Machine Learning for Biological Inference Across Spatial, Multimodal, and Clinical Dimensions in Precision Medicine

By

Raphael-Donatus Kirchgaessner

A DISSERTATION

Presented to the Department of Biomedical Engineering

and the Oregon Health & Science University School of Medicine

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

August 2025

Table of Contents

List of Figures

List of Tables

List of Abbreviations

GMM	Gaussian Mixture Model
RS	Random Sampling
IM	Isolation Model
MCRM	Multi Class Regression model
CM	Composite Model
MCC	Matthews Correlation Coefficient
SC	Sample Counts
R	Repetitions
LumA	Luminal A
TMB	Tumor Mutational Burden
ICB	Immune Checkpoint Blockage
ML	Machine Learning
TME	Tumor Microenvironment
mIF	Multiplex Immunofluorescence
iNOS, eNOS, nNOS	Nitric Oxide Synthases
PD-L1, LAG-3	Immune Checkpoint Markers
NO	Nitric Oxide
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
PFS	Progression Free Survival
TLS	Tertiary Lymphoid Structures
RCN	Recurrent Cellular Neighborhood
TIL	Tumor-infiltrating Lymphocytes
LGBM	Light Gradient Boosting Machine
AE	AutoEncoder
LOGO-CV	Leave-One-Group-Out Cross-Validation
FFPE	Formalin-Fixed Paraffin-Embedded
RT	Room Temperature
FDR	False Discovery Rate
LOOCV	Leave-one-out Cross Validation
WSI	Whole Slide Image
OLS	Ordinary Least Squares
RECIST	Response Evaluation Criteria in Solid Tumors
BRCA	Breast Cancer
BLCA	Bladder Cancer
LUAD	Lung Adenocarcinoma
STAD	Stomach Adenocarcinoma
THCA	Thyroid Cancer

COAD	Colon Adenocarcinoma
TCGA	The Cancer Genome Atlas
H&E	Hematoxylin and Eosin stain
sBERT	Sentence-BERT
VAE	Variational Autoencoder
KL	Kullback-Leibler
OCR	Optical Character Recognition
RAG	Retrieval Augmented Generation
HNSW	Hierarchical Navigable Small Worlds
DL	Deep Learning
CLIP	Contrastive Language-Image Pre-Training
LLM	Large Language Model
AI	Artificial Intelligence
MAE	Mean Absolute Error
AMI	Adjusted Mutual Information
ARI	Adjusted Rand Index
EN	Elastic Net
IP	In-Patient
AP	Across-Patients
MTI	Multiplex Tissue Imaging
t-CyCIF	Tissue Cyclic Immunofluorescence
OHSU	Oregon Health & Science University
IRB	Institutional Review Board
HTAN	Human Tumor Atlas Network
HIER	Heat Induced Epitope Retrieval
TSA	Tyramine Signal Amplification
DBMS	Database Management Systems
NIH	National Institutes of Health
HR+	Hormone Receptor positive
µm	Micrometer
PCA	Principal Component Analysis
GDC	Genomic Data Commons

Acknowledgements

I am deeply thankful to everyone who supported me in earning my degree. First, I want to thank my mentor, Dr. Jeremy Goecks, for his ongoing support and guidance throughout my studies. Your feedback, patience, and confidence in my abilities helped me grow both as a scientist and personally. I will always remember our meetings fondly.

Thanks also to Dr. Kyle Ellrott, who welcomed me into his lab without hesitation and offered guidance whenever needed. Your support helped me overcome many obstacles along the way. I also appreciate the members of my Dissertation Advisory Committee, Dr. Daniel Zuckerman and Dr. Xubo Song, for their consistent feedback and suggestions that elevated my research.

Finally, I am grateful to my family for their unwavering support. My partner, Kaya, not only helped with illustrations for my research but also supported me in all aspects of life. My children, Karoline and Ryoko, have enriched my life in ways I never expected. Thanks to Monika, Lea, and Thomas for their ongoing support and confidence, and to Patrick for his friendship and much-needed escapes from work. I am truly thankful for everyone's endless support throughout this journey.

Abstract

The success of precision medicine in oncology depends on our ability to extract actionable insights from complex, high-dimensional datasets.

The promise of precision medicine in oncology hinges on the ability to extract actionable insights from complex, high-dimensional datasets. Advances in single-cell technologies, multiplex tissue imaging (MTI), and multi-omics assays have enabled unprecedented characterization of tumors and their microenvironments. However, translating these data into clinically meaningful inference remains a central challenge. This is due to issues such as missing values, heterogeneous data modalities, and limited interpretability of predictive models. This dissertation develops and evaluates machine learning approaches tailored to address these barriers across three critical dimensions of precision oncology: spatial biology, multimodal integration, and clinical outcome prediction.

First, to address missing data in spatial proteomics, a novel imputation strategy was introduced for MTI datasets. This method leverages the spatial architecture and protein expression profiles of individual cells to reconstruct missing values. The resulting improvements in data completeness enhance downstream tasks such as phenotype classification, facilitating more accurate reconstruction of cellular organization and tumor heterogeneity.

Second, multimodal data integration was approached through a scalable embedding aggregation framework that synthesizes diverse data types, including gene expression, histopathology, somatic mutations, and clinical metadata, into unified patient-level representations. By applying simple vector summation across modality-specific embeddings, this approach enables interpretable and effective classification of cancer type, subtype, and tumor mutational burden (TMB). These results underscore the utility of low-complexity strategies for integrative modeling in biomedical research.

Third, in the clinical domain, machine learning models were applied to identify features predictive of response to immune checkpoint blockade (ICB) in advanced melanoma. Emphasis was placed on model reproducibility and interpretability to support future biomarker discovery. While not yet suitable for clinical deployment, these models provide a foundation for stratifying patients based on molecular and clinical correlates of therapeutic response.

Collectively, this work demonstrates how machine learning can support biological inference across spatial, multimodal, and clinical dimensions, offering both methodological contributions and translational insight. By improving data quality, enabling interpretable integration, and identifying clinically relevant features, this dissertation advances core capabilities essential for the realization of precision medicine in oncology.

Chapter 1 Introduction

1.1 Background and Motivation

1.1.1 Precision Medicine

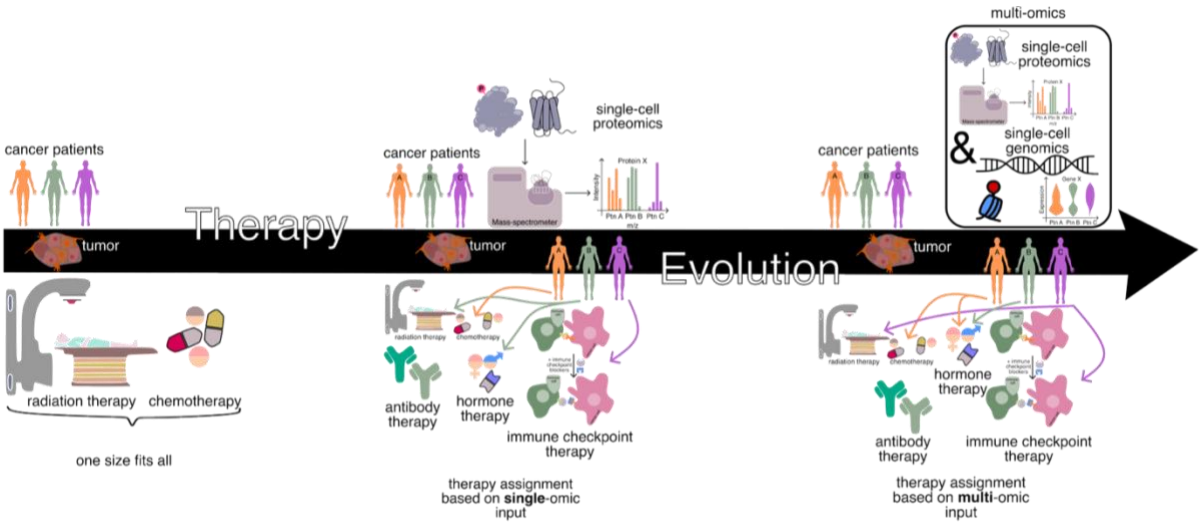


Figure 1.1 Evolution of therapeutic strategies. Initial treatment paradigms followed a "one-size-fits-all" approach, wherein all patients received similar therapies regardless of individual biological differences. With the advent of high-throughput sequencing technologies, therapeutic strategies shifted toward single-omics-guided approaches, enabling treatment decisions based on genomic, transcriptomic, or proteomic profiles, individually. More recently, the field has moved toward integrated multi-omics frameworks, which incorporate diverse layers of molecular information to provide a comprehensive view of the patient's tumor biology and facilitate truly personalized treatment strategies.

Precision medicine has redefined the therapeutic landscape of oncology by shifting from standardized treatment protocols to individualized strategies guided by the molecular and cellular characteristics of each patient's tumor. Historically, cancer treatment followed a one-size-fits-all model, an approach long embedded in medical practice. Therein, interventions were developed based on average responses within large patient populations. While this population-based framework was foundational to early therapeutic advances, it often failed to account for inter-individual variability, resulting in limited

efficacy for many patients [1]. The advent of next-generation sequencing (NGS) enabled the molecular profiling of tumors at unprecedented resolution, allowing clinicians to tailor therapies to the unique genetic and molecular features of individual patients, a concept now central to precision medicine. This paradigm shift departs from uniform treatment approaches by leveraging genomic, transcriptomic, proteomic, and spatial data to guide targeted therapeutic interventions [2–4]. The overarching goal is to enhance efficacy and minimize toxicity, thereby improving patient outcomes across diverse cancer types. Precision oncology has already yielded transformative advances in several malignancies, including non-small cell lung cancer and melanoma, where targeted therapies have significantly altered disease trajectories [5,6]. In clinical practice, precision oncology is commonly operationalized through molecular profiling of tumors to identify biologically relevant alterations that guide treatment decisions. This typically involves analyzing genomic and transcriptomic data to detect molecular features that may predict sensitivity or resistance to specific therapies. These molecular insights are used to stratify patients into subgroups most likely to benefit from targeted interventions, and to inform the selection of diagnostic assays, therapeutic agents, and clinical trial eligibility.

However, fully realizing the potential of precision therapy necessitates moving beyond single-modality analyses. It requires the integration of diverse molecular and cellular layers, including transcriptomic, proteomic, spatial, and immune-related data, to more comprehensively capture the complexity of individual patients, tumor biology, and therapeutic responses [7,8]. For example, Budczies et al. demonstrated that although tumor mutational burden (TMB) can serve as a biomarker for predicting benefit from

immune checkpoint inhibitors, its discriminatory power in distinguishing responders from non-responders remains limited [9]. In contrast, Khader et al. demonstrated that models trained on multimodal data outperformed those trained on unimodal inputs, highlighting the added predictive value of integrative approaches [10].

However, achieving this integration is particularly challenging due to the heterogeneity of these datasets, differences in resolution, scale, and modality, as well as technical limitations such as sparsity, missing values, and batch effects. These data-driven challenges are compounded by biological complexities, including extensive inter- and intra-tumoral heterogeneity, highly individualized tumor–immune interactions, and the limited availability of robust, predictive biomarkers to guide therapeutic decision-making and patient stratification [11,12].

Recent advances in high-throughput single-cell technologies, such as single-cell RNA sequencing, spatial transcriptomics, and multiplexed protein profiling, have enabled detailed characterization of the tumor microenvironment at unmatched resolution [13–15]. These tools have uncovered the diversity of cell types and states within tumors, as well as their spatial organization and functional dynamics. However, leveraging such data for precision oncology remains hindered by several obstacles. Technical noise, sparsity, and missing values are common across modalities, and integrating diverse data types into coherent, interpretable frameworks suitable for downstream analysis and prediction is methodologically demanding [16–18].

To address these challenges, latest efforts have focused on developing computational approaches that can impute missing molecular measurements, integrate heterogeneous single-cell and spatial data into unified embeddings [19,20] and support downstream analyses such as cell type classification, and patient stratification [21–23]. These methodological advances are essential for realizing the full potential of single-cell technologies in guiding personalized cancer diagnostics and treatment.

In this dissertation, seven primary cancer types are investigated: breast cancer (BRCA), bladder cancer (BLCA), thyroid cancer (THCA), stomach cancer (STAD), lung adenocarcinoma (LUAD), colon cancer (COAD), and melanoma. The selection of these malignancies was driven by a combination of data availability, biological diversity (Chapters 2 and 3), and their relevance to the specific research focus (Chapter 4). For example, the breast cancer cohort from the Human Tumor Atlas Network (HTAN) and the melanoma cohort from the Moffitt Cancer Center provided comprehensive, high-quality datasets well suited for integrative analyses. In particular, the HTAN breast cancer dataset included both pre- and post-treatment samples, offering valuable insights into disease progression and molecular alterations at the single-cell level. Similarly, the melanoma dataset was derived from a cohort with detailed, high-fidelity annotations, providing a robust foundation for downstream data science and machine learning applications. The inclusion of additional cancer types was intended to encompass a broad range of biological similarities and distinctions, thereby facilitating a more heterogeneous and representative examination of molecular and spatial patterns across diverse malignancies.

1.1.2 Spatial and Multimodal data in Biomedicine

Advancements in molecular profiling, imaging technologies, and integrative data strategies have significantly expanded the capacity to generate complex biomedical datasets that capture multiple dimensions of biological systems [24,25]. Among these, spatial and multimodal data modalities have emerged as particularly powerful in elucidating disease mechanisms, tissue architecture, and therapeutic response at unprecedented resolution [26–28]. Spatial data preserve the physical context of biological signals, offering insights into the architectural organization of tissues and the tumor microenvironment. Techniques such as multiplexed immunofluorescence allow quantification of protein expression and morphology within native spatial frameworks [29] , supporting the identification of cellular neighborhoods [30]. These localized microenvironments, together with broader tissue architecture and the overall composition of the tumor immune microenvironment (TIME), play a central role in shaping disease progression and therapeutic outcomes. Spatial interactions among malignant cells, stromal elements, and infiltrating immune populations contribute to tumor heterogeneity and influence key processes such as immune evasion, response to targeted or immunomodulatory therapies. Therefore, preserving and interrogating spatial context is essential for accurately characterizing tumor biology and for advancing spatially informed strategies in precision oncology. Beyond their standalone utility, spatial datasets are increasingly integrated with molecular measurements, such as transcriptomic, proteomic, or genomic data, to form multimodal frameworks in precision oncology. In this context, spatial data provide critical architectural and microenvironmental

context, while molecular data capture the underlying biological states of cells. Together, this combination enables a more comprehensive representation of tumor ecosystems. Thus, spatial plus molecular data effectively define a multimodal paradigm, wherein both dimensions are essential for unraveling the complexity of disease progression, therapeutic response, and cellular heterogeneity in cancer.

Multimodal data refer to datasets that integrate information from different biological, clinical, or technological sources. These can include combinations of genomic, transcriptomic, proteomic, imaging, and electronic health record (EHR) data. When harmonized appropriately, multimodal data integration enables a more comprehensive representation of biological systems and patient phenotypes, offering enhanced potential for accurate diagnosis, disease subtyping, and personalized treatment strategies. Despite the growing emphasis on multimodal approaches in contemporary research, the majority of studies continue to rely predominantly on single-modality datasets [31,32]. This reliance can obscure critical interdependencies across molecular, cellular, and spatial domains, thereby limiting the discovery of integrative biomarkers or mechanisms of disease. Furthermore, many machine learning and computational frameworks remain modality-specific, failing to exploit the synergistic value inherent in cross-modal associations [31,32].

Overcoming these limitations requires the development of integrative analytical frameworks capable of aligning heterogeneous data types while preserving modality-specific nuances and uncovering shared biological signals. However, the inherent heterogeneity, high dimensionality, and interdependence of spatial and multimodal

biomedical data pose significant analytical challenges. Modern single-cell technologies, for instance, routinely produce data encompassing tens of thousands to millions of cells, each annotated with extensive molecular, phenotypic, and spatial features. The sheer scale and complexity of multimodal biomedical datasets render both manual evaluation and integrative analysis exceedingly challenging. In particular, inconsistencies in data resolution, widespread patterns of missingness, and sampling biases complicate efforts to derive robust and reproducible biological insights. Among these challenges, missing data stands out as a critical barrier to discovery. First, it can distort the underlying structure of the biological system under investigation, leading to biased or incomplete interpretations. Second, the inability to accurately infer missing values hinders the reconstruction of a coherent and comprehensive biological narrative. As a result, essential patterns may be masked or misrepresented, thereby limiting the potential for novel insights and reducing the reliability of downstream analyses.

These challenges are compounded by the biological heterogeneity both within and across diseases. Even among patients diagnosed with the same condition, underlying molecular profiles can differ substantially, reflecting diverse mutational, microenvironmental contexts, or host factors. This variation makes it difficult to detect subtle yet biologically meaningful signals, such as nuanced differences in gene expression, which may remain undetected through conventional analysis.

Together, these complexities underscore the need for advanced computational frameworks capable of integrating, modeling, and interpreting spatial and multimodal datasets in a scalable and biologically informed manner. Addressing these complexities requires the

development and application of robust computational frameworks, particularly those rooted in machine learning, that can learn meaningful representations across diverse data types while preserving biological interpretability [33].

1.1.3 Machine Learning / Deep Learning

Recent advances in high-throughput technologies have introduced unprecedented complexity into biomedical datasets, including spatial, multimodal, and high-dimensional data with heterogeneous structures and pervasive missingness. These challenges underscore the need for computational approaches capable of modeling nonlinear relationships, integrating diverse data modalities, and extracting biologically meaningful insights from complex systems. Machine learning (ML), a subset of artificial intelligence (AI), offers scalable and flexible methods that are particularly well suited to address these demands by capturing latent structure and subtle variation in biomedical data.

ML enables computational systems, referred to as models, to learn from high-dimensional data that are often beyond the scope of human cognitive processing. These models can identify complex patterns, make predictions, and support decision-making with minimal human intervention. ML approaches are broadly categorized into supervised and unsupervised learning. Supervised learning involves training models on labeled datasets, making it especially useful for tasks such as classification, e.g., predicting whether a biopsy sample is malignant or benign. In contrast, unsupervised learning operates on unlabeled data and is primarily used to uncover latent structures or groupings within

datasets, such as identifying patient subpopulations with shared clinical or molecular characteristics [34,35].

The growth of ML applications in oncology has been largely propelled by the exponential expansion of high-resolution biological, clinical, and molecular datasets. This surge is enabled by technologies that capture spatial, multimodal, and patient-specific dimensions with increasing precision and scale. As a result, ML has become a transformative tool in both biomedical research and clinical oncology [36], offering capabilities that often exceed the limits of traditional statistical approaches in handling complex, high-dimensional data [37].

Prior to the adoption of ML, researchers relied heavily on classical statistical and multivariate techniques such as principal component analysis (PCA), k-means clustering, and distance-based methods (e.g., Euclidean metrics). Patient stratification typically depended on predefined clinical or molecular subtypes, such as estrogen receptor (ER) status or HER2 expression. In contrast, modern ML methods allow researchers to move beyond fixed categories by learning nuanced, nonlinear relationships and identifying emergent patterns between samples, ultimately enabling a more refined understanding of disease phenotypes and therapeutic responses.

The rapid development of ML architectures, particularly those based on deep learning (DL), has equipped researchers with a powerful and adaptable toolkit for addressing a wide range of biomedical questions. DL refers to a class of neural network models with multiple layers that can learn abstract, hierarchical representations from complex datasets. Among

these, autoencoders (AEs) are commonly employed for dimensionality reduction in high-dimensional contexts such as single-cell RNA-seq. Graph Neural Networks (GNNs) are well suited for modeling relational and spatially structured data, including patient similarity graphs. Transformer-based models excel in handling sequence data and integrating heterogeneous modalities.

Although not a DL model per se, Light Gradient Boosting Machines (LGBMs) are frequently used in biomedical ML pipelines due to their strong performance on structured datasets[38–40] . Light Gradient Boosting Machines (LGBMs) are an efficient implementation of gradient boosting decision trees that leverage histogram-based algorithms and leaf-wise tree growth strategies to optimize both computational speed and predictive accuracy. As non-linear models, LGBMs are capable of capturing complex, high-dimensional relationships within data, similar to deep learning approaches, which makes them particularly well suited for biomedical applications. Their ability to handle missing values, support categorical features natively, and incorporate robust regularization techniques enables effective modeling of datasets that often exhibit sparsity, noise, and heterogeneity. Furthermore, LGBMs provide built-in mechanisms for assessing feature importance, thereby enhancing model interpretability, an essential consideration in clinical and translational research contexts. Consequently, LGBMs are frequently employed for classification and regression tasks involving clinical metadata, mutation profiles, and other structured molecular features, serving both as strong baselines and as complementary models to deep learning frameworks in biomedical data analysis.

The choice of model should be carefully aligned with the specific biological question, data structure, and interpretability requirements of a given study. For example, AEs can compress large feature spaces while preserving key signals; GNNs can capture relationships in spatial omics data; transformers can unify disparate data modalities in a single framework; and LGBMs can deliver robust predictive performance in clinical outcome modeling.

ML and DL has been successfully applied across a broad spectrum of oncology-related tasks, including prognosis, risk stratification, treatment planning, personalized medicine, cancer subtyping, and multi-omics integration. In proteomics-centric applications, ML approaches, including probabilistic models and deep generative frameworks, have proven particularly effective at addressing missing or sparse data by learning robust latent representations and enabling accurate data imputation [41]. DL models have also contributed to predicting therapeutic outcomes, detecting metastases, stratifying cancer subtypes, and integrating diverse molecular and clinical datasets, enhancing both the resolution and translational impact of oncology research [36,42,43].

1.2 Multiplex Tissue Imaging

1.2.1 Brief Overview

Multiplex tissue imaging (MTI) are a set of single-cell spatial proteomics and transcriptomics assays for highly detailed profiling of biological tissues. With MTI, single-

cell abundance levels and spatial distribution of 10-150 proteins and/or 500-2000 RNAs can be quantified simultaneously [44,45]. MTI enables characterization of individual cells as well as tissue organization, and MTI has been used in studies of healthy tissue [46], COVID [47], cancer [48] and other diseases [49–51]. There are many MTI platforms, including multiplex immunofluorescence (mIF) [52], with its variants; cyclic immunofluorescence (CycIF) [29], CO-Detection by indEXing (CODEX) [53], CosMx [54], Xenium [55] and multiplex immunohistochemistry [56]. MTI has been used to generate large datasets in NIH consortia such as the NIH Human BioMolecular Atlas Program [57] and the NCI Cancer Moonshot Human Tumor Atlas Network [58]. MTI is also an increasingly common assay in cancer [59], where it has proven important for quantifying tumor spatial organization and microenvironment heterogeneity [60] and connecting these features to cancer subtypes, prognosis, and therapy response [61,62].

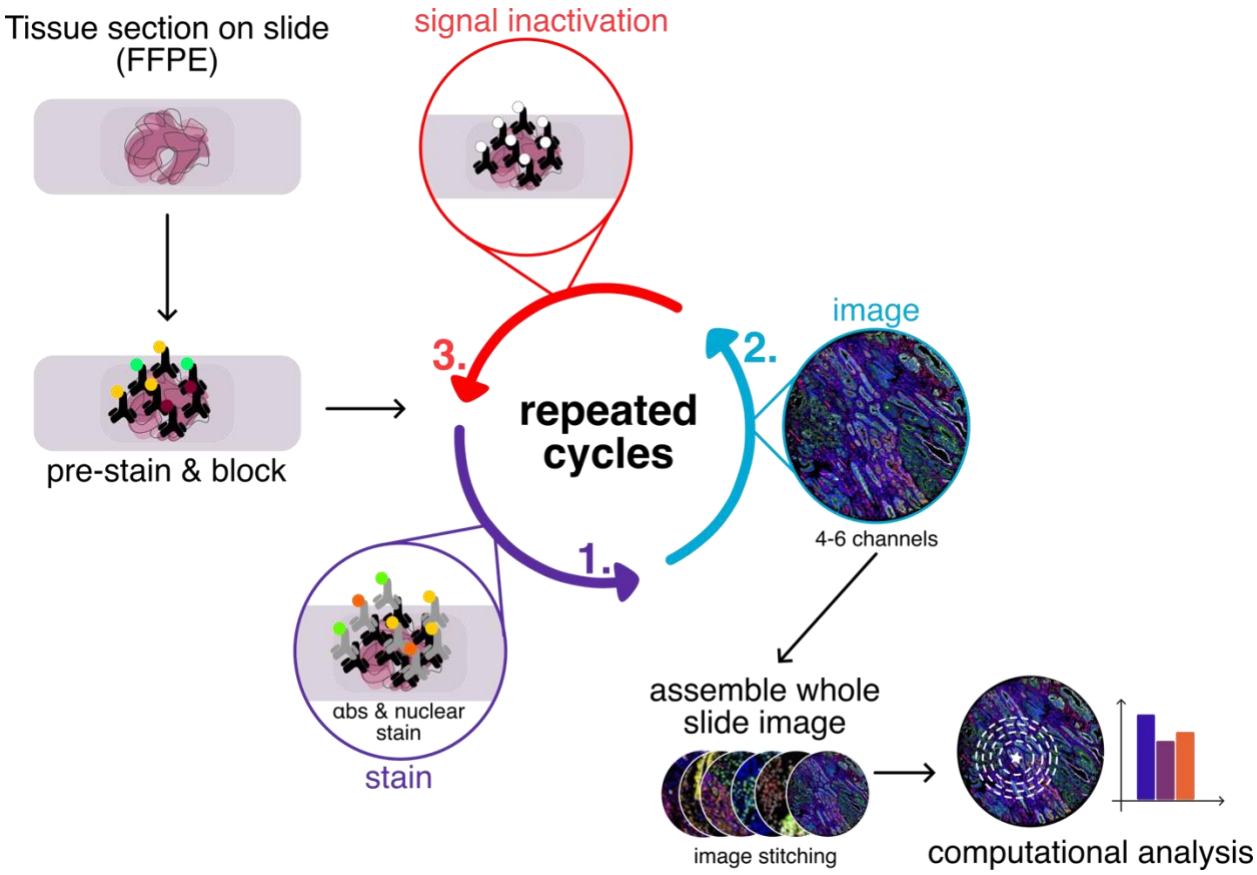


Figure 1.2 Schematic overview of a multiplex tissue imaging workflow.

Tissue sections, e.g. formalin-fixed paraffin-embedded (FFPE) human tissue, are first pre-treated to reduce background and autofluorescence caused by non-specific antibody binding. Each imaging cycle consists of staining the sample for target proteins of interest (1), followed by high-resolution fluorescence imaging (2). After imaging, the stains are inactivated (3), and the cycle is repeated to enable multiplexed imaging across dozens of markers. The iterative nature of the process allows spatially resolved, single-cell-level protein profiling across complex tissue architecture. After all rounds have been completed, the slide images can be assembled and stitched together and used for further computational analysis.

One commonly used MTI platform, tissue-based cyclic immunofluorescence (T-CyCIF), operates on formalin-fixed paraffin-embedded (FFPE) human tissue sections. Prior to cyclic staining, sections are incubated with secondary antibodies to reduce tissue autofluorescence arising from non-specific antibody binding. Each imaging cycle begins with the application of one to four primary antibodies targeting specific proteins of interest, along with a nuclear stain. The sample is then imaged, chemically bleached to remove

fluorophores, and the cycle is repeated. This iterative process (Figure **1.2**) enables highly multiplexed imaging of protein markers across the same tissue section. Upon completion of all imaging cycles, the sample may optionally undergo hematoxylin and eosin (H&E) staining to provide additional histopathological context.

1.2.2 Challenges of multiplex tissue imaging

Several key factors limit the usefulness of MTI. Only 10-150 proteins and/or several thousand RNAs can be assayed in a single experiment, and hence the information obtained from a single experiment is bounded. T-CyCIF is generally restricted to the detection of up to ~60 protein targets due to practical constraints on the number of staining and imaging cycles. Further, MTI assays can suffer from several technical issues that reduce the information obtained, including tissue loss or folding, probe failure, illumination artifacts, or errors in downstream image processing. These limitations greatly impact MTI data quality and substantially reduce the overall utility of MTI.

1.2.3 Clinical Relevance

MTI technologies, such as t- CyCIF, have become pivotal in advancing the field of precision oncology due to their ability to simultaneously visualize dozens of biomarkers at subcellular resolution. This high-dimensional imaging capability allows for detailed characterization of the molecular and cellular composition of tumors, thereby facilitating the design of patient-specific treatment strategies. Such personalized therapies are increasingly prioritized in cancer care, as they tend to produce fewer side effects and result

in improved survival outcomes compared to traditional approaches like chemotherapy and radiotherapy.

The clinical value of MTI lies not only in its molecular resolution but also in its capacity to preserve spatial context. Techniques like t-CyCIF enable comprehensive profiling of the tumor microenvironment (TME), revealing the spatial organization and interaction of diverse cell types within the tumor [63]. This spatial information is critical, as the architecture of the TME is known to influence tumor progression, immune evasion, recurrence, and metastatic potential [64]. By capturing both phenotypic and spatial complexity, MTI provides insights into patient-specific tumor biology that are essential for informing targeted therapeutic interventions.

Moreover, tumors are characterized by an accumulation of somatic mutations and patient-specific molecular alterations, which can serve as the foundation for therapeutic targeting. MTI, through platforms such as t-CyCIF, contributes to the identification of these clinically actionable features by enabling a spatially resolved and multiplexed analysis of both tumor cells and their surrounding microenvironments. This detailed mapping supports the development of more precise, effective, and individualized treatment plans, ultimately enhancing clinical outcomes and survival rates in cancer patients.

1.3 Heterogenous Embeddings

1.3.1 Brief overview

In modern machine learning, the integration of diverse data modalities into unified embedding vectors has become a standard practice, significantly improving model performance in fields such as natural language processing, genomics, and computer vision. By fusing heterogeneous data, these models generate richer representations that capture multiple dimensions of complex systems. However, this integration often comes at the cost of interpretability, as the contribution of individual modalities becomes obscured within the composite embeddings. Understanding the provenance and influence of each data source is critical for improving model transparency and ensuring that clinically or biologically relevant signals are preserved.

Biomedical data in precision oncology is inherently multi-modal, encompassing genomic and transcriptomic profiles, histopathological imaging, and unstructured clinical notes. A common approach to multi-modal representation learning involves constructing a shared embedding space in which all data modalities are jointly encoded. For example, methods such as Contrastive Language-Image Pretraining (CLIP) [65] have been successfully applied to align language and image representations in a common space, enabling applications such as the text-to-image generation seen in DALL·E [66].

Our objective was to evaluate whether a simplified aggregation strategy could be employed for integrating multi-modal patient data, specifically RNA expression profiles, H&E-stained histopathology images, clinical text annotations and somatic mutations, into a single embedding space that facilitates search, clustering, and retrieval of related cancer records. Our primary motivation was to assess whether these heterogeneous data types could be combined without the need for complex manifold alignment or joint training procedures. Instead, we sought to generate independent uni-modal embeddings for each data type and subsequently merge them into a unified representation, retaining the capacity to trace each component back to its respective modality. This approach emphasizes interpretability and modularity, enabling more transparent exploration of integrated biomedical data while supporting downstream analytic tasks.

1.3.2 Vector Database

Vector databases are increasingly critical for the storage and retrieval of multimodal unstructured data, including large text documents, medical images, and procedural videos. In clinical contexts, these data types correspond to diagnostic annotations, histopathological biopsy images, and potentially videos from endoscopic or surgical procedures [67]. Through a process known as vectorization, complex and high-dimensional features of such data can be efficiently encoded into vector representations. This transformation enables noise reduction while retaining the most salient information [67].

The growing reliance on applications such as reverse image search, content-based retrieval, and recommendation systems has amplified the demand for effective vector

management. Vector databases, specialized database management systems (DBMS) designed for handling high-dimensional vector data, serve this need by enabling rapid similarity searches across large collections of embedded data points [67–69]. To retrieve or query information from a database management system (DBMS), a variety of techniques can be employed, with similarity-based search emerging as a particularly powerful approach. This method leverages similarity scores, quantitative measures of the closeness between two or more feature vectors, to identify records that share common characteristics or patterns. Such vector-based querying is especially valuable in domains where exact matches are uncommon and nuanced similarity is more informative, including image retrieval, natural language processing, and recommendation systems. In recent years, these techniques have shown increasing promise in healthcare, particularly within the field of precision medicine. As illustrated in Figure **1.3**, a vector database comprising patient-specific feature embeddings can be queried using a new patient's vector representation. This allows the identification of patients with similar biological profiles, potentially reflecting comparable disease subtypes or molecular mechanisms. The corresponding patient identifiers can then be cross-referenced with electronic health records (EHRs) to retrieve prior treatment responses, enabling more informed therapeutic

decisions. This approach has the potential to accelerate clinical decision-making and improve patient outcomes by tailoring treatments based on biologically grounded patient similarity.

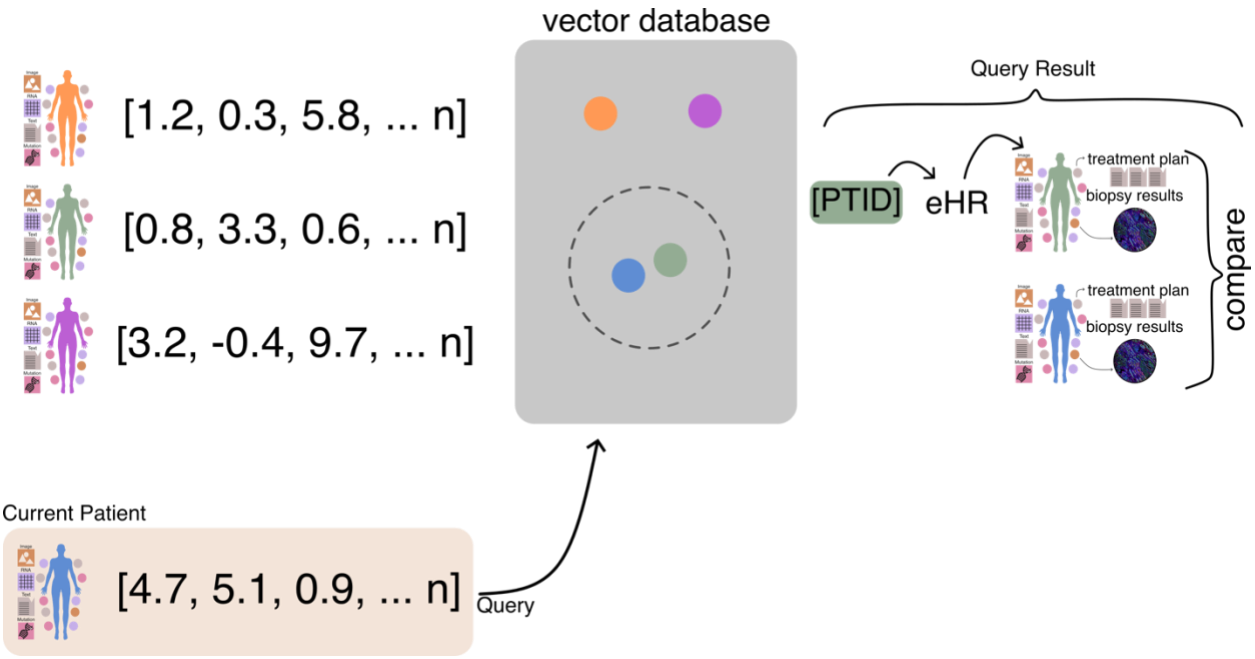


Figure 1.3 Potential schematic vector database usage in a health care setting. Overview of the vector database query process. Each patient is represented as a vector in a high-dimensional mathematical space. A query vector, derived from a new case, is used to retrieve the most similar patient vectors from the database. These retrieved patient profiles can then be linked to corresponding electronic health records (eHR) to inform potential therapeutic responses, guide treatment decisions, and support further clinical insights.

1.3.3 Clinical Relevance

The ability to integrate and store multimodal patient data within vector databases significantly advances the goals of precision medicine. Unlike traditional approaches that rely on unimodal comparisons, such as genomic data alone, multimodal analysis considers a combination of data types, including imaging, clinical text, and molecular profiles. Storing these heterogeneous data as vectorized representations enables efficient

retrieval and comparison at a deeper, more holistic level. As a result, patient similarity assessments can be performed across multiple biological and clinical dimensions, leading to more accurate identification of clinically relevant cohorts. This capability supports data-driven treatment decisions, facilitates the discovery of novel disease subtypes, and enhances the potential for personalized therapeutic strategies by leveraging patterns found in patients with shared multimodal profiles.

1.4 Response to Immune Checkpoint Blockage

1.4.1. Brief overview

The immune system inherently possesses the capability to recognize and eliminate cancer cells. However, this immune response can be suppressed by inhibitory receptors and their ligands [70]. Cancer cells exploit these inhibitory pathways to evade immune destruction by expressing corresponding ligands, thus avoiding detection and clearance by the immune system [30]. Therapeutic intervention by blocking these immune checkpoints, specifically through antibodies targeting programmed death receptor-1 (PD-1) and its ligand PD-L1, can restore and reinvigorate immune activity against cancer cells (Figure **1.4**). This therapeutic strategy is termed immune checkpoint blockade (ICB).

The introduction of immune checkpoint blockade (ICB) therapies has revolutionized cancer treatment by harnessing the body's immune system to recognize and eliminate tumor cells [71]. These therapies have demonstrated clinical benefit across multiple malignancies,

including lung, bladder, renal, and head and neck cancers [72], marking a paradigm shift from traditional cytotoxic and targeted approaches toward immunomodulatory strategies. Among these cancer types, melanoma has emerged as a particularly notable success story. Historically associated with poor prognosis and limited treatment options, advanced melanoma once carried a median overall survival of less than one year. However, with the advent of ICB therapies, particularly inhibitors targeting the programmed cell death protein 1 (PD-1) and its ligand PD-L1, long-term survival rates have dramatically improved, with a substantial proportion of patients now achieving durable responses and multi-year survival [73]. Similarly, anti-cytotoxic T-lymphocyte-associated protein 4 (anti-CTLA-4) checkpoint blockade has become an established therapeutic option for patients with advanced melanoma (stage III and stage IV) [74].

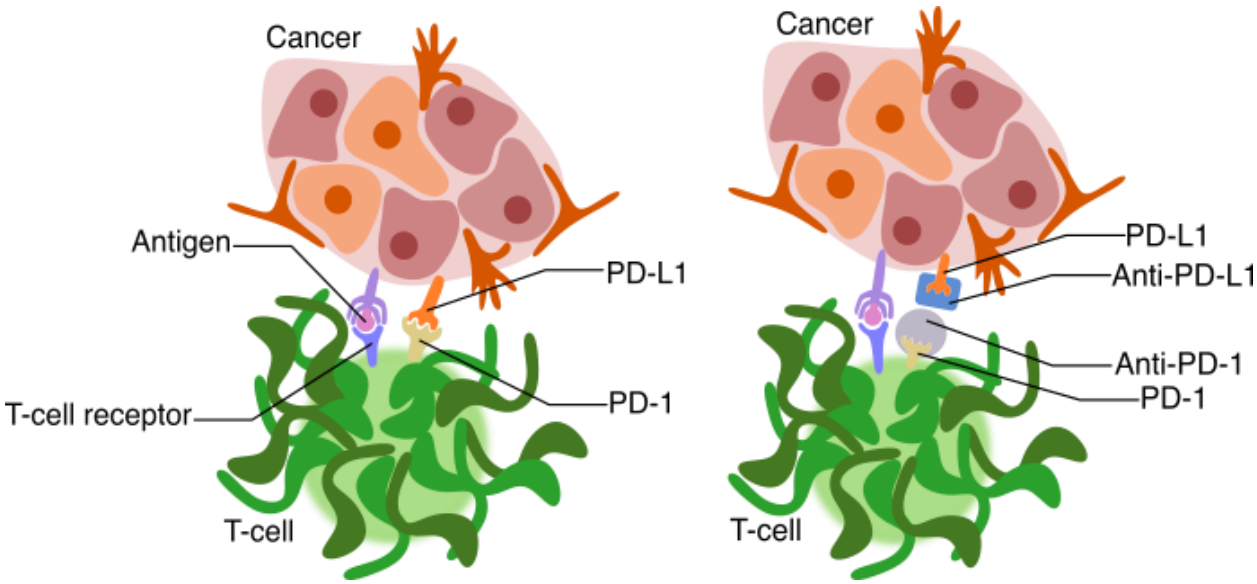


Figure 1.4 Schematic immune checkpoint blockage. The left panel illustrates an interaction between a T-cell and a cancer cell. Tumor cells express antigens recognized by the T-cell receptor (TCR), but also express PD-L1, which binds to the PD-1 receptor on T-cells, leading to T-cell inactivation and immune evasion. The right panel demonstrates the mechanism of immune checkpoint blockade. Therapeutic antibodies targeting PD-1 and PD-L1 (Anti-PD-1 and Anti-PD-L1) disrupt this inhibitory interaction, restoring T-cell activity and enabling immune-mediated tumor destruction. This strategy underlies the clinical use of immune checkpoint inhibitors in cancer immunotherapy.

Despite these advancements, the therapeutic efficacy of checkpoint inhibitors remains limited, benefiting only approximately 30%-40% of melanoma patients [75–79]. Consequently, patients who do not respond to these treatments experience critical delays, potentially reducing the likelihood of successful outcomes through alternative therapeutic strategies. This highlights the urgent need for predictive biomarkers and enhanced treatment protocols to identify and effectively manage non-responsive patients promptly.

1.4.2 Melanoma

Chapter 4 presents a focused study on melanoma, an aggressive and often lethal form of skin cancer originating from melanocytes, the pigment-producing cells of the epidermis [80] Melanoma is distinguished by its high mutational burden, driven largely by ultraviolet (UV) radiation-induced DNA damage, which results in a diverse array of neoantigens. This high degree of tumor immunogenicity renders melanoma particularly susceptible to immune surveillance and, consequently, an ideal candidate for immunotherapeutic strategies. Among these, immune checkpoint blockade (ICB) therapies, such as anti-CTLA-4 and anti-PD-1/PD-L1 antibodies, have demonstrated remarkable clinical efficacy in melanoma, often leading to durable responses in a subset of patients [81]. However, therapeutic resistance and heterogeneous response profiles remain major challenges, underscoring the need for a deeper understanding of the tumor microenvironment, immune infiltration dynamics, and spatial heterogeneity.

1.4.3 Clinical Relevance

There remains a critical need to refine strategies for identifying robust and clinically actionable biomarkers that can predict therapeutic response across a range of cancer treatments, including but not limited to immune checkpoint blockade (ICB). In advanced melanoma, the identification of predictive biomarkers holds particular clinical significance, as it enables early stratification of patients most likely to benefit from ICB therapies, thereby facilitating timely and personalized treatment decisions. More broadly, such predictive frameworks are equally applicable to other immunotherapies, targeted agents, and combination regimens, especially as spatially resolved and multimodal omics datasets become increasingly available.

Machine learning–based approaches provide a powerful means to address this challenge, as they are capable of integrating high-dimensional spatial, molecular, and clinical data to uncover complex, non-linear relationships associated with therapeutic outcomes. Early identification of non-responders would enable rapid transition to alternative treatment strategies, reducing exposure to ineffective therapies and minimizing associated toxicities. This precision-guided framework not only enhances treatment efficacy but also conserves healthcare resources and improves patient quality of life.

While ICB represents an important proof-of-concept, the underlying methodology, combining spatial omics profiling with predictive modeling, is broadly generalizable. It can be applied to diverse therapeutic contexts to elucidate mechanisms of response and

resistance, ultimately advancing the broader goal of personalized oncology through data-driven, mechanistically informed prediction.

Chapter 2 Imputing single-cell protein abundance in multiplex tissue imaging

This chapter has been formatted for inclusion in this dissertation from the manuscript " Imputing single-cell protein abundance in multiplex tissue imaging" by Raphael Kirchgaessner, Cameron Watson, Allison Creason, Kaya Keutler and Jeremy Goecks,

Published in Nature Communications (2025) [41]. The author of this dissertation is the first author of this manuscript and used single-cell gated data generated by MCMICRO [82] to conduct the computational experiments, the results of which were used to generate all figures in this manuscript.

Introduction

Multiplex tissue imaging (MTI) are a set of single-cell spatial proteomics and transcriptomics assays for highly detailed profiling of biological tissues. With MTI, single-cell abundance levels and spatial distribution of 10-150 of proteins and/or 500-2000 RNAs can be quantified simultaneously [44,45]. MTI enables characterization of individual cells as well as tissue organization, and MTI has been used in studies of healthy tissue [46], COVID [47], cancer [48], and other diseases [49–51]. There are many MTI platforms, including cyclic immunofluorescence (CycIF) [29], CO-Detection by indEXing (CODEX) [53], CosMx [54], Xenium [55] multiplex immunohistochemistry [56]. MTI has been used to generate large datasets in NIH consortia such as the NIH Human BioMolecular Atlas Program [57] and the NCI Cancer Moonshot Human Tumor Atlas Network [58]. MTI is also

an increasingly common assay in cancer [59], where it has proven important for quantifying tumor spatial organization and microenvironment heterogeneity [60] and connecting these features to cancer subtypes, prognosis, and therapy response [61,62].

However, several key factors limit the usefulness of MTI. Only 10-150 proteins and/or several thousand RNAs can be assayed in a single experiment, and hence the information obtained from a single experiment is bounded. Further, MTI assays can suffer from several technical issues that reduce the information obtained, including tissue loss or folding, probe failure, illumination artifacts, or errors in downstream image processing. These limitations greatly impact MTI data quality and substantially reduce the overall utility of MTI. To mitigate these limitations and improve utility of MTI, machine learning and deep learning approaches can be used to computationally increase the numbers of proteins/RNAs available from MTI and mitigate assay failures. Computationally increasing, or imputing, additional data by filling in missing data with predicted values is already common in other molecular assays, such as single-cell RNA sequencing (scRNA) [83–90], bulk genomics [91] and bulk transcriptomics [92]. While imputation has been applied to MTI images [93,94], to the best of our knowledge imputation on MTI single-cell datasets has not been explored. Imputation has been applied to MTI image data [95,96], being able to reconstruct protein expression in images. However, imputing single-cell data is especially valuable because single-cell datasets require fewer computational resources to process than images and can be readily integrated with other molecular datasets.

In this study, we applied machine learning (ML) and deep learning (DL) methods to impute protein abundance in tissue-based cyclic immunofluorescence (t-CyCIF) [29] datasets obtained from breast cancer tissues. Because t-CyCIF is an open and quantitative multiplexed tissue imaging assay, it is ideally suited for imputation. We evaluated the performance of ML and DL methods to predict protein abundance levels in t-CyCIF single-cell datasets that included 20 proteins. Three distinct ML/DL approaches, regularized linear regression, gradient-boosted trees, and autoencoders, were used to impute single-cell protein abundance values across both patients and timepoints. Spatial information was introduced to improve imputation results. To demonstrate a biological application of imputed single-cell protein abundance, we used imputed data to predict whether single cells were more likely to come from pre-treatment or post-treatment breast cancer biopsies Overall, our results demonstrate that accurate imputation is possible for many proteins, that spatial information significantly improves imputation results, and that imputed protein values are useful in a biological application.

Results

Study Cohort and Analysis Overview

The multiplexed tissue imaging single-cell datasets used in this study were generated using a 20-plex t-CyCIF [29] applied to a cohort of hormone receptor-positive (HR+), HER-2 negative metastatic breast cancer biopsies. t-CyCIF is a unique multiplexed tissue imaging

assay that has been shown to provide robust and repeatable quantifications of protein concentrations across a range of biological samples.

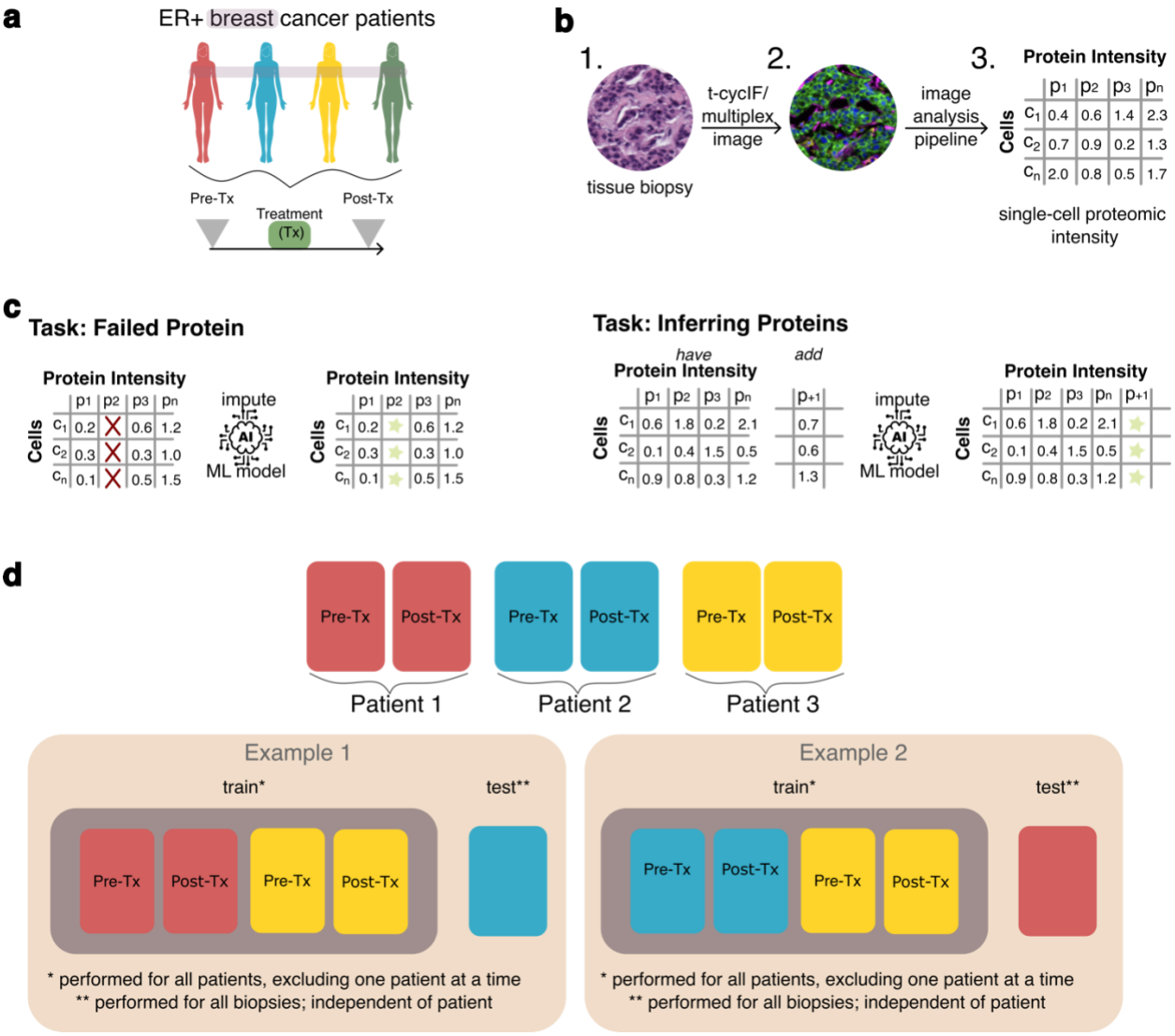


Figure 2.1 Overview of dataset, study motivations, and analysis approaches. a: Biopsies were obtained from four HR+ breast cancer patients before and after the same standard-of-care therapy for a total of eight biopsies. b: each biopsy was assayed using the multiplexed tissue imaging assay t-CyCIF to quantify abundance levels of 20 proteins and then processed using an image analysis pipeline to create single-cell feature tables (total number of cells identified: 475359); c: the key tasks addressed by this work are imputing failed proteins and inferring additional proteins not present in an multiplex tissue imaging (MTI) experiment; d: approaches for training and testing ML models for imputing proteins across patients.

The tissue biopsies and datasets are part of the NCI Cancer Moonshot Human Tumor Atlas Network [58] and have detailed associated clinical metadata. Our dataset includes a total of eight biopsies derived from four patients (Figure **2.1**a) that received a CDK4/6 inhibitor in combination with endocrine therapy, which is a common combination therapy in metastatic HR+ breast cancer. Each patient contributed a pair of biopsies, a pre-treatment biopsy and a biopsy taken at the time of tumor progression.

Image stacks collected from t-CyCIF were processed using the MCMICRO image analysis pipeline [82] to generate single-cell feature tables (Figure **2.1**b). Each row in the table is a single cell identified in the image, and the table columns are the protein abundance levels calculated via mean pixel intensity per cell. In total 475,359 single cells were identified across all biopsies, with an average of 59,400 cells per biopsy. To perform in-patient evaluation, either the pre-treatment or the post-treatment biopsy was used for training a machine learning model while the remaining biopsies were used for testing model performance. Biopsy timing was not used in this study. In total 16 proteins were shared between all biopsies, including eight proteins for identifying cell types (lineage proteins) and eight proteins for characterizing cellular functional states (functional proteins) (Table 2.1).

Table 2.1: Overview of proteins assayed using t-CyCIF and their use as functional or lineage proteins. Lineage proteins are used to identify cell types whereas functional proteins are used to characterize cell function.

Lineage Protein	Functional Protein
CD45	Ki67
αSMA	pERK
eCadherin	PR
CK19	EGFR
CK14	p21
CK17	pRB
CK7	AR
Vimentin	HER2

The imputation task in this study was to predict protein abundance levels for a withheld protein or set of proteins. Preprocessing was performed to remove all columns from the datasets except protein intensities, followed by a min-max scaling approach, which maps values between zero and one. Thus, model error is in the range [0,1] where lower error represents better performance. For each machine learning experiment, one or more proteins were withheld and used as the target variable(s) for the predictive model, and the remaining protein abundances were used as input features for the model. This task simulates the key application for imputation in MTI: computationally increasing proteins not originally included in an MTI assay or inferring protein levels where the assay failed (Figure **2.1**c). Three machine learning methods were used for imputation: elastic-net (EN) regularized linear regression [97], light gradient-boosting machine (LGBM) [98], and neural network autoencoders (AE) [99].

These algorithms offer different advantages for addressing the complexities of our dataset and research objectives. Elastic Net (EN) is a linear model that effectively handles high-dimensional data like that in our MTI datasets by using regularization to manage

multicollinearity and select relevant proteins. However, EN requires a separate model to predict each protein, and this is time-consuming. Light Gradient Boosting Machines (LGBM) use a non-linear approach for developing predictive models and are amongst the most efficient and best performing methods for tabular data like our MTI datasets. Like EN, LGBM also requires a model for each protein. Autoencoders (AEs) can learn non-linear relationships, reduce dimensionality, and denoise data, allowing a single model to impute multiple proteins at once, although their compression techniques may lead to some loss of precision. Autoencoders were chosen for this study due to their ability to reduce dimensionality and denoise data while preserving essential information. This helps in data imputation and improving data quality before further analysis. EN and LGBM are straightforward to implement and handle linear and non-linear relationships, respectively, while AEs provide efficient preprocessing and multi-protein imputation.

Imputation model training and evaluation were conducted using a leave-one-out cross-validation (LOOCV) approach (**Figure 2.1**d). In this methodology, each patient was considered a single data point, whereby a model was trained on all biopsies except those from one patient. Subsequently, the model's performance was assessed using the biopsies from the patient excluded during training. This LOOCV approach was chosen to prevent data leakage from biopsies associated with the same patient as the test biopsy, thereby closely approximating real-world application scenarios. Model performance was calculated by averaging the mean absolute error (MAE) scores across all runs of the model on a particular train-test dataset split. Statistical evaluations were carried out using the

Mann-Whitney U test, and multiple hypothesis tests were adjusted using the Benjamini-Hochberg correction method.

Protein abundance imputation with elastic net and light gradient-boosting machines

To establish baseline performance of our imputation models, we conducted a test using mean imputation, where values were imputed by using the mean protein abundance value in the training dataset. Using mean values for imputation serves as a null or baseline model to determine if a machine learning model provides genuine improvements over a simple heuristic. The EN model outperformed the null model by an average of 0.078 MAE indicating that the EN model demonstrated superior performance compared to the null model (Figure **2.2**a). This performance difference was statistically significant, with an average adjusted p-value less than 0.0001 for all proteins. Proteins CK17 and Ki67 were most accurately imputed with MAE of 0.05. Proteins for which the imputation MAE exceeded 0.2 included CK19, ER, CK14, and PR.

Using Light Gradient Boosting Machine (LGBM) yielded improved imputation accuracy compared to EN (Figure **2.2**b). Like the EN, LGBM performance for the same 12 of 16 proteins was between 0.05 and 0.20 MAE. LGBM imputation accuracy for CK19 and ER are like the EN and greater than 0.2 MAE. Overall, LGBM displayed more accurate imputation results than EN (Table **2.2**) both in terms of mean and standard deviation.

Table 2.2: Mean and standard deviation of performance for EN, LGBM and AE.

Model	EN	LGBM	AE Single	AE Multi
Mean (Std)	0.11 (0.07)	0.10 (0.06)	0.13 (0.09)	0.13 (0.09)

To provide a comprehensive overview of performance, a *mean of all proteins* column is included to show the average imputation accuracy across all proteins for each model (Figure **2.2**a, Figure **2.2**b). While using LGBM improved imputation accuracy compared to the EN, some proteins still exhibit a high MAE, such as CK19 and ER. These proteins exhibited high variance (Supplementary Table **2.6**), presenting a significant challenge for accurate imputation. Supplementary Figure 2.9 shows protein abundance distributions of selected proteins with especially high or low variance to illustrate why imputation is difficult for proteins such as CK19 and ER that exhibit high variance. To further evaluate imputation performance, we calculated the single-cell level correlation between ground truth and imputed protein expression. Correlation ranged from 0.4 (CK19, PR) and 0.8 (EGFR) with an average of 0.56 indicating moderate to strong alignment of the imputed data with the ground truth data (Supplementary Figure **2.10**, Supplementary Figure 2.11). We also evaluated imputation accuracy within patients by modifying the LOOCV approach described above. The modified within-patients LOOCV approach included one biopsy from each patient in the training dataset and used the remaining biopsy from the same patient for testing.

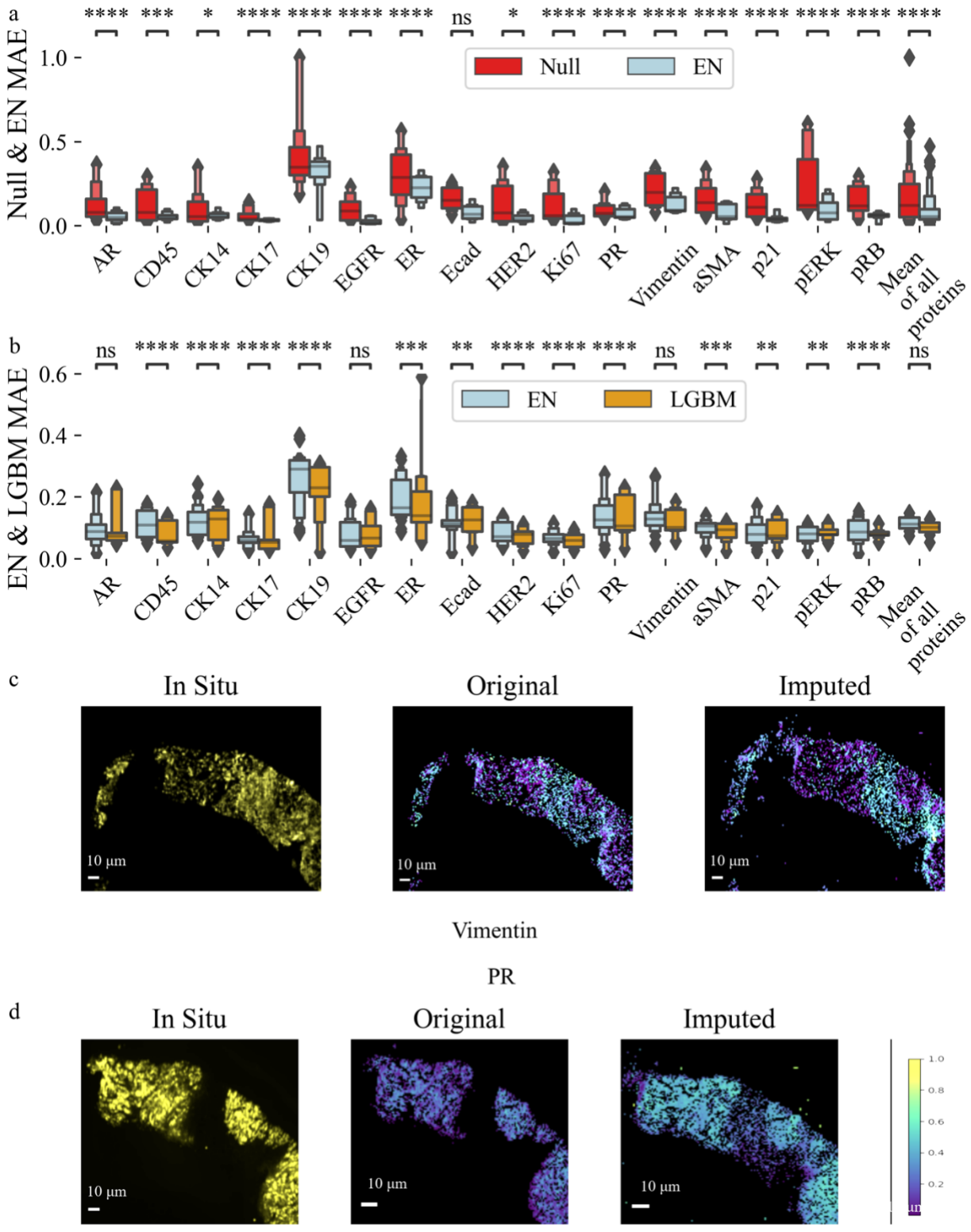


Figure 2.2 Imputation results for null model and elastic net and Light GBM machine learning models across patients. a: Imputation results for all proteins demonstrates improved mean absolute error (MAE) by using Elastic Net (EN) compared to a null model. b: Imputation results using EN & Light GBM (LGBM) show low MAE for imputation for 12

out of 16 available proteins. c-d: Visualization of in situ protein expression, ground-truth single-cell abundance from the image processing pipeline, and imputed single-cell abundance for proteins Vimentin and PR. Results were created using n= 475359 single cells. We used 30 replicates with different train & test splits to validate performance metrics. p-values were calculated using a two-sided Mann-Whitney test and the Benjamini-Hochberg procedure for multiple testing comparisons. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-value: ns: not significant, p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-0

Surprisingly, imputation accuracy in the across-patient LOOCV approach was higher than imputation accuracy in the within patients for some proteins (Supplementary Figure **2.12**). We hypothesize that this performance difference may be attributable to the more diverse training dataset in the across-patient approach. This diverse training dataset may enable imputation models to better handle heterogeneity across patients.

We further assessed imputation performance using additional metrics. Side-by-side visualization of in situ imaging from the original assay, the ground truth single-cell protein abundance values calculated via image processing, and the imputed data show the same tissue structural patterns in all three modalities (Figure **2.2**c, Figure **2.2**d, Supplementary Figure **2.13**). This visual alignment of tissue structure demonstrates that the imputed data preserves the biological structure present in the raw images and the ground truth single-cell data. The adjusted rand index (ARI) and silhouette scores were also calculated using

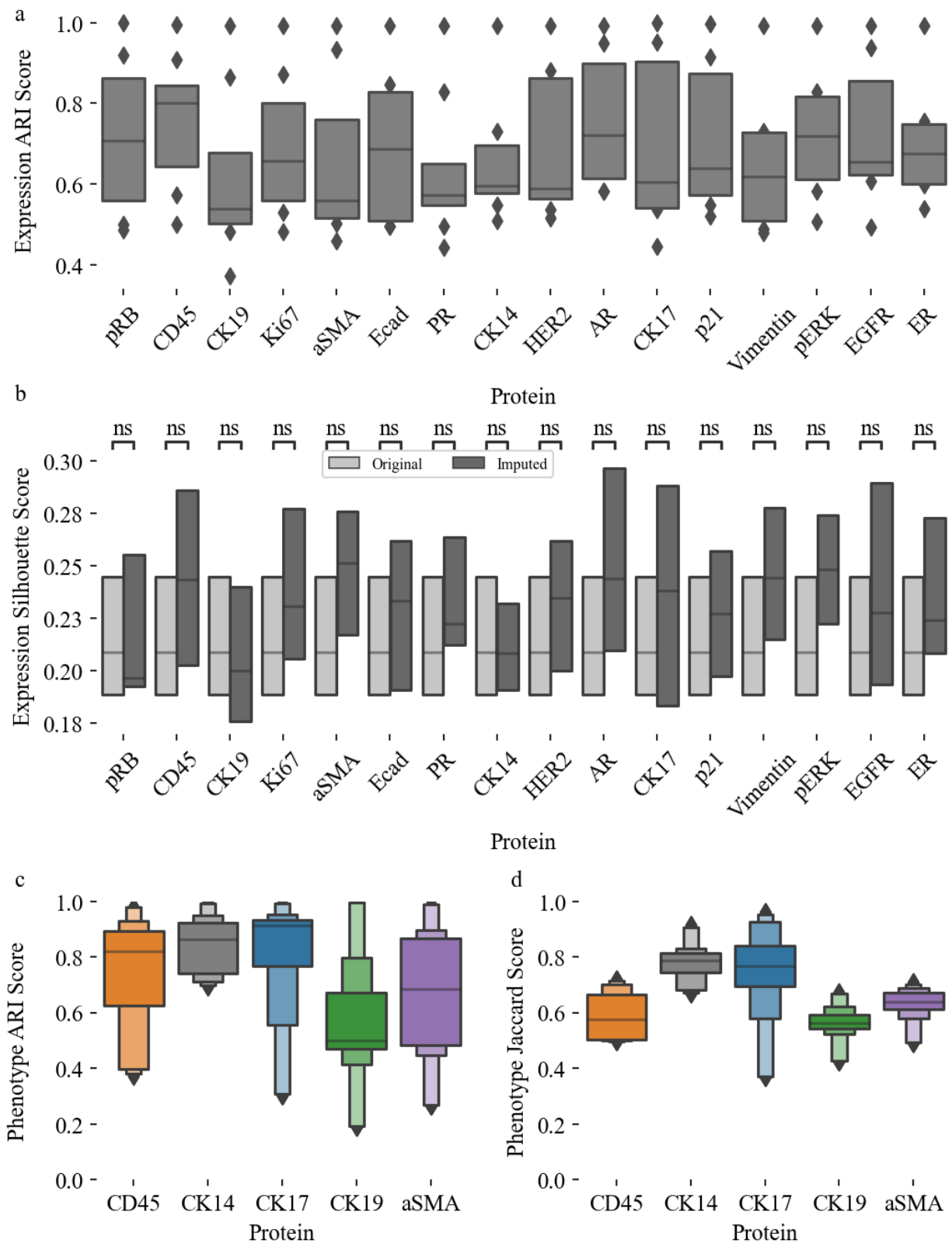

a
Expression ARI Score
1.0
0.8
0.6
0.4
pRB
CD45
CK19
Ki67
aSMA
Ecad
PR
CK14
HER2
AR
CK17
p21
Vimentin
pERK
EGFR
ER
Protein
b
Expression Silhouette Score
ns
Original
Imputed
0.30
0.28
0.25
0.23
0.20
0.18
Protein
c
Phenotype ARI Score
1.0
0.8
0.6
0.4
0.2
0.0
CD45
CK14
CK17
CK19
aSMA
Protein
d
Phenotype Jaccard Score
Protein

Figure 2.3 Cluster metrics and Phenotype calling results between original and imputed values. a: Adjusted Rand Index (ARI) scores demonstrating high similarity between ground truth and imputed data clustering. b: Silhouette scores for single-cell protein abundance clustering improve when using imputed values, indicating improved clustering when using imputed data. c: Adjusted Rand Index (ARI) Scores for cell phenotype matching using ground truth and imputed data showing moderate to strong overlap. d: Jaccard scores show moderate to strong overlap between phenotypes using ground truth and imputed protein expression data. Results were created using n= 475359 single cells. We used 30 replicates with different train & test splits to validate performance metrics. p-values were calculated using a two-sided Mann-Whitney test and the Benjamini-Hochberg procedure for multiple testing comparisons. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values ns: not significant, p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-0

ground-truth and imputed single-cell protein expression values. ARI measures agreement between two clustering results, with a minimum of zero for no clustering agreement and one for perfect agreement. ARI values between ground truth and imputed data range from 0.6 to 0.77 with an average of 0.69, indicating strong cluster agreement between ground truth and imputed protein expression values (Figure **2.3**a). The silhouette score measures cluster quality by assessing how tightly data points in a cluster are grouped. Silhouette scores of ground truth and imputed data show that the cluster quality improved by an average of 1.13% using the imputed data. The silhouette score for imputed data decreased for only CK19 (Figure **2.3**b).

We also performed a single-cell phenotype analysis using the ground truth and imputed data. For each dataset, we assigned phenotypes using the MCMICRO tool suite [82]. ARI between assigned phenotypes from ground truth data and assigned phenotypes from imputed data averaged 0.72, indicating strong to very strong overlap of phenotypes (Figure **2.3**c). Further, we assessed how well an LGBM model can assign phenotypes using either ground truth protein expression data or ground truth data replaced with imputed values for one protein. The model was trained and evaluated using the same across-patient cross

validation approach as our imputation models. To assess the consistency of phenotype predictions based on imputed data, we computed the Jaccard score between predicted phenotypes derived from ground truth data and those obtained using imputed protein values. Since only a subset of proteins in our panel were used for phenotype predictions, the analysis was restricted to these proteins. Jaccard scores ranged from 0.5 to 0.8, with a mean of 0.66, indicating moderate to strong agreement between ground truth- and imputation-based phenotype predictions (Figure **2.3**d). The Adjusted Mutual Info Score (AMI) as well as silhouette scores (**Supplementary Figure 2.14**) were computed for ground truth and predicted phenotypes as well. AMI scores average 0.73 across all proteins, demonstrating high overlap between ground truth and imputed phenotypes. Silhouette scores from ground truth phenotypes range from 0.25 and 0.3, while scores from imputed phenotypes are comparable but somewhat lower, with a range of 0.2 to 0.28 (**Supplementary Figure 2.14**).

Protein abundance imputation using autoencoders and all model comparisons

An autoencoder (AE) is a deep learning neural network for accurate reconstruction of high-dimensional data that include two distinct components: (1) an encoder network that maps a high-dimensional input to a lower-dimensional representation in a latent space and (2) a decoder network that reconstructs the original high-dimensional input from the low-dimensional latent space representation. The goal of an AE is to perform information-

preserving dimensionality reduction of its input to the latent space so that it can then accurately reconstruct the input from the latent representation. AEs have been successfully used for imputation in various biological domains, including single-cell RNA [88,100–102], genomics [91] and more [103]. Unlike LGBM and EN models, AEs can impute multiple features simultaneously due to their ability to fully reconstruct the entire input data. Leveraging this capability, we conducted both single-protein and multi-protein imputation experiments based on the order of protein assays during t-CyCIF's multiple imaging rounds. T-CyCIF involves multiple rounds to stain, incubate, and capture images. Proteins were sequentially removed from each round, and the AE was trained and evaluated for each set of proteins. Initially, proteins from the first round were removed, and the AE was trained and evaluated. This process was then repeated for all proteins in the second round, and so on. To maintain simplicity, no other pairings of proteins were made beyond the rounds.

The AE was trained using biopsies from three patients, including both pre- and post-treatment samples, with biopsies from a fourth patient reserved for validation. Aggregating all biopsy data allowed the AE to develop an internal representation focused on minimizing reconstruction error. During the imputation phase, we initially replaced the target protein's values with the mean expression levels across the dataset. The modified dataset was processed through the AE, which performed continuous cycles of encoding and decoding to iteratively refine the imputed values. For each cycle, the AE replaced the ground truth protein values with the decoded data from the previous cycle. This iterative process was repeated 10 times, as each protein required a different number of optimal iterations for

accurate imputation. We used the mean expression from iterations five to ten as the final imputed value (**Figure 2.4**a).

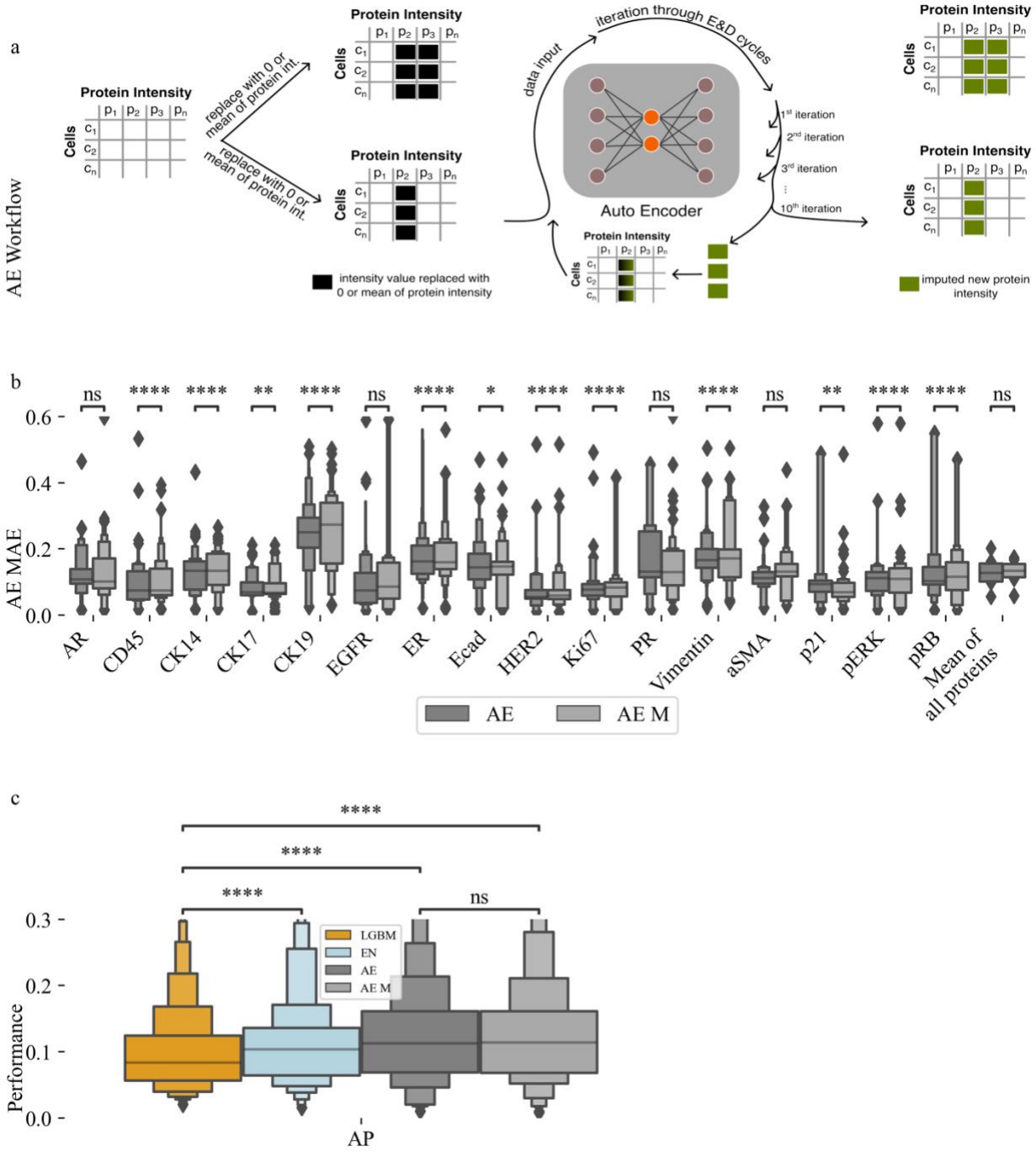


Figure 2.4 Autoencoder imputation results and performance comparison between machine learning models. a: The autoencoder (AE) is trained and then uses an iterative approach to impute single or multiple proteins. To start, proteins to be imputed are replaced with either zero or the mean of the intensity values in the training set. Then, the autoencoder is

used iteratively to predict protein intensities using output values as new input values for each iteration. b: AE single- and multi-protein imputation performance. c: performance comparison between all evaluated machine learning (ML) models shows similar performance overall and that light gradient boosting machine (LGBM) performs best, followed by Elastic Net (EN) and finally AE. There is no significant difference between single and multi-protein imputation performance for AE. Results were created using n= 475359 single cells. We used 30 replicates with different train & test splits to validate performance metrics. p-values were calculated using a two-sided Mann-Whitney test and the Benjamini-Hochberg procedure for multiple testing comparisons. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values: ns: not significant p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-0

AEs accurately imputed proteins in both single- and multi-protein experiments (**Figure 2.4**b). Imputation accuracy of CK19 levels is between 0.15-0.35 MAE, while imputation of the best performing proteins, CK17 and p21, is between 0.05 and 0.10 MAE. Like the EN and LGBM models, imputation performance is worst for the proteins with the most variable abundance levels in our breast cancer cohort, including CK19, ER, and PR. We next compared performance for all three machine learning models used for imputation. Overall, LGBM performed best, followed by the EN and the AEs. These performance differences are consistent between models (**Figure 2.4**c). However, performance differences between the models are relatively modest, with the LGBM achieving a mean accuracy of 0.10 MAE, followed by the EN with a mean accuracy of 0.11 MAE, and the AEs with a mean accuracy of 0.13 MAE (**Figure 2.4**c). Autoencoder performance differences between the imputation of single proteins and that of multi-proteins are minimal.

To further evaluate the performance of imputation in MTI using machine learning, we performed imputation on an additional t-CyCIF dataset from the Human Tumor Atlas Network [58]. This dataset was taken from a breast cancer tissue microarray and included two tissue cores from each of 26 breast cancer tumors. On average, each core included

approximately 9850 cells. Unlike the cancers in our main analysis cohort, these cancers are primary disease rather than metastatic and represent all different subtypes of breast cancer. This data is publicly available at the NCI Human Tumor Atlas Portal (see Data Availability section), and the same primary image processing analysis pipeline used in our main analysis was used to generate a single-cell dataset for imputation. From this single-cell dataset of a breast cancer tissue microarray, we extracted the proteins shared with the primary breast cancer dataset discussed previously.

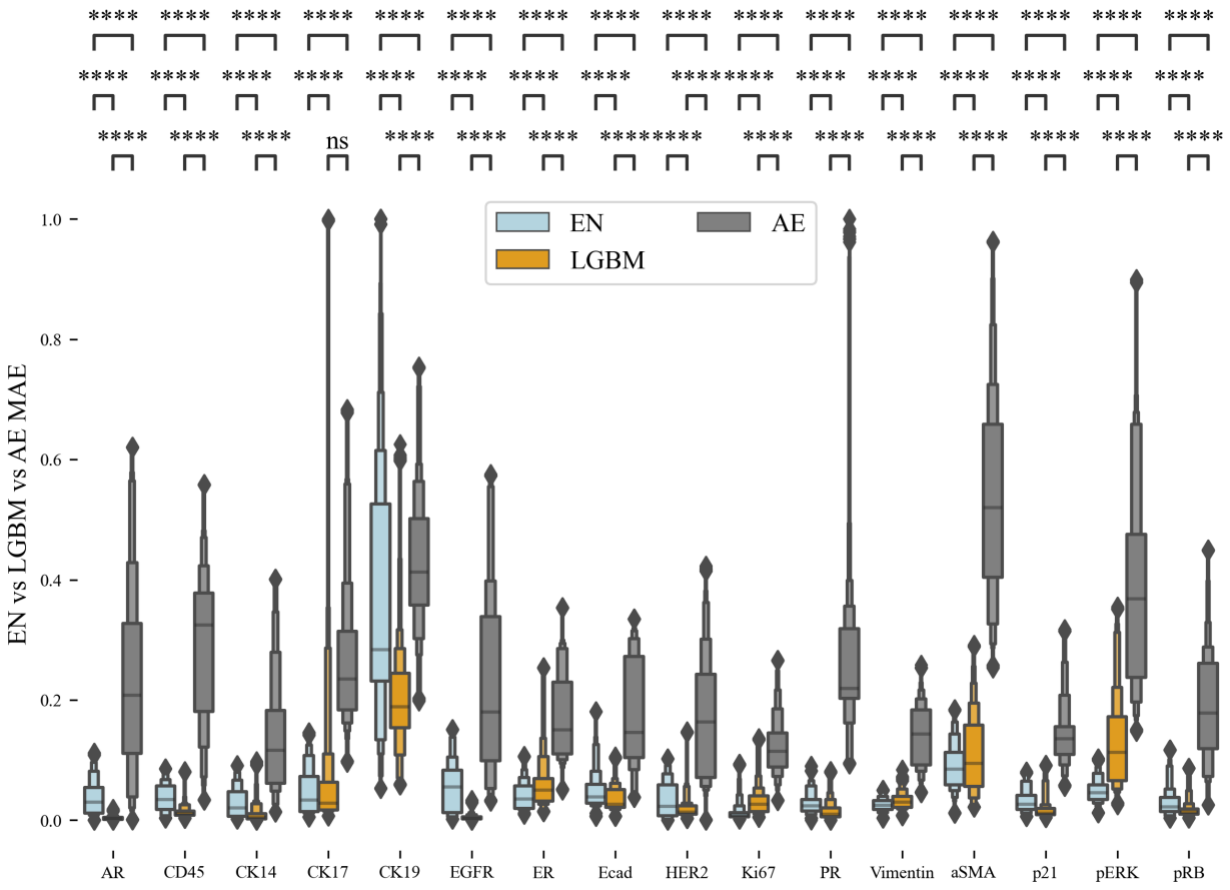


Figure 2.5 Imputation performance of EN, LGBM, and AE machine learning models on an independent t-CyCIF dataset. Dataset was obtained from a breast cancer tissue microarray that includes two cores each from 26 tumors. Imputation results are similar to those obtained in our primary cohort and dataset, showing that our imputation methods are applicable beyond the primary cohort to other cohorts and datasets. Results were created using n= 475359 single cells. We used 30 replicates with different train & test splits to validate performance metrics. p-values were calculated using a two-sided Mann-Whitney test and the Benjamini-Hochberg procedure for multiple testing comparisons. Each

boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values: ns: p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-0

We then replicated our imputation experiments using the EN, LGBM, and AE models, employing the same LOOCV approach as before. Results from this analysis show accurate imputation from EN and LGBM models with approximately the same level of overall accuracy as in our main dataset. Accuracy of the AE models was significantly lower in this dataset as compared to accuracy in our main dataset, but it is unclear why this drop in performance occurred. The LGBM model outperformed all other models, while the AE performed better than the EN but worse than the LGBM (Figure **2.5**, Supplementary Table 2).

Using cellular spatial information to improve imputation

A key advantage of MTI datasets is that the spatial coordinates of each cell are known, making it possible to quantify spatial information around individual cells. We hypothesized that the spatial information available in t-CyCIF could be used to improve imputation performance. To test this hypothesis, we quantified the spatial cellular context surrounding a target cell (cell of interest) by calculating the mean protein abundance of neighboring cells. The average abundance levels of all proteins, excluding the target protein designated for imputation (as its measurement was assumed to be unavailable), were computed across neighboring cells. These averaged neighbor abundances were then incorporated

into the prior set of input features, resulting in a combined feature set comprising both single-cell protein abundances and the mean protein abundances of neighboring cells. (Figure **2.6**a). Importantly, successful imputation did not depend on the presence of the same protein in neighboring cells, as the target protein was assumed to be missing not only from the cell of origin but also from all other cells within the biopsy.

When no neighboring cells were detected, a value of zero was assigned for neighbors' protein abundances. Radii of 15, 30, 60, 90 and 120 micrometers (μm) were used to identify neighboring cells and assess the impact of using different sizes of radii on imputation performance. Only features for one radius setting were used for training a model, and hence a single set of spatial features was included as input for a predictive model.

A radius of 15 μm captures most of the immediate neighbors of a cell, whereas larger radii capture the extended neighborhood of a cell. We evaluated imputation accuracy using added spatial information in only LGBM and AEs because LGBM performed better than EN and AEs can perform multi-protein imputation. Using spatial information improved overall imputation accuracy for LGBM (Figure **2.6**b, Supplementary Figure **2.15**), single protein AE (Figure **2.7**a, Supplementary Figure **2.16**) and multi-protein AE (Figure **2.7**b, Supplementary Figure **2.17**). Importantly, imputation accuracy for proteins that had proven difficult to impute well due to their very high levels of variance (see Supplementary Figure 2.9) was improved significantly with spatial information (Figure **2.7**a). In particular, imputation of CK19, ER, and PR was much more accurate with spatial information. LGBM performance also improved for other proteins such as CK17, CD45, Ecad, ASMA and p21. Performance of the AE achieved improvements in single-protein imputation for most proteins, with CK19

showing the greatest improvement (Figure **2.6**a). Multi-protein imputation also benefited from spatial information integration (Figure **2.7**b).

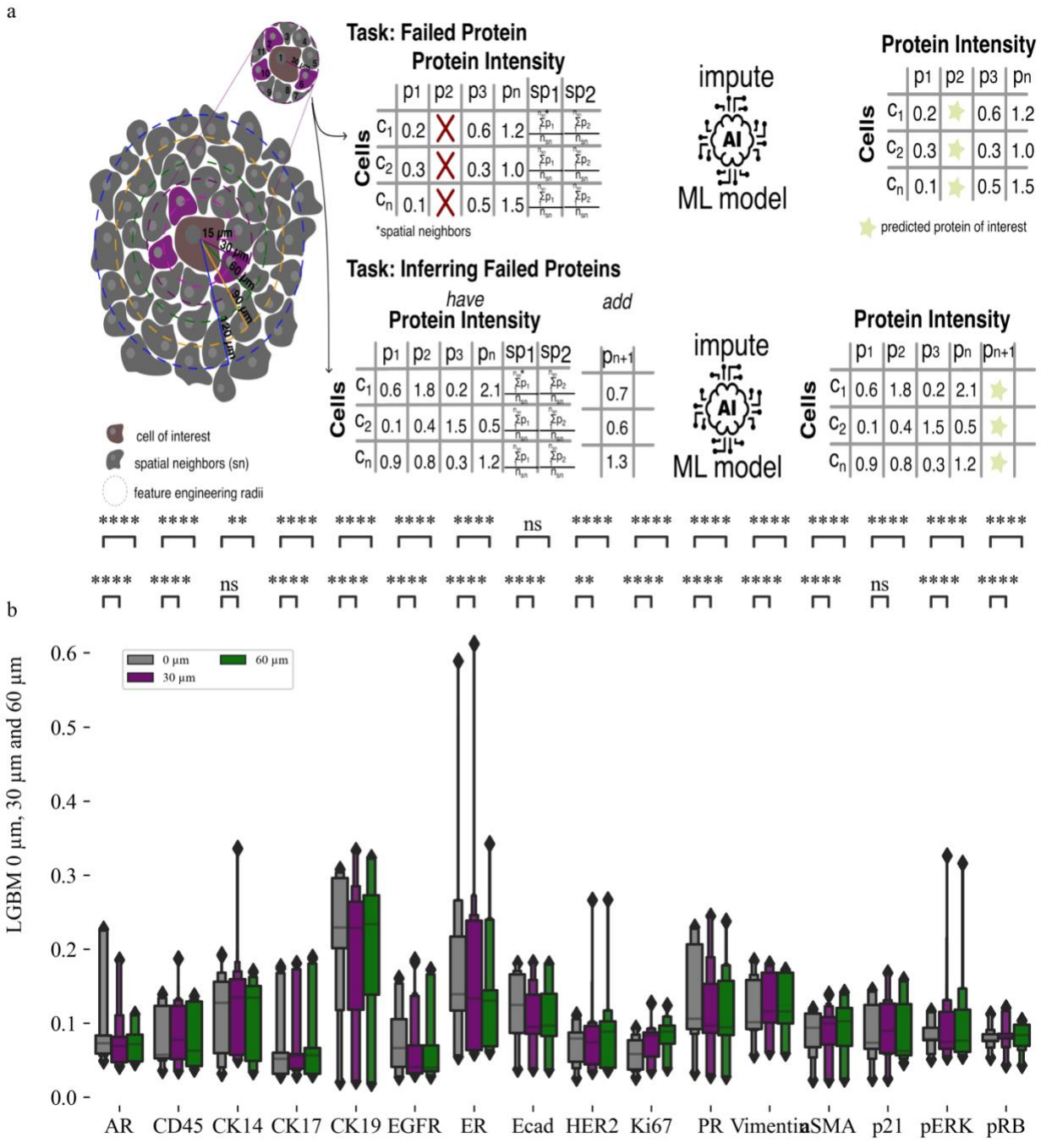


Figure 2.6 Using spatial information improves imputation performance for LGBM. a: Schematic for creating a feature table based on spatial neighbors found in selected radii. Exemplary 15 µm radius is shown. Red marks the cell of interest (or origin) and protein abundance levels of cells in its neighborhood are averaged to get neighborhood abundance levels.

b: Light gradient boosting machine (LGBM) imputation results across patients with mean absolute error (MAE) scores for 0 µm, 30 µm, 60 µm reveal significant improvement for several proteins such as EGFR, ER, ECAD and PR. Results were created using n= 475359 single cells. We used 30 replicates with different train & test splits to validate performance metrics. p-values were calculated using a two-sided Mann-Whitney test and the Benjamini-Hochberg procedure for multiple testing comparisons. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values: ns: p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-04

However, performance gains were not as pronounced compared to the single protein imputation model. Aligned with prior research [104], imputation accuracy generally improves up to a certain neighborhood radius and then plateaus or declines (Table 3, Supplementary Figure **2.15**, Supplementary Figure **2.16**, Supplementary Figure **2.17**). However, the LGBM does not show the same improvement up until a certain radius, but instead remains largely steady, with a peak performance observed using 60 µm (Table 2.3, Supplementary Figure **2.16**). Imputation performance is improved by incorporating spatial information (Supplementary Table **2.5**).

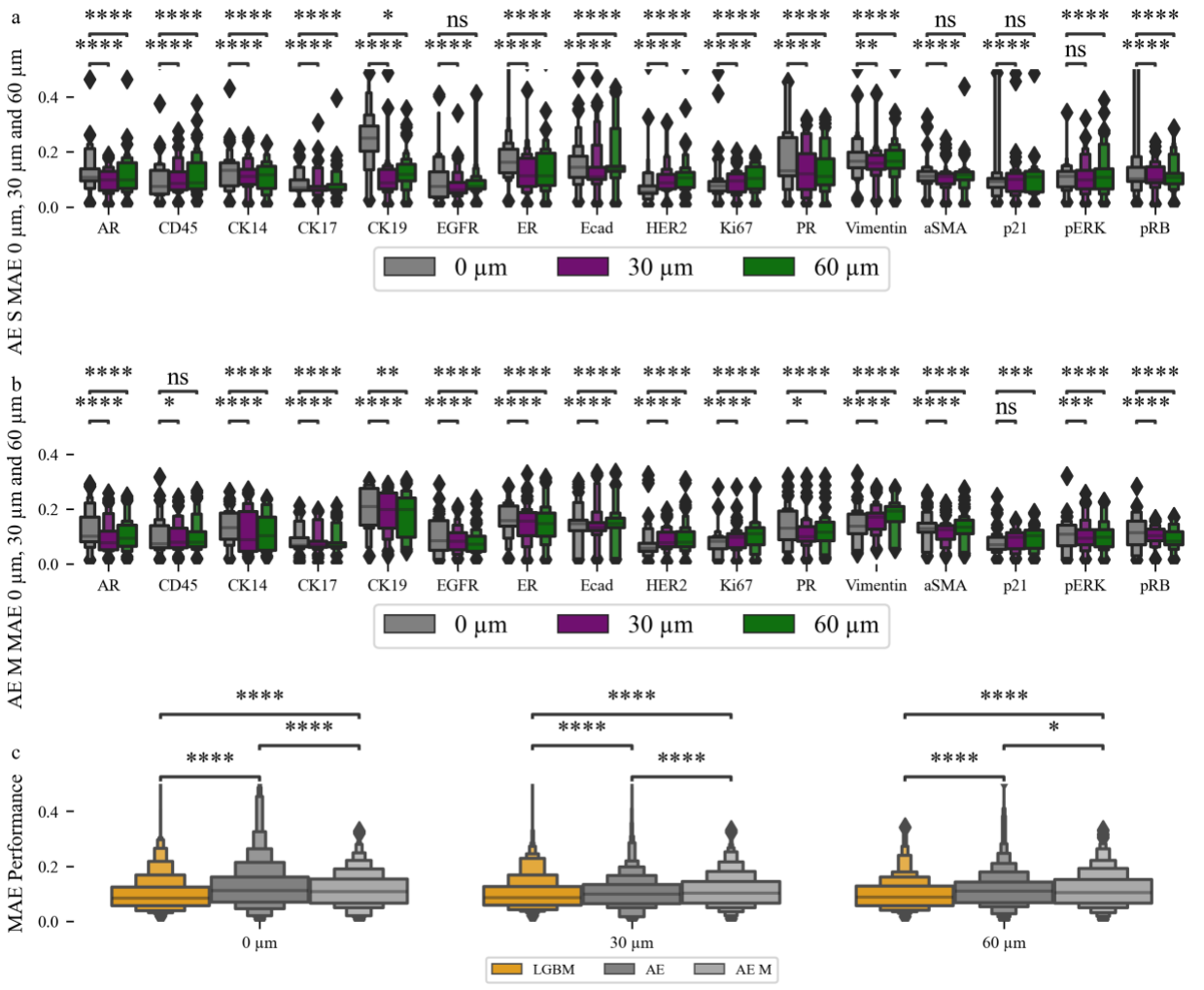


Figure 2.7 Using spatial information improves imputation performance. a: Single protein imputation mean absolute error (MAE) for 0 µm, 30 µm and 60 µm leads to improved imputation accuracy for proteins such as AR, CK14, CK19, ER and more. Proteins for which imputation improved when using spatial information are in bold and underlined. b: Multi-protein imputation MAE scores for 0 µm, 30 µm and 60 µm and leads to improved imputation accuracy for proteins such as AR, CK14, CK19, ER and more. c: Comparison of light gradient boosting machine (LGBM) and autoencoder (AE) imputation performance for 0,30 and 60µm shows similar performance of all models. Results were created using n= 475359 single cells. We used 30 replicates with different train & test splits to validate performance metrics. p-values were calculated using a two-sided Mann-Whitney test and the Benjamini-Hochberg procedure for multiple testing comparisons. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values: ns: p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-04

Using Imputation to Predict Treatment Timepoints of Breast Cancer Cells

To evaluate the utility of imputed single-cell protein values from our machine learning models, we used these imputed values to predict whether cells were in pre-treatment or post-treatment timepoints. Using a machine learning classifier that predicts whether single cells are most likely to come from a pre- or post-treatment, we compared classifier accuracy using three different training datasets: (1) ground truth data; (2) ground truth data with a protein's values removed; and (3) ground truth data with a protein's values imputed. The dataset used for this analysis comprised our primary cohort, consisting of four pre-treatment and four post-treatment biopsies. For modeling, we selected the non-spatial LightGBM (LGBM) model. While the spatial variant demonstrated superior performance overall, it required substantial computational resources. To balance performance and efficiency, the non-spatial LGBM model was used for this analysis.

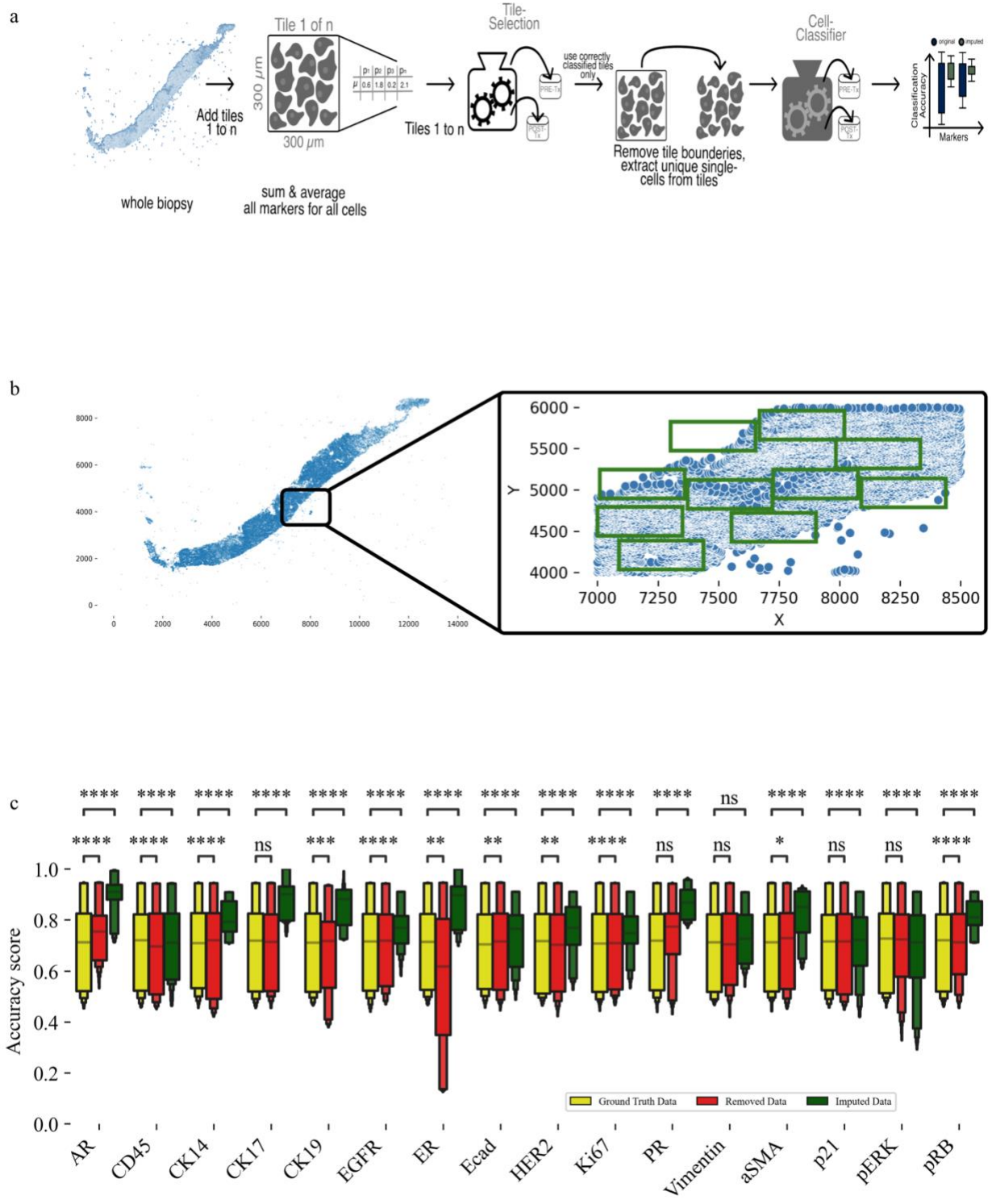


Figure 2.8 Experimental setup and validation for using imputed values to predict treatment timepoints for single cells. a: An initial tile classifier was used to identify tissue strongly associated with treatment timepoints. Next, cells in tissue associated with treatment timepoints were to train a cell classifier to identify whether cells came from pre-treatment or post-treatment biopsies. b: Complete biopsy overview, with a zoomed in view depicting green squares which

show tiles strongly associated with treatment timepoints. c: Classification accuracy (higher bar is better) of the cell classifier shows improved performance using imputed values as compared to performance using ground truth or removed protein values. 13125 tiles were used to run the models, with a replicate count of 30. p-values were calculated using a two-sided Mann-Whitney test and the Benjamini-Hochberg procedure for multiple testing comparisons. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values: ns: p <= 1.00e+00 *: 1.00e-02 < p <= 5.00e-02 **: 1.00e-03 < p <= 1.00e-02 ***: 1.00e-04 < p <= 1.00e-03 ****: p <= 1.00e-04

Initially poor classifier performance was observed because not all biopsy tissues exhibited strong signals associated with a treatment timepoint. To address this issue, we developed a two-step process using two machine learning classifiers (Figure **2.8**a).

First, we selected 300 µm x 300 µm tiles for each biopsy that were associated with treatment timepoint by using a tile machine learning classifier (Figure **2.8**b). The average protein expression of all cells within a tile was used as input to this tile classifier, and the tiles correctly predicted as pre- or post-treatment by the classifier were selected and used for further analysis. Importantly, the protein to be imputed was removed from the input to this classifier. After the tiles were selected, the second step was performed using cells from the selected tiles. In this step, single-cell imputation was used to produce values for a single protein, and the ground truth plus imputed data was used for timepoint prediction using a single-cell classifier. This single-cell classifier was trained to classify single cells as either pre-treatment or post-treatment based on protein abundance levels. Accuracy of this single-cell classifier was compared using both imputed values and the ground truth values for each protein.

To prevent data leakage, we employed LOOCV across patients. For each iteration, a biopsy was designated as the test set, while the remaining biopsies, excluding those from the same patient, were used for training the models. We then compared classification accuracy using the ground truth protein abundance data, a control dataset with a protein's values removed, and an imputed dataset where a single protein's expression was replaced with imputed values. Imputation was performed using the LGBM model because it was highly accurate in prior analyses. No spatial data was used in the imputation model for simplicity and so that a comparison between ground truth and imputed data was straightforward.

Overall, classification accuracy from models using imputed data met or exceeded accuracy from models using ground truth data (Figure **2.8**c). On average, classification accuracy increased by 8.93% when using imputed data compared to the original ground truth data. We hypothesize that this improvement is attributable to the denoising effect inherent to the autoencoder architecture used for imputation. During the encoding and decoding process, the autoencoder likely suppressed technical noise present in the raw data, resulting in cleaner, more biologically relevant feature representations. Consequently, the classifier trained on imputed data benefited from reduced noise interference, enhancing its ability to detect meaningful patterns and improving predictive performance. Reducing noise not only enhances data quality but also complements the potential effects of upsampling via imputation. Upsampling, amplification of meaningful patterns in the data, has been observed in imputation applications previously [105–107]. Classification accuracy using the control datasets with a protein's values removed were

mixed. For ten proteins, removing their values led to equal or reduced accuracy, but accuracy slightly increased when removing six proteins. When removing a protein increases a model's accuracy, this suggests that the protein is introducing noise to the predictions and possibly harming the classifiers' ability to accurately classify cells. Accuracy of the control models provides additional evidence that the imputation process may be removing noise from the ground truth data, which in turn may help explain higher classification accuracy with imputed data.

Discussion

In this study, we utilized machine learning models to accurately impute single-cell protein abundance levels in breast cancer tissue using datasets obtained from the t-CyCIF multiplexed tissue imaging (MTI) assay. Our datasets comprised eight biopsies from a cohort of four metastatic breast cancer patients, facilitating the training and evaluation of these models. Within a range of [0,1], the imputation performance for most proteins exhibited a mean absolute error (MAE) between 0.05 and 0.15. However, proteins with high variance in our cohort, such as CK19 and ER, were more challenging to impute, with MAE ranging from 0.15 to 0.35. Additional evaluation approaches, in-situ visualization, adjusted rand index (ARI), and silhouette scores, also provide evidence that the imputed values are highly comparable to ground truth data. Further, a single-cell phenotyping analysis showed that phenotypes called via imputed and ground truth data are very similar. These results

demonstrate that imputed data maintains key information needed for biological applications.

The LGBM model, a gradient-boosted regression tree approach, showed modestly better accuracy than an Elastic Net (EN) model or a deep learning autoencoder (AE). Incorporating spatial features into the models, represented by neighboring cells' protein abundance levels, enhanced their accuracy and reduced the average MAE by 0.02. This improvement was particularly significant for proteins with high variance that were otherwise difficult to impute. This use of spatial information complements recent research indicating that cell communication may vary and requires careful evaluation using multiple cellular neighborhoods [104]. Our results are concordant with this observation as they show a similar pattern of improved protein performance using a diverse set of radii.

While the LGBM shows the overall best performance, there are tradeoffs to consider when choosing a machine learning model for single-cell protein abundance imputation in MTI datasets. Traditional ML models such as LGBM and EN can only impute one protein per model, which requires training and storing a model for each protein to be imputed. Using a single model for each protein is time and cost inefficient. In contrast, an AE can impute multiple proteins at once and even all proteins included in their training data, requiring only a single training session and model. While AEs perform marginally worse than LGBM and EN for protein imputation, their capability for multi-protein imputation offers an advantage in reduced training time and cost. Multi-protein imputation, as opposed to sequential imputation, also models inter-protein relationships, and potentially yields more biologically pertinent relationships to explore.

We have demonstrated robust performance of our imputation methods and biological significance of imputed values. Using an independent MTI dataset from a cohort of 26 breast cancers that included all major subtypes of the disease, our imputation methods showed similar performance to that in our primary cohort. Imputation results across these datasets suggest that our machine learning methods are versatile and can potentially be used in other MTI datasets.

To demonstrate the biological significance of our imputed protein abundance data, classification models were built with the imputed data and used to predict whether individual cells originated from pre-treatment or post-treatment biopsies. Accuracy of these models was high, indicating that single cells can often be classified as associated with a particular disease state. This aligns with previous results that similarly show how single cells can be classified based on disease or treatment state [108–110]. Surprisingly, using imputed data improved classification accuracy compared to using ground truth data. Results from classification models using ground truth data or control data with a protein removed suggest that imputed data may be denoising the ground truth data and leading to improved accuracy. This denoising may reduce errors in the data obtained from the multiplex tissue imaging assay or the primary image analysis workflow used to generate the ground truth data. It is also possible that imputation is enhancing patterns in the ground truth data via upsampling [111,112], and this upsampling via imputation is leading to improved classification accuracy using the imputed data.

Limitations of this work include a focus on protein abundance rather than RNA expression, the small number of proteins used for imputation, the cohort composition of metastatic breast cancers, and the small sample size. These analyses demonstrate that it is possible to impute protein abundance in MTI, but imputation of RNA expression has not been explored. This analysis used the sixteen proteins that were shared amongst all biopsies, and it is uncertain if other proteins can be imputed as accurately as these sixteen. This analysis also focused on breast cancer biopsies and diseased tissue, and imputation results may be different in healthy tissue or in other diseases. A study like ours would benefit from using a larger and more diverse cohort and from different MTI assays with more proteins. Different and larger datasets would help establish the robustness and generalizability of our imputation methods.

In summary, this study demonstrates that machine learning can effectively impute biologically meaningful single-cell protein abundance levels using MTI datasets. Our results provide a foundation for future applications of machine-learning imputed data in single-cell MTI datasets. One potential application is imputation of additional single-cell and cellular neighborhood features, which can in turn aid in understanding tissue ecosystems. Another future application is the use of imputed datasets to predict biomedical outcomes such as tissue response to perturbations or, in the case of disease, response to therapy.

Methods

Ethical Statement

All biospecimens and data were collected under the single-center, observational study Molecular Mechanisms of Tumor Evolution and Resistance to Therapy (IRB#16113). The study was reviewed and approved by the 508 Oregon Health & Science University (OHSU) Institutional Review Board (IRB). All datasets used for training and testing machine learning models are publicly available. Data generation and handling were conducted in accordance with the National Cancer Institute Human Tumor Atlas Network (HTAN) data standards, which are publicly accessible at: https://humantumoratlas.org/standards. All datasets used for training and testing machine learning models are publicly available.

Experimental Setup

The BOND RX Automated IHC/ISH Stainer was used to bake FFPE slides at 60°C for 30 minutes, to dewax the sections using the Bond Dewax solution at 72°C, and for antigen retrieval using Epitope Retrieval 1 (Leica™) solution at 100°C for 20 minutes. Slides underwent multiple cycles of antibody incubation, imaging, and fluorophore inactivation. All antibodies were incubated overnight at 4°C in the dark. Slides were stained with Hoechst 33342 for 10 minutes at room temperature in the dark following antibody incubation in every cycle. Coverslips were wet-mounted using 200 µL of 10% Glycerol in PBS prior to imaging. Images were acquired using a 20x objective (0.75 NA) on a CyteFinder

slide scanning fluorescence microscope (RareCyte Inc. Seattle WA). Fluorophores were inactivated using a 4.5% H_2O_2, 24 mM NaOH/PBS solution and an LED light source for 1 hour.

The detailed protocol is available in protocols.io (dx.doi.org/10.17504/protocols.io.bjiukkew).

Data Preparation

The source files include X and Y spatial coordinates and bio-morphological information (orientation, area, extent, etc.) for each cell. These features are removed for the initial imputation experiments, which solely rely on protein information.

To prepare the available data for the machine learning models and deep learning networks, we used Min-Max Scaling to scale features to be in the [0,1] range.

Statistical Validity:

For robust statistical validity, we conducted more than 30 experiments ($n > 30$) for each protein imputation and each model.

Elastic Net

Elastic Net regression experiments were conducted using the ElasticNetCV implementation from the scikit-learn library, which performs internal cross-validation to select optimal model parameters. Prior to modeling, all data were normalized. Patient-wise leave-one-out cross validation was used. One patient and all data from that patient was placed into a test dataset, and the remaining data from all other patients was used as the training dataset. This approach was done for each patient to create a training and test dataset for each patient that prevented data leakage. The model was trained on the training set and evaluated on the corresponding test set using performance metrics such as mean absolute error (MAE). A separate model was trained for each protein, treating each protein as an individual prediction target.

To ensure statistical robustness and reproducibility, all experiments were repeated more than 30 times ($n > 30$), each with a distinct random seed.

Light GBM

To set up a training and evaluation pipeline for our Light GBM [98] model, we used the Ludwig [113] platform, which enables “End-to-end machine learning pipelines” in a low code environment. Ludwig models were configured using a YAML-based configuration file specifying the input features, proteins in this study, and the target variable to be imputed. To streamline and scale this process, we automated model setup and execution using a combination of shell scripts and Makefiles. Consistent with all other experiments, train and

test sets were generated under a strict constraint: patients selected for evaluation were excluded from the training data and used exclusively for testing. Each run was initialized with a unique random seed to support reproducibility. Performance metrics generated by Ludwig were logged and retained for downstream analysis.

Autoencoder

An autoencoder-based approach was employed to iteratively impute missing protein values. First, the preprocessed source data were loaded, and train-test splits were generated following the same constraint applied throughout the study: patient data designated for evaluation were excluded from the training set and used solely for testing. To impute a specific protein, its values were initially replaced with the mean of all available entries for that protein.

The prepared dataset was then passed through the autoencoder, which performed a full encoding and decoding cycle (Figure **2.4**). The resulting output was stored as an intermediate representation for future reference and downstream analysis. From this output, the imputed values for the target protein were extracted and used to replace the initial mean values. This process was repeated for a total of 10 iterations, forming a 10-step iterative imputation pipeline.

Final imputation performance was assessed using the last five iterations: for each patient, the mean imputed value across these five decoding steps was calculated and used to compute the final mean absolute error (MAE).

To ensure statistical robustness, each model configuration was run a minimum of 30 times, each initialized with a distinct random seed to generate reproducible yet variable results.

Statistics & Reproducibility:

Each model was evaluated using cross-fold validation with at least 30 iterations to ensure robust performance estimation. To assess differences in performance between the original and imputed data, we conducted two-sided Mann-Whitney U tests. To account for multiple comparisons, we applied the Benjamini-Hochberg correction. No statistical method was used to predetermine sample size.

No datapoints were excluded from the analyses; the entire available single-cell dataset was used. The models were not trained on the data they were later tested on. Blinding was not needed nor applied during model training and evaluation. Investigators were not blinded to allocation during experiments and outcome assessment.

Data availability

The single-cell spatial cyclic immunofluorescence data analyzed in this study and results generated from model predictions have been deposited in the Dataverse database under DOI https://doi.org/10.7910/DVN/RBIJSQ.

Table 4 lists the HTAN Biopsy and Biospecimen IDs used in this study. More information on these biopsies and biospecimens is available at https://humantumoratlas.org/explore

All TMA data is available through the HTAN Data Portal as part of the HTAN TNP-TMA Project (https://data.humantumoratlas.org/)

Code availability

The source code of this work is freely available in the GitHub repository.

https://github.com/goeckslab/MTIProteinImputation

Acknowledgements

This research was supported by the National Cancer Institute (NCI) of the National Institutes of Health grants U24CA231877, U24CA284167, and U2CCA233280 and by funding from the Prospect Creek Foundation to the OHSU SMMART (Serial Measurement of Molecular and Architectural Responses to Therapy) Program.

Contributions

R.K and J.G both conceived the study. C.W. and A.C. processed the primary image data to produce the single-cell datasets analyzed in this work. R.K. developed the methods, wrote the code, and performed the analysis. C.W., A.C., and J.G. provided feedback and suggestions on methods, code, and analyses. K.K. designed illustrations used in the figures and provided feedback on the overall figure design. All authors read and approved the final paper.

Competing interests

The authors declare no competing interests.

Tables

Table 2.3: Mean and standard deviation of performance for LGBM and AEs for different radii. Mean and (standard deviation) for each model are listed. 0μm is considered as baseline without any use of spatial information.

Network	**0 μm**	**30 μm**	**60 μm**
LGBM	0.10 (0.06)	0.10 (0.06)	0.10 (0.05)
AE Single Protein	0.13 (0.09)	0.10 (0.06)	0.11 (0.06)
AE Multi Protein	0.12 (0.09)	0.11 (0.06)	0.12 (0.07)

Table 2.4: Human Tumor Atlas (HTAN) biopsy and biospecimen IDs

HTAN Biopsy ID	**HTAN Biospecimen ID**
9 2 1	HTA9_2_11
9 2 2	HTA9_2_21
9 3 1	HTA9_3_11
9 3 2	HTA9_3_21
9 14 1	HTA9_14_6
9 14 2	HTA9_14_14
9 15 1	HTA9_15_7
9 15 2	HTA9_15_15

Figures

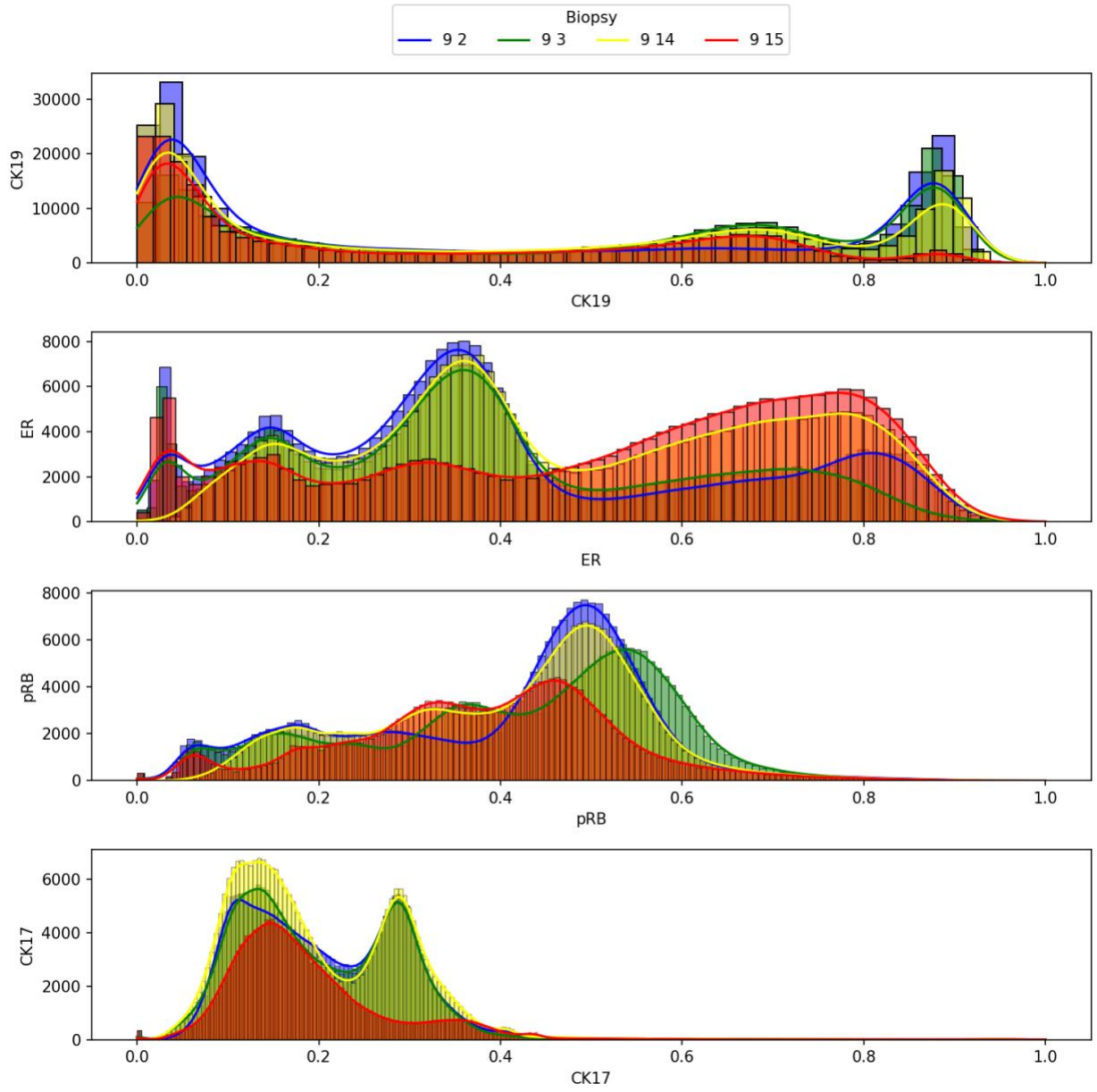
Supplementary Table 2.5 Mean and (Standard Deviation) of all observed radii as well as the baseline for LGBM, AE Single and Multi-Imputation using an Across-Patient Setup.

Network	0 µm/ Baseline	15 µm	30 µm	60 µm	90 µm	120 µm
LGBM	0.099 (0.060)	0.101 (0.059)	0.101 (0.059)	0.097 (0.055)	0.104 (0.062)	0.106 (0.068)
AE Single Protein	0.128 (0.089)	0.100 (0.064)	0.105 (0.059)	0.113 (0.062)	0.120 (0.076)	0.119 (0.075)
AE Multi Protein	0.120 (0.087)	0.115 (0.068)	0.112 (0.062)	0.117 (0.068)	0.120 (0.081)	0.121 (0.081)

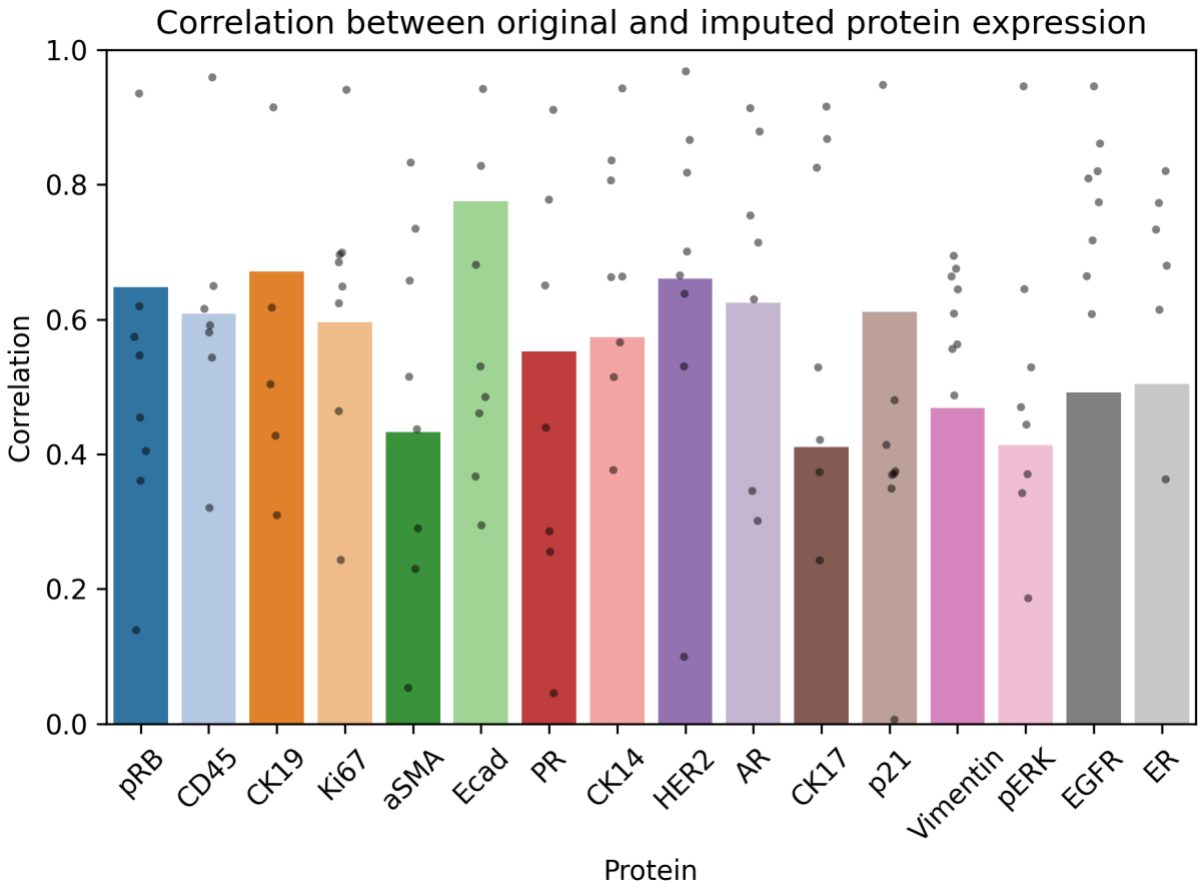
Supplementary Table 2.6 Variance for each observed protein. CK19 shows to highest variance across all patients, followed by Vimentin and ER. Lowest Variance can be observed for Proteins p21, PR and CK17.

Protein	Variance
CK19	1.28
Vimentin	0.42
ER	0.32
pERK	0.24

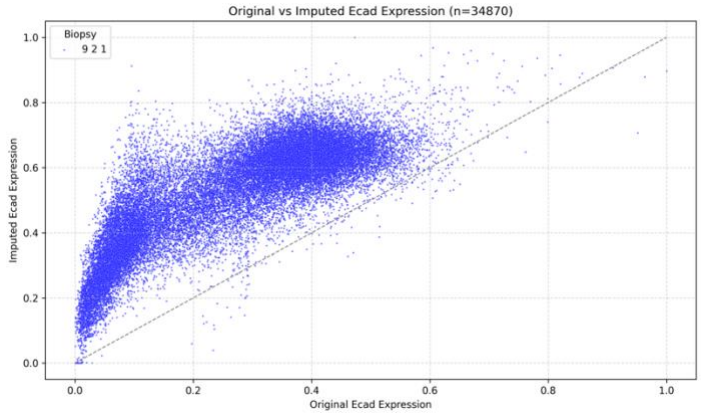
aSMA	0.22
pRB	0.18
Ecad	0.14
AR	0.13
CK14	0.11
CD45	0.09
HER2	0.09
Ki67	0.08
EGFR	0.06
P21	0.05
PR	0.05
CK17	0.05

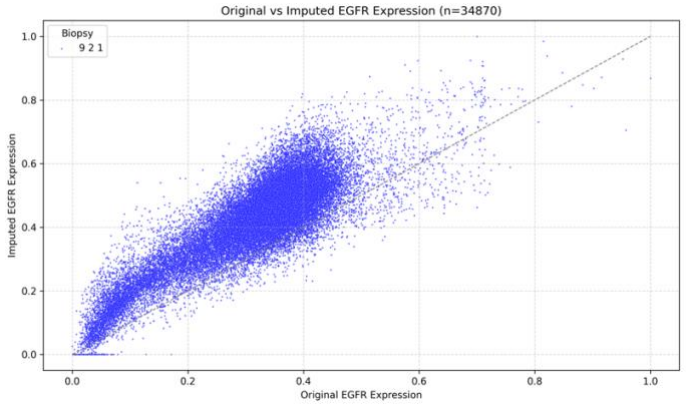


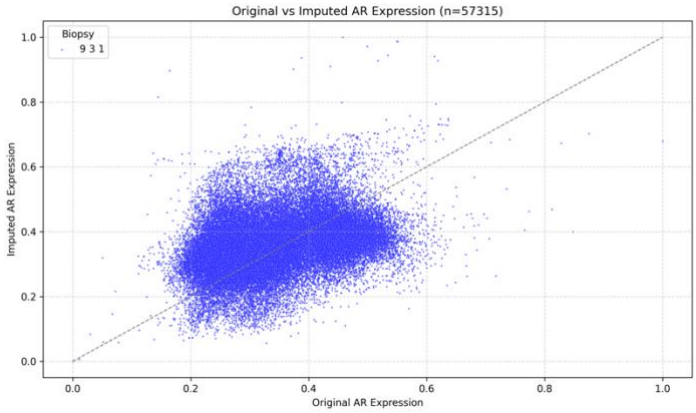
Supplementary Figure 2.9: Protein expression distribution for four proteins (CK19, ER, pRB, CK17). Proteins CK19 and ER experience the highest variance. The distribution is heterogeneous across both patients and proteins.

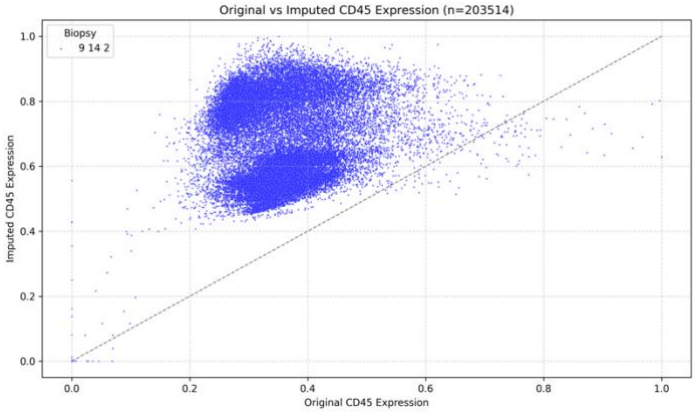


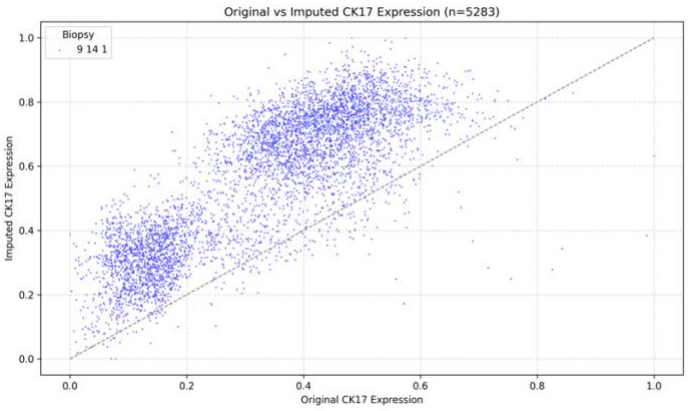
Supplementary Figure 2.10 Pearson correlation coefficients between original and imputed protein abundance values show strong correlations for most proteins. The highest correlation can be observed for EGFR ($r \geq 0.75$) indicating very strong relation. Notable exceptions are CK19 and PR with a correlation coefficient of $r \approx 0.4$, indicating moderate relation.

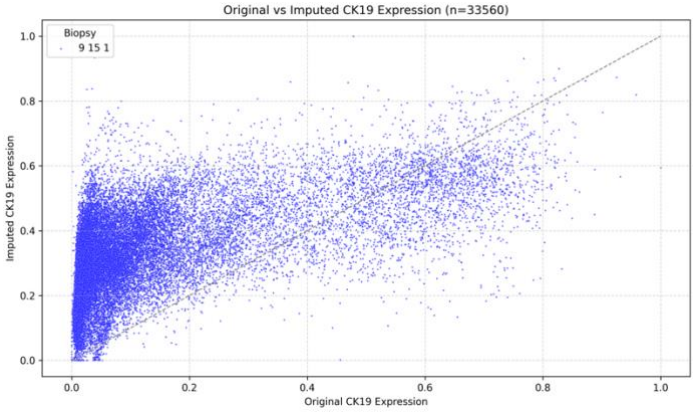

Original vs Imputed Ecad Expression (n=34870)
Biopsy
9 2 1
Imputed Ecad Expression
Original Ecad Expression
0.0
0.2
0.4
0.6
0.8
1.0

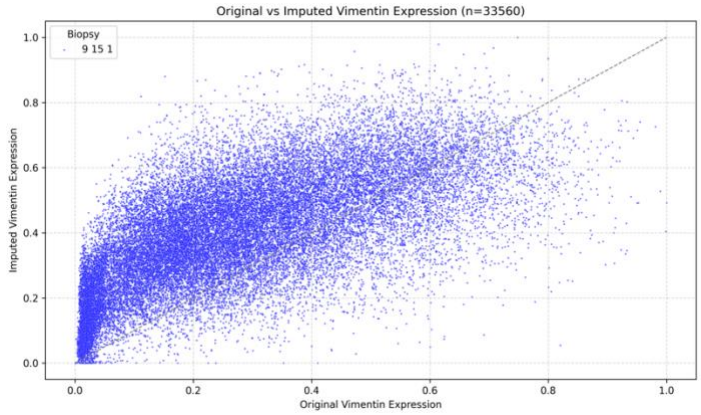

Original vs Imputed EGFR Expression (n=34870)
Biopsy
9 2 1
Imputed EGFR Expression
Original EGFR Expression
0.0
0.2
0.4
0.6
0.8
1.0

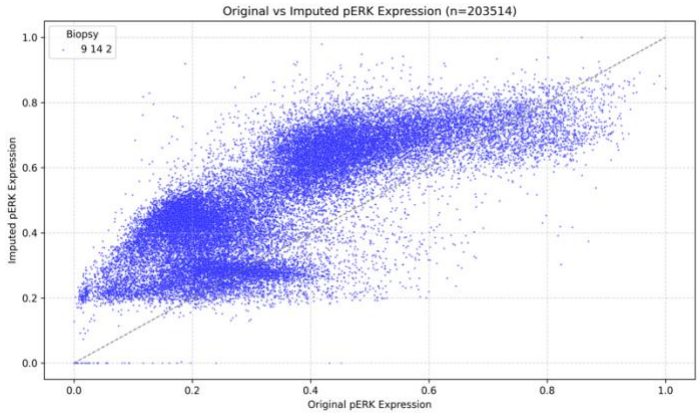

Original vs Imputed AR Expression (n=57315)
Biopsy
9 3 1
Imputed AR Expression
Original AR Expression
0.0
0.2
0.4
0.6
0.8
1.0

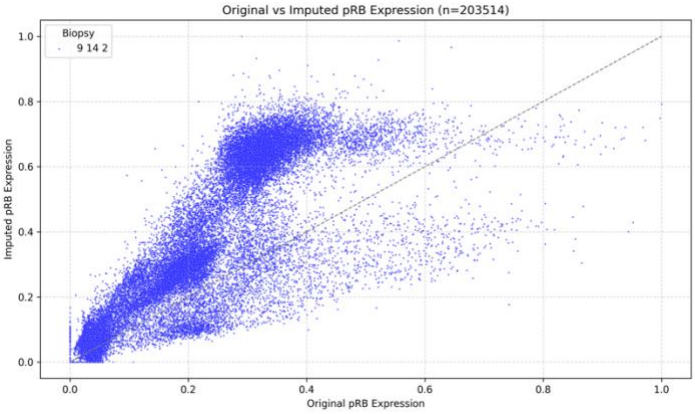

Original vs Imputed CD45 Expression (n=203514)
Biopsy
9 14 2
Imputed CD45 Expression
Original CD45 Expression
0.0
0.2
0.4
0.6
0.8
1.0

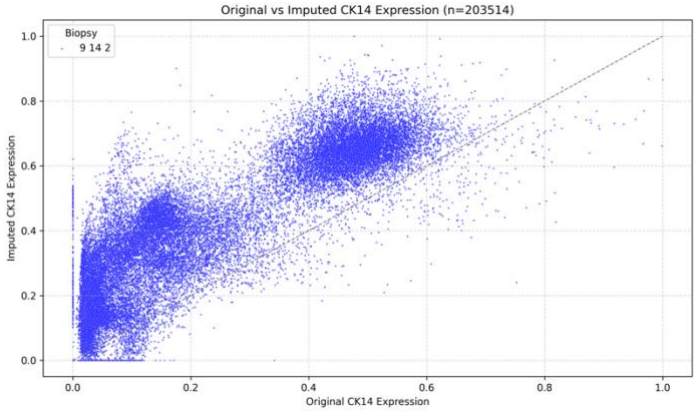

Original vs Imputed CK17 Expression (n=5283)
Biopsy
9 14 1
Imputed CK17 Expression
Original CK17 Expression
0.0
0.2
0.4
0.6
0.8
1.0

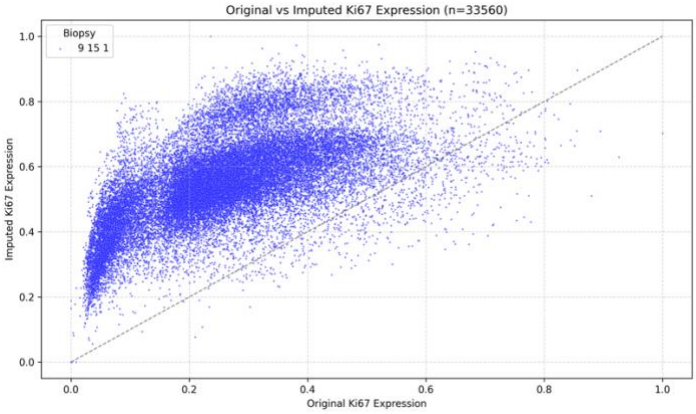

Original vs Imputed CK19 Expression (n=33560)
Biopsy
9 15 1
Imputed CK19 Expression
Original CK19 Expression
0.0
0.2
0.4
0.6
0.8
1.0

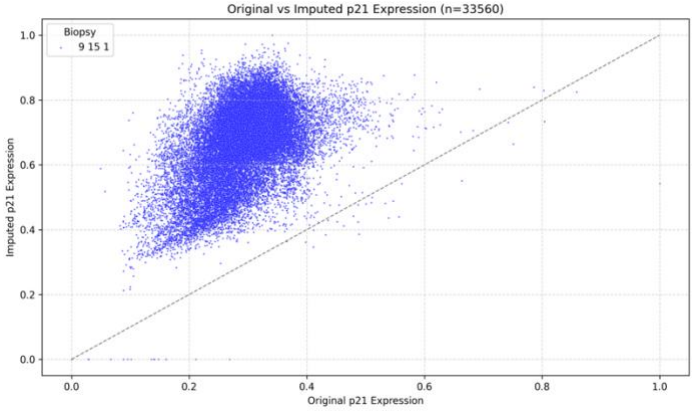

Original vs Imputed Vimentin Expression (n=33560)
Biopsy
9 15 1
Imputed Vimentin Expression
Original Vimentin Expression

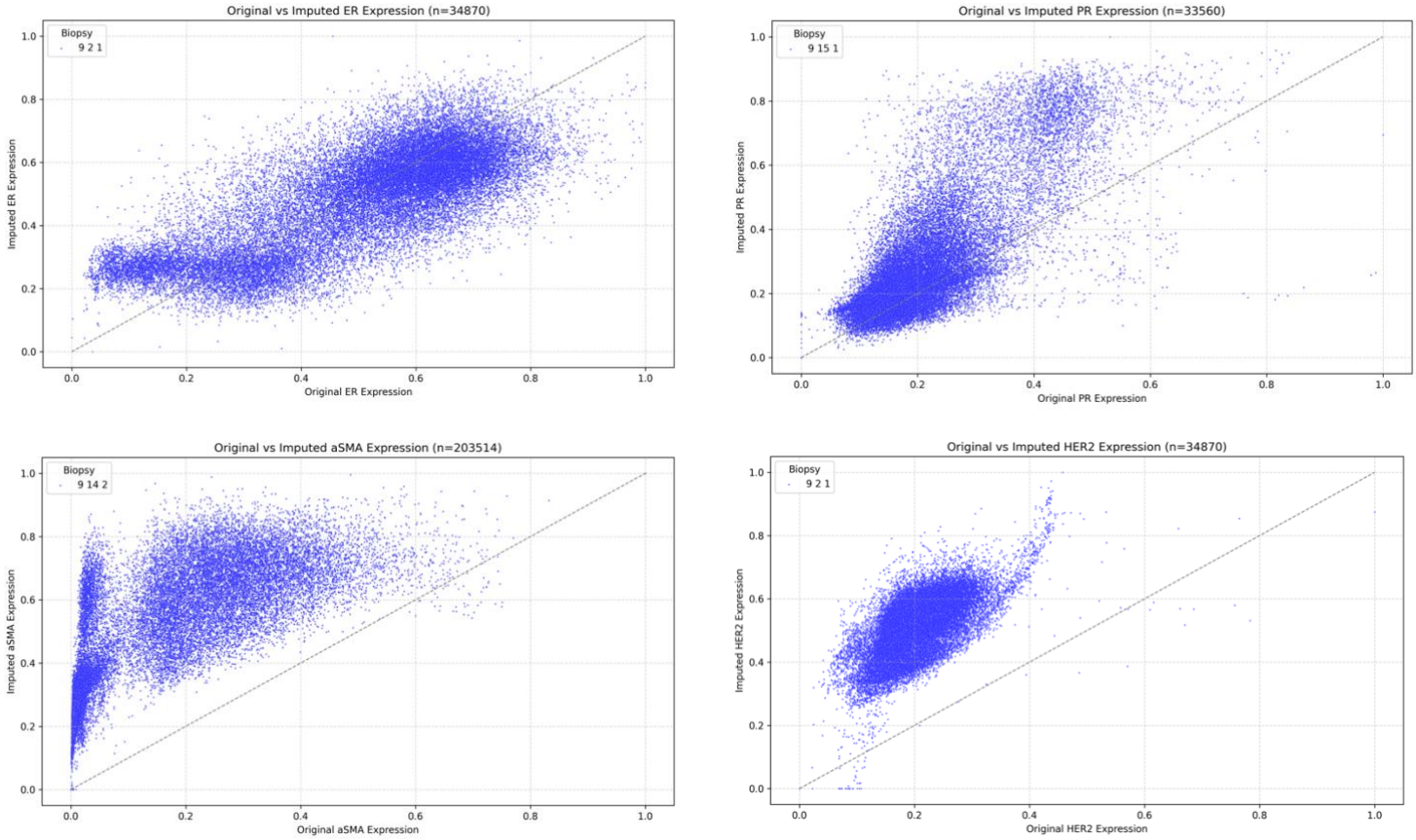

Original vs Imputed pERK Expression (n=203514)
Biopsy
9 14 2
Imputed pERK Expression
Original pERK Expression


Original vs Imputed pRB Expression (n=203514)
Biopsy
9 14 2
Imputed pRB Expression
Original pRB Expression

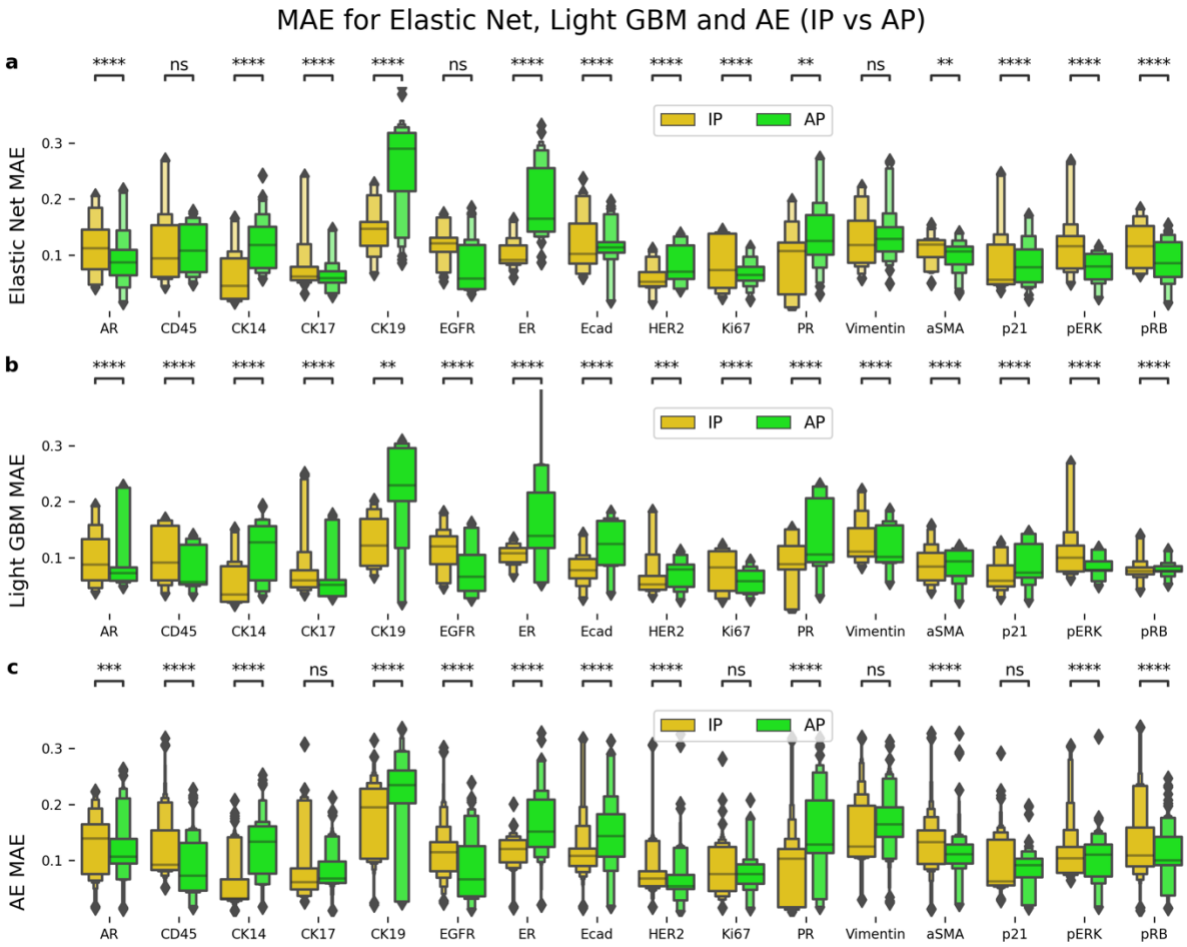

Original vs Imputed CK14 Expression (n=203514)
Biopsy
9 14 2
Imputed CK14 Expression
Original CK14 Expression


Original vs Imputed Ki67 Expression (n=33560)
Biopsy
9 15 1
Imputed Ki67 Expression
Original Ki67 Expression


Original vs Imputed p21 Expression (n=33560)
Biopsy
9 15 1
Imputed p21 Expression
Original p21 Expression



Supplementary Figure 2.11 Representative scatter plots illustrating the relationship between ground truth and imputed data across all proteins demonstrate moderate to strong correlation. The red line represents the theoretical line of perfect agreement between original and imputed values.



Supplementary Figure 2.12 Performance comparison between In Patient (IP) and Across Patient (AP).In Patient (IP, golden) is compared to Across Patient (AP, green) across models. a) Elastic Net (top), b) Light GBM (middle) and c) Autoencoder (AE, bottom). For means of comparison the Mean Absolute Error (MAE) was calculated. Statistical analysis using Mann-Whittney-Wilcoxon and multi-hypothesis testing using Benjamini-Hochberg correction. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values:

ns: not significant, $p \leq 1.00e+00$

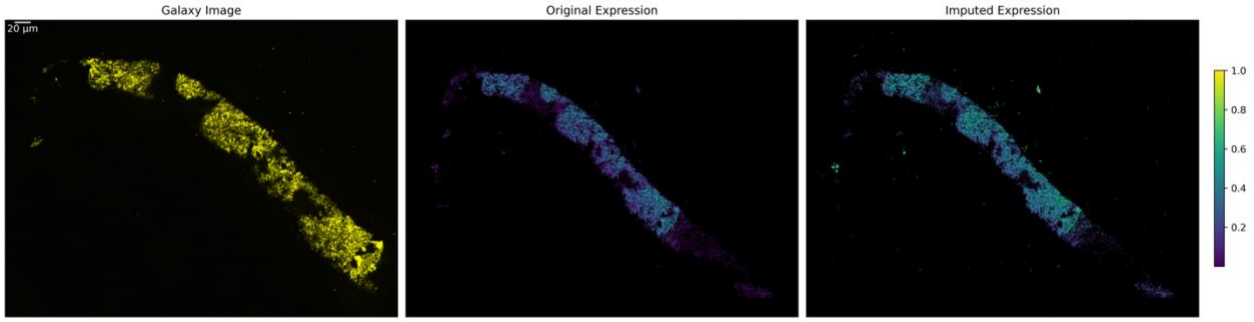
*: $1.00e-02 < p \leq 5.00e-02$

**: $1.00e-03 < p \leq 1.00e-02$

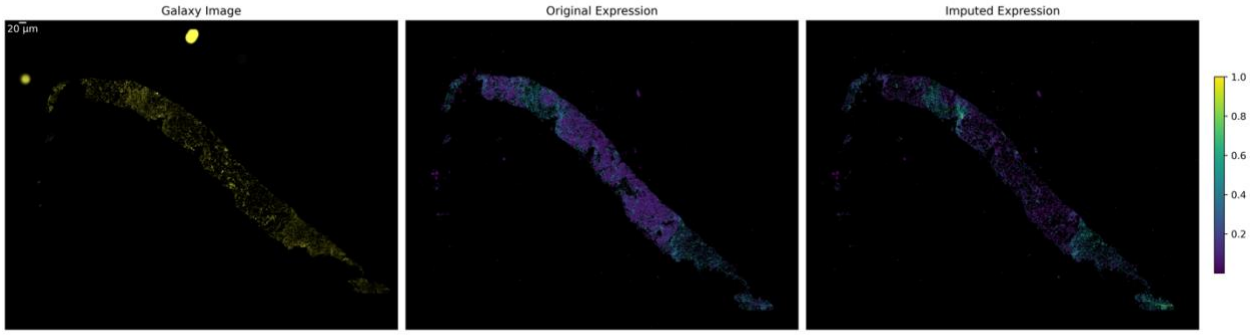
***: $1.00e-04 < p \leq 1.00e-03$

****: $p \leq 1.00e-04$

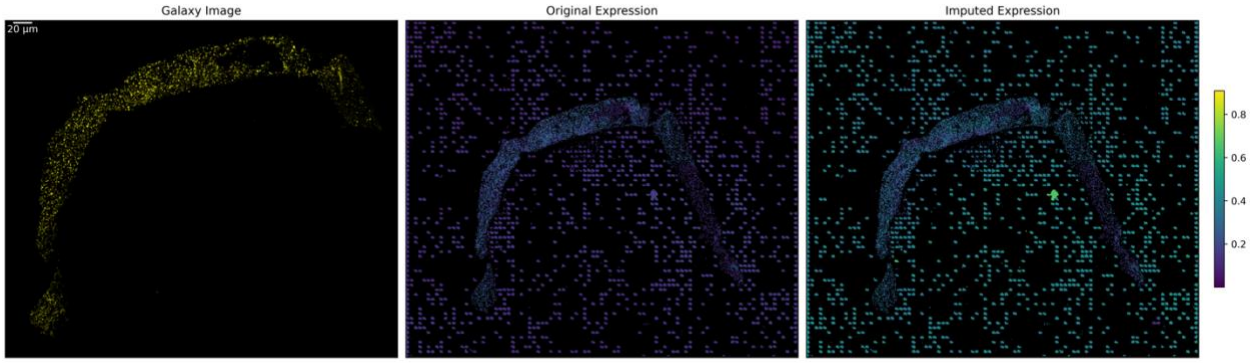
9 2 1 - PR



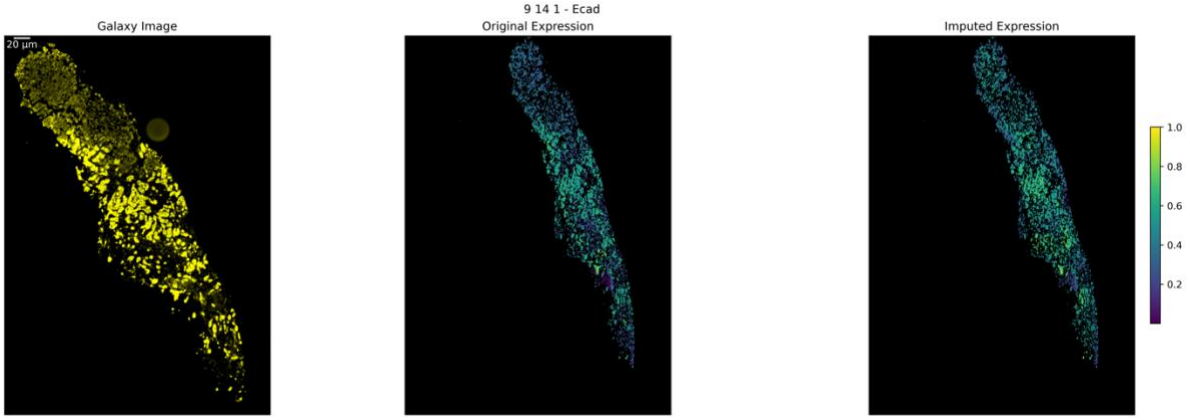
9 2 1 - Vimentin

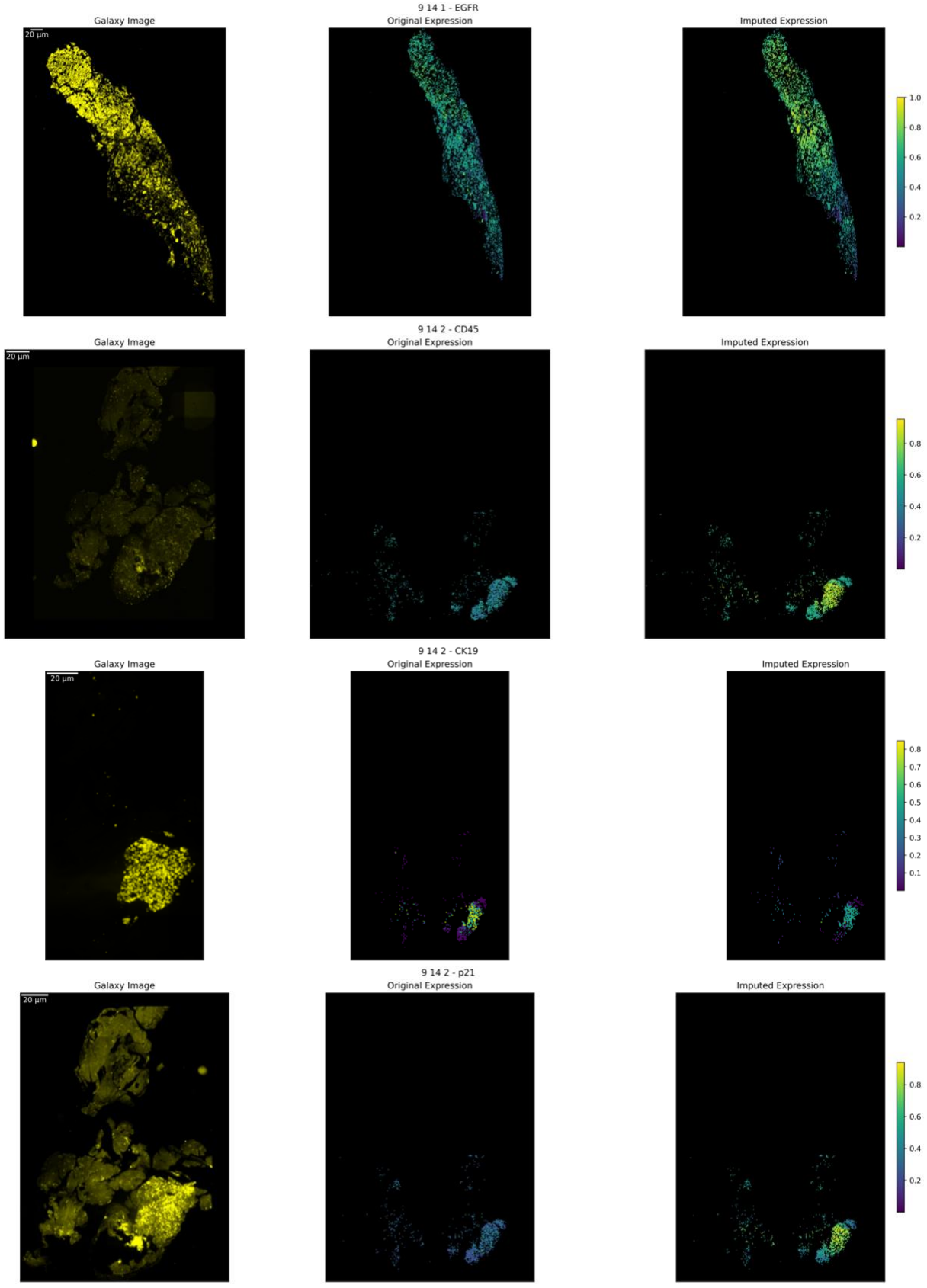


9 3 2 - pRB

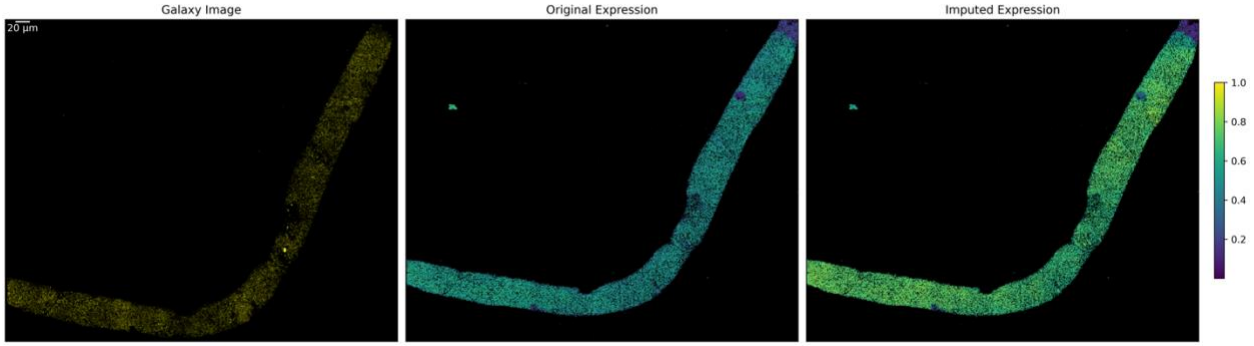


9 14 1 - Ecad

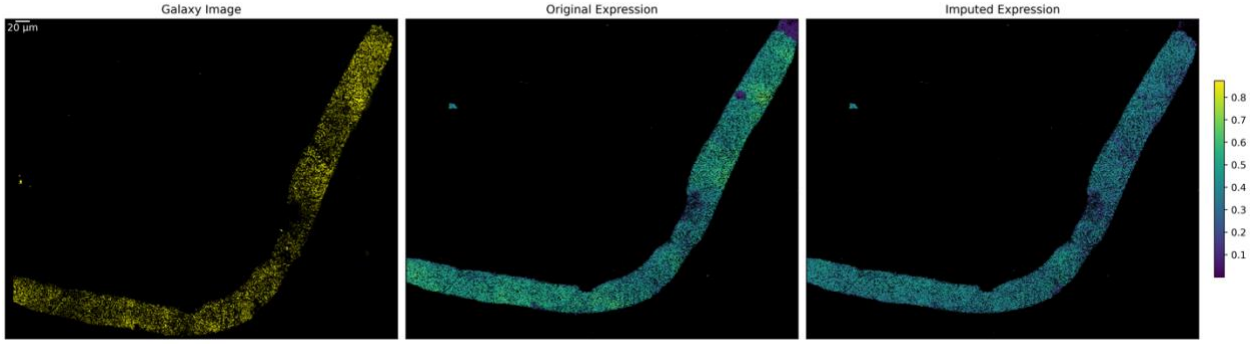


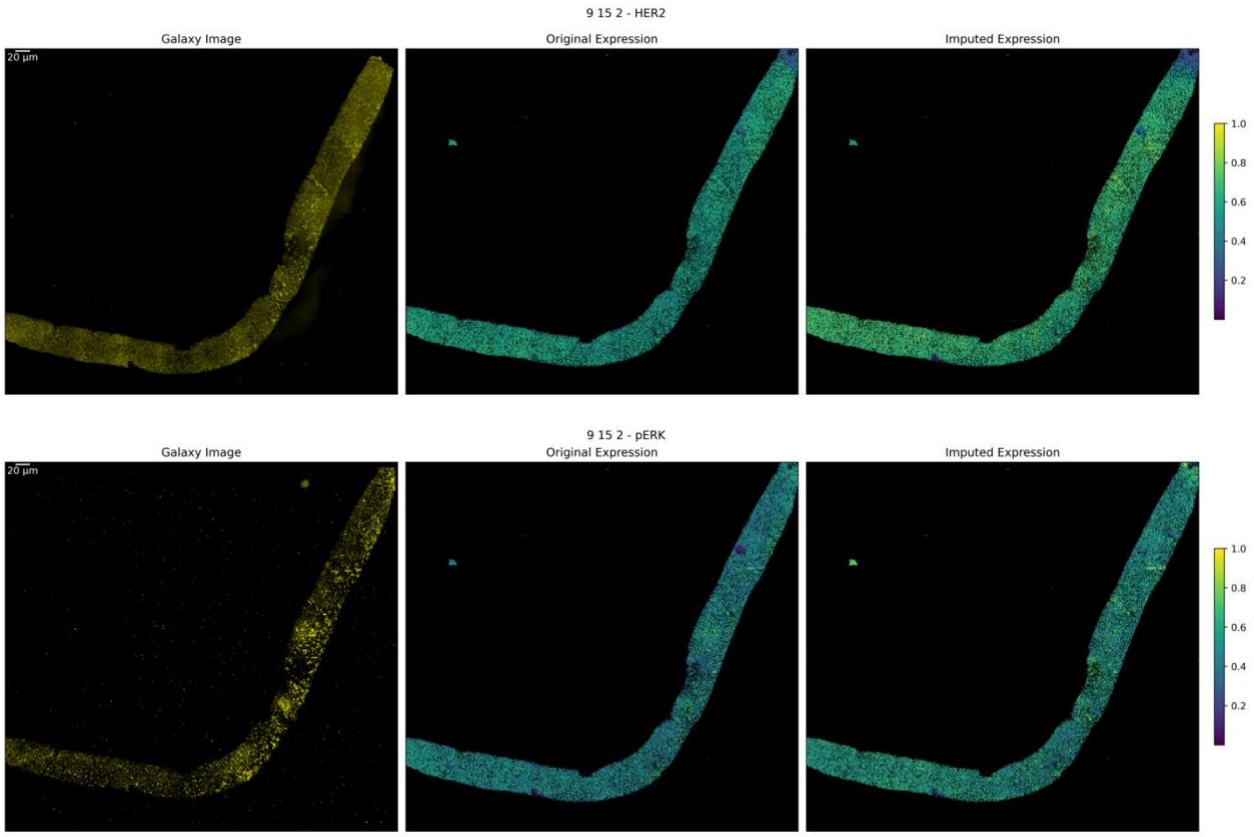

Galaxy Image
20 µm
9 14 1 - EGFR
Original Expression
Imputed Expression
1.0
0.8
0.6
0.4
0.2
Galaxy Image
20 µm
9 14 2 - CD45
Original Expression
Imputed Expression
0.8
0.6
0.4
0.2
Galaxy Image
20 µm
9 14 2 - CK19
Original Expression
Imputed Expression
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Galaxy Image
20 µm
9 14 2 - p21
Original Expression
Imputed Expression
0.8
0.6
0.4
0.2

9 15 2 - AR

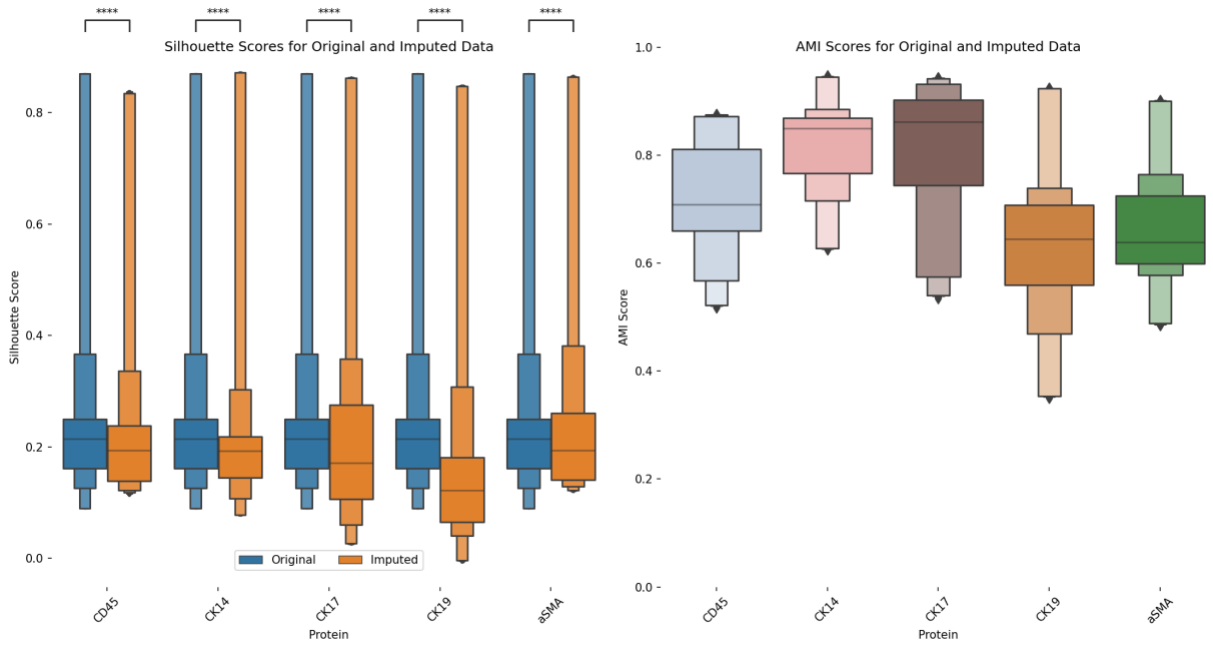


9 15 2 - ER





Supplementary Figure 2.13 In Situ vs Original vs Imputed expression data. A selection of protein expression is shown with in situ visualization, alongside the original and imputed protein expression data. The imputed protein expression successfully mirrors the overall structural patterns of the original data.



Supplementary Figure 2.14 Silhouette & AMI scores for phenotype calling. Silhouette scores indicate a slight decrease in clustering performance for the proteins involved in phenotype calling, with differences ranging from 0.01 (CD45) to 0.05 (CK19). AMI scores reveal substantial information overlap between clusters of original and imputed data for these proteins, with the highest score exceeding 0.8 for CK14 and CK17, and the lowest score remaining above 0.6 for CK19. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values:

ns: not significant, p ≤ 1.00e+00

*: 1.00e-02 < p ≤ 5.00e-02

**: 1.00e-03 < p ≤ 1.00e-02

***: 1.00e-04 < p ≤ 1.00e-03

****: p ≤ 1.00e-04

Supplementary Figure 2.15 Performance comparison between baseline (0 µm) and increasing spatial distances of 15, 30, 60, 90 and 120 µm. Light GBM results comparing increasing spatial distances to the baseline (0 µm in grey) across proteins. a) Baseline compared to 15 µm (magenta) and 30 µm (plum). b) Baseline compared to 60 µm (green) and 90 µm (yellow) and c) Baseline compared to 120 µm (red). For means of comparison the Mean Absolute Error (MAE) was calculated. Statistical analysis using Mann-Whittney-Wilcoxon and multi-hypothesis testing using Benjamini-Hochberg correction. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values:

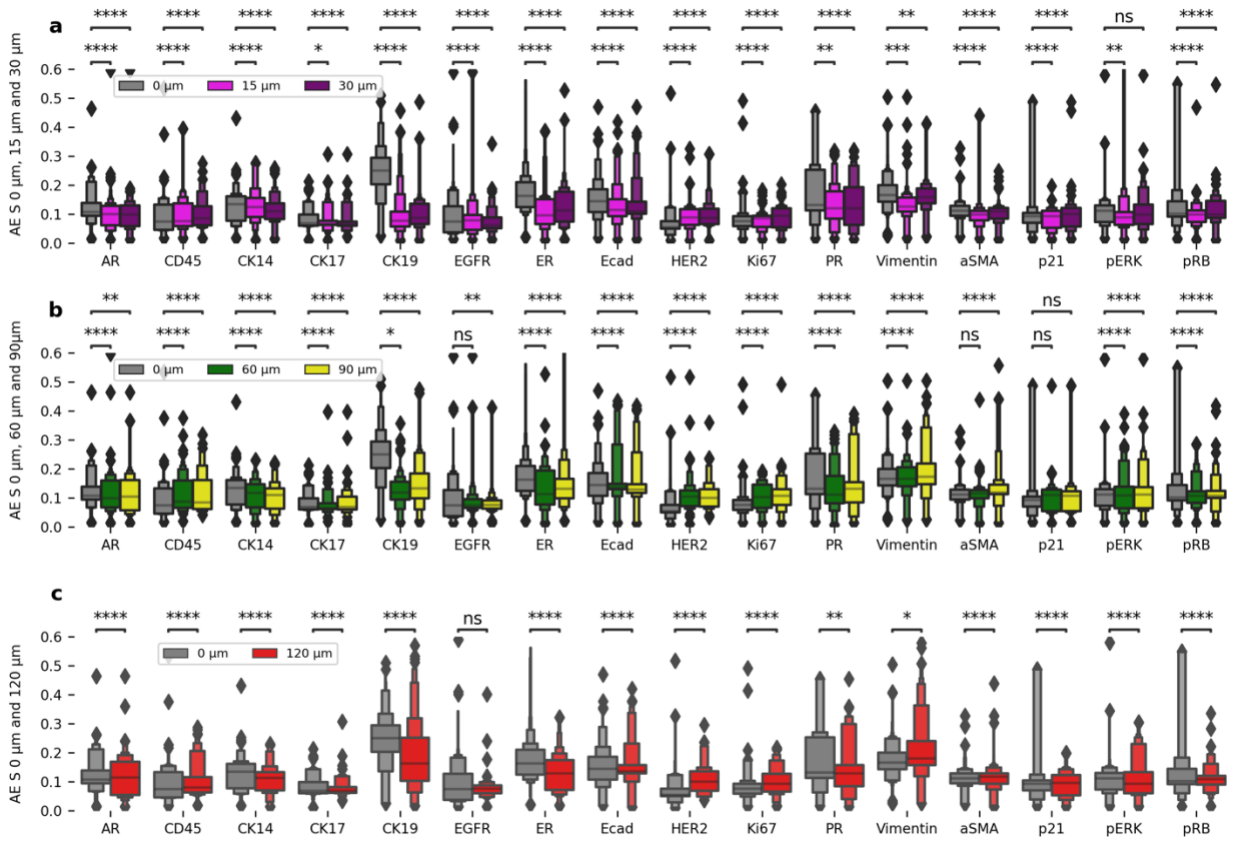
ns: not significant, $p \leq 1.00e+00$

*: $1.00e-02 < p \leq 5.00e-02$

**: $1.00e-03 < p \leq 1.00e-02$

***: $1.00e-04 < p \leq 1.00e-03$

****: $p \leq 1.00e-04$



Supplementary Figure 2.16 Performance comparison between baseline (0 µm) and increasing spatial distances of 15, 30, 60, 90 and 120 µm. Autoencoder (AE) results comparing increasing spatial distances to the baseline (0 µm in grey) across proteins. a) Baseline compared to 15 µm (magenta) and 30 µm (plum). b) Baseline compared to 60 µm (green) and 90 µm (yellow) and c) Baseline compared to 120 µm (red). For means of comparison the Mean Absolute Error (MAE) was calculated. Statistical analysis using Mann-Whittney-Wilcoxon and multi-hypothesis testing using Benjamini-Hochberg correction. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values:

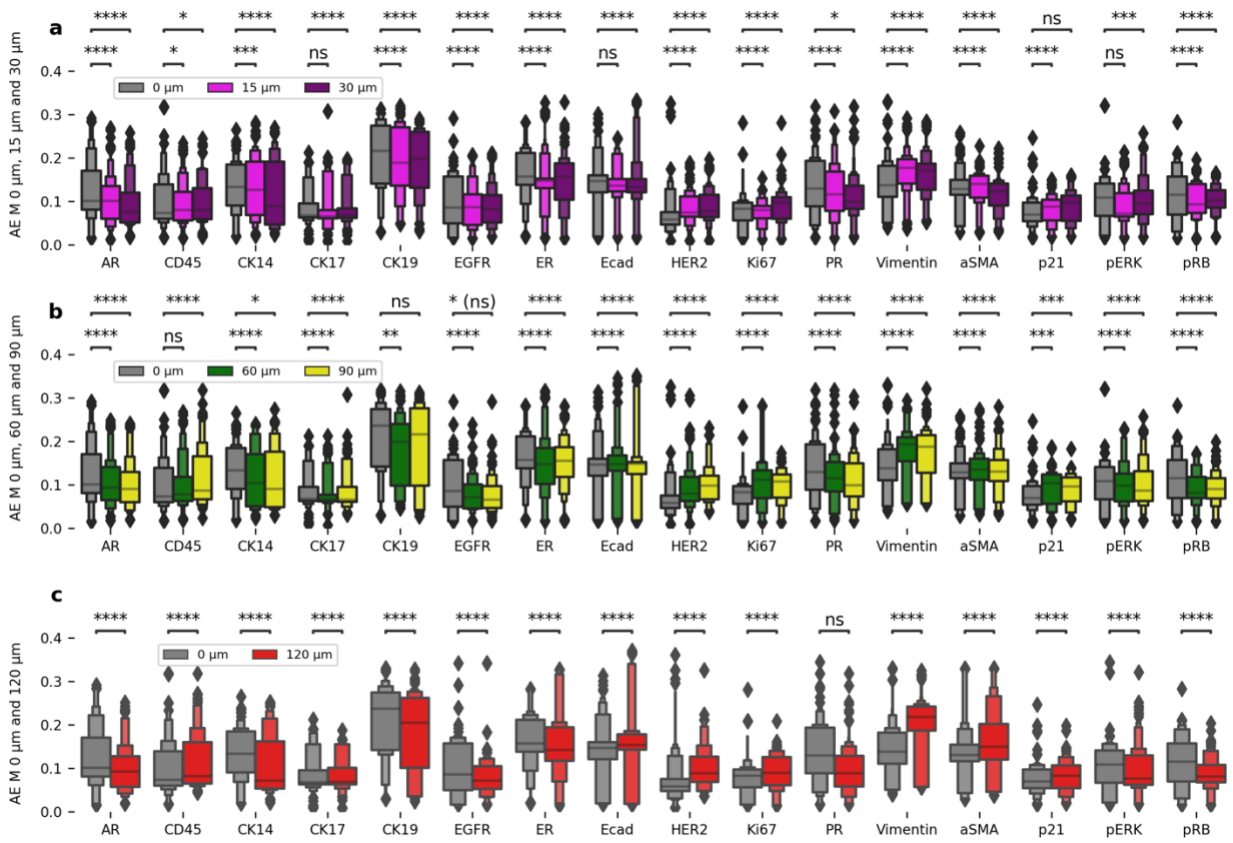
ns: not significant, $p \leq 1.00e+00$

*: $1.00e-02 < p \leq 5.00e-02$

**: $1.00e-03 < p \leq 1.00e-02$

***: $1.00e-04 < p \leq 1.00e-03$

****: $p \leq 1.00e-04$



Supplementary Figure 2.17 Performance comparison between baseline (0 µm) and increasing distances of 15, 30, 60, 90 and 120 µm. Autoencoder M (AE M) results using multi protein imputation comparing increasing spatial distances to the baseline (0 µm in grey) across proteins. a) Baseline compared to 15 µm (magenta) and 30 µm (plum). b) Baseline compared to 60 µm (green) and 90 µm (yellow) and c) Baseline compared to 120 µm (red). For means of comparison the Mean Absolute Error (MAE) was calculated. Statistical analysis using Mann-Whittney-Wilcoxon and multi-hypothesis testing using Benjamini-Hochberg correction. Each boxenplot displays nested boxes corresponding to progressively smaller quantile ranges. The central, widest box represents the interquartile range (25th–75th percentiles), capturing the middle 50% of the data. Narrower boxes above and below reflect increasingly extreme quantiles (e.g., 12.5th–87.5th, 6.25th–93.75th), providing a detailed view of distribution tails. Outliers beyond the outermost quantile range are shown as diamonds.

p-values:

ns: not significant, p ≤ 1.00e+00

*: 1.00e-02 < p ≤ 5.00e-02

**: 1.00e-03 < p ≤ 1.00e-02

***: 1.00e-04 < p ≤ 1.00e-03

****: p ≤ 1.00e-04

Chapter 3 Aggregating multimodal cancer data across unaligned embedding spaces maintains tumor of origin signal

This chapter has been formatted for inclusion in this dissertation from the manuscript

" Aggregating multimodal cancer data across unaligned embedding spaces maintains tumor of origin signal" by Raphael Kirchgaessner, Kaya Keutler, Layaa Sivakumar, Xubo Song, Kyle Ellrott, submitted to Bioinformatics (2025). The author of this dissertation is the first author of this manuscript and used RNA, H&E, annotations & somatic mutations data derived from the TCGA database to conduct the computational experiments, the results of which were used to generate Main Figures 1,2,3,4,5,6.

Motivation:

As AI techniques are applied to biomedical and genomic data, one of the first questions is how the data is encoded and represented, as the first step in many AI based pipelines is encoding the source data into an embedding space [114–116]. Embedding spaces are lower dimensional manifolds where relative relationships of elements are maintained, for example similar elements are placed closer to each other. In modern machine learning, combining diverse data types into single embedding vectors, using aggregation functions such as summation, is a common practice [117,118]. These aggregations significantly enhance model performance in areas such as natural language processing, genomics and computer vision [119–122]. This fusion of heterogeneous data allows for richer representations, capturing multiple facets of complex datasets. However, this process often obscures the origins of the data, making it challenging to interpret or analyze the specific contributions of individual modalities [123]. Understanding the sources of these combined embeddings is crucial for improving model interpretability and ensuring that critical information is preserved.

Importantly, there is an open question about how the aggregation of non-aligned embedding vectors will affect recognition. In creating embedding spaces different dimensions tend to encode complex concepts, with the relative value associated with the position in the axis. In language models, concepts such as gender or the concept of royalty can be associated with different axes. The difference between the words 'king' and a 'queen' moves along the gender axis, while the difference between 'man' and 'king' moves

along the royalty axis. Importantly, because these different axes were created in coordinated optimization, the summation of two vectors, each one representing movements in singular axis, produce a coherent combined effect, i.e. adding a vector in the direction of gender to one with the direction of royalty will move from the word 'man' to 'queen' [124]. This property has widely been taken advantage of in Large Language Models (LLM) and graph neural networks. In the case of multi-modal data, where embedding spaces have been created with no consideration of other data encoding, there is no guarantee that the summation of vectors will not cause interference between concepts encoded in different spaces. There is the possibility that if the major axes in different embedding spaces conflict with each other, any recognizable signal could be lost in the aggregation. We proposed a set of experiments to measure the effect of interference between different embedding spaces when aggregated together.

Biomedical data related to precision oncology comes in a variety of different forms including genomic and transcriptomic profiles, imaging data and clinical notes [125–127]. One strategy would be to create a multi-modal embedding space, where all data modalities are commonly encoded, using a method such as the Contrastive Language-Image Pre-Training (CLIP) algorithm [128] which helped to build the text to image conversion of DALL-E [129]. However, we were interested to see if a more simplistic strategy for multi-modal embedding integration could be deployed for enabling search, clustering and recognition of connected cancer records. Our primary motivation was to investigate whether it is possible to combine heterogeneous patient data, such as RNA transcription levels, H&E images, and patient clinical text annotations, into a unified embedding while still retaining the

ability to distinguish the sources that contributed to the aggregated vector. Importantly, we wished to determine if this analysis could be done without coordinating the various embedding manifolds, meaning that each of the various embedding spaces are created independently and only organized using the uni-modal source information.

Graph learning algorithms have been proposed as effective tools for encoding graph-structured information [130,131]. However, most existing methods have been demonstrated using graphs constructed from a single data modality. In the case of complex biological data, there may be multiple data modalities, and thus multiple embedding spaces, that can be connected together in a structured graph. In order for these methods to be effective, it must be possible to aggregate vectors from multiple data modalities, without losing all relevant information to noise. This is possible in single-modality embedding spaces, however whether or not this capacity is maintained when aggregating across uncoordinated embeddings is less well known.

In addition, we evaluated whether the summed embeddings maintain enough information integrity to be stored and efficiently retrieved from vector databases. Use of embedding methods for creating vector indices of data allows for semantically aware search of document stores. Metrics such as cosine similarity, Euclidean distance, and dot product were used to assess whether the aggregated embeddings could still be meaningfully compared and queried. This aspect of the study is highly relevant, as the ability to store and retrieve embeddings in vector databases is critical for real-world applications, including personalized medicine, where fast, reliable retrieval of patient-specific data can lead to better treatment outcomes.

Lastly, we aimed to assess whether these summed embeddings, generated from heterogeneous multi-modal data, could effectively classify cancer types. This would demonstrate that despite the complexity of the data fusion, the resulting embeddings still contain enough discriminative power to distinguish between different cancer types. This investigation could have significant implications for cancer diagnostics, enabling more accurate and interpretable classification models in the future.

Approach

To investigate whether simple aggregation of embeddings from unaligned latent spaces can yield meaningful representations for downstream tasks, we developed two complementary sampling and evaluation strategies.

The first strategy focused on modality-level sampling to address the composition recognition problem. In this setting, the goal was to determine whether it is possible to infer which data modalities (e.g., RNA, H&E, clinical annotations) were used to construct a given aggregate vector. To generate the training data for this task, we randomly sampled embeddings from different modalities without regard to patient identity, and combined them to form synthetic aggregate vectors. A recognition model was then trained to predict the constituent modalities present in each input vector. This approach allowed us to assess whether naive aggregation preserves sufficient modality-specific structure to support decomposition, despite the embeddings originating from unaligned latent spaces.

The second strategy was designed to assess the clinical and biological utility of the aggregated embeddings. Here, the task was to predict cancer-relevant attributes, including

cancer type (e.g., BRCA, BLCA), tumor mutational burden (TMB), and molecular subtype, based on patient-level aggregate vectors. To construct these vectors, we performed patient-level sampling, where each aggregate vector was composed of heterogeneous embeddings corresponding to the same individual. These included representations derived from RNA-seq profiles, H&E images, clinical annotations, and somatic mutation data. This setup enabled us to evaluate whether the aggregated vectors retained enough biological signal to be useful for clinically relevant prediction tasks.

All embeddings were generated using either Variational Autoencoders (VAEs) or Sentence-BERT (sBERT) (Figure 3.1a). sBERT is a transformer-based model that extends the original BERT architecture by incorporating a siamese network structure, enabling efficient generation of semantically meaningful sentence-level embeddings through cosine similarity comparisons. It is particularly well-suited for capturing contextual relationships in textual or sequence-based data. In contrast to Variational Autoencoders (VAEs), which learn a continuous and probabilistic latent space optimized for reconstruction and generative modeling, sBERT produces deterministic and discrete embeddings that are optimized for semantic similarity tasks. Each embedding represented a datapoint tied to a specific patient or modality. The only coordination between different embedding models was dimensionality alignment: all vectors were projected into a shared 768-dimensional space to match the output size of the sBERT model. To prepare the embeddings for use with traditional machine learning pipelines and vector databases, we applied structured random sampling and aggregation procedures. This transformation was critical for enabling efficient storage, retrieval, and analysis of heterogeneous data types using conventional

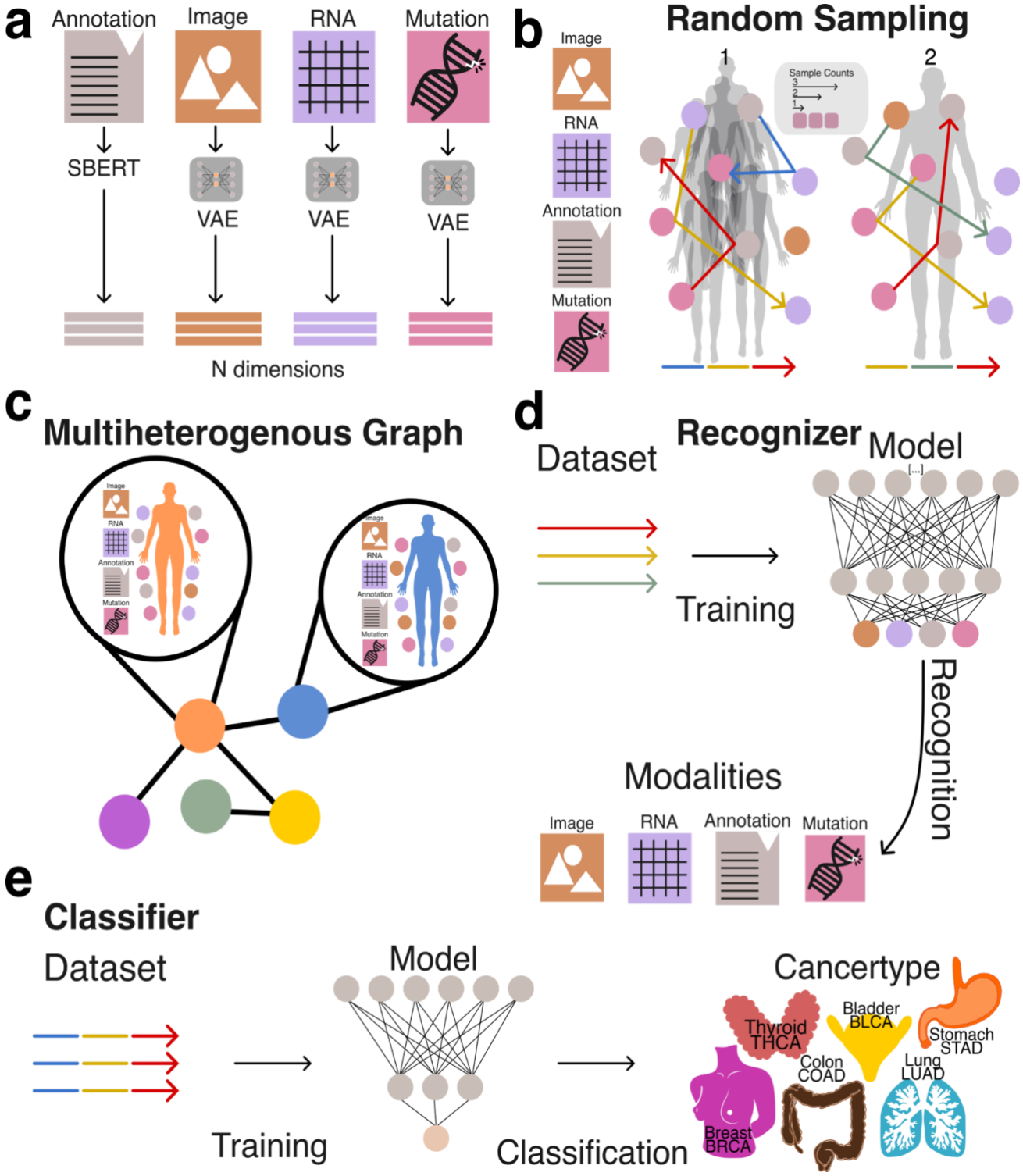


Figure 3.1 Overview of embedding generation and network architecture. a: Modalities (RNA, Image, Mutation & Annotation) used to generate embeddings by using either SBERT or a VAE. **b:** Illustration of random samplings performed over multiple patients (1) and one patient (2), demonstrating different sampling counts. **c:** Schematic representation of a heterogeneous graph incorporating patients and their associated modalities. **d:** Schematic of the recognizer experimental setup to recognize multiple modalities, including image, RNA, annotations, and mutation data. **e:** Schematic of the

classifier experiment setup which predicts the patient's cancer type, classified as BRCA, LUAD, BLCA, THCA, COAD, or STAD.

vector-based methods. By encapsulating diverse multimodal representations into a unified vector format (Figure **3.1**b, Figure **3.1**c), we facilitated systematic evaluation of whether naive aggregation preserves informative structure across both tasks.

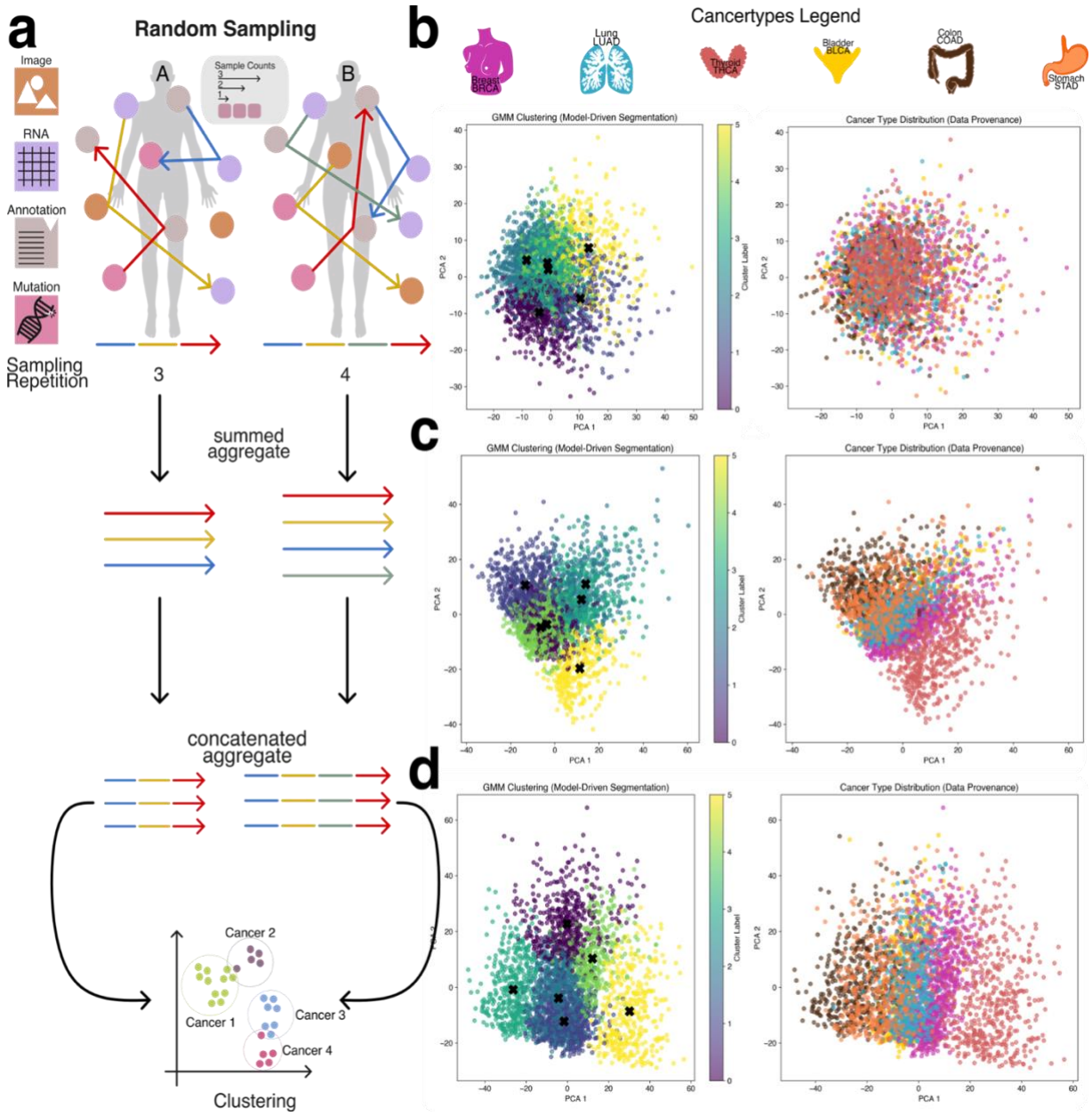
Together, these two strategies allowed us to evaluate the extent to which simple vector aggregation from unaligned modality-specific embeddings can support downstream machine learning applications, including modality composition recognition and clinically meaningful phenotype prediction (Figure **3.1**d, Figure 3.1e).

Results

Unsupervised clustering

To evaluate the intrinsic separability of cancer types without relying on complex or computationally intensive deep learning models, we applied a Gaussian Mixture Model (GMM) clustering algorithm to patient-level aggregate embeddings generated via the second sampling strategy (patient-level sampling) (Figure 3.2a). GMM is an unsupervised learning method that models data as a mixture of Gaussian distributions, allowing it to group similar data points by estimating their underlying probabilistic structure. The number of GMM components was set equal to the number of cancer types in the dataset (BRCA, BLCA, STAD, COAD, THCA, LUAD), enabling a direct assessment of whether embeddings corresponding to different cancer types could be naturally grouped.

Figure 3.2 Unsupervised clustering of summed and concatenated embeddings. a: Schematic depiction of patient-specific embeddings obtained by aggregating and concatenating vectors derived from random samplings. Random samplings performed on the graph structure generate vectors concatenated into a single embedding vector for subsequent usage in a gaussian mixture model (GMM) clustering. b: PCA representation of aggregated and contacted



embeddings using a sample count of three and a sample count repeat of three, shows little differentiation between cancer types. c: Increasing both sample count and repetition by one to four both PCA and GMM clustering shows a clearer separation between cancer types. d: Using a sample count and repeat of 5 further enhances cancer type separation with THCA (far right cluster) observing the strongest separation.

Using a low sampling configuration, specifically, a sample count of three and sampling repeats of three, the resulting clusters were largely indistinct, with significant overlap between cancer types and no clear boundaries between groups (Figure 3.2b). However, as both the sample count and sampling repeats increased, clustering quality improved markedly. Increasing the number of modality embeddings used in aggregation (i.e., higher sample counts) and repeating the sampling process multiple times allowed the aggregated embeddings to better capture patient-specific variation and inter-modality relationships. This led to progressively clearer separation between cancer types (Figure 3.2c, Figure 3.2d), with THCA emerging as the most distinctly clustered cancer type, while other cancers such as BLCA, STAD, and COAD exhibited closer proximity and partial overlap.

The clearest clustering performance was achieved with a sample count of five and five sampling repeats, which resulted in well-defined cluster boundaries and improved alignment between GMM components and cancer labels (Figure 3.2b-d, Supplementary Figure **3.8** - Supplementary Figure **3.10**). Notably, when cancer types appeared between two GMM clusters, they often exhibited biological similarities to both, reinforcing the non-linearity of the underlying molecular patterns. These results demonstrate that even in the absence of supervised learning, cancer type identity is at least partially encoded in the aggregated embedding space, and that unsupervised clustering performance is strongly influenced by the richness and redundancy introduced through the sampling process.

Identification of Component Embeddings in Aggregated Representations

In this portion of the study, we evaluated the ability of machine learning models to accurately identify the underlying data modalities and cancer types used to generate each aggregate embedding. Successful recognition of the constituent modalities and cancer type supports our hypothesis that, despite the naive combination of embeddings, sufficient structural information is preserved to enable meaningful downstream analyses. To support this hypothesis we implemented two modeling strategies: a simple approach, where the model was tasked with distinguishing between the base modalities (e.g., RNA, image, mutations, and annotations), and a cancer-specific approach, which further required the model to identify the cancer type from which the embeddings were derived (Figure 3.3a).

To generate datasets for training and evaluating the recognition model, we created datasets ranging from three to ten random sampling steps, each containing 15,000 data points (Figure 3.3b). To assess model generalizability, we constructed a composite dataset by integrating all datasets containing between three and ten constituents, resulting in a total of 90,000 summed embeddings (Figure 3.3c). To establish a baseline, we employed a multiclass logistic regression model (MCRM) to evaluate the feasibility of recognizing the composition of aggregate embeddings, aiming to determine whether the model could accurately distinguish the base modalities used to generate each aggregate vector and to evaluate whether more complex models are necessary. Additionally, we developed a deep learning model (DL) to leverage a more expansive feature space for representation learning.

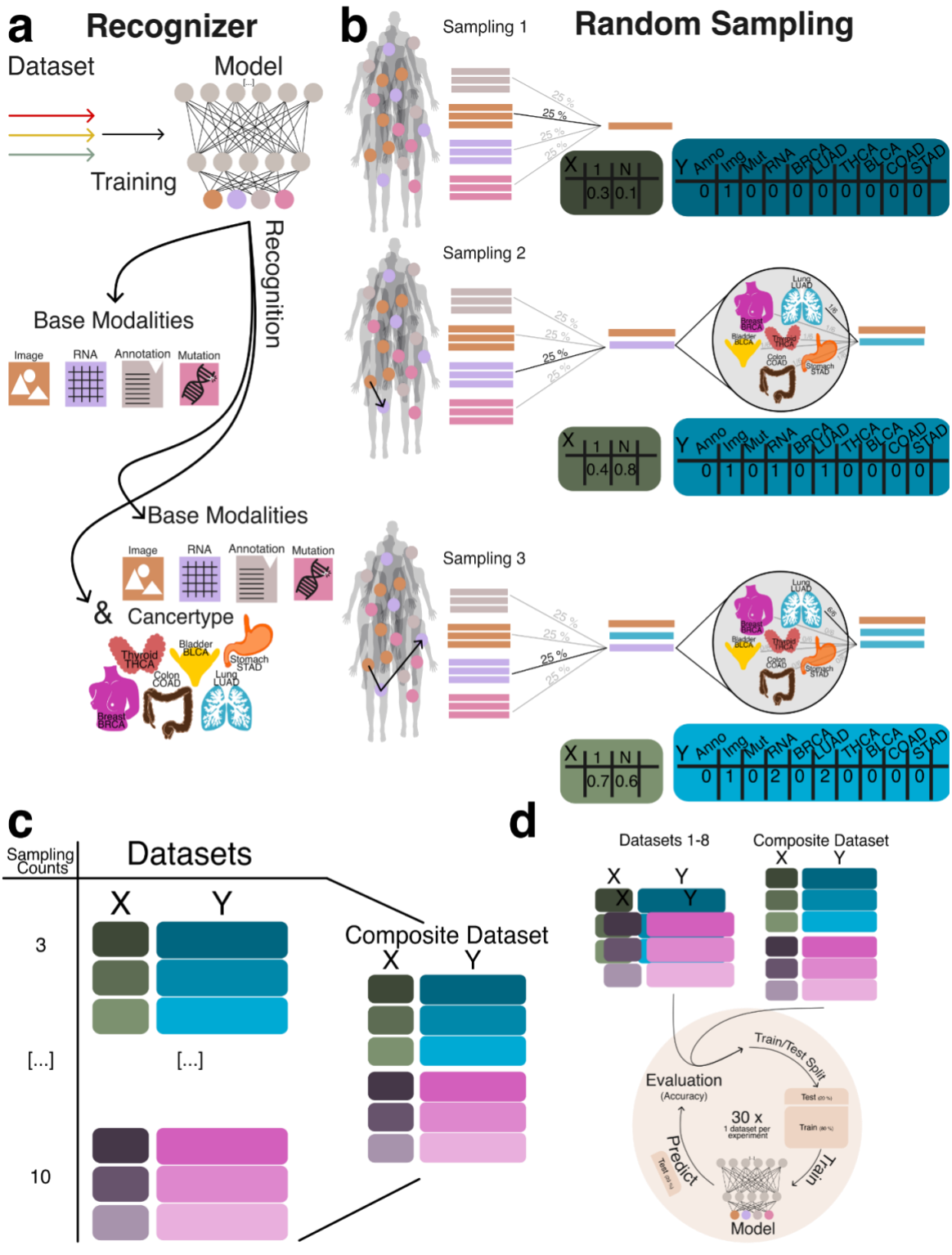

a
Recognizer
Dataset
Model
Training
Recognition
Base Modalities
Image
RNA
Annotation
Mutation
Base Modalities
Image
RNA
Annotation
Mutation
& Cancertype
Thyroid THCA
Bladder BLCA
Stomach STAD
Colon COAD
Lung LUAD
Breast BRCA
b
Random Sampling
Sampling 1
25 %
X 1 N 0.3 0.1
Y Anno Img Mut RNA BRCA LUAD THCA BLCA COAD STAD 0 1 0 0 0 0 0 0 0 0
Sampling 2
25 %
X 1 N 0.4 0.8
Y Anno Img Mut RNA BRCA LUAD THCA BLCA COAD STAD 0 1 0 1 0 1 0 0 0 0
Sampling 3
25 %
X 1 N 0.7 0.6
Y Anno Img Mut RNA BRCA LUAD THCA BLCA COAD STAD 0 1 0 2 0 2 0 0 0 0
c
Sampling Counts
Datasets
X
Y
3
[...]
10
Composite Dataset
d
Datasets 1-8
Composite Dataset
Train/Test Split
Evaluation
(Accuracy)
30 x
1 dataset per experiment
Test
Train
Predict
Train
Model

Figure 3.3 Embedding generation for the recognizer network. a: Recognizer setup including a simple recognizer trained to identify base modalities (RNA, image, somatic mutations, and annotations), and a cancer-specific recognizer designed to additionally distinguish the cancer type from which embeddings are derived. b: Random Sampling (RS) process shown for a sample count of 3. The process begins by selecting a random embedding (X) and recording its modality and cancer type in Y. A second embedding is sampled from the same cancer type, summed with the first, and tracked. A third embedding is then added, completing the aggregated vector while continuing to track contributing modalities. c: Composite dataset generation by combining all datasets generated using sample counts ranging from 3 to 10. d: Both isolation and composite datasets are used for training and evaluating model performance, with training and testing conducted using at least n > 30 independent runs per condition.

Both models were trained and evaluated on datasets split into training and test sets, with each experiment count of n >= 30 to ensure robustness (Figure 3.3d). Given the dataset's class imbalance and zero inflation, we evaluated model performance using the Matthews correlation coefficient (MCC), balanced accuracy, and F1 score. MCC was chosen as the primary metric due to its suitability for imbalanced data [132,133]. Figure 3.4a illustrates model performance under the simple aggregation approach, which corresponds to the first strategy described earlier. In this setting, we trained isolated models, where each model was exposed only to embeddings generated using a fixed number of randomly sampled modality vectors (referred to as sample count, e.g., three). Sample count serves as a proxy for aggregation complexity, with higher counts introducing more heterogeneity into each embedding. Both the MCRM baseline (BL) and deep learning (DL) models achieved high Matthews correlation coefficient (MCC), accuracy, and F1 scores at lower sample counts. However, as the number of aggregated vectors increased, performance differences became more pronounced, particularly for somatic mutation and annotation embeddings. The MCC for the MCRM BL model declined below 0.8 for somatic mutations and below 0.6 for annotations, while the DL model maintained stable performance close to 0.9 across all

sample counts. Figure 3.4b compares two deep learning models: an isolation model, trained separately on embeddings generated from a fixed sample count, and a composite model, trained on a dataset comprising all sample counts. Across most modalities, both models performed similarly. However, for annotation embeddings, the composite model consistently underperformed relative to the isolation model. The composite model exhibited Matthews Correlation Coefficient (MCC) values ranging from 0.2 to 0.6, whereas the isolation model achieved consistently higher and more stable MCC values. This suggests that the isolation model is more robust when trained on inputs with uniform structure. We hypothesize that this discrepancy arises from the nature of the embeddings and the challenges associated with aggregating heterogeneous representations. Specifically, annotation embeddings were generated using Sentence-BERT (sBERT), which produces semantically structured representations, while all other modality embeddings were derived from Variational Autoencoders (VAEs), which are optimized for different latent structures. This mismatch introduces incompatibility during aggregation, particularly in the composite model, where embeddings from different modalities are combined.

At lower sample counts (e.g., three), the model can still extract meaningful cross-modal patterns, as the small number of heterogeneous embeddings may retain enough coherence. However, as the number of samples increases, differences in feature distribution and representational structure across modalities become more pronounced. The model's ability to integrate these divergent signals diminishes, resulting in a noticeable performance drop between three and eight input samples.

Interestingly, a slight recovery in model performance is observed for the annotation embeddings once the sample count exceeds eight. This trend is unique to this modality, which was generated using sBERT, and likely reflects the semantic rather than geometric nature of its latent structure. Beyond this threshold, aggregating a larger and more diverse set of embeddings may provide additional contextual cues that help the model partially mitigate modality-specific noise. In this regime, redundancy and complementarity among samples could allow the model to identify weak but consistent semantic patterns across embeddings. Nonetheless, the improvement remains modest, suggesting that the inherent incompatibility between sBERT-derived representations, which capture semantic similarity without enforcing spatial continuity, and VAE-based embeddings, which are optimized for continuous, manifold-like organization, continues to limit cross-modal generalization. Figure 3.4c presents results from the specific-cancer approach, where isolated models were trained using a fixed number of samples per cancer type (e.g., three samples per cancer). Despite this constraint, the DL model maintained stable performance across cancer types. In contrast, the MCRM BL model exhibited significant performance degradation under these conditions, with the effect particularly pronounced for cancer types such as lung adenocarcinoma (LUAD) and bladder cancer (BLCA), where MCC dropped to 0.2 when trained on a sample count of 10. Figure 3.4d mirrors the composite versus isolation comparison within the specific-cancer context, again evaluating only deep learning models. Consistent with the findings in Figure 3.4b, annotation embeddings exhibited the most pronounced performance drop in the composite setting. The composite model failed to achieve MCC values above 0.6 for annotation embeddings, with

performance falling below 0.4 for sample counts between 5 and 8. In contrast, the isolation model maintained consistently high performance, with MCC values remaining above 0.9 across all sample counts.

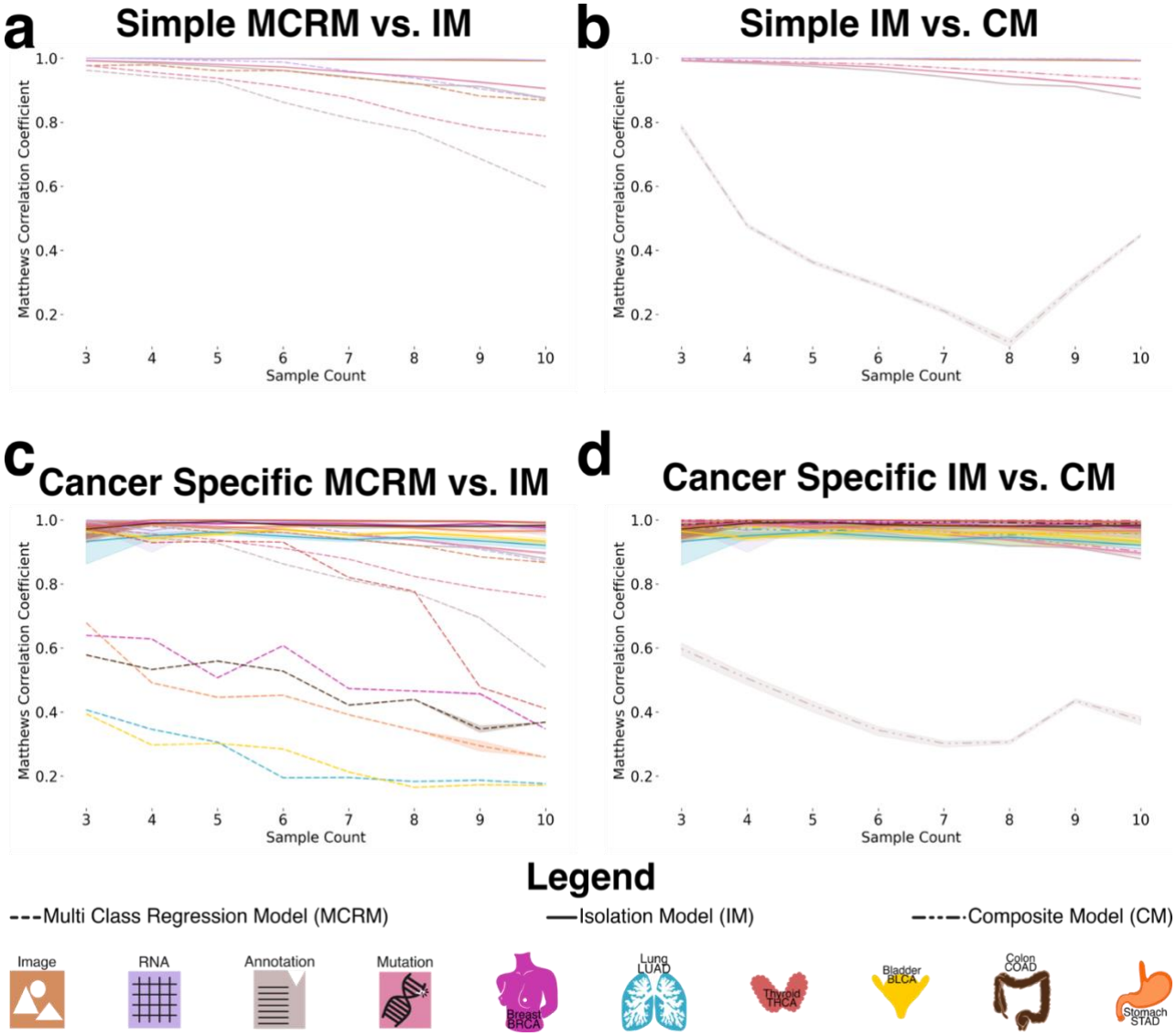


Figure 3.4 Performance of deep learning and baseline models across aggregation strategies and sampling contexts. a: Performance of the simple aggregation approach using isolation models (IM) across modalities, evaluated at sample counts ranging from 3 to 10. The multi-class logistic regression model (MCRM), used as a baseline, achieves MCC values between 0.6 and 0.8, while the DL-based IM consistently outperforms it, achieving MCC values above 0.9. b: Comparison between composite models (CM) and isolation models (IM) under the simple aggregation setting. Performance remains high across most modalities for both models, except for the annotation modality, where the CM shows a marked decline in MCC with increasing sample count, dropping as low as 0.4. c: Performance under the specific-cancer approach comparing cancer-specific IMs to the MCRM baseline. The baseline shows poor performance for several cancer types, particularly BLCA, LUAD, and STAD, with MCC values between 0.2 and 0.4. In contrast, the DL-based IMs achieve strong and consistent performance across all cancer types (MCC > 0.9). d: Comparison between cancer-specific composite models (CM) and isolation models (IM). While most modalities maintain high performance (MCC > 0.85) across both

models, the annotation modality shows poor performance in the CM across all sample counts, consistent with trends observed in panel b.

We attribute this pattern to the same factors observed in Figure **3.4**b, with an important clarification regarding the nature of the underlying embedding spaces. Annotation embeddings are generated using sBERT, which produces deterministic embeddings, each input is mapped to a fixed point in a high-dimensional semantic space without an explicit distribution. In contrast, embeddings from other modalities are derived from Variational Autoencoders (VAEs), which learn probabilistic latent spaces and encourage structured, continuous representations. The VAE architecture supports smooth interpolation and aggregation of latent vectors, properties that are not inherently preserved in the sBERT embedding space. As a result, when annotation embeddings (derived from sBERT) are combined with VAE-based embeddings in the composite setting, particularly across varying sample sizes and cancer types, the fundamental differences in how these models organize their latent spaces likely lead to misalignment. While sBERT produces deterministic embeddings structured for semantic similarity in a discrete embedding space, VAEs generate continuous, probabilistic latent representations shaped by reconstruction and prior constraints. This misalignment in embedding space organization may introduce feature incompatibilities that degrade overall model performance. Isolation models, by contrast, operate on more uniformly structured inputs and are thus less affected by this latent space mismatch, enabling them to maintain stable performance even when incorporating heterogeneous modality types.

To further evaluate model robustness, we introduced a controlled noise perturbation framework in which a defined proportion of modality embeddings within each aggregated

vector was systematically replaced with random Gaussian noise. This procedure preserved the total number of embeddings per vector, maintaining the original sample count, and enabled assessment of the model's ability to differentiate true modality embeddings from noise. The deep learning (DL) model was trained exclusively on clean (noise-free) data and subsequently evaluated on perturbed datasets. Performance was assessed using Matthews correlation coefficient (MCC), F1 score, and balanced accuracy to quantify the impact of noise on the model's ability to correctly infer embedding composition (Figure 3.5 a).

Due to the MCRM baseline model's limited performance in previous experiments, this analysis was conducted exclusively with the DL model. We evaluated both the simple aggregation approach and the specific-cancer approach, comparing model performance across increasing noise levels Figure 3.5b-e present results across noise conditions ranging from 10% to 50% embedding replacement.

Under the simple aggregation approach using isolation models (Figure 3.5b, c), performance remained robust at lower noise levels. Mutation embeddings, in particular, demonstrated high resilience, maintaining MCC values above 0.8 when 10% of the input was replaced with noise. However, at 50% noise, mutation performance declined to approximately 0.6 MCC, suggesting reduced but still meaningful predictive capacity. In contrast, annotation embeddings, derived from the sBERT model rather than VAE-based encoders, exhibited consistently lower MCC values, even under minimal noise, reflecting their greater sensitivity to perturbation and underlying latent space incompatibility.

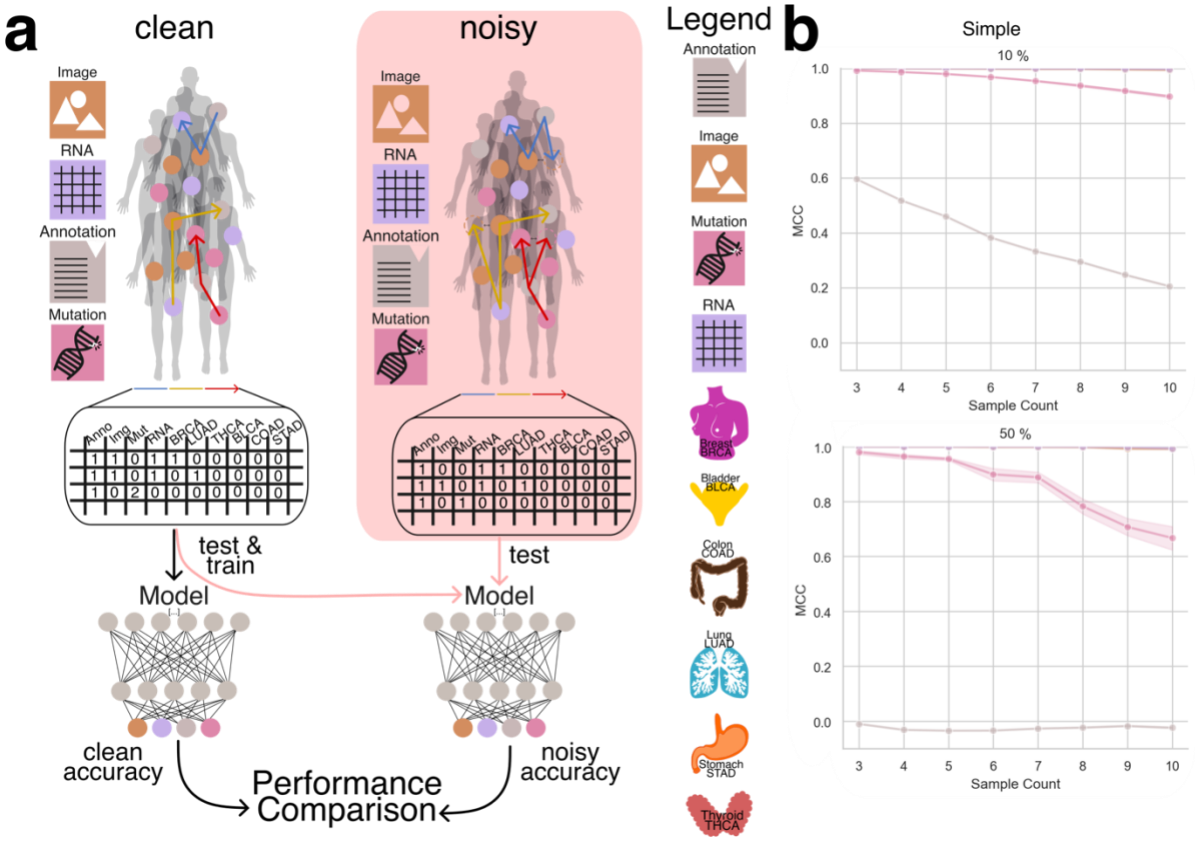

a
clean
noisy
Image
RNA
Annotation
Mutation
Anno
Img
Mut
RNA
BRCA
LUAD
THCA
BLCA
COAD
STAD
test & train
test
Model
clean accuracy
noisy accuracy
Performance Comparison
Legend
Annotation
Image
Mutation
RNA
Breast BRCA
Bladder BLCA
Colon COAD
Lung LUAD
Stomach STAD
Thyroid THCA
b
Simple
10 %
50 %
MCC
Sample Count
c
Simple Composite
10 %
50 %
MCC
Sample Count
e
Specific Cancer Composite Model
10 %
MCC
Sample Count
d
Specific Cancer
10 %
50 %
MCC
Sample Count

Figure 3.5 Performance of deep learning recognizer models under controlled noise perturbation. a: Overview of the experimental setup: models were trained on clean (noise-free) aggregated embeddings and evaluated on test data with controlled Gaussian noise introduced at 10% and 50% of embedding positions. Noise was applied while preserving the original sample count to assess robustness in distinguishing true modality embeddings from noise. **b:** Performance of the simple aggregation approach using isolation models (IM) under noisy conditions. At 10% noise, the model maintains high MCC values for RNA and image modalities, while performance for the mutation modality decreases moderately and the annotation modality shows poor performance even at low noise levels. At 50% noise, annotation performance remains consistently low, and mutation modality performance declines progressively with increasing sample count. **c**: Performance of the specific-cancer approach using isolation models (IM) at 10% and 50% noise. Models demonstrate stable MCC values across most modalities at 10% noise, with the exception of annotations. At 50% noise, annotation performance further deteriorates, and mutation embeddings exhibit notable decline for sample counts greater than six. **d:** Performance of the simple aggregation composite model (CM) under noisy conditions. Despite the presence of noise, the model achieves high MCC values (>0.9) for most modalities at both 10% and 50% noise. The annotation modality continues to underperform across all conditions. Notably, mutation embeddings show improved stability compared to the simple isolation model. e: Performance of the specific-cancer composite model (CM) under 10% and 50% noise. The model maintains strong MCC performance across most modalities and cancer types, except for the annotation modality, which consistently exhibits poor performance across both noise levels.

In the specific-cancer setting using composite models (Figure 3.5d, Figure 3.5e), annotation embeddings continued to perform poorly across all noise levels. However, mutation embeddings showed enhanced stability compared to the isolation setting, with MCC values exceeding 0.9 even at higher noise levels. This suggests that composite training across a broader dataset may help the model generalize more effectively in the presence of noise, at least for certain modalities.

These results underscore the DL model's resilience to input perturbations and highlight the importance of both aggregation strategy and embedding origin in determining robustness. While noise degrades performance in a modality-dependent manner, the ability to retain high MCC values, particularly in the mutation modality, demonstrates the potential of deep learning to support reliable composition recognition even under challenging conditions.

Tumor of Origin Identification

The classification task in this study aimed to predict specific cancer types from aggregated embeddings. While the dataset remained consistent with previous experiments, this task employed the second sampling strategy, patient-level sampling, to ensure that each aggregated embedding corresponded uniquely to a single patient. Each embedding integrated heterogeneous modality representations, including RNA-seq, H&E image features, clinical annotations, and somatic mutation profiles, reflecting the structure of real-world clinical datasets where multi-modal information is available and is ideally combined for tasks such as cancer type prediction.

To construct the aggregated embeddings, we performed multiple rounds of random sampling (sample repetitions) per patient. In each round, a subset of that patient's modality-specific embeddings was randomly selected and summed to produce a single embedding of fixed dimensionality (e.g., 768 dimensions). To capture additional intra-patient variability and enhance representational richness, this sampling process was repeated multiple times. The resulting sampled embeddings were then concatenated to form a final, patient-specific aggregate embedding. For example, if three sampling rounds were performed, the final embedding has a dimensionality of $3 \times 768 = 2304$. This approach preserved both the diversity of intra-patient modality contributions and a standardized input structure suitable for downstream machine learning analysis (Figure 3.6a).

Prior to evaluating the classification models, we performed a cluster analysis to determine whether the aggregated embeddings inherently captured cancer type–specific structure, that is, whether patients with the same cancer type tended to cluster together. This analysis was motivated by the premise that vector database retrieval systems, in which embedding similarity underlies search and retrieval operations, depend on meaningful clustering of patients, as these clusters implicitly define inter-point distances. By comparing intra-cancer and inter-cancer distances, we assessed whether these embeddings alone could effectively discriminate between different cancer types (Supplementary Figure 3.11). The results demonstrated that patient embeddings from the same cancer type exhibited greater similarity to one another than to embeddings from different cancer types. Collectively, these findings suggest that even in the absence of supervised learning, the aggregated embeddings preserve biologically informative

structure pertinent to cancer identity and may thus hold utility for retrieval-based clinical applications

Tumor classification using a deep learning model

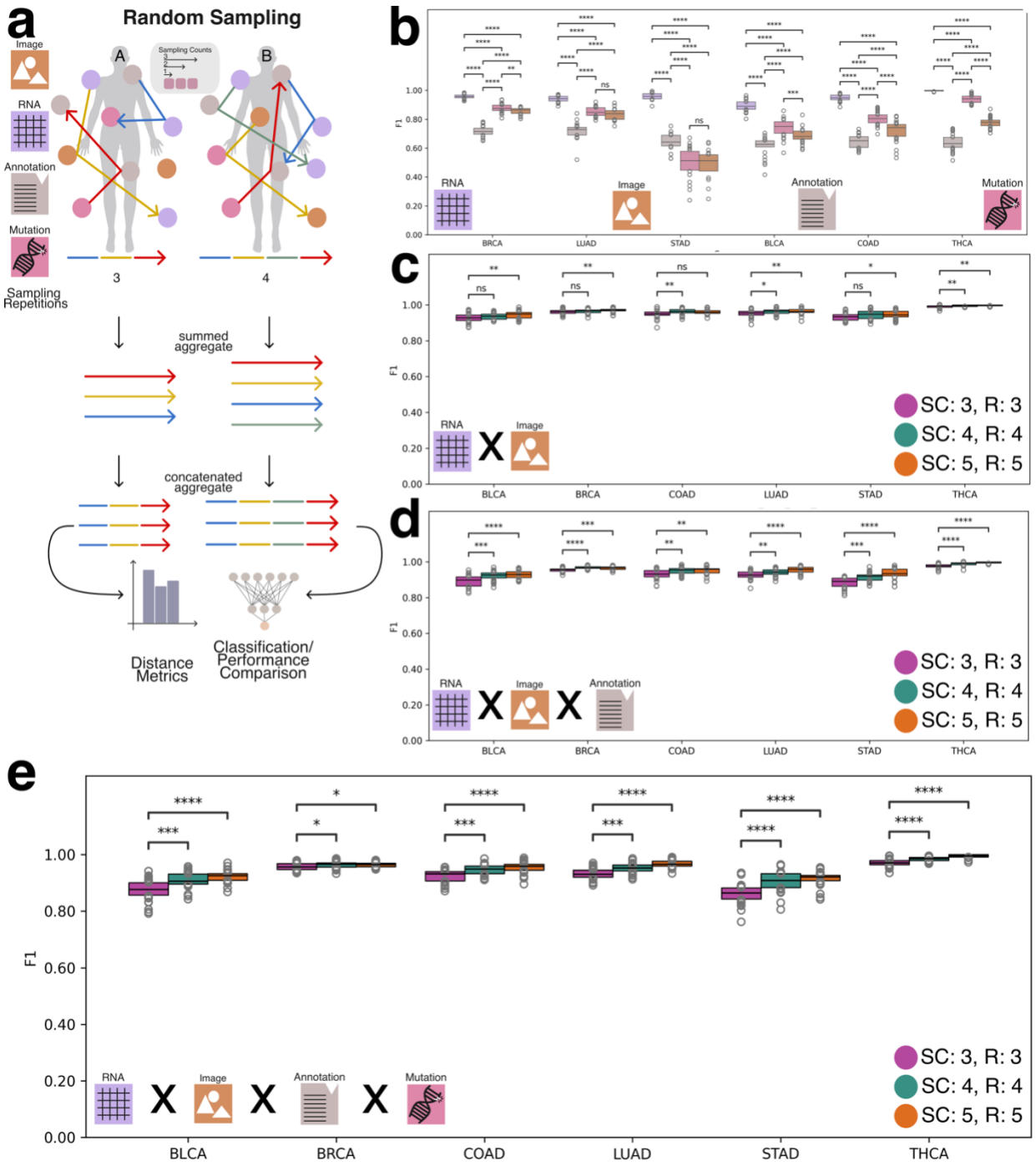


Figure 3.6 Classifier configuration and performance metrics. a: Schematic overview of patient-specific embeddings generated through aggregation and concatenation of vectors obtained from random samplings of the graph structure. Each random sampling yields a vector representation that is concatenated into a unified embedding used for downstream distance metric calculations. **b**: Single-modality classification performance (F1 scores) across RNA, image, mutation, and annotation embeddings shows that RNA consistently achieves the highest performance across all cancer types, whereas image and mutation embeddings display variable performance, with particularly low F1 scores observed in STAD. **c**: The combination of RNA and image embeddings exhibits consistently high classification performance across increasing sample counts (SC) and sample repetitions (SR). **d**: Integration of RNA, image, and annotation embeddings yields F1 scores comparable to the baseline. **e**: Incorporating all available modalities further improves performance as SC and/or SR increase, with F1 scores consistently exceeding 0.8.

p-values: ns: not significant $p \leq 1.00e+00$; *: $1.00e-02 < p \leq 5.00e-02$; **: $1.00e-03 < p \leq 1.00e-02$; ***: $1.00e-04 < p \leq 1.00e-03$; ****:$p \leq 1.00e-04$.

For the deep learning–based classification task, the concatenated aggregate embeddings were used as input features to the classifier network. These embeddings integrated information from multiple random samplings of each patient's modality-specific data, thereby capturing both the heterogeneity and redundancy inherent across modalities. To establish a baseline, models were initially trained and evaluated using single-modality embeddings (e.g., RNA) (Figure 3.6b). Model performance, assessed using multiple evaluation metrics (with F1 scores shown), revealed that the RNA modality achieved the highest performance, whereas other modalities exhibited variable performance across cancer types. Notably, the mutation and image modalities demonstrated the lowest performance for stomach adenocarcinoma (STAD), with F1 scores of approximately 0.5.

Subsequently, we conducted combination experiments in which different modalities, ranging from two, to three, and finally all available modalities, were integrated. To further examine the impact of sampling design, we evaluated how classification performance varied as a function of sample count and sampling repetition. As illustrated in **Figure 3.6** c–f (F1) and **Figure 3.7** (b,d,f), (MCC) increasing either the number of embeddings per sample

or the number of sampling iterations consistently enhanced model performance. The lowest overall performance was observed for bladder cancer (BLCA) and stomach adenocarcinoma (STAD). Analysis of the confusion matrices (**Figure 3.7** a,c,e) indicated that misclassifications frequently occurred between lung adenocarcinoma (LUAD) and BLCA, as well as between colon adenocarcinoma (COAD) and STAD, patterns consistent with previously reported molecular similarities among these cancer types [21].

These results suggest that aggregating additional heterogeneous information from a patient, regardless of the specific modality, enhances patient-level representation and improves the model's ability to differentiate between cancer types.

To extend this framework to additional predictive tasks, we evaluated the feasibility of using the same aggregated embeddings to classify cancer subtypes (Supplementary Figure **3.13**, Supplementary Figure **3.14**) and to predict tumor mutational burden (TMB) (Supplementary Figure **3.15**, Supplementary Figure **3.16**). The model achieved reasonable performance across both tasks, further supporting the utility of integrating multiple unaligned modalities to construct a robust, patient-specific embedding suitable for a wide range of clinically relevant predictions.

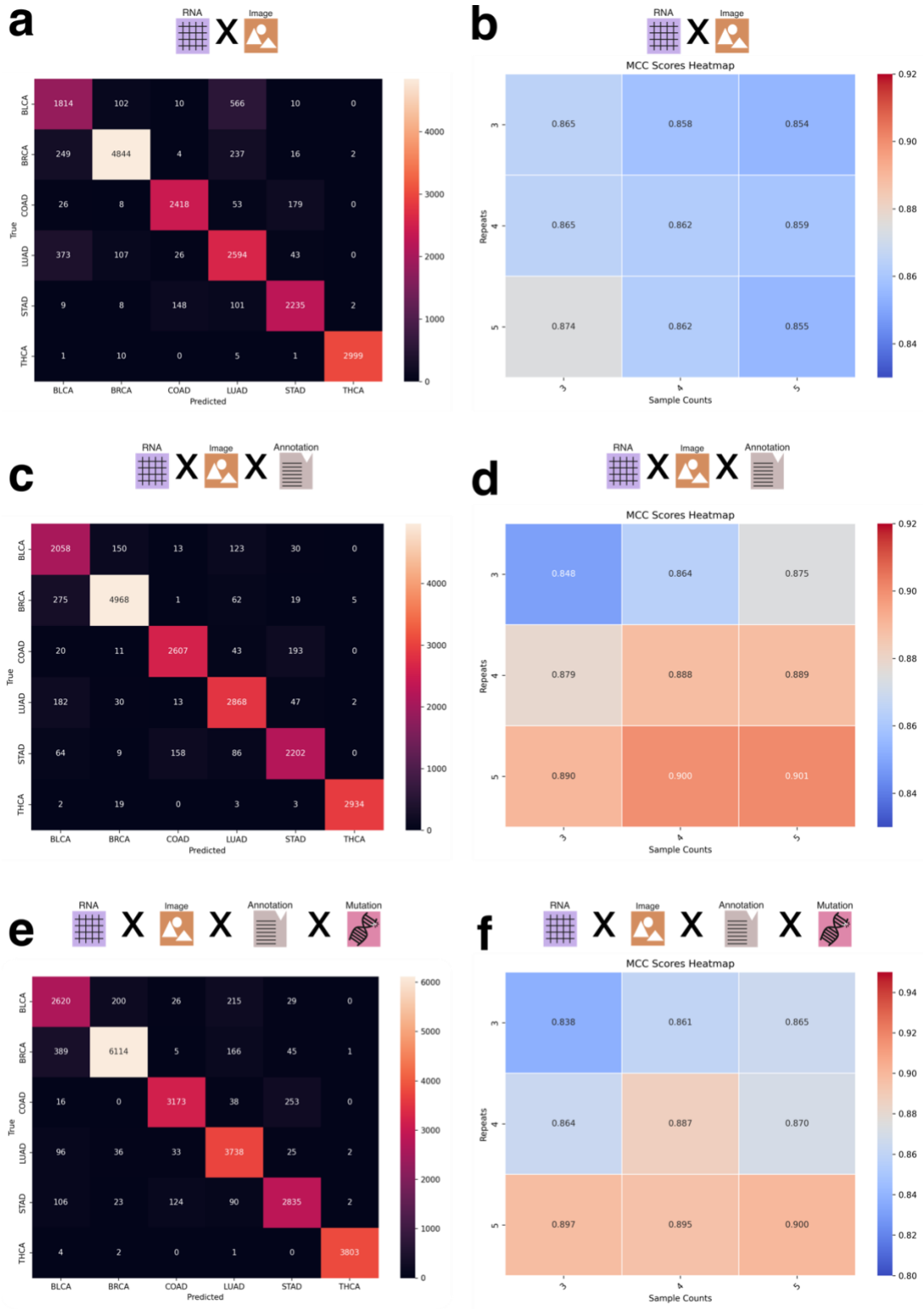


Figure 3.7 Confusion matrix and heatmaps show model performance for different combinations of repetitions and sample counts. a: Confusion matrix showing model confusion for classifications using image and rna embeddings. **b**: MCC scores heatmap shows increased performance when increasing repeats. **c**: Confusion matrix showing model confusion using rna, image and annotation embeddings. **d**: MCC scores heatmap shows improved performance when

increasing either sample counts or repetitions. **e**: Confusion matrix, showing model confusion when all modalities are being used to train and evaluate the model. **f**: MCC heatmap shows improved performance when increasing either sample count or repetitions.

Discussion & Future Work

Our findings demonstrate that the performance of both the recognizer and the classifier model indicates that embeddings from diverse embedding spaces can be effectively combined to describe local subgraphs without the need of a shared latent space. This approach enables the successful identification of both the composition of aggregated embeddings and the cancer type, sub type or tumor mutational burden associated with them.

Furthermore, our findings demonstrate that in a heterogeneous graph, aggregates covering specific regions can serve as a secondary index, enabling the linkage of multimodal records to individual entities. Furthermore, our results highlight the feasibility of using heterogeneous networks for graph neural network transformations, expanding the potential applications of these architectures. Vector databases have gained significant attention as the foundation of Retrieval-Augmented Generation (RAG). Algorithms such as Hierarchical Navigable Small Worlds (HNSW) [134] allow for the rapid indexing of vector based data. This work demonstrates that despite mixing uncoordinated embedding spaces, aggregate vectors still contain sufficient data to be identifiable regarding their modality composition. This will allow for indexing of complex heterogeneous data graphs, not only at the per node level, but also at the neighborhood level. Records in a complex patient information system

can be rapidly compared, even if the assays available are not equivalent between each individual.

We hypothesize that the observed performance degradation in the annotation modality stems from fundamental differences in the embedding generation methods. While all other modalities were encoded using variational autoencoders (VAEs), annotation embeddings were generated using Sentence-BERT (sBERT). This architectural mismatch likely contributed to the reduced performance of the annotation embeddings, particularly in recognition tasks and under composite training conditions. VAEs are designed to construct a continuous, structured latent space by enforcing a probabilistic distribution over the learned representations. This structure promotes smooth interpolation, supports meaningful aggregation, and enhances compatibility across embeddings, properties that are critical for robust downstream learning. In contrast, sBERT produces deterministic embeddings optimized for semantic similarity, without enforcing continuity or geometric smoothness in the latent space. As a result, sBERT-derived embeddings may exhibit a more fragmented or irregular structure, with weaker alignment to the representations learned by VAEs. This mismatch likely becomes more problematic as the sample count increases in aggregated embeddings, introducing feature conflicts that the model struggles to reconcile. Specifically, when a small number of embeddings (e.g., a sample count of three) are combined, the model may still learn relatively coherent patterns between modalities, even in the presence of latent space heterogeneity. However, as additional embeddings are aggregated, the lack of alignment between the sBERT and VAE embedding spaces amplifies inconsistency, degrading performance, particularly in composite models where

embeddings from a wide range of sample counts and cancer types are mixed. This lack of alignment refers to the mismatch in the underlying geometric and statistical structure of the embedding spaces: sBERT representations are organized primarily by semantic similarity, where distances reflect contextual meaning rather than continuous numerical relationships, while VAE embeddings are structured to preserve smooth, manifold-like continuity in the latent space. As a result, linear operations such as averaging or concatenating embeddings across these spaces can distort the relative relationships among samples, effectively introducing noise rather than informative cross-modal signal. Consequently, the model struggles to integrate these heterogeneous representations, leading to reduced predictive stability at intermediate sample counts. Interestingly, at higher sample counts (e.g., nine or more), we observed a partial recovery in model performance, possibly reflecting a compensatory effect from increased sample diversity. We hypothesize that this is due to the emergence of higher-order patterns and redundancies across modalities. As more diverse data is included in the aggregation, the model may begin to abstract away modality-specific inconsistencies and instead learn robust decision boundaries that generalize across latent spaces. Nevertheless, the persistent underperformance of annotation embeddings, especially in composite settings, underscores the importance of selecting compatible embedding strategies when integrating unaligned modalities. These findings suggest that the structural properties of embedding spaces, particularly whether they support compositionality, smoothness, and alignment, play a critical role in determining model robustness and generalization in multi-modal learning frameworks.

Future work could explore alternative embedding strategies for annotations, such as fine-tuned VAEs or hybrid approaches that combine semantic and latent-space regularization, to improve consistency across modalities. Follow up research could pivot towards a heterogeneous data graph used for benchmarking. Identifying a multimodal data graph with complex relationships, structures and well-defined prediction tasks will take considerable effort. This paper has demonstrated that the basic premise of multimodal embedding aggregation is viable and is able to hold information without interfering signals immediately collapsing into noise. We plan to apply this technique to encoding tumor evolution patterns, including encoded information representing tri-nucleotide patterns, structured somatic mutation timing, and sub clonal mutation clustering. Other experiments could include multi-modal single cell data encoding, capturing transcriptomic and methylation data across various cell states.

Acknowledgements

The research supporting this publication was supported by NCI GDAN 5U24CA264007, NIH U54HG012517 and NHGRI U24HG010263. The research reported in this publication used computational infrastructure supported by the Office of Research Infrastructure Programs, Office of the Director, of the National Institutes of Health under Award Number S10OD034224. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Methods

Creation of Dataset

To obtain the necessary data, we utilized The Cancer Genome Atlas (TCGA), a comprehensive repository containing molecular and clinical data from over 10,000 cancer patients across 33 cancer types. For this study, we focused on six cancers: Breast Cancer (BRCA), Bladder Cancer (BLCA), Lung Adenocarcinoma (LUAD), Stomach Adenocarcinoma (STAD), Thyroid Cancer (THCA), and Colon Adenocarcinoma (COAD). TCGA provides extensive genomic insights, including RNA expression profiles, somatic mutations, and imaging data such as H&E-stained slides, enabling a robust analysis of cancer heterogeneity and multi modal embedding aggregations.

Data Preparation

For each cancer type, we gathered the corresponding H&E images, patient annotations, gene mutations, and RNA readouts. Specifically, we obtained data from patients diagnosed with either Breast (BRCA), Bladder (BLCA), Lung (LUAD), Thyroid (THCA), Colon (COAD) or Stomach (STAD) cancer across these six cancer types. (Table **3.1**, Table **3.2**)

While image, RNA, and mutation data were readily available in formats like CSV and TIFF files, patient annotations were only provided in PDF format, making them less accessible for analysis. To facilitate processing, we converted the PDF files into text. For cases where the PDFs contained handwritten notes, we used Optical Character Recognition (OCR) to extract the textual information. The extracted text was then segmented by sentences or

periods, producing multiple distinct text segments for each patient. Each segment was then transformed into separate embeddings, allowing us to create a comprehensive set of embeddings for every individual patient. This multi-layered embedding strategy ensured that all relevant data modalities were captured and integrated accurately for further analysis.

Recognizer Network Architecture

We developed a deep learning-based recognizer network to identify the constituent modality embeddings within an aggregated embedding vector. The objective of this model was to determine whether it is possible to recover the composition of an aggregated embedding. specifically, which modality types (RNA, H&E, mutation, or annotation) were included, based solely on the vector representation. This task reflects a compositional recognition problem, where the challenge lies in disentangling heterogeneous embeddings that originate from unaligned latent spaces.

The recognizer network architecture consists of a series of shared layers followed by modality-specific output branches. The shared layers, common to all input embeddings, are designed to extract generalizable features from the aggregated vectors. These layers include standard components such as fully connected layers with activation functions, batch normalization, and dropout for regularization. After this shared feature extraction stage, the network branches into multiple modality-specific layers, each tailored to predict the presence or absence of a particular embedding type within the input vector. This

modular structure allows the network to maintain generalization while preserving sensitivity to distinct modality-specific features.

To train the recognizer network, we used aggregated embeddings constructed from real patient data across four modalities: RNA-seq, H&E image embeddings, somatic mutations, and clinical annotations. The training data was generated using the modality-level sampling strategy (first strategy), in which aggregated embeddings were composed by randomly sampling modality embeddings without regard to patient identity. This setup allowed us to assess the network's ability to learn modality-specific signatures in a mixed latent space.

We evaluated the recognizer network under two complementary experimental settings. In the simple aggregation approach, models were trained and tested on embeddings generated using fixed sample counts (isolation models), while in the composite setting, the recognizer was trained on a mixture of sample counts to assess robustness to input heterogeneity. This design enabled us to quantify the recognizer's capacity to infer embedding composition under both controlled and variable aggregation conditions.

Embedding Generation

To capture the unique characteristics of each modality, RNA expression, somatic mutations, histopathology (H&E) images, and patient-level annotations, we generated modality-specific embeddings designed to reside in a shared representational space. Given the inherent heterogeneity in feature types and dimensionalities across these data

sources, our objective was to project each modality into a compatible latent space that retained biological relevance while enabling downstream integration.

For RNA, mutation, and image data, we employed Variational Autoencoders (VAEs) to learn compact, biologically meaningful latent representations. For patient-level annotations, we leveraged a pretrained Sentence-BERT (sBERT)model, which produces 768-dimensional embeddings optimized for semantic similarity in textual data [135]. To align all modality representations, each VAE was configured to output embeddings of 768 dimensions, matching the fixed dimensionality of sBERT. This uniform representation allowed for straightforward aggregation and cross-modality comparison.

Each VAE was trained separately on data from all six cancer types to ensure generalizability and to preserve the biological variability inherent across different tumor types. After training, embeddings were extracted using the encoder component of each VAE, enabling compression of modality-specific input into a biologically informative latent space.

For somatic mutations, the mutation VAE (NETVAE) was trained on a one-hot encoded mutation matrix. Training was performed with a batch size of 256 and incorporated an early stopping criterion with a patience of 10 epochs to prevent overfitting. A maximum of 50 training epochs was allowed. The resulting encoder produced a 768-dimensional latent vector that effectively captured mutational patterns across patients.

For RNA expression data, a dedicated VAE was trained using RNA-seq data from The Cancer Genome Atlas (TCGA). Training was performed in two stages: an initial pretraining phase followed by fine-tuning. A warm-up strategy was applied to the Kullback–Leibler (KL)

divergence term in the VAE loss function to improve training stability. Specifically, KL loss was initialized at zero and gradually increased according to an annealing schedule governed by a κ (kappa) parameter. In our implementation, κ was set to 1, and the KL loss weight (β) was initialized at 0, resulting in a one-epoch warm-up phase. This annealing technique is known to promote more stable convergence by allowing the model to learn meaningful reconstructions before imposing latent space regularization.

For H&E histopathology images, whole-slide images were partitioned into non-overlapping 256×256 pixel tiles. Tiles predominantly containing background were removed using a simple RGB pixel thresholding heuristic. Remaining tiles were processed with Prov-Gigapath, a whole-slide foundation model pre-trained on large-scale pathology datasets [136]. Tile-level inference produced 1536-dimensional embeddings, which were then truncated to the final 768 dimensions. This truncation strategy was informed by internal benchmarking, which indicated that the latter half of the embedding vector retained sufficient discriminative power for downstream tasks. The resulting tile-level embeddings were used in all subsequent analyses.

Embedding Aggregation

To construct a comprehensive dataset for training and evaluating recognition models, we generated aggregated embeddings composed of three to ten constituent embeddings per vector. This range ensured exposure to varying levels of aggregation complexity, enabling the model to generalize across different embedding combinations. The embedding construction process followed two distinct sampling strategies: the simple aggregation

approach and the specific-cancer approach, corresponding to our previously described first and second strategies, respectively.

For the simple aggregation approach, all available embeddings across modalities, RNA, H&E, mutations, and annotations, were first loaded into memory. A random sampling procedure was applied, selecting a specified number of embeddings (e.g., sample count of 3 to 10) regardless of their patient or cancer type of origin. The selected embeddings were summed to produce a new 768-dimensional aggregated embedding, and the modalities contributing to the aggregation were tracked and stored as ground truth labels for supervised training.

In the specific-cancer approach, the sampling procedure was constrained to enforce cancer-type specificity. After loading all available embeddings, random sampling was performed such that all selected embeddings originated from the same cancer type (e.g., only embeddings derived from BRCA patients). As in the simple approach, the selected embeddings were summed to form a 768-dimensional vector. Both the modalities involved in the aggregation and the corresponding cancer type were recorded as ground truth labels. This design enabled us to evaluate the recognizer model's performance under both general and cancer-specific conditions, while preserving control over the composition and structure of the aggregated embeddings.

Training the Recognizer Network

To evaluate whether the composition of aggregated embeddings could be recovered, we trained a deep learning-based recognizer network on vectors composed of three to ten

constituent embeddings. These embeddings were generated through the simple aggregation approach, with additional models trained under cancer-type-specific sampling to assess performance under more biologically constrained conditions.

Aggregated embeddings were constructed by randomly selecting and summing modality-specific embeddings drawn from RNA expression, H&E images, somatic mutations, and patient-level annotations. This procedure produced 768-dimensional vectors, each labeled with the set of contributing modalities and, when applicable, the associated cancer type. The resulting dataset formed the input for training isolation recognizer models, each trained on a specific sample count, as well as a composite recognizer model, trained on the full range of sample counts (3–10) within a unified dataset.

Training Procedure

The recognizer network was trained using supervised learning. The input to the model was the aggregated embedding vector, and the output was a binary vector indicating the presence or absence of each modality in the aggregation. The model was trained using the Adam optimizer with an initial learning rate of 0.001, and the mean squared error (MSE) loss function was used to penalize incorrect predictions of modality presence.

The network architecture consisted of an initial set of shared layers designed for general feature extraction, followed by modality-specific output branches that performed binary classification for each of the four modalities. To improve generalization and stability, the architecture incorporated fully connected layers, dropout, and batch normalization.

To prevent modality imbalance during training, we explicitly controlled the probability of selecting each modality during the random sampling process. For example, when using four modalities, each was assigned a uniform selection probability of 25%, ensuring equal representation across training batches and avoiding bias toward more frequently represented modalities.

Validation and Testing

The dataset was split into training, validation, and test sets using an 80/20 split, followed by an 80/20 subdivision of the training set into training and validation subsets. The validation set was used during training to monitor model generalization and tune hyperparameters, while the test set was held out entirely for final performance evaluation.

Composite Recognizer Model

In addition to training individual isolation models for each sample count, we trained a composite recognizer model using a merged dataset containing all aggregated embeddings with sample counts from 3 to 10. This model was designed to assess whether a single network could generalize across varying levels of aggregation complexity and modality combinations. The same training architecture and optimization parameters were applied to the composite model, enabling direct comparison with the isolation-based models.

Metric calculation for recognizer network evaluation

The evaluation of the recognizer network required special consideration due to the sparse structure of the training data, which was generated using the first sampling strategy,

modality-level sampling. In this approach, embeddings were randomly selected across modalities without regard to patient identity, and aggregated to create synthetic vectors composed of three to ten constituent embeddings. Because any given modality may or may not have been included in a given aggregated vector, the resulting label matrix contained a high proportion of zeros. Additionally, constraints in the data generation process (e.g., limiting the number of embeddings or excluding certain modalities) led to systematically enforced zero entries in specific columns.

This sparsity posed challenges for standard classification metrics such as accuracy, F1 score, and Matthews correlation coefficient (MCC). A trivial model that always predicts absence (i.e., zeros for all modality labels) could achieve high accuracy, especially in test sets with many modality-absent samples. To ensure correct and informative evaluation, we stratified metric computation for each modality (RNA, H&E, somatic mutations, and clinical annotations) by separating test samples into two groups:

1. Zero-labeled samples, in which a modality was absent from the aggregated vector
2. Non-zero-labeled samples, in which a modality was present and thus had to be correctly recognized

Performance metrics, MCC, accuracy, precision, recall, and F1 score, were computed independently within each group. This allowed us to assess the recognizer model's capacity to both correctly detect present modalities and avoid false positives for absent ones, preventing inflated performance due to label imbalance.

For models trained under noise-perturbed conditions, we applied the same stratification logic. These datasets were derived from clean aggregated vectors (generated via modality-level sampling) with Gaussian noise introduced to replace a subset of embeddings. While the total number of embeddings per vector remained unchanged, the substitution of valid modality embeddings with noise created further sparsity in the ground truth labels. In this setting, models predicting only absent modalities could again appear to perform well based on raw accuracy, despite lacking real compositional understanding. As with the standard setting, excluding zero-only rows and stratifying metrics enabled a more meaningful evaluation of the model's robustness.

By applying these adjustments, we ensured that the reported metrics provided a reliable assessment of the recognizer network's performance in identifying the constituent modality composition of aggregated embeddings, both under standard conditions and in the presence of noise.

Gaussian Mixture Models

To enhance the visualization of information retrieval, we employed Gaussian Mixture Models (GMMs). GMMs are probabilistic models that represent complex data distributions as a combination of multiple Gaussian components, each weighted to capture distinct patterns or clusters. This approach is widely used in unsupervised learning for clustering and anomaly detection. In our analysis, we constructed a GMM with six components to evaluate its ability to distinguish between the six cancer types in our dataset. The primary goal was to compare a 2D principal component analysis (PCA) visualization with GMM

clustering to assess their alignment. While biological non-linearity prevents a clear-cut separation between cancer types, distinct cluster boundaries emerged, highlighting the potential of GMMs in capturing underlying structures in high-dimensional biological data.

Classification Model

The classification model was trained using a supervised learning approach, with input features consisting of aggregated and concatenated patient-level embeddings constructed via the second sampling strategy (patient-level sampling). Each embedding was paired with a ground truth label corresponding to one of three classification targets: cancer type, cancer subtype, or tumor mutational burden (TMB).

Data were split using an 80/20 train-test split, with training and testing set ratios held consistent across all experimental runs. To ensure reliable and robust performance estimation, the model was trained and evaluated across at least 30 independent runs. This repeated training procedure reduced the impact of stochastic variability introduced by random initialization and data shuffling, and enabled the calculation of stable performance metrics across runs. Results were aggregated to report average performance, allowing for a more accurate and reproducible assessment of the model's predictive capacity.

Tumor Mutational Burden (TMB) Calculation

To calculate the tumor mutational burden (TMB) for each patient, we used the one-hot encoded somatic mutation file, which was used to generate the mutation embeddings, where each gene's mutation status was encoded as either mutated (1) or not mutated (0).

For each patient, we counted the number of mutated genes by summing all values across the remaining columns.

To standardize the calculation, we assumed an exonic coverage of 30 megabases (Mb) and computed TMB as the number of mutations per megabase using this formula: *TMB = Number of Mutations / 30*. [137]

We then classified TMB into high and low categories based on a predefined threshold. A TMB value of 0.5 mutations/Mb or higher was classified as high (1), while values below this threshold were classified as low (0).

If no mutations were detected for a patient, TMB was set to 0, and the classification was assigned a separate category (2) to account for cases where no mutations were present.

This approach ensures a consistent and interpretable quantification of tumor mutational burden across patients while enabling classification into biologically relevant groups.

Tumor Subtypes

We utilized publicly available resources to map patients and cancer types to their corresponding cancer subtypes. Specifically, subtype annotations were derived using data and tools from the Genomic Data Commons (GDC) [138] and the subtype mapping file provided in the associated repository (https://github.com/NCICCGPO/gdan-tmp-models/blob/main/tools/cancer2name.json). Subtype classification was performed only for cancer types where a sufficient number of annotated samples (n > 100) were available to support reliable supervised learning.

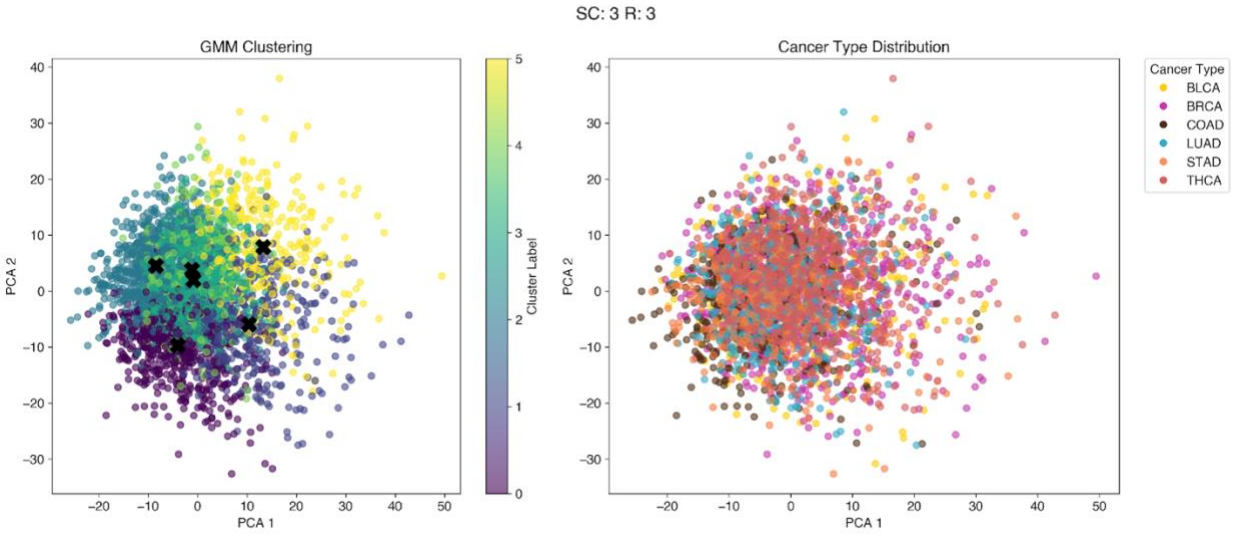
Figures

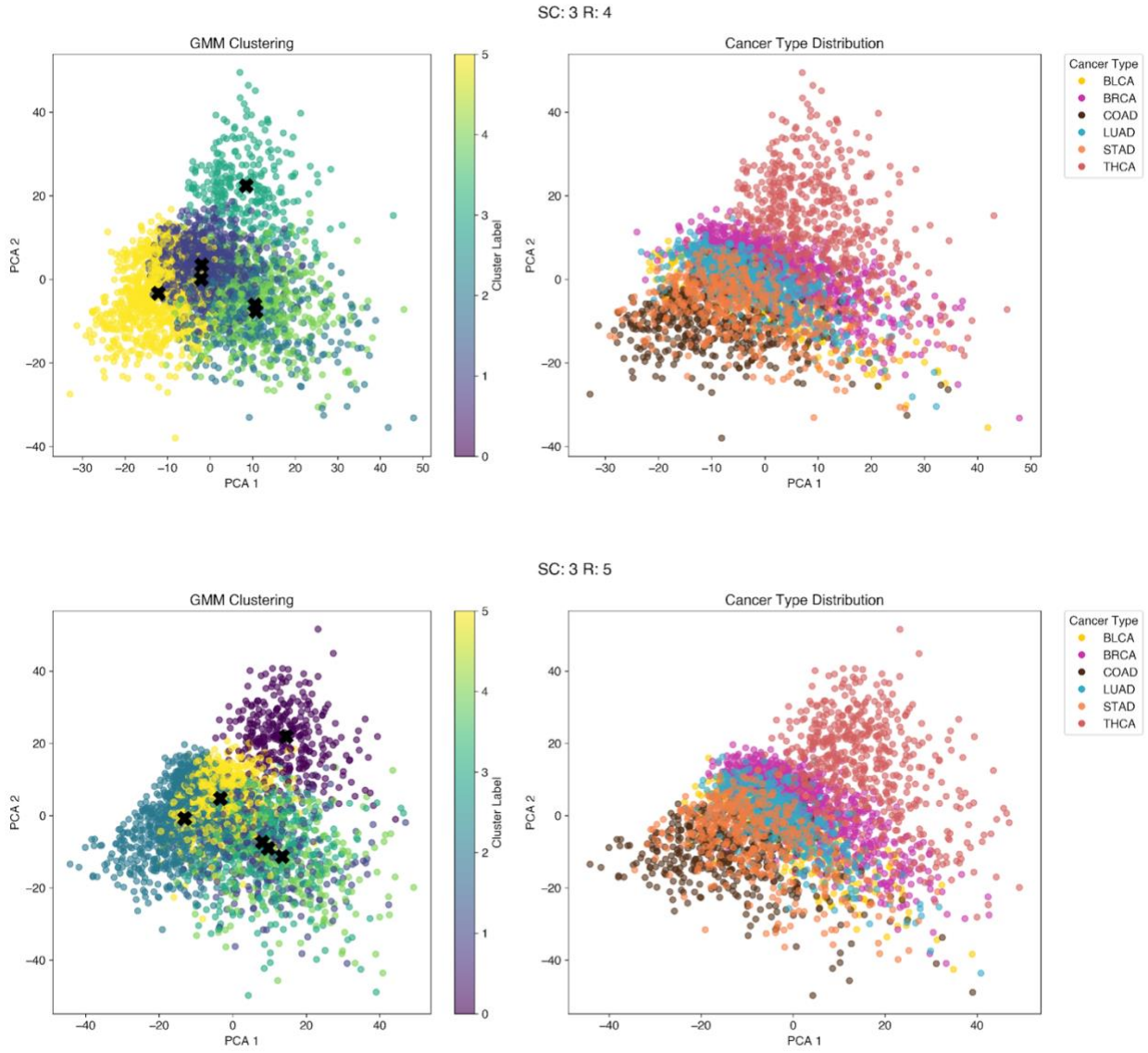
Table 3.1 Number of patients available per data modality. RNA data includes the largest cohort with 3,558 patients, while mutation data includes the smallest cohort with 3,189 patients.

Modality	Patients
RNA	3558
Mutation	3189
H&E	3442
Annotations	3443

Table 3.2 Detailed distribution of available embeddings per cancer type across data modalities. Each modality includes at least one embedding per patient. RNA and mutation modalities consistently provide one embedding per patient. In contrast, annotations and H&E image modalities often include multiple embeddings per patient, with the image modality exhibiting the widest distribution.

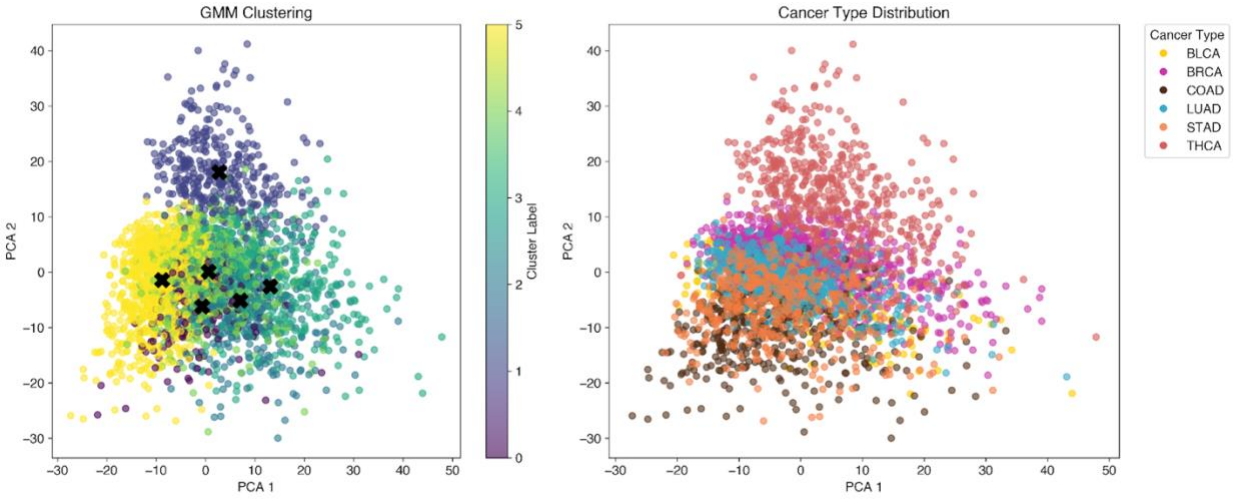
Modality	BRCA	BLCA	COAD	STAD	THCA	LUAD
RNA	1000	430	514	453	572	589
Mutation	1003	407	404	431	435	509
H&E	329400	123600	138000	132900	152100	156600
Annotations	69366	40236	14980	12595	30136	37452



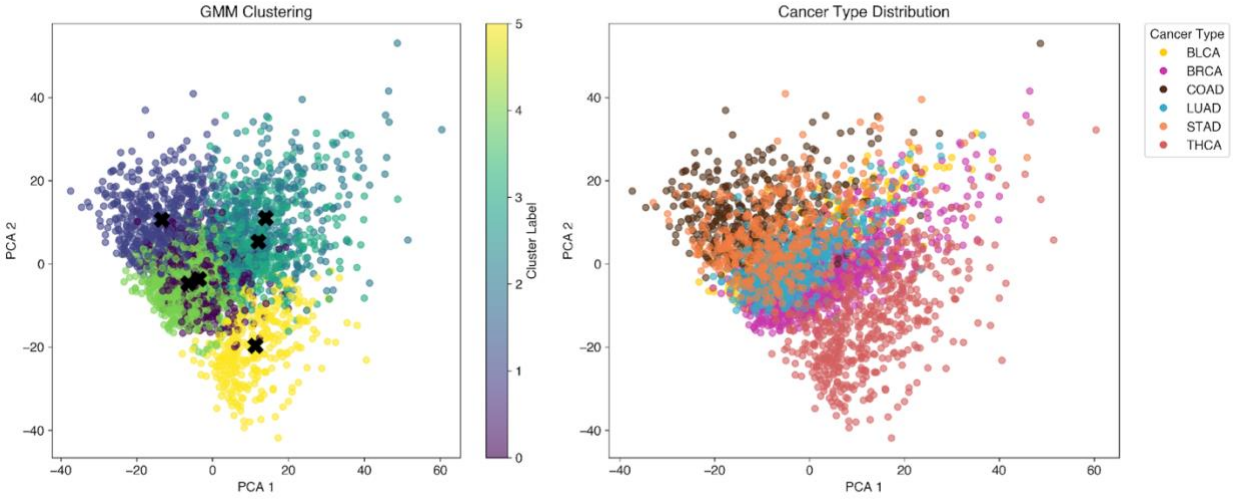


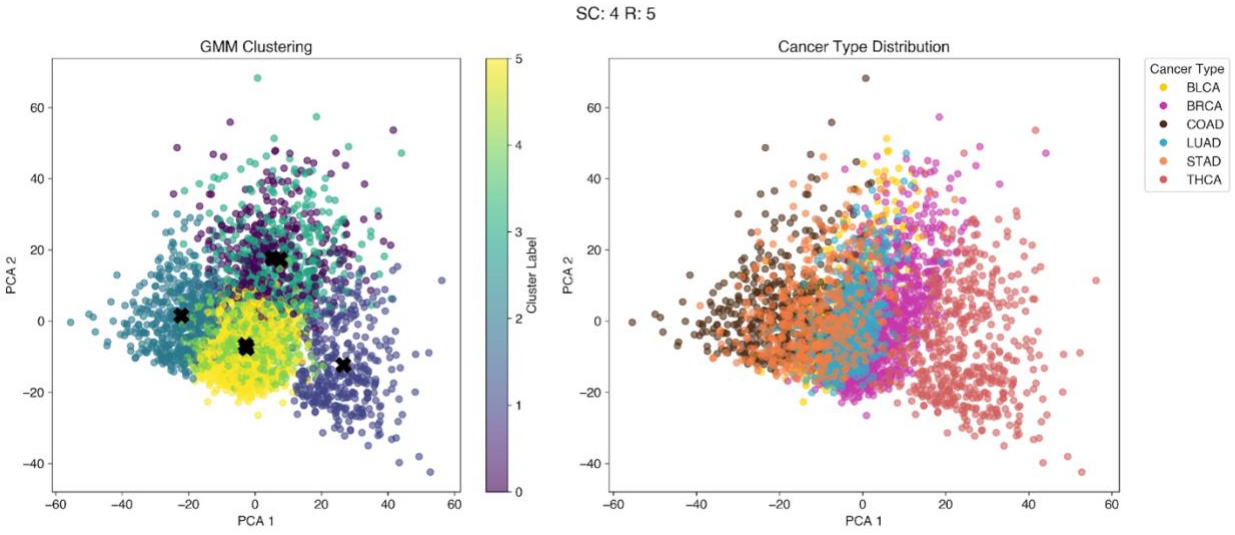
Supplementary Figure 3.8 Clustering performance using 3 repetitions. Clustering performance with a fixed sample count of 3 and increasing sampling counts from 3 to 5 demonstrates progressive enhancement in cluster separation. With 3 repetitions, clustering lacks clear separation, as reflected by the Gaussian Mixture Model (GMM) clustering. Increasing the repetition count improves both cancer type and GMM clustering, resulting in greater separation and more distinct cluster formation.

SC: 4 R: 3

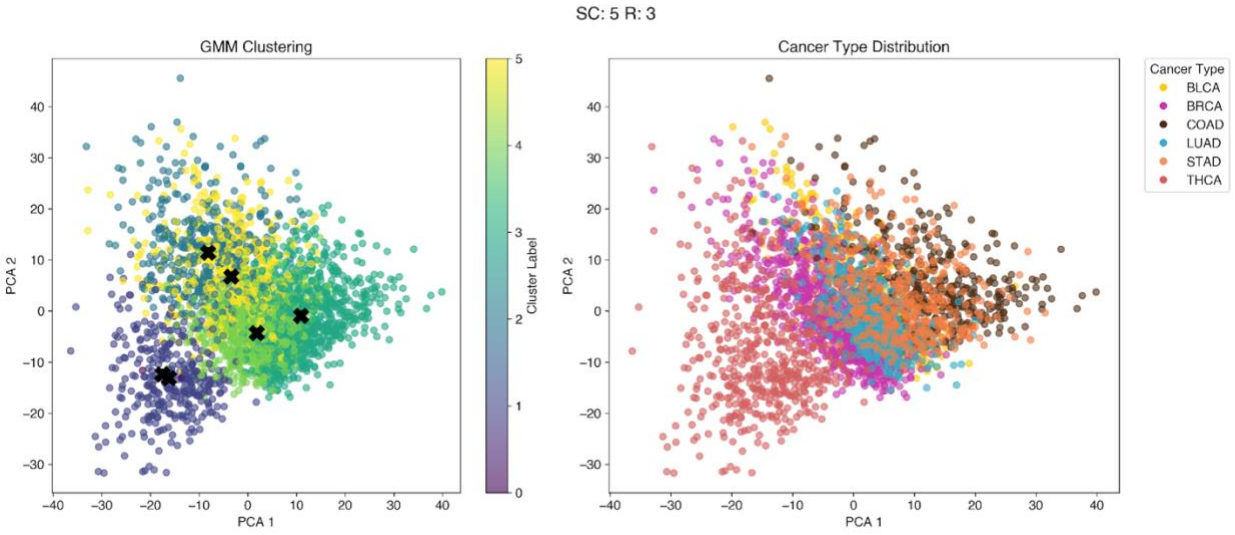

GMM Clustering
PCA 1
PCA 2
Cluster Label
Cancer Type Distribution
Cancer Type
BLCA
BRCA
COAD
LUAD
STAD
THCA

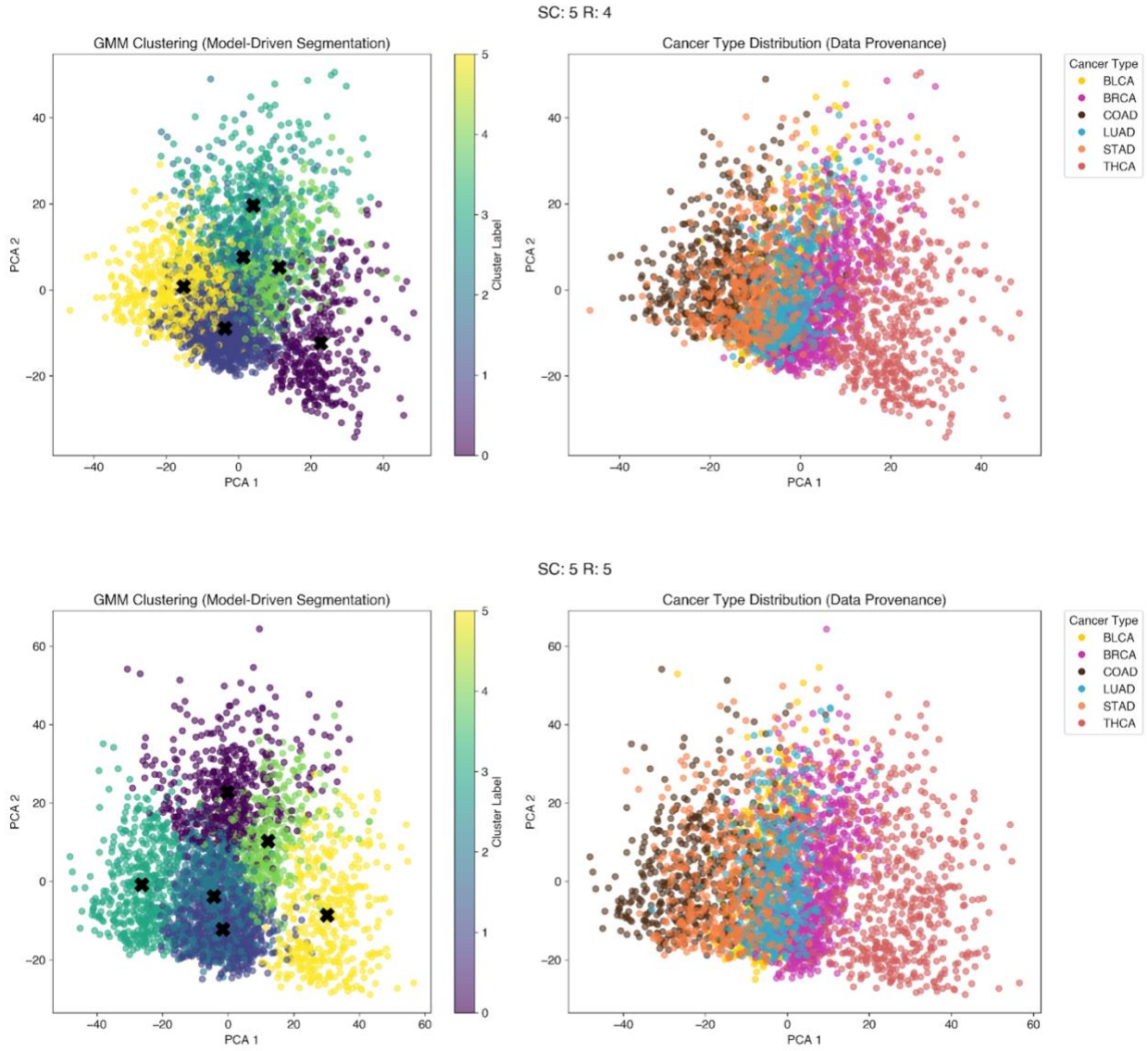
SC: 4 R: 4


GMM Clustering
PCA 1
PCA 2
Cluster Label
Cancer Type Distribution
Cancer Type
BLCA
BRCA
COAD
LUAD
STAD
THCA

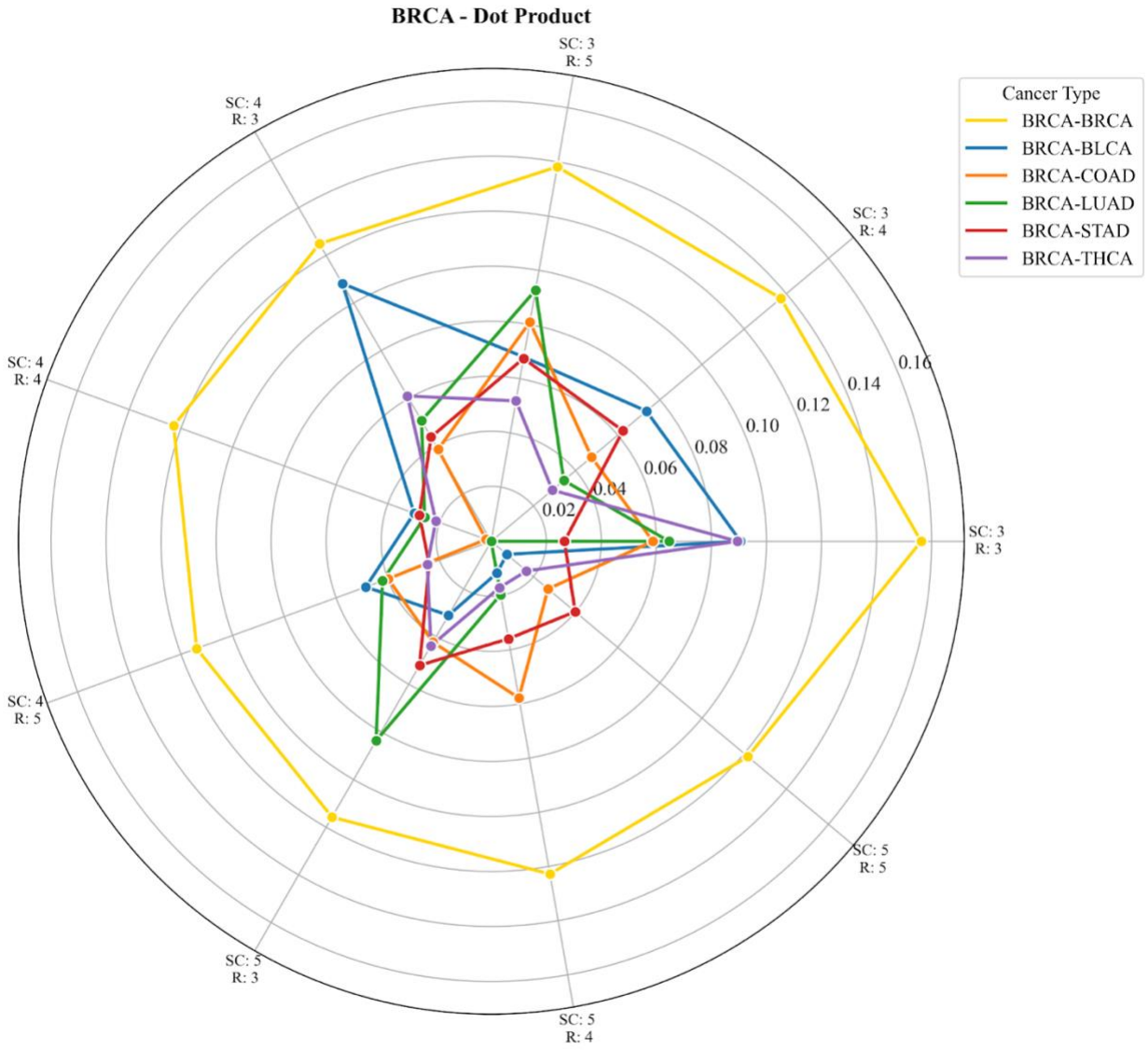


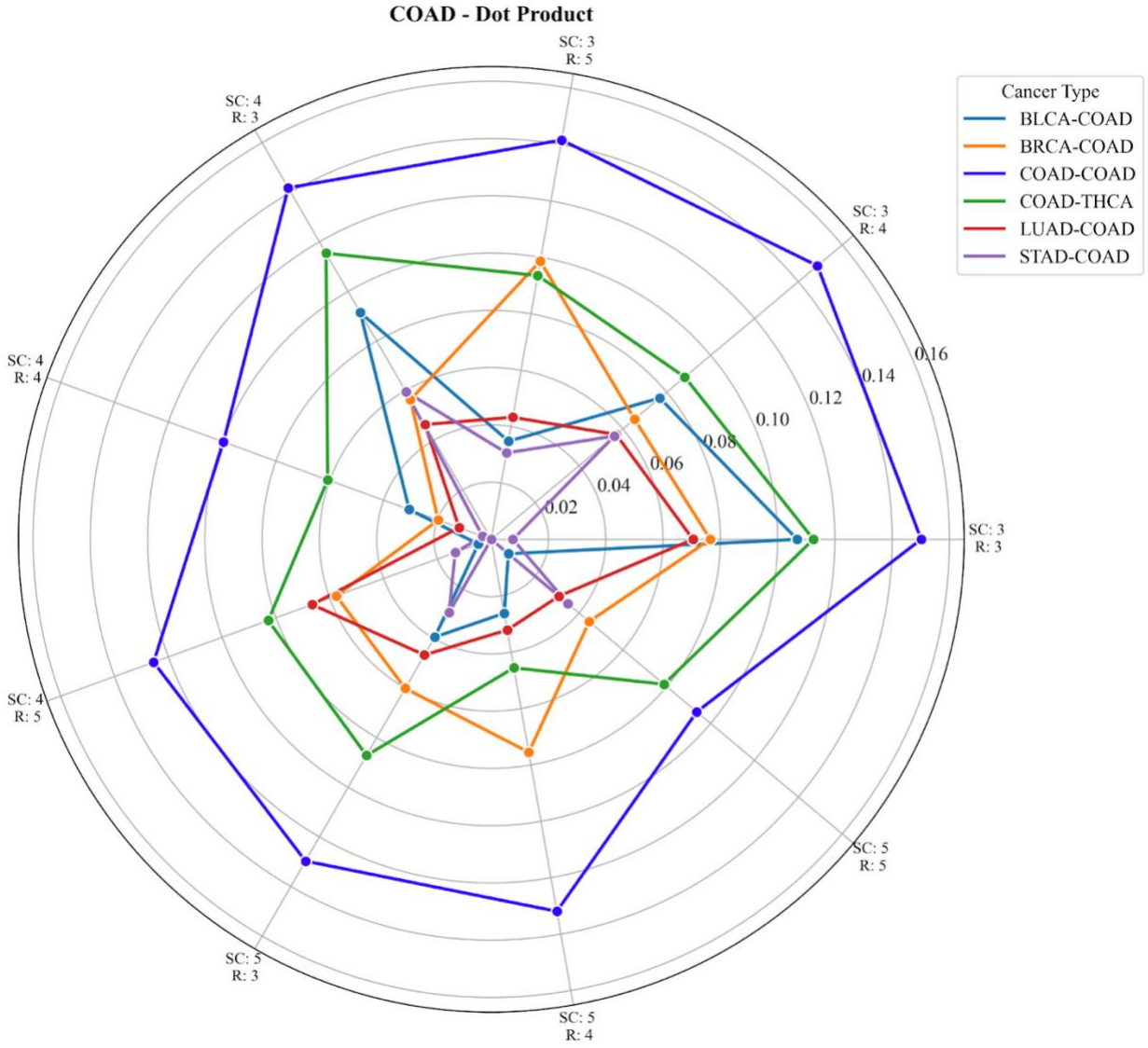
Supplementary Figure 3.9 Clustering performance using 4 repetitions. Clustering performance with a fixed sample count of 4 and varying repetition counts between 3 and 5 shows enhanced clustering quality with increasing repetition. As repetition count increases, both Gaussian Mixture Model (GMM) clustering and cancer type clustering exhibit improved alignment and separation, indicating more robust cluster formation.

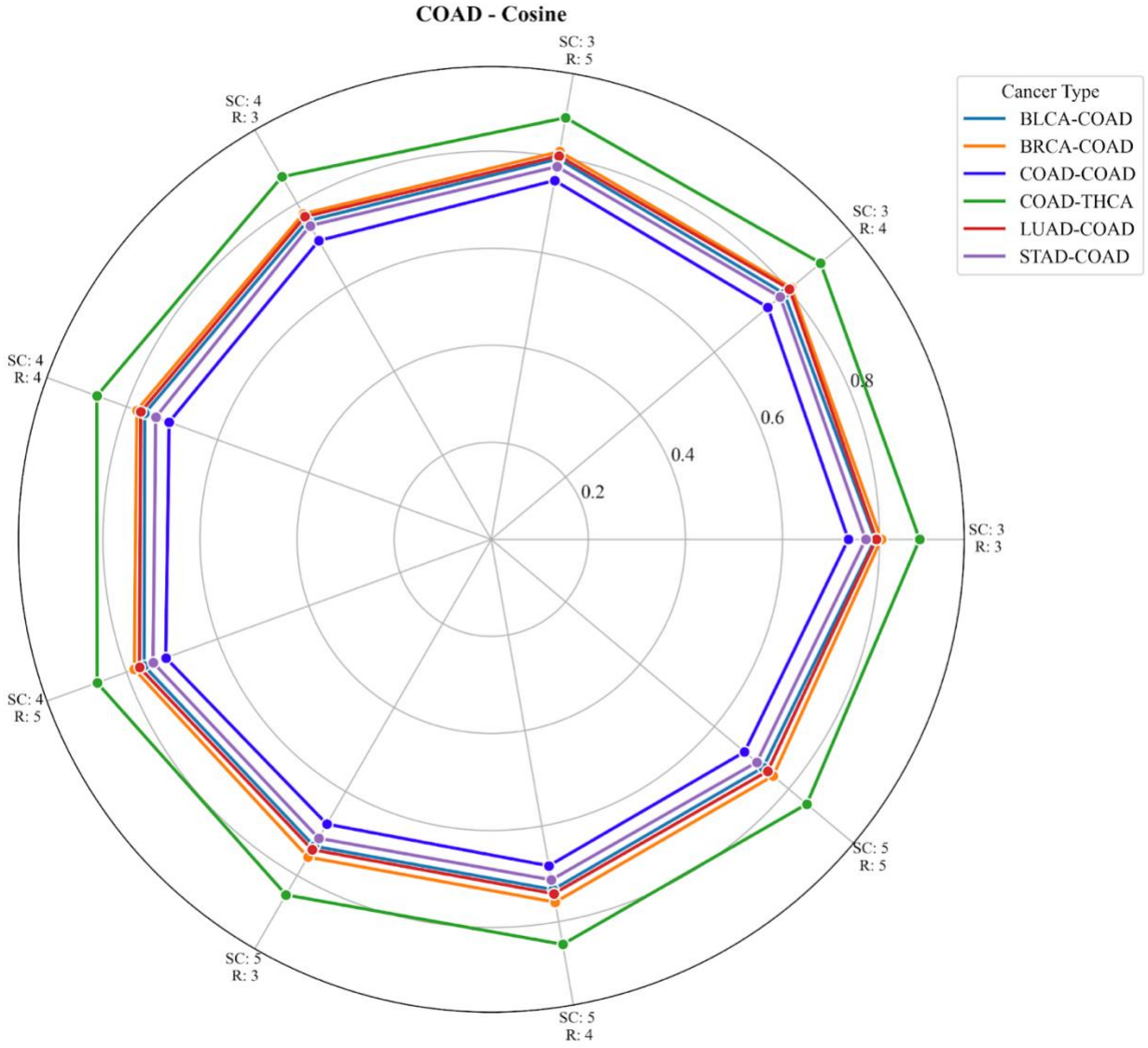


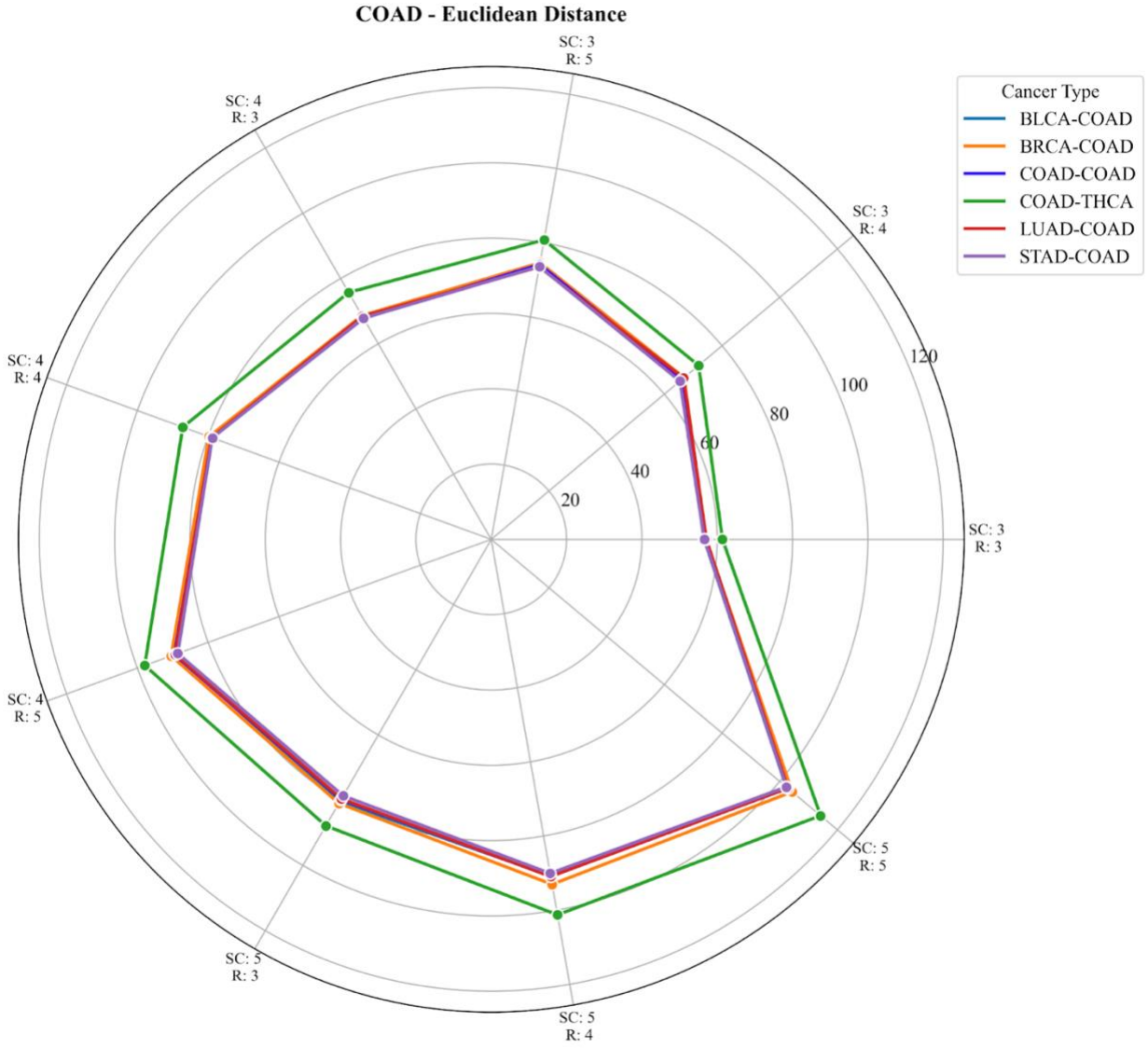


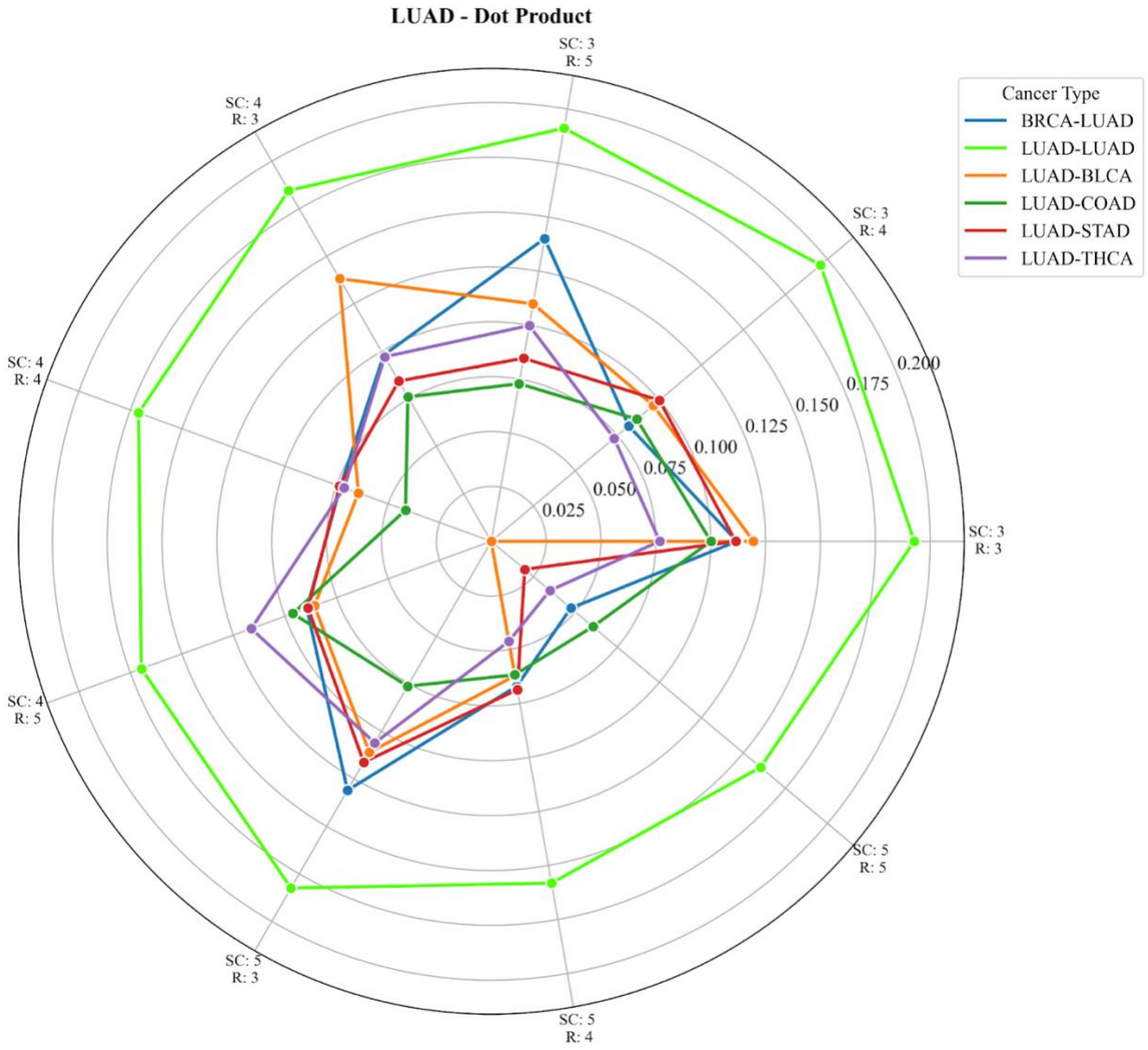
Supplementary Figure 3.10 Clustering performance using 5 repetitions. Clustering performance with a fixed sample count of 5 and varying repetition counts between 3 and 5 demonstrates improved clustering with increasing repetition. Additionally, the Gaussian Mixture Model (GMM) unsupervised clustering aligns well with the cancer subtype clusters, indicating strong agreement between predicted and true cluster structures.

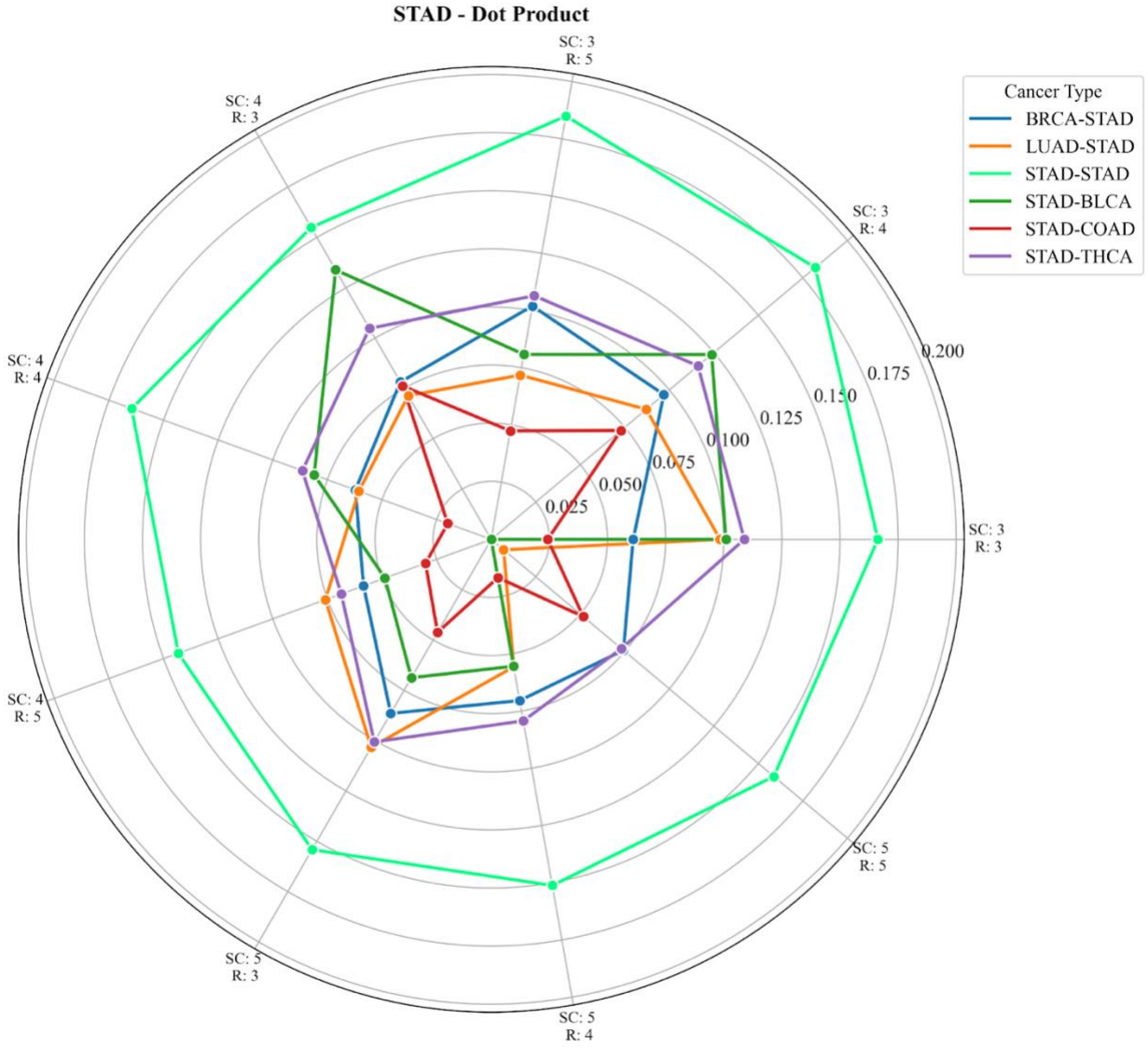

BRCA - Dot Product
SC: 3
R: 5
SC: 3
R: 4
SC: 3
R: 3
SC: 4
R: 5
SC: 4
R: 4
SC: 4
R: 3
SC: 5
R: 5
SC: 5
R: 4
SC: 5
R: 3
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Cancer Type
BRCA-BRCA
BRCA-BLCA
BRCA-COAD
BRCA-LUAD
BRCA-STAD
BRCA-THCA

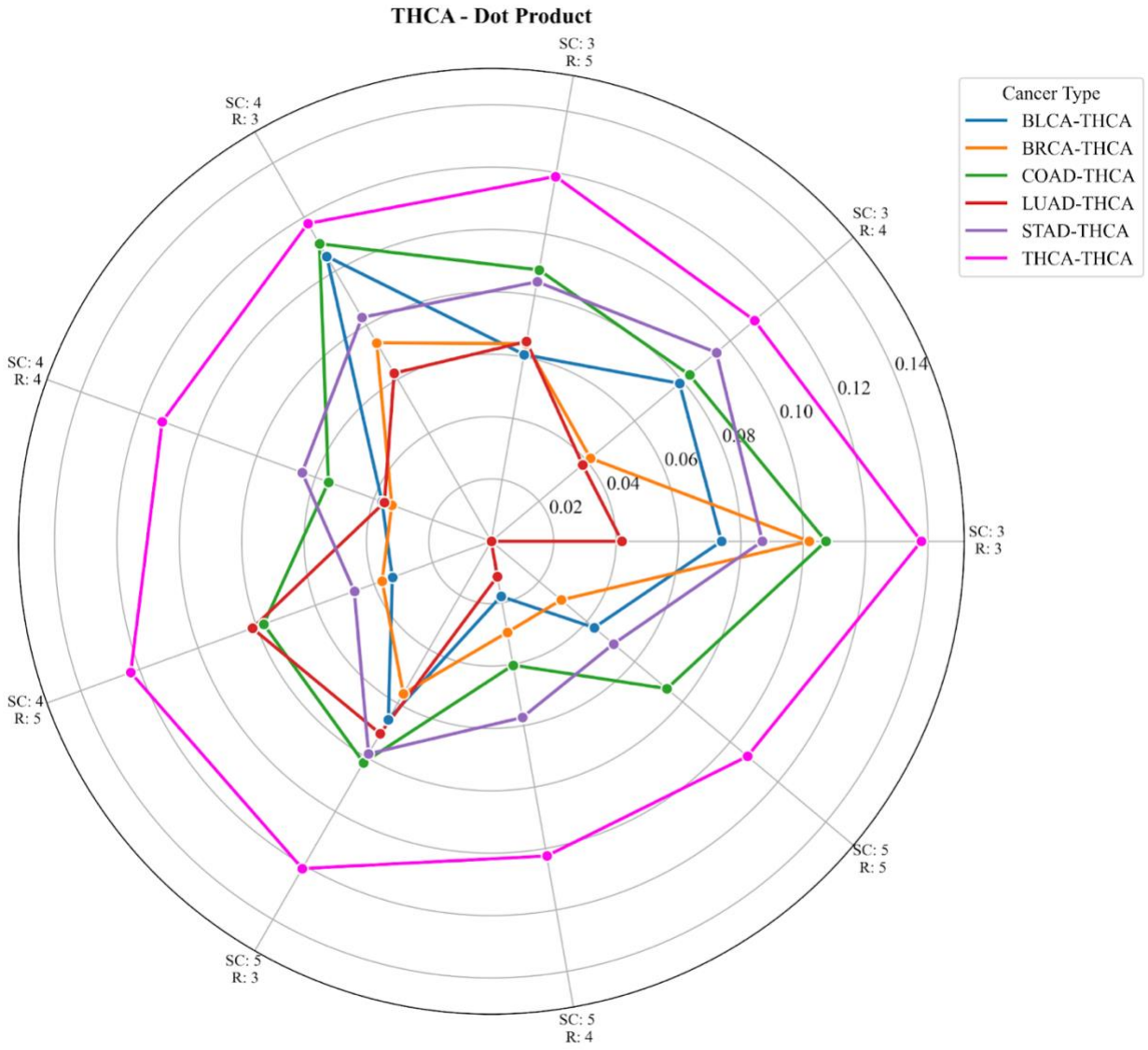

COAD - Dot Product
SC: 3
R: 5
SC: 3
R: 4
SC: 3
R: 3
SC: 4
R: 3
SC: 4
R: 4
SC: 4
R: 5
SC: 5
R: 3
SC: 5
R: 4
SC: 5
R: 5
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Cancer Type
BLCA-COAD
BRCA-COAD
COAD-COAD
COAD-THCA
LUAD-COAD
STAD-COAD

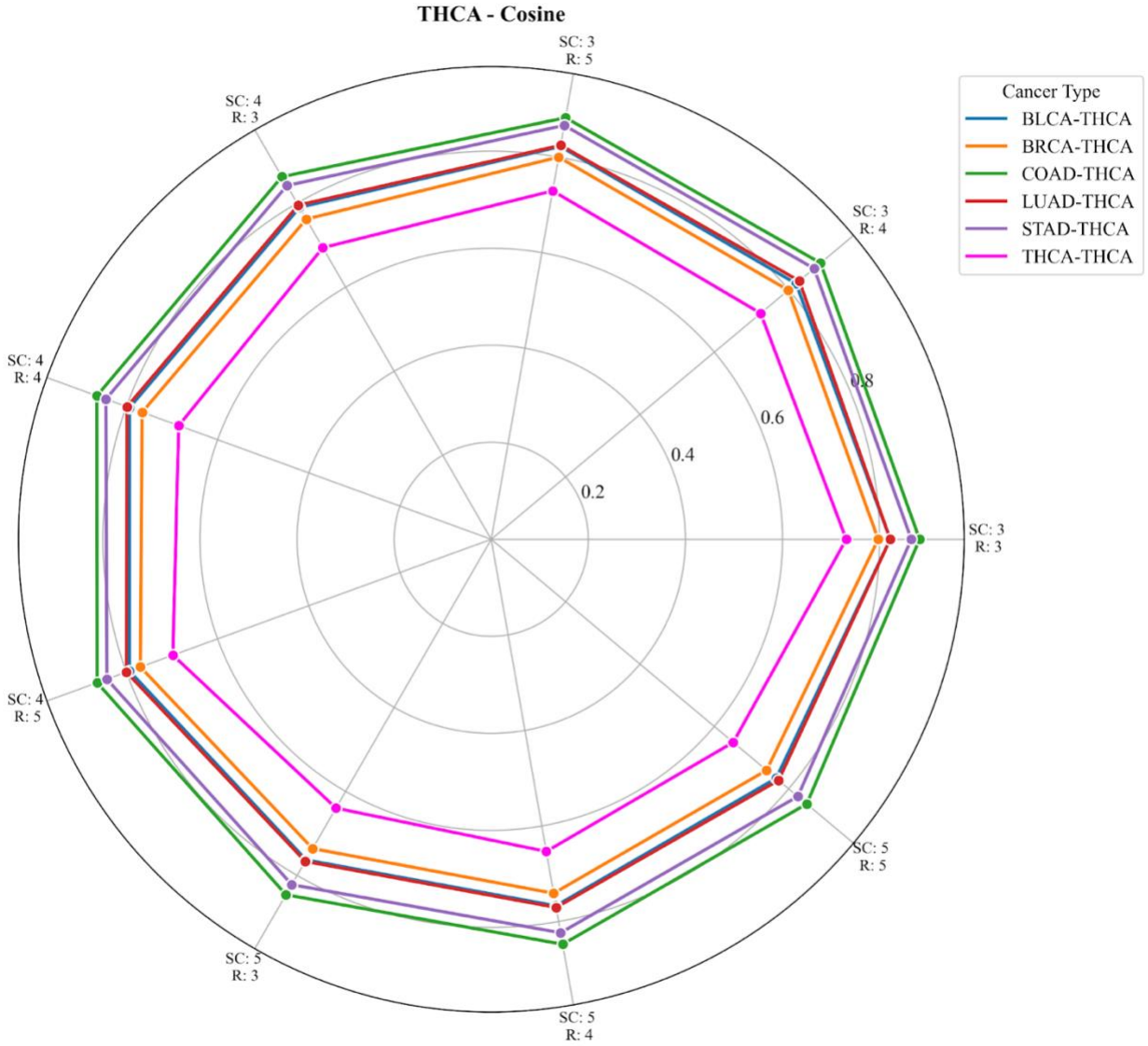

COAD - Cosine
SC: 3
R: 5
SC: 3
R: 4
SC: 3
R: 3
SC: 4
R: 3
SC: 4
R: 4
SC: 4
R: 5
SC: 5
R: 5
SC: 5
R: 3
SC: 5
R: 4
0.2
0.4
0.6
0.8
Cancer Type
BLCA-COAD
BRCA-COAD
COAD-COAD
COAD-THCA
LUAD-COAD
STAD-COAD


COAD - Euclidean Distance
SC: 3
R: 5
SC: 3
R: 4
SC: 3
R: 3
SC: 4
R: 5
SC: 4
R: 4
SC: 4
R: 3
SC: 5
R: 5
SC: 5
R: 4
SC: 5
R: 3
20
40
60
80
100
120
Cancer Type
BLCA-COAD
BRCA-COAD
COAD-COAD
COAD-THCA
LUAD-COAD
STAD-COAD

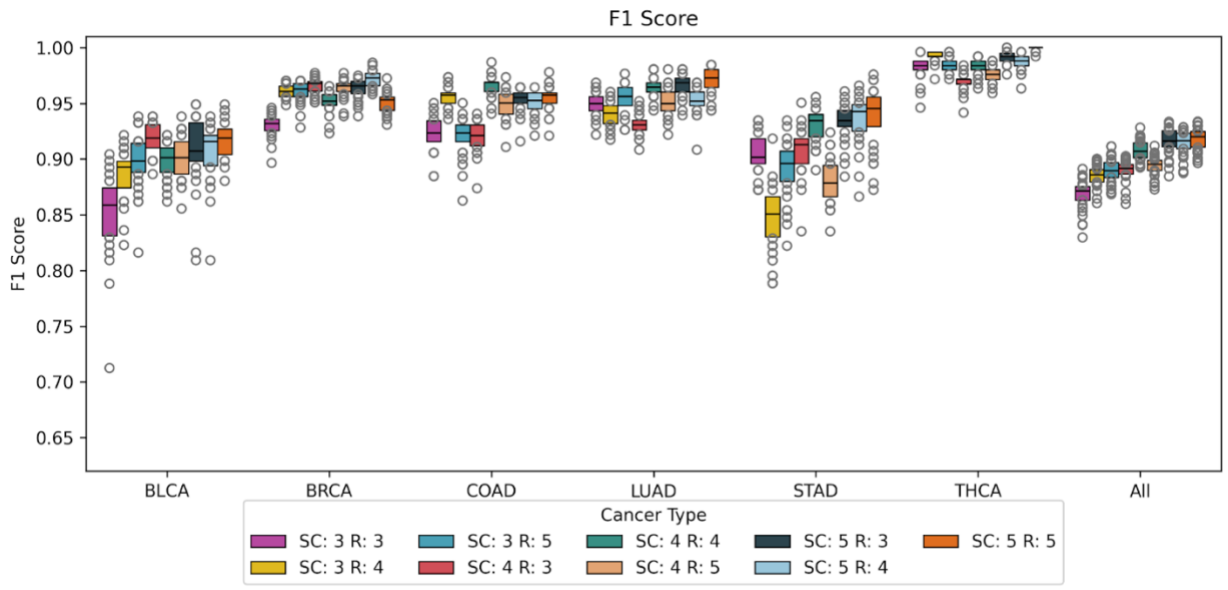

LUAD - Dot Product
SC: 3
R: 5
SC: 3
R: 4
SC: 3
R: 3
SC: 4
R: 3
SC: 4
R: 4
SC: 4
R: 5
SC: 5
R: 5
SC: 5
R: 3
SC: 5
R: 4
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Cancer Type
BRCA-LUAD
LUAD-LUAD
LUAD-BLCA
LUAD-COAD
LUAD-STAD
LUAD-THCA


STAD - Dot Product
SC: 3
R: 5
SC: 3
R: 4
SC: 3
R: 3
SC: 4
R: 3
SC: 4
R: 4
SC: 4
R: 5
SC: 5
R: 3
SC: 5
R: 4
SC: 5
R: 5
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Cancer Type
BRCA-STAD
LUAD-STAD
STAD-STAD
STAD-BLCA
STAD-COAD
STAD-THCA

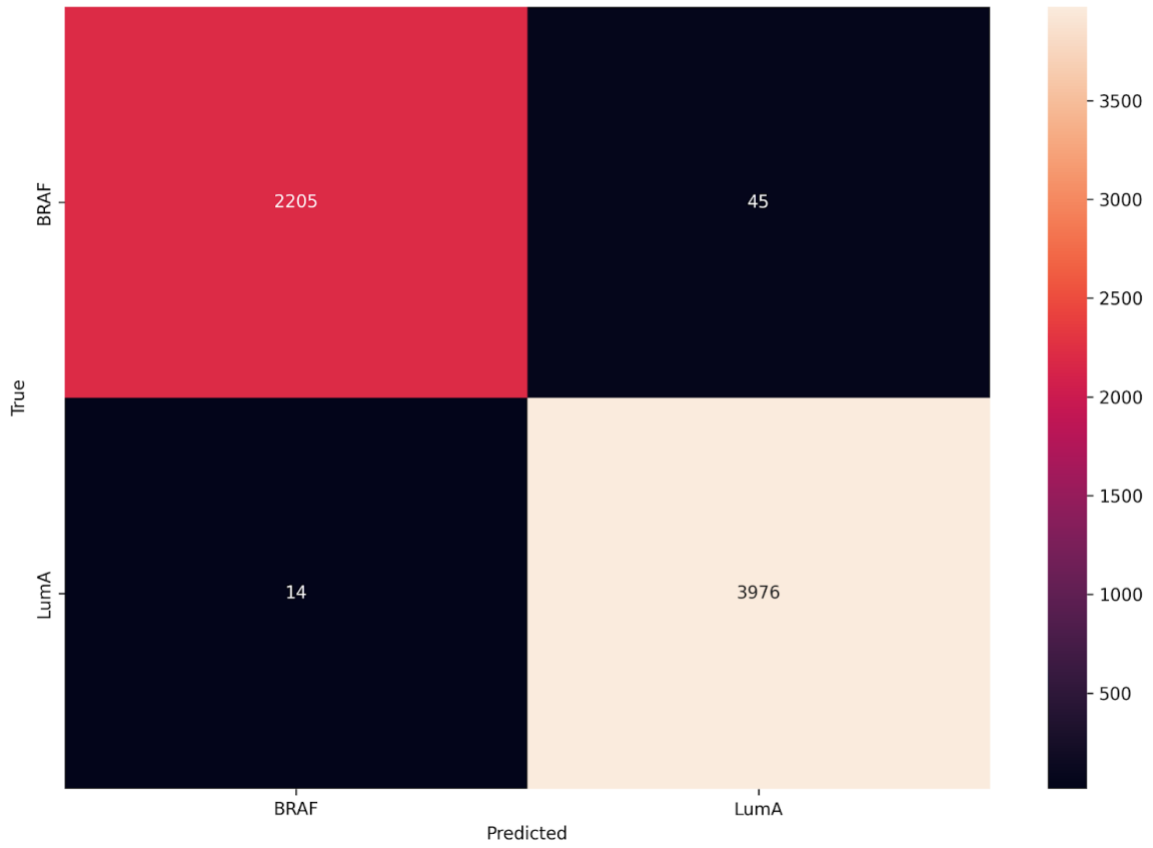

THCA - Dot Product
SC: 3
R: 5
SC: 3
R: 4
SC: 3
R: 3
SC: 4
R: 3
SC: 4
R: 4
SC: 4
R: 5
SC: 5
R: 5
SC: 5
R: 3
SC: 5
R: 4
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Cancer Type
BLCA-THCA
BRCA-THCA
COAD-THCA
LUAD-THCA
STAD-THCA
THCA-THCA



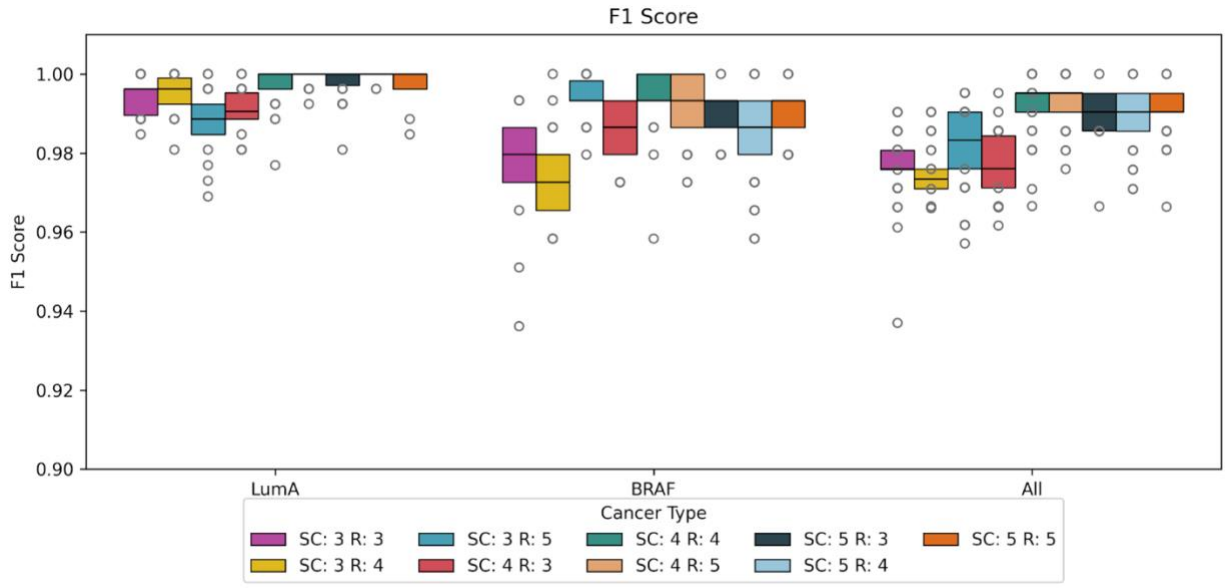
Supplementary Figure 3.11 Pairwise distance calculations between cancer types. Pairwise distance calculations between cancer types using Euclidean distance, cosine similarity, and dot product metrics demonstrate clear separation between intra- and inter-cancer distances. Cosine similarity and Euclidean distance are normalized between 0 and 1, whereas the dot product is only bounded at 0 with no defined upper limit.



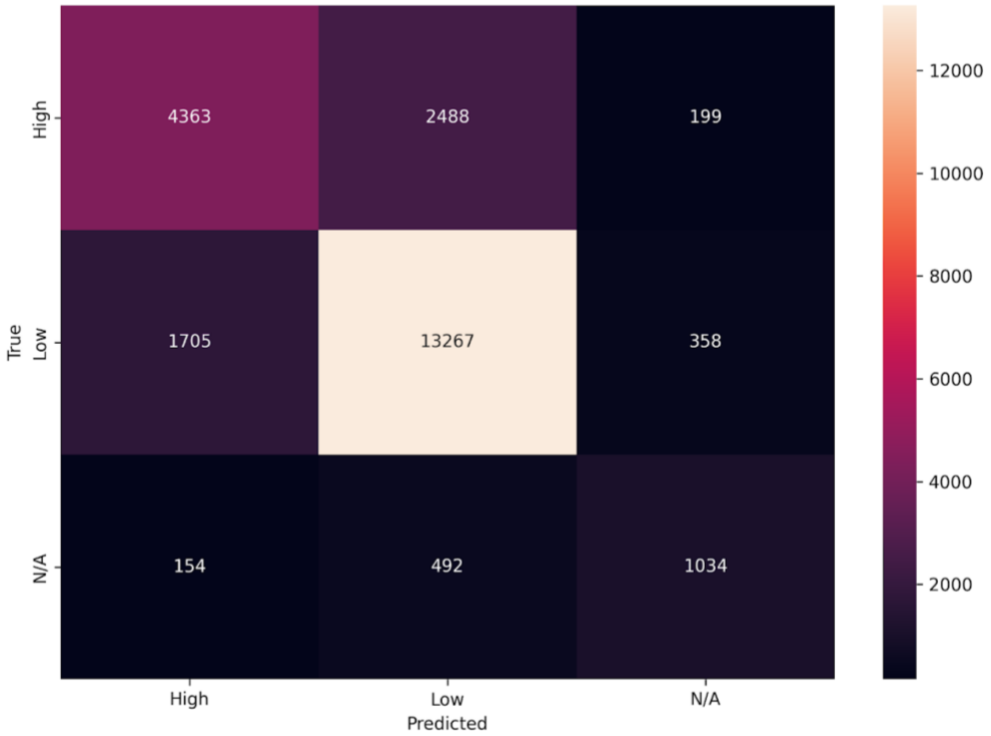
Supplementary Figure 3.12 F1 scores depicting classification performance. F1 scores depicting classification performance across six different cancer types, evaluated over varying sample counts (SC) and repetitions (R), indicate that increasing either sample count or repetition enhances classification performance across all cancer types. The overall average F1 score across all cancers ranges between 0.90 and 0.95, demonstrating robust and consistent performance.



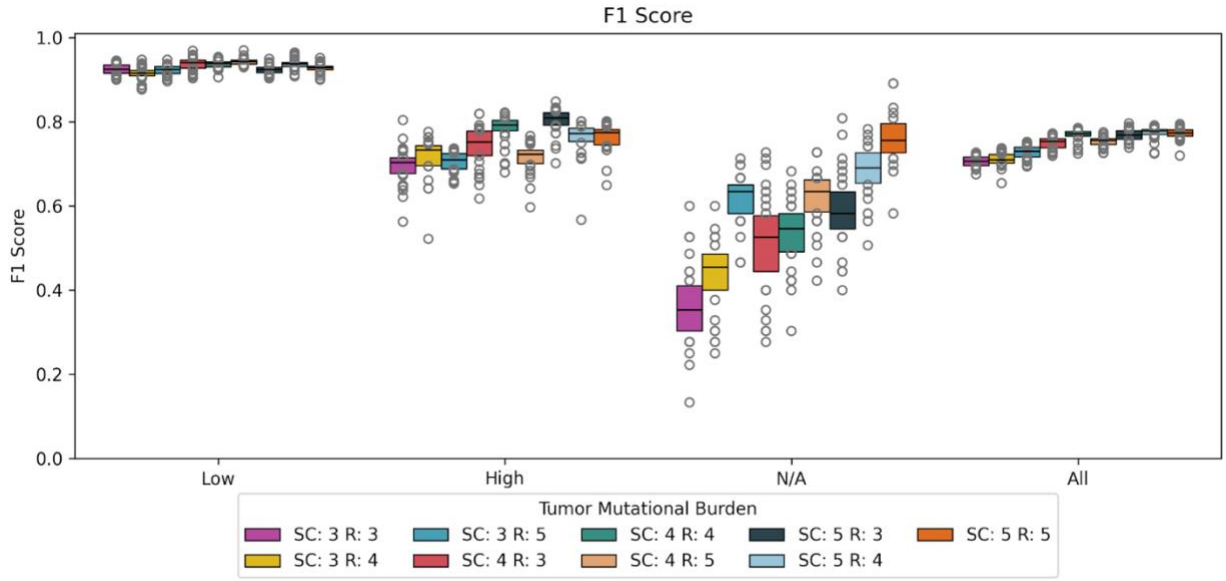
Supplementary Figure 3.13 True versus predicted subtype confusion matrix. Confusion matrix illustrating true versus predicted classifications for Luminal A (LumA) and BRAF cancer subtypes. Misclassifications are minimal, with a higher occurrence of BRAF samples being incorrectly classified as LumA.



Supplementary Figure 3.14 F1 performance for subtype classification. F1 scores for subtype classification of Luminal A (LumA) and BRAF compared with overall F1 performance across varying sample counts (SC) and repeats (R), ranging from 3 to 5. Performance remains consistently high, with F1 scores ≥ 0.9 across all conditions.



Supplementary Figure 3.15 True versus predicted tumor for tumor mutational burden confusion matrix. Confusion matrix depicting true versus predicted classifications of Low, High, and N/A tumor mutational burden (TMB). Elevated misclassifications are observed for the N/A class, predominantly predicted as Low TMB. Additionally, minor misclassifications are noted for the N/A class when predicted as High TMB.



Supplementary Figure 3.16 F1 scores classification performance for tumor mutation burden. F1 scores for Low, High, and N/A tumor mutational burden (TMB) classes, along with an overall average across all classes, are evaluated over varying sample counts (SC) and repetitions (R). The Low TMB class consistently achieves high performance with F1 scores ≥ 0.9, while the High TMB class maintains moderate performance ranging between 0.7 and 0.8. Notably, the N/A class exhibits poor performance with F1 scores < 0.4 when using a sample count of 3 and repetition of 3. However, increasing the sample count and repetition progressively improves N/A class performance, reaching an F1 score of approximately 0.8.

Chapter 4 Predicting anti-PD-1 immune checkpoint blockade response in melanoma patients with spatially aware machine learning models

This chapter has been formatted for inclusion in this dissertation from the manuscript " Predicting anti-PD-1 immune checkpoint blockade response in melanoma patients with spatially aware machine learning models " by Alyssa Pybus, Raphael Kirchgaessner, Jonathan Nguyen, Carlos Moran Segura, Jeremy Goecks, and Joseph Markowitz, submitted to Npj Precision Oncology (2025). The author of this dissertation is a shared first author of this manuscript and used multiplex immunofluorescence assays to conduct the computational experiments, the results of which were used to generate Main Figure 6 and Supplementary Figures 11 & 12.

Abstract

There is an acute need to accurately identify patients with advanced melanoma who are most likely to respond to anti-PD1 immune checkpoint blockade (ICB) therapy. While anti-PD1 therapy can be highly effective in advanced melanoma patients, only 30-40% of patients respond well. In this study, we apply single-cell spatial proteomics together with statistical and machine learning (ML) methods to successfully predict advanced melanoma patient response to anti-PD1 ICB in a cohort of 12 patients with >8 million cells. While no single biomarker is sufficient to predict ICB response in our cohort, ML models integrating multiple molecular features accurately predict response in 11 of 12 patients. A recurrent cellular neighborhood analysis revealed a tumor-infiltrating lymphocytes niche that was present in the tumors of most responders. This neighborhood, tumor microenvironment immune cell composition, and levels of nitric oxide synthases were all important features used by our ML models to make accurate predictions. Optimal predictive performance by our ML models, a ROC AUC of 0.76, was achieved when using all molecular features, including cellular spatial relationships, but limiting our analysis to only immune-rich tissue regions.

Introduction

Patients with advanced melanoma (stage III and stage IV) show a 30-40% response rate to anti-PD-1 immune checkpoint blockade (ICB) drugs pembrolizumab and nivolumab [75–78]. Importantly, patients who respond to ICB often show a durable response to therapy

that can last for years. For example, a recent follow-up of the KEYNOTE-006 trial reports that pembrolizumab has improved median melanoma-specific survival in advanced disease to over four years [76]. Unfortunately, there are no accurate biomarkers for predicting long-term response to immunotherapy in melanoma [139–143]. Well-studied biomarkers sometimes used by oncologists to assess the likelihood of melanoma patient response to ICB, such as tumor mutational burden [144] (TMB) and PD-L1 expression [145], have not been shown to be predictive. Given this landscape, there remains a tremendous opportunity to develop new approaches for predicting ICB response in advanced melanoma. Improving the predictive accuracy of ICB response would substantially improve the treatment of this patient cohort. Patients likely to respond to therapy could be given ICB as a first-line therapy, whereas patients unlikely to respond could pursue alternative therapy options earlier [79,146–148] .

New approaches to improve predictions of ICB response in advanced melanoma and other cancers have used a variety of omics approaches. Genomics, transcriptomics, proteomics, and even metabolomics and radiomics have been shown to have predictive power, and research in these areas is ongoing [149] . Recently, single-cell spatial proteomics has proven especially promising for understanding tumor microenvironments (TME) and linking TME features to clinical attributes such as patient response to therapy. Spatial proteomics has been used to make novel insights about the tumor microenvironment (TME), shedding light on cellular heterogeneity, tumor progression, and tumor-immune dynamics across a wide range of cancers [150,151]. Several studies have found that spatial localization of T-cell infiltration and interactions between lymphocytes, macrophages, and PD-L1 within the

tumor compartment can be used to predict melanoma patient response to ICBs [151–156]. Specifically, spatial co-localization of PD-L1+ cells and macrophages with CD8+ T-cells and tumor cells correlated with favorable outcomes to ICB therapy [153,154]. Machine learning has become an important tool for analyzing cancer spatial proteomics and transcriptomics datasets because these datasets are quite complex [157,158].Recent applications of machine learning to TME spatial omics datasets include predicting transition to invasive breast cancer [159] , understanding and predicting tumor response to a novel immunotherapy [160], and tailoring therapy based on predictions of tumor response to therapy [161].

This study builds on spatial proteomics applications in cancer and uses statistical and machine learning (ML) methods to predict response to ICB response in advanced melanoma. We assayed pretreatment advanced melanoma biopsies using two 8-plex multiplex immunofluorescence (mIF) panels consisting of lymphoid or myeloid markers, nitric oxide synthases (iNOS, eNOS, nNOS), immune checkpoint markers (PD-L1, LAG-3), and malignant melanoma tumor cell marker SOX10. Nitric Oxide (NO) dependent processes are associated with either death of melanoma cells or progression of the disease [162]. Therefore, these panels were developed to investigate the dichotomy of NO in melanoma. Our data set includes more than 8 million cells from 12 advanced melanoma tumors from separate patients. We observe no strong univariate biomarkers that predict response, but ML models accurately predict ICB response for 11 of 12 patients in our cohort. A detailed analysis of >1700 1mm^2 tissue tiles reveals that the most accurate predictions are made using compositional and spatial information from immune-rich areas

of the tumors. Our most accurate ML models have a Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) of 0.76. The most important TME features used by the ML models to make accurate predictions were immune and nitric oxide synthases features in immune-rich areas of the tumors. Our results demonstrate how targeted spatial proteomics analysis may play a role in precision medicine for advanced melanoma and future clinical trials.

Approach

A study using single-cell spatial proteomics to assess ICB response in advanced melanoma

Our study cohort consists of twelve stage IV melanoma patients who underwent a surgical tumor resection or biopsy and then received anti-PD1 immune checkpoint blockade (ICB) therapy, either pembrolizumab or nivolumab (Table **4.1**). Four patients experienced a progression-free survival (PFS) of >600 days with RECIST responses of complete or partial response. We term this group the ICB responder group. A non-responder group of eight patients showed tumor progression within at most 161 days (Figure **4.1**a). To assess the single-cell spatial structure of patient tumors, multiplex immunofluorescence (mIF) was performed on each patient resection or biopsy. Two custom mIF panels of eight markers each were used: a lymphoid panel (CD3, CD8, CD20, PD-L1, LAG-3, SOX10, iNOS, nNOS) and a myeloid panel (CD11c, CD14, CD34, MHCII, N-Cadherin, SOX10, iNOS, eNOS) (Figure **4.1**b). Together these panels and markers provide single-cell spatial information on

malignant cells, lymphoid and myeloid immune cells, and cells expressing nitric oxide synthase proteins.

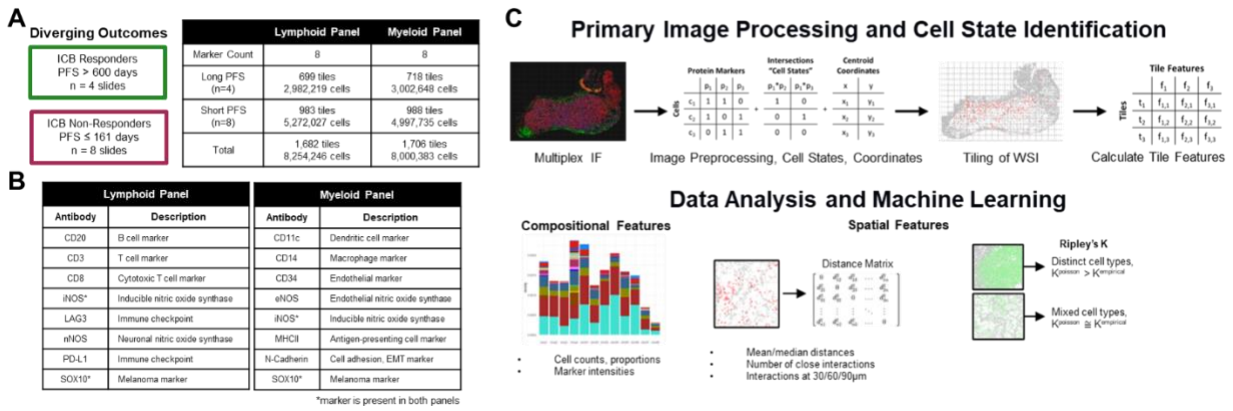


Figure 4.1 Experimental Overview. Our study includes twelve advanced melanoma patients who underwent anti-PD1 immune checkpoint blockade (ICB) after tumor resection/biopsy. (A) Four patients experienced complete or partial response to ICB with >600 days of progression-free survival ("responders") while eight were assessed as progressive disease with progression within 161 days of the start of treatment ("non-responders"). Tumor sections collected prior to treatment underwent multiplex immunofluorescence (mIF) with custom lymphoid and myeloid marker melanoma panels. Whole-slide images were sectioned into 1mm by 1mm tiles for analysis. (B) Protein markers included in each mIF panel. A variety of immune-related markers were probed as well as melanoma marker SOX10, cell adhesion marker N-Cadherin, and nitric oxide synthases iNOS, nNOS, and eNOS. (C) Whole-slide images underwent primary image processing, cell state assignment, and 1mm by 1mm image tiling (see Methods) followed by statistical and machine learning analyses using compositional features (cell state proportions and specific cell subpopulations in a tile) and spatial features (normalized counts of distances between cells within defined radii, Ripley's K spatial statistics).

Images from the mIF panels underwent primary image processing, including cell segmentation and intensity thresholding to produce single-cell annotations for binary protein expression of each of the eight markers (Figure **4.1**c, Methods). Cell states were defined as the 20 most prevalent protein marker combinations among the entire cell population, plus lone expression of any single protein marker. In total 8,254,246 cells were identified across all patient samples and 22 cell states were identified with the lymphoid

mIF panel (8,000,383 cells and 20 cell states with the myeloid mIF panel). Lymphoid and myeloid mIF panels were analyzed separately.

We then tiled the images into smaller sections and applied statistical and machine learning approaches to understand and predict ICB therapy response differences between responders and non-responders in our cohort. To capture spatial information at the local level, we separated each whole slide image into several square tiles of 1mm2 (Figure **4.1**c), totaling approximately 3 million cells across 699 tiles from the responder cohort and 5 million cells across 983 tiles from the non-responder cohort (Supplementary Figure **4.7**). Each tile was characterized using both compositional and spatial features. Compositional features include (a) proportions of cell states within each tile, such as the proportion of CD3+ CD8+ cells in a tile, and (b) proportions of marker positivity within a cell state, such as the proportion of CD8+ cells all CD3+ cells in a tile. Tile spatial features quantified include average distances between cell states, the relative number of interactions between cell states within specific distances (20, 30, 60, 90 μm), and measures of cell state mixing using Ripley's K spatial statistics (variance-stabilized to Ripley's L). Finally, we used statistical and machine learning methods to analyze tile compositional and spatial features to quantify differences between responders and non-responders (Figure **4.1**c). The differences identified by these quantitative approaches provide insight into mechanisms of response for anti-PD-1 ICB therapy in advanced melanoma and suggest approaches for predicting response to ICB therapy.

Results

Characterizing compositional and spatial cellular features among responders and non-responders

To evaluate patient-to-patient variation in cell state composition, we first quantified cell state counts and proportions at the whole-slide level (Figure **4.2**a-d for the lymphoid panel, Supplementary Figure **4.8** for the myeloid panel). Total cell counts and proportions varied widely from slide to slide, and the two biopsies (S7, S8) had very few cells. Among responders, we noted two "immunoreactive" samples (L1, L4) with high PD-L1+ and CD8+ cell proportions and two "immune-cold" ICB responders (L2, L3). All non-responder samples were immune cold as well (Figure **4.2**e-f). No single cell state proportion was statistically significantly different between responders and non-responders, including PD-L1+ cells (Figure **4.2**e-f for lymphoid panel, Supplementary Figure **4.8**,Supplementary Figure **4.9** for myeloid panel, and additional lymphoid details). Visual spatial distributions of top cell states for each mIF panel did not reveal any clear patterns (Supplementary Figure **4.10** for the lymphoid panel and Supplementary Figure **4.11** for the myeloid panel). Tertiary lymphoid structures (TLS) were not observed in any of the 12 patient samples.

Next, we assessed compositional and spatial tile features to capture trends at the local tissue level. Hierarchical clustering of cell state proportions within 1mm2 tiles across all samples showed separation of most ICB responder tiles from most non-responder tiles (Figure **4.3**a, immunoreactive and immune-cold responder tiles cluster to the right), indicating potential for pre-treatment ICB response classification from the combined

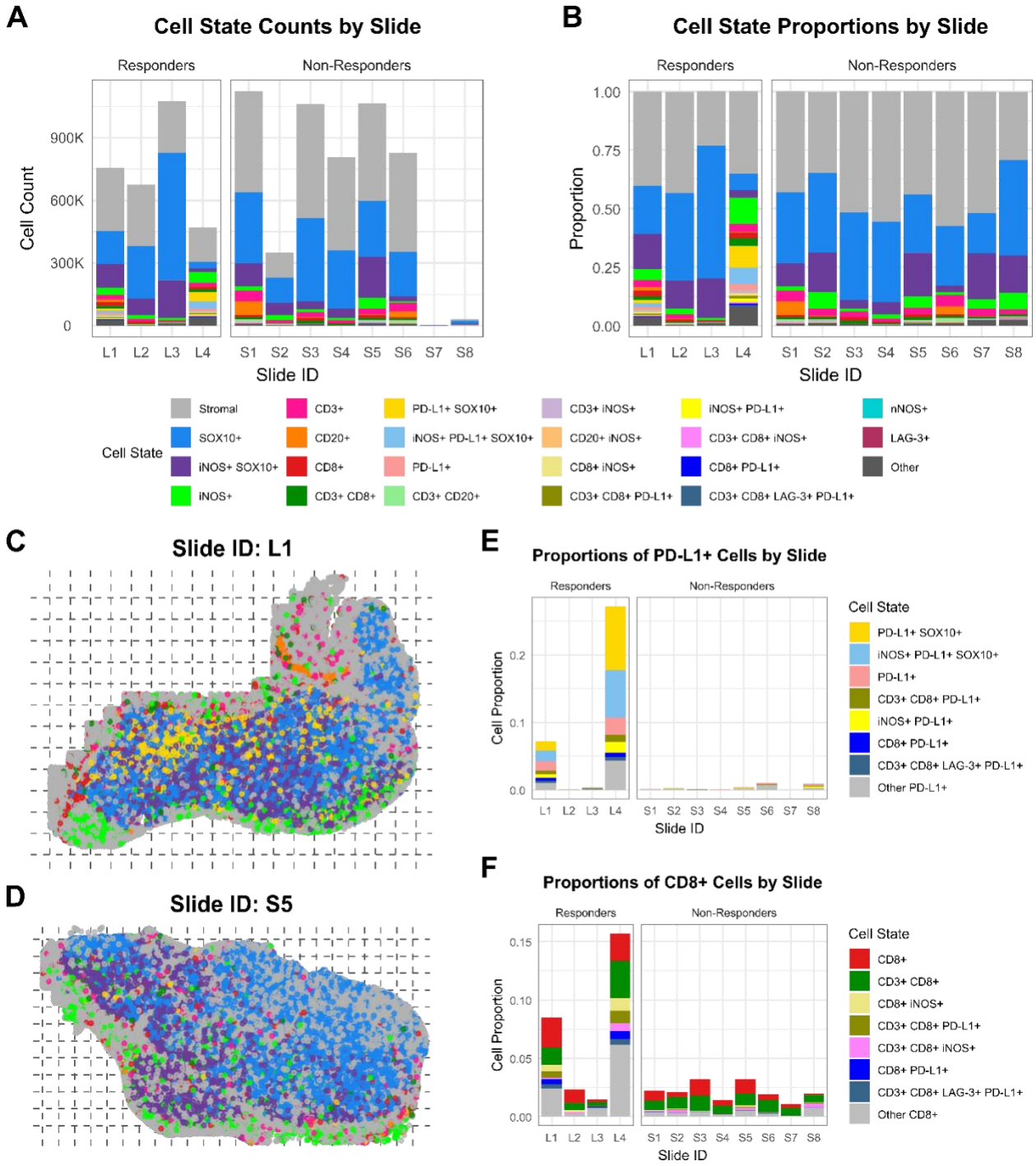


Figure 4.2 Cell state composition by slide, lymphoid mIF panel. (A-B) Stacked bar charts displaying cell counts (A) and proportions (B) of each of 22 cell states (plus an "Other" category) by slide for all used slides in the study. Bars representing ICB responders are displayed on the left (L1-L4) of each chart and bars for non-responders are on the right (S1-S8). Cells lacking any of the eight protein markers are labeled Stromal. (C-D) Whole-slide visualizations of the spatial distribution of the top 8 cell states (Stromal, SOX10+, iNOS+ SOX10+, iNOS+, CD3+, CD20+, CD8+, CD3+ CD8+) and any PD-L1+ cells (yellow) for one responder sample (C, slide L1) and one non-responder sample (D, S5). Cells are plotted by their computed centroids. Dashed lines indicate tile x and y limits. Tiles are 1mm by 1mm in size. (E-F) Stacked bar charts displaying cell state proportions for all cell states containing PD-L1 (E) or CD8 (F) by slide. All cells which contain PD-L1 or

CD8 but are not captured within the 22 defined cell states are represented in gray for “Other PD-L1+” or “Other CD8+”. Elevated proportions of PD-L1+ and CD8+ cells are found in the immunoreactive responder samples (L1/L4) compared to immune-cold responders (L2/L3) and non-responders (S1-S8).

information across our 22 defined cell states of the lymphoid mIF panel. In concordance with the slide-level analysis, we found several cell state proportions that distinguish the two immunoreactive responders’ tiles from the tiles of non-responders and immune-cold responders (Figure **4.3**a). Key distinguishing features include cell state proportions containing cytotoxic T-cell marker CD8 and immune checkpoint marker PD-L1 (Figure **4.3**b). However, these cell state proportions alone fail to distinguish all ICB responders from non-responders because some responders are immunoreactive while some are immune-cold, similar to non-responders. Hierarchical clustering of cell state proportions from the myeloid mIF panel showed poorer separation by tile response class (Supplementary Figure **4.12**a) and revealed several elevated cell state proportions in immunoreactive responder tiles versus immune-cold responders and non-responders, including CD14+, NCad+, and eNOS+ cells (Supplementary Figure **4.12**b). Conversely, MHCII+ cell proportions were elevated in non-responder tiles.

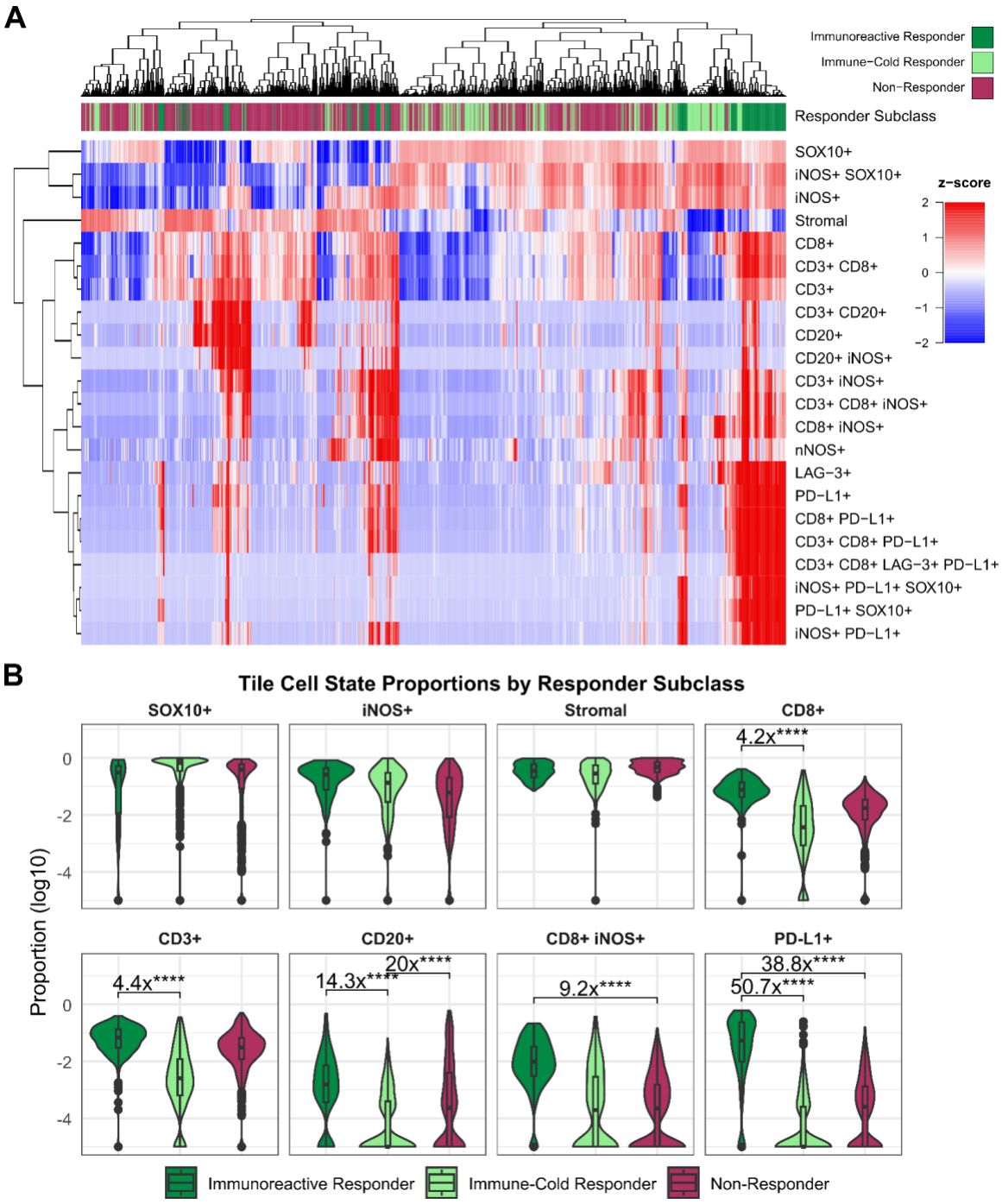


Figure 4.3 Cell state compositional tile features distinguish immunoreactive responders, immune-cold responders, and non-responders to ICB therapy. Lymphoid panel mIF whole slide images from each patient are separated into 1mm by 1mm tiles and proportions of each of 22 defined cell states are calculated within each tile. (A) A heatmap displaying the relative proportions of each cell state (rows) within all 1,682 tiles (columns). Proportions are z-scored across rows to represent the relative expression of each protein (red to blue for high to low expression). The color bar above the heatmap denotes the ICB response subclass of each sample: Immunoreactive Responders (L1/L4), Immune-Cold Responders (L2/L3), and Non-Responders (S1-S8). Rows and columns are clustered by Euclidean distance. Tiles from ICB responders generally cluster separately from those of non-responders, though major differences exist between immunoreactive and immune-cold responder tiles. (B) The tile proportions of SOX10+, iNOS+, Stromal, CD8+, CD3+, CD20+, CD8+ iNOS+, and

PD-L1+ cell states (left to right, top to bottom) are displayed according to ICB response subclass: Immunoreactive Responders (L1/L4), Immune-Cold Responders (L2/L3), and Non-Responders (S1-S8). Significant differences (FDR-corrected Wilcoxon rank-sum p-values) with a fold-change of 4x or greater between response classes are displayed (****p<0.0001). As most markers show differing distributions between immunoreactive and immune-cold responder tiles, no single marker proportion shows the ability to distinguish between all responders (immunoreactive and immune-cold) and non-responders.

We next conducted a correlation analysis between individual protein markers in both responder and non-responder tiles to determine co-occurring or mutually exclusive cell type pairs (Supplementary Figure **4.13**). Within the lymphoid panel data (Supplementary Figure **4.13**a, Supplementary Figure **4.13**b), responders showed increased correlation between tile proportions of CD8+ cells and both PD-L1+ (R=0.5 vs R=0.27) and LAG-3+ cells (R=0.53 vs R=0.31) compared to non-responders (Supplementary Figure **4.13**b, left two plots), suggesting co-localization of CD8+ cytotoxic T cells with immune checkpoint markers PD-L1 and LAG-3 in the pre-treatment TME contributes to successful ICB response. We also note higher positive correlation of B cell marker CD20 with T cell marker CD3 cell proportions in non-responder tiles (R=0.61 vs R=0.44, Supplementary Figure **4.13**b, rightmost plot), indicating a potential immunosuppressive role for B cells in the non-responder population in the absence of TLSs. Correlations from the myeloid panel data (Supplementary Figure **4.13**c, Supplementary Figure **4.13**d) showed higher correlation of cell adhesion marker NCad with monocyte/macrophage marker CD14 (R=0.72 vs R=0.06), nitric oxide synthase eNOS (R=0.21 vs R=-0.04), and dendritic cell marker CD11c (R=0.22 vs R=-0.08) in responder tiles compared to non-responder tiles (Supplementary Figure **4.13**d).

To identify the significant features distinguishing ICB responders and non-responders, we conducted statistical testing using the Wilcoxon rank-sum test, which treats tiles as independent observations, as well as a mixed effects model that accounts for slide-to-slide variation among tiles. While caution is needed in interpreting the results of a Wilcoxon rank-sum test due to violation of the independent samples assumption, this approach allows for descriptive characterization of differences between ICB responders and non-responders at the local tissue level. We then complement these results with a mixed effects model which accounts for slide-to-slide variation to assess potential population-level differences across our limited sample set.

Wilcoxon rank-sum analysis of ICB responder versus non-responder tiles revealed 54 significantly different compositional features (Figure 4.4a). Cell proportions higher in responder tiles as compared to non-responder tiles include iNOS+ PD-L1+ SOX10+ cells (p=2.0e-40, 157x fold change), PD-L1+ SOX10+ cells (p=4.3e-33, 162x fold change), iNOS+ cells (p=9.4e-25, 1.7x fold change), and CD8+ iNOS+ cells (p=3.4e-23, 5.4x fold change) (all Wilcox p-values are adjusted for false discovery rate using the Benjamini Hochberg procedure). Cell proportions higher in non-responder tiles include stromal cells (p=7.5e-29, 1.4x fold change), CD3+ cells (1.7e-18, 1.2x fold change), and CD3+ CD20+ cells (p=1.0e-4, 2.9x fold change). We also found 71 significantly different Ripley's L 20 μm spatial features (Figure 4.4b). In responder tiles, there is increased clustering of stromal cells between other stromal cells (p=1.8e-46, 1.2x fold change), iNOS+ cells (p=2.5e-45, 1.2x fold change), CD8+ cells (p=7.8e-18, 1.1x fold change), and CD3+ cells (p=1.3e-15, 1.1x fold change) compared to non-responder tiles. On the other hand, clustering at 20 μm

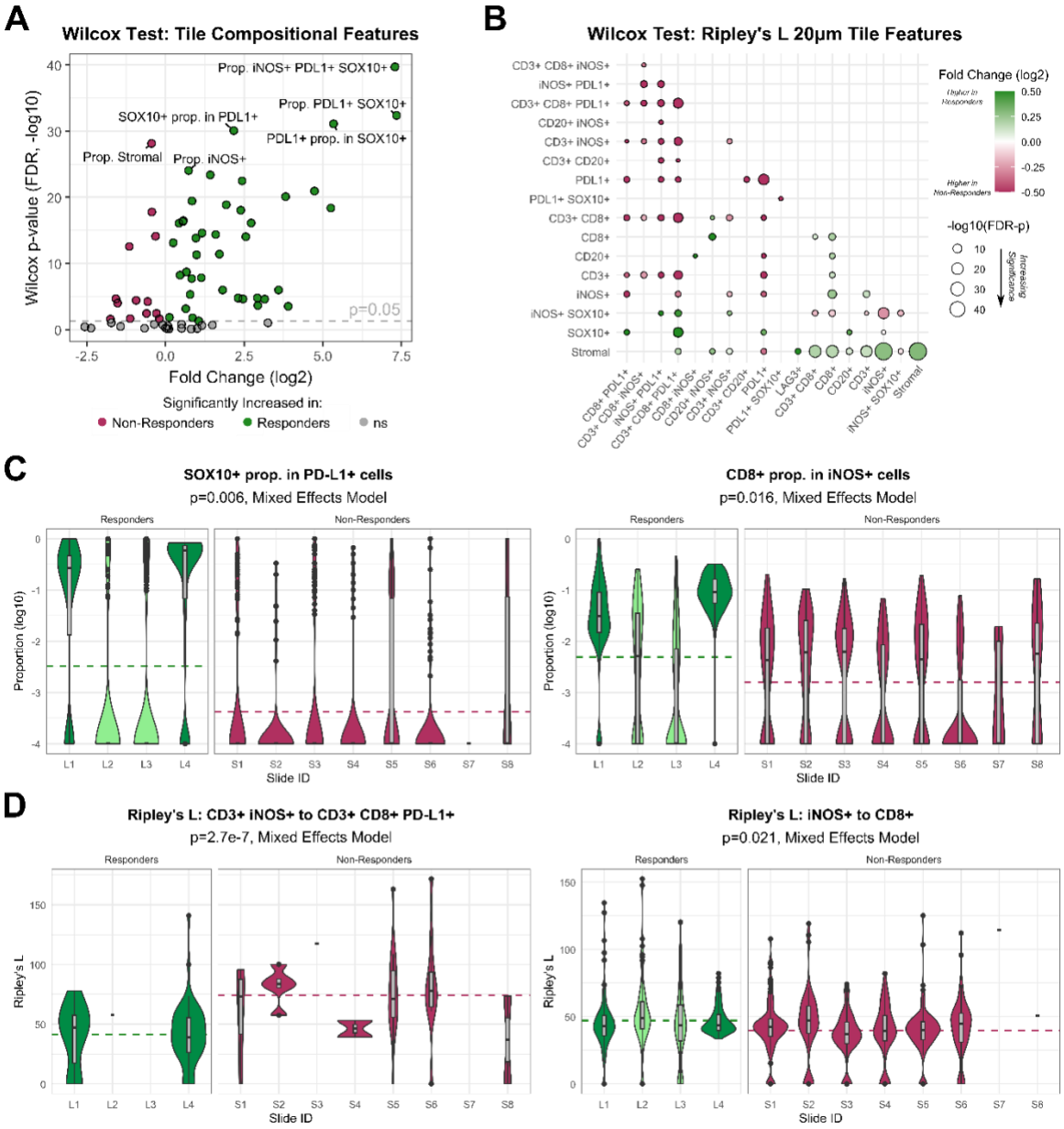


Figure 4.4 Univariate analysis of ICB responders vs non-responders reveals differentially expressed compositional and spatial tile features. Wilcoxon rank-sum test (A, B) and mixed effects modeling (C, D) were performed to identify compositional and spatial tile features from lymphoid panel mIF data in pre-treatment tumor samples capable of distinguishing ICB responders and non-responders. (A) Wilcoxon rank sum test was applied to 72 compositional tile features including cell state proportions within the total tile population and proportions of each marker within subpopulations of each other marker. P-values were corrected for false discovery rate (FDR) using the Benjamini-Hochberg procedure. The volcano plot shows increasing significance on the vertical axis (-log10 of the FDR-adjusted p-values) and effect size on the horizontal axis (log2 of the mean fold-change). Positive fold-change values represent features increased within the ICB responder group. (B) Wilcoxon rank-sum test was applied to 168 Ripley's L spatial tile features calculated at 20 µm distances between cells and p-values were FDR-adjusted as above. The bubble plot shows bubbles of increasing size for increasing significance (-log10 of the FDR-adjusted p-values) colored by effect size (log2 of the mean fold-change, with green and maroon indicating increased values in responder or non-responder tiles, respectively) for cell state pairs indicated on the horizontal and vertical axes. Only significant differences are represented with a bubble (if a bubble is missing in the lower diagonal, p>0.05). (C) To account for slide-to-slide variation within tiles, we also created mixed effects models using slide ID as a random effect to be controlled. Out of 72 compositional features, 8 resulted in p<0.05 though none remained significant after FDR-adjustment. The tile distributions of the top two

significant features are displayed by slide and response group on the horizontal axis and log10(proportion) values on the vertical axis. Dotted lines denote the average value of the feature across all tiles of each response class. Mixed effects model p-values displayed in the subtitles are unadjusted. (D) Same as (C) but for Ripley's L spatial features calculated at 20 μm distances. Out of 62 features, 5 resulted in $p<0.05$ with just Ripley's L values between CD3+ iNOS+ and CD3+ CD8+ PD-L1+ cells remaining significant after FDR-adjustment (p=2.7e-7, FDR-p=1.7e-5). Only Ripley's L features with non-missing values in at least 10 of the 12 slides were included in the analysis.

was increased in non-responder tiles between iNOS+ cells and iNOS+ SOX10+ cells (p=9.2e-15, 1.2x fold change), PD-L1+ cells with each other (p=4.0e-14, 2.0x fold change), and CD3+ CD8+ PD-L1+ cells with each other (p=6.6e-11, 2.4x fold change) and CD3+ CD8+ cells (2.8e-10, 1.6x fold change).

Compositional features most different between responder and non-responder tiles from the myeloid panel include the proportion of MHCII+ cells in CD14+ cells (p=1.2e-55, 19.7x fold change), the proportion of MHCII+ cells (p=3.6e-54, 8.0x fold change), and the proportion of MHCII+ cells in CD11c+ cells (p=5.7e-44, 4.5x fold change), all increased in non-responder tiles (Supplementary Figure **4.14**a). Spatial features of top significance in the myeloid panel also show increased Ripley's L clustering of stromal cells with various other cell types (eNOS+, CD34+ eNOS+, CD14+, iNOS+, CD34+ cells) in responder tiles, and we also observed significantly increased clustering of NCad+ cells with NCad+ SOX10+ cells (p=3.1e-14, 1.6x fold change) in non-responder tiles, among other features (Supplementary Figure **4.14**b).

To account for the effect of slide-to-slide variation, we also calculated mixed effects models using slide ID as a random effect and each tile feature as a fixed effect, one model per feature. The two top compositional features distinguishing responders (n=4) from non-responders (n=8) in the mixed effects model include the proportion of SOX10+ cells within

the PD-L1+ subpopulation (p=0.006) as well as the proportion of CD8+ cells within the iNOS+ subpopulation (p=0.016), though neither comparison remained significant after correction for false-discovery rate (FDR) among the 72 examined compositional features (Figure 4.4c). Among Ripley's L spatial features, we found significantly increased clustering within 20 μm between CD3+ iNOS+ cells and CD3+ CD8+ PD-L1+ cells in non-responders compared to non-responders (p=2.7e-7, FDR-p=1.7e-5), with no other features remaining significant after FDR-adjustment (Figure 4.4d). Although not significant after adjustment, we observed trends of increased clustering between iNOS+ and CD8+ cells in responders compared to non-responders. Similar mixed effects modeling of features from the myeloid panel did not reveal and significant differences between response groups, though we observed trends of increased proportions of CD14+ CD34+ cells and CD34+ proportion in SOX10+ cells within responder tiles (Supplementary Figure **4.14**c) and trends of increased stromal to eNOS+ cell clustering in responder tiles and increased CD34+ to CD14+ SOX10+ cell clustering in non-responder tiles (Supplementary Figure **4.14**d).

Recurrent cellular neighborhood analysis reveals tumor-infiltrating lymphocytes niche in ICB responders

To quantify local tissue spatial architectures, we conducted a recurrent cellular neighborhood (RCN) analysis across all tumor samples. Neighborhoods were defined as the collection of all cells with centroids within 60 μm of a central index cell, in concordance

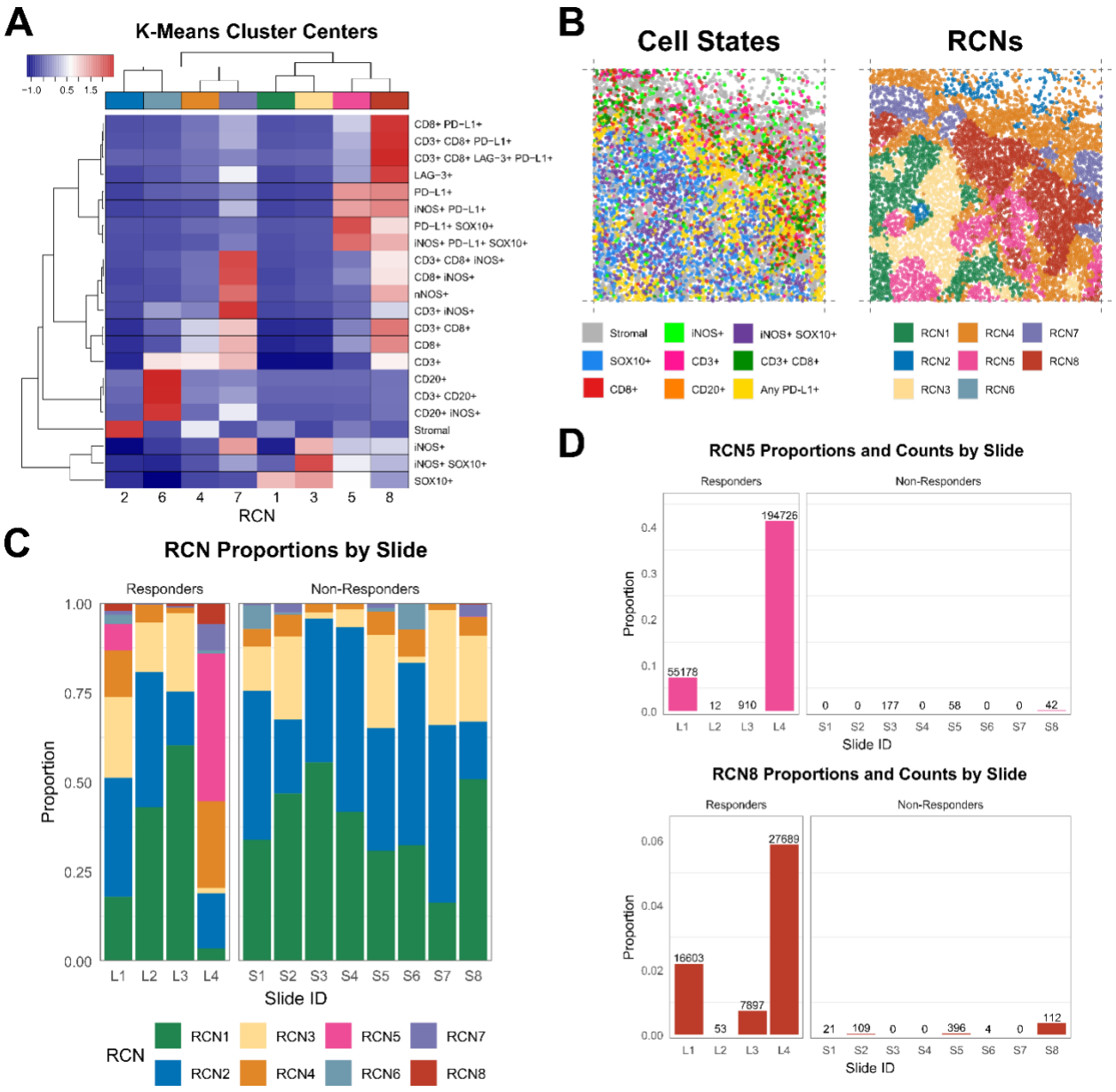


Figure 4.5 Recurrent cellular neighborhoods analysis. Recurrent cellular neighborhoods (RCNs) were determined via k-means clustering of all lymphoid mIF panel cellular neighborhoods across 12 slides. (A) K-means clustering resulted in eight distinct cluster centers which define representative cell state proportion compositions for each RCN. Each column represents an RCN while rows represent cellular neighborhood cell state proportions. Red to blue indicates relatively high to low expression of the corresponding cell state proportion within each RCN. Rows and columns are clustered using Euclidean distance. RCNs are labeled according to size, with RCN1 representing the largest cluster (3,172,458 neighborhoods) down to RCN8, the smallest (52,884 neighborhoods). (B) A single tile from slide L1 shown with cell states (left) and computed RCN labels (right). (C) Stacked bar chart showing the proportion of cellular neighborhoods within each slide corresponding to each RCN label. Bars representing ICB responders are displayed on the left (L1-L4) and bars for non-responders are on the right (S1-S8). (D) Bar charts showing the proportion of RCN5- and RCN8-labeled cellular neighborhoods within each slide with labels for total count of neighborhoods displayed above.

with similar studies examining tumor immune microenvironment cell-cell interactions [160,163,164]. We defined neighborhoods for each of the ~8M total cells in our dataset, calculated proportions of each cell state within each neighborhood, then conducted k-means clustering to identify recurrent patterns (see Methods). Our analysis of the lymphoid mIF panel yielded eight distinct recurrent cellular neighborhoods (Figure **4.5**a, Figure **4.5**b) that were present in varying degrees across all twelve patients (Figure **4.5**c, Supplementary Figure **4.15**). In concordance with the PD-L1+ cell state proportions in Figure **4.2**b, there was strong enrichment of the PD-L1-enriched cluster RCN5 in the immunoreactive responders (L1/L4) but very limited representation of this RCN among immune-cold responders (L2/L3) and non-responders (Figure **4.5**d). Interestingly, RCN8 proportions among all neighborhoods within each patient corresponded well with ICB response status: three ICB responders showed the highest RCN8 proportions (Figure **4.5**d). RCN8 is enriched for the cytotoxic T-cell marker CD-8, the immune checkpoint markers LAG-3 and PD-L1, the malignant melanoma marker SOX10, and the enzyme iNOS, representing potential regions of tumor-infiltrating lymphocytes (TILs). RCN analysis of the myeloid mIF panel data created clusters of distinct combinations of cell types (Supplementary Figure **4.16**a) but did not show differential enrichment of RCNs by patient response class (Supplementary Figure **4.16**b, Supplementary Figure **4.16**c).

Machine learning models accurately predict patient ICB response

To build accurate predictive models of patient ICB response that use all available features, we developed and evaluated 13 machine learning (ML) models. Three sets of features were used to create three distinct kinds of ML models: 1) “Compositional” models that use only compositional features, proportions and subpopulation proportions, for predictions; 2) “Combined” models that use compositional features and spatial features, RCN proportions and Ripley’s L between cell states at 20 μm, for predictions; and 3) “Immune-High” models that, due to our findings on the importance of “TILs RCN” in the previous section, uses compositional and spatial features from a subset of tiles with high proportions (≥ 0.02) of CD8+ cells for making predictions.

Supplementary Figure **4.17** provides performance information from our assessment of different model architectures, spatial feature subset addition, and immune-high tile filtering which informed our final model design. Based on this analysis, LightGBM [165] models performed the best, and we used this approach for all final evaluations.

We evaluated our models using Leave-One-Group-Out Cross-Validation (LOGO-CV), leaving one patient (group) out for each fold of cross validation over 100 iterations, resulting in 1,200 training-validation splits per model type. Performance metrics were computed for each LOGO-CV iteration by aggregating predictions from every left-out patient set (see Methods). SHAP [166] was used to estimate the importance of features in making predictions for each ML model.

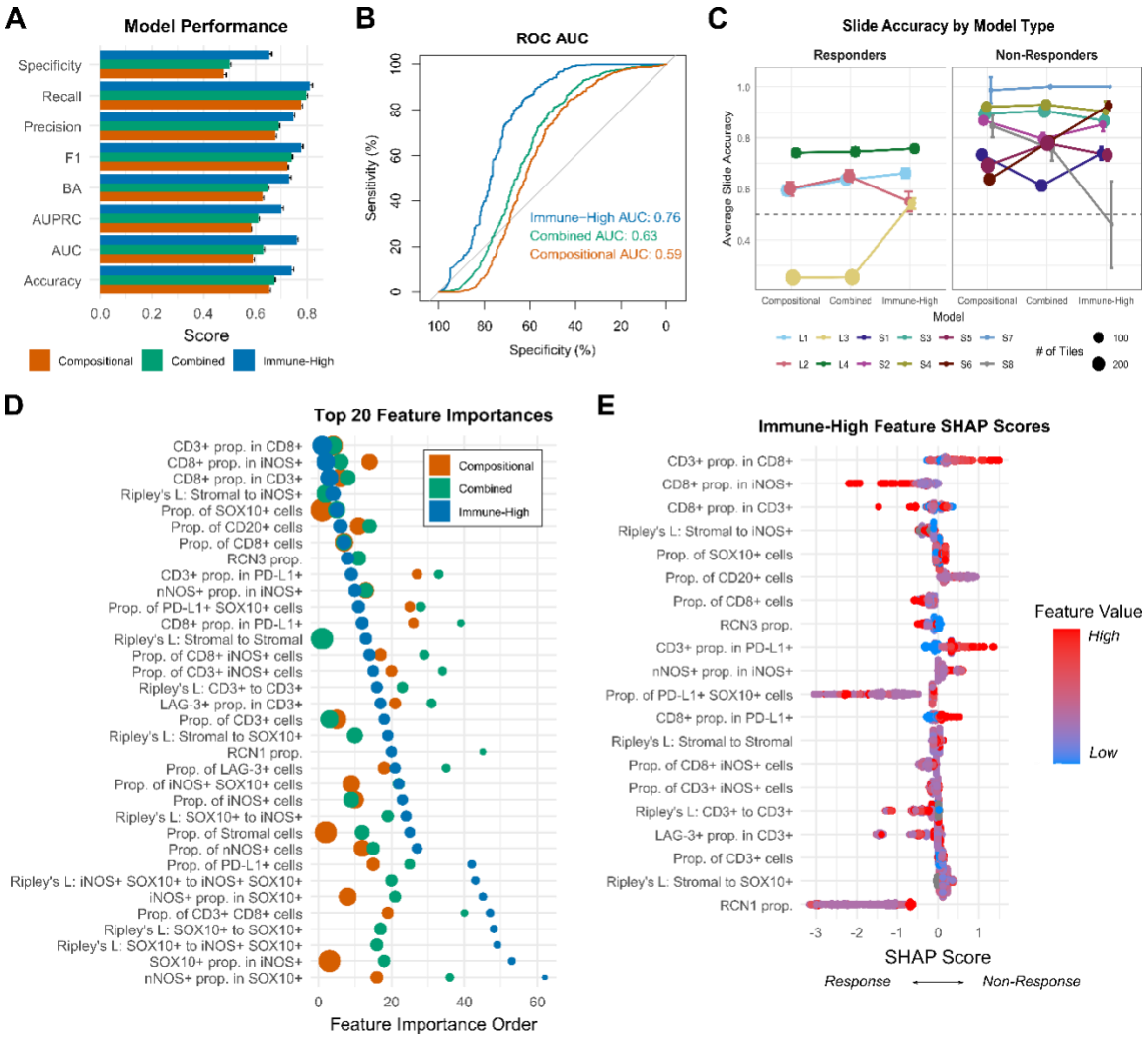


Figure 4.6 Models trained on immune-high tiles show improved classification performance and increased importance of immune-related features. Machine learning model performance metrics calculated from 100 iterations of 12-fold leave-one-out cross validation (LOOCV, leaving one of 12 patients out each time) for the 1) compositional feature only all-tile models ("Compositional"), 2) the combined compositional and spatial feature all-tile models ("Combined"), and 3) the combined compositional and spatial feature 2% CD8-filtered tiles model ("Immune-High") with data from the lymphoid mIF panel. (A) Across all performance metrics, the Immune-High model type performs best followed by the Combined then Compositional models. Bars and error bars represent the average and standard deviation of each metric across 100 iterations of 12x LOOCV. (B) The receiver operating characteristic (ROC) curve shows model tradeoffs between sensitivity (true positive rate) and specificity (true negative rate). Area under the curve (AUC) represents the model's ability to distinguish between classes, with higher values corresponding to better performance. The Immune-High model shows the highest AUC (AUC=0.76), followed by the Combined (AUC=0.63) and the Compositional models (AUC=0.59). (C) Tile accuracy by slide for each model type. Points represent average accuracy and error bars represent standard deviation across all 100 iterations of each leave-one-out model. Point size represents the relative number of tiles used for the model test set, with smaller tile sets for the immune-high model (includes only tiles with ≥ 2% CD8+ cells). The dashed gray line indicates an accuracy of 50%. If slide prediction of patient ICB response is determined by the majority class prediction of its constituent tiles, the immune-high model is the only model which correctly predicts all tumor resection samples (L1-L4, S1-S6), though it usually incorrectly classifies S8, a core biopsy from a non-responder. (D) The order of average feature importance values for each model type (Compositional, Combined, and Immune-High). Data points are colored by model type and sized by their relative average feature importance value. A feature importance order of 1 indicates the feature had the highest feature importance value within that model type. All features within the top 20 average feature importance for any model type are shown. (E) Shapley Additive exPlanation (SHAP) scores from a

representative Immune-High model iteration of 12x LOOCV show the relative effect (SHAP score, x-axis) of each feature (y-axis) alongside its relative feature value (red to blue indicates high to low feature values) for each predicted tile (data points). Negative to positive SHAP scores indicate that the associated feature value pushed the model towards predicting ICB response or non-response respectively, while SHAP scores close to zero indicate little to no effect on prediction.

The Combined model outperformed the Compositional model with consistent and moderately better performance in all metrics (Figure 4.6a). These results demonstrate the value of incorporating spatial features with compositional features as Combined model performance was 0.01 to 0.04 higher than the Compositional model, with the receiver operating characteristic curve (ROC AUC) increasing from 0.59 to 0.63 (Figure 4.6b). While the improvements in model performance are modest, the addition of spatial features resulted in significantly increased patient-specific tile prediction accuracy for 2 of 4 responders and 5 of 8 non-responders ($p<0.05$, Wilcoxon rank-sum, Figure 4.6). Notable compositional features appearing in the top 10 average feature importances for both Compositional and Combined model types included the proportions of stromal cells, malignant cells, nitric oxide synthase iNOS+ cells, T cells, and cytotoxic T cells (Figure 4.6d-e). Important spatial features included Ripley's L statistics for the self-clustering of stromal cells; the clustering of stromal cells with iNOS+, SOX10+ (melanoma), and CD3+ cells; the self-clustering of iNOS+ melanoma cells; and the self-clustering of melanoma cells (Figure 4.6d, Figure 4.6e). These findings demonstrate the importance of both compositional and spatial information among immune cells, iNOS+ cells, and stromal cells in the tumor microenvironment for making accurate predictions of patient response.

Next, we compared the performance of our Immune-High model that was built using a subset of tiles with the proportion of CD8+ cells ≥ 0.02 with our Combined model that used all tiles. The Immune-High model performed substantially better than the Combined model

(Figure 4.6a, Figure 4.6b), including a 0.13 ROC AUC increase to 0.76 compared to 0.63 for the Combined model as well as a 0.15 improvement in specificity to 0.65 as compared to 0.50 for the Combined model.

We made patient-level response predictions by using the majority prediction for a patient's tiles. For the Immune-High model, only tiles meeting the CD8+ cells proportion threshold were used. Patient-level prediction accuracy was significantly improved in the Immune-High vs Combined model for 3 of 4 responders and 3 of 8 non-responders. Accuracy substantially increased for samples S1 and L3 and decreased for only one core biopsy sample S8 (fold change > 1.2x, $p<0.001$). Accuracy was minimally changed for 7 patients (fold change < 1.1x), including the other core biopsy S7 (Figure 4.6c). Importantly, "Immune-High" was the only model capable of attaining ≥50% tile accuracy for all ICB responders as well as all but one non-responder. The incorrectly predicted non-responder was a core biopsy with only six immune-high tiles (S8). The "Immune-High" model revealed increased importance of cells with co-expression of T cell markers CD3 and CD8, increased importance of the immune checkpoint marker PD-L1, and decreased importance of stromal cell information relative to the Combined and Compositional models (Figure 4.6d,Figure 4.6e). Together, these results suggest that immune-active regions contain especially important compositional and spatial information for determining a patient's likelihood of response to ICB.

Lastly, we trained and evaluated ICB response prediction models using the myeloid mIF panel data (Supplementary Figure **4.18**). Myeloid panel models performed worse than lymphoid panel models, especially when using compositional features alone. Similar to our findings with the lymphoid models, the addition of spatial information improved all measured performance metrics (Supplementary Figure **4.18**a, Supplementary Figure **4.18**b) and most tile accuracy rates by slide (Supplementary Figure **4.18**c). Top features included stromal cell clustering among themselves and with tumor cells (SOX10+) as well as proportions of CD34+ cells, macrophage (CD14+), iNOS+ cells, and antigen-presenting cells (MHCII+) (Supplementary Figure **4.18**d-e).

Discussion

There is an acute clinical need to better predict patient response to immune checkpoint blockade (ICB) in advanced melanoma and understand the biological mechanisms associated with ICB response. In this study of pre-anti-PD1 therapy stage IV melanoma patients, we generated a spatial single-cell targeted proteomics dataset using a novel multiplex immunofluorescence (mIF) panel. Statistical analyses of this dataset revealed novel tumor microenvironment (TME) features associated with anti-PD1 response, and our machine learning (ML) models correctly predicted patient response in 11 of 12 patients. Our study yielded three key findings. First, machine learning models that combined multiple TME features to predict patient response outperformed predictions made using all single TME features, including one current standard of care biomarker, PD-L1 positivity

[167–169]. Second, our analysis demonstrates the importance of spatial analysis in classifying ICB response. A recurrent cellular neighborhood resembling a tumor-infiltrating lymphocytes (TILs) niche was elevated in three of four ICB responder patients, and several spatial features showed significant differences in ICB responders and non-responders. Third, our machine learning models were much more accurate in predicting ICB response when trained only on immune-rich regions. Tile-level classification accuracy jumped from 67% to 74%, and ROC AUC rose from 0.63 to 0.76. Machine learning models trained on immune-rich regions accurately predicted response in all four patients that responded to ICB therapy and in seven of eight patients that did not respond to therapy. Therefore, this model is potentially useful for predicting both responses and failures and additional cohorts should be analyzed prior to application in the biomarker directed clinical trial setting.

While current FDA-approved biomarkers for ICB response in melanoma do not consider spatial tumor organization (PD-L1 expression, tumor mutational burden), our work and other recent studies show the potential of incorporating spatially derived biomarkers for improving prediction of response. Previous studies have identified spatial relationships between T cell and macrophage subtypes with various immune checkpoint markers and tumor cell markers [151,153–156]. In our study, the TIL-like recurrent cellular neighborhood elevated in three of four ICB responder patients was enriched for cytotoxic T cell marker CD8, immune checkpoint markers LAG-3 and PD-L1, malignant melanoma tumor cell marker SOX10, and inducible nitric oxide synthase iNOS. Similarly, correlation of tile cell state proportions suggested increased co-localization of cytotoxic T cells with immune

checkpoint markers LAG-3 and PD-L1 in responders compared to non-responders. These results are concordant with several studies highlighting the potential of using tumor-infiltrating lymphocytes as a predictive biomarker of ICB response in melanoma[170,171] . We also found significantly increased clustering of SOX10+ cells with PD-L1+ T cells in ICB responders, mirroring a recent study showing close interactions between tumor cells and PD-L1+ T cells are associated with favorable ICB response [154]. Lastly, our ML models performed substantially better when including spatial features (0.59 to 0.63 AUC in lymphoid, 0.52 to 0.60 AUC in myeloid), with the greatest performance improvement observed for spatial features computed at a close distance (20 μm).

There are no current standard-of-care biomarkers for ICB response in melanoma. PD-L1 and TMB are insufficient because there are many patients who respond to ICB despite low PD-L1 and/or TMB levels [144,148,172,173]. The shortcomings of single-biomarker predictions of ICB response likely stem from high tumoral heterogeneity typical of advanced melanoma [167–169]. In our cohort, responders included both those with “immunoreactive” TMEs characterized by high PD-L1 and CD8 expression but also those with “immune-cold” TMEs that more closely resembled non-responders. Our analysis shows that the integration of multiple features leads to more accurate prediction of ICB response prediction. Unsupervised hierarchical clustering of lymphoid mIF panel TME features correctly grouped most tiles from ICB responders to a single cluster while any single feature proved insufficient for grouping. In addition, our ICB response ML classifier integrated and weighted TME features to obtain 92% patient accuracy, with optimal performance achieved when training the classifier only on immune-rich tissue regions. This

improvement suggests similar models or biomarkers may benefit from limiting analysis to immune-rich regions of the TME, potentially removing noise from less informative regions.

Nitric oxide and nitric oxide synthase have been implicated in both pro- and anti-tumor effects based on concentration and the cell type of expression [162,174–178]. Within our cohort, iNOS expression in cytotoxic T cells was associated with ICB response as indicated by significantly increased cell type proportions by our univariate analysis as well as high feature importance within our ICB response ML model. We also found increased levels of iNOS+ tumor cells were associated with response. Spatially, an iNOS+ malignant tumor recurrent cellular neighborhood (RCN3) emerged within the top 10 most important features, with higher proportions associated with anti-PD-1 response. Together, these findings support evidence of iNOS as a potential pre-treatment biomarker of anti-PD-1 response in melanoma, particularly when expressed within T cells. It is noted that there is extensive literature of iNOS being a poor prognostic factor for melanoma patients due to its effects on the TME as reviewed elsewhere [162], though initial studies were done in the era prior to checkpoint blockade [179]. Further research is needed to fully elucidate its dichotomous role in tumor progression and/or immune activation both before and during immunotherapy.

A notable limitation of our study is the small cohort size of ICB responder patients (n=4). This limitation is magnified by the high level of variability of several key features within the responder cohort (the dichotomy between immunoreactive and immune-cold responders, particularly in PD-L1 and CD8 expression), variations in tumor collection site (skin, intestine, lung, and lymph node), and the use of resections plus core-needle biopsies.

However, the wealth of data and our tiling approach to focus on local tissue microenvironments, 14 unique protein markers, several hundred spatial interaction features, more than 8 million cells, and ~1700 tiles, allow for meaningful analysis despite our small patient cohort. Future work should expand on our exploratory study to validate these findings with larger patient populations. This work is also limited by the relatively small number of protein markers within a single mIF panel, preventing detailed immune cell subtyping and quantification of spatial interactions between both lymphoid and myeloid cell types.

In summary, our study demonstrates how single-cell spatial proteomics, together with machine learning, can accurately predict advanced melanoma patient response to anti-PD-1 therapy. Both spatial information and immune-rich tumor microenvironment regions proved to be important for machine learning models to make accurate predictions. This study provides the foundation for expanded analyses and clinical trials that use single-cell spatial proteomics and machine learning to better predict melanoma patient response to ICB therapy.

Methods

Sample Information (Dr Markowitz)

Tumor samples were collected from Stage IV advanced melanoma patients at Moffitt Cancer Center 1-12 months prior to the start of anti-PD1 immunotherapy under IRB protocol MCC18583. Descriptive information about the patient cohort appears in Table **4.1**.

Samples were preserved in formalin-fixed paraffin-embedded (FFPE) tissue blocks and sectioned for multiplex immunofluorescence imaging.

Multiplex Immunofluorescence (Dr Markowitz)

Formalin-fixed and paraffin-embedded (FFPE) tissue samples were immunostained using the AKOYA Biosciences OPAL™ 7-Color Automation IHC kit (Waltham, MA) on the BOND RX autostainer (Leica Biosystems, Vista, CA). The OPAL™ 7-color kit uses tyramide signal amplification (TSA)-conjugated to individual fluorophores to detect various targets within the multiplex assay. Sections were baked at 65°C for three hours then transferred to the BOND RX (Leica Biosystems). All subsequent steps (ex., deparaffinization, antigen retrieval) were performed using an automated OPAL™ IHC procedure (AKOYA). OPAL™ staining of each antigen occurred as follows: heat induced epitope retrieval (HIER) was achieved with Citrate pH 6.0 buffer for 20min at 95°C before the slides were blocked with AKOYA blocking buffer for 10 min. The slides were incubated with primary antibodies at room temperature (RT) for 60 min followed by the OPAL™ HRP polymer and one of the OPAL™ fluorophores during the final TSA step. Individual antibody complexes are stripped after each round of antigen detection.

We developed two panels with the following antibodies (summarized in Supplementary Table **4.2** and Supplementary Table **4.3**):

- Lymphoid Panel: iNOS (Thermo Fisher, 4E5, 1:50, dye 570), CD20 (DAKO, L26, HIER-EDTA pH 9.0, 1:150, dye480), PD-L1 (CST, E1L3N, HIER- EDTA pH 9.0, 1:150, dye540), LAG-3 (CST, D2G40, HIER-EDTA pH 9.0, 1:300, dye 690), CD8 (DAKO, C8/

144B, HIER- EDTA pH 9.0, 1:100, dye520), nNOS (CST, C7D7, HIER-EDTA pH 9.0, 1:300, dye 480), SOX10 (Biocare, BC34, HIER- EDTA pH 9.0, 1:50, dye650) and CD3 (DAKO, Rb poly, HIER- EDTA pH 9.0, dye780).

- Myeloid Panel: iNOS (Thermo Fisher, 4E5, HIER- EDTA pH 9.0, 1:50, dye570), eNOS (CST, D8A6N, HIER- EDTA pH 9.0, 1:50, dye620), N-Cadherin (CST, D4R1H, HIER-EDTA pH 9.0, 1:50, dye480), CD14 (Abcam, LPSR/2386, HIER- EDTA pH 9.0, 1:75, dye540), CD34 (Abcam, EP373Y, HIER- EDTA pH 9.0, 1:250, dye520), MHCII (Dako, M0775, HIER- EDTA pH 9.0, 1:150, dye690), SOX10 (Biocare, BC34, HIER- EDTA pH 9.0, 1:50, dye650) and CD11c (Abcam, EPR4421, HIER- Citrate pH 6.5, 1:150, dye480).

After the final stripping step, DAPI counterstain is applied to the multiplexed slide and is removed from BOND RX for coverslipping with ProLong Diamond Antifade Mountant (Thermo Fisher Scientific). All slides were imaged with the PhenoImager HT Imaging System.

Quantitative Image Analysis (Jonathon and Carlos)

Multi-layer TIFF images were exported from InForm (AKOYA) and loaded into HALO (Indica Labs, New Mexico) for quantitative image analysis. The tissue was first segmented into individual cells using the DAPI cell nuclei stain using a proprietary algorithm with the HALO image analysis platform. For each marker, a positivity threshold within the nucleus or cytoplasm was determined per marker based on published staining patterns and

intensities for each specific antibody. The per-cell analysis was then exported to provide the marker status (positive or negative) and fluorescent intensity of every individual cell within each image. The resolution of each image was 0.4992 µm per pixel.

Slide Tiling

To capture spatial information at the local tissue level, tiling was performed on all slides to define regions of 1mm by 1mm with a minimum of 100 cells per tile. Cells were assigned to each tile based on the location of their centroid, xcentroid = (xmin + xmax) / 2 and ycentroid = (ymin + ymax) / 2.

Tile Features

Compositional Features

Cell State Proportions

Cell states are defined as the top 20 occurring combinations of protein markers across all cells from all slides, plus individual markers outside of the top 20 combinations. For each cell state, proportions of all cells expressing the respective combination of markers (inclusive of cells expressing those plus other markers) are calculated by dividing cell state counts by the total number of cells within the tile.

Subpopulation marker proportions

Marker proportions are further analyzed as percentages within defined subpopulations. For each tile, subpopulations are defined as all cells which display a specific marker (ex. all SOX10+ cells, all iNOS+ cells). Within each subpopulation, the proportion of that cell population with every other individual marker is calculated (ex. the proportion of SOX10+ cells that are PD-L1+).

Recurrent cellular neighborhood (RCN) analysis

Within each tile, cellular neighborhoods are calculated for every cell. Each cell serves as a "seed cell" and its neighborhood is defined as the population of cells within a 60 μm distance. The neighborhood is then quantified as the proportion of each cell state (as defined above in "Cell State Proportions") within the total cell count of that neighborhood. Using the matrix of all 8M+ cellular neighborhoods and the 20+ cell state proportion features, we then define recurrent cellular neighborhoods (RCNs) through k-means clustering using the stats package in R. First, we identify the optimal number of clusters (k) via the "elbow method", increasing k until the within-cluster sum of squares does not decrease. Then, we conduct k-means clustering and assign each neighborhood to an RCN cluster (k=8 RCNs for the lymphoid panel, k=9 RCNs for the myeloid panel). Proportions of each RCN are calculated for each tile to be used as compositional tile features.

Spatial Features

Interaction Features

Interaction features are calculated to reflect the spatial relationships between cell states across multiple distances. For each pair of cell states within a tile, a distance matrix is calculated containing all the Euclidean distances between each cell of cell state A to every cell of cell state B. To calculate interaction scores, we count the total number of cell state A to cell state B distances less than 20, 30, 60, or 90 μm then normalize by the sum of the proportions of cell state A and cell state B.

Ripley's L Features

Ripley's L features reflect whether cell state pairs show discrete or mixed spatial patterns. Ripley's K function describes the expected number of cells of cell state A within a specified radius from a typical cell of cell state B. The R package spatstat [180] is used to calculate Ripley's K distributions for all cell state pairs within a tile while accounting for edge effects with isotropic border correction. These distributions are then normalized for increased variance at higher radius values by taking the square root of Ripley's K divided by pi, defined as Ripley's L. Empirically-derived Ripley's L values (Liso) are then compared with theoretical values for randomly-mixed, Poisson-distributed cell state populations (Ltheo) to determine whether the cell state pair is uniformly mixed (Liso ≅ Ltheo), clustered (Liso > Ltheo), or distinct (Liso < Ltheo). Extracted tile features include: 1) Ripley's L values (Liso) at 20, 30, 60, and 90 μm, 2) "delta L", Liso - Ltheo at 20, 30, 60, and 90 μm representing the degree of spatial clustering (positive) or repulsion (negative) at each distance, and 3) the

“cluster score”, integrating delta L from 0 to 100 μm to represent the average clustering/repulsion behavior.

Univariate statistics

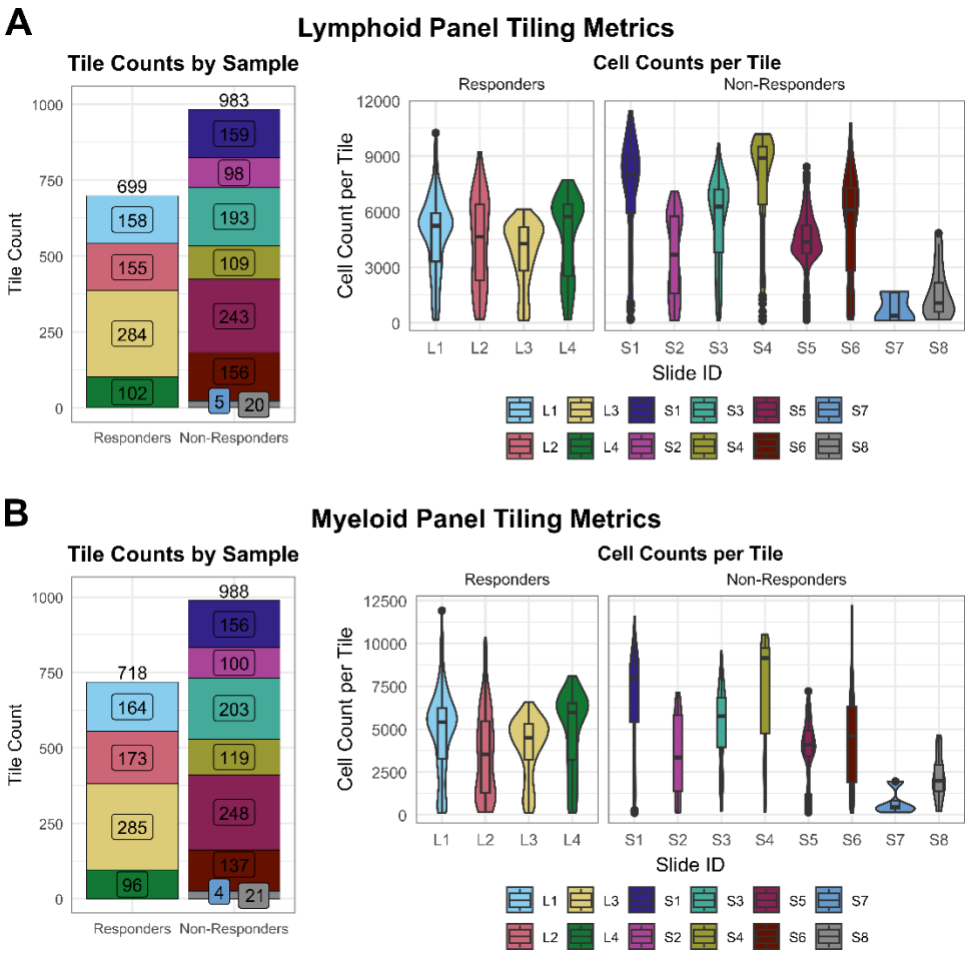
Data analysis and visualization was performed using the R programming language with functions from the stats package and the tidyverse [181] collection of packages. Wilcoxon rank-sum tests (also known as the Mann-Whitney U test) are performed for each feature comparing long PFS tiles and short PFS tiles. P-values are then corrected for false discovery rate using the Benjamini Hochberg procedure.

Machine learning models

Binary classification machine learning models were created using PyCaret in Python to predict tissue sections from long progression free survival patients (ICB responders) versus tissue sections from short progression free survival patients (ICB non-responders). Feature selection was performed to optimize classification performance out of the XXX total features defined as outlined above. More info on feature selection. We used a leave-one-patient-out cross-validation approach, creating 12 separate models each with all tiles from one patient left out of the training set and used for testing. This prevents the models from learning patient-specific patterns and artificially improving model accuracy. We aggregated test set prediction metrics across all cross-validation models to report an overall confusion matrix and summary statistics of accuracy, area under the receiver operating characteristic

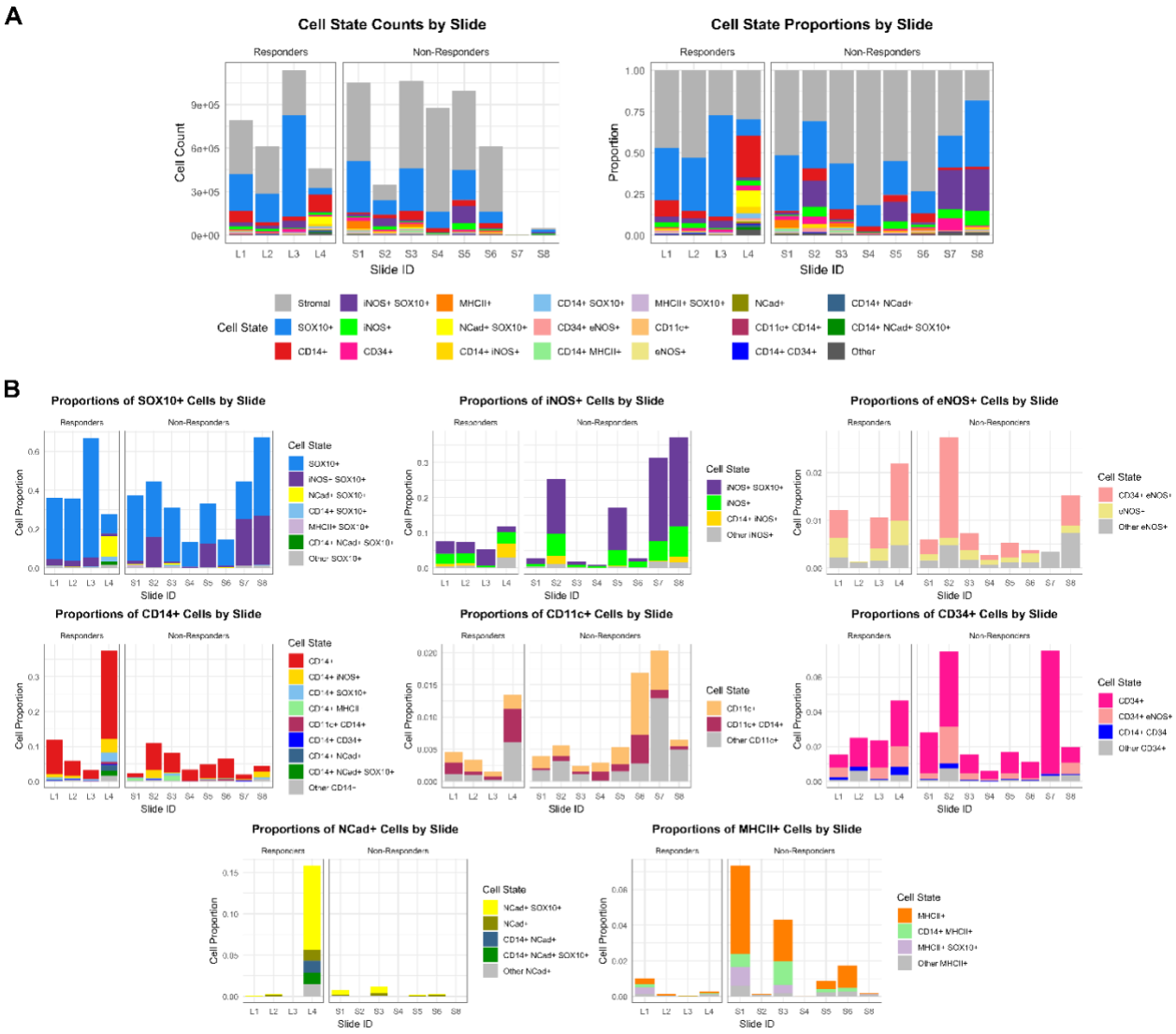
curve (ROC AUC), precision, and recall and f1. We used a stratified k fold of 10 to evaluate the model. Every machine learning experiment was run at least n >= 30 to ensure statistical significance.

Figures

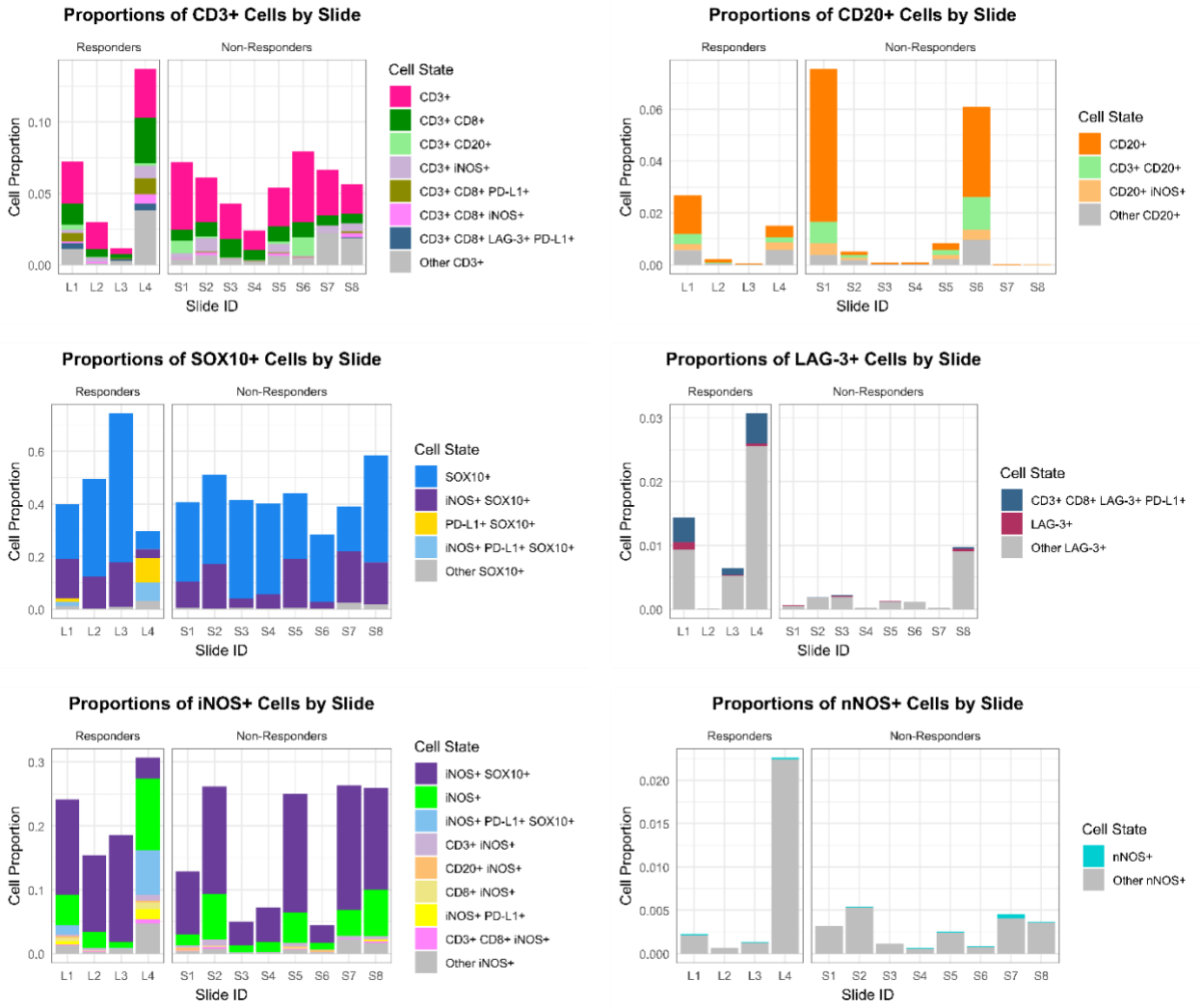


Supplementary Figure 4.7 Tile counts by sample and cell counts per tile. Whole slide images (WSIs) of surgically resected tumors (L1-L4, S1-S6) or tumor biopsies (S7-S8) were separated into 1mm-by-1mm square tiles with a minimum of 100 cells per tile. (A) Lymphoid mIF panel WSIs: the number of tiles per resected tumor ranged from 98 to 284, with biopsies contributing 5 (S7) and 20 (S8) tiles (bar chart on the left). In total, the dataset consists of 1,682 tiles: 699 tiles from ICB responder samples and 983 tiles from non-responder samples. Sections of the bar chart are colored by sample corresponding to the legend at the bottom of panel B. Cell counts per tile ranged from a minimum of 100 to over 10,000 (violin/boxplot on the right). Most resected tumor samples have a median of between 4,000 and 6,000 cells per tile, with biopsies showing fewer cells per tile. (B) Myeloid panel mIF WSIs: very similar numbers of tiles to lymphoid panel, with

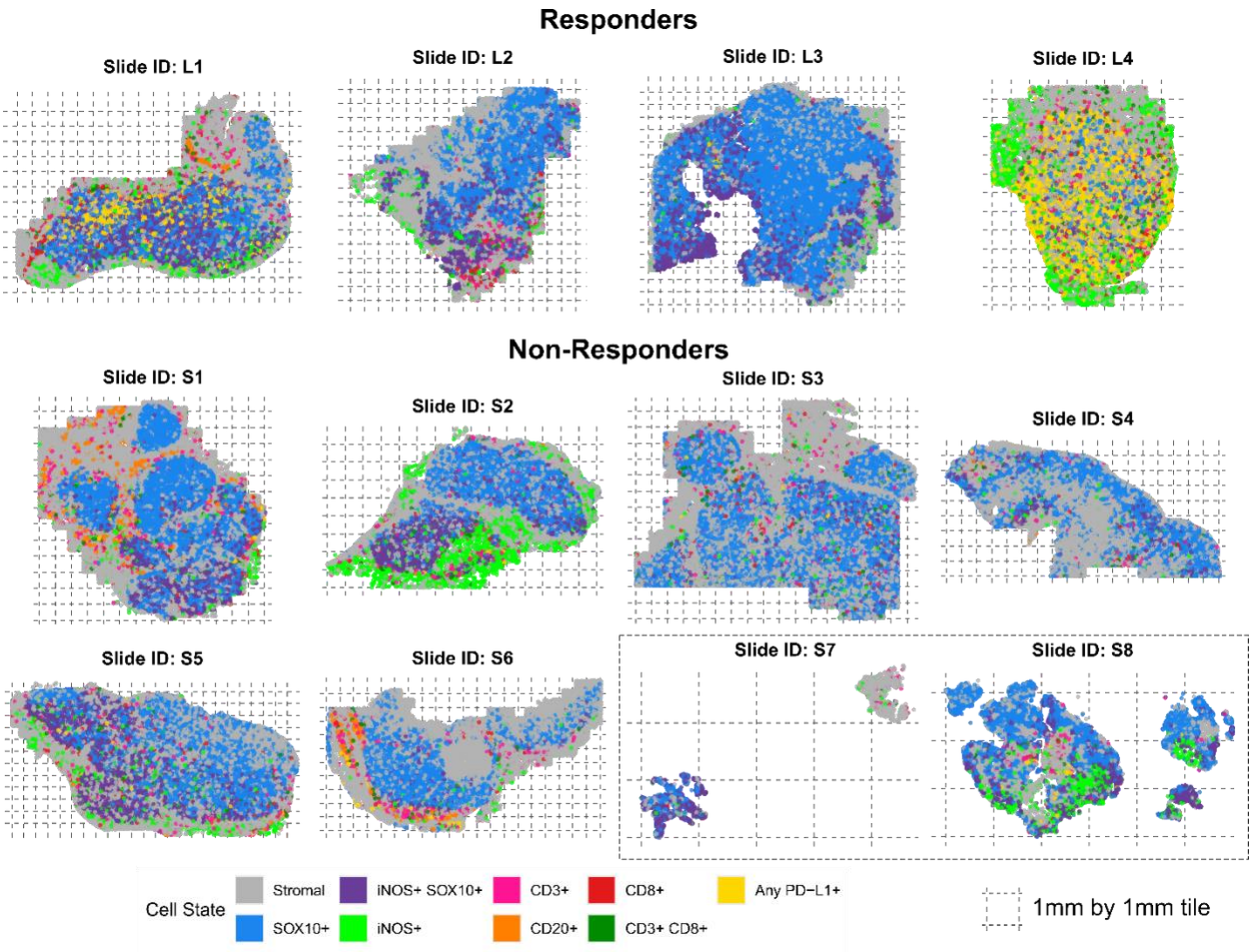
718 from responders and 988 from non-responders (left, 1706 total tiles). Cell counts per tile also showed high concordance with lymphoid panel: 100 to over 10,000 cells per tile (medians around 5,000) and biopsies with fewer cells per tile.



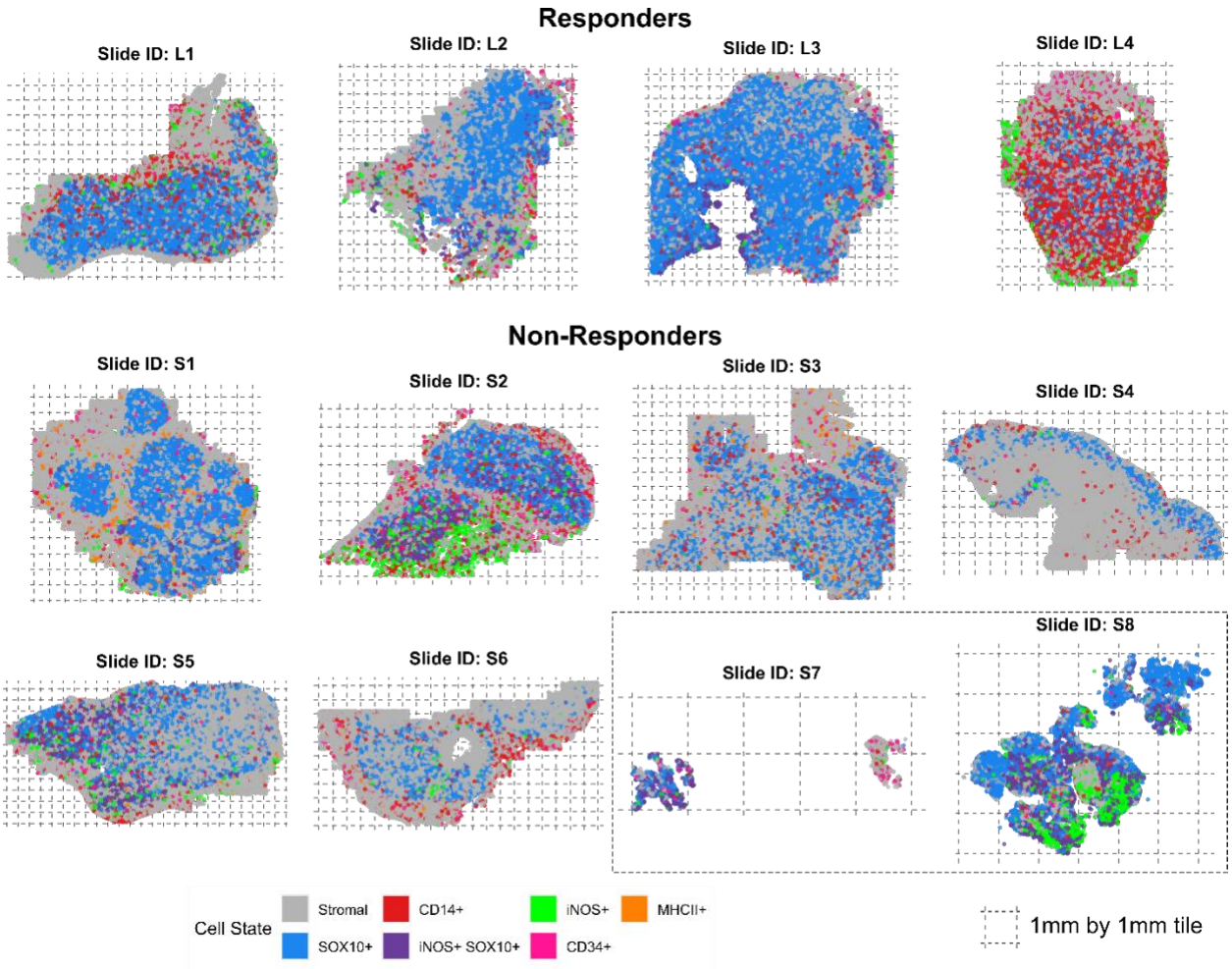
Supplementary Figure 4.8 Cell state counts and proportions, myeloid panel. (A) Stacked bar charts displaying cell counts (left) and proportions (right) of each of 20 cell states (plus an "Other" category) by slide. Bars representing ICB responders are displayed on the left (L1-L4) of each chart and bars for non-responders are on the right (S1-S8). Cells lacking any of the eight protein markers are labeled Stromal. (B) Stacked bar charts displaying cell state proportions for all cell states containing SOX10, iNOS, eNOS, CD14, CD11c, CD34, NCad, or MHCII (left to right, top to bottom). All cells which contain the specified marker but are not captured within the 20 defined cell states are represented in gray.



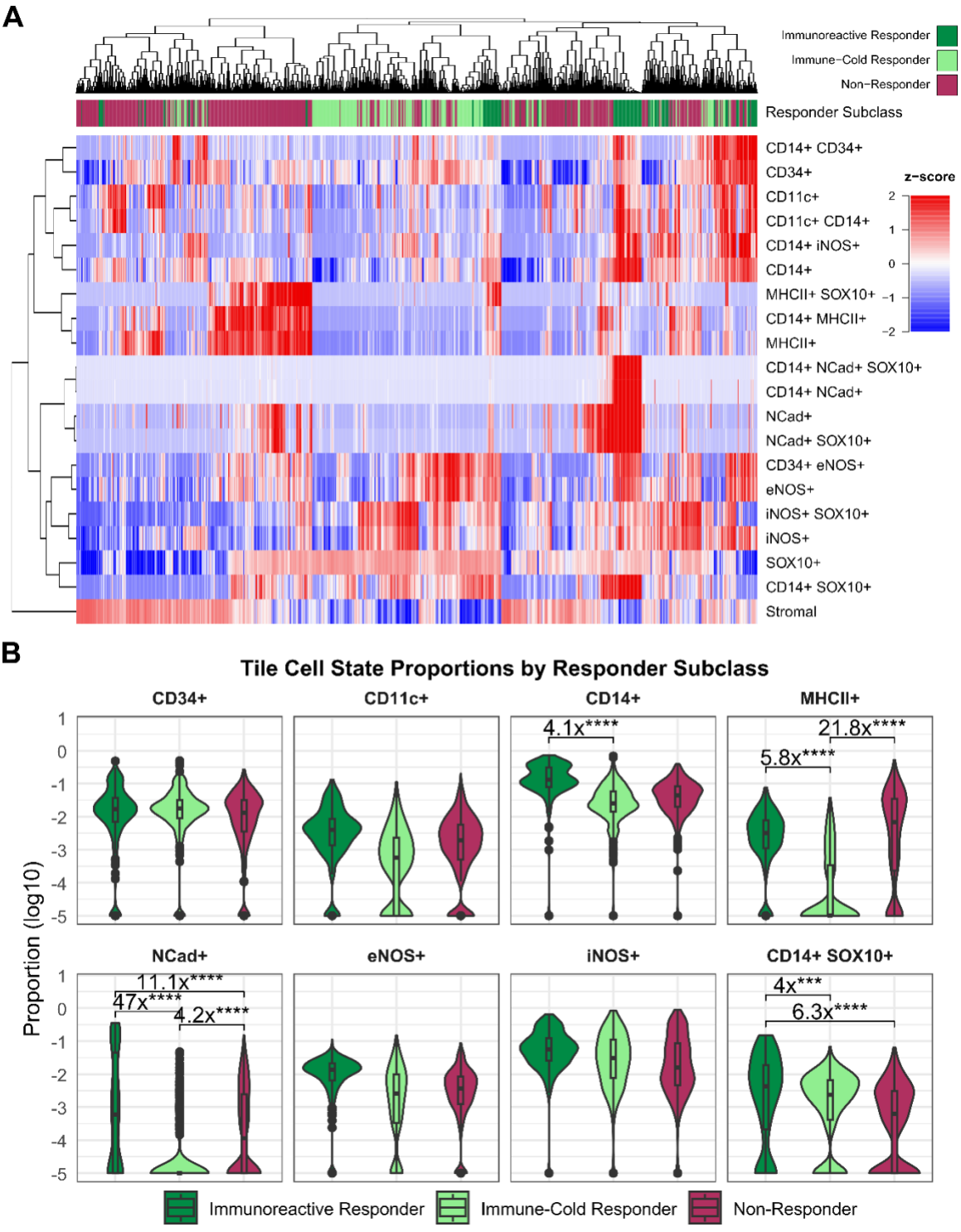
Supplementary Figure 4.9 Cell state proportions of select protein markers, lymphoid panel. Stacked bar charts displaying cell state proportions for all cell states containing CD3, CD20, SOX10, LAG-3, iNOS, or nNOS (left to right, top to bottom). All cells which contain the specified marker but are not captured within the 22 defined cell states are represented in gray.



Supplementary Figure 4.10 Cell state slide visualizations, lymphoid panel. Whole-slide visualizations of the spatial distribution of the top 8 cell states (Stromal, SOX10+, iNOS+ SOX10+, iNOS+, CD3+, CD20+, CD8+, CD3+ CD8+) and any PD-L1+ cells (yellow). Cells are plotted by their computed centroids. Dashed lines indicate tile x and y limits. Tiles are 1mm by 1mm in size. The box around samples S7 and S8 indicates these samples are tumor core biopsies; all other samples are tumor resections.

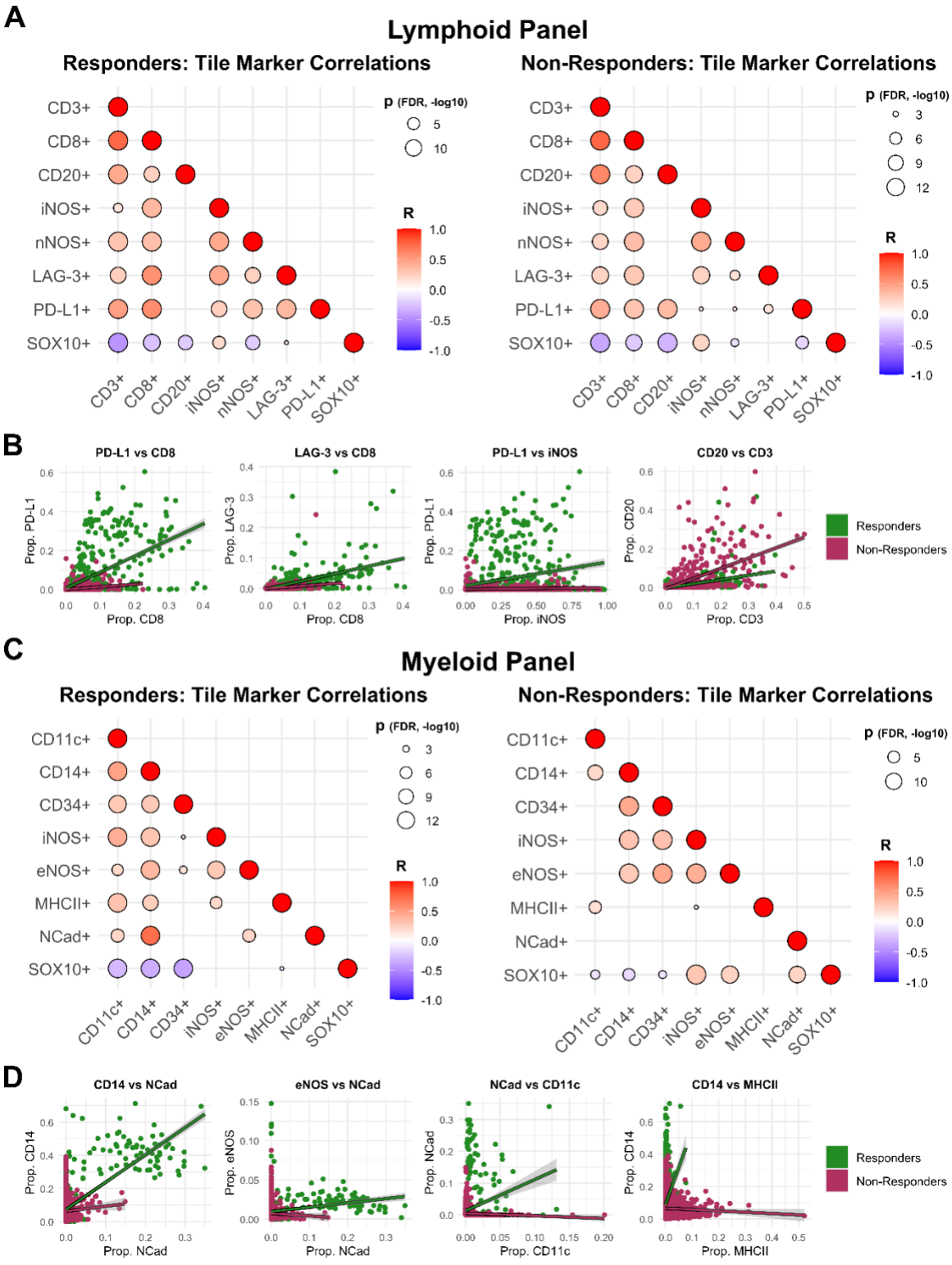


Supplementary Figure 4.11 Cell state slide visualizations, myeloid panel. Whole-slide visualizations of the spatial distribution of the top 7 cell states (Stromal, SOX10+, CD14+, iNOS+ SOX10+, iNOS+, CD34+, and MHCII+). Cells are plotted by their computed centroids. Dashed lines indicate tile x and y limits. Tiles are 1mm by 1mm in size. The box around samples S7 and S8 indicates these samples are tumor core biopsies; all other samples are tumor resections.



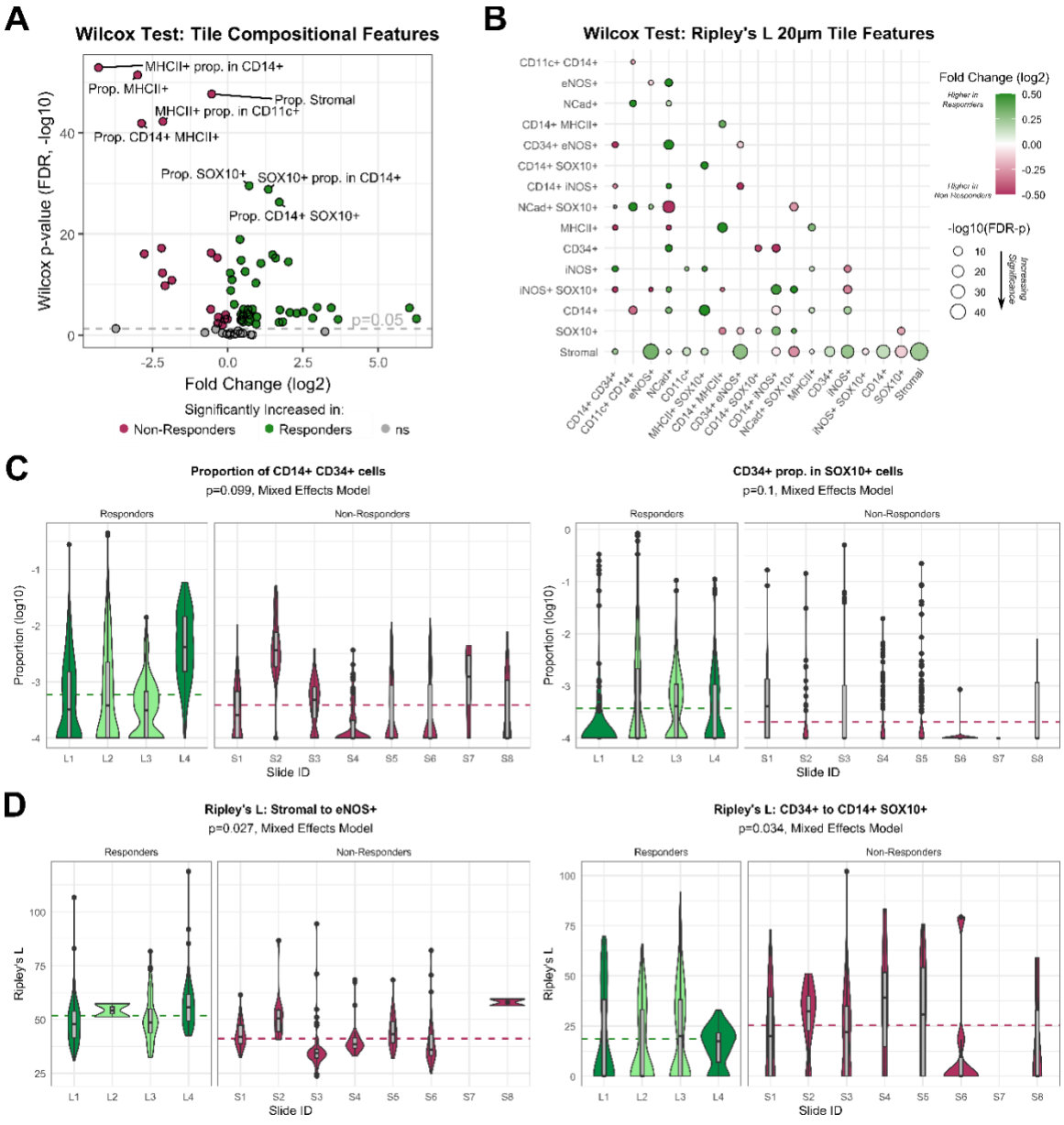
Supplementary Figure 4.12 Myeloid panel cell state compositional tile features. Myeloid panel mIF whole slide images from each patient are separated into 1mm by 1mm tiles and proportions of each of 20 defined cell states are calculated within each tile. (A) A heatmap displaying the relative proportions of each cell state (rows) within all 1,706 tiles (columns). Proportions are z-scored across rows to represent the relative expression of each protein (red to blue for high to low expression). The color bar above the heatmap denotes the ICB response subclass of each sample: Immunoreactive Responders (L1/L4), Immune-Cold Responders (L2/L3), and Non-Responders (S1-S8). Rows and columns are clustered by Euclidean distance. (B) The tile proportions of CD34+, CD11c+, CD14+, MHCII+, NCad+,

eNOS+, iNOS+, and CD14+ SOX10+ cell states (left to right, top to bottom) are displayed according to ICB response subclass: Immunoreactive Responders (L1/L4), Immune-Cold Responders (L2/L3), and Non-Responders (S1-S8). Significant differences (FDR-corrected Wilcoxon rank-sum p-values) with a fold-change of 4x or greater between response classes are displayed (****p<0.0001, ***p<0.001).



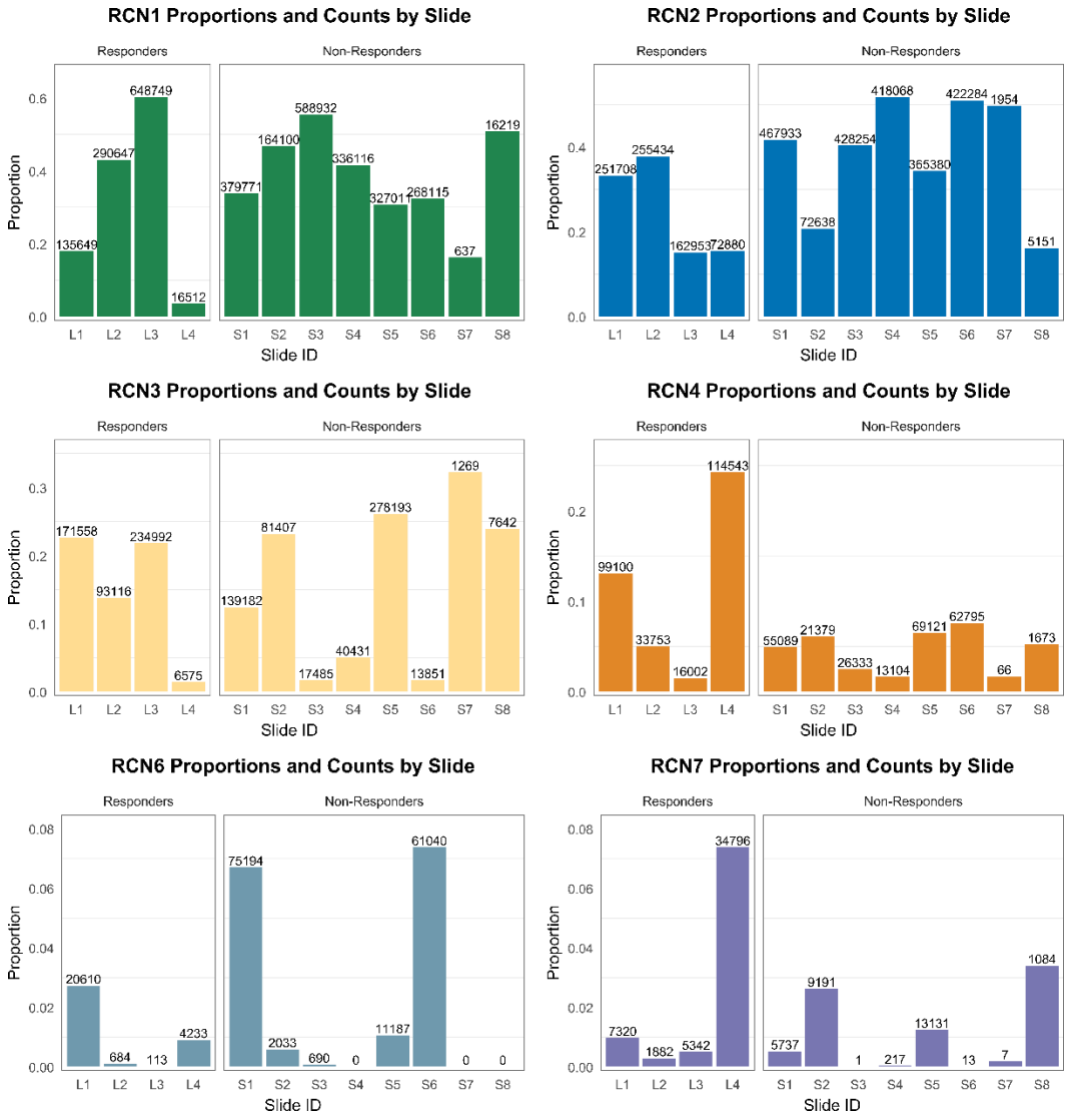
Supplementary Figure 4.13 Tile proportion correlations between protein markers differ between responders and non-responders. Ordinary least squares (OLS) regression was performed between the tile proportions of each individual marker within all responder tiles or all non-responder tiles from WSIs probed with the lymphoid mIF panel (A-B) and the myeloid mIF panel (C-D). (A, C) OLS regression statistics are displayed for each pair of markers, sized by statistical

significance (small to large indicating increasing significance by the -log10(p-value)) and colored by Pearson's R correlation coefficient (red to white to blue corresponding to positive correlation, no correlation, and inverse correlation). (B, D) Scatter plots show differences in correlations for select pairs of individual markers within tiles from ICB responders (green) or non-responders (maroon). Solid lines indicate OLS regression for each response class with shading for 90% confidence intervals. For example, tile proportions of immune checkpoint marker PD-L1 and cytotoxic T cell marker CD8 show a greater degree of correlation within responder tiles versus non-responders (B, leftmost plot).

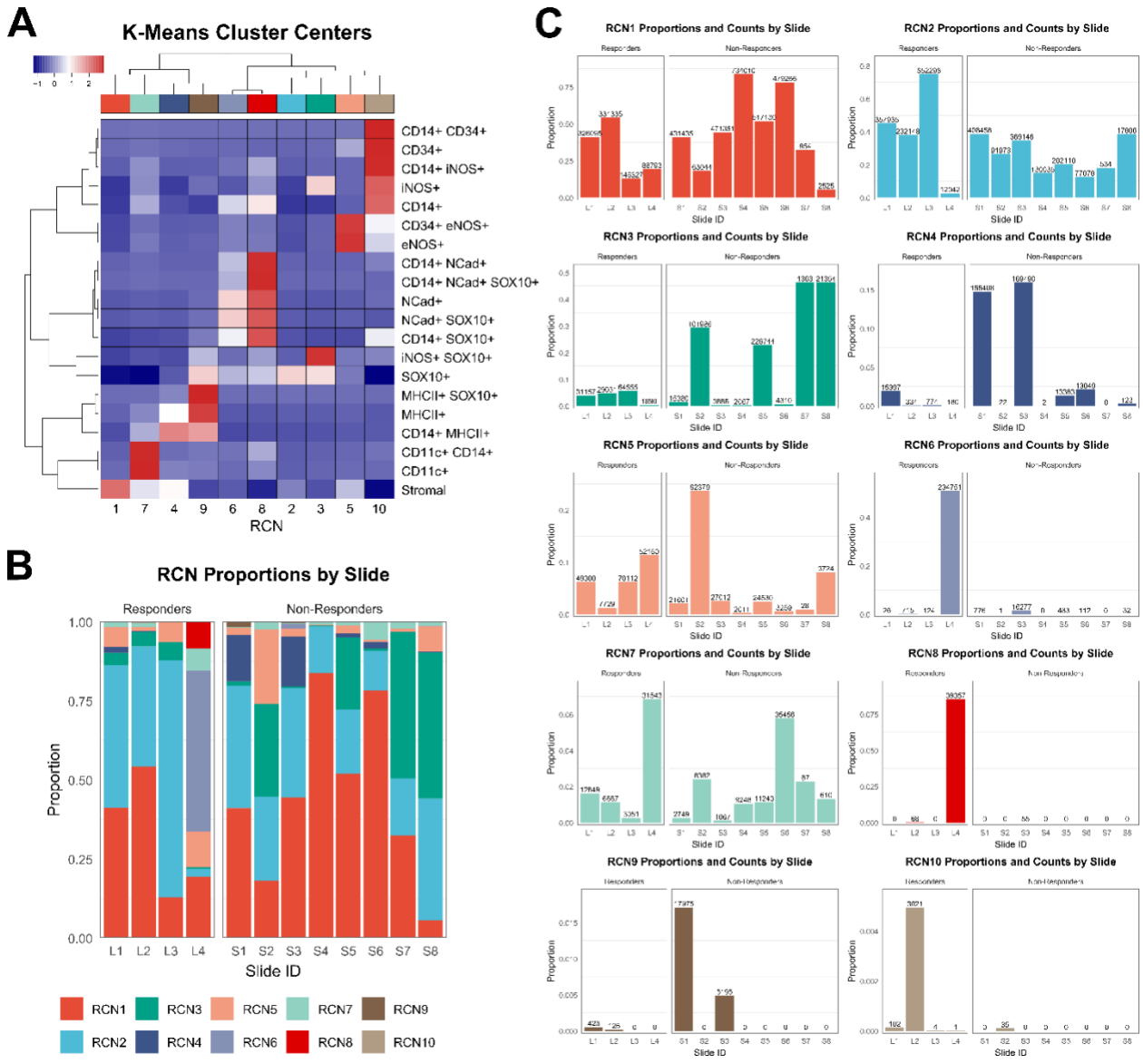


Supplementary Figure 4.14 Univariate analysis of ICB responder vs non-responder tile features from myeloid panel. Wilcoxon rank-sum test (A, B) and mixed effects modeling (C, D) were performed to identify compositional and spatial tile features from myeloid panel mIF data in pre-treatment tumor samples capable of distinguishing ICB responders and non-responders. (A) Wilcoxon rank sum test was applied to 76 compositional tile features including cell state proportions within the total tile population and proportions of each marker within subpopulations of each

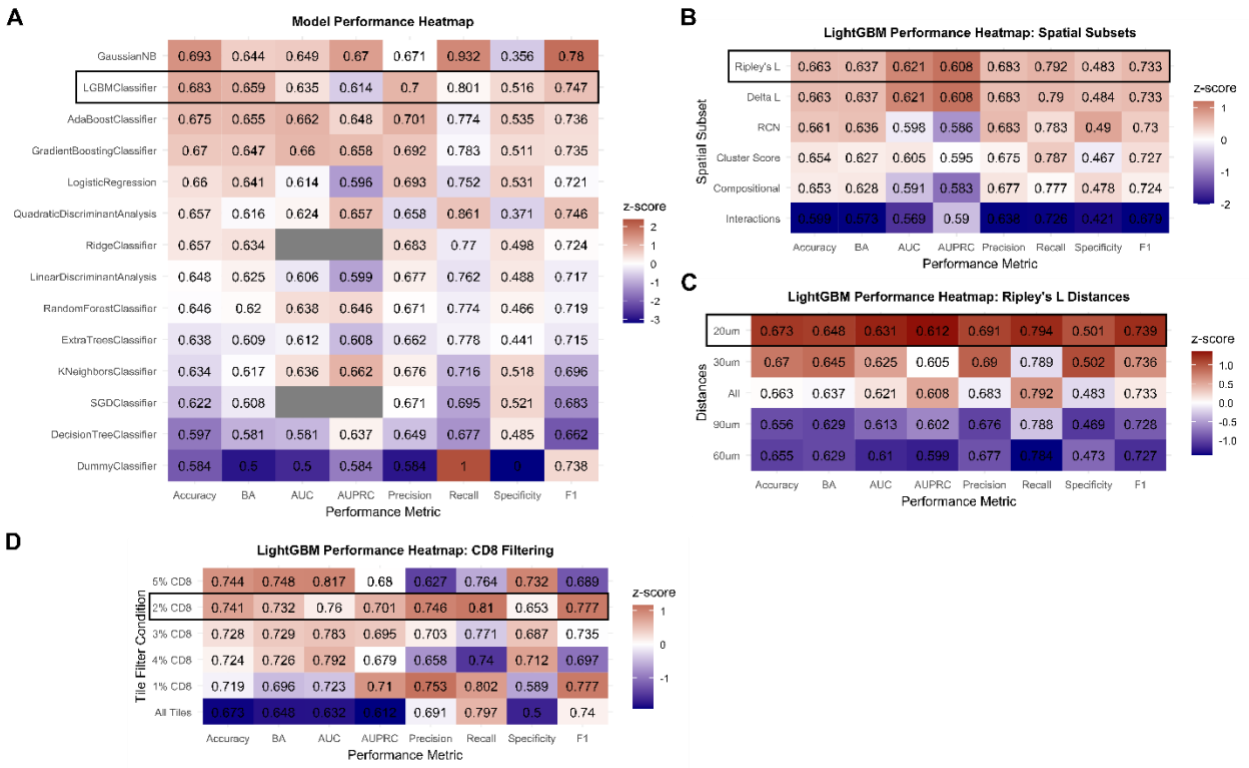
other marker. P-values were corrected for false discovery rate (FDR) using the Benjamini-Hochberg procedure. The volcano plot shows increasing significance on the vertical axis (-log10 of the FDR-adjusted p-values) and effect size on the horizontal axis (log2 of the mean fold-change). Positive fold-change values represent features increased within the ICB responder group. (B) Wilcoxon rank-sum test was applied to 169 Ripley's L spatial tile features calculated at 20 μm distances between cells and p-values were FDR-adjusted as above. The bubble plot shows bubbles of increasing size for increasing significance (-log10 of the FDR-adjusted p-values) colored by effect size (log2 of the mean fold-change, with green and maroon indicating increased values in responder or non-responder tiles, respectively) for cell state pairs indicated on the horizontal and vertical axes. Only significant differences are represented with a bubble (if a bubble is missing in the lower diagonal, $p>0.05$). (C) To account for slide-to-slide variation within tiles, we also created mixed effects models using slide ID as a random effect to be controlled. None of the 76 compositional features showed significant differences between response groups. The tile distributions of the top two features are displayed by slide and response group on the horizontal axis and log10(proportion) values on the vertical axis. Dotted lines denote the average value of the feature across all tiles of each response class. Mixed effects model p-values displayed in the subtitles are unadjusted. (D) Same as (C) but for Ripley's L spatial features calculated at 20 μm distances. Out of 91 features, 2 resulted in $p<0.05$ but neither remained significant after FDR-adjustment. Only Ripley's L features with non-missing values in at least 10 of the 12 slides were included in the analysis.



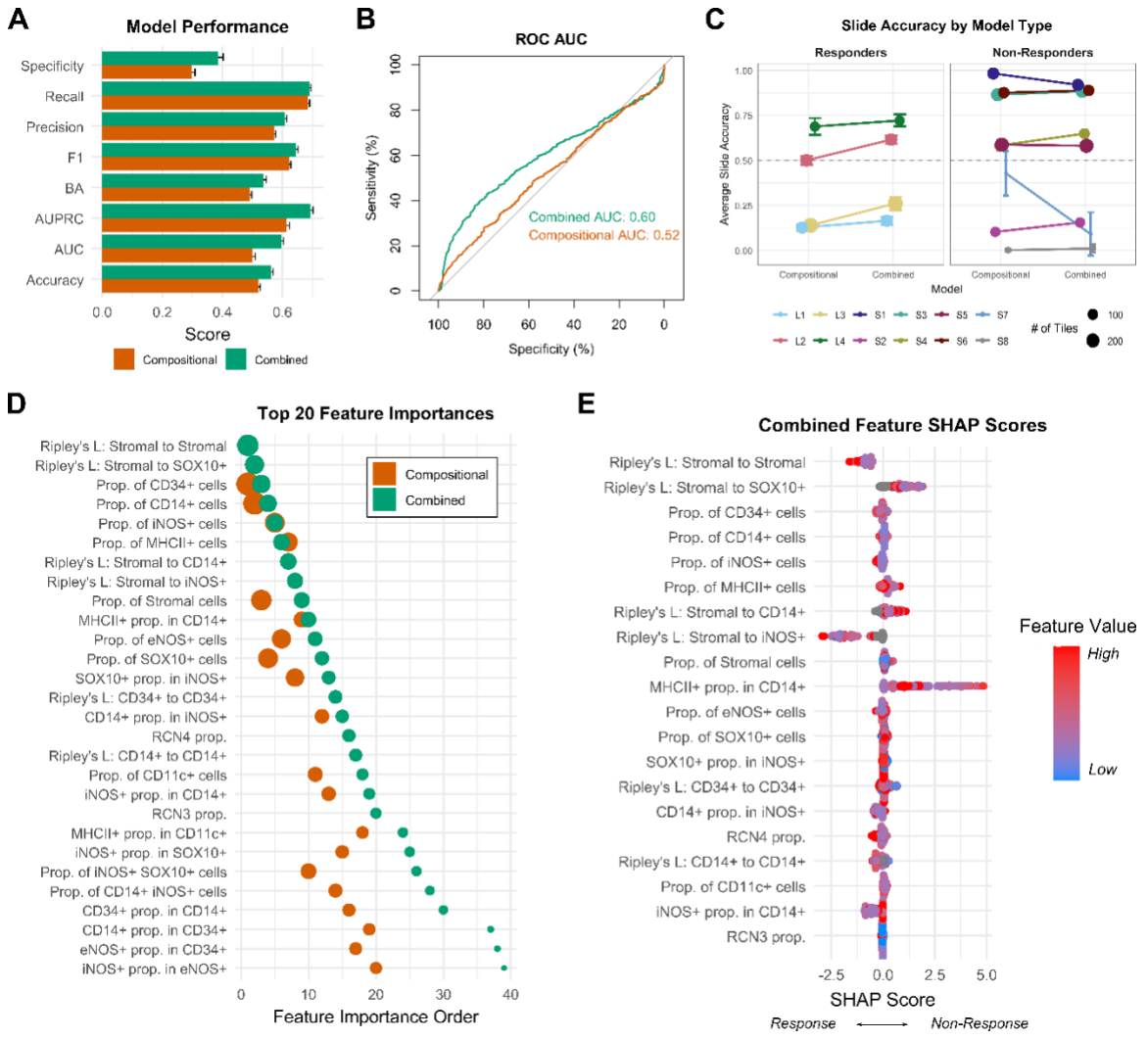
Supplementary Figure 4.15 Proportions and counts of recurrent cellular neighborhoods by slide. Bar charts showing the proportion of RCNs 1, 2, 3, 4, 6, and 7 (left to right, top to bottom) within each slide with labels for total count of RCN neighborhoods displayed above.



Supplementary Figure 4.16 Recurrent cellular neighborhood analysis of myeloid mIF panel. Recurrent cellular neighborhoods (RCNs) were determined via k-means clustering of all myeloid mIF panel cellular neighborhoods across 12 slides. (A) K-means clustering resulted in ten distinct cluster centers which define representative cell state proportion compositions for each RCN. Each column represents an RCN while rows represent cellular neighborhood cell state proportions. Red to blue indicates relatively high to low expression of the corresponding cell state proportion within each RCN. Rows and columns are clustered using Euclidean distance. RCNs are labeled according to size, with RCN1 representing the largest cluster (3,592,265 neighborhoods) down to RCN10, the smallest (3,163 neighborhoods). (B) Stacked bar chart showing the proportion of cellular neighborhoods within each slide corresponding to each RCN label. Bars representing ICB responders are displayed on the left (L1-L4) and bars for non-responders are on the right (S1-S8). (C) Bar charts showing the proportion of RCN-labeled cellular neighborhoods within each slide with labels for total count of neighborhoods displayed above.



Supplementary Figure 4.17 Machine learning model optimization. To ensure optimal prediction of ICB response by machine learning, we tested multiple ML architectures, feature subsets, and tile filtering conditions. Each model was evaluated based on average performance metrics across 10x iterations of 12-fold leave-one-out cross validation (LOOCV, leaving one of 12 patients out each time) using data from the lymphoid mIF panel. (A) We tested 13 ML architectures (plus a dummy classifier which always predicts the majority class, non-responders) and found optimal accuracy and F1 scores from the Gaussian Naïve Bayes (GaussianNB) model. However, due to its remarkably low specificity (true negative rate, prediction of ICB responders) we selected the Light Gradient-Boosting Machine (LightGBM) model which showed the highest balanced accuracy between classes. (B) To determine which spatial features provided the most useful information for classification performance, we compared performance metrics across models containing all compositional features (proportions and subpopulation proportions) plus either Ripley's L, delta L, RCN proportion, cluster score, or interaction features (see Methods for feature descriptions). Including interaction features decreased model performance, while model performance was at its best with the inclusion of Ripley's L features. (C) To determine the impact of Ripley's L distances on classification performance, we assessed models containing all compositional features plus just Ripley's L features calculated at 20, 30, 60, or 90 μm, or all four distance subsets. Limiting Ripley's L features to values calculated for longer distances (60 and 90 μm) resulted in relatively worse performance versus including just those calculated at 20 and 30 μm, with optimal performance for including just 20 μm Ripley's L features (accounting for close interactions between cell states). (D) To determine whether regions with increased immune cell presence are particularly informative for ICB response classification compared to all tissue regions, we created models using our "Combined" feature subset (compositional features, Ripley's L 20 μm features, and RCN proportions) for either all tiles or tiles containing at least 1, 2, 3, 4, or 5% CD8+ cell proportions. While the strictest filter of 5% CD8+ cells yielded the best overall accuracy, this model suffered from decreases in precision and recall, leading us to select the 2% CD8+ filter as the superior model. Every level of CD8 filtering from 1-5% resulted in increased accuracy, balanced accuracy, AUC, AUPRC, and specificity.



Supplementary Figure 4.18 ICB response classification ML models with myeloid mIF panel. Machine learning model performance metrics calculated from 100 iterations of 12-fold leave-one-out cross validation (LOOCV, leaving one of 12 patients out each time) for the 1) compositional feature only all-tile models ("Compositional") and 2) the combined compositional and spatial feature all-tile models ("Combined") with data from the myeloid mIF panel. (A) The Combined model type performed better than the Compositional model. Bars and error bars represent the average and standard deviation of each ML performance metric across 100 iterations of 12x LOOCV. (B) The receiver operating characteristic (ROC) curve shows model tradeoffs between sensitivity (true positive rate) and specificity (true negative rate). Area under the curve (AUC) represents the model's ability to distinguish between classes, with higher values corresponding to better performance. The Combined model showed a higher AUC than the Compositional model (AUC=0.60 vs AUC=0.52). (C) Tile accuracy by slide for each model type. Points represent average accuracy and error bars represent standard deviation across all 100 iterations of each leave-one-out model. Point size represents the relative number of tiles used for the model test set. The dashed gray line indicates an accuracy of 50%. (D) The order of average feature importance values for each model type Data points are colored by model type and sized by their relative average feature importance value. A feature importance order of 1 indicates the feature had the highest feature importance value within that model type. All features within the top 20 average feature importance for either model type are shown. (E) Shapley Additive exPlanation (SHAP) scores from a representative Combined model iteration of 12x LOOCV show the relative effect (SHAP score, x-axis) of each feature (y-axis) alongside its relative feature value (red to blue indicates high to low feature values) for each predicted tile (data points). Negative to positive SHAP scores indicate that the associated feature value pushed the model towards predicting ICB response or non-response respectively, while SHAP scores close to zero indicate little to no effect on prediction.

Table 4.1 Cohort characteristics. Demographic and treatment information for the study cohort of twelve stage IV melanoma patients at Moffitt Cancer Center. All patients received anti-PD1 immune checkpoint blockade (ICB) therapy (pembrolizumab or nivolumab) following tumor resection/biopsy. Four patients (L1-L4) experienced progression-free survival (PFS) of >600 days (all had ongoing PFS by the end of the study) while eight (S1-S8) showed progression within 161 days. RECIST = Response Evaluation Criteria in Solid Tumors.

*Indicates ongoing progression-free survival at time of study conclusion.

✢Indicates core biopsies (S7/S8), all other samples are surgically resected tumors.

Slide ID	Therapy	Response (RECIST)	Age at IO	Sex	Metastatic Stage	PFS (Days)	Resection Site
L1	Pembrolizumab	Partial response	61	M	M1c	911*	Intestine
L2	Pembrolizumab	Complete response	61	F	M1c	805*	Skin
L3	Pembrolizumab	Partial response	57	F	M1b	826*	Skin
L4	Nivolumab	Partial response	60	M	M1c	615*	Lung
S1	Pembrolizumab	Progressive disease	44	M	M1c	84	Lymph node
S2	Pembrolizumab	Progressive disease	41	F	M1c	148	Lymph node
S3	Pembrolizumab	Progressive disease	58	M	M1c	161	Intestine
S4	Pembrolizumab	Progressive disease	68	M	M1c	161	Intestine
S5	Pembrolizumab	Progressive disease	79	F	M1a	81	Lymph node
S6	Pembrolizumab	Progressive disease	66	M	M1c	155	Intestine
S7	Nivolumab	Progressive disease	64	F	M1c	79	Lung✢
S8	Pembrolizumab	Progressive disease	69	M	M1c	84	Lymph node✢

Supplementary Table 4.2 Lymphocyte mIF panel antibody information

Reagent	Lot#	Product#	Opal	Clone	Dilution	Manufacturer	AR
Target Antibody 1 ***iNOS***	YA3808684	**MA5-17139**	OPAL-570	4E5	**1:50**	Thermo Fisher	**ER2**
Target Antibody 2 ***CD20***	20042864	**M0755**	OPAL-480	L26	**1:150**	Dako	**ER2**
Target Antibody 3 ***PD-L1***	18	**13684s**	OPAL-540	E1L3N	**1:75**	CST	**ER2**
Target Antibody 4 ***LAG-3***	6	**15372s**	OPAL-690	D2G40	**1:75**	CST	**ER2**
Target Antibody 5 ***CD8***	41389238	**M7103**	OPAL-520	C8/144B	**1:50**	Dako	**ER2**
Target Antibody 6 ***nNOS***	2	**4231s**	OPAL-620		**1:50**	CST	**ER2**
Target Antibody 7 ***SOX10***	06022A-2	**ACI3099A**	OPAL-650	BC34	**1:50**	Biocare	**ER2**
Target Antibody 8 ***CD3***	41315552	**A0452**	OPAL-780	Rb poly	**1:100**	Dako	**ER2**
			TSA+ Dig				

Supplementary Table 4.3 Myeloid mIF panel antibody information

Reagent	Lot#	Product#	Opal	Clone	Dilution	Manufacturer	AR
Target Antibody 1 ***iNOS***	YA3808684	**MA5-17139**	OPAL-570	4E5	**1:50**	Thermo Fisher	**ER2**
Target Antibody 2 ***eNOS***	1	**35362s**	OPAL-620	D8A6N	**1:50**	CST	**ER2**
Target Antibody 3 ***N-Cadherin***	6	**13116s**	OPAL-480	D4R1H	**1:50**	CST	**ER2**
Target Antibody 4 ***CD14***	GR3421205-1	**ab238089**	OPAL-540	LPSR/2386	**1:75**	Abcam	**ER2**
Target Antibody 5 ***CD34***	GR3240236-11	**ab81289**	OPAL-520	EP373Y	**1:250**	Abcam	**ER2**
Target Antibody 6 ***MHCII***	20045582	**M0775**	OPAL-690	CR3/43	**1:150**	Dako	**ER2**
Target Antibody 7 ***SOX10***	06022A-2	**ACI3099A**	OPAL-650	BC34	**1:50**	Biocare	**ER2**
Target Antibody 8 ***CD11c***	GR3334379-3	**ab52632**	OPAL-780	EP1347Y	**1:100**	Abcam	**ER1**
			TSA+ Dig				

Chapter 5 Discussion

5.1 Common themes

Although Chapters 2, 3, and 4 focus on distinct problem domains, namely, marker imputation in multiplexed tissue imaging, multimodal embedding-based patient classification, and therapeutic response prediction in melanoma, several unifying themes emerge across these studies. A central objective throughout is the elucidation of underlying patient biology to better evaluate treatment responses and identify biologically meaningful similarities among patients that may inform therapeutic strategies. In essence, understanding why patients respond differently to treatment requires a detailed characterization of both inter-patient heterogeneity and shared biological features that may underpin common therapeutic vulnerabilities.

This overarching objective is most clearly reflected in Chapter 4, which focuses on predicting therapeutic response based on underlying biomarkers. Chapter 3 further contributes to this goal by examining whether aggregated multimodal embeddings preserve sufficient biological signal to enable downstream analyses, such as cancer type and subtype classification, and whether such embeddings continue to exhibit stratification by cancer type. Chapter 2 supports this broader aim by presenting a method to recover failed markers that would otherwise be lost, potentially compromising the evaluation of patient biology and obscuring meaningful biological similarities across patients.

Ultimately, gaining insight into the underlying biology requires a deeper understanding of model predictions; thus, model interpretability is indispensable for drawing biologically and clinically meaningful conclusions.. Without interpretability, it becomes difficult, if not impossible, to determine the rationale behind a model's decision-making process. For instance, a model might rely on the expression levels of specific protein markers to predict therapeutic response, or it may utilize distinct segments of a multimodal embedding. To facilitate model interpretability, explanation methods such as SHAP (SHapley Additive exPlanations) [166] were employed to quantify the contribution of individual features or embedding components to predictive outcomes. Such interpretability is essential not only for biological insight but also for evaluating the reliability and biomedical relevance of model outputs. In parallel, rigorous performance evaluation remains a critical aspect of validating model findings. Across all chapters, performance metrics were selected based on the specific biological hypothesis under investigation, as well as their established usage in the literature.

Across all chapters, a central objective was the extraction of biologically informative and computationally tractable features to enhance predictive performance in clinically relevant tasks. Chapters 2 and 3 addressed this by employing advanced dimensionality reduction strategies, including Autoencoders (AE) and Variational Autoencoders (VAE). These methods were chosen for their capacity to model non-linear relationships in high-dimensional data, and to learn compact, denoised latent representations that retain biologically salient structure.

In Chapter 2, the use of AE and VAE was particularly advantageous for handling transcriptomic data characterized by sparsity and measurement noise. The learned embeddings not only reduced dimensionality but also implicitly denoised the data. This was reflected in improved classification accuracy in the biopsy timepoint prediction task. Notably, models trained on the latent representations of imputed data achieved comparable or superior performance to those trained on raw, noise-prone input, suggesting that the latent space captured critical temporal or pathological signals relevant to disease progression.

In Chapter 3, these dimensionality reduction techniques were applied to two fundamentally different data modalities: high-resolution hematoxylin and eosin (H&E) stained tissue sections and single-cell RNA-seq data. Despite their differing structures, both datasets benefited from latent space modeling. For H&E images, the VAE reduced spatial complexity while preserving morphological features critical for tissue classification. For scRNA-seq, VAE-derived representations preserved their biological information, which in turn supported robust cell-type classification and integrative modeling across samples.

In contrast, Chapter 4 employed a spatial tiling strategy to address the challenge of heterogeneity within whole-slide biopsy images. Rather than reducing dimensionality in the abstract feature space, this approach reduced spatial complexity by partitioning images into localized tiles. This allowed for more targeted learning of region-specific features and enabled spatially resolved predictions, which would be masked in whole-slide aggregation. The tiling approach not only improved predictive accuracy but also facilitated interpretability, as model outputs could be traced to specific anatomical subregions.

Together, these approaches reflect a consistent effort across chapters to balance model complexity, predictive accuracy, and biological interpretability. Each method was selected to align with the structure of the data and the underlying biological questions, with measurable gains in downstream performance and retention of biologically meaningful patterns. Nonetheless, each approach has limitations, latent space methods may obscure direct interpretability, and tiling may neglect global context. These trade-offs are addressed in the discussion sections of each chapter. In total, a common set of computational and biological themes emerges from the studies presented in this dissertation. While each chapter addresses a distinct problem, ranging from marker imputation to multimodal classification and therapeutic response prediction, all share a foundational goal: to interpret patient-specific biology in a way that enables precise and informative clinical insights. The complexity of the datasets used, including high-dimensional imaging and transcriptomic data, necessitates advanced computational strategies not only to reduce dimensionality but also to preserve signal relevant to biological interpretation. As such, this work underscores the importance of model interpretability, the careful selection of performance metrics tied to biological hypotheses, and the implementation of architectures capable of learning from incomplete, noisy, or heterogeneous data.

From a biological perspective, these studies emphasize the critical need to understand the features that differentiate patients as well as those that connect them. Whether through imputing missing markers, aggregating multimodal representations, or dissecting spatial patterns in tissue architecture, each approach attempts to uncover the underlying

structure of disease heterogeneity. The variability in patient biology remains a fundamental challenge in developing generalizable predictive models, yet the results presented here demonstrate that, when carefully designed, computational tools can begin to map this landscape. Importantly, this dissertation does not attempt to define a universal biological fingerprint, but rather illustrates how nuanced, data-driven approaches can move us closer to that goal by identifying patterns that hold across contexts while accommodating the complexity of individual cases.

In sum, this body of work provides a framework for integrating computational modeling with biological interpretation across spatial, multimodal, and clinical dimensions. It highlights both the potential and limitations of current methodologies and sets the stage for future studies aimed at deepening our understanding of patient-specific features that influence disease progression and treatment response.

5.2 Translational and Clinical Opportunities

Building upon the common computational and biological themes identified across Chapters 2, 3, and 4, the findings presented in this dissertation highlight several avenues for advancing machine learning–enabled tools toward translational and clinical utility. Central to this work is the development of flexible, biologically informed, and data-driven methodologies that operate across spatial, multimodal, and clinical contexts to better characterize patient-level heterogeneity. While the analyses remain primarily exploratory, they establish a conceptual and technical framework that may inform future applications in diagnostic classification, therapeutic decision-making, and personalized oncology. By

emphasizing interpretability, biological relevance, and methodological adaptability, this dissertation contributes foundational strategies that can support the integration of computational modeling into clinically actionable workflows.

Chapter 2 highlights how addressing missing data in spatial proteomics assays can directly improve the utility of these technologies in clinical settings. Proteomic imaging is increasingly used to characterize the tumor microenvironment (TME) and guide treatment decisions, but is often limited by technical artifacts and sample degradation. The imputation strategies developed here, particularly those incorporating spatial context, enhance data completeness and reliability, which is essential for downstream use in diagnostic algorithms or predictive models. Moreover, the observed marker-specific variation in optimal spatial parameters suggests that biologically informed tuning may improve clinical model development. These findings suggest that imputation not only recovers lost signal but may also denoise high-dimensional data in ways that could improve robustness and interpretability in translational pipelines.

The work in Chapter 3 illustrates the potential for multimodal embeddings to support integrated patient similarity analyses across data types frequently used in clinical oncology, including histology, gene expression, somatic mutation profiles, and clinical metadata. As cancer diagnosis and treatment increasingly rely on the convergence of diverse data modalities, the ability to perform unified patient comparisons across these sources can inform more nuanced stratification strategies and case-based reasoning. Embedding-based vector representations can be readily incorporated into scalable similarity search systems, such as clinical dashboards or decision-support platforms,

without requiring retraining or labor-intensive harmonization. This capability may support applications such as identifying patients with similar molecular profiles or treatment histories, selecting relevant clinical trials, or prioritizing patients for targeted therapies.

Chapter 4 underscores the potential of machine learning to improve therapeutic decision-making in immuno-oncology, specifically through early prediction of response to immune checkpoint blockade (ICB) in advanced melanoma. In a clinical landscape where timely and effective treatment selection is critical, predictive models that integrate spatial and molecular features could help identify patients most likely to benefit from immunotherapy. By moving beyond traditional biomarkers to consider spatial interactions and high-dimensional protein patterns, this work offers a path toward more individualized treatment planning. The interpretability of these models further enhances their translational value by pointing to specific cell types or microenvironmental structures associated with therapeutic response, insights that could inform biomarker development, treatment combinations, or future clinical trials.

Together, these findings suggest several translational directions. Embedding-based representations could be deployed in clinical decision-support systems to enable real-time, patient-specific insights. Imputation tools for spatial omics may be integrated into digital pathology workflows to ensure completeness and interpretability of tissue-based assays. And similarity-based retrieval using multimodal embeddings could assist in treatment selection or cohort matching for research and trial enrollment. As spatial and multimodal technologies continue to enter clinical practice, these computational

strategies offer a pathway to make complex biological data more accessible, interpretable, and actionable in the clinical setting.

5.3 Future directions

This dissertation explores the application of machine learning to biological inference across spatial, multimodal, and clinical dimensions. The models utilized in this work, including LightGBM (LGBM), autoencoders (AEs), and standard feed-forward neural networks, illustrate the potential of data-driven methodologies in a range of biomedical contexts. Despite their utility, these models exhibit several limitations. For instance, the experiments detailed in Chapter 2 employed elastic net (EN) regression and LGBM models for protein imputation. Both approaches are constrained to imputing one protein at a time, necessitating repeated model retraining. In our study, this requirement resulted in 16 separate training procedures, significantly increasing computational burden. Moreover, the development of a dedicated model for each marker is contingent upon the availability of corresponding training data.

In Chapter 3, models were trained on four distinct data modalities: hematoxylin and eosin (H&E) stained images, somatic mutation profiles, expert annotations, and transcriptomic data. While these modalities capture complementary aspects of the biological state, they do not encompass the full spectrum of available information. Proteomic and genomic profiles, among other data types, remain unincorporated, despite their potential to offer a more comprehensive representation of the underlying biology. The integration of such

additional modalities could substantially enhance the granularity and accuracy of patient-specific biological inference. However, as the number and complexity of data modalities continue to grow, traditional modeling approaches become increasingly limited in their scalability and adaptability. This challenge highlights the need for more flexible, unified frameworks capable of leveraging heterogeneous inputs without extensive task-specific retraining.

One particularly promising direction to address the issue of imputing missing data involves the development of foundation models; large-scale models trained across diverse and heterogeneous datasets to enable generalizable representations and broad applicability across downstream tasks [182–185]. The models in Chapter 2 required intensive retraining, which possibly could be prevented by using a foundation model. In Chapter 3, for instance, a model was constructed using all available data; however, its training was restricted to a limited subset of six cancer types. This constraint inherently limits the model's capacity for generalization. In contrast, a foundation model trained across a broader and more diverse set of cancer types and data modalities could offer substantial advantages.

Such models are particularly valuable in complex and heterogeneous diseases like cancer, where traditionally, separate task-specific models must be trained for each cohort or prediction objective. Foundation models provide a unified framework that can capture both shared and distinct biological patterns across contexts. Even the same arguments that apply to specialized models, such as uncovering meaningful similarities and differences between cancer types, can be extended to foundation models, but at a broader and potentially population-wide scale. As such, developing and evaluating foundation models

across large, diverse cohorts holds considerable promise for advancing both the predictive performance and biological interpretability of machine learning in biomedical research.

The foundation model paradigm may also be extended to the domain of data imputation. In Chapter 2, the focus was on imputing protein marker expression using a limited subset of available markers, where at most three markers were imputed simultaneously based on the remaining observed markers in the dataset. The models developed in this context were constrained by the specific marker combinations present during training and, as a result, could not generalize beyond this predefined subset. In contrast, a foundation model trained across a broad set of protein markers could enable marker-agnostic imputation, offering the ability to infer missing values for any marker, without requiring retraining for each new configuration. Such an approach would substantially increase the flexibility and scalability of imputation in high-dimensional proteomic data.

Similarly, the predictive modeling work presented in Chapter 4 could benefit from foundation model development. As larger and more heterogeneous patient cohorts become available, foundation models have the potential to capture a wider spectrum of biological variability, enhancing the model's capacity to generalize. The increased diversity in training data would allow such models to learn more robust representations, improving both accuracy and reliability when applied to independent cohorts. Thus, the adoption of foundation model strategies across multiple components of this dissertation's methodological framework represents a natural and promising extension for future research.

Another avenue for future development builds on the groundwork laid in Chapter 3, which demonstrated that it is feasible to aggregate embeddings from heterogeneous and unaligned embedding spaces into a unified representation, while still preserving predictive signal for downstream tasks such as cancer type and subtype classification, as well as tumor mutational burden (TMB) estimation.

These aggregated embeddings could serve as inputs to Graph Neural Networks (GNNs), enabling the construction of patient- or sample-level graphs in which nodes represent individuals and edges capture biological, clinical, or embedding-derived similarities [186,187]. Such GNN-based models would facilitate subpopulation discovery through unsupervised clustering on graph-structured data, as well as enable supervised classification tasks such as predicting therapeutic response or identifying cohort-specific differences. By incorporating relational structure among patients, these approaches could yield more nuanced and context-aware representations than models that treat each sample independently.

In addition to graph-based modeling, a more focused but impactful future direction involves refining the representation of patient annotations. In Chapter 2, patient-level textual annotations were embedded using Sentence-BERT (sBERT). However, the results suggest that sBERT may not be optimally suited for capturing the domain-specific semantics of biomedical or clinical text for downstream tasks. Fine-tuning language models on clinically annotated corpora or constructing task-specific embeddings with supervised objectives has the potential to substantially enhance the informativeness and modality alignment of textual representations. These approaches can improve the

integration of text-derived features in multimodal frameworks by ensuring that the resulting embeddings are semantically aligned with biological and clinical data. Similar strategies have been employed in the context of tumor mutation trees, where embedding models trained on mutation data have been shown to improve clustering performance and effectively capture the complex interdependencies inherent in somatic mutation data [188].

Building upon the findings presented in Chapter 3, which demonstrated the feasibility of aggregating vectors across unaligned latent spaces, a natural next step involves the development of a comprehensive vector database composed of these aggregated representations. Such a resource would enable integrative, multi-modal investigations of patient subgroups across diverse omics layers and clinical data.

This proposed vector database would serve as a foundational infrastructure for holistic subgroup discovery, facilitating the identification and characterization of clinically and biologically meaningful clusters. These subgroups could subsequently be interrogated for their internal heterogeneity, potentially revealing finer-grained disease states or subtypes not apparent through unimodal analysis. Iterative refinement and subdivision of such clusters may uncover latent phenotypes, contributing to a deeper mechanistic understanding of disease progression.

To translate such a system from conceptual design to reliable clinical application, rigorous validation and methodological safeguards are essential. To ensure the reliability and generalizability of such findings, validation using independent clinical cohorts will be

essential. External datasets provide a means to assess the reproducibility of subgroup definitions and to confirm that observed patterns are not artifacts of cohort-specific sampling or preprocessing biases. Validation across cohorts with distinct demographic, clinical, or technical characteristics will further strengthen the biological interpretability and translational value of the identified subgroups.

Moreover, given that certain modalities inherently capture spatially structured information, it is important to incorporate statistical considerations specific to spatial data analysis. Spatial dependencies between neighboring observations violate assumptions of independence and can affect statistical inference. Approaches such as in silico spatial power analysis, as outlined by Baker et al.[189], can help determine sampling requirements, optimize field-of-view size, and ensure sufficient power to detect spatial relationships or cell–cell interactions. Incorporating such frameworks into the analytical design will enhance the robustness of spatially informed findings and ensure that detected spatial patterns are statistically meaningful rather than products of sampling variance.

This vector database framework could support the development of real-time clinical decision support systems. By dynamically updating a patient's vector representation as new data becomes available, *e.g.*, the integration of an additional imaging vector, the patient's position within the low-dimensional space would shift accordingly. This dynamic repositioning may provide clinicians with a more accurate and up-to-date visualization of

disease trajectory, thereby enhancing the ability to monitor treatment response and inform personalized therapeutic strategies.

Establishing such a system will require careful consideration of data harmonization, privacy-preserving computation, and interpretability. Nonetheless, the potential for advancing precision medicine through an adaptive, multi-omics-informed patient representation system is considerable and warrants further investigation.

5.4. Concluding remarks

The application of machine learning and deep learning to biological inference holds transformative potential for understanding complex disease processes and improving patient care. This dissertation advances that vision by demonstrating how machine learning can be systematically applied across spatial, multimodal, and clinical dimensions to uncover meaningful biological patterns and inform clinical decision-making. Specifically, the work presented here contributes new methods for imputing missing markers in multiplexed tissue imaging assays, explores the feasibility of integrating unaligned latent spaces for multimodal patient characterization, and applies predictive modeling to understand patient response to immune checkpoint blockade therapies.

Together, these studies contribute to the evolving framework of precision oncology by enhancing the ability to model patient similarity, stratify treatment response, and extract clinically actionable insights from high-dimensional biomedical data. More broadly, this dissertation establishes a foundation for future applications of machine learning across

diverse domains of cancer research, reinforcing the role of spatial context, multimodal integration, and clinical relevance in building more accurate and interpretable computational models for biomedical discovery and therapeutic innovation.

1. Parvizpour S, Beyrampour-Basmenj H, Razmara J, Farhadi F, Shamsir MS. Cancer treatment comes to age: from one-size-fits-all to next-generation sequencing (NGS) technologies. BioImpacts. 2024;14:29957. https://doi.org/10.34172/bi.2023.29957

2. Sbitan L, Alzraikat N, Tanous H, Saad AM, Odeh M. From one size fits all to a tailored approach: integrating precision medicine into medical education. BMC Med Educ. 2025;25:90. https://doi.org/10.1186/s12909-024-06138-y

3. Kucherlapati R. Impact of Precision Medicine in Oncology. The Cancer Journal. 2023;29:1–2. https://doi.org/10.1097/PPO.0000000000000642

4. Kalia M. Personalized oncology: Recent advances and future challenges. Metabolism. 2013;62:S11–4. https://doi.org/10.1016/j.metabol.2012.08.016

5. Rituraj, Pal RS, Wahlang J, Pal Y, Chaitanya M, Saxena S. Precision oncology: transforming cancer care through personalized medicine. Medical Oncology. 2025;42:246. https://doi.org/10.1007/s12032-025-02817-y

6. Bertolaccini L, Casiraghi M, Uslenghi C, Maiorca S, Spaggiari L. Recent advances in lung cancer research: unravelling the future of treatment. Updates Surg. 2024;76:2129–40. https://doi.org/10.1007/s13304-024-01841-3

7. Lin R. AI-Driven Personalized Healthcare: Leveraging Multimodal Data for Precision Medicine. 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). IEEE; 2024. p. 693–7. https://doi.org/10.1109/WI-IAT62293.2024.00112

8. Horgan D, Tanner M, Aggarwal C, Thomas D, Grover S, Basel-Salmon L, et al. Precision Oncology: A Global Perspective on Implementation and Policy Development. JCO Glob Oncol. 2025; https://doi.org/10.1200/GO-24-00416

9. Budczies J, Kazdal D, Menzel M, Beck S, Kluck K, Altbürger C, et al. Tumour mutational burden: clinical utility, challenges and emerging improvements. Nat Rev Clin Oncol. 2024;21:725–42. https://doi.org/10.1038/s41571-024-00932-9

10. Khader F, Müller-Franzes G, Wang T, Han T, Tayebi Arasteh S, Haarburger C, et al. Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters: A Case for Transformers. Radiology. 2023;309. https://doi.org/10.1148/radiol.230806

11. Naithani N, Atal AT, Tilak TVSVGK, Vasudevan B, Misra P, Sinha S. Precision medicine: Uses and challenges. Med J Armed Forces India. 2021;77:258–65. https://doi.org/10.1016/j.mjafi.2021.06.020

12. Perelló-Reus CM, Rubio-Tomás T, Cisneros-Barroso E, Ibargüen-González L, Segura-Sampedro JJ, Morales-Soriano R, et al. Challenges in precision medicine in pancreatic cancer: A focus in cancer stem cells and microbiota. Front Oncol. 2022;12. https://doi.org/10.3389/fonc.2022.995357

13. Truong T, Kelly RT. What's new in single-cell proteomics. Curr Opin Biotechnol. 2024;86:103077. https://doi.org/10.1016/j.copbio.2024.103077

14. Chang X, Zheng Y, Xu K. Single-Cell RNA Sequencing: Technological Progress and Biomedical Application in Cancer Research. Mol Biotechnol. 2024;66:1497–519. https://doi.org/10.1007/s12033-023-00777-0

15. Liang A, Kong Y, Chen Z, Qiu Y, Wu Y, Zhu X, et al. Advancements and applications of single-cell multi-omics techniques in cancer research: Unveiling heterogeneity and paving the way for precision therapeutics. Biochem Biophys Rep. 2024;37:101589. https://doi.org/10.1016/j.bbrep.2023.101589

16. Yang X, Huang K, Yang D, Zhao W, Zhou X. Biomedical Big Data Technologies, Applications, and Challenges for Precision Medicine: A Review. Global Challenges. 2024;8. https://doi.org/10.1002/gch2.202300163

17. Abdelaziz EH, Ismail R, Mabrouk MS, Amin E. Multi-omics data integration and analysis pipeline for precision medicine: Systematic review. Comput Biol Chem. 2024;113:108254. https://doi.org/10.1016/j.compbiolchem.2024.108254

18. Ge S, Sun S, Xu H, Cheng Q, Ren Z. Deep learning in single-cell and spatial transcriptomics data analysis: advances and challenges from a data science perspective. Brief Bioinform. 2025;26. https://doi.org/10.1093/bib/bbaf136

19. Chen H, Ryu J, Vinyard ME, Lerer A, Pinello L. SIMBA: single-cell embedding along with features. Nat Methods. 2024;21:1003–13. https://doi.org/10.1038/s41592-023-01899-8

20. Tang Z, Chen G, Chen S, He H, Huang J, Dong T, et al. Modal-NexT: Towards unified heterogeneous cellular data integration. Information Fusion. 2026;125:103479. https://doi.org/10.1016/j.inffus.2025.103479

21. Meyer L, Jackson HW, Eling N, Zhao S, Usui G, Dakhli H, et al. A stratification system for breast cancer based on basoluminal tumor cells and spatial tumor architecture. Cancer Cell. 2025; https://doi.org/10.1016/j.ccell.2025.06.019

22. Chu X, Li X, Zhang Y, Dang G, Miao Y, Xu W, et al. Integrative single-cell analysis of human colorectal cancer reveals patient stratification with distinct immune evasion mechanisms. Nat Cancer. 2024;5:1409–26. https://doi.org/10.1038/s43018-024-00807-z

23. Zheng Y, Sadée C, Ozawa M, Howitt BE, Gevaert O. Single-cell multimodal analysis reveals tumor microenvironment predictive of treatment response in non–small cell lung cancer. Sci Adv. 2025;11. https://doi.org/10.1126/sciadv.adu2151

24. Watson ER, Taherian Fard A, Mar JC. Computational Methods for Single-Cell Imaging and Omics Data Integration. Front Mol Biosci. 2022;8. https://doi.org/10.3389/fmolb.2021.768106

25. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer. 2022;22:114–26. https://doi.org/10.1038/s41568-021-00408-3

26. Chen J, Larsson L, Swarbrick A, Lundeberg J. Spatial landscapes of cancers: insights and opportunities. Nat Rev Clin Oncol. 2024;21:660–74. https://doi.org/10.1038/s41571-024-00926-7

27. Lee S, Kim G, Lee J, Lee AC, Kwon S. Mapping cancer biology in space: applications and perspectives on spatial omics for oncology. Mol Cancer. 2024;23:26. https://doi.org/10.1186/s12943-024-01941-z

28. Chen JH, Gainor JF. Spatial biology captures the effects of neoadjuvant chemo-immunotherapy in lung cancer. Nat Genet. 2025;57:6–8. https://doi.org/10.1038/s41588-024-01992-4

29. Lin J-R, Izar B, Wang S, Yapp C, Mei S, Shah PM, et al. Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. 2018; https://doi.org/10.7554/eLife.31657.001

30. Schürch CM, Bhate SS, Barlow GL, Phillips DJ, Noti L, Zlobec I, et al. Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. Cell. 2020;182:1341-1359.e19. https://doi.org/10.1016/j.cell.2020.07.005

31. Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell. 2022;40:1095–110. https://doi.org/10.1016/j.ccell.2022.09.012

32. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer. 2022;22:114–26. https://doi.org/10.1038/s41568-021-00408-3

33. Acharya D, Mukhopadhyay A. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. Brief Funct Genomics. 2024;23:549–60. https://doi.org/10.1093/bfgp/elae013

34. Kourou K, Exarchos KP, Papaloukas C, Sakaloglou P, Exarchos T, Fotiadis DI. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Comput Struct Biotechnol J. Elsevier B.V.; 2021. p. 5546–55. https://doi.org/10.1016/j.csbj.2021.10.006

35. Caroline Diana N. Adoption of machine learning in diagnosis and treatment in healthcare; A systematic literature review [Internet]. International Academic Journal of Health. 2023. https://iajournals.org/articles/iajhmn_v2_i1_380_399.pdf

36. Unger M, Kather JN. Deep learning in cancer genomics and histopathology. Genome Med. BioMed Central Ltd; 2024. https://doi.org/10.1186/s13073-024-01315-6

37. Nam Y, Kim J, Jung S-H, Woerner J, Suh EH, Lee D, et al. Harnessing Artificial Intelligence in Multimodal Omics Data Integration: Paving the Path for the Next Frontier in Precision Medicine. Annu Rev Biomed Data Sci. 2024;7:225–50. https://doi.org/10.1146/annurev-biodatasci-102523-103801

38. Ramalingam K, Yadalam PK, Ramani P, Krishna M, Hafedh S, Badnjević A, et al. Light gradient boosting-based prediction of quality of life among oral cancer-treated patients. BMC Oral Health. 2024;24:349. https://doi.org/10.1186/s12903-024-04050-x

39. Zhao X, Yang V, Zou J, Jia S, Chen A, Ranasinghe P. Precision-calibrated LightGBM machine learning model to predict serious adverse events in oncology patients using FAERS. Journal of Clinical Oncology. 2025;43:12017–12017. https://doi.org/10.1200/JCO.2025.43.16_suppl.12017

40. Goswami B, Bhuyan MK, Alfarhood S, Safran M. Classification of Oral Cancer Into Pre-Cancerous Stages From White Light Images Using LightGBM Algorithm. IEEE Access. 2024;12:31626–39. https://doi.org/10.1109/ACCESS.2024.3370157

41. Kirchgaessner R, Watson C, Creason A, Keutler K, Goecks J. Imputing single-cell protein abundance in multiplex tissue imaging. Nat Commun. 2025;16:4747. https://doi.org/10.1038/s41467-025-59788-x

42. Wen B, Zeng WF, Liao Y, Shi Z, Savage SR, Jiang W, et al. Deep Learning in Proteomics. Proteomics. Wiley-VCH Verlag; 2020. https://doi.org/10.1002/pmic.201900335

43. Zhu W, Xie L, Han J, Guo X. The application of deep learning in cancer prognosis prediction. Cancers (Basel). MDPI AG; 2020. https://doi.org/10.3390/cancers12030603

44. Francisco-Cruz A, Parra ER, Tetzlaff MT, Wistuba II. Multiplex Immunofluorescence Assays. Methods in Molecular Biology. Humana Press Inc.; 2020. p. 467–95. https://doi.org/10.1007/978-1-4939-9773-2_22

45. Sheng W, Zhang C, Mohiuddin TM, Al-Rawe M, Zeppernick F, Falcone FH, et al. Multiplex Immunofluorescence: A Powerful Tool in Cancer Immunotherapy. Int J Mol Sci [Internet]. 2023;24:3086. https://doi.org/10.3390/ijms24043086

46. Neumann EK, Patterson NH, Rivera ES, Allen JL, Brewer M, deCaestecker MP, et al. Highly multiplexed immunofluorescence of the human kidney using co-detection by indexing. Kidney Int. Elsevier; 2022;101:137–43. https://doi.org/10.1016/J.KINT.2021.08.033

47. Werlein C, Ackermann M, Stark H, Shah HR, Tzankov A, Haslbauer JD, et al. Inflammation and vascular remodeling in COVID-19 hearts. Angiogenesis. Springer Science and Business Media B.V.; 2023;26:233–48. https://doi.org/10.1007/s10456-022-09860-7

48. Lewis SM, Asselin-Labat ML, Nguyen Q, Berthelet J, Tan X, Wimmer VC, et al. Spatial omics and multiplexed imaging to explore cancer biology. Nature Methods 2021 18:9 [Internet]. Nature Publishing Group; 2021 [cited 2023 Jul 10];18:997–1012. https://doi.org/10.1038/s41592-021-01203-6

49. Sepe JJ, Gardner RT, Blake MR, Brooks DM, Staffenson MA, Betts CB, et al. Therapeutics That Promote Sympathetic Reinnervation Modulate the Inflammatory Response After Myocardial Infarction VISUAL ABSTRACT. 2022 [cited 2023 Oct 1]; https://doi.org/10.1016/j.jacbts.2022.04.009

50. Kitko CL, Arora M, DeFilipp Z, Abu Zaid M, Di Stasi A, Radojcic V, et al. Axatilimab for Chronic Graft-Versus-Host Disease After Failure of at Least Two Prior Systemic Therapies:

Results of a Phase I/II Study. J Clin Oncol [Internet]. 2022;41:1864–75. https://doi.org/10.1200/JCO.22

51. McCaffrey EF, Donato M, Keren L, Chen Z, Delmastro A, Fitzpatrick MB, et al. The immunoregulatory landscape of human tuberculosis granulomas. Nat Immunol. Nature Research; 2022;23:318–29. https://doi.org/10.1038/s41590-021-01121-x

52. Sheng W, Zhang C, Mohiuddin TM, Al-Rawe M, Zeppernick F, Falcone FH, et al. Multiplex Immunofluorescence: A Powerful Tool in Cancer Immunotherapy. Int J Mol Sci. 2023;24:3086. https://doi.org/10.3390/ijms24043086

53. Black S, Phillips D, Hickey JW, Kennedy-Darling J, Venkataraaman VG, Samusik N, et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. Nat Protoc. Nature Research; 2021. p. 3802–35. https://doi.org/10.1038/s41596-021-00556-8

54. Lewis ZR, Phan-Everson T, Geiss G, Korukonda M, Bhatt R, Brown C, et al. Subcellular characterization of over 100 proteins in FFPE tumor biopsies with CosMx Spatial Molecular Imager. Cancer Res. 2022;82:3878–3878. https://doi.org/10.1158/1538-7445.AM2022-3878

55. Janesick A, Shelansky R, Gottscho AD, Wagner F, Rouault M, Beliakoff G, et al. High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. 2022; https://doi.org/10.1101/2022.10.06.510405

56. Tsujikawa T, Kumar S, Borkar RN, Azimi V, Thibault G, Chang YH, et al. Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity

Associated with Poor Prognosis. Cell Rep. Elsevier B.V.; 2017;19:203–17. https://doi.org/10.1016/j.celrep.2017.03.037

57. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature. Nature Publishing Group; 2019. p. 187–92. https://doi.org/10.1038/s41586-019-1629-x

58. Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, et al. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell. Cell Press; 2020. p. 236–49. https://doi.org/10.1016/j.cell.2020.03.053

59. Tan WCC, Nerurkar SN, Cai HY, Ng HHM, Wu D, Wee YTF, et al. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. Cancer Commun. John Wiley and Sons Inc.; 2020. p. 135–53. https://doi.org/10.1002/cac2.12023

60. Blise KE, Sivagnanam S, Banik GL, Coussens LM, Goecks J. Single-cell spatial architectures associated with clinical outcome in head and neck squamous cell carcinoma. NPJ Precis Oncol. Nature Research; 2022;6. https://doi.org/10.1038/s41698-022-00253-z

61. Friebel E, Kapolou K, Unger S, Núñez NG, Utz S, Rushing EJ, et al. Single-Cell Mapping of Human Brain Cancer Reveals Tumor-Specific Instruction of Tissue-Invading Leukocytes. Cell. Cell Press; 2020;181:1626-1642.e20. https://doi.org/10.1016/j.cell.2020.04.055

62. Steele NG, Carpenter ES, Kemp SB, Sirihorachai VR, The S, Delrosario L, et al. Multimodal mapping of the tumor and peripheral blood immune landscape in human pancreatic cancer. Nat Cancer. Nature Research; 2020;1:1097–112. https://doi.org/10.1038/s43018-020-00121-4

63. Bollhagen A, Bodenmiller B. Highly Multiplexed Tissue Imaging in Precision Oncology and Translational Cancer Research. Cancer Discov. 2024. p. 2071–88. https://doi.org/10.1158/2159-8290.CD-23-1165

64. Fu T, Dai LJ, Wu SY, Xiao Y, Ma D, Jiang YZ, et al. Spatial architecture of the immune microenvironment orchestrates tumor immunity and therapeutic response. J Hematol Oncol. BioMed Central Ltd; 2021. https://doi.org/10.1186/s13045-021-01103-4

65. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning [Internet]. PMLR; 2021. p. 8748–63. https://proceedings.mlr.press/v139/radford21a.html

66. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-Shot Text-to-Image Generation. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning [Internet]. PMLR; 2021. p. 8821–31. https://proceedings.mlr.press/v139/ramesh21a.html

67. Pan JJ, Wang J, Li G. Vector Database Management Techniques and Systems. Proceedings of the ACM SIGMOD International Conference on Management of Data.

Association for Computing Machinery; 2024. p. 597–604. https://doi.org/10.1145/3626246.3654691

68. Han Y, Liu C, Wang P. A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge.

69. Taipalus T. Vector database management systems: Fundamental concepts, use-cases, and current challenges. Cogn Syst Res. Elsevier B.V.; 2024;85. https://doi.org/10.1016/j.cogsys.2024.101216

70. Topalian SL, Drake CG, Pardoll DM. Immune checkpoint blockade: A common denominator approach to cancer therapy. Cancer Cell. Cell Press; 2015. p. 450–61. https://doi.org/10.1016/j.ccell.2015.03.001

71. Sun Q, Hong Z, Zhang C, Wang L, Han Z, Ma D. Immune checkpoint therapy for solid tumours: clinical dilemmas and future trends. Signal Transduct Target Ther. Springer Nature; 2023. https://doi.org/10.1038/s41392-023-01522-4

72. Lopez-Beltran A, Cimadamore A, Blanca A, Massari F, Vau N, Scarpelli M, et al. Immune checkpoint inhibitors for the treatment of bladder cancer. Cancers (Basel). MDPI AG; 2021. p. 1–16. https://doi.org/10.3390/cancers13010131

73. Carlino MS, Larkin J, Long G V. Immune checkpoint inhibitors in melanoma. The Lancet. 2021;398:1002–14. https://doi.org/10.1016/S0140-6736(21)01206-X

74. Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, et al. Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. New England Journal of

Medicine. Massachusetts Medical Society; 2010;363:711–23. https://doi.org/10.1056/nejmoa1003466

75. Wolchok JD, Chiarion-Sileni V, Rutkowski P, Cowey CL, Schadendorf D, Wagstaff J, et al. Final, 10-Year Outcomes with Nivolumab plus Ipilimumab in Advanced Melanoma. New England Journal of Medicine [Internet]. 2025;392:11–22. https://doi.org/10.1056/NEJMoa2407417

76. Long G V., Carlino MS, McNeil C, Ribas A, Gaudy-Marqueste C, Schachter J, et al. Pembrolizumab versus ipilimumab for advanced melanoma: 10-year follow-up of the phase III KEYNOTE-006 study. Annals of Oncology. Elsevier Ltd; 2024; https://doi.org/10.1016/j.annonc.2024.08.2330

77. Rizzetto G, De Simoni E, Molinelli E, Offidani A, Simonetti O. Efficacy of Pembrolizumab in Advanced Melanoma: A Narrative Review. Int J Mol Sci. Multidisciplinary Digital Publishing Institute (MDPI); 2023. https://doi.org/10.3390/ijms241512383

78. Hamid O, Robert C, Daud A, Hodi FS, Hwu WJ, Kefford R, et al. Five-year survival outcomes for patients with advanced melanoma treated with pembrolizumab in KEYNOTE-001. Annals of Oncology. Oxford University Press; 2019;30:582–8. https://doi.org/10.1093/annonc/mdz011

79. Morad G, Helmink BA, Sharma P, Wargo JA. Hallmarks of response, resistance, and toxicity to immune checkpoint blockade. Cell. Elsevier B.V.; 2021. p. 5309–37. https://doi.org/10.1016/j.cell.2021.09.020

80. Schadendorf D, van Akkooi ACJ, Berking C, Griewank KG, Gutzmer R, Hauschild A, et al. Melanoma. The Lancet. 2018;392:971–84. https://doi.org/10.1016/S0140-6736(18)31559-9

81. Morrison C, Pabla S, Conroy JM, Nesline MK, Glenn ST, Dressman D, et al. Predicting response to checkpoint inhibitors in melanoma beyond PD-L1 and mutational burden. J Immunother Cancer. 2018;6:32. https://doi.org/10.1186/s40425-018-0344-8

82. Schapiro D, Sokolov A, Yapp C, Chen Y-A, Muhlich JL, Hess J, et al. MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. Nat Methods [Internet]. 2021; https://doi.org/10.1038/s41592-021-01308-y

83. Kharchenko P V., Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. Nature Publishing Group; 2014;11:740–2. https://doi.org/10.1038/nmeth.2967

84. Gong W, Kwak IY, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: Imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics. BioMed Central Ltd.; 2018;19. https://doi.org/10.1186/s12859-018-2226-y

85. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. Cell Press; 2018;174:716-729.e27. https://doi.org/10.1016/j.cell.2018.05.061

86. Tran D, Tran B, Nguyen H, Nguyen T. A novel method for single-cell data imputation using subspace regression. Sci Rep. Nature Research; 2022;12. https://doi.org/10.1038/s41598-022-06500-4

87. He Y, Yuan H, Wu C, Xie Z. DISC: A highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning. Genome Biol. BioMed Central; 2020;21. https://doi.org/10.1186/s13059-020-02083-3

88. Talwar D, Mongia A, Sengupta D, Majumdar A. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. Sci Rep. Nature Publishing Group; 2018;8. https://doi.org/10.1038/s41598-018-34688-x

89. Chen Y, Wang Y, Chen Y, Cheng Y, Wei Y, Li Y, et al. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. Nat Commun. Nature Research; 2022;13. https://doi.org/10.1038/s41467-022-34550-9

90. Xu C, Cai L, Gao J. An efficient scRNA-seq dropout imputation method using graph attention network. BMC Bioinformatics. BioMed Central Ltd; 2021;22. https://doi.org/10.1186/s12859-021-04493-x

91. Qiu YL, Zheng H, Gevaert O. Genomic data imputation with variational auto-encoders. Gigascience [Internet]. 2020;9. https://doi.org/10.1093/gigascience/giaa082

92. Patruno L, Maspero D, Craighero F, Angaroni F, Antoniotti M, Graudenzi A. A review of computational strategies for denoising and imputation of single-cell transcriptomic data. Brief Bioinform. Oxford University Press; 2021. https://doi.org/10.1093/bib/bbaa222

93. Sims Z, Chang YH. A Masked Image Modeling Approach to Cyclic Immunofluorescence (CyCIF) Panel Reduction and Marker Imputation. 2023; https://doi.org/10.1101/2023.05.10.540265

94. Ternes L, Dane M, Labrie M, Mills G, Gray J, Heiser L, et al. ME-VAE: Multi-Encoder Variational AutoEncoder for Controlling Multiple Transformational Features in Single Cell Image Analysis. 2021; https://doi.org/10.1101/2021.04.22.441005

95. Pati P, Karkampouna S, Bonollo F, Compérat E, Radic M, Spahn M, et al. Multiplexed tumor profiling with generative AI accelerates histopathology workflows and improves clinical predictions. 2023; https://doi.org/10.1101/2023.11.29.568996

96. Wu E, Trevino AE, Wu Z, Swanson K, Kim HJ, D'Angio HB, et al. 7-UP: Generating in silico CODEX from a small set of immunofluorescence markers. PNAS Nexus. National Academy of Sciences; 2023;2. https://doi.org/10.1093/pnasnexus/pgad171

97. Su X, Yan X, Tsai CL. Linear regression. Wiley Interdiscip Rev Comput Stat. 2012;4:275–94. https://doi.org/10.1002/wics.1198

98. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree [Internet]. 2017. https://github.com/Microsoft/LightGBM.

99. Zhai J, Zhang S, Chen J, He Q. Autoencoder and Its Various Variants. 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE; 2018. p. 415–9. https://doi.org/10.1109/SMC.2018.00080

100. Lotfollahi M, Dony L, Agarwala H, Theis F. Out-of-distribution prediction with disentangled representations for single-cell RNA sequencing data. 2020; https://doi.org/10.1101/2021.09.01.458535

101. Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biol. BioMed Central Ltd; 2020;21. https://doi.org/10.1186/s13059-020-02132-x

102. Grønbech CH, Vording MF, Timshel PN, Sønderby CK, Pers TH, Winther O. scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics [Internet]. Oxford Academic; 2020 [cited 2021 Nov 13];36:4415–22. https://doi.org/10.1093/BIOINFORMATICS/BTAA293

103. McCoy JT, Kroon S, Auret L. Variational Autoencoders for Missing Data Imputation with Application to a Simulated Milling Circuit. IFAC-PapersOnLine. 2018;51:141–6. https://doi.org/10.1016/j.ifacol.2018.09.406

104. Fischer DS, Schaar AC, Theis FJ. Modeling intercellular communication in tissues using spatial graphs of cells. Nat Biotechnol. Nature Research; 2022; https://doi.org/10.1038/s41587-022-01467-z

105. Gondara L, Wang K. MIDA: Multiple Imputation Using Denoising Autoencoders. 2018. p. 260–72. https://doi.org/10.1007/978-3-319-93040-4_21

106. Tran T, Le U, Shi Y. An effective up-sampling approach for breast cancer prediction with imbalanced data: A machine learning model-based comparative analysis. PLoS One. 2022;17:e0269135. https://doi.org/10.1371/journal.pone.0269135

107. Patruno L, Maspero D, Craighero F, Angaroni F, Antoniotti M, Graudenzi A. A review of computational strategies for denoising and imputation of single-cell transcriptomic data. Brief Bioinform. 2020; https://doi.org/10.1093/bib/bbaa222

108. De Biasi S, Lo Tartaro D, Neroni A, Rau M, Paschalidis N, Borella R, et al. Immunosenescence and vaccine efficacy revealed by immunometabolic analysis of SARS-CoV-2-specific cells in multiple sclerosis patients. Nat Commun. 2024;15:2752. https://doi.org/10.1038/s41467-024-47013-0

109. Cannoodt R, Saelens W, Deconinck L, Saeys Y. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. Nat Commun. 2021;12:3942. https://doi.org/10.1038/s41467-021-24152-2

110. Ren T, Chen C, Danilov A V., Liu S, Guan X, Du S, et al. Supervised learning of high-confidence phenotypic subpopulations from single-cell data. Nat Mach Intell. 2023;5:528–41. https://doi.org/10.1038/s42256-023-00656-y

111. Zheng W, Min W, Wang S. TsImpute: an accurate two-step imputation method for single-cell RNA-seq data. Bioinformatics. 2023;39. https://doi.org/10.1093/bioinformatics/btad731

112. Sun ED, Ma R, Zou J. SPRITE: improving spatial gene expression imputation with gene and cell networks. Bioinformatics. 2024;40:i521–8. https://doi.org/10.1093/bioinformatics/btae253

113. Molino P, Dudin Y, Miryala SS. Ludwig: a type-based declarative deep learning toolbox. 2019; http://arxiv.org/abs/1909.07930

114. Waqas A, Tripathi A, Ramachandran RP, Stewart PA, Rasool G. Multimodal data integration for oncology in the era of deep neural networks: a review. Front Artif Intell. Frontiers Media SA; 2024;7:1408843.

115. Madan S, Lentzen M, Brandt J, Rueckert D, Hofmann-Apitius M, Fröhlich H. Transformer models in biomedicine. BMC Med Inform Decis Mak. Springer Science and Business Media LLC; 2024;24:214.

116. Gema AP, Grabarczyk D, De Wulf W, Borole P, Alfaro JA, Minervini Pasquale and Vergari A, et al. Knowledge graph embeddings in the biomedical domain: are they useful? A look at link prediction, rule learning, and downstream polypharmacy tasks. Bioinform Adv. Oxford University Press (OUP); 2024;4:vbae097.

117. Makarov I, Kiselev D, Nikitinsky N, Subelj L. Survey on graph embeddings and their applications to machine learning problems on graphs. PeerJ Comput Sci. PeerJ Inc.; 2021;7:1–62. https://doi.org/10.7717/peerj-cs.357

118. Xu M. Understanding graph embedding methods and their applications. SIAM Rev Soc Ind Appl Math. Society for Industrial & Applied Mathematics (SIAM); 2021;63:825–53.

119. Ma F, Xue H, Wang G, Zhou Y, Rao F, Yan S, et al. Multi-Modal Generative Embedding Model. arXiv [csCV]. 2024;

120. Refahi M, Sokhansanj BA, Mell Joshua C and Brown JR, Yoo H, Hearne G, Rosen GL. Enhancing nucleotide sequence representations in genomic analysis with contrastive optimization. Commun Biol. Springer Science and Business Media LLC; 2025;8:517.

121. Wang C, Xu H, Zhang X, Wang L, Zheng Z, Liu H. Convolutional embedding makes hierarchical vision transformer stronger. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2022. p. 739–56.

122. Jush FK, Truong T, Vogler S, Lenga M. Medical image retrieval using pretrained embeddings. 2024 IEEE International Symposium on Biomedical Imaging (ISBI). IEEE; 2024. p. 1–5.

123. Eijpe A, Lakbir S, Cesur ME, Oliveira SP, Abeln S, Silva W. Disentangled and Interpretable Multimodal Attention Fusion for cancer survival prediction. arXiv [csCV]. 2025;

124. Mikolov T, Yih W-T, Zweig G. Linguistic regularities in continuous space word representations. North Am Chapter Assoc Comput Linguistics. 2013;746–51.

125. Winchester DP, Stewart AK, Phillips JL, Ward EE. The national cancer data base: past, present, and future. Ann Surg Oncol. Springer Science and Business Media LLC; 2010;17:4–7.

126. Lee J-S. Exploring cancer genomic data from the cancer genome atlas project. BMB Rep. Korean Society for Biochemistry and Molecular Biology - BMB Reports; 2016;49:607–11.

127. Thennavan A, Beca F, Xia Y, Recio SG, Allison K, Collins LC, et al. Molecular analysis of TCGA breast cancer histologic types. Cell Genom. Elsevier BV; 2021;1:100067.

128. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. arXiv [csCV]. 2021;

129. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-to-image generation. arXiv [csCV]. 2021;

130. Talati N, Jin D, Ye H, Brahmakshatriya A, Dasika G, Amarasinghe S, et al. A deep dive into understanding the random walk-based temporal graph learning. 2021 IEEE International Symposium on Workload Characterization (IISWC). IEEE; 2021. p. 87–100.

131. Nikolentzos G, Vazirgiannis M. Random walk graph neural networks. Neural Inf Process Syst. 2020;33.

132. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. Springer Science and Business Media LLC; 2021;14:13.

133. Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. PLoS One. Public Library of Science (PLoS); 2012;7:e41882.

134. Malkov YA, Yashunin DA. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. IEEE Trans Pattern Anal Mach Intell. 2020;42:824–36. https://doi.org/10.1109/TPAMI.2018.2889473

135. Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv [csCL]. 2019;

136. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, et al. A whole-slide foundation model for digital pathology from real-world data. Nature. Springer Science and Business Media LLC; 2024;630:181–8.

137. Zhou C, Chen S, Xu F, Wei J, Zhou X, Wu Z, et al. Estimating tumor mutational burden across multiple cancer types using whole-exome sequencing. Ann Transl Med. AME Publishing Company; 2021;9:1437.

138. Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets.

139. Davis AA, Patel VG. The role of PD-L1 expression as a predictive biomarker: an analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors. J Immunother Cancer. 2019;7:278. https://doi.org/10.1186/s40425-019-0768-9

140. Madore J, Vilain RE, Menzies AM, Kakavand H, Wilmott JS, Hyman J, et al. PD-L1 expression in melanoma shows marked heterogeneity within and between patients: implications for anti-PD-1/PD-L 1 clinical trials. Pigment Cell Melanoma Res. 2015;28:245–53. https://doi.org/10.1111/pcmr.12340

141. Mandalà M, Merelli B, Massi D. PD-L1 in Melanoma: Facts and Myths. Melanoma Manag. 2016;3:187–94. https://doi.org/10.2217/mmt-2016-0013

142. Maher NG, Vergara IA, Long G V., Scolyer RA. Prognostic and predictive biomarkers in melanoma. Pathology. 2024;56:259–73. https://doi.org/10.1016/j.pathol.2023.11.004

143. Ma W, Liu W, Zhong J, Zou Z, Lin X, Sun W, et al. Advances in predictive biomarkers for melanoma immunotherapy. Holistic Integrative Oncology. 2024;3:48. https://doi.org/10.1007/s44178-024-00121-9

144. Marcus L, Fashoyin-Aje LA, Donoghue M, Yuan M, Rodriguez L, Gallagher PS, et al. FDA Approval Summary: Pembrolizumab for the Treatment of Tumor Mutational Burden–High Solid Tumors. Clinical Cancer Research. 2021;27:4685–9. https://doi.org/10.1158/1078-0432.CCR-21-0327

145. de Visser KE, Joyce JA. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. Cancer Cell. 2023;41:374–403. https://doi.org/10.1016/j.ccell.2023.02.016

146. Topalian SL, Taube JM, Anders RA, Pardoll DM. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. Nat Rev Cancer. 2016;16:275–87. https://doi.org/10.1038/nrc.2016.36

147. Nishino M, Ramaiya NH, Hatabu H, Hodi FS. Monitoring immune-checkpoint blockade: response evaluation and biomarker development. Nat Rev Clin Oncol. 2017;14:655–68. https://doi.org/10.1038/nrclinonc.2017.88

148. Chang T-G, Cao Y, Sfreddo HJ, Dhruba SR, Lee S-H, Valero C, et al. LORIS robustly predicts patient outcomes with immune checkpoint blockade therapy using common clinical, pathologic and genomic features. Nat Cancer. 2024;5:1158–75. https://doi.org/10.1038/s43018-024-00772-7

149. Valenti F, Falcone I, Ungania S, Desiderio F, Giacomini P, Bazzichetto C, et al. Precision Medicine and Melanoma: Multi-Omics Approaches to Monitoring the Immunotherapy Response. Int J Mol Sci. 2021;22:3837. https://doi.org/10.3390/ijms22083837

150. Quail DF, Walsh LA. Revolutionizing cancer research with spatial proteomics and visual intelligence. Nat Methods. 2024;21:2216–9. https://doi.org/10.1038/s41592-024-02542-w

151. Williams HL, Frei AL, Koessler T, Berger MD, Dawson H, Michielin O, et al. The current landscape of spatial biomarkers for prediction of response to immune checkpoint inhibition. NPJ Precis Oncol. 2024;8:178. https://doi.org/10.1038/s41698-024-00671-1

152. Toki MI, Merritt CR, Wong PF, Smithy JW, Kluger HM, Syrigos KN, et al. High-Plex Predictive Marker Discovery for Melanoma Immunotherapy–Treated Patients Using Digital Spatial Profiling. Clinical Cancer Research. 2019;25:5503–12. https://doi.org/10.1158/1078-0432.CCR-19-0104

153. Antoranz A, Van Herck Y, Bolognesi MM, Lynch SM, Rahman A, Gallagher WM, et al. Mapping the Immune Landscape in Metastatic Melanoma Reveals Localized Cell–Cell Interactions That Predict Immunotherapy Response. Cancer Res. 2022;82:3275–90. https://doi.org/10.1158/0008-5472.CAN-22-0363

154. Kim S, Kim JR, Lee JH, Moon S-H, In Jo S, Bae D-J, et al. Differential RNA expression of immune-related genes and tumor cell proximity from intratumoral M1 macrophages in acral lentiginous melanomas treated with PD-1 blockade. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2022;1868:166516. https://doi.org/10.1016/j.bbadis.2022.166516

155. Attrill GH, Owen CN, Ahmed T, Vergara IA, Colebatch AJ, Conway JW, et al. Higher proportions of CD39+ tumor-resident cytotoxic T cells predict recurrence-free survival in patients with stage III melanoma treated with adjuvant immunotherapy. J Immunother Cancer. 2022;10:e004771. https://doi.org/10.1136/jitc-2022-004771

156. Johnson DB, Bordeaux J, Kim JY, Vaupel C, Rimm DL, Ho TH, et al. Quantitative Spatial Profiling of PD-1/PD-L1 Interaction and HLA-DR/IDO-1 Predicts Improved Outcomes of Anti–PD-1 Therapies in Metastatic Melanoma. Clinical Cancer Research. 2018;24:5250–60. https://doi.org/10.1158/1078-0432.CCR-18-0309

157. Mi H, Sivagnanam S, Ho WJ, Zhang S, Bergman D, Deshpande A, et al. Computational methods and biomarker discovery strategies for spatial proteomics: a review in immuno-oncology. Brief Bioinform. 2024;25. https://doi.org/10.1093/bib/bbae421

158. Mou M, Pan Z, Lu M, Sun H, Wang Y, Luo Y, et al. Application of Machine Learning in Spatial Proteomics. J Chem Inf Model. 2022;62:5875–95. https://doi.org/10.1021/acs.jcim.2c01161

159. Risom T, Glass DR, Averbukh I, Liu CC, Baranski A, Kagel A, et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. Cell. 2022;185:299-310.e18. https://doi.org/10.1016/j.cell.2021.12.023

160. Blise KE, Sivagnanam S, Betts CB, Betre K, Kirchberger N, Tate BJ, et al. Machine Learning Links T-cell Function and Spatial Localization to Neoadjuvant Immunotherapy and Clinical Outcome in Pancreatic Cancer. Cancer Immunol Res. 2024;12:544–58. https://doi.org/10.1158/2326-6066.CIR-23-0873

161. McNamara KL, Caswell-Jin JL, Joshi R, Ma Z, Kotler E, Bean GR, et al. Spatial proteomic characterization of HER2-positive breast tumors through neoadjuvant therapy predicts response. Nat Cancer. 2021;2:400–13. https://doi.org/10.1038/s43018-021-00190-z

162. Jimenez J, Dubey P, Carter B, Koomen JM, Markowitz J. A metabolic perspective on nitric oxide function in melanoma. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2024;1879:189038. https://doi.org/10.1016/j.bbcan.2023.189038

163. Schrom EC, McCaffrey EF, Sreejithkumar V, Radtke AJ, Ichise H, Arroyo-Mejias A, et al. Spatial Patterning Analysis of Cellular Ensembles (SPACE) discovers complex spatial organization at the cell and tissue levels. 2023. https://doi.org/10.1101/2023.12.08.570837

164. Maus RLG, Leontovich AA, Moore RM, Becher L, Nevala WK, Flotte TJ, et al. Resolving the Heterogeneous Tumor-Centric Cellular Neighborhood through Multiplexed, Spatial Paracrine Interactions in the Setting of Immune Checkpoint Blockade. Cancer Research Communications. 2022;2:78–89. https://doi.org/10.1158/2767-9764.CRC-21-0146

165. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 3149–57.

166. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 4768–77.

167. Fattore L, Ruggiero CF, Liguoro D, Mancini R, Ciliberto G. Single cell analysis to dissect molecular heterogeneity and disease evolution in metastatic melanoma. Cell Death Dis. 2019;10:827. https://doi.org/10.1038/s41419-019-2048-5

168. Zhang C, Shen H, Yang T, Li T, Liu X, Wang J, et al. A single-cell analysis reveals tumor heterogeneity and immune environment of acral melanoma. Nat Commun. 2022;13:7250. https://doi.org/10.1038/s41467-022-34877-3

169. Grzywa TM, Paskal W, Włodarski PK. Intratumor and Intertumor Heterogeneity in Melanoma. Transl Oncol. 2017;10:956–75. https://doi.org/10.1016/j.tranon.2017.09.007

170. Vargas GM, Shafique N, Xu X, Karakousis G. Tumor-infiltrating lymphocytes as a prognostic and predictive factor for Melanoma. Expert Rev Mol Diagn. 2024;24:299–310. https://doi.org/10.1080/14737159.2024.2312102

171. Straker RJ, Krupp K, Sharon CE, Thaler AS, Kelly NJ, Chu EY, et al. Prognostic Significance of Primary Tumor-Infiltrating Lymphocytes in a Contemporary Melanoma Cohort. Ann Surg Oncol. 2022;29:5207–16. https://doi.org/10.1245/s10434-022-11478-4

172. Wang X, Lamberti G, Di Federico A, Alessi J, Ferrara R, Sholl ML, et al. Tumor mutational burden for the prediction of PD-(L)1 blockade efficacy in cancer: challenges and opportunities. Annals of Oncology. 2024;35:508–22. https://doi.org/10.1016/j.annonc.2024.03.007

173. Huang AC, Zappasodi R. A decade of checkpoint blockade immunotherapy in melanoma: understanding the molecular basis for immune sensitivity and resistance. Nat Immunol. 2022;23:660–70. https://doi.org/10.1038/s41590-022-01141-1

174. Garg SK, Sun J, Kim Y, Whiting J, Sarnaik A, Conejo-Garcia JR, et al. Dichotomous Nitric Oxide–Dependent Post-Translational Modifications of STAT1 Are Associated with Ipilimumab Benefits in Melanoma. Cancers (Basel). 2023;15:1755. https://doi.org/10.3390/cancers15061755

175. Garg SK, Welsh EA, Fang B, Hernandez YI, Rose T, Gray J, et al. Multi-Omics and Informatics Analysis of FFPE Tissues Derived from Melanoma Patients with Long/Short Responses to Anti-PD1 Therapy Reveals Pathways of Response. Cancers (Basel). 2020;12:3515. https://doi.org/10.3390/cancers12123515

176. Garg SK, Ott MJ, Mostofa AGM, Chen Z, Chen YA, Kroeger J, et al. Multi-Dimensional Flow Cytometry Analyses Reveal a Dichotomous Role for Nitric Oxide in Melanoma

Patients Receiving Immunotherapy. Front Immunol. 2020;11. https://doi.org/10.3389/fimmu.2020.00164

177. Tanese K, Grimm EA, Ekmekcioglu S. The role of melanoma tumor-derived nitric oxide in the tumor inflammatory microenvironment: Its impact on the chemokine expression profile, including suppression of CXCL10. Int J Cancer. 2012;131:891–901. https://doi.org/10.1002/ijc.26451

178. Yarlagadda K, Hassani J, Foote IP, Markowitz J. The role of nitric oxide in melanoma. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2017;1868:500–9. https://doi.org/10.1016/j.bbcan.2017.09.005

179. Ekmekcioglu S, Ellerhorst J, Smid CM, Prieto VG, Munsell M, Buzaid AC, et al. Inducible nitric oxide synthase and nitrotyrosine in human metastatic melanoma tumors correlate with poor survival. Clin Cancer Res. 2000;6:4768–75.

180. Baddeley A, Rubak E, Turner R. Spatial Point Patterns Methodology and Applications with R. 2016.

181. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4:1686. https://doi.org/10.21105/joss.01686

182. Truhn D, Eckardt J-N, Ferber D, Kather JN. Large language models and multimodal foundation models for precision oncology. NPJ Precis Oncol. 2024;8:72. https://doi.org/10.1038/s41698-024-00573-2

183. Xiang J, Wang X, Zhang X, Xi Y, Eweje F, Chen Y, et al. A vision–language foundation model for precision oncology. Nature. 2025;638:769–78. https://doi.org/10.1038/s41586-024-08378-w

184. Pai S, Bontempi D, Hadzic I, Prudente V, Sokač M, Chaunzwa TL, et al. Foundation model for cancer imaging biomarkers. Nat Mach Intell. 2024;6:354–67. https://doi.org/10.1038/s42256-024-00807-9

185. Skourti E. Foundation models in clinical oncology. Nat Cancer. 2024;5:1790–1790. https://doi.org/10.1038/s43018-024-00837-7

186. Gogoshin G, Rodin AS. Graph Neural Networks in Cancer and Oncology Research: Emerging and Future Trends. Cancers (Basel). 2023;15:5858. https://doi.org/10.3390/cancers15245858

187. Waqas A, Tripathi A, Naeini M, Stewart P, Schabath MB, Rasool G. Using Patient Embeddings From Foundation Models for Enhanced Survival Analysis in Lung Squamous Cell Carcinoma. Am J Respir Crit Care Med. 2025;211:A3108–A3108. https://doi.org/10.1164/ajrccm.2025.211.Abstracts.A3108

188. Baciu-Drăgan M-A, Beerenwinkel N. Oncotree2vec — a method for embedding and clustering of tumor mutation trees. Bioinformatics. 2024;40:i180–8. https://doi.org/10.1093/bioinformatics/btae214

189. Baker EAG, Schapiro D, Dumitrascu B, Vickovic S, Regev A. In silico tissue generation and power analysis for spatial omics. Nat Methods. Nature Research; 2023;20:424–31. https://doi.org/10.1038/s41592-023-01766-6