64, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True) ReflectionPad2d((3, 3, 3, 3))	No
7	Conv2d(64, 3, kernel_size=(7,7), stride=(1,1)) Tanh()	No

Layer	Discriminators	Shared?
1	Conv2d(11, 64, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1)) LeakyReLU(negative_slope=0.2, inplace=True)	No
2	Conv2d(64, 128, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1)) InstanceNorm2d(128, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
3	Conv2d(128, 256, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1)) InstanceNorm2d(256, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
4	Conv2d(256, 512, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1)) InstanceNorm2d(512, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
5	Conv2d(512, 1, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))	No

ResBlock
ReflectionPad2d((1, 1, 1, 1)) Conv2d(N, N, kernel_size=(K, K), stride=(S, S)) InstanceNorm2d(N, eps=1e-05, momentum=0.1, affine=False) ReLU(inplace=True) ReflectionPad2d((1, 1, 1, 1)) Conv2d(N, N, kernel_size=(K, K), stride=(S, S)) InstanceNorm2d(N, eps=1e-05, momentum=0.1, affine=False)

TABLE 3.2: XAE model architecture. Layers are represented in PyTorch pseudocode.

Chapter 4

Toward single-cell data analysis across multiplex tissue imaging platforms

— What if the answers are wrong?
— Just stir the pile until they start looking right.

xkcd

4.1 Abstract

The emergence of megascale single-cell multiplex tissue imaging (MTI) datasets necessitates reproducible, scalable, and robust tools for cell phenotyping and spatial analysis. We developed open-source, graphics processing unit (GPU)-accelerated tools for intensity normalization, phenotyping, and microenvironment characterization. We deploy the toolkit on a human breast cancer (BC) tissue microarray stained by cyclic immunofluorescence and benchmark our cell phenotypes against a published MTI dataset. Finally, we demonstrate an integrative analysis revealing BC subtype-specific features.

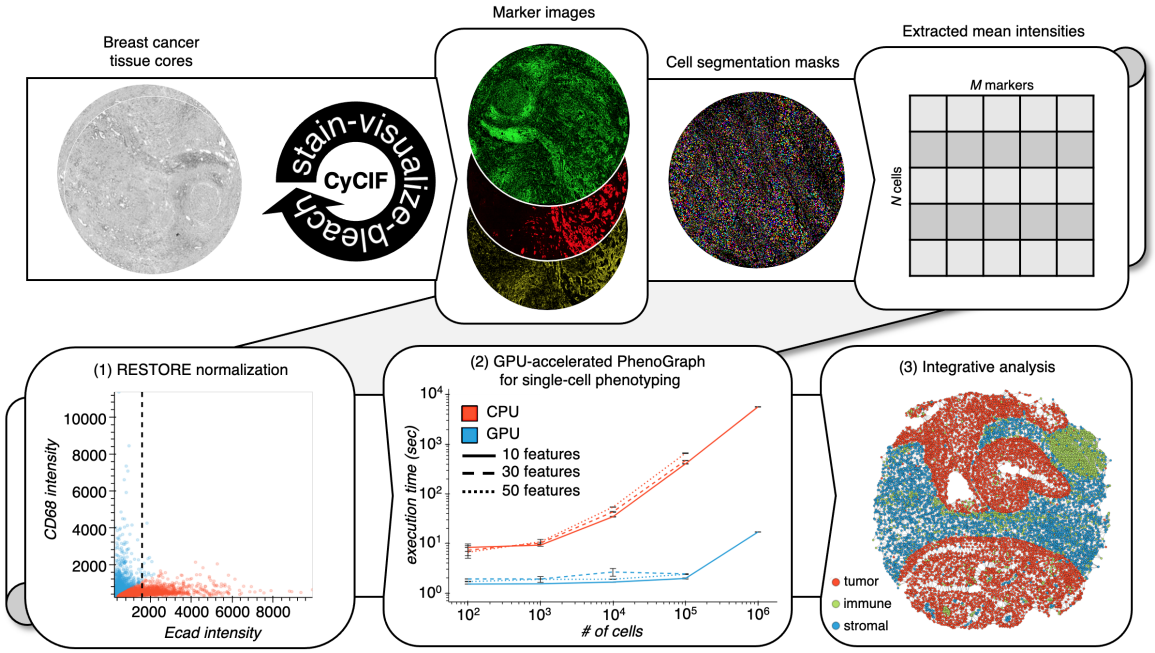


FIGURE 4.1: Overview of CyCIF analysis workflow. Once TMA cores are stained by CyCIF, cells are segmented and cell mean intensities are extracted, normalized, then used to define cell phenotypes for analyses of tissue composition and architecture. Box (1) shows an example of RESTORE normalization of single-cell Ecad intensities using mutually-exclusive expression of CD68 to derive a normalization factor for Ecad. Box (2) shows benchmarking results for CPU and GPU implementations of PhenoGraph for phenotyping of simulated single-cell datasets. Compared to the legacy CPU implementation, our GPU implementation of PhenoGraph is orders of magnitude faster at scale. Error bars show standard deviation of three replicate executions. Box (3) shows the spatial layout of high-dimensional cell phenotypes in a representative tissue core.

4.2 Introduction

Multiplex tissue imaging (MTI) methods like cyclic immunofluorescence (CyCIF) [60, 30], CODEX [37], multiplex immunohistochemistry (mIHC) [111], imaging mass cytometry (IMC) [36], and multiplex ion beam imaging [4] enable measurements of the expression and spatial distribution of tens of markers in tissues, and have facilitated our understanding of the interactions and relationships among distinct cell types in diverse tissue microenvironments. Nevertheless, for MTI to reach its full potential as a research

paradigm, numerous computational challenges must be overcome, including (1) reproducible normalization of single-cell intensity measurements to enable intra- and inter-sample comparisons; (2) robust cell phenotyping at megascale to enable comparison–and soon compilation–of MTI datasets from different platforms; and (3) the development of insightful spatial features to characterize the microenvironment of the tissue or disease of interest, and so enable discrimination between tissues that vary over important clinical parameters.

To address these challenges, we present (see Figure 4.1): (1) a broadened application of our data-intrinsic normalization method [17], which leverages the mutually exclusive expression pattern of marker pairs in MTI stain panels to estimate normalization factors without subjective and time-consuming manual gating; (2) a distributed and graphics processing unit (GPU)-accelerated implementation of PhenoGraph [58], the popular graph-based algorithm for subpopulation detection in high-dimensional single-cell data; and (3) an integrative analysis using this toolkit on ∼1.3 million cells from a 180-sample, pan-subtype human breast cancer (BC) tissue microarray (TMA) dataset (Figure 4.2) stained by CyCIF using a marker panel that characterizes tumor, immune, and stromal compartments (Table 4.1). Through consideration of both tissue composition and architecture, we identify features independent from hormone receptor (HR) and human epidermal growth factor receptor 2 (HER2) expression which discriminate between the canonical BC subtypes.

4.3 Results

For RESTORE normalization [17] of each TMA core, we leverage the fact that tumor, immune, and stromal cells exhibit mutually exclusive expression of cell type-specific markers, and use a graph-based clustering to define positive and negative cells and normalization factors (Figure 4.3A). Putative reference and mutually-exclusive (ME) marker pairs used

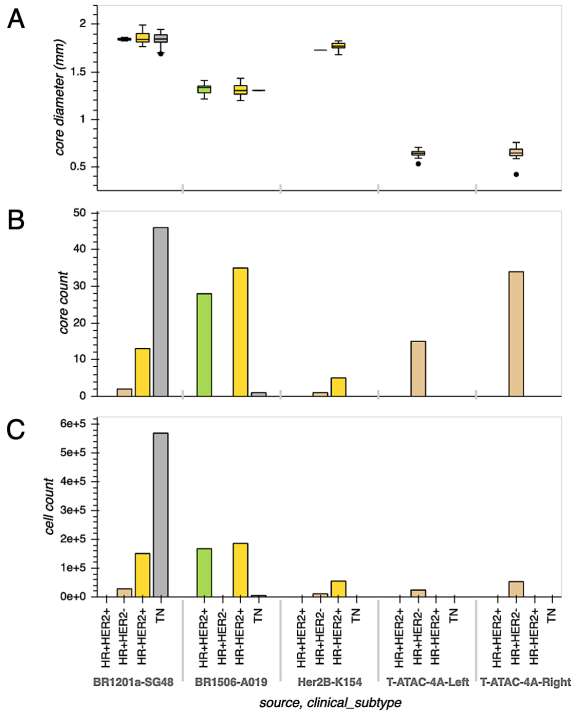


FIGURE 4.2: Overview of TMA composition. (A) Core diameters split by TMA source and BC subtype. (B) Core count split by TMA source and BC subtype. (C) Cell count split by TMA source and BC subtype.

for normalization can be found in Table 4.2. Also see section 4.5.3 for implementation details.

When the raw expression vectors of all cells across TMAs are embedded by *t*-stochastic neighbor embedding (*t*-SNE) [70], cells are segregated based on TMA source (Figure 4.3B, left), mainly due to batch effect and in part due to subtype bias within TMAs (Figure 4.2B-C). Following normalization, shared cell types between TMAs are co-embedded (Figure 4.3B, right) and cell expression of immune, tumor, and stromal markers is segregated (Figure 4.4C), a validation of the normalization process.

To define cell types among the ∼1.3 million cells in the normalized feature table, we first attempted to use the central processing unit (CPU)-based version of the widely-used algorithm PhenoGraph [58], but found it to be inefficient at this scale. To overcome this computational bottleneck, we re-implemented PhenoGraph to be executable on GPUs.

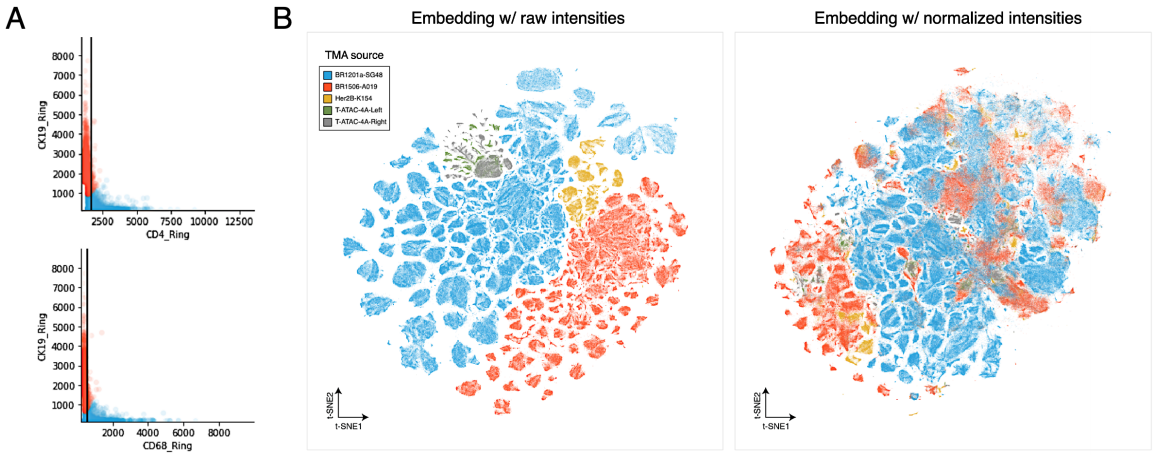


FIGURE 4.3: Cell mean intensity normalization across TMAs. (A) Example of RESTORE normalization [17] of CD4 and CD68 cell mean intensities for a single TMA core. Cell mean intensities of these immune markers are plotted against the cell mean intensity of epithelial CK19, a mutually exclusive marker. Cells are partitioned into positive (blue) and negative (red) populations. Black lines represent the computed normalization factors. Cells to the right of each line are above the background intensity level for that immune marker. (B) *t*-SNE embeddings of all cells using either raw (left) or normalized (right) cell mean intensities for all markers. Cells are colored according to the TMA from which they originate. A strong batch effect is observed before normalization, leading to partitioning according to TMA of origin. Following normalization, cell phenotypes shared between TMAs are co-localized. However, some TMA-specific partitioning remains due to subtype-specific marker bias within TMAs. The coordinates used in the *t*-SNE plots at right are the same as those used in Figure 4.5A, where cells are colored by metacluster. In the plot in Figure 4.5A, it is clear that the HR+ and HER2+ tumor cells aggregated in the lower left correspond to the HR+ and HER2+ TMA cores from the BR1506, Her2B, and T-ATAC cohorts from Figure 4.3.

Using the Python libraries RAPIDS [88] and CuPy [79] to parallelize and accelerate several of PhenoGraph's computations (see subsection 4.5.4), we observed multiple orders of magnitude improvement in the algorithm's speed without sacrificing clustering quality (Figure 4.1, box (2)). Our PhenoGraph implementation identified diverse tumor, immune, and stromal cell types across tissues and BC subtypes (Figure 4.5). To define phenotypes shared across tissues, metaclusters of similar phenotypes were aggregated based on the hierarchical clustering of phenotypes based on their mean marker expression. While tissues from all BC subtypes contained similar populations of immune, stromal, and endothelial cells, differences between BC subtypes were largely driven by variable tumor cells expression of luminal and basal cytokeratins, HER2, and the hormone receptors ER and PgR, as

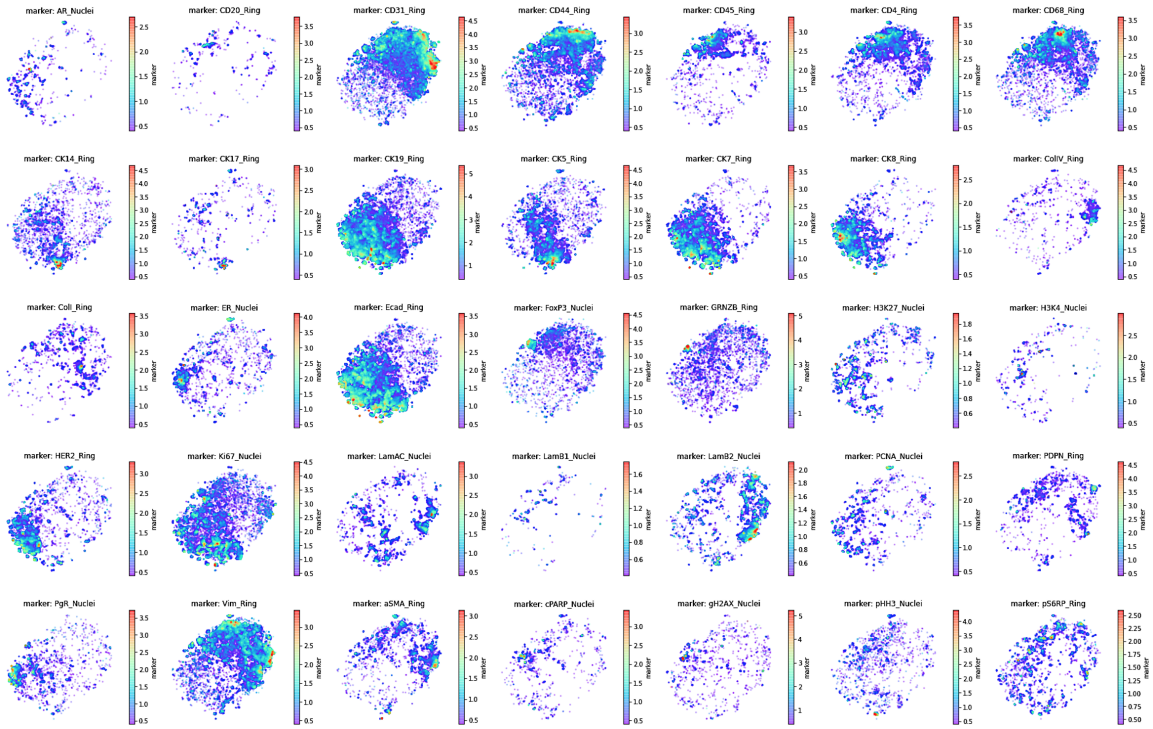


FIGURE 4.4: Marker intensity distribution over normalized cells. *t*-SNE normalized embedding from Figure 4.3, faceted by marker, showing only cells that have a mean intensity above the normalization factor for that marker. Color scale is log-transformed normalized intensity.

previously reported [50]. We further assessed the robustness of our identified cell phenotypes by either subsampling tissue cores (Figure 4.6A-C), or by applying $\pm 20\%$ noise to normalization factors for each marker for each core (Figure 4.6D) and found that phenotypes are identified as they are sampled and are robust to minor variation in normalization factors.

With the growth of MTI in the cancer research and translational communities, there is an acute need for robust and integrative analyses of MTI data across platforms and cohorts [97]. In a step toward addressing that need, we validated our identified cell types through comparison with a recently published survey of BC by IMC [50]. While the total number of cells from each BC subtype varied between the Basel (IMC) and OHSU (CyCIF) cohorts (Figure 4.7A), there was substantial overlap between the marker panels used for

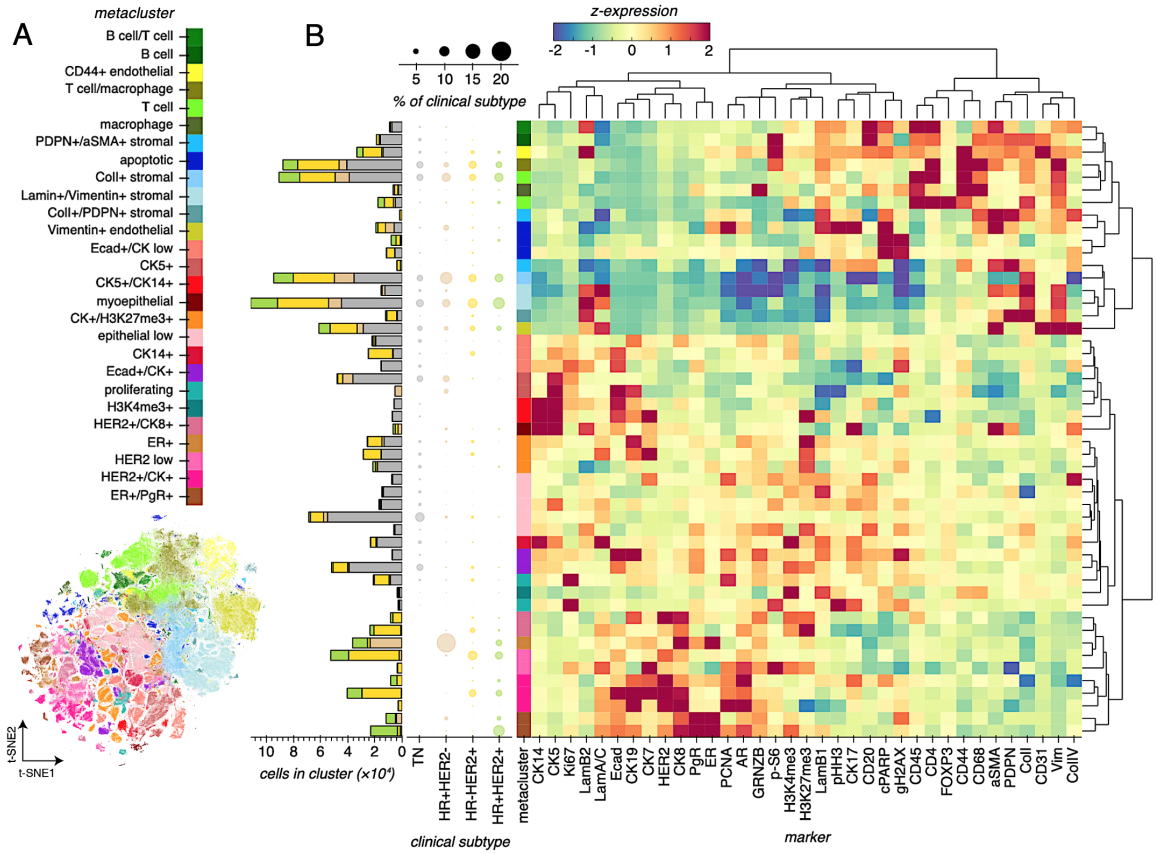


FIGURE 4.5: Defining single-cell phenotypes across breast cancer clinical subtypes. (A) *t*-SNE embedding of full single-cell CyCIF dataset colored by cell phenotype metacluster. (B) Hierarchical clustering of PhenoGraph clusters and CyCIF markers. The colorscale represents the *z*-scored marker expression. The scatter plot displays how each BC subtype is composed, where point size represents the percentage of that BC subtype that is composed of that cluster. The bar plot represents that absolute number of cells belonging to each cluster and BC subtype.

each MTI platform (Figure 4.7B). By aligning the cell phenotypes independently detected by PhenoGraph in each cohort (Figure 4.8C), we found highly-correlated clusters for stromal, immune, basal, and proliferating cell types, among others (Figure 4.8D), suggesting that shared cell types could be matched across cohorts and MTI platforms, a necessary step for data integration. We note that differences between cohort cell types may reflect the differences between cohort composition with respect to BC subtype. Consistent between cohorts and platforms, tumor cells differed more between samples than did immune, stromal, and endothelial cells (Figure 4.9).

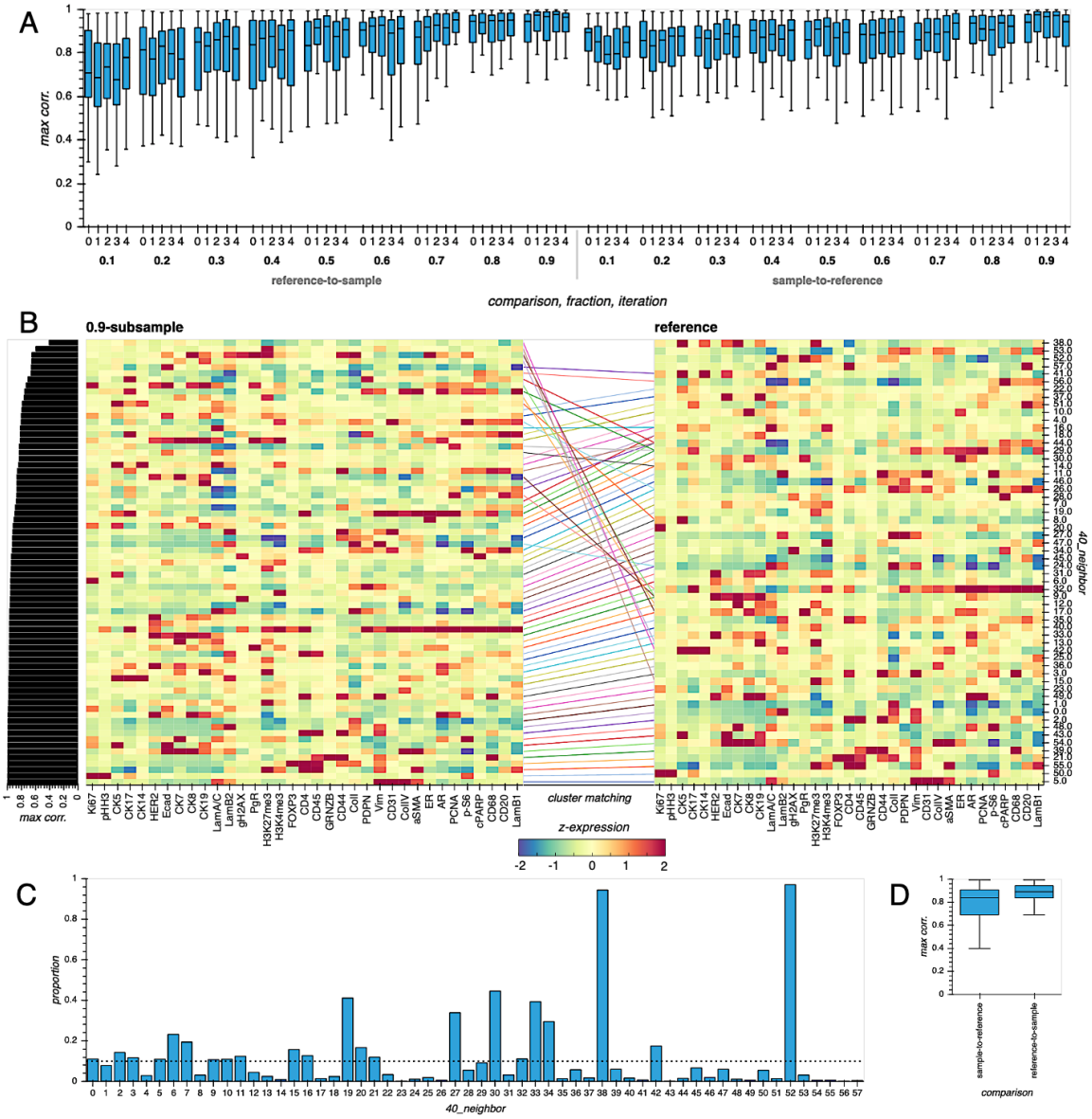


FIGURE 4.6: Validation of BC cell phenotype robustness. (A) Maximum Pearson's correlations between full reference cell phenotypes and those derived using PhenoGraph on cells from a random fractional sample of TMA cores, iterated five times at each fraction level. (B) Phenotype matching between the full reference phenotypes and those derived from a 90%-subsampled fraction of TMA cores. Phenotypes are ordered by increasing matching correlation. Matching phenotypes are linked by a line, and lines are colored to discriminate between adjacent or overlapping links. 40_neighbor represents the PhenoGraph cluster labels since we set $k = 40$ when defining the k-nearest neighbor graph in the PhenoGraph routine. The colorbar indicates z-scored marker expression. (C) The proportion of cells from each reference phenotype that correspond to the 10% of TMA cores held out from the 90%-subsampled fraction from (B). Unmatched phenotypes in the full reference correspond in part to cells from held-out cores. The dotted line marks the 10% threshold. (D) Maximum Pearson's correlation between full reference phenotypes and those derived using PhenoGraph on normalized mean intensities from all TMA cores, but with $\pm$20% random noise added to each marker in each core.

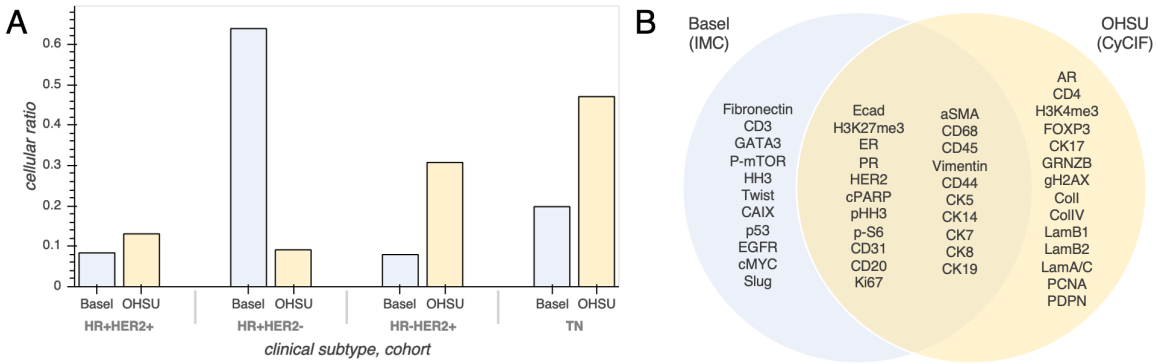


FIGURE 4.7: Comparison of IMC and CyCIF BCTMA datasets. (A) Cellular ratio highlighting compositional differences between Basel [50] and OHSU (this work) cohorts with respect to BC subtype. (B) The intersection of the IMC and CyCIF marker panels used to stain tissues from the Basel and OHSU cohorts, respectively.

Although BC is appreciated as a genetically and morphologically heterogeneous disease, its clinical subtyping is based on the expression of relatively few markers—in particular, tumor cell expression of the hormone receptors for estrogen and progesterone (ER and PgR, respectively, or HR, collectively), and human epidermal growth factor 2 (HER2)—which is insufficient to explain differences in treatment response within each subtype [87]. Recent studies using MTI to interrogate intact BC tissues have found that the spatial contexture of the BC microenvironment can improve our ability to predict clinical outcome [50, 53]. However, because these studies have either focused on disease risk or a single BC clinical subtype, here we focused on compositional and spatial features which differentiate between subtypes.

At the composition level, we first considered the tumor cell differentiation states of BC subtypes through their expression of luminal and basal cytokeratins (CKs). While CK+ cells in HR-/HER2+, HR+/HER2-, and HR+/HER2+ tissues primarily expressed luminal CKs 19, 8, and 7, CK+ cells in TN tissues exhibited significantly greater differentiation state heterogeneity (Figure 4.10A), as they expressed many different combinations of luminal and basal cytokeratins (Figure 4.10B). This differentiation state heterogeneity is consistent with the genetic and histological heterogeneity of triple negative BC (TNBC) described

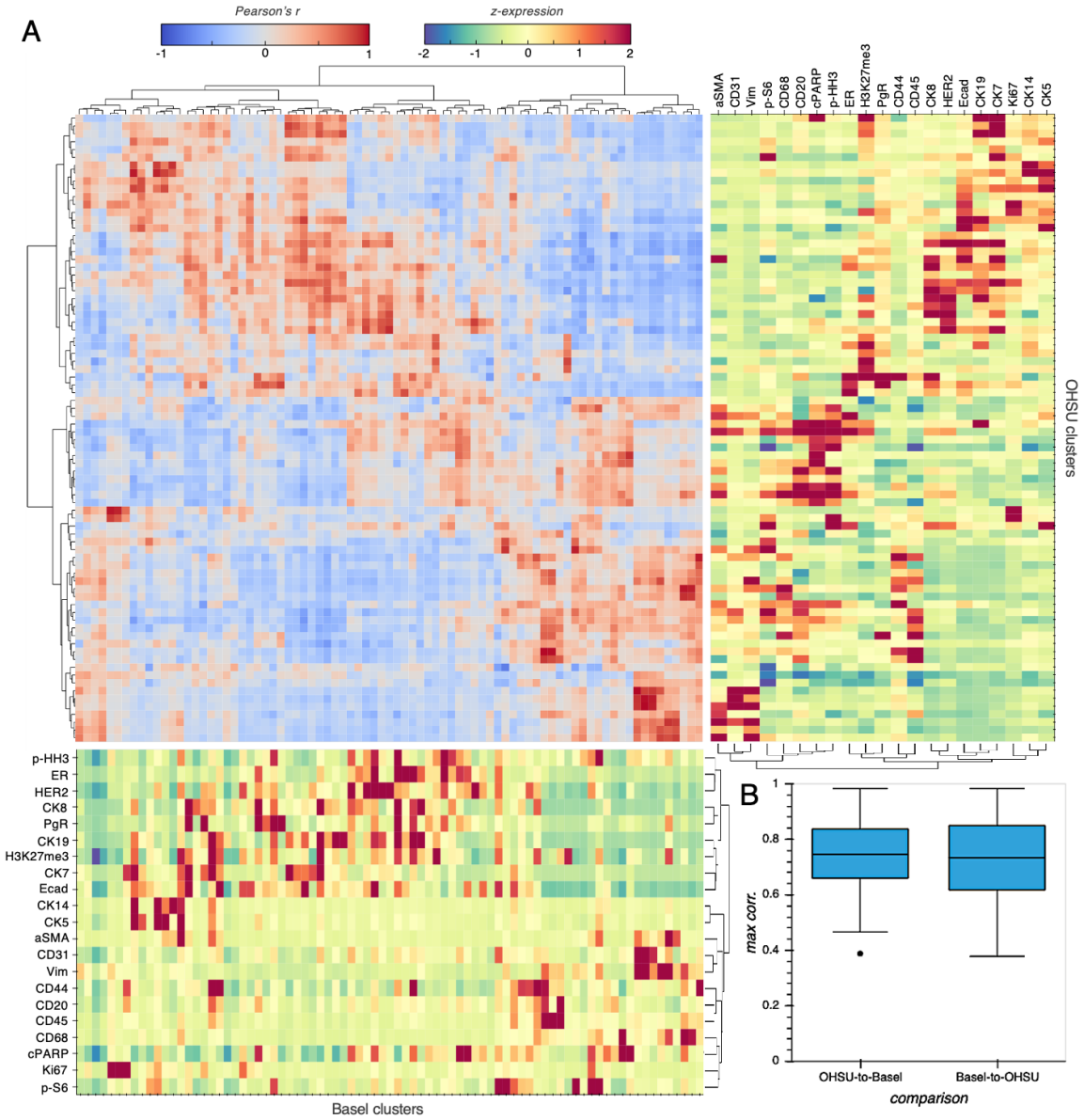


FIGURE 4.8: Cross-platform benchmarking of BC cell phenotypes. (A) PhenoGraph cluster matching between Basel and OHSU cohorts. Using only the intersecting markers, cells from each cohort were independently clustered using PhenoGraph with the same parameterization, then cohort clusters were pairwise correlated and hierarchically clustered based on the resulting correlation structure. We identified highly-correlated clusters between cohorts, including those corresponding to epithelial, immune, stromal, endothelial, and proliferating cell populations. (B) Maximum Pearson's correlation corresponding to inter-cohort cluster matches.

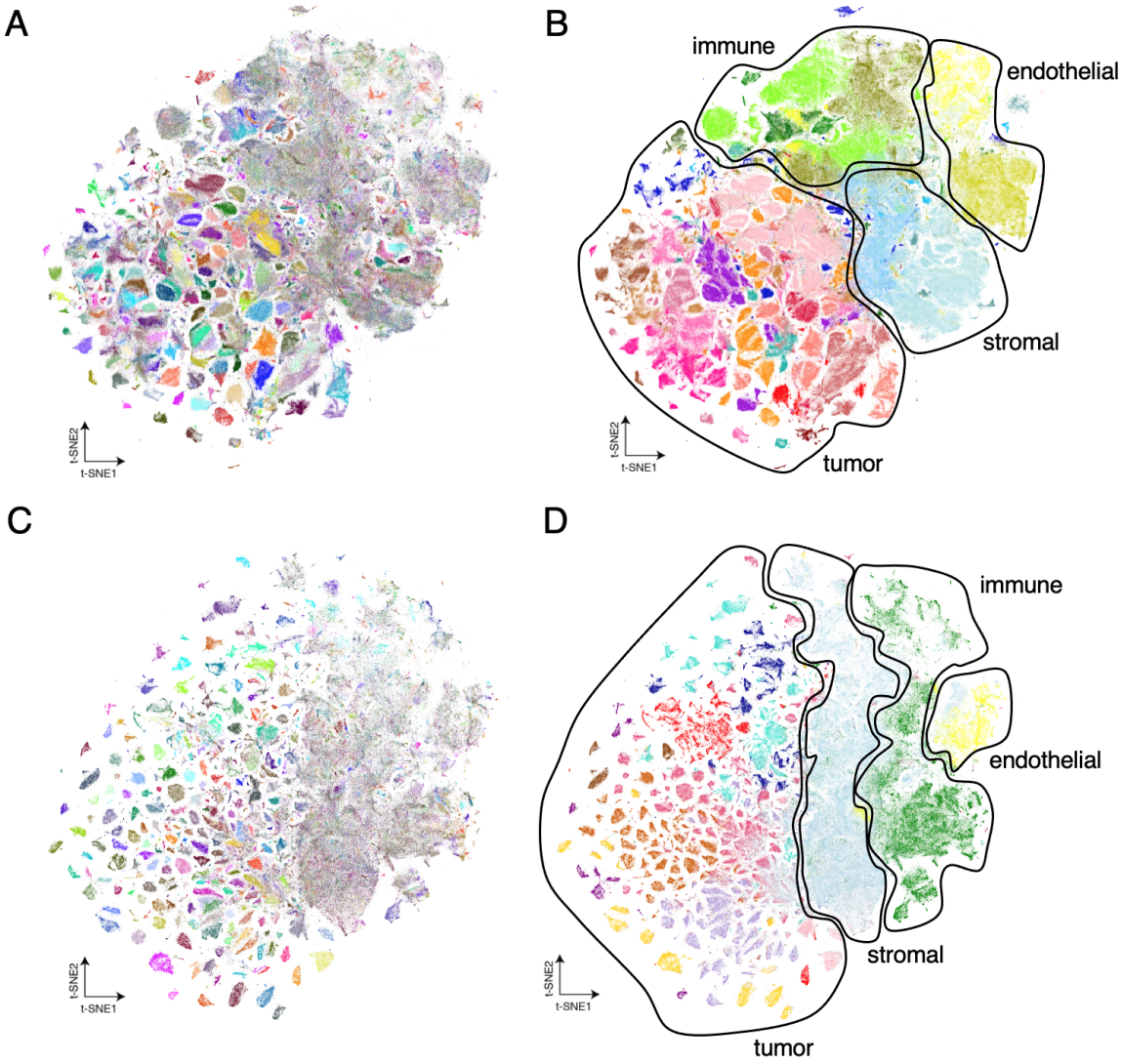


FIGURE 4.9: For both OHSU and Basel cohorts, tumor cells differ more between samples than immune, stromal, or endothelial cells. (A) The same *t*-SNE embedding of the OHSU dataset from Figure 4.5A, but with cells colored based on the unique tissue core from which they are derived. (B) The same *t*-SNE embedding of the OHSU dataset from Figure 4.5A, but with annotations indicating immune, stromal, endothelial, and tumor phenotypic regions. (C) A *t*-SNE embedding of the Basel dataset [50] derived using the same parameters as were used for the OHSU *t*-SNE embedding, with cells colored based on the unique tissue core from which they are derived. (D) The same *t*-SNE embedding as in (C), but with annotations indicating immune, stromal, endothelial, and tumor phenotypic regions.

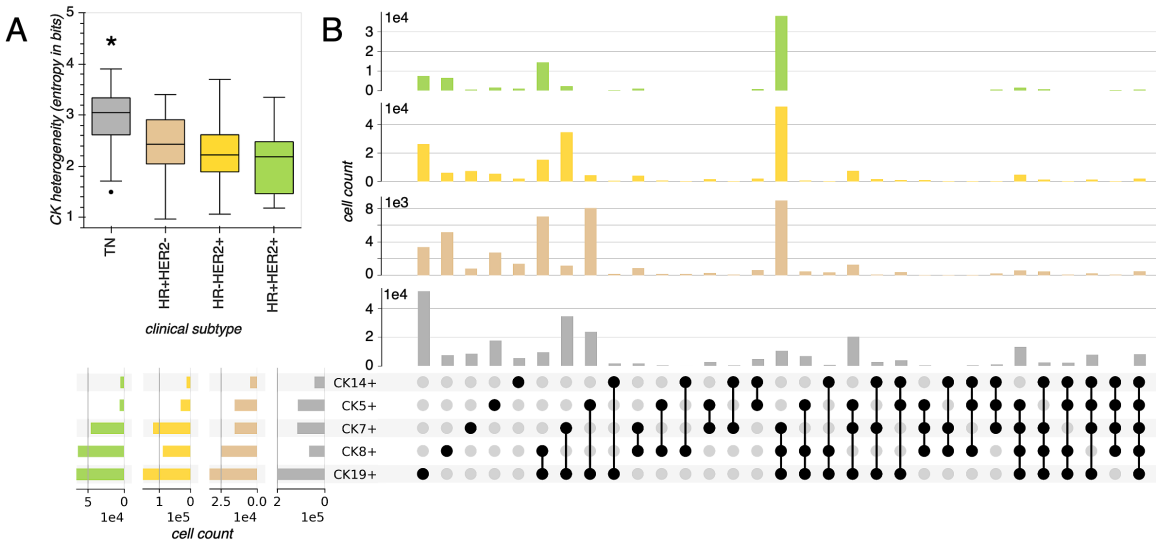


FIGURE 4.10: Epithelial differentiation heterogeneity across BC subtypes. (A) Box plot displaying cytokeratin (CK) expression heterogeneity across BC subtypes, where each box represents the distribution of tissue cores from a BC subtype, and each core is summarized based on the entropy of the distribution of CK+ cell types contained within it. Groupwise comparisons were made using one-way ANOVA with pairwise Tukey post-hoc test (TN, $n = 47$; HR+HER2-, $n = 52$; HR-HER2+, $n = 53$; HR+HER2+, $n = 28$). *$P < 0.001$ for all TN comparisons with other BC subtypes. (B) UpSet plot summarizing the distribution of CK+ cell types across BC subtypes, considering each CK alone (left margin) or in combination (upper margin).

in other studies [42, 8]. We next determined the composition of each tissue core with respect to the cell metaclusters we defined above. Hierarchical clustering of cores based on their cell metacluster densities highlighted the broad variability of cellular composition within and between BC subtypes (Figure 4.11A-B). When the cell metaclusters were further aggregated into immune, stromal, and tumor cell types (see section 4.5.8), we found the HR+HER2- tissues to have lower overall immune cell density than the other BC subtypes, and no differences in stromal or tumor cell density between subtypes (Figure 4.11C).

Recognizing that cell density measurements fail to capture the organization of cells in each tissue, we next characterized the spatial architectures of BC subtypes by building cell neighborhood graphs for each tissue. Given the recent evidence that the quantity and diversity of BC tumor cell interactions with other cell types can inform disease outcome [50,

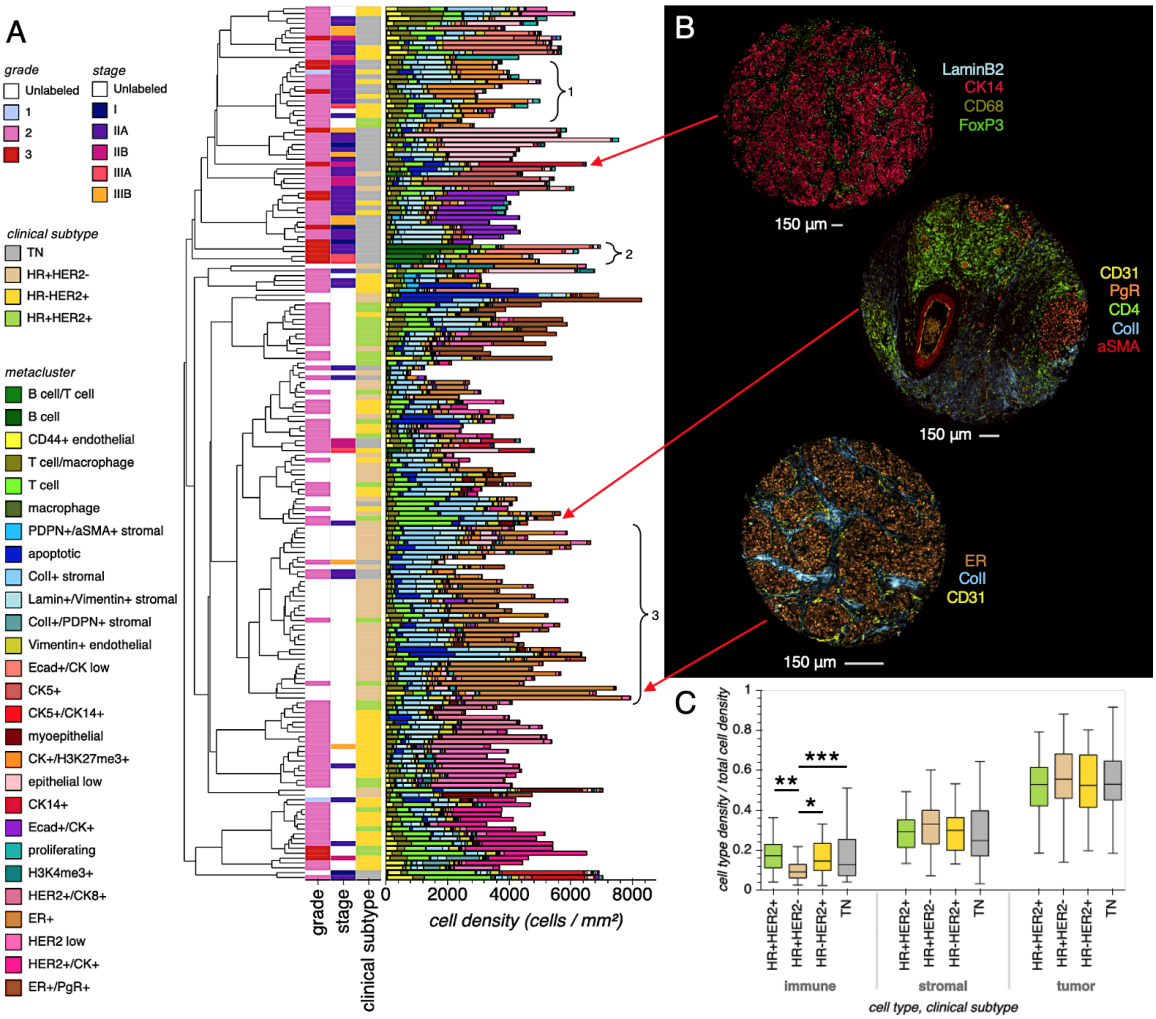


FIGURE 4.11: (A) Cell phenotype density across tissue cores. Bar plot where each bar represents a TMA core, the full bar height represents its total cell density, and each colored segment represents the density of a particular cell metacluster. Bars are hierarchically clustered based on cell metacluster densities. Each bar is labeled with its corresponding subtype, stage, and grade, if a label is available. The inset brackets indicate (1) cores with abundant H3K27me3+ tumor cells, which could indicate a mechanism of HR repression in some TN and HR-/HER2+ tissues [21]; (2) cores with abundant infiltrating B cells (TIL-B), consistent with association found between TIL-B and high-grade, HR- BC [34]; and (3) cores with relatively low immune density, consistent with the finding that HR+/HER- tissues are immunologically cold compared to TN and HER2+ tissues [3, 125]. (B) A selection of representative tissue cores. (C) The immune, stromal, and tumor densities of tissue cores from each BC subtype. Groupwise comparisons were made using one-way Welch ANOVA and Games-Howell post-hoc test. $*P = 0.034$, $**P = 0.035$, $***P = 0.079$ (TN, $n = 47$; HR+HER2-, $n = 52$; HR-HER2+, $n = 53$; HR+HER2+, $n = 28$).

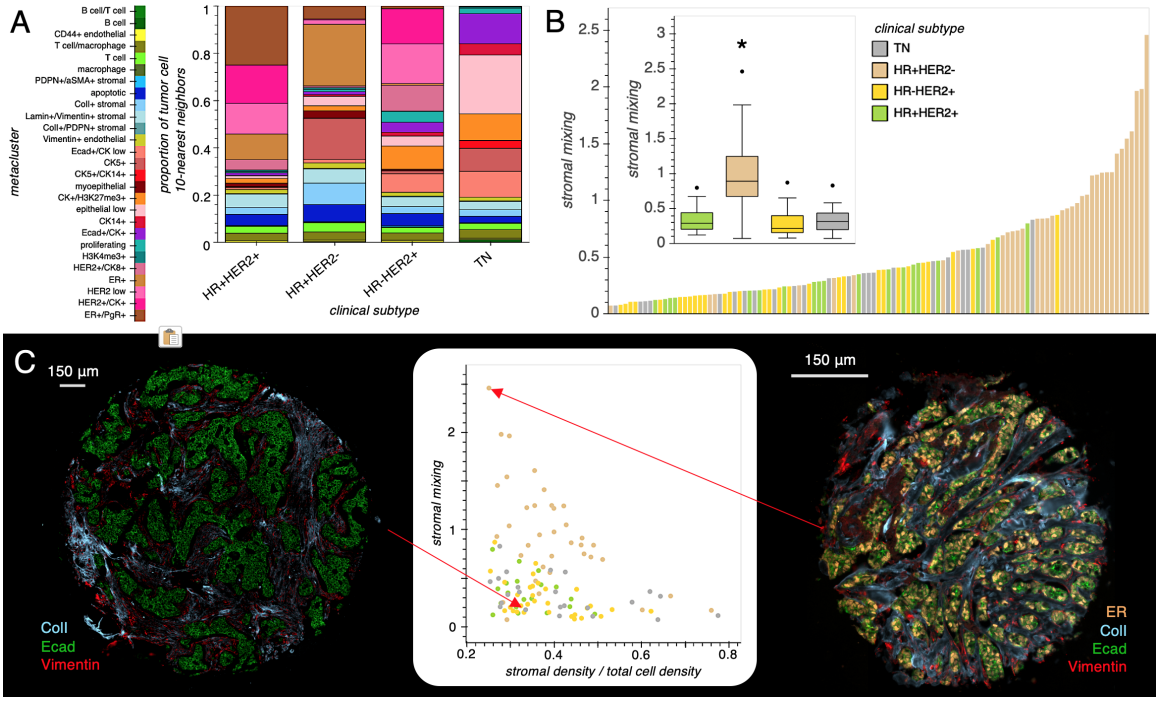


FIGURE 4.12: Breast cancer cellular composition belies tumor-stromal interaction.(A) Stacked bar plots displaying the proportion of tumor cell 10-nearest neighbors for each BC subtype. Each colored bar segment represents the proportion of tumor cell neighbors that are comprised of the corresponding cell metacluster. (B) Bar plot representation of tissue core stromal mixing, where only cores with greater than 0.25 stromal fraction are shown. Cores are ordered based on increasing stromal mixing. Inset shows box plot comparing stromal mixing over BC subtype. Groupwise comparisons were made using one-way Welch ANOVA and Games-Howell post-hoc test. *P <0.001 for all HR+HER2- comparisons (TN, $n = 47$; HR+HER2-, $n = 52$; HR-HER2+, $n = 53$; HR+HER2+, $n = 28$). (C) Scatter plot displaying stromal density versus stromal mixing and images of representative cores with similar stromal density but different stromal mixing.

53], we first identified the neighboring cells to each tumor cell and compared the composition of tumor cell neighborhoods across BC subtypes (Figure 4.12A). Though most tumor cell interactions (∼70-80%) are with other tumor cells typical to their BC subtype, we observed increased tumor-stromal interaction in the HR+HER2- subtype. When considering tissues that contain an appreciable population of stromal cells (tissues comprised of at least 25% stromal cells), we confirmed that there was significantly more stromal mixing with tumor cells in HR+HER2- tissues (Figure 4.12B). Importantly, stromal mixing can vary

widely between tissues in spite of their similar stromal density (Figure 4.12C), highlighting the importance of the spatial context that is preserved in intact tissues. Since malignant epithelial cells can suppress fibroblast maturation and thus promote fibroblast aromatase activity [12], ER+HER2- tumors likely favor more from proximal fibroblasts as a source of growth-inducing estrogen than other BC subtypes, and may even act to maintain tumor microenvironments with high stromal mixing [11].

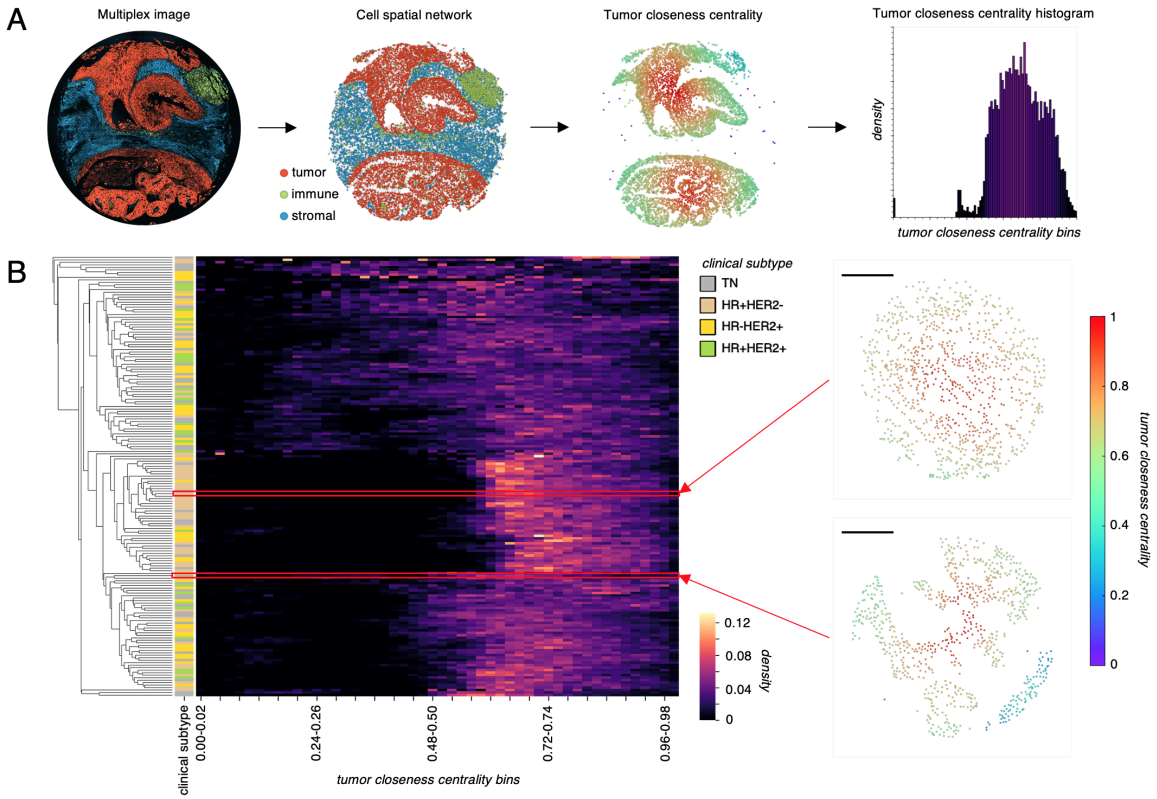


FIGURE 4.13: Graph-based characterization of tumor architecture discriminates HR+/HER2- tumors. (A) Overview of tumor architecture characterization. A spatial graph is defined over tumor cells, where each tumor cell is connected to others within a 65 µm radius from its centroid, and the closeness centrality is then measured over this graph. Each core is then summarized as a histogram of centrality values. (B) Hierarchical clustering of cores based on their tumor closeness centrality histograms. The upper and lower outset tumor graphs correspond to the right and left tissue core images in Figure 4.12C, respectively. Scale bar in tumor graphs is 150 µm.

We reasoned that differences in tumor-stromal interaction might translate into detectable differences between BC subtypes based on their tumor architectures alone.

To characterize the tumor architecture of each tissue, we constructed tumor architecture graphs over which we computed the closeness centrality for each tumor cell, which quantifies the relative closeness of that tumor cell to all other tumor cells in the tissue (Figure 4.13A). Consistent with the stromal mixing trend observed above, HR+HER2- tissues had a mean tumor closeness centrality significantly greater than tissues from the other BC subtypes (Figure 4.13B and Figure 4.14), which is in part a reflection of HR+HER2- tumor cell nests tending to be separated by narrower streams of stromal cells than tumor nests in tissues from other BC subtypes (Figure 4.12C). In summary, by analyzing BC tissues with spatially resolved MTI, we have identified inter-cell phenotype (stromal mixing) and intra-cell phenotype (tumor closeness centrality) interactions which can be leveraged to help discriminate between canonical BC subtypes on a basis other than receptor expression.

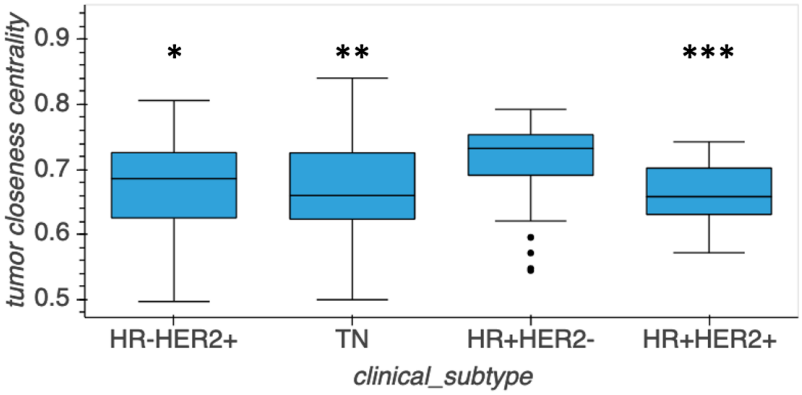


FIGURE 4.14: Tumor closeness centrality increased in HR+/HER2- tumors. Comparison of tumor closeness centrality between BC subtypes. Groupwise comparisons of mean tumor closeness centrality were made using one-way Welch ANOVA and Games-Howell post-hoc test. $*P = 0.016$, $**P = 0.0099$, $***P = 0.0010$ (TN, $n = 47$; HR+HER2-, $n = 52$; HR-HER2+, $n = 53$; HR+HER2+, $n = 28$).

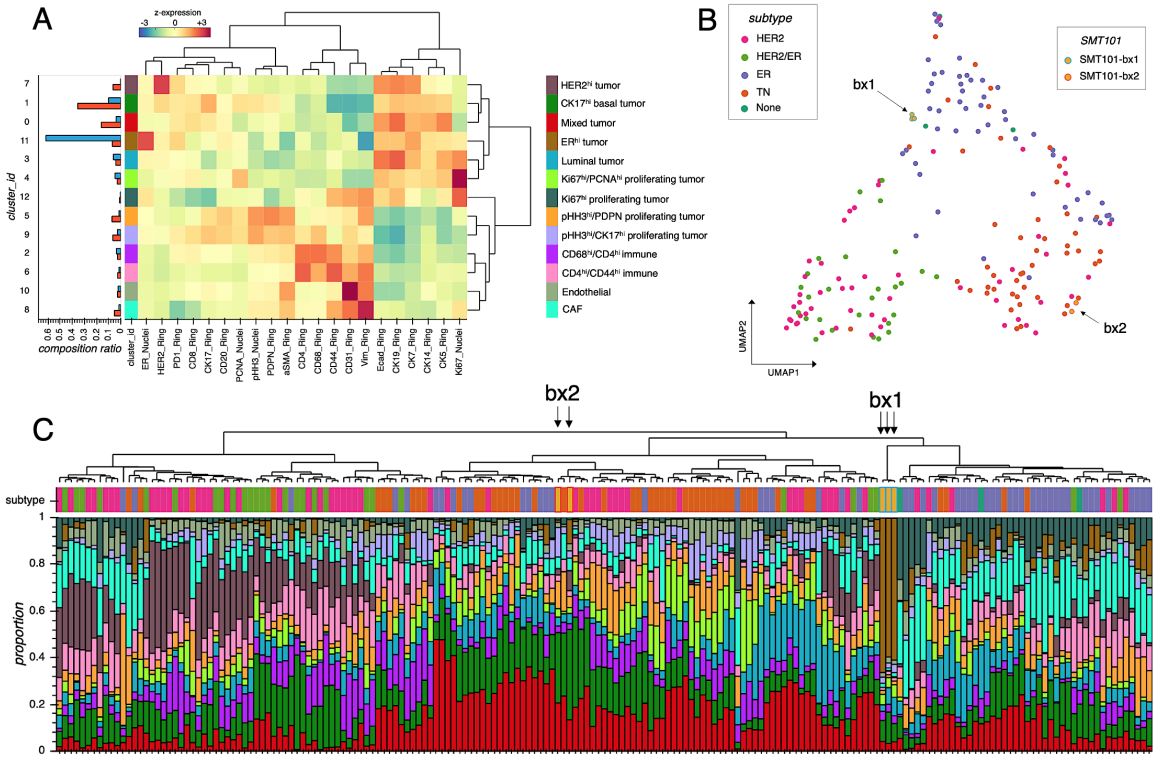


FIGURE 4.15: Using a BC cell type dictionary to put clinical samples in context. Here we demonstrate in a report mock-up how our analysis framework could be used in a clinical setting like the ongoing SMMART trial [51]. (A) In the SMARRT trial, patients have tissue samples taken throughout their treatment. By integrating the CyCIF data of these clinical samples with the BCTMA CyCIF data, we can put them in a pan-subtype context and are able to assess how cell type composition differs after treatment, as shown in the bar plot at left. Blue and red bars correspond to biopsy 1 and biopsy 2, respectively, i.e. pre- and post-treatment. (B) A UMAP projection of all SMARRT and BCTMA tisues based on cell type composition can indicate disease subtype, transition, or switching. (C) Tissues can be hierarchically clustered based on cell type composition to reveal population-level features and trends. Stacked bar plot colors correspond to the cell types defined in (A). The hierarchical clustering of subtype labels provides an alternative representation of tissue similarity compared to the UMAP projection.

4.4 Discussion

This work is motivated by an understanding that the spatial context of the tumor microenvironment in intact cancer tissues enables a more granular definition of disease, and—we hope—the design of more personalized and effective treatments. With spatially-resolved MTI, our analysis makes clear that the cellular composition of BC tissue can belie important aspects of its spatial architecture. Ongoing work involves validating these findings

in a cohort with more extensive clinical annotation to assess their significance to disease outcome between and within BC subtypes. One potential application of our BC cell phenotypes is as a reference "dictionary" for putting clinical samples in context (Figure 4.15). While the BC cell phenotypes and architectural features we have derived will be assets to future BC studies, our generic toolkit can be used stand-alone or integrated with existing toolkits [100] to improve the efficiency and reproducibility of analytics for any single-cell measurement platform.

4.5 Methods

4.5.1 Acquisition of breast cancer tissue microarrays (TMAs)

The tissues used in this study are a compilation of multiple TMAs: BR1201a-SG48 (US Biomax Inc., https://www.biomax.us/BR1201at), BR1506-A019 (US Biomax Inc., https://www.biomax.us/tissue-arrays/Breast/BR1506), Her2B-K154 (US Biomax Inc., https://www.biomax.us/tissue-arrays/Breast/Her2B), and the TransATAC TMAs T-ATAC-4A-Left and T-ATAC-4A-Right [28]. All tissues that were successfully stained and imaged were included in the study, representing 180 tissue cores from 128 patients.

4.5.2 Cyclic immunofluorescence (CyCIF) staining of tissues

Tissue preparation

Formalin-fixed paraffin-embedded (FFPE) human tissues were received mounted on adhesive slides. The slides were baked overnight in an oven at 55 °C (Robbin Scientific, Model 1000) and an additional 30 minutes at 65 °C (Clinical Scientific Equipment, NO. 100). Tissues were deparaffinized with xylene and rehydrated with graded ethanol baths. Two step antigen retrieval was performed in the Decloaking Chamber (Biocare Medical)

using the following settings: set point 1 (SP1), 125 °C, 30 seconds; SP2: 90 °C, 30 seconds; SP limit: 10 °C. Slides were further incubated in hot Target Retrieval Solution, pH 9 (Agilent, S236784-2) for 15 minutes. Slides were then washed in two brief changes of diH2O (∼2 seconds) and once for 5 minutes in 1x phosphate buffered saline (PBS), pH 7.4 (Fisher, BP39920). Sections were blocked in 10% normal goat serum (NGS, Vector S-1000), 1% bovine serum albumin (BSA, Sigma A7906) in PBS for 30 minutes at 20 °C in a humid chamber, followed by PBS washes. Primary antibodies were diluted in 5% NGS, 1% BSA in 1x PBS and applied overnight at 4 °C in a humid chamber, covered with plastic coverslips (Bio-Rad, SLF0601). Following overnight incubation, tissues were washed 3 x 10 min in 1x PBS. Coverslips (Corning; 2980-243 or 2980-245) were mounted in Slowfade Gold plus DAPI mounting media (Life Technologies, S36938).

Fluorescence microscopy

Fluorescently stained slides were scanned on the Zeiss AxioScan.Z1 (Zeiss, Germany) with a Colibri 7 light source (Zeiss). The filter cubes used for image collection were DAPI (Zeiss 96 HE), Alexa Fluor 488 (AF488, Zeiss 38 HE), AF555 (Zeiss 43 HE), AF647 (Zeiss 50) and AF750 (Chroma 49007 ET Cy7). The exposure time was determined individually for each slide and stain to ensure good dynamic range but not saturation. Full tissue scans were taken with the 20x objective (Plan-Apochromat 0.8NA WD=0.55, Zeiss) and stitching was performed in Zen Blue image acquisition software (Zeiss).

Quenching fluorescence signal

After successful scanning, slides were soaked in 1x PBS for 10–30 minutes in a glass Coplin jar, waiting until glass coverslip slid off without agitation. Quenching solution containing 20 mM sodium hydroxide (NaOH) and 3% hydrogen peroxide (H2O2) in 1x PBS was freshly prepared from stock solutions of 5 M NaOH and 30% H2O2, and each slide placed

in 10 ml quenching solution. Slides were quenched under incandescent light, for 30 minutes for FFPE tissue slides. Slides were then removed from chamber with forceps and washed 3 x 2 min in 1x PBS. The next round of primary antibodies was applied, diluted in blocking buffer as previously described, and imaging and quenching were repeated over ten rounds for FFPE tissue slides.

4.5.3 Data pre-processing

Cell segmentation and mean intensity extraction

Cell segmentation and mean intensity extraction were performed as previously described [30]. The nuclei and cells segmentation are performed using mathematical morphology. The process starts by segmenting the nuclei:

1. The DAPI image contrast is equalized using contrast-limited adaptive histogram equalization to remove illumination and staining irregularities.

2. The equalized DAPI image is cleaned by removing noise and artifacts as well as flattening the texture using an alternative sequential filter (alternation of opening and closing with structuring elements of increasing size).

3. A white top-hat filter is applied to separate the nuclei from the remaining background.

4. Area openings and closings (opening/closing based on the surface instead of a structuring element) are performed to flatten nuclei texture.

5. An ultimate opening is employed to find nuclei centers.

6. Nuclei centers are used as seeds in a watershed algorithm applied on the Sobel gradient of the original image, which provides the final nuclei segmentation.

7. For cell segmentation, nuclear segmentation masks are used as seeds in another watershed algorithm applied on a gradients combination of the markers CD44, CD45, CK7, CK19, and E-cadherin.

Mean intensities for each cell were extracted from the biologically-relevant compartment for each marker, i.e. mean intensities for markers with known nuclear (cytoplasmic) localization were extracted from nuclear (cytoplasmic) segmentation masks (Table 4.2). Cytoplasmic segmentation masks were computed by subtracting nuclear segmentation masks from full cell body segmentation masks.

Single-cell intensity normalization

Normalization factors for single-cell mean intensities were computed as previously described [17] using the putative mutually-exclusive marker pairs in Table 4.2. Normalization factors are computed for each pair of reference and mutually-exclusive markers, and the median of these factors is used to normalize each raw single-cell mean intensity vector for each CyCIF marker and each TMA core. Raw intensities were normalized using the equation:

$$\hat{\mathbf{x}}_{i,j} = \frac{\mathbf{x}_{i,j} - min(\mathbf{x}_{i,j})}{\phi_{i,j} - min(\mathbf{x}_{i,j})}, \tag{4.1}$$

where $\hat{\mathbf{x}}_{i,j}$ and $\mathbf{x}_{i,j}$ are the normalized and raw single-cell mean intensity vectors for CyCIF marker i for all cells in tissue core j, respectively, and $\phi_{i,j}$ is the corresponding normalization factor determined as described above. Therefore, cells with a normalized intensity greater than 1 are considered to be above the background intensity level.

4.5.4 Single-cell phenotyping

Algorithm selection

The number of single-cell phenotyping algorithms has rapidly proliferated alongside the development of single-cell cytometry platforms [124, 65]. With so many options from

which to choose, the appropriate choice of algorithm depends on the questions one hopes to ask of the data being considered. For instance, prior biological knowledge can be leveraged by supervised or semi-supervised algorithms to bias identification toward known cell types of interest. In contrast, unsupervised algorithms identify cell types by leveraging only the internal data structure, making them the algorithms of choice in discovery-based studies where relatively little is known about the underlying biology. Among the unsupervised algorithms, PhenoGraph [58] and FlowSOM [119] stand out for their abilities to precisely identify known cell types with high cluster coherence—i.e. high (low) inter-(intra)-cluster variance—and without the incorporation of prior biological knowledge [65]. One notable tradeoff between these two CPU-executed algorithms is runtime, with FlowSOM having a considerably faster runtime compared to PhenoGraph [124]. However, because PhenoGraph was used to define breast cancer cell types in a related study [50], we opted to use PhenoGraph in the current study to enable a fair comparison of the cell types defined between studies. Furthermore, the relatively slow runtime of CPU-executed PhenoGraph motivated our GPU-accelerated implementation.

GPU acceleration of PhenoGraph

Given a cell-by-feature dataframe, the PhenoGraph algorithm [58] consists of two primary steps: (1) defining a k-nearest neighbor graph over all cells that is then refined by computing the Jaccard similarity measure over graph edges, and (2) partitioning the graph into discrete cell phenotypes through optimization of partition modularity, such that cells in the same partition are more connected to each other than to cells of another partition. In the official version of PhenoGraph (`https://github.com/dpeerlab/PhenoGraph`), these steps are implemented using a combination of Python and C++ libraries that execute on CPU. In Figure 4.1 box (2), we show that PhenoGraph execution time increases exponentially with increasing dataset size, taking approximately 3 hours to process a synthetic 1

million cell-by-10 feature dataset. Most MTI datasets measure tens of features, but CPU-based PhenoGraph was unable to fully process the 1 million cell-by-30- and 50-feature synthetic datasets in the 8 hours allotted for the experiment. We see such computational bottlenecks–which would be even further constricted when compiling multiple MTI or cytometry datasets–as a major obstacle to current studies and future meta-studies of high-dimensional MTI datasets, where rapid iteration will be essential to the validation of cross-platform data integration techniques.

Owing to recent advances in GPU computing and its ever-broadening adoption in machine learning research, there now exist accelerated GPU-based analogs of many Python scientific computing libraries [88, 79], including those with which the CPU-based PhenoGraph is implemented. Some of these libraries even allow computation to be distributed across multiple GPUs [88, 26]. We employed two of such libraries, CuPy [79] and RAPIDS [88], to accelerate each step of the PhenoGraph algorithm and enable distributed computing over multiple GPUs. For example, for a synthetic dataset containing 50,000 samples and 50 features, the GPU implementation realizes a 354-fold speed up in the graph building and refinement step (97.3 seconds for CPU vs. 0.275 seconds for GPU) and a 141-fold speed up in the Louvain partitioning step (11.2 seconds for CPU vs. 0.0795 seconds for GPU).

With our GPU-based implementation, it is now possible to phenotype cells in megascale cytometry datasets in seconds-to-minutes rather than hours-to-days, and without subsampling. Moreover, our GPU implementation of PhenoGraph is competitive with FlowSOM in terms of execution time (Figure 4.16), which until now was one of the primary motivations for choosing FlowSOM over PhenoGraph [124, 65].

We have packaged our GPU implementation into a Python library called *grapheno* and have adopted the API from the official CPU implementation of PhenoGraph found at `https://github.com/dpeerlab/phenograph`. To run a simple clustering of synthetic data:

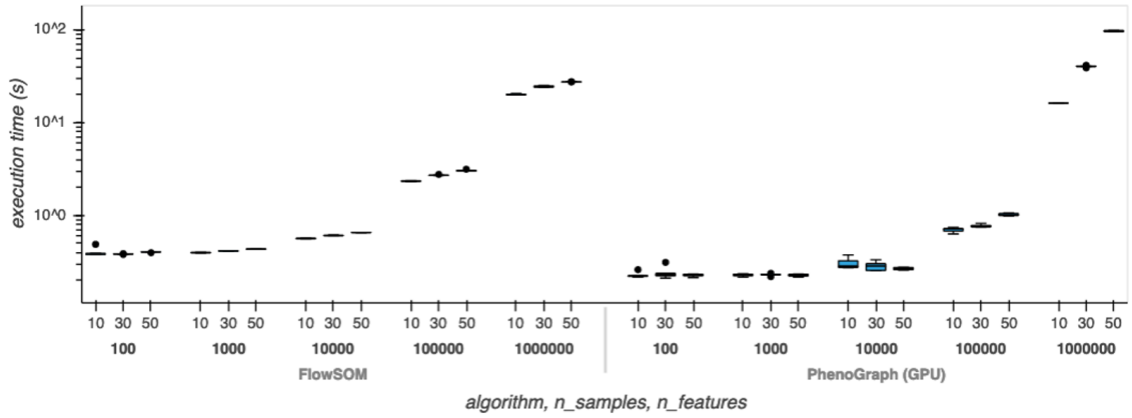


FIGURE 4.16: Comparison of execution time between FlowSOM and our GPU implementation of PhenoGraph.

```
import cudf
import cuml
import grapheno

X, _ = cuml.make_blobs()
X = cudf.DataFrame.from_records(X)
communities, G, Q = grapheno.cluster(X)
```

In practice, X can be any single-cell dataframe with cells as rows and features as columns. For a dataframe with N cells, *communities* will be a vector of length N specifying the cluster label for each cell. G is a RAPIDS graph object representing the relationships between cells that were used for clustering. Q is the modularity score for the clustering result defined by *communities*. Installation instructions and additional details about our implementation can be found at https://gitlab.com/eburling/grapheno.

Benchmarking CPU and GPU implementations of PhenoGraph

To ensure that our GPU implementation of PhenoGraph produced results consistent with the CPU implementation, we benchmarked each using synthetic datasets which varied in terms of number of samples and number of features (Figure 4.17).

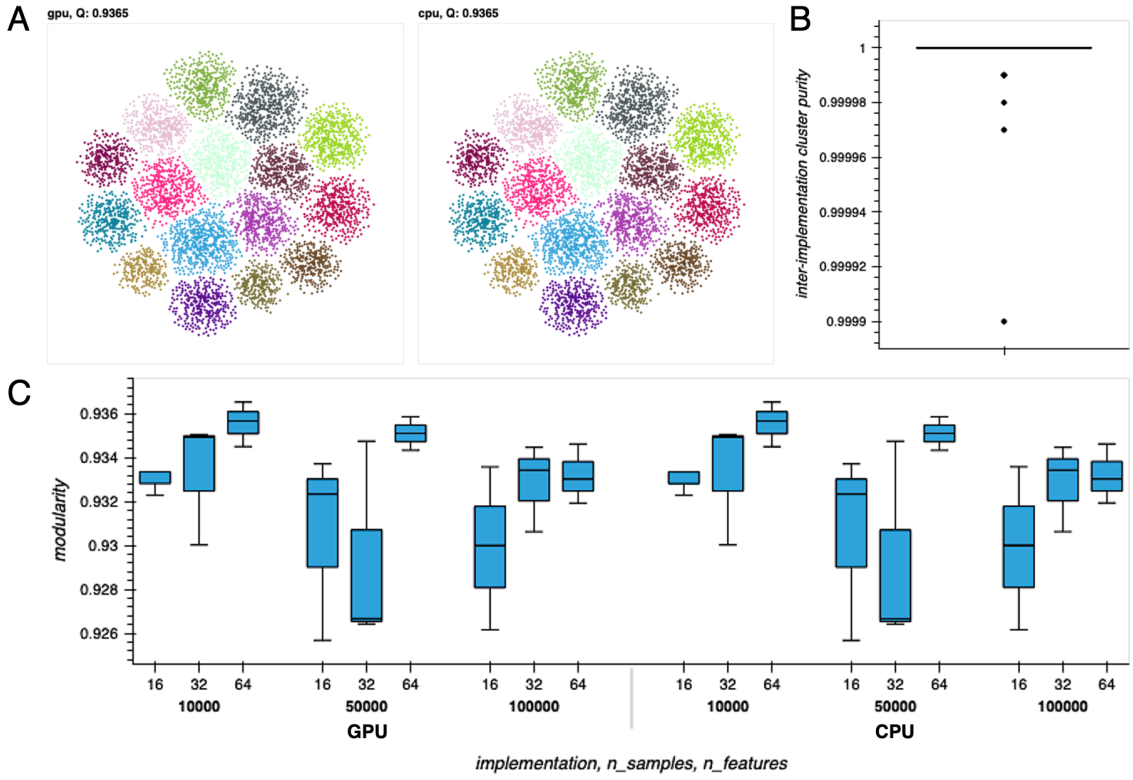


FIGURE 4.17: Benchmarking CPU and GPU implementations of PhenoGraph. (A) Scatter plots showing *t*-SNE embeddings of synthetic data, colored based on cluster label defined by either implementation. The modularity Q for each clustering result is shown above the plot. (B) Boxplot showing the distribution of GPU vs.CPU cluster purity for synthetic data. (C) Boxplots showing the distribution of clustering modularities for either implementation for the same synthetic data generated using the indicated number of samples and number of features. No significant difference was observed in modularity between implementations (ANOVA, $P = 0.999$).

Synthetic data were generated using the make_classification function from the RAPIDS library, using the settings n_classes=16, class_sep=4, and weights=dirichlet(ones(16)*5) to randomly scale the proportion of samples in each class. Examples of clustering results for

each implementation are shown in Figure 4.17A. To check for differences in clustering results, we compared the purity and modularity of clusters derived using either CPU or GPU implementations and the same parameterization ($k = 40$). Purity measures the percentage of correctly clustered objects—in this case the percentage of agreement between clustering results from each implementation—and is calculated as:

$$Purity = \frac{1}{N}\sum_{i=1}^{m} max_j|c_i \cap t_j| \tag{4.2}$$

where N is the total number of objects, m is the number of clusters from the CPU implementation, c_i is the i-th cluster from the CPU implementation, and t_j is the cluster from the GPU implementation which has the maximum number of objects from the cluster c_i. The CPU and GPU implementations yielded almost identical clustering results, i.e. $>= 0.9999$ inter-implementation cluster purity for all combinations of n_sample and n_features (Figure 4.17B). Modularity (Q) is the quantity optimized by the Louvain algorithm used by PhenoGraph to partition k-nearest neighbor graphs of cells into clusters of similar cells. As anticipated based on the tight agreement between cluster results between implementations, we detected no significant difference between CPU and GPU clustering modularity (Figure 4.17C). Very slight differences between clustering results correspond to a maximum of 0.003% of cells being mis-matched between implementations and can be attributed to the degeneracy of Louvain partitioning.

Phenotyping and metacluster annotation

Apart from the benchmarking experiment described in Figure 4.1 box (2), we use only our GPU-based implementation of PhenoGraph throughout this work. Single-cell phenotypes were defined based on single-cell mean intensity for the 35-marker CyCIF panel (Table 4.1). Prior to application of PhenoGraph, data were 99.9th-percentile normalized and arcsin transformed (cofactor = 5). Following [50], PhenoGraph was parameterized

(k=40) to over-cluster the data and detect rare cell types. PhenoGraph clustering was followed by aggregation of phenotypes into metaclusters based on hierarchical clustering of phenotype mean marker intensities and to preserve known biological variation.

Robustness of derived cell phenotypes

To assess PhenoGraph clustering robustness to sampling shift, PhenoGraph clusters were derived using random subsets of tissues of varying cardinality, from 10% to 90% of all tissues, and compared the z-scored mean marker intensities of the PhenoGraph-derived clusters from the full reference dataset and each subset using pairwise Pearson's correlation. Even with heavy subsampling, the median of the maximum correlations between matching clusters from reference-to-sample comparisons held at $\sim$0.75 (Figure 4.6A), indicating that we are defining a robust core set of cell phenotypes. Indeed, the major variation between reference and subsample clusters appeared to be sample-specific tumor cell phenotypes from the tissues held out from each subsample (Figure 4.6B-C), suggesting that PhenoGraph defines robust cell phenotypes that are shared across tissues and is capable of detecting new phenotypes as they are added to the dataset. To assess the robustness of our derived phenotypes to variability in normalization, we also simulated $\pm$20% measurement noise by multiplying the normalized cell intensity vectors for each tissue and marker by a scaling factor drawn uniformly at random from the range [0.8,1.2] and compared the z-scored mean marker intensities of the PhenoGraph-derived clusters from the clean reference and noisy datasets using pairwise Pearson's correlation. Even with these significant perturbations to the intensity profiles of cells, the median of the maximum correlations between matching clean and noisy clusters held above 0.8 (Figure 4.6D), indicating cluster robustness to differentials in preanalytical variables like tissue fixation or autofluorescence which can affect measured IF intensity across a TMA.

4.5.5 t-stochastic neighbor embedding (*t*-SNE)

To enable visualization, the full 35-feature single-cell dataset was reduced to 2 dimensions using the RAPIDS implementation of *t*-SNE [70] with default parameters except perplexity = 60. Prior to *t*-SNE processing, data were 99.9th-percentile normalized and arcsin transformed (cofactor = 5). Plots containing *t*-SNE embeddings of the full ∼1.3 million-cell dataset were created using Datashader (https://github.com/holoviz/datashader).

4.5.6 Cross-platform breast cancer cell phenotype validation

The imaging mass cytometry (IMC) dataset [50] used for our cell phenotype validation experiment was retrieved from https://zenodo.org/record/3518284. To make a fair comparison between the OHSU (CyCIF) and Basel (IMC) datasets, we independently ran PhenoGraph on each using the same parameters (k=40, Louvain partitioning) and only the overlapping features between IMC and CyCIF stain panels (Figure 4.7B). The phenotypes derived from each platform were then cross-correlated to identify inter-platform phenotype matches. Using the clustermap function from seaborn [123], the cross-correlation matrix was then hierarchically clustered with Ward linkage and used to sort the matching clusters between the two cohort heatmaps.

4.5.7 Statistical analyses

For the groupwise comparisons in Figure 4.10A, Figure 4.11C, Figure 4.12B, we first tested the assumption of homogeneity of variances using the bartlett function from the Python package SciPy [120]. When the assumption was (not) met, we made groupwise comparisons using one-way ANOVA with Tukey-HSD post-hoc test using the pairwise_tukey (one-way Welch ANOVA with Games-Howell post-hoc test using the pairwise_gameshowell) function from the Python package pingouin [118].

4.5.8 Tissue composition analyses

Epithelial differentiation heterogeneity

Cells from each core were labeled as positive for each cytokeratin if their mean intensity was greater than the normalization factor computed for that cytokeratin for that core. The plot from Figure 4.10B was generated using UpSetPlot (https://github.com/jnothman/UpSetPlot). The CK heterogeneity of each core was computed by measuring the Shannon entropy of its distribution of CK-expressing cells. The homogeneity of variances assumption was met, so comparison of the CK heterogeneity over BC subtypes was made using the pairwise_tukey function from pingouin [118].

Aggregation of immune, stromal, and tumor cell phenotypes

To enable high-level comparison of cell phenotype distribution over BC subtypes (Figure 4.11C and Figure 4.12B), cell metaclusters were aggregated into immune, stromal, and tumor groups with the following metacluster membership:

- immune = [B cell/T cell, B cell, T cell/macrophage, T cell, macrophage]
- stromal = [CD44+ endothelial, PDPN+/aSMA+ stromal, ColI+ stromal, Lamin+/Vimentin+ stromal, ColI+/PDPN+ stromal, Vimentin+ stromal, Vimentin+ endothelial]
- tumor = [Ecad+/CK low, CK5+, CK5+/CK14+, myoepithelial, CK+/H3K27me3+, epithelial low, CK14+, Ecad+/CK+, proliferating, apoptotic, H3K4me3+, HER2+/CK8+, ER+, HER2 low, HER2+/CK+, ER+/PgR+]

Cell phenotype density

The density of each of the 27 cell metaclusters in each tissue core was measured by counting the number of cells of each metacluster in the core, then dividing the count by the area

of the convex hull defined by the cell centroids of the core. Tissue cores were then hierarchically clustered based on their z-scored cell metacluster densities using the clustermap function with Ward linkage from seaborn [123]. The homogeneity of variances assumption was not met, so comparisons of immune, stromal, and tumor cell densities over BC subtypes were made using the pairwise_gameshowell function from pingouin [118].

4.5.9 Tissue architecture analyses

Tumor cell neighborhood interactions

To characterize the microenvironments of tumor cells across BC subtypes, we identified the cell metaclusters of the 10 nearest cells within 65 µm (double the median of the minimum tumor-stromal distances across all tissue cores) of each tumor cell. Tumor cells were then split based on the BC subtype of the tissue from which they were derived, and counts for each metacluster were summed over all tumor cells such that each metacluster could be represented as a proportion of the total tumor neighborhood for each BC subtype.

To measure the extent of tumor-stromal cell interactions in each tissue core, we computed their stromal mixing scores, an adaptation of a previously described cell-cell mixing score [53]. To focus on cores that had substantial stromal composition, we first selected cores which are comprised of at least 25% stromal cells and for each we defined a 10-nearest neighbor spatial graph over all cells in that core. Second, we removed edges between cells with an interaction distance greater than 65 µm. Finally, we computed the stromal mixing score for tissue core j as:

$$(\text{stromal mixing})_j = \frac{(\text{\# tumor-stromal interactions})_j}{(\text{\# stromal-stromal interactions})_j}. \tag{4.3}$$

Tumor graph centrality

To characterize tumor architecture in each tissue core, we considered the spatial interactions between tumor cells only. To account for variation in tissue core diameter (Figure 4.2A) which would affect the scale of spatial graph characteristics, we subsampled large diameter cores to be equal in size to the smallest diameter cores by only considering cells within the 300 µm-radius circle drawn about the centroid of each core. With the spatially-subsampled cores, we first construct a 4-nearest neighbor spatial graph over all tumor cells in each core. Here we use $k = 4$ rather than $k = 10$ to construct a sparser graph since we are focusing on tumor cells only. Over this graph we compute the Wasserman-Faust closeness centrality of each cell using the closeness_centrality function from the Python package networkx [41]. The Wasserman-Faust closeness centrality of cell u is computed as:

$$C_{WF}(u) = \frac{n-1}{N-1}\frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}, \tag{4.4}$$

where $d(v,u)$ is the shortest-path distance between cells v and u, n is the number of cells that can reach u, and N is the number of cells in the graph. For heatmap visualization, the distribution of tumor cell centrality for each core was max-normalized, converted into a 50-bin histogram over range = (0,1), then hierarchically clustered using the clustermap function from seaborn [123] with Jensen-Shannon distance and average linkage.

4.5.10 Plotting and visualization

Unless otherwise noted, all plots were generated using Holoviews [98] with either the Bokeh [10] or matplotlib [47] backends. Images of CyCIF-stained tissue cores were generated using napari [108].

4.5.11 Computing hardware

The GPU-accelerated PhenoGraph implementation was developed and deployed on the NVIDIA V100 GPU with 32 GB memory, but the *grapheno* Python library can be compiled to work with any NVIDIA GPU.

4.5.12 Data availability

Data and code with be made available through `zenodo.com` and `gitlab.com` upon publication.

Marker	Description	Vendor	Vendor #	Dye	Clone
CD20	B cell	Abcam	ab198941	AF488	EP459Y
CD4	T cell	Abcam	ab196147	AF647	EPR6855
CD44	cell adhesion	Abcam	ab216647	AF750	EPR1013Y
CD45	lymphocyte	Abcam	ab214437	AF750	EP322Y
CD68	macrophage	Biolegend	916104	AF555	KP1
FOXP3	regulatory T cell	Biolegend	320102	AF750	206D
GRNZB	proteolysis	Abcam	ab219803	AF750	EPR20129-217
CK5	basal cytokeratin	Abcam	ab193894	AF488	EP1601Y
CK14	basal cytokeratin	Abcam	ab212547	AF555	LL002
CK17	basal cytokeratin	Abcam	ab185032	AF488	EP1623
CK7	luminal cytokeratin	Abcam	ab185048	AF488	EPR1619Y
CK8	luminal cytokeratin	Abcam	ab192467	AF488	EP1628Y
CK19	luminal cytokeratin	Biolegend	628502	AF750	A53-B/A2
Ecad	cell adhesion	Abcam	ab201499	AF750	EP700Y
AR	hormone receptor	Sigma	06-680-AF555	AF555	polyclonal
ER	hormone receptor	Abcam	ab205851	AF647	EPR4097
PgR	hormone receptor	Abcam	ab199455	AF750	YR85
HER2	receptor tyrosine kinase	Santa Cruz	sc-33684	AF555	3B5
aSMA	myoepithelia	Santa Cruz	sc-32251	AF488	1A4
CD31	endothelia	Abcam	ab218582	AF647	EPR3094
Vimentin (Vim)	mesenchyme	CST	9854	AF488	D21H3
ColI	extracellular matrix	Abcam	ab215969	AF750	EPR7785
ColIV	extracellular matrix	ThermoFisher/eBioscience	51-9871-82	AF647	1042
LamA/C	nuclear structure	Sigma	SAB4200236	AF750	4C11
LamB1	nuclear structure	Abcam	ab194106	AF488	EPR8985(B)
LamB2	nuclear structure	Abcam	ab200427	AF647	EPR9701(B)
H3K4me3	epigenetic activation	CST	11960	AF555	C42D8
H3K27me3	epigenetic repression	CST	5499	AF488	C36B11
PDPN	migration signalling	Biolegend	916606	AF555	polyclonal
cPARP	apoptosis	CST	6894	AF555	D64E10
gH2AX	replication stress	Abcam	ab195189	AF647	EP854(2)Y
Ki67	proliferation	CST	12075	AF647	D3B5
PCNA	proliferation	CST	8580	AF488	PC10
pHH3	mitosis	CST	3465	AF488	D2C8
p-S6	translational activation	CST	3985	AF555	D57.2.2E

TABLE 4.1: Antibody panel used for CyCIF staining of tissues.

Reference	**ME markers**						
AR_Nuclei	CK5_Ring	FOXP3_Nuclei	ColIV_Ring				
aSMA_Ring	CK14_Ring	CD45_Ring	CK7_Ring	CK5_Ring	CK19_Ring		
CD20_Ring	CK14_Ring	CK7_Ring	CK5_Ring	CK19_Ring			
CD31_Ring	CK5_Ring	CK19_Ring	CK14_Ring	CK7_Ring	Ecad_Ring		
CD4_Ring	CK19_Ring	CK7_Ring	CK14_Ring	CK5_Ring	Ecad_Ring		
CD44_Ring	CK14_Ring	CK7_Ring	CK5_Ring	CK19_Ring	CD31_Ring		
CD45_Ring	CK19_Ring	CK7_Ring	CK14_Ring	CK5_Ring	CK8_Ring	CD31_Ring	
CD68_Ring	CK19_Ring	CK7_Ring	CD31_Ring	CK14_Ring			
CK14_Ring	CD31_Ring	CD68_Ring	Vim_Ring	aSMA_Ring	CD20_Ring	CD45_Ring	
CK17_Ring	CD31_Ring	CD68_Ring	Vim_Ring	ColI_Ring	CD45_Ring		
CK5_Ring	CD31_Ring	CD68_Ring	Vim_Ring	CD4_Ring	CD45_Ring		
CK19_Ring	CD68_Ring	CD4_Ring	CD31_Ring	CD45_Ring			
CK7_Ring	CD68_Ring	CD4_Ring	CD31_Ring	CD45_Ring	FOXP3_Nuclei		
CK8_Ring	CD68_Ring	CD4_Ring	CD31_Ring	CD45_Ring			
ColI_Ring	CD45_Ring	CK19_Ring	CK7_Ring	CK14_Ring	CK5_Ring		
ColIV_Ring	CK19_Ring	CK7_Ring	CK14_Ring	CK5_Ring	CD68_Ring	FOXP3_Nuclei	
cPARP_Nuclei	Ki67_Nuclei	CK5_Ring	CD31_Ring	CD68_Ring	CK14_Ring		
Ecad_Ring	CD68_Ring	CD4_Ring	CD31_Ring				
ER_Nuclei	CD68_Ring	CD4_Ring	CD31_Ring	FOXP3_Nuclei			
FOXP3_Nuclei	CK19_Ring	CK7_Ring	CK5_Ring	CK14_Ring	CD31_Ring	CK8_Ring	
gH2AX_Nuclei	CK8_Ring	CK14_Ring	CK5_Ring	CK7_Ring			
GRNZB_Ring	CK19_Ring	CK7_Ring	CK5_Ring	CD31_Ring	CK14_Ring	aSMA_Ring	
H3K27me3_Nuclei	CD31_Ring	CD68_Ring	CD44_Ring				
H3K4me3_Nuclei	CD31_Ring	CD68_Ring	CD44_Ring	CK19_Ring			
HER2_Ring	CD68_Ring	CD44_Ring	CD31_Ring	Vim_Ring	CD4_Ring		
Ki67_Nuclei	cPARP_Nuclei						
LamAC_Nuclei	CD68_Ring	CD44_Ring	CK19_Ring	CD45_Ring			
LamB1_Nuclei	CD68_Ring	CD44_Ring	CK19_Ring	CD45_Ring	CK14_Ring	CK7_Ring	CD31_Ring
LamB2_Nuclei	CD68_Ring	CD44_Ring	CK19_Ring	CD31_Ring	CK7_Ring	CK14_Ring	
PCNA_Nuclei	CK7_Ring	CD45_Ring	CD31_Ring	CD68_Ring	CK14_Ring	LamB2_Nuclei	
PDPN_Ring	CK19_Ring	CK7_Ring	CK14_Ring	CK5_Ring	CD31_Ring	CD68_Ring	
PgR_Nuclei	CD68_Ring	CD4_Ring	CD31_Ring	CD20_Ring	aSMA_Ring	Vim_Ring	
pHH3_Nuclei	CD31_Ring	CK5_Ring	CK19_Ring	CK14_Ring	GRNZB_Ring		
p-S6_Ring	CK19_Ring	CK5_Ring	CK7_Ring	CK14_Ring			
Vim_Ring	CK19_Ring	CK7_Ring	CD68_Ring	CD45_Ring	Ecad_Ring		

TABLE 4.2: Putative reference and mutually-exclusive (ME) marker pairs used for RESTORE normalization of cell mean intensities. Each marker name indicates from which compartment its mean intensity was extracted. "Ring" indicates that a marker's intensity was extracted from the ring-shaped cytoplasmic segmentation masks derived by subtracting the "Nuclei" segmentation masks from "Cell" segmentation masks.

Chapter 5

Conclusion

5.1 Thesis summary

Without question, MTI is playing and will continue to play an essential role in the advancement of cancer systems biology and precision medicine, and the work we present herein supports that advancement. The key contributions of this dissertation are:

1. The development of a virtual staining paradigm which enables fast and accurate prediction of protein distribution in digitized histology slides, and a framework for quantitative cohort selection based on histological features.

2. The extension of the virtual staining paradigm into the third dimension through application on a 3D tumor atlas, and a framework for quantitative ROI selection based on the integration of histological and molecular features.

3. The development of a GPU-accelerated single-cell analysis framework which enables reproducible, scalable, and robust analysis of data across MTI platforms, both in terms of cell composition and spatial arrangement in associated tissues.

Toward the original goal of offsetting some of the challenges of MTI, the work presented in this dissertation is preliminary and will require validation in larger cohorts before consideration of use in research or clinical settings.

5.2 Significance and commercial potential

5.2.1 Significance of virtual staining with SHIFT

The virtual staining paradigm we present through SHIFT could make an impact in both research and clinical settings, where the quality of patient care is limited by the efficiency and accessibility of imaging assessments. According to the Centers for Medicare and Medicaid Services—the federal agencies which regulate all clinical laboratory testing on human specimens through the Clinical Laboratory Improvement Amendments (CLIA)—there are 260,000 laboratory entities in compliance with CLIA in operation today in the U.S. [1], many of which rely on IHC/IF to better understand the pathology of their specimens. The Immunostains Laboratory at Mayo Clinic alone is responsible for performing ∼198,000 IHC tests each year [2], each of which can cost hundreds of dollars and take up to a day to process before being interpreted by a pathologist. Considering projections that the U.S. will be 5,700 pathologists short of the projected need of 20,000 by 2030 [94], SHIFT could help to overcome this cost and bottleneck by allowing fewer pathologists to handle more cases.

The speed with which SHIFT is able to deliver accurate virtual IF information make it an appealing preliminary, auxiliary, or alternative technique to traditional IF or CyCIF in clinical trials which strive for near-real-time monitoring of patient disease and response to therapy. For example, the metastatic BC-focused Serial Measurements of Molecular and Architectural Responses to Therapy (SMMART) trials currently being conducted at Oregon Health & Science University (OHSU) are integrating measurements from a battery of genomic, proteomic, and imaging assays, including CyCIF, to guide treatment on as fine a timescale as possible [51]. Currently, it takes four weeks from time of biopsy to analyze, interpret, and report on a patient's disease status, with CyCIF being one of the most time-intensive assays in the battery with a lead time of at least three weeks. Once we determine which stains can be accurately inferred, the virtual IF staining provided by

SHIFT could substitute for the real stains. If SHIFT were integrated into the SMMART via a computational pipeline like MCMICRO [101], the lead time for CyCIF could be reduced by at least one day—the time required for one cycle of CyCIF—for every three virtual IF stains that SHIFT is able to deliver. This would be a major step toward the SMMART trial goal of shortening the total analytics lead time to ten days from time of biopsy.

Similarly, SHIFT could enable underserved medical communities currently lacking expensive IHC/IF imaging technologies to engage with IF data to improve diagnostic workflows, providing savings in both time and money while improving patient care. We believe our approach will be particularly useful in the developing world, where access to IHC/IF imaging and expertise is sparse. For example, the Moi Teaching and Referral Hospital (MTRH) is one of Kenya's two tertiary care facilities and serves approximately half of Kenya's population, or 20 million people. Alarmingly, a 2016 assessment of the MTRH found that its Department of Pathology was staffed by four general pathologists and six laboratory technicians and that only one member of the Department of Immunology had extensive prior experience with IHC [84]. Since H&E staining is part of the diagnostic routine at MTRH, an automated tool like SHIFT could help with triaging the disproportionately large population which its clinical departments serve.

Concretely, any laboratory capable of producing digital images of H&E-stained tissue sections could benefit from the virtual staining paradigm of SHIFT. As such, we see SHIFT as an opportunity to simultaneously democratize and economize advanced imaging technologies in histopathology workflows.

5.2.2 Commercial potential of virtual staining with SHIFT

We envision that SHIFT could be delivered to users via a web-service targeted towards clinical pathologists, histologists, oncologists, clinical lab directors as well as personnel involved in antibody validation. The digital nature of the SHIFT framework makes it highly amenable as a web-based service. For a recurring subscription fee, users would

have access to a web interface that connects them to SHIFT models. From their personal workstation, a user will be able to upload a digital H&E image and select one or more desired markers of interest for generative inference. The digital H&E image would then be uploaded to a graphical processing unit (GPU)-equipped server that will automatically execute necessary pre-processing of the uploaded image (e.g. rescaling, background removal, and normalization) and apply the requested SHIFT models to infer the markers of interest. The generated images will then be returned back to the user through download via their internet connection. High-resolution scans of H&E images are variable in memory footprint, but average around 1 gigabyte per image, which necessitates a reasonably efficient internet connection to transfer through the web. An alternative commercialization strategy would allow download of the SHIFT models to the user's local computer through a desktop or mobile application. This approach would eliminate the need for uploading H&E images through the web, but would increase the inference time of the model if the user's workstation is not equipped with one or more GPUs. The rise of low-cost high-quality smartphone cameras and edge devices that are capable of DL model inference for pathology applications [67, 31] make this delivery approach attractive since it would allow SHIFT to be used in areas with limited resources or unreliable internet access. SHIFT models could also be embedded into the software of automated slide stainers and scanners to augment a pathologist's decision about which real stains to apply to a sample (Figure 5.1).

Our approach to generating spatial estimations of markers is highly cost-effective relative to established imaging technologies. For this reason, we will be able to provide SHIFT to users at a very competitive price. The cost of receiving a single H&E image over the internet, running inference on a pre-trained SHIFT model and transmitting the virtual IF results is small relative to the cost of traditional IF imaging technologies, staining antibodies, and requisite technical training. We anticipate that the savings afforded by SHIFT will occupy a highly competitive position in the marketplace by reducing the cost of generating IF images by several orders of magnitude. A modest 0.1% market penetration would

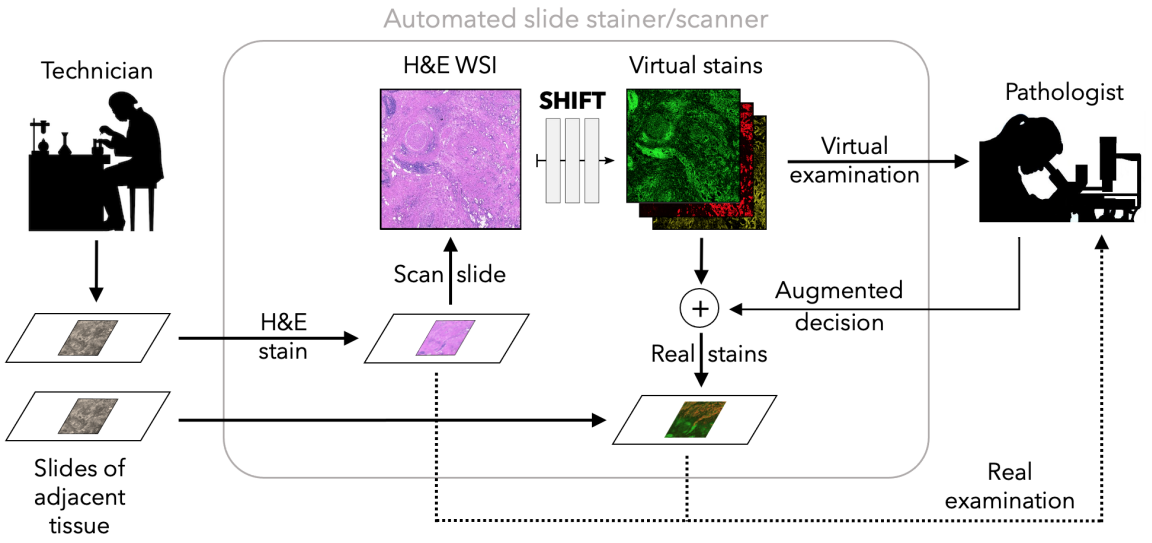


FIGURE 5.1: Theoretical application for SHIFT in stain prioritization and automation.

enable entry into approximately 260 CLIA labs in the USA which are expected to process on average 10 samples a week, or approximately ($10 \times 52 \times 260 =$) 135,200 runs per year. As typical multiplex IHC staining costs approximately $500 per run, and IF imaging costs approximately $3000 per run (depending on antibody stains, etc.) we project current technologies to serve this market to cost somewhere between $67 million and $270 million per year. At OHSU, individual samples can cost up to $25,000/tissue sample using a state-of-the-art MultiOmyx platform provided by GE.

Because the SHIFT framework is entirely computational, the costs required to generate SHIFT images are only those necessary to transmit the images and run them through a trained learning model. The cost of generating SHIFT images is therefore limited to the computational costs of running a server and GPU with an internet connection, which would cost approximately $300/month operating at full capacity. We could therefore afford to offer SHIFT for a fraction of the cost of typical tissue staining. To undercut the cost of typical imaging by a factor of 10, we would require between $50 and $300 dollars per inference run, depending on the needs of the user. This revenue, with an additional annual subscription fee, will provide necessary funding to generate additional H&E/CyCIF

data with which to train additional models, react to the needs of the user, and reinforce the utility of computational inference across diverse medical imaging market segments.

5.3 Ultimate vision of this dissertation

The sooner a pathologist understands a cancer, the sooner they can treat it. Rapid determination of disease subtype and phenotypic response to therapy are required to deliver an appropriate treatment and inform therapeutic strategy in the context of disease evolution and drug resistance. Virtual staining and other GPU-accelerated analytics could revolutionize these pathological determinations by virtue of being faster, cheaper, and in some cases more reliable than traditional methods, but they will each require thorough validation. We hope that the work presented in this dissertation will serve as a framework for the early development and validation of the next generation of biomarkers in cancer systems biology.

Bibliography

[1] URL: https://www.cms.gov/regulations-and-guidance/legislation/clia?redirect=/clia/.

[2] URL: https://www.mayoclinic.org/departments-centers/laboratory-medicine-pathology/overview/specialty-groups/anatomic-pathology/services/immunostains-laboratory.

[3] H. R. Ali, S. E. Glont, F. M. Blows, E. Provenzano, S. J. Dawson, B. Liu, L. Hiller, J. Dunn, C. J. Poole, S. Bowden, and et al. "PD-L1 protein expression in breast cancer is rare, enriched in basal-like tumours and associated with infiltrating lymphocytes". In: *Annals of Oncology* 26.7 (2015), 1488–1493. ISSN: 0923-7534. DOI: 10.1093/annonc/mdv192.

[4] Michael Angelo, Sean C. Bendall, Rachel Finck, Matthew B. Hale, Chuck Hitzman, Alexander D. Borowsky, Richard M. Levenson, John B. Lowe, Scot D. Liu, Shuchun Zhao, and et al. "Multiplexed ion beam imaging (MIBI) of human breast tumors". In: *Nature medicine* 20.4 (2014), 436–442. ISSN: 1078-8956. DOI: 10.1038/nm.3488.

[5] Ricard Argelaguet, Anna S. E. Cuomo, Oliver Stegle, and John C. Marioni. "Computational principles and challenges in single-cell data integration". In: *Nature Biotechnology* (2021). ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-021-00895-7. URL: http://www.nature.com/articles/s41587-021-00895-7.

[6] Aharon Azulay and Yair Weiss. "Why do deep convolutional networks generalize so poorly to small image transformations?" In: *arXiv:1805.12177 [cs]* (2018). arXiv: 1805.12177. URL: http://arxiv.org/abs/1805.12177.

[7] Vivian Barak, Helena Goike, Katja W. Panaretakis, and Roland Einarsson. "Clinical utility of cytokeratins as tumor markers". In: *Clinical Biochemistry* 37.7 (2004), 529–540. ISSN: 00099120. DOI: 10.1016/j.clinbiochem.2004.05.009.

[8] Giampaolo Bianchini, Justin M. Balko, Ingrid A. Mayer, Melinda E. Sanders, and Luca Gianni. "Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease". In: *Nature Reviews Clinical Oncology* 13.11 (2016), 674–690. ISSN: 1759-4774, 1759-4782. DOI: 10.1038/nrclinonc.2016.66.

[9] Natalia Y. Bilenko and Jack L. Gallant. "Pyrcca: Regularized Kernel Canonical Correlation Analysis in Python and Its Applications to Neuroimaging". In: *Frontiers in Neuroinformatics* 10 (2016). DOI: 10.3389/fninf.2016.00049.

[10] Bokeh Development Team. *Bokeh: Python library for interactive visualization*. 2020. URL: https://bokeh.org/.

[11] Heather M. Brechbuhl, Jessica Finlay-Schultz, Tomomi M. Yamamoto, Austin E. Gillen, Diana M. Cittelly, Aik-Choon Tan, Sharon B. Sams, Manoj M. Pillai, Anthony D. Elias, William A. Robinson, and et al. "Fibroblast Subtypes Regulate Responsiveness of Luminal Breast Cancer to Estrogen". In: *Clinical Cancer Research* 23.7 (2017), 1710–1721. ISSN: 1078-0432, 1557-3265. DOI: 10.1158/1078-0432.CCR-15-2851.

[12] Serdar E. Bulun, Dong Chen, Irene Moy, David C Brooks, and Hong Zhao. "Aromatase, breast cancer and obesity: a complex interaction". In: *Trends in endocrinology and metabolism: TEM* 23.2 (2012), 83–89. ISSN: 1043-2760. DOI: 10.1016/j.tem.2011.10.003.

[13] Juan C. Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, Joseph D. Barry, Harmanjit Singh Bansal, Oren Kraus, and et al. "Data-analysis strategies for image-based cell profiling". In: *Nature Methods* 14.99 (2017), 849–863. ISSN: 1548-7105. DOI: 10.1038/nmeth.4397.

[14] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". In: *Nature Medicine* (2019), p. 1. ISSN: 1546-170X. DOI: 10.1038/s41591-019-0508-1.

[15] Raúl Catena, Alaz Özcan, Laura Kütt, Alex Plüss, IMAXT Consortium, Peter Schraml, Holger Moch, and Bernd Bodenmiller. *Highly multiplexed molecular and cellular mapping of breast cancer tissue in three dimensions using mass tomography*. 2020. DOI: 10.1101/2020.05.24.113571. URL: http://biorxiv.org/lookup/doi/10.1101/2020.05.24.113571.

[16] Y. H. Chang, G. Thibault, O. Madin, V. Azimi, C. Meyers, B. Johnson, J. Link, A. Margolin, and J. W. Gray. "Deep learning based Nucleus Classification in pancreas histological images". In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2017, 672–675. DOI: 10.1109/EMBC.2017.8036914.

[17] Young Hwan Chang, Koei Chin, Guillaume Thibault, Jennifer Eng, **Erik A. Burlingame**, and Joe W. Gray. "RESTORE: Robust intEnSiTy nORmalization mEthod for multiplexed imaging". In: *Communications Biology* 3.11 (2020), 1–9. ISSN: 2399-3642. DOI: 10.1038/s42003-020-0828-1.

[18] Young Hwan Chang, **Erik A. Burlingame**, Geoffrey Schau, and Joe W. Gray. *Translation of images of stained biological material*. 2020. URL: https:

//patents.google.com/patent/WO2020142461A1/en?q=erik+burlingame&inventor=Erik+BURLINGAME.

[19] Po-Hsuan Cameron Chen, Krishna Gadepalli, Robert MacDonald, Yun Liu, Kunal Nagpal, Timo Kohlberger, Jeffrey Dean, Greg S. Corrado, Jason D. Hipp, and Martin C. Stumpe. "Microscope 2.0: An Augmented Reality Microscope with Real-time Artificial Intelligence Integration". In: *arXiv:1812.00825 [cs]* (2018). arXiv: 1812.00825. URL: http://arxiv.org/abs/1812.00825.

[20] Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. "Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis". In: *arXiv:1912.08937 [cs, q-bio]* (2020). arXiv: 1912.08937. URL: http://arxiv.org/abs/1912.08937.

[21] Xiaohua Chen, Hanyang Hu, Lin He, Xueyuan Yu, Xiangyu Liu, Rong Zhong, and Maoguo Shu. "A novel subtype classification and risk of breast cancer by histone modification profiling". In: *Breast Cancer Research and Treatment* 157.2 (2016), 267–279. ISSN: 1573-7217. DOI: 10.1007/s10549-016-3826-8.

[22] Eric M Christiansen, Samuel J Yang, D Michael Ando, Lee L Rubin, Philip Nelson, Steven Finkbeiner, Eric M Christiansen, Samuel J Yang, D Michael Ando, Ashkan Javaherian, and et al. "In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images Resource In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images". In: *Cell* 173.3 (2018), 792–803.e19. DOI: 10.1016/j.cell.2018.03.040.

[23] N. C. F. Codella, Q. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith. "Deep learning ensembles for melanoma recognition in dermoscopy images". In: *IBM Journal of Research and Development* 61.4/5 (2017), 5:1–5:15. ISSN: 0018-8646. DOI: 10.1147/JRD.2017.2708299.

[24] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho. "End-to-End Adversarial Retinal Image Synthesis". In: *IEEE Transactions on Medical Imaging* 37.3 (2018), 781–791. ISSN: 0278-0062. DOI: 10.1109/TMI.2017.2759102.

[25] Heather D. Couture, Lindsay A. Williams, Joseph Geradts, Sarah J. Nyante, Ebonee N. Butler, J. S. Marron, Charles M. Perou, Melissa A. Troester, and Marc Niethammer. "Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype". In: *npj Breast Cancer* 4.1 (2018), p. 30. ISSN: 2374-4677. DOI: 10.1038/s41523-018-0079-1.

[26] Dask Development Team. *Dask: Library for dynamic task scheduling*. 2016. URL: https://dask.org.

[27] Matthew S. Dietz, Thomas L. Sutton, Brett S. Walker, Charles E. Gast, Luai Zarour, Sidharth K. Sengupta, John R. Swain, Jennifer Eng, Michael Parappilly, Kristen Limbach, and et al. "Relevance of Circulating Hybrid Cells as a Non-Invasive Biomarker for Myriad Solid Tumors". In: *bioRxiv* (2021), p. 2021.03.11.434896. DOI: 10.1101/2021.03.11.434896.

[28] Mitch Dowsett, Craig Allred, Jill Knox, Emma Quinn, Janine Salter, Chris Wale, Jack Cuzick, Joan Houghton, Norman Williams, Elizabeth Mallon, and et al. "Relationship Between Quantitative Estrogen and Progesterone Receptor Expression and Human Epidermal Growth Factor Receptor 2 (HER-2) Status With Recurrence in the Arimidex, Tamoxifen, Alone or in Combination Trial". In: *Journal of Clinical Oncology* 26.7 (2008), 1059–1065. ISSN: 0732-183X, 1527-7755. DOI: 10.1200/JCO.2007.12.9437.

[29] Jeyapradha Duraiyan, Rajeshwar Govindarajan, Karunakaran Kaliyappan, and Murugesan Palanisamy. "Applications of immunohistochemistry". In: *Journal of*

Pharmacy & Bioallied Sciences 4.Suppl 2 (2012), S307–S309. ISSN: 0976-4879. DOI: 10.4103/0975-7406.100281.

[30] Jennifer Eng, Guillaume Thibault, Shiuh-Wen Luoh, Joe W. Gray, Young Hwan Chang, and Koei Chin. "Cyclic Multiplexed-Immunofluorescence (cmIF), a Highly Multiplexed Method for Single-Cell Analysis". In: *Biomarkers for Immunotherapy of Cancer: Methods and Protocols*. Methods in Molecular Biology. Springer, 2020, 521–562. ISBN: 978-1-4939-9773-2.

[31] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.76397639 (2017), 115–118. ISSN: 1476-4687. DOI: 10.1038/nature21056.

[32] Henrik Failmezger, Sathya Muralidhar, Antonio Rullan, Carlos E. de Andrea, Erik Sahai, and Yinyin Yuan. "Topological Tumor Graphs: A Graph-Based Spatial Model to Infer Stromal Recruitment for Immunosuppression in Melanoma Histology". In: *Cancer Research* 80.5 (2020), 1199–1209. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-19-2268.

[33] Michael Gadermayr, Laxmi Gupta, Barbara M. Klinkhammer, Peter Boor, and Dorit Merhof. "Unsupervisedly Training GANs for Segmenting Digital Pathology with Automatically Generated Annotations". In: *International Conference on Medical Imaging with Deep Learning*. PMLR, 2019, 175–184. URL: http://proceedings.mlr.press/v102/gadermayr19a.html.

[34] Soizic Garaud, Laurence Buisseret, Cinzia Solinas, Chunyan Gu-Trantien, Alexandre de Wind, Gert Van den Eynden, Celine Naveaux, Jean-Nicolas Lodewyckx, Anaïs Boisson, Hughes Duvillier, and et al. "Tumor-infiltrating B cells signal functional humoral immune responses in breast cancer". In: *JCI Insight* 4.18 (2019). ISSN:

2379-3708. DOI: 10.1172/jci.insight.129641. URL: https://insight.jci.org/articles/view/129641.

[35] Michael J. Gerdes, Christopher J. Sevinsky, Anup Sood, Sudeshna Adak, Musodiq O. Bello, Alexander Bordwell, Ali Can, Alex Corwin, Sean Dinn, Robert J. Filkins, and et al. "Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue". In: *Proceedings of the National Academy of Sciences* 110.29 (2013), 11982–11987. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1300136110.

[36] Charlotte Giesen, Hao A. O. Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J. Schüffler, Daniel Grolimund, Joachim M. Buhmann, Simone Brandt, and et al. "Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry". In: *Nature Methods* 11.44 (2014), 417–422. ISSN: 1548-7105. DOI: 10.1038/nmeth.2869.

[37] Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. "Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging". In: *Cell* 174.4 (2018), 968–981.e15. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.07.010.

[38] Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. "Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging". In: *Cell* 174.4 (2018), 968–981.e15. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.07.010.

[39] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks". In: *arXiv:1406.2661 [cs, stat]* (2014). arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661.

[40] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, and B. Yener. "Histopathological Image Analysis: A Review". In: *IEEE Reviews in Biomedical Engineering* 2 (2009), 147–171. ISSN: 1937-3333, 1941-1189. DOI: 10.1109/RBME.2009.2034865.

[41] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11 –15.

[42] Bisong Haupt, Jae Y. Ro, and Mary R. Schwartz. "Basal-like Breast Carcinoma: A Phenotypically Distinct Entity". In: *Archives of Pathology and Laboratory Medicine* 134.1 (2010), 130–133. ISSN: 0003-9985. DOI: 10.1043/1543-2165-134.1.130.

[43] Narayan Hegde, Jason D. Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J. Cai, Mahul B. Amin, Craig H. Mermel, and et al. "Similar image search for histopathology: SMILY". In: *npj Digital Medicine* 2.1 (2019), p. 56. ISSN: 2398-6352. DOI: 10.1038/s41746-019-0131-z.

[44] Caitlin A. Hester, Mathew M. Augustine, Michael A. Choti, John C. Mansour, Rebecca M. Minter, Patricio M. Polanco, Matthew R. Porembka, Sam C. Wang, and Adam C. Yopp. "Comparative outcomes of adenosquamous carcinoma of the pancreas: An analysis of the National Cancer Database". In: *Journal of Surgical Oncology* 118.1 (2018), 21–30. ISSN: 1096-9098. DOI: 10.1002/jso.25112.

[45] Dorit S. Hochbaum and Anu Pathria. "Analysis of the greedy approach in problems of maximum k-coverage". In: *Naval Research Logistics (NRL)* 45.6 (1998), 615–627. ISSN: 1520-6750. DOI: 10.1002/(SICI)1520-6750(199809)45:6<615::AID-NAV5>3.0.CO;2-5.

[46] Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, and Qingling Sun. "Deep learning for image-based cancer detection and diagnosis-A survey". In: *Pattern Recognition* 83 (2018), 134–149. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2018.05.014.

[47] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[48] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *arXiv:1611.07004 [cs]* (2016). arXiv: 1611.07004. URL: http://arxiv.org/abs/1611.07004.

[49] Christopher R. Jackson, Aravindhan Sriharan, and Louis J. Vaickus. "A machine learning algorithm for simulating immunohistochemistry: development of SOX10 virtual IHC and evaluation on primarily melanocytic neoplasms". In: *Modern Pathology* 33.99 (2020), 1638–1648. ISSN: 1530-0285. DOI: 10.1038/s41379-020-0526-z.

[50] Hartland W. Jackson, Jana R. Fischer, Vito R. T. Zanotelli, H. Raza Ali, Robert Mechera, Savas D. Soysal, Holger Moch, Simone Muenst, Zsuzsanna Varga, Walter P. Weber, and et al. "The single-cell pathology landscape of breast cancer". In: *Nature* 578.7796 (2020), 615–620. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1876-x.

[51] Brett E. Johnson, Allison L. Creason, Jayne M. Stommel, Jamie Keck, Swapnil Parmar, Courtney B. Betts, Aurora Blucher, Christopher Boniface, Elmar Bucher, **Erik A. Burlingame**, and et al. "An Integrated Clinical, Omic, and Image Atlas of an Evolving Metastatic Breast Cancer". In: *bioRxiv* (2020), p. 2020.12.03.408500. DOI: 10.1101/2020.12.03.408500.

[52] R. Kannan, S. Vempala, and A. Veta. "On clusterings-good, bad and spectral". In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*. 2000, 367–377. DOI: 10.1109/SFCS.2000.892125.

[53] Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, Allison Kurian, David Van Valen, Robert West, and et al. "A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging". In: *Cell* 174.6 (2018), 1373–1387.e19. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2018.08.039.

[54] Ashley Kiemen, Alicia M. Braxton, Mia P. Grahn, Kyu Sang Han, Jaanvi Mahesh Babu, Rebecca Reichel, Falone Amoa, Seung-Mo Hong, Toby C. Cornish, Elizabeth D. Thompson, and et al. "In situ characterization of the 3D microanatomy of the pancreas and pancreatic cancer at single cell resolution". In: (2020). DOI: 10.1101/2020.12.08.416909. URL: http://biorxiv.org/lookup/doi/10.1101/2020.12.08.416909.

[55] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv:1412.6980 [cs]* (2017). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

[56] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *arXiv:1312.6114 [cs, stat]* (2013). arXiv: 1312.6114. URL: http://arxiv.org/abs/1312.6114.

[57] Leeor Langer, Yoav Binenbaum, Leonid Gugel, Moran Amit, Ziv Gil, and Shai Dekel. "Computer-aided diagnostics in digital pathology: automated evaluation of early-phase pancreatic cancer in mice". In: *International Journal of Computer Assisted Radiology and Surgery* 10.7 (2015), 1043–1054. ISSN: 1861-6429. DOI: 10.1007/s11548-014-1122-9.

[58] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, and et al. "Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis". In: *Cell* 162.1 (2015), 184–197. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.05.047.

[59] Jia-Ren Lin, Mohammad Fallahi-Sichani, Jia-Yun Chen, and Peter K. Sorger. "Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging". In: *Current protocols in chemical biology* 8.4 (2016), 251–264. ISSN: 2160-4762. DOI: 10.1002/cpch.14.

[60] Jia-Ren Lin, Benjamin Izar, Shu Wang, Clarence Yapp, Shaolin Mei, Parin M Shah, Sandro Santagata, and Peter K Sorger. "Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes". In: *eLife* 7 (2018). Ed. by Arup K Chakraborty, Arjun Raj, Carsten Marr, and Péter Horváth, e31657. ISSN: 2050-084X. DOI: 10.7554/eLife.31657.

[61] Jia-Ren Lin, Shu Wang, Shannon Coy, Madison Tyler, Clarence Yapp, Yu-An Chen, Cody N. Heiser, Ken S. Lau, Sandro Santagata, and Peter K. Sorger. "Multiplexed 3D atlas of state transitions and immune interactions in colorectal cancer". In: (2021). DOI: 10.1101/2021.03.31.437984. URL: http://biorxiv.org/lookup/doi/10.1101/2021.03.31.437984.

[62] Jonathan T. C. Liu, Adam K. Glaser, Kaustav Bera, Lawrence D. True, Nicholas P. Reder, Kevin W. Eliceiri, and Anant Madabhushi. "Harnessing non-destructive 3D pathology". In: *Nature Biomedical Engineering* 5.33 (2021), 203–218. ISSN: 2157-846X. DOI: 10.1038/s41551-020-00681-x.

[63] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. "Unsupervised Image-to-Image Translation Networks". In: *arXiv:1703.00848 [cs]* (2018). arXiv: 1703.00848. URL: http://arxiv.org/abs/1703.00848.

[64] Peng Liu, Silvia Liu, Yusi Fang, Xiangning Xue, Jian Zou, George Tseng, and Liza Konnikova. "Recent Advances in Computer-Assisted Algorithms for Cell Subtype Identification of Cytometry Data". In: *Frontiers in Cell and Developmental Biology* 8 (2020). ISSN: 2296-634X. DOI: 10.3389/fcell.2020.00234. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7198724/.

[65] Xiao Liu, Weichen Song, Brandon Y. Wong, Ting Zhang, Shunying Yu, Guan Ning Lin, and Xianting Ding. "A comparison framework and guideline of clustering methods for mass cytometry data". In: *Genome Biology* 20.1 (2019), p. 297. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1917-7.

[66] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q. Nelson, Greg S. Corrado, and et al. "Detecting Cancer Metastases on Gigapixel Pathology Images". In: *arXiv:1703.02442 [cs]* (2017). arXiv: 1703.02442. URL: http://arxiv.org/abs/1703.02442.

[67] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. "Data-efficient and weakly supervised computational pathology on whole-slide images". In: *Nature Biomedical Engineering* (2021), 1–16. ISSN: 2157-846X. DOI: 10.1038/s41551-020-00682-w.

[68] Steve Lu, Julie E. Stein, David L. Rimm, Daphne W. Wang, J. Michael Bell, Douglas B. Johnson, Jeffrey A. Sosman, Kurt A. Schalper, Robert A. Anders, Hao Wang, and et al. "Comparison of Biomarker Modalities for Predicting Response to PD-1/PD-L1 Checkpoint Blockade: A Systematic Review and Meta-analysis". In: *JAMA Oncology* (2019). DOI: 10.1001/jamaoncol.2019.1549. URL: https://jamanetwork.com/journals/jamaoncology/fullarticle/2738418.

[69] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), 2579–2605. ISSN: ISSN 1533-7928.

[70] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), 2579–2605. ISSN: ISSN 1533-7928.

[71] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. "A method for normalizing histology slides for quantitative analysis". In: *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009* (2009), 1107–1110. ISSN: 9781424439324. DOI: 10.1109/ISBI.2009.5193250.

[72] Anant Madabhushi and George Lee. "Image analysis and machine learning in digital pathology: Challenges and opportunities". In: *Medical Image Analysis*. 20th anniversary of the Medical Image Analysis journal (MedIA) 33 (2016), 170–175. ISSN: 1361-8415. DOI: 10.1016/j.media.2016.06.037.

[73] MATLAB. *9.3.0.713579 (R2017b)*. The MathWorks Inc., 2017.

[74] B. W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2 (1975), 442–451. ISSN: 0005-2795. DOI: 10.1016/0005-2795(75)90109-9.

[75] Christopher R. Merritt, Giang T. Ong, Sarah Church, Kristi Barker, Gary Geiss, Margaret Hoang, Jaemyeong Jung, Yan Liang, Jill McKay-Fleisch, Karen Nguyen, and et al. "High multiplex, digital spatial profiling of proteins and RNA in fixed tissue using genomic detection methods". In: *bioRxiv* (2019). DOI: 10.1101/559021. URL: http://biorxiv.org/lookup/doi/10.1101/559021.

[76] Mehdi Mirza and Simon Osindero. "Conditional Generative Adversarial Nets". In: *arXiv:1411.1784 [cs, stat]* (2014). arXiv: 1411.1784. URL: http://arxiv.org/abs/1411.1784.

[77] S. Naik, Scott Doyle, Anant Madabhushi, John E Tomaszewski, and Michael D Feldman. "Automated Gland Segmentation and Gleason Grading of Prostate Histology

by Integrating Low-, High-level and Domain Specific Information". In: *Workshop on Microscopic Image Analysis with Applications in Biology*. 2007.

[78] Shivang Naik, Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology". In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2008, 284–287. DOI: 10.1109/ISBI.2008.4540988.

[79] Ryosuke Okuta, Yuya Unno, Daisuke Nishino, Shohei Hido, and Crissman Loomis. "CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations". In: *Proceedings of Workshop on Machine Learning Systems (LearningSys) in the Thirty-first Annual Conference on Neural Information Processing Systems (NeurIPS)* (2017), p. 7.

[80] Sebastian Otalora, Roger Schaer, Manfredo Atzori, Oscar Alfonso Jimenez del Toro, and Henning Muller. "Deep learning based retrieval system for gigapixel histopathology cases and open access literature". In: *bioRxiv* (2018). DOI: 10.1101/408237. URL: http://biorxiv.org/lookup/doi/10.1101/408237.

[81] Chawin Ounkomol, Sharmishtaa Seshamani, Mary M. Maleckar, Forrest Collman, and Gregory R. Johnson. "Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy". In: *Nature Methods* 15.11 (2018), 917–920. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-018-0111-2.

[82] Wei Ouyang, Andrey Aristov, Mickaël Lelek, Xian Hao, and Christophe Zimmer. "Deep learning massively accelerates super-resolution localization microscopy". In: *Nature Biotechnology* 36.5 (2018), 460–468. ISSN: 1546-1696. DOI: 10.1038/nbt.4106.

[83] Feng Pan, Wei Wang, A.K.H. Tung, and Jiong Yang. "Finding Representative Set from Massive Data". In: *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 2005, 338–345. ISBN: 978-0-7695-2278-4. DOI: 10.1109/ICDM.2005.69. URL: http://ieeexplore.ieee.org/document/1565697/.

[84] Kirtika Patel, R Matthew Strother, Francis Ndiangui, David Chumba, William Jacobson, Cecelia Dodson, Murray B Resnic, Randall W Strate, and James W Smith. "Development of immunohistochemistry services for cancer care in western Kenya: Implications for low- and middle-income countries." In: *African journal of laboratory medicine* 5.1 (2016), 187–187. DOI: 10.4102/ajlm.v5i1.187.

[85] Yash Patel, Srikar Appalaraju, and R. Manmatha. "Deep Perceptual Compression". In: *arXiv:1907.08310 [cs, eess]* (2019). arXiv: 1907.08310. URL: http://arxiv.org/abs/1907.08310.

[86] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), 2825-2830. ISSN: 1533-7928.

[87] Aleix Prat, Estela Pineda, Barbara Adamo, Patricia Galván, Aranzazu Fernández, Lydia Gaba, Marc Díez, Margarita Viladot, Ana Arance, and Montserrat Muñoz. "Clinical implications of the intrinsic molecular subtypes of breast cancer". In: *The Breast*. 14th St.Gallen International Breast Cancer Conference – Proceedings Book 24 (2015), S26–S35. ISSN: 0960-9776. DOI: 10.1016/j.breast.2015.07.008.

[88] Sebastian Raschka, Joshua Patterson, and Corey Nolet. "Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence". In: *arXiv preprint arXiv:2002.04803* (2020).

[89] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. "Color transfer between images". In: *IEEE Computer Graphics and Applications* 21.5 (2001), 34–41. ISSN: 1558-1756. DOI: 10.1109/38.946629.

[90] Sandy Reiß, Stefan Tomiuk, Jutta Kollet, Jan Drewes, Wolfgang Brück, Melanie Jungblut, and Andreas Bosio. "Characterization and classification of glioblastoma multiforme using the novel multiparametric cyclic immunofluorescence analysis

system MACSima". In: *Cancer Research* 79.13 Supplement (2019), 245–245. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/1538-7445.SABCS18-245.

[91] David L. Rimm. "What brown cannot do for you". In: *Nature Biotechnology* 24.88 (2006), 914–916. ISSN: 1546-1696. DOI: 10.1038/nbt0806-914.

[92] Yair Rivenson, Tairan Liu, Zhensong Wei, Yibo Zhang, Kevin de Haan, and Aydogan Ozcan. "PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning". In: *Light: Science & Applications* 8.1 (2019), p. 23. ISSN: 2047-7538. DOI: 10.1038/s41377-019-0129-y.

[93] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydın, Jonathan E. Zuckerman, Thomas Chong, Anthony E. Sisk, and et al. "Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning". In: *Nature Biomedical Engineering* (2019), p. 1. ISSN: 2157-846X. DOI: 10.1038/s41551-019-0362-y.

[94] Stanley J. Robboy, Sally Weintraub, Andrew E. Horvath, Bradden W. Jensen, C. Bruce Alexander, Edward P. Fody, James M. Crawford, Jimmy R. Clark, Julie Cantor-Weinberg, Megha G. Joshi, and et al. "Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply". In: *Archives of Pathology & Laboratory Medicine* 137.12 (2013), 1723–1732. ISSN: 1543-2165. DOI: 10.5858/arpa.2013-0200-OA.

[95] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *arXiv:1505.04597 [cs]* (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[96] Orit Rozenblatt-Rosen, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E. Rood, Orr Ashenberg, Ethan Cerami, Robert J. Coffey, Emek Demir, and et al. "The Human Tumor Atlas Network: Charting Tumor Transitions

across Space and Time at Single-Cell Resolution". In: *Cell* 181.2 (2020), 236–249. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.03.053.

[97] Orit Rozenblatt-Rosen, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E. Rood, Orr Ashenberg, Ethan Cerami, Robert J. Coffey, Emek Demir, and et al. "The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution". In: *Cell* 181.2 (2020), 236–249. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.03.053.

[98] Philipp Rudiger, Jean-Luc Stevens, James A. Bednar, Bas Nijholt, Andrew, Chris B, Vasco Tenner, Jon Mease, Achim Randelhoff, maxalbert, Markus Kaiser, ea42gh, stonebig, jordansamuels, henriqueribeiro, John Bampton, Scott Lowe, Daniel Stephan, arabidopsis, Yuval Langer, Lukas Barth, Justin Bois, Julia Signell, Florian LB, Irv Lustig, Andrew Tolmie, Almar Klein, Benjamin W. Portner, Anthony Monthe, and Anar Z. Yusifov. *holoviz/holoviews: Version 1.12.7*. Version v1.12.7. Nov. 2019. DOI: 10.5281/zenodo.3551257. URL: https://doi.org/10.5281/zenodo.3551257.

[99] A. C. Ruifrok and D. A. Johnston. "Quantification of histochemical staining by color deconvolution". In: *Analytical and Quantitative Cytology and Histology* 23.4 (2001), 291–299.

[100] Denis Schapiro, Hartland W. Jackson, Swetha Raghuraman, Jana R. Fischer, Vito R. T. Zanotelli, Daniel Schulz, Charlotte Giesen, Raúl Catena, Zsuzsanna Varga, and Bernd Bodenmiller. "histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data". In: *Nature Methods* 14.99 (2017), 873–876. ISSN: 1548-7105. DOI: 10.1038/nmeth.4391.

[101] Denis Schapiro, Artem Sokolov, Clarence Yapp, Jeremy L. Muhlich, Joshua Hess, Jia-Ren Lin, Yu-An Chen, Maulik K. Nariya, Gregory J. Baker, Juha Ruokonen, and et al. "MCMICRO: A scalable, modular image-processing pipeline for multiplexed

tissue imaging". In: (2021). DOI: 10.1101/2021.03.15.435473. URL: http://biorxiv.org/lookup/doi/10.1101/2021.03.15.435473.

[102] Geoffrey Schau, Erik Burlingame, and Young Hwan Chang. "DISSECT: DISentangle SharablE ConTent for Multimodal Integration and Crosswise-mapping". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. 2020, 5092–5097. DOI: 10.1109/CDC42340.2020.9304354.

[103] Geoffrey Schau, **Erik A. Burlingame**, and Young Hwan Chang. "DISSECT: DISentangle SharablE ConTent for Multimodal Integration and Crosswise-mapping". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. 2020, 5092–5097. DOI: 10.1109/CDC42340.2020.9304354.

[104] Geoffrey F. Schau, Hassan Ghani, **Erik A. Burlingame**, Guillaume Thibault, Joe W. Gray, Christopher Corless, and Young Hwan Chang. "Transfer Learning for Inference of Metastatic Origin from Whole Slide Histology". In: *bioRxiv* (2021). DOI: 10.1101/2021.04.21.440864.

[105] Geoffrey F. Schau, **Erik A. Burlingame**, Guillaume Thibault, Tauangtham Anekpuritanang, Ying Wang, Joe W. Gray, Christopher Corless, and Young Hwan Chang. "Predicting primary site of secondary liver cancer with a neural estimator of metastatic origin". In: *Journal of Medical Imaging* 7.1 (2020), p. 012706. ISSN: 2329-4302, 2329-4310. DOI: 10.1117/1.JMI.7.1.012706.

[106] Caglar Senaras, Muhammad Khalid Khan Niazi, Berkman Sahiner, Michael P. Pennell, Gary Tozbikian, Gerard Lozanski, and Metin N. Gurcan. "Optimized generation of high-resolution phantom images using cGAN: Application to quantification of Ki67 breast cancer images". In: *PLOS ONE* 13.5 (2018), e0196846. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0196846.

[107] M. Sinn, C. Denkert, J. K. Striefler, U. Pelzer, J. M. Stieler, M. Bahra, P. Lohneis, B. Dörken, H. Oettle, H. Riess, and et al. "α-Smooth muscle actin expression and

desmoplastic stromal reaction in pancreatic cancer: results from the CONKO-001 study". In: *British Journal of Cancer* 111.10 (2014), 1917–1923. ISSN: 1532-1827. DOI: 10.1038/bjc.2014.495.

[108] Nicholas Sofroniew, Kira Evans, Talley Lambert, Juan Nunez-Iglesias, Ahmet Can Solak, Kevin Yamauchi, Genevieve Buckley, Tony Tung, Grzegorz Bokota, Peter Boone, Jeremy Freeman, Hagai Har-Gil, Loic Royer, Shannon Axelrod, jakirkham, Reece Dunham, Pranathi Vemuri, Mars Huang, Hector, Bryant, Ariel Rokem, Justin Kiggins, Hugo van Kemenade, Heath Patterson, Guillaume Gay, Eric Perlman, Davis Bennett, Christoph Gohlke, Bhavya Chopra, and Alexandre de Siqueira. *napari/napari: 0.3.0*. Version v0.3.0. May 2020. DOI: 10.5281/zenodo.3785908. URL: https://doi.org/10.5281/zenodo.3785908.

[109] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. "Cellpose: a generalist algorithm for cellular segmentation". In: *Nature Methods* 18.11 (2021), 100–106. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01018-x.

[110] Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Rachelle Riggers, Joe W. Gray, John Muschler, and Young Hwan Chang. "VISTA: VIsual Semantic Tissue Analysis for pancreatic disease quantification in murine cohorts". In: *Scientific Reports* 10.11 (2020), p. 20904. ISSN: 2045-2322. DOI: 10.1038/s41598-020-78061-3.

[111] Takahiro Tsujikawa, Sushil Kumar, Rohan N. Borkar, Vahid Azimi, Guillaume Thibault, Young Hwan Chang, Ariel Balter, Rie Kawashima, Gina Choe, David Sauer, and et al. "Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis". In: *Cell Reports* 19.1 (2017), 203–217. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.03.037.

[112] Takahiro Tsujikawa, Sushil Kumar, Rohan N. Borkar, Vahid Azimi, Guillaume Thibault, Young Hwan Chang, Ariel Balter, Rie Kawashima, Gina Choe, David

Sauer, and et al. "Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis". In: *Cell Reports* 19.1 (2017), 203–217. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.03.037.

[113] A. Udrea and G. D. Mitra. "Generative Adversarial Neural Networks for Pigmented and Non-Pigmented Skin Lesions Detection in Clinical Images". In: *2017 21st International Conference on Control Systems and Computer Science (CSCS)*. 2017, 364–368. DOI: 10.1109/CSCS.2017.56.

[114] **Erik A. Burlingame**, Jennifer Eng, Guillaume Thibault, Koei Chin, Joe W. Gray, and Young Hwan Chang. "Toward reproducible, scalable, and robust data analysis across multiplex tissue imaging platforms". In: *Cell Reports Methods* 0.0 (2021). ISSN: 2667-2375. DOI: 10.1016/j.crmeth.2021.100053. URL: https://www.cell.com/cell-reports-methods/abstract/S2667-2375(21)00101-6.

[115] **Erik A. Burlingame**, Adam A. Margolin, Joe W. Gray, and Young Hwan Chang. "SHIFT: speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks". In: *Proceedings of SPIE–the International Society for Optical Engineering* 10581 (2018). ISSN: 0277-786X. DOI: 10.1117/12.2293249. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6166432/.

[116] **Erik A. Burlingame**, Mary McDonnell, Geoffrey F. Schau, Guillaume Thibault, Christian Lanciault, Terry Morgan, Brett E. Johnson, Christopher Corless, Joe W. Gray, and Young Hwan Chang. "SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning". In: *Scientific Reports* 10.11 (2020), p. 17507. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74500-3.

[117] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. "Structure-Preserving Color Normalization and Sparse Stain Separation for

Histological Images". In: *IEEE transactions on medical imaging* 35.8 (2016), 1962–1971. ISSN: 1558-254X. DOI: 10.1109/TMI.2016.2529665.

[118] Raphael Vallat. "Pingouin: statistics in Python". In: *Journal of Open Source Software* 3.31 (2018), p. 1026. ISSN: 2475-9066. DOI: 10.21105/joss.01026.

[119] Sofie Van Gassen, Britt Callebaut, Mary J. Van Helden, Bart N. Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data". In: *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* 87.7 (2015), 636–645. ISSN: 1552-4930. DOI: 10.1002/cyto.a.22625.

[120] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[121] Du Wang, Chaochen Gu, Kaijie Wu, and Xinping Guan. "Adversarial neural networks for basal membrane segmentation of microinvasive cervix carcinoma in histopathology images". In: *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*. Vol. 2. 2017, 385–389. DOI: 10.1109/ICMLC.2017.8108952.

[122] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), 600–612. ISSN: 1057-7149. DOI: 10.1109/TIP.2003.819861.

[123] Michael Waskom and the seaborn development team. *mwaskom/seaborn*. Version latest. Sept. 2020. DOI: 10.5281/zenodo.592845. URL: https://doi.org/10.5281/zenodo.592845.

[124] Lukas M. Weber and Mark D. Robinson. "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data". In: *Cytometry Part A* 89.12 (2016), 1084–1096. ISSN: 1552-4930. DOI: https://doi.org/10.1002/cyto.a.23030.

[125] Hallie Wimberly, Jason R. Brown, Kurt Schalper, Herbert Haack, Matthew R. Silver, Christian Nixon, Veerle Bossuyt, Lajos Pusztai, Donald R. Lannin, and David L. Rimm. "PD-L1 Expression Correlates with Tumor-Infiltrating Lymphocytes and Response to Neoadjuvant Chemotherapy in Breast Cancer". In: *Cancer Immunology Research* 3.4 (2015), 326–332. ISSN: 2326-6066, 2326-6074. DOI: 10.1158/2326-6066.CIR-14-0133.

[126] N. Wissozky. "Ueber das Eosin als Reagens auf Hämoglobin und die Bildung von Blutgefässen und Blutkörperchen bei Säugethier- und Hühnerembryonen". In: *Archiv für mikroskopische Anatomie* 13.1 (1877), 479–496. ISSN: 0176-7364. DOI: 10.1007/BF02933947.

[127] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. "A deep learning-based multi-model ensemble method for cancer prediction". In: *Computer Methods and Programs in Biomedicine* 153 (2018), 1–9. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2017.09.005.

[128] Weisi Xie, Adam Glaser, Nicholas Reder, Nadia Postupna, Chenyi Mao, Can Koyuncu, Patrick Leo, Robert Serafin, Hongyi Huang, Anant Madabhushi, and et al. "Abstract PO-017: Annotation-free 3D gland segmentation with generative image-sequence translation for prostate cancer risk assessment". In: *Clinical Cancer*

Research 27.5 Supplement (2021), PO–PO–017. ISSN: 1078-0432, 1557-3265. DOI: 10.1158/1557-3265.ADI21-PO-017.

[129] Xin Yi, Ekta Walia, and Paul Babyn. "Generative Adversarial Network in Medical Imaging: A Review". In: *Medical Image Analysis* 58 (2019). arXiv: 1809.07294, p. 101552. ISSN: 13618415. DOI: 10.1016/j.media.2019.101552.

[130] Yinyin Yuan. "Spatial Heterogeneity in the Tumor Microenvironment". In: *Cold Spring Harbor Perspectives in Medicine* 6.8 (2016), a026583. ISSN: 2157-1422. DOI: 10.1101/cshperspect.a026583.

[131] Dana Carmen Zaha. "Significance of immunohistochemistry in breast cancer". In: *World Journal of Clinical Oncology* 5.3 (2014), p. 382. ISSN: 2218-4333. DOI: 10.5306/wjco.v5.i3.382.

[132] Farhad Ghazvinian Zanjani, Svitlana Zinger, Babak Ehteshami Bejnordi, Jeroen A W M van der Laak, and Peter H. N. de With. "Stain normalization of histopathology images using generative adversarial networks". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, 573–577. DOI: 10.1109/ISBI.2018.8363641.