Enhancing Biomedical Imaging with AI: Compression, Prediction, and Multi-Modal Integration for Clinical Advancement

Mohammad Sadegh Nasr

Enhancing Biomedical Imaging with AI: Compression, Prediction, and Multi-Modal Integration for Clinical Advancement

A DISSERTATION PRESENTED
BY
MOHAMMAD SADEGH NASR
TO
THE FACULTY OF THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
COMPUTER AND INFORMATION SCIENCES

THE UNIVERSITY OF TEXAS AT ARLINGTON
ARLINGTON, TEXAS

DECEMBER, 2023

Enhancing Biomedical Imaging with AI: Compression, Prediction, and Multi-Modal Integration for Clinical Advancement

Supervising Professors: Prof. Manfred Huber, Dr. Jacob Luber

Mohammad Sadegh Nasr

Abstract

This dissertation delves into the enhancement of biomedical image analysis through the deployment of artificial intelligence methodologies, focusing on the transition from theoretical innovation to practical clinical utility. Spanning four cornerstone projects, the work encapsulates the development of predictive models for spatial transcriptomics, efficient image compression for cancer pathology slides, and critical evaluations of histopathology slide search engines. The first project employs Random Forest Regression and spatial point processes to forecast cell distribution patterns, thereby offering a novel perspective on gene expression in embryogenesis at a single-molecule resolution. The second venture introduces a Variational Autoencoder (VAE) that sets a new precedent in histopathology imaging with a significant compression ratio, maintaining diagnostic reliability. Lastly, the third project assesses the performance of leading histopathology slide search engines, establishing a benchmark for their clinical application and suggesting enhancements for future integration. Together, these projects pave the way for AI-driven approaches to be woven into the fabric of clinical practice, signaling a transformative leap in the utility of biomedical imaging and multi-channel data interpretation

Acknowledgments

As I stand at the culmination of my PhD journey, I am imbued with profound gratitude and deep appreciation for those who have been my pillars of support, sources of inspiration, and constant companions in this endeavor.

I extend my deepest thanks to my PhD advisors, Dr. Manfred Huber and Dr. Jacob Luber. Dr. Huber, your selflessness, kindness, and vast knowledge have been invaluable, while Dr. Luber, you've been not just an excellent advisor but also a true friend, consistently providing the means for my success. I am also grateful to my considerate committee members, Dr. Vassilis Athitsos, Dr. Farhad Kamangar, and Dr. Deokgun Park, for your guidance and support. Dr. Marion Ball and Dr. Gabriela Wilson of MICHI, your love, care, and support have been extraordinary. Dr. Bahram Khalili and Ms. Ginger Dickens, my graduate advisors, your assistance in navigating educational challenges and meeting my needs has been invaluable. Ms. Sherry Gother, your diligence in managing my administrative needs has been greatly appreciated.

My colleagues at Luberlab – Jitenga, Thomas, Helen, Michael, Jillur Rahman, Parisa, Amir, Jai, Anahita, Paul, Lucy, and Cody – your collaboration and friendship have made every project a memorable and enriching experience. Solmaz Shariat Torbaghan, my manager at Varian Medical Systems, your mentorship and support have been vital to my growth. To my colleagues at Varian Medical Systems – Parin, Salar, Pouya, Ricky, and Oskar – your insights and camaraderie have been invaluable. My internship at MD Anderson Cancer Center, under the mentorship of Dr. Jia Wu and Dr. Pingjun Chen, was an enriching experience. Dr. Senthilkumar Damodaran, your guidance through my research challenges has been a beacon of light.

I am profoundly grateful to Dr. Seyyed Mohammad Bagher Malaek, my master's degree advisor, who profoundly shaped my character and taught me critical thinking. Dr. Ali Abedian and Dr. Kaveh Ghorbanian, my college professors, your wisdom has always steered me towards excellence. Mostafa Malekian, my spiritual guide and confidant, your presence in my life is an immeasurable gift. Eng. Khalil Davari, I am forever thankful for your wise and beneficial life lessons. Reza Khalou, my high school physics teacher, your impact on my personality and worldview continues to resonate with me.

To my dearest friends since college, Soheil and Motaharreh, your comforting presence has been a source of solace and strength. To my college friends, Sina and Farshid, your friendship has been invaluable. Dr. Ehsan Irajizad and Dr. Faezeh Darbaniyan, your hospitality, advice, and unwavering support in times of trouble have been deeply appreciated. Armin, Hamidreza, and your families, your trustworthy and selfless friendship has been a treasure. To my cherished friends – Bahram and Shirin, Mohsen and Nadia, Farnaz and Armin, Hossein and Maygol, Negin, and Nima and Zahra – your support from the first day of my doctorate studies has been invaluable.

Sima Amin and her family – Esmat, Masoud, Seyamak, Romina, Shabnam, and Shadi – thank you for your unwavering support and for making your home a haven. Nima Mirpour, along with her children Doctors Saeed and Fahimeh Sasan, your open hearts and homes, immense help, and advice have been a constant support. Hassan and Mohammad Samini, knowing you throughout my life has been a fortune. Mohammad, your selflessness and modesty are exemplary. Hassan, your support in my education abroad has been crucial to my journey. Eng. Seiavash Mir and your lovely family, thank you for always being there for us.

Last but not definitely the least, my deepest thanks go to my beloved wife, Bahar. Your sacrifices, unwavering support, and love have been the bedrock of my journey. Your presence has been a source of strength and joy, making every challenge surmountable and every achievement sweeter. My parents, Marziyeh and Ali, to whom I owe an immeasurable debt of gratitude. Your sacrifices, unlimited care, love, and support have shaped me into the person I am today. I cannot thank you enough. Zahra and Mohammadreza, my dear mother-in-law and father-in-law, your belief in me and constant support and kindness have always been a powerful force. Sadra, my brother-in-law, thank you for always bringing joy into our lives and for enlightening me with your knowledge of biological concepts. Dr. Hamidreza Safikhani, thank you for your professional and emotional support. A special thanks to my aunt Hamideh for her constant care. My gratitude extends to all my family members, including my grandmothers Masoumeh and Fatemeh, my late grandfather Mohammadreza, my wife's grandmothers Masoumeh and Homa, my aunt Akram, my uncles Mohammad and Hassan, and all their families.

To all of you, I am eternally grateful. Your unique roles in my journey have made this achievement as much yours as it is mine. Thank you for being part of my story.

I dedicate this dissertation to my wife and parents: Bahar, Marziyeh, and Ali.

Contents

Listing of figures

1

Introduction and Overview

At the heart of this dissertation lies the exploration of artificial intelligence's transformative role in biomedical imaging, a domain where innovation, technology, and medical science converge to redefine the boundaries of diagnosis, prognosis, and treatment strategies. This introductory chapter sets the stage for a comprehensive journey through a series of studies that collectively aim to enhance the analysis and prediction capabilities in biomedical imaging. Each paper within this dissertation not only stands as an individual testament to the advancements in the field but also contributes to a cohesive narrative that addresses critical challenges in spatial transcriptomics, cancer pathology, and the integration of multi-modal data. Here, we delineate the common problems these studies seek to solve, the unique contributions of the dissertation writer and co-authors, and the context within which this research is situated in the broader landscape of medical imaging and AI. By providing a synthesis of the existing literature and the gaps our research aims to fill, this chapter offers a roadmap for the insights and innovations that unfold in the subsequent sections, encapsulating the essence of the dissertation's contribution to the ever-evolving field of biomedical imaging.

1.1 Introduction to the Dissertation Theme

The advent of artificial intelligence (AI) and its integration into biomedical imaging has marked a transformative era in medical diagnostics, research, and treatment strategies. Biomedical imaging stands at the forefront of modern clinical practice, offering a window into the complex workings of the human body. Its evolution, particularly through the integration of AI, has been nothing short of revolutionary, reshaping how clinicians approach diagnosis, treatment planning, and patient monitoring. The application of AI in biomedical imaging has led to significant improvements in image quality, diagnostic accuracy, and the efficiency of image analysis.

1.1.1 Overview of AI in Biomedical Imaging

AI algorithms have been instrumental in enhancing the resolution and clarity of medical images, allowing for more precise identification of pathological changes in tissues. Technologies like deep learning have shown exceptional promise in identifying patterns in imaging data that are often subtle and complex, well beyond the capabilities of traditional imaging techniques[45,30]. This advancement is pivotal in fields such as radiology and pathology, where accurate image interpretation is critical for diagnosing diseases like cancer, neurological disorders, and cardiovascular diseases[16,49].

Moreover, AI-driven tools have streamlined the process of medical imaging, reducing the time required for image analysis and interpretation. This efficiency is particularly beneficial in high-volume clinical settings, where rapid and accurate analysis of a large number of images is crucial for patient care[44,50]. AI's capacity to process and analyze vast datasets rapidly not only aids in early disease detection but also enables personalized medicine by facilitating the identification of disease subtypes and the prediction of treatment responses[39,41].

The transformative impact of AI on biomedical imaging is also evident in its role in advancing non-invasive diagnostic techniques. For example, AI has played a significant role in the development and refinement of techniques such as magnetic resonance imaging (MRI) and computed tomography (CT), enabling more detailed and accurate visualization of internal structures without the need for invasive procedures[12,27]. This advancement not only enhances patient comfort but also reduces the risks associated with invasive diagnostic methods.

Biomedical imaging also plays a crucial role in spatial transcriptomics, a rapidly emerging field that combines gene expression data with spatial context. This integration allows for the visualization and analysis of the spatial distribution of transcripts within tissue sections, providing insights into the complex cellular architecture of tissues and the spatial heterogeneity of gene expression[47,37]. In cancer pathology, biomedical imaging serves as an indispensable tool for detecting and characterizing tumors. The detailed images obtained through advanced imaging techniques facilitate the identification of cancerous cells and structures, enabling pathologists to make accurate diagnoses and tailor treatment strategies to individual patients[13,31].

In the realm of histopathology slide analysis, biomedical imaging is pivotal for the examination of tissue samples at the microscopic level. AI-enhanced imaging techniques have significantly improved the accuracy and efficiency of detecting pathological changes, such as the presence of tumor cells or the assessment of tumor margins[17,10]. These advancements not only accelerate the diagnostic process but also ensure greater consistency and precision in histopathological evaluations.

The convergence of biomedical imaging with AI technologies in these areas represents a paradigm shift, offering opportunities to enhance our understanding of complex biological processes and diseases. It enables the medical community to move beyond traditional diagnostic methods and embrace more precise, personalized approaches to patient care.

1.1.2 Challenges in the Field

While AI in biomedical imaging has brought about significant advancements, the field also faces several challenges that need to be addressed to fully realize its potential. One of the primary challenges is the need for large, diverse, and high-quality datasets for training AI models. The accuracy and robustness of AI algorithms heavily depend on the volume and variety of data they are trained on. However, acquiring such extensive datasets is often hindered by issues related to patient privacy, data sharing regulations, and the inherent variability in medical imaging due to different equipment and protocols across institutions[36].

Another significant challenge is the interpretability and explainability of AI models in medical imaging. Many advanced AI algorithms, particularly deep learning models, operate as 'black boxes,' making it difficult for clinicians to understand how these models arrive at a particular diagnosis or prediction. This lack of transparency can lead to trust issues and reluctance in adopting AI-driven diagnostic tools in clinical practice[15,4].

Furthermore, integrating AI into clinical workflows poses its own set of challenges. The healthcare industry often faces barriers in terms of infrastructure, funding, and technical expertise required to implement and maintain AI solutions effectively. Additionally, there is a need for significant training and adaptation among healthcare professionals to efficiently utilize AI-enhanced tools in their daily practice[1].

Lastly, addressing the ethical and legal considerations surrounding AI in healthcare is crucial. This includes ensuring patient confidentiality, addressing potential biases in AI algorithms, and navigating the legal implications of AI-driven medical decisions[11,5].

1.1.3 Advancing Predictive and Analytical Techniques in Biomedical Imaging

This dissertation tries to address some of these challenges by delving into the nuances of biomedical imaging in areas like spatial transcriptomics, cancer pathology, and histopathology slide analysis, underscoring the pivotal role of AI in enhancing predictive accuracy and optimizing image compression for comprehensive insights into complexity of cancer.

The significance of AI extends beyond mere image analysis, venturing into the realm of predictive modeling, a facet crucial for understanding disease progression and therapeutic outcomes. Predictive models in spatial transcriptomics, for instance, are essential for deciphering the spatial distribution of gene expression within tissues, offering invaluable insights into cellular interactions and functions in a spatial context[43]. These models aid in interpreting the complex spatial arrangements of cells, which are pivotal for understanding various biological processes and disease pathologies, including cancer development.

Furthermore, the integration of AI in image compression, particularly for cancer pathology slides, addresses the significant challenge of managing and analyzing the enormous datasets typical in digital pathology[35,22]. High-resolution whole-slide images (WSIs) of pathological samples generate vast amounts of data, necessitating efficient compression techniques to facilitate storage, transmission, and analysis. Variational Autoencoders (VAEs), for example, have emerged as a promising solution, offering a balance between compression efficiency and image reconstruction fidelity[21,48].

Finally, histopathology slide analysis, another cornerstone of modern diagnostics, benefits immensely from AI-driven search engines and analytical tools. These tools not only expedite the diagnostic process but also enhance the accuracy and reproducibility of histopathological assessments[30,3]. The incorporation of machine learning algorithms in the analysis of histopathology slides enables the identification of subtle patterns and features that might be overlooked by the human eye, thereby supporting more accurate diagnoses and prognostic evaluations.

In conclusion, the application of AI in biomedical imaging is a rapidly evolving field, with significant implications for medical research and clinical practice. This dissertation aims to contribute to this field by exploring and advancing AI techniques in predictive modeling, image compression, and histopathology slide analysis.

1.2 Overview of Included Papers

In this dissertation, we explore a series of studies that collectively enhance the analysis and prediction capabilities in biomedical imaging. Each paper contributes unique insights and methodologies, addressing different yet interconnected aspects of this vast field. From the intricate analysis of spatial transcriptomics to the nuanced evaluation of cancer pathology, these papers collectively push the boundaries of what is achievable through the application of AI in biomedical imaging. This section provides a succinct overview of each paper, highlighting their individual contributions while weaving a common thread that underscores their collective impact in advancing the field.

1.2.1 Paper 1: Predicting the Future States of Gene Expression

The first paper in our exploration, "Predicting Future States with Spatial Point Processes in Single Molecule Resolution Spatial Transcriptomics," presents a pioneering approach to understanding cellular behavior at the molecular level. This study addresses a critical challenge in spatial transcriptomics: predicting the future distribution of cells expressing specific genes. Leveraging the power of Random Forest Regression, we developed a predictive model that operates with high accuracy and resolution.

In spatial transcriptomics, understanding how genes are expressed spatially within a cell and how this expression changes over time is crucial for unraveling complex biological processes. Traditional methods often fall short in capturing these dynamic changes with the needed precision. This paper introduces an innovative pipeline that combines Ripley's K-function with spatial point processes, providing a detailed view of gene expression patterns at a single-molecule resolution. The approach is exemplified through the study of the Sog-D gene in the Drosophila embryogenesis process, offering fresh insights into how cells control gene expression over time.

The significance of this paper lies not just in its technical novelty but also in its practical implications. By accurately predicting the future states of gene expression, researchers can gain a deeper understanding of developmental biology, disease progression, and even potential therapeutic targets. This work stands as a testament to the power of integrating computational methods with biological data, opening new avenues for research in genomics and beyond. A sample prediction made by this model is showcased in 1.1.

The common thread that links this paper to the others in this dissertation is its focus on enhancing the predictive capabilities in biomedical imaging. Just as the other papers explore new frontiers in cancer pathology, histopathology, and multi-modal data analysis, this study advances our ability to forecast biological processes with greater precision and detail, underscoring the transformative impact of AI in the realm of biomedical imaging.

1.2.2 Paper 2: Clinically Relevant Histopathology Slide Compression

The second paper titled "Clinically Relevant Latent Space Embedding of Cancer Histopathology Slides Through Variational Autoencoder Based Image Compression" marks a significant advancement in the field of digital pathology. This paper addresses the critical challenge of efficiently managing the vast amounts of data generated by high-resolution cancer histopathology slides. The sheer size of these datasets poses substantial challenges in terms

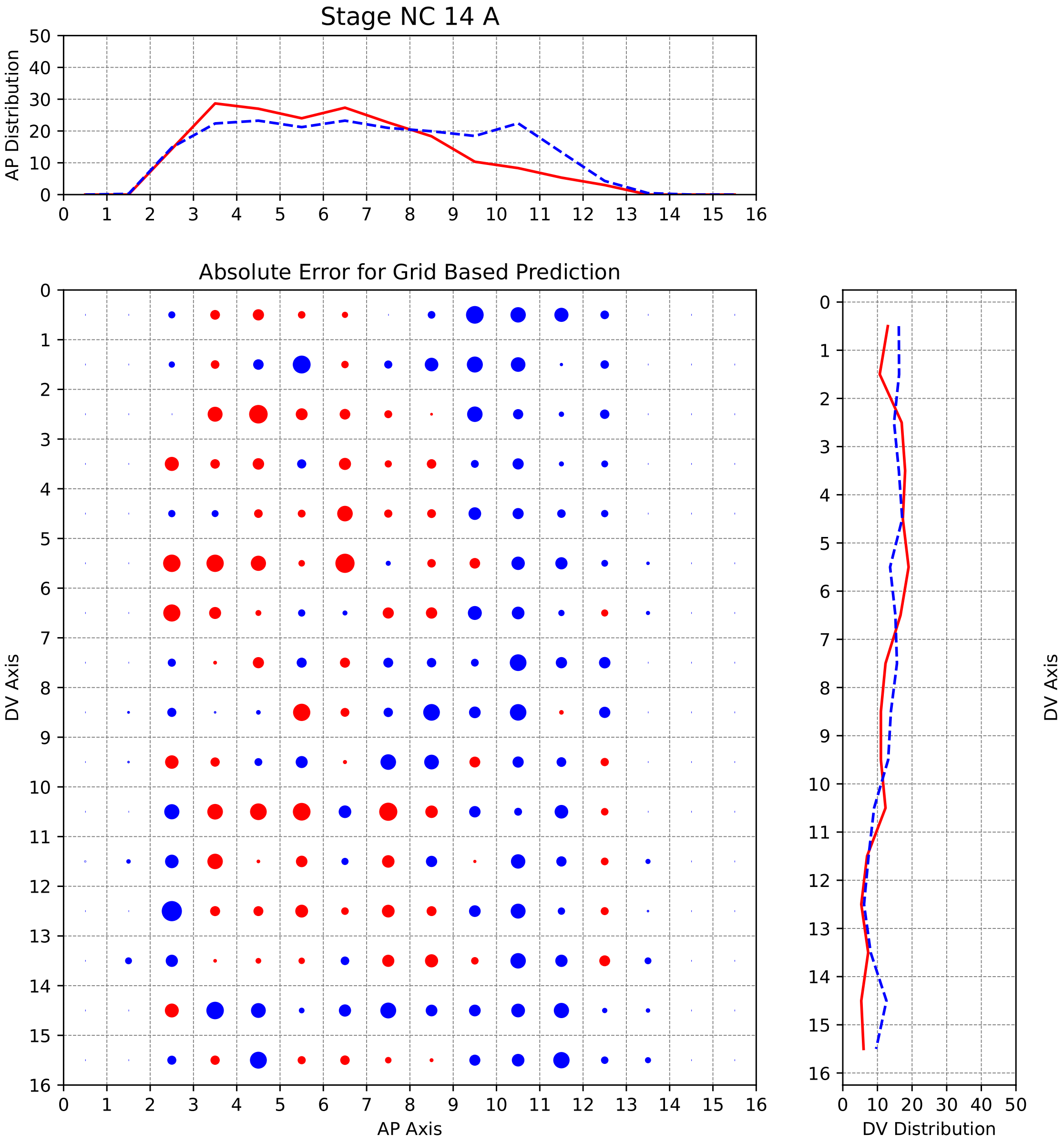


Figure 1.1: The predicted distribution of active cell for stage *NC 14 A* for the sample with the best accuracy based on mean absolute error values. The top and right plot show the distribution of active cells along the anterior to posterior (AP) and Dorsal to Ventral (DV) axes, respectively. The solid red lines are true distributions, and the blue dotted lines are predicted distributions. The middle plot shows the absolute error in each grid.

of storage, retrieval, and analysis. To tackle these issues, we present a novel approach utilizing a Variational Autoencoder (VAE) for image compression.

The core of this paper revolves around the development of a VAE-based training pipeline that achieves high compression ratios while maintaining the integrity and clinical relevance of the histopathological images. This is particularly crucial in cancer diagnostics, where the fidelity of images is paramount for accurate analysis and diagnosis. The VAE model developed in this study not only compresses images efficiently but also ensures that the compressed images retain critical histological features necessary for clinical diagnosis and research.

One of the most important aspects of this study is the generation and visualization of embeddings from the compressed latent space. These embeddings demonstrate that the compressed data maintains crucial clinical information, which can be used for rapid and accurate searches of large histopathological image databases. Such capability, if perfected, has the potential to revolutionize how pathologists and researchers access and analyze histopathology slides, especially in large-scale studies. The UMAP plot of these embeddings are illustrated in 1.2. Also, reconstructed slides at different compression ratios are presented in 1.3 for a sample breast tissue.

The paper aligns seamlessly with the overarching theme of the dissertation by enhancing the analysis and prediction capabilities in biomedical imaging. While the first paper focuses on predicting cellular behavior at the molecular level, this paper contributes to the efficient handling and analysis of large-scale histopathological data. Together, these studies underscore the diverse yet cohesive applications of AI in biomedical imaging, each addressing different facets of the challenges inherent in the field.

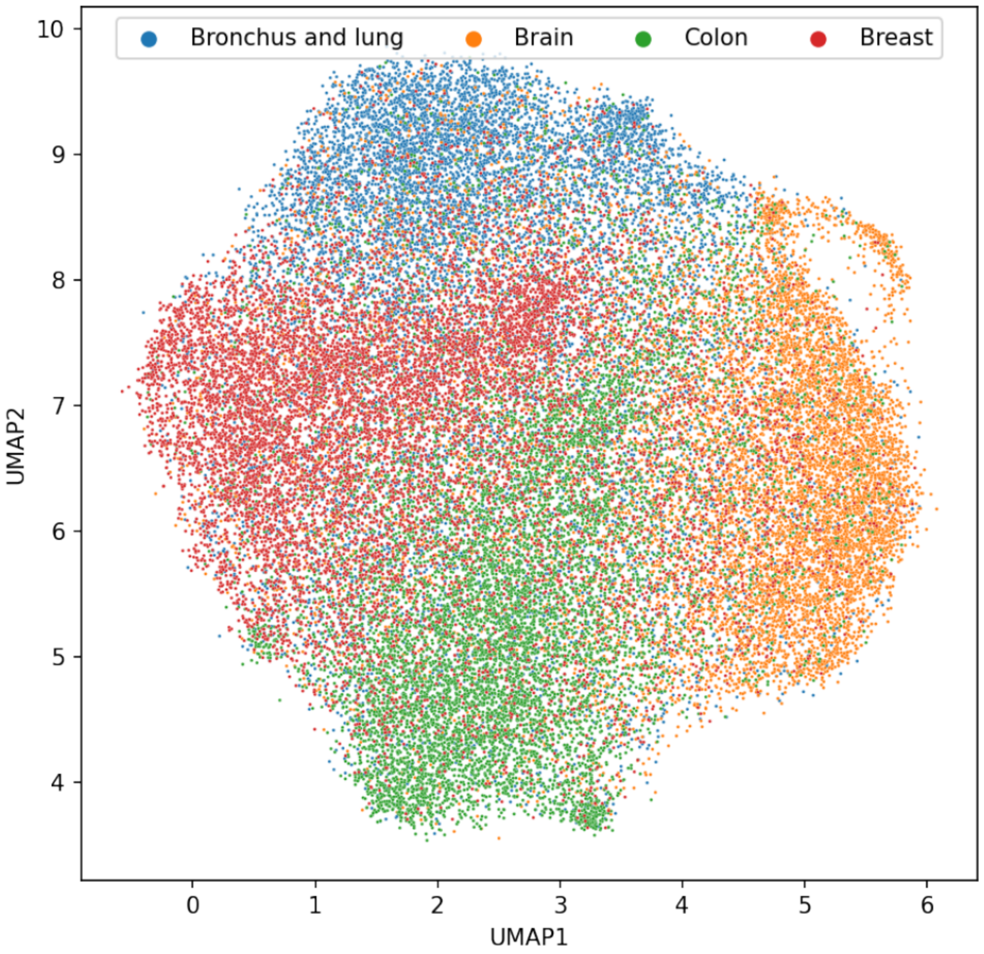


Figure 1.2: The UMAP plot of embedding space demonstrates clinical information preservation (©2023 IEEE).

The innovations presented in this paper are not just technical feats; they hold profound implications for cancer research and diagnosis. By enabling more efficient storage and retrieval of cancer pathology slides and preserving their clinical utility even in a compressed format, this work paves the way for more streamlined and effective diagnostic processes. It exemplifies the potential of AI to transform the way medical data is handled, making it more accessible and usable for clinicians and researchers alike.

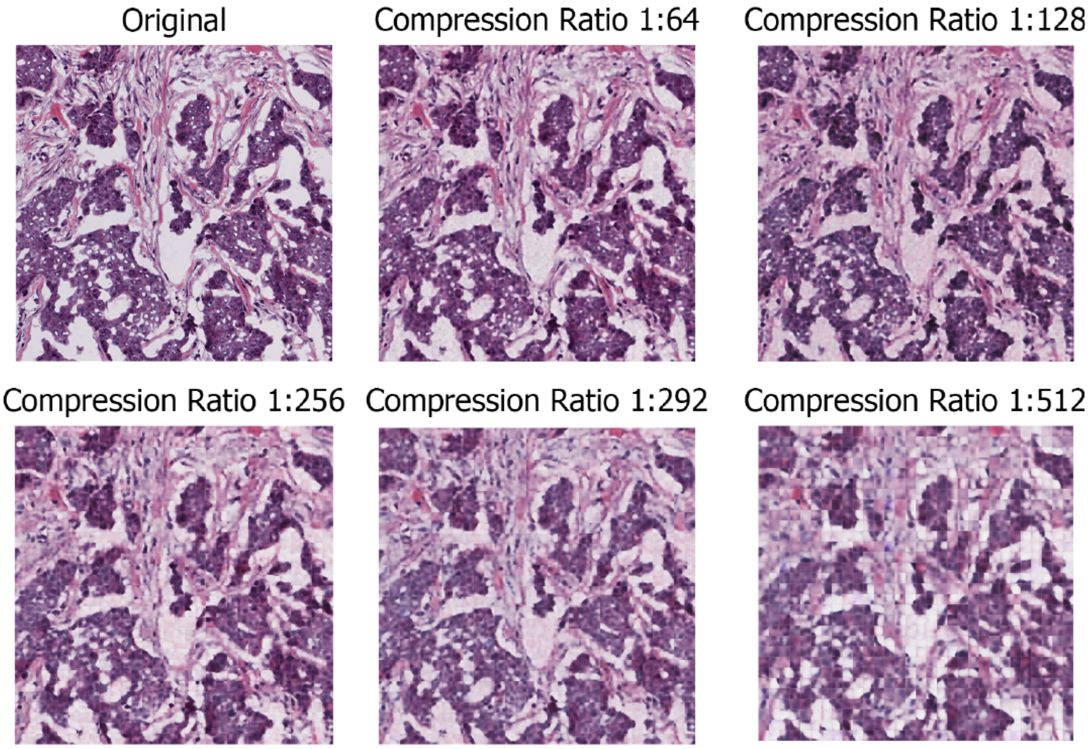


Figure 1.3: The reconstruction results for breast cancer tissues at 5 different compression ratios (©2023 IEEE).

1.2.3 Paper 3: Readiness of Histopathology Slide Search for Clinic

The third paper in this series, titled ”Histopathology Slide Indexing and Search: Are We There Yet?” delves into a critical area of digital pathology - the indexing and retrieval of histopathology slides. This paper presents an in-depth analysis of the current state and advancements in the technologies used for managing and searching vast repositories of histopathology images. It provides a comprehensive review of the methodologies employed in indexing these slides and evaluates the effectiveness of various search algorithms within this domain.

Histopathology slides are integral to the diagnosis and study of diseases, especially cancer. However, the sheer volume of slides produced in clinical and research settings presents significant challenges in terms of storage, management, and retrieval. Efficient indexing and search mechanisms are crucial for leveraging these vast datasets effectively. This paper scrutinizes the progress made in this field, assessing how far we have come in terms of technology and what gaps still exist.

One key focus of the paper is the evaluation of the effectiveness of current search engines designed for histopathology slides. We explore various dimensions of these search engines, including their accuracy, efficiency, and the algorithms that power them. They discuss the challenges faced, such as the need for high precision in search results, handling the variability in slide preparation and imaging, and the integration of these systems into clinical workflows. A sample patch retrieval analysis is shown in 1.4.

The paper's significance lies in its critical analysis of a fundamental aspect of digital pathology. While the previous papers in this dissertation discuss predictive modeling and efficient data compression, this paper addresses the practical aspects of handling and utilizing the resultant data effectively. It highlights the need for robust and intelligent search engines that can handle the complexity and scale of histopathology datasets.

The common thread linking this paper to the others is the overarching goal of enhancing the utility and accessibility of biomedical imaging data. Just as the other papers propose methods to predict biological processes and compress large datasets effectively, this study addresses the subsequent challenge of retrieving and utilizing this data efficiently. It underscores the necessity of continued innovation in AI and machine learning to develop more sophisticated tools for managing the ever-growing repositories of medical imaging data, thereby contributing to the broader objectives of improving diagnostic accuracy and advancing medical research.

In summary, ”Histopathology Slide Indexing and Search: Are We There Yet?” is a pivotal contribution to the field, offering insights into the current capabilities and limitations of histopathology slide indexing and search technologies. It sets the stage for future advancements in this area, calling for ongoing research and development to address the remaining challenges and fully harness the potential of digital pathology.

1.3 Description of the Problem

The realm of biomedical imaging, particularly when augmented with artificial intelligence, presents a host of complex challenges that are crucial to address for advancing medical science and patient care. This dissertation, through its included studies, tackles several of these pivotal challenges:

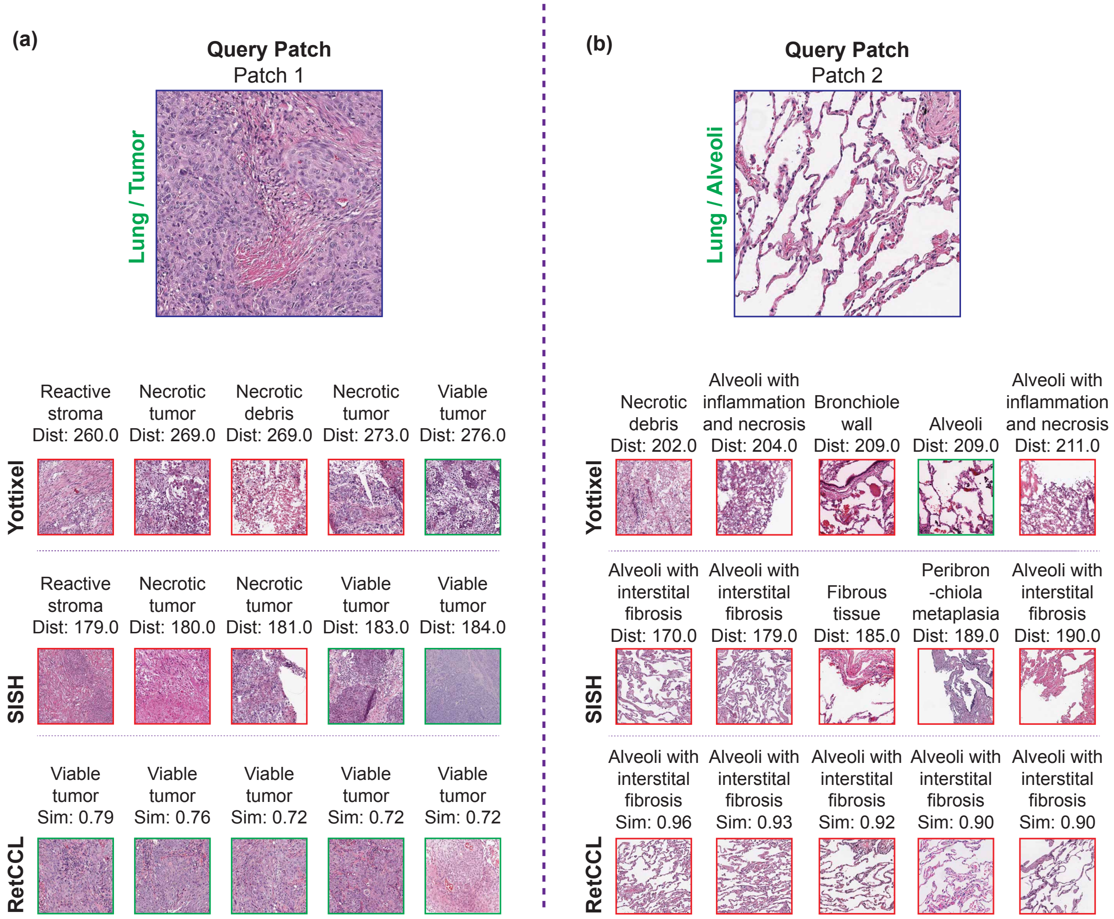

(a)
Query Patch
Patch 1
Lung / Tumor
Reactive stroma Dist: 260.0
Necrotic tumor Dist: 269.0
Necrotic debris Dist: 269.0
Necrotic tumor Dist: 273.0
Viable tumor Dist: 276.0
Yottixel
Reactive stroma Dist: 179.0
Necrotic tumor Dist: 180.0
Necrotic tumor Dist: 181.0
Viable tumor Dist: 183.0
Viable tumor Dist: 184.0
SISH
Viable tumor Sim: 0.79
Viable tumor Sim: 0.76
Viable tumor Sim: 0.72
Viable tumor Sim: 0.72
Viable tumor Sim: 0.72
RetCCL
(b)
Query Patch
Patch 2
Lung / Alveoli
Necrotic debris Dist: 202.0
Alveoli with inflammation and necrosis Dist: 204.0
Bronchiole wall Dist: 209.0
Alveoli Dist: 209.0
Alveoli with inflammation and necrosis Dist: 211.0
Yottixel
Alveoli with interstital fibrosis Dist: 170.0
Alveoli with interstital fibrosis Dist: 179.0
Fibrous tissue Dist: 185.0
Peribron -chiola metaplasia Dist: 189.0
Alveoli with interstital fibrosis Dist: 190.0
SISH
Alveoli with interstital fibrosis Sim: 0.96
Alveoli with interstital fibrosis Sim: 0.93
Alveoli with interstital fibrosis Sim: 0.92
Alveoli with interstital fibrosis Sim: 0.90
Alveoli with interstital fibrosis Sim: 0.90
RetCCL

Figure 1.4: Results of patch retrieval for two patches from sample patches from paper 3. Correct labels are printed in green to the left of query patches. Green border means correct label; red border means wrong label.

1.3.1 Challenges in Analyzing and Predicting Spatial Transcriptomics and Cancer Pathology

The study of spatial transcriptomics and cancer pathology requires not only the analysis of vast amounts of complex biological data but also the accurate prediction of how these data evolve over time. In spatial transcriptomics, the challenge lies in mapping and interpreting the spatial arrangement of gene expression within tissues, which is key to understanding cellular functions and interactions in a given biological context. This requires sophisticated tools capable of handling high-dimensional data and extracting meaningful patterns from them[38,7].

In cancer pathology, the difficulty intensifies with the need to differentiate between subtle morphological features that distinguish benign from malignant cells and among various cancer subtypes. The complexity of tumor biology, including its heterogeneity and the variability in its presentation, makes accurate diagnosis and prognosis a challenging task[14].

1.3.2 The Need for Robust Image Compression Methods

In the digital era of pathology, the transition from glass slides to digital slides generates massive datasets, especially when dealing with high-resolution whole-slide images (WSIs). These datasets necessitate robust image compression methods to enable efficient storage, transmission, and analysis. However, traditional image compression techniques often result in loss of critical information, making them unsuitable for medical applications where precision is paramount[2].

1.3.3 The Complexity of Histopathology Slide Analysis

Histopathology slide analysis is integral to disease diagnosis and research. The challenge here is twofold: first, in the accurate segmentation and classification of pathological features from slide images, and second, in the development of systems capable of handling the sheer scale and variability of histopathology data. This demands algorithms that are not only precise but also adaptable to various staining techniques, tissue types, and disease states[22,30].

In conclusion, these challenges form the basis of the problems addressed in this dissertation. Each of the included studies tackles a specific aspect of these challenges, contributing to the overarching goal of enhancing the capabilities in biomedical imaging and predictive analysis through the application of advanced AI techniques.

1.4 Contributions of the Dissertation Writer and Co-Authors

The body of work presented in this dissertation is the result of collaborative efforts between the dissertation writer and various co-authors. Each paper reflects a synergy of expertise from different fields, bringing together innovative ideas, technical skill sets, and domain-specific knowledge. Below is a detailed breakdown of the contributions made by the dissertation writer and the co-authors for each paper:

1.4.1 Paper 1: Predicting the Future States of Gene Expression

Dissertation Writer's Contributions

1. **Conceptualization and Methodology**: Originated the idea of applying Random Forest Regression for predicting future states in spatial transcriptomics.

2. **Implementation**: Developed the Random Forest Regression model and executed the hypothesis testing, including analysis.

3. **Data Analysis and Visualization**: Conducted and visualized the grid search for the optimal grid size (Fig. 2), created the layout and code for plotting results (Fig. 3), and generated bootstrap plots (Fig. 4).

Co-First Authors' Contributions

- **Parisa Boodaghi Malidarreh**: Played a pivotal role in writing most of the paper, procuring imaging samples, interpreting results, and leading the project.

- **Biraaj Rout**: Managed bulk data analysis and feature extraction, developed an automated data pipeline, and coordinated with the biology department for data acquisition. Also ran the high throughput segmentation pipline for cell segmentation.

- **Priyanshi Borad**: Responsible for imaging and video capturing under lab conditions.

- **Jillur Rahman Saurav**: Assisted in defining the prediction problem and translating biological aspects into computational solutions.

1.4.2 Paper 2: Clinically Relevant Histopathology Slide Compression

Dissertation Writer's Contributions

1. **Data Acquisition and Scripting**: Authored the script for downloading the large dataset and developed the deep learning pipeline in Python.

2. **Technical Implementation**: Parallelized the code for multiple GPU usage, created slurm scripts for supercomputer execution, and generated Figures 1 and 2. Authored the methods section of the paper.

Co-First Author's Contributions

- **Amir Hajighasemi**: Conducted experiments on the CIFAR-10 dataset and designed hypothesis testing to elucidate the model's superior performance on histopathology images. Contributed to parts of the analysis and methods sections.

1.4.3 Paper 3: Readiness of Histopathology Slide Search for Clinic

Dissertation Writer's Contributions

1. **Original Concept and Design**: Initiated the paper idea and designed experiments for testing various search engines.

2. **Software Development and Analysis**: Wrote the complete source code, implemented RetCCL from scratch, conducted ablation studies, performed statistical analysis, and generated Figures 2 and 3. Authored the methods section.

Co-First Author's Contributions

- **Dr. Helen Shang**: Managed the medical analysis, patient data interpretation, and most of the paper writing. Facilitated access to human samples from UCLA.

These contributions underscore the collaborative nature of scientific research, where each member brings unique skills and insights. The dissertation writer, through their direct involvement in conceptualizing, developing, and implementing these studies, has demonstrated a deep understanding and capability in AI applications in biomedical imaging. This collaborative effort has significantly advanced the field, addressing some of the most pressing challenges in biomedical imaging and predictive analysis.

1.5 Review of Prior Work and Literature Review

The field of biomedical imaging, particularly in the context of AI-enhanced analysis and prediction, stands at the forefront of modern medical research and practice. This section delves into a comprehensive review of prior work and literature that has paved the way for the current research presented in this dissertation. Each paper included in this work builds upon a rich foundation of previous studies, addressing critical aspects of spatial transcriptomics, cancer pathology, and image compression. By reviewing and contextualizing the existing literature, we aim to highlight the evolutionary trajectory of these research domains and underscore the significant gaps our current research endeavors to fill. This literature review not only serves to situate our contributions within the broader research landscape but also to demonstrate the progressive nature of these fields, where each advancement brings us closer to more precise, efficient, and predictive capabilities in biomedical imaging and analysis.

1.5.1 Paper 1: Predicting the Future States of Gene Expression

In ”Predicting Future States with Spatial Point Processes in Single Molecule Resolution Spatial Transcriptomics,” we delve into the frontier of understanding spatial gene expression patterns in embryogenesis, a critical aspect of modern genomics and developmental biology. The introduction of the paper lays a foundation for the study by highlighting key advancements and challenges in the field.

Comprehensive Review of Existing Research

Recent technological strides have enabled the capture of high-resolution images during the embryogenesis process, which are pivotal for studying gene expression patterns[24,9]. The Drosophila embryo, in particular, has been a model organism for understanding how enhancers control gene expression in a complex and dynamic manner[33,42]. However, the rapid advancement in genetic and live imaging techniques has outpaced the development of analytical methods capable of extracting the wealth of information contained within these datasets[29].

The paper discusses the challenge of systematically assessing mutant enhancer phenotypes. We developed a quantitative approach using enhancer-driven MS2-yellow reporter constructs, captured through in vivo imaging, to provide insights into the timing, levels, and spatial domains of expression[34]. These advancements in imaging technology necessitate novel methods for efficient prediction and analysis of spatial gene expression data.

The concept of RNA velocity, defined as the time derivative of gene expression, was introduced as a novel way to estimate the future state of individual cells in standard scRNA-seq protocols[26]. Furthermore, methodologies capturing spatial proteomics data to predict cancer patient survival, utilizing tools like Ripley's K-function for spatial feature analysis, have inspired the proposed pipeline in this study[8].

Contextualizing the Papers within the Broader Research Landscape

This paper is positioned at the intersection of advanced imaging techniques and computational analysis. We acknowledge the rich history of developmental biology research and the current technological capabilities that enable the study of gene expression dynamics with increased temporal resolution. However, they also identify a gap in the ability to predict and analyze these dynamics efficiently, especially in the context of spatial transcriptomics.

Highlighting the Gaps Addressed by the Current Research

The research addresses the critical gap in predictive analysis of spatial gene expression data. While previous methods have provided static snapshots of gene expression, this study introduces a dynamic perspective, offering a tool analogous to RNA velocity but tailored for spatially resolved developmental biology. The introduction of a Random Forest Regression model, combined with temporally resolved spatial point processes, marks a significant advancement in the field. It exemplifies the transition from static to dynamic analysis of gene expression, an essential step in understanding complex biological processes like embryogenesis.

The literature review in this paper underscores the necessity for innovative computational approaches to keep pace with rapid advancements in imaging technology. By addressing the need for dynamic predictive models in spatial transcriptomics, the study makes a substantial contribution to the field, bridging a crucial gap between data acquisition and data analysis.

1.5.2 Paper 2: Clinically Relevant Histopathology Slide Compression

The second paper, ”Clinically Relevant Latent Space Embedding of Cancer Histopathology Slides Through Variational Autoencoder Based Image Compression,” focuses on addressing the challenges of managing and analyzing large-scale histopathological image data. The introduction section provides an extensive review of the current state of histopathological image analysis and the need for efficient image compression methods in the field.

Comprehensive Review of Existing Research

Histopathological images, particularly those derived from cross-sectional tissue microscopy, play a crucial role in diagnosing various diseases and conditions [13,14]. The introduction of Hemotoxylin and Eosin (H&E) staining has been a significant advancement, enabling the discernment of nuclear and cytoplasmic structures for identifying carcinomal regions in excised tissue from cancer patients [40]. With the creation of large databases like the NIH Genomic Data Commons (GDC), containing tens of thousands of Whole Slide Images (WSIs), the need for efficient data management has become more pressing [20].

However, traditional compression methods like JPEG2000 have limitations in maintaining the usability of images for histopathological classification beyond certain compression ratios [25,23]. This has prompted the exploration of neural networks, particularly Variational Auto Encoders (VAEs), which have shown higher efficiency and fidelity in compressing image data while retaining critical information necessary for medical applications [46,19].

Contextualizing the Papers within the Broader Research Landscape

The paper is situated at a critical juncture in the field of digital pathology, where the growing size of histopathological datasets demands innovative solutions for data compression and retrieval. The authors highlight the advancements in VAE technology and its application in medical image analysis, demonstrating its superiority over traditional compression methods [18,32].

Highlighting the Gaps Addressed by the Current Research

The research addresses a significant gap in the field of histopathological image analysis - the need for a compression method that balances high efficiency with the preservation of clinically relevant information. The introduction of a VAE-based approach for the compression and indexing of WSIs presents a novel solution to this problem. It not only facilitates the efficient storage and retrieval of large-scale histopathological data but also ensures the clinical utility of the compressed images, a crucial aspect in cancer diagnostics and research.

The literature review in this paper underscores the importance of developing advanced computational methods to keep pace with the increasing scale of histopathological datasets. By focusing on the efficient compression and retrieval of these images, the study makes a substantial contribution to the field, bridging a vital gap between data acquisition and clinical application.

1.5.3 Paper 3: Readiness of Histopathology Slide Search for Clinic

The paper ”Histopathology Slide Indexing and Search: Are We There Yet?” presents a comprehensive examination of the current state of histopathology slide indexing and retrieval systems. This review focuses on the evolution of these systems, their technological advancements, and the existing challenges in the field.

Comprehensive Review of Existing Research

Histopathology, a critical domain in medical diagnostics, has undergone a significant transformation with the digitization of slides. This shift has prompted the development of automated search and retrieval systems, which are essential for managing the burgeoning volume of digital slides. Among the notable advancements are end-to-end systems like Yottixel[20], SISH[6], RetCCL[51], and HSHR[28]. Yottixel, for example, introduced an innovative approach of processing large-scale WSIs by using a DenseNet-based feature extractor on mosaic tiles, rather than the entire WSI, marking a significant improvement in handling large volumes of data efficiently.

Contextualizing the Papers within the Broader Research Landscape

This research fits into a broader landscape where the efficiency and accuracy of histopathology slide search engines are critical. These systems not only enhance the diagnostic process but also play a vital role in research and education. The advancements in CBIR systems, feature extraction techniques, and integration with AI and machine learning have marked significant strides in this field.

Highlighting the Gaps Addressed by the Current Research

The paper identifies gaps in existing methodologies, notably the need for more sophisticated, accurate, and user-friendly systems for histopathology slide indexing and search. Despite advancements, challenges persist in handling the scale of digital pathology, ensuring the clinical relevance of search results, and integrating these systems into existing workflows. This study addresses these gaps, providing insights and directions for future research.

2

Paper 1: Predicting the Future States of Gene Expression

2.1 Introductory Comments

The study presented in this chapter, ”Predicting Future States with Spatial Point Processes in Single Molecule Resolution Spatial Transcriptomics,” explores the intricate process of embryogenesis in Drosophila, focusing on the spatial and temporal patterns of gene expression. Central to this exploration are the Anterior-Posterior (AP) and Dorsal-Ventral (DV) axes, which play a crucial role in the development of the embryo.

Significance of the AP and DV Axes in Development

The AP and DV axes are fundamental in determining the body plan of an organism during embryonic development. In Drosophila, as in many other organisms, these axes are established very early in embryogenesis and dictate the spatial arrangement of tissues and organs. The AP axis runs from the head (anterior) to the tail (posterior) of the organism, while the DV axis runs from the back (dorsal) to the belly (ventral) side.

Understanding the gene expression patterns along these axes is crucial for unraveling the complex mechanisms that guide embryonic development. Genes expressed along these axes determine the positional information of cells, influencing their fate and function in the developing organism. Any alterations or disruptions in the expression patterns along these axes can lead to developmental abnormalities.

Approach and Relevance of the Study

This study employs a sophisticated approach using Random Forest Regression combined with spatial point processes, enabling the prediction of future states of cell distribution along the AP and DV axes. By leveraging high-resolution imaging data, the research provides insights into the dynamic nature of gene expression at different stages of embryogenesis.

The ability to predict the distribution of cells expressing specific genes is not just a technical achievement but also a significant step forward in understanding the developmental biology of organisms. This research contributes to a deeper understanding of how organisms develop from a single cell to a complex system of tissues and organs, shedding light on the fundamental processes that underpin life itself.

The following pages will present the full manuscript of ”Predicting Future States with Spatial Point Processes in Single Molecule Resolution Spatial Transcriptomics,” including the title page, abstract, main body, references, and any supplementary material, formatted according to the requirements of the publisher and renumbered for consistency within the thesis/dissertation.

PREDICTING FUTURE STATES WITH SPATIAL POINT PROCESSES IN SINGLE MOLECULE RESOLUTION SPATIAL TRANSCRIPTOMICS

Parisa Boodaghi Malidarreh*, Biraaj Rout*, **Mohammad Sadegh Nasr***, Priyanshi Borad*, Jillur Rahman Saurav*, Jai Prakash Veerla, Kelli Fenelon, Theodora Koromila, and Jacob M. Luber

Submitted to the 2024 IEEE International Symposium on Biomedical Imaging (ISBI)

The first page of the article begins on the next page.

* Indicates co-first authors.

PREDICTING FUTURE STATES WITH SPATIAL POINT PROCESSES IN SINGLE MOLECULE RESOLUTION SPATIAL TRANSCRIPTOMICS

Parisa Boodaghi Malidarreh[⋆,1,4] *Biraaj Rout*[⋆,1,4] *Mohammad Sadegh Nasr*[⋆,1,4]
Priyanshi Borad[⋆,2,4] *Jillur Rahman Saurav*[⋆,1,4] *Jai Prakash Veerla*[1,4]
Kelli Fenelon[2,4] *Theodora Koromila*[†,2,4] *Jacob M. Luber*[†,1,3,4]

[1] Department of Computer Science and Engineering, University of Texas at Arlington
[2] Department of Biology, University of Texas at Arlington
[3] Department of Bioengineering, University of Texas at Arlington
[4] Multi-Interprofessional Center for Health Informatics, University of Texas at Arlington

ABSTRACT

In this paper, we introduce a pipeline based on Random Forest Regression to predict the future distribution of cells that are expressed by the Sog-D gene (active cells) in both the Anterior to posterior (AP) and the Dorsal to Ventral (DV) axis of the Drosophila in embryogenesis process. This method provides insights about how cells and living organisms control gene expression in super resolution whole embryo spatial transcriptomics imaging at sub cellular, single molecule resolution. A Random Forest Regression model was used to predict the next stage active distribution based on the previous one. To achieve this goal, we leveraged temporally resolved, spatial point processes by including Ripley's K-function in conjunction with the cell's state in each stage of embryogenesis, and found average predictive accuracy of active cell distribution. This tool is analogous to RNA Velocity for spatially resolved developmental biology, from one data point we can predict future spatially resolved gene expression using features from the spatial point processes.

Index Terms— Random Forest, Regression, Dorpsophila, Sog-D, Ripley's K-function, transcriptomics, embryogenesis

1. INTRODUCTION

Recent technological advances have made it possible to capture high resolution images from embryogenesis process that help researchers to study gene expression patterns.[1, 2]. One of the major challenges of the modern genomics era is to better understand how gene expression is regulated to support spatiotemporal outputs that change over the course of development. The early Drosophila embryo has served as a paradigm for how enhancers control patterning and has demonstrated that the patterning process is complex and dynamic. It is known that multiple, transiently acting enhancers act sequentially to support changing outputs of expression for some genes[2, 3, 4], whereas other genes are controlled by enhancers that act over a longer period and support changing spatial outputs over time. For example, expression of the gene short gastrulation (sog) is driven by at least two co-acting enhancers that support temporally dynamic expression. Live imaging experiments offer the capacity to analyze gene expression dynamics with increased temporal resolution and linear quantification. However, genetic and live imaging techniques have outpaced analysis techniques to harvest the bountiful information contained within real-time movies of transcriptional dynamics with modern methods confined to static parameter cell and transcript tracking methods [1, 5, 6]. To assess these mutant enhancer phenotypes systematically, we developed a quantitative approach to measure the spatiotemporal outputs of enhancer-driven MS2-yellow reporter constructs as captured by in vivo imaging to provide information about the timing, levels, and spatial domains of expression. Using transgenic fly lines, we conducted live imaging of the GFP signal associated with the MS2 stem-loop reporter sequence. This MS2 cassette contains 24 repeats of a DNA sequence that produces an RNA stem loop when transcribed. The stem-loop structure is specifically bound by the phage MS2 coat protein (MCP). MCP fused to GFP binds to MS2-containing transcripts (i.e., sog_Distal.MS2) producing a strong green signal within the nuclei of Drosophila embryos at sites of nascent transcript production. In this system, the nuclear GFP signal is only observed as a single dot for every nucleus corresponding to nascent transcription of the one copy of the MS2-containingreporter transgene site integrated into the genome. Furthermore, the nuclear periphery is marked by a fusion of RFP to nuclear pore protein (Nup-RFP) [7]. The imaging protocol was optimized to provide spatial information across the entire dorsal-ventral (DV) axis of embryos with the fastest temporal resolution that also retains

[⋆] Equal contribution.

[†] Responsible authors. Email: jacob.luber@uta.edu, theodora.koromila@uta.edu

embryo viability. In brief, embryos were imaged on Zeiss LSM 900 continuously over the course of 2hr at an interval of 30s per scan (twice as fast compared to previous studies). Importantly, this imaging protocol is not phototoxic to embryos. Because spatial outputs likely change in time across the embryo for many gene expression patterns, we developed an image processing approach to collect detailed information in both time and space by capturing one lateral half of the embryos. With this qualified imaging dataset, our goal was to predict the distribution of active cell in each stage of the embryo development. Several methods have been proposed for the efficient prediction of temporal variables. Authors in [8] proposed a novel concept called RNA velocity, which is defined as the time derivative of the gene expression. This concept allows for the estimation of the future state of individual cells in standard scRNA-seq protocoles. In [9], authors proposed a method to capture spatial proteomics data to map cell states in order to predict cancer patient survival. They utilized the Ripley's K-function for capturing spatial features which inspired us in our proposed pipeline. We developed a feature extraction method and analysis pipeline that can be used to predict the future distribution of cells in which the Sog-D gene is expressed.

2. METHODS

We generated super resolution live imaging data expressing *sog* gene (control) and *sog-D* gene (case) in early embryo of *Drosophila* (9 case, 4 control). We conduct pre-processing, feature extraction, training, and testing Fig.1. Both the training and testing phases incorporate identical pre-processing and feature extraction steps. The videos shows real time images from embryonic development, which were manually given stage development labels: NC 13 early, NC 13 late, NC 14 A, NC 14 B, NC 14 C, NC 14 D. In the pre-processing step, we used a generalist, deep learning-based segmentation method called Cellpose, which can precisely segment cells in each frame of the embryo development. Active cells were identified based on prevalance of green pixels indicative of gene expression within the cell, and the active mask underwent feature extraction. During this stage, the masked images underwent a gridding procedure with a predetermined size. Subsequently, the entire imaging dataset was transformed into a tabular format, taking into account the spatial information of each cell. We utilized four different metrics to capture both local and global features in a frame including m1, m2 for both AP and DV axes, Ripley's k-function, and n (total number of cells in each grid). Here, m1 and m2 denote the first and second moments, respectively, capturing the distribution of active cells at each stage. Furthermore, Ripley's k-function was employed to analyze spatial correlation and quantify deviations from a random spatial distribution. Equation 1 illustrates the formula for calculating Ripley's k-function. Where, A is the area under each window with constant radius, n is the number of data points, d_{ij} is the distance between two points, and e_{ij} is an edge correction weight. Then, the tabular data went through two steps of averaging on each stage and time correcting. Since our goal is to predict the distribution of active cells in each stage and we have different number of frames for each stage, we averaged the whole feature values based on each stage. Also, to account for temporal alignment, we implemented a one-stage shift in features, where we utilized the features from the previous stage in prediction of the current stage. Following the completion of the feature extraction process, the dataset undergoes preparation for training a random forest regression model, a supervised learning algorithm. The outcome of this pipeline is the count of active cells within each grid at a given stage, determined by the features from the preceding stage. Subsequent to training the model, its performance is evaluated using test data. During testing, all pre-processing and feature extraction steps are replicated, and the pre-trained random forest regression model is employed to forecast the count of active cells for each grid across various stages.

$$\hat{K}_r = \frac{A}{n(n-1)} \sum_{i=1}^{n} \sum_{i=1, j\neq i}^{n} 1(d_{ij} \leq r) e_{ij} \qquad (1)$$

3. EXPERIMENT AND RESULTS

3.1. Main study

As outlined in the methodology section, during the feature extraction phase, square grids were applied to images, and the number of active cells within each grid was predicted. The key challenge was selecting the optimal grid size to enhance performance on test data. Consequently, we replicated the entire process of pre-processing and feature extraction for four distinct grid sizes: 250, 125, 62.5, and 31.25 (where the grid size of 'n' indicates the division of the entire image into n*n squares). We used three different metrics to calculate the model performance on test data for different grid sizes which are rmse (root mean squared error), mae (mean absolute error), and Kullback-Leibler (KL) Divergence. Fig.2 shows the experiment for different grid sizes. Our analysis revealed the same increasing trend in both rmse and mae as the grid size increases from 31.25 to 250 which indicated that a smaller grid size corresponds to a lower error. KL Divergence, which we also utilized as a metric, measures how one probability distribution diverges from a second one. Thus, the smaller value for it shows that two distributions are closer to each other. We used this criterion to see how well the pipeline can capture the trends in the active cells distribution. The KL Divergence for these four different grid sizes showed the different trend. Increasing the grid size from 31.25 to 250 yielded a decrease in KL Divergence. We had two options, the first one was to select 31.25 based on the lower rmse and mae. However, the problem was the average size of the cell was approximately 36 so if we set the grid size to 31.25 we have just one cell

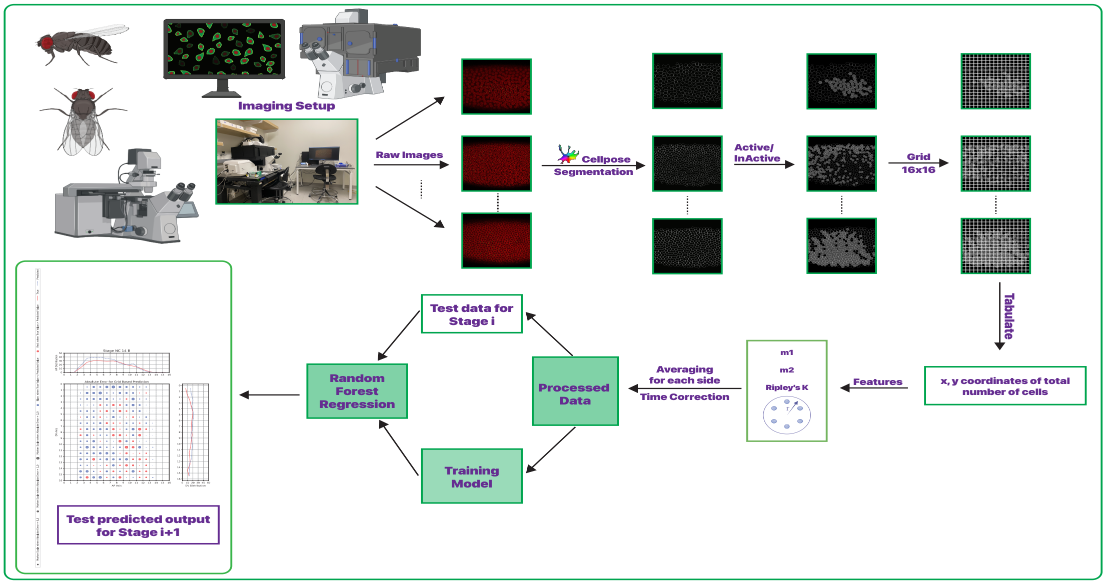


Fig. 1. Implemented pipeline, starting with the imaging process, followed by subsequent stages involving pre-processing, feature extracting, training ,and testing. These steps collectively aim to predict the distribution of active cells for the next stage.

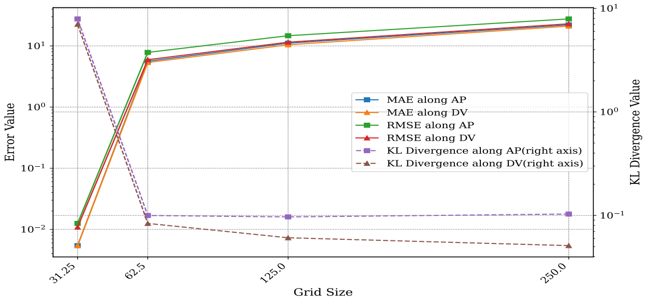


Fig. 2. The experiment for grid search to find the optimal grid size

in each grid which changes the problem to a classification of active or inactive for each grid which was not our purpose. Another option was to select the optimal grid size based on KL Divergence, which finally, We selected the grid size of 62.5 over 31.25. The decision of selecting 63.5 over 125.0 although the 125 had lower KL Divergence, is attributed to the computational constraints of calculating Ripley's k-function for larger grid sizes in our setup.

In subsequent experiment, we conducted an ablation study to discern the relative importance of features, identifying those deemed crucial for inclusion in the final release and those that may be omitted. Table 1 indicates the performance of different combinations of features. It can be concluded that features of the first row including Ripley's k-function and n are the most important features that we used them for training and testing the pipeline. All reported mae values underwent the K-fold cross validation method to mitigate the influence of random results.

To visualize the performance of the pipeline with selected features and parameters we tested the pre-trained model on test dataset. Fig 3 shows the distribution of active cell for the best, median and the worst prediction based on the average mae values.

Feature list	**mae**
n, Ripley's k-function	4.53
m2_DV, n, Ripley's k-function	4.73
m1_DV, n, Ripley's k-function	4.75
m1_DV, m2_AP, n, Ripley's k-function	4.77
m2_AP, n, Ripley's k-function	4.77

Table 1. The average mae value on K-fold cross validation over test dataset for different combinations of features for ablation study.

3.2. Case and control study

As, we had 4 videos for case (transgenic) and 9 for control, we randomly selected 3 videos from each group for training and 1 for testing. Then, we averaged the AP_mae, DV_mae, and mean_mae for whole case and control experiments and calculated the difference between case and control for each of these metrics and the results were 1.86, -0.689, and 0.58 respectively. We also utilized cross-validation to avoid overfitting. These results show there is a difference between the performance of our pipeline on case and control in AP_mean and mean_mae. In other words, our method works better in predicting along AP axis and the mean of AP and DV on control data in comparison with the case one. However, the negative difference between case and control for DV_mae indicates that the pipeline works better in predicting the distribution on DV axis of case compared to control. In order to To substantiate this assertion, we conducted two additional experiments: First, we leveraged Mixed-Effects modelling, which can account for both fixed effects (like the group:

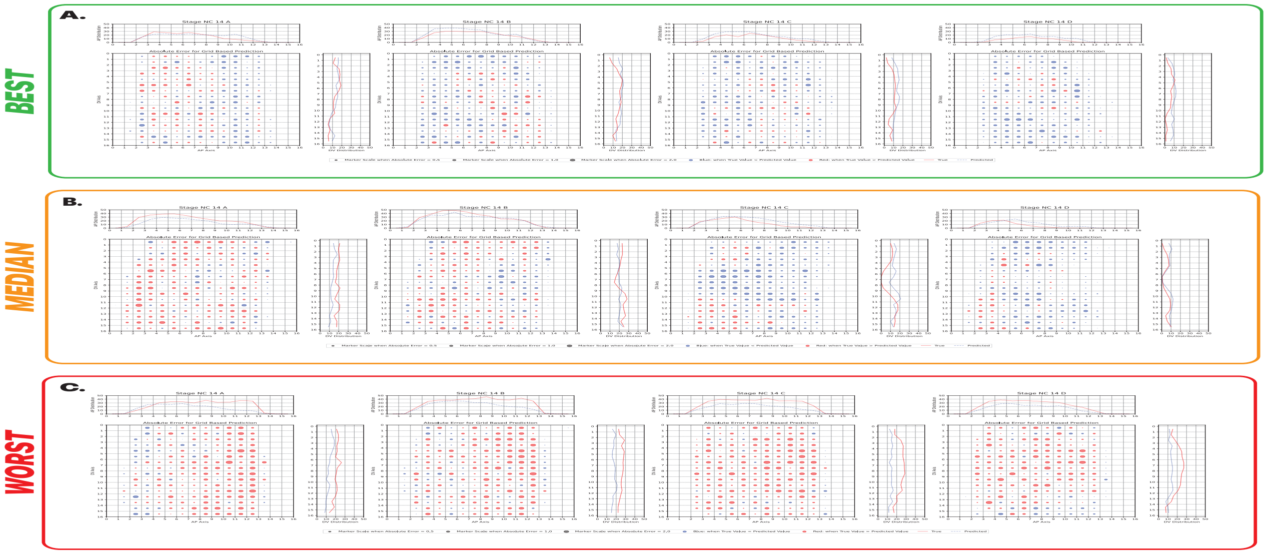


Fig. 3. The distribution of active cell for the best (A), median (B), and worst (C) accuracy based on mae values. For each A, B, and C from left to right stages are NC 14 A-D. For each stage the top and right plot shows the distribution of active cells along AP and DV axis respectively. The middle plot shows the absolute error in each grid.

case or control) and random effects (like the variation within videos and stages). The mixed-effects model can help in understanding the influence of these fixed and random effects on our dependent variables like DV_mae, AP_mae, mean_mae. The goal is to understand whether there is a significant difference in any metrics between the case and control groups, accounting for the variability introduced by different stages. The control group has, on average, a lower AP_mae compared to the case by about 1.828 units with the P_value of 0.003. It shows based on this test, there is a statistically significant difference in AP_mae between case and control groups. However, the result for DV_mae shows the control group has higher value by 0.714 units and 0.231 P_value. Also, the result for mean_mae indicates control has higher value by -0.557 units and 0.347 P_value. Two latter results for DV_mae and mean_mae cannot indicate any significant difference between case and control because of the high P_values. In addition, we implemented another empirical hypothesis testing called Bootstrap method. Bootstrap methods can be used to estimate the distribution of our metrics under the null hypothesis. To implement the bootstrap, we used the same metrics as previous method. we drew samples from the original dataset with replacement, to create a new dataset. Then, for each bootstrap sample, we computed the statistics of interest which are DV_mae, AP_mae, and mean_mae. By analyzing the this bootstrap distribution we can find the confidence intervals for each metrics. Fig 4 shows the Bootstrap distribution of mean difference in AP_mae, DV_mae, and mean_mae. It indicates that with 95% confidence interval the mean difference of AP_mae, (AP_mae(case) - AP_mae(control)) was between [0.69061964 3.11528348]. It can be concluded that with 95% confidence interval the AP_mae for case is at least 0.69061964 units higher than case, which means the performance of the pipeline is better for control outperforms case one. These ranges for DV_mae and mean_mae are respectively, [-1.65878863 0.27041668] and [-0.33784703 1.5450897]. It can be seen that for DV_mae and mean_mae the ranges include zero means the performance of control can be better, equal, or worse than case. The results with Bootstrap method confirms the results derived from mixed effects method, which makes sense given that large amounts of training data are needed to model transgenic effects.

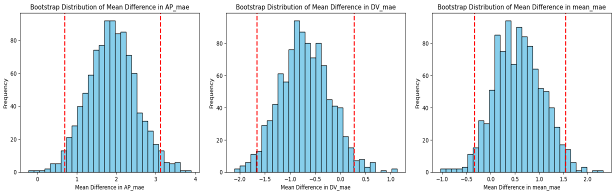


Fig. 4. The Bootstrap Distribution of Mean Difference in AP_mae, DV_mae, and mean_mae between case and control in 1000 iterations.

4. CONCLUSION

Our work presents several key contributions. Firstly, we have developed a novel and optimized imaging technology that delivers spatial information throughout the entire DV axis of an embryo. Secondly, we introduce an automated pipeline that effectively discriminates cell types with high accuracy. Lastly, our approach enables the accurate prediction of the stage-level distribution of active cells, based on data from the preceding stage.

5. COMPLIANCE WITH ETHICAL STANDARDS

All animal experiments were approved by the UTA IACUC review board. This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of my institution.

6. ACKNOWLEDGMENTS

This work was supported by the Cancer Prevention and Research Institute of Texas (CPRIT) Recruitment of First-Time, Tenure-Track Faculty Members Grant (RR220015) (JML) and University of Texas System Rising STARs award (JML).

7. REFERENCES

[1] Theodora Koromila and Angelike Stathopoulos, "Distinct roles of broadly expressed repressors support dynamic enhancer action and change in time," *Cell reports*, vol. 28, no. 4, pp. 855–863, 2019.

[2] Leslie Dunipace, Abbie Saunders, Hilary L Ashe, and Angelike Stathopoulos, "Autoregulatory feedback controls sequential action of cis-regulatory modules at the brinker locus," *Developmental cell*, vol. 26, no. 5, pp. 536–543, 2013.

[3] Hannah K Long, Sara L Prescott, and Joanna Wysocka, "Ever-changing landscapes: transcriptional enhancers in development and evolution," *Cell*, vol. 167, no. 5, pp. 1170–1187, 2016.

[4] Michael W Perry, Jacques P Bothma, Ryan D Luu, and Michael Levine, "Precision of hunchback expression in the drosophila embryo," *Current biology*, vol. 22, no. 23, pp. 2247–2252, 2012.

[5] Bomyi Lim, Tyler Heist, Michael Levine, and Takashi Fukaya, "Visualization of transvection in living drosophila embryos," *Molecular cell*, vol. 70, no. 2, pp. 287–296, 2018.

[6] Anthony Birnie, Audrey Plat, Cemil Korkmaz, and Jacques P Bothma, "Precisely timed regulation of enhancer activity defines the binary expression pattern of fushi tarazu in the drosophila embryo," *Current Biology*, 2023.

[7] Tanguy Lucas, Teresa Ferraro, Baptiste Roelens, Jose De Las Heras Chanes, Aleksandra M Walczak, Mathieu Coppey, and Nathalie Dostatni, "Live imaging of bicoid-dependent transcription in drosophila embryos," *Current biology*, vol. 23, no. 21, pp. 2135–2139, 2013.

[8] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al., "Rna velocity of single cells," *Nature*, vol. 560, no. 7719, pp. 494–498, 2018.

[9] Monica T Dayao, Alexandro Trevino, Honesty Kim, Matthew Ruffalo, H Blaize D'Angio, Ryan Preska, Umamaheswar Duvvuri, Aaron T Mayer, and Ziv Bar-Joseph, "Deriving spatial features from in situ proteomics imaging to enhance cancer survival analysis," *Bioinformatics*, vol. 39, no. Supplement_1, pp. i140–i148, 2023.

3

Paper 2: Clinically Relevant Histopathology Slide Compression

3.1 Introductory Comments

In this chapter, we delve into the paper titled ”Clinically Relevant Latent Space Embedding of Cancer Histopathology Slides Through Variational Autoencoder Based Image Compression," which addresses the challenges of managing and analyzing large-scale histopathological data. Central to this study are the concepts of Variational Autoencoders (VAEs) and Uniform Manifold Approximation and Projection (UMAP), both of which represent significant advancements in the field of machine learning and data representation.

3.1.1 Understanding Variational Autoencoders (VAEs)

A Variational Autoencoder (VAE) is a type of deep learning model that's particularly effective for unsupervised learning of complex data distributions. VAEs are designed to compress data into a latent, or hidden, space and then reconstruct the original data from this compressed representation. The key aspect of VAEs lies in their ability to model the latent space in a way that encourages efficient, continuous, and structured data representation. This makes VAEs highly suitable for tasks like image compression, where the goal is to reduce the dimensionality of the data while retaining its critical features.

In the context of histopathology slides, the VAE model facilitates the compression of high-resolution images into a manageable size, making it easier to store, process, and analyze these large datasets. The model's ability to reconstruct images from the latent space ensures that crucial diagnostic information is not lost during compression.

3.1.2 Role of Uniform Manifold Approximation and Projection (UMAP)

UMAP is a dimensionality reduction technique that is particularly useful for visualizing high-dimensional data in a lower-dimensional space. In this study, UMAP is employed to visualize and understand the latent space created by the VAE. This visualization provides insights into how different types of histopathological data are represented and clustered in the latent space.

UMAP’s strength lies in its ability to maintain the local and global structure of high-dimensional data, making it an excellent tool for exploring patterns and relationships in complex datasets. In the context of histopathology slide compression, UMAP helps demonstrate that even after significant data reduction, the latent representations preserve essential clinical and biological information.

The following pages will present the full manuscript of ”Clinically Relevant Latent Space Embedding of Cancer Histopathology Slides Through Variational Autoencoder Based Image Compression,” including the title page, abstract, main body, references, and any supplementary material. This presentation will adhere to the publisher’s formatting requirements and be seamlessly integrated into the overall structure of the thesis/dissertation.

CLINICALLY RELEVANT LATENT SPACE EMBEDDING OF CANCER HISTOPATHOLOGY SLIDES THROUGH VARIATIONAL AUTOENCODER BASED IMAGE COMPRESSION[1]

Mohammad Sadegh Nasr*, Amir Hajighasemi*, Paul Koomey, Parisa Boodaghi Malidarreh, Michael Robben, Jillur Rahman Saurav, Helen H Shang, Manfred Huber, and Jacob M. Luber

Nasr, Mohammad Sadegh, Amir Hajighasemi, Paul Koomey, Parisa Boodaghi Malidarreh, Michael Robben, Jillur Rahman Saurav, Helen H Shang, Manfred Huber, and Jacob M. Luber. "Clinically Relevant Latent Space Embedding of Cancer Histopathology Slides Through Variational Autoencoder Based Image Compression." In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), 1–5, 2023. https://doi.org/10.1109/ISBI53787.2023.10230343.

The first page of the article begins on the next page.

* Indicates co-first authors.

CLINICALLY RELEVANT LATENT SPACE EMBEDDING OF CANCER HISTOPATHOLOGY SLIDES THROUGH VARIATIONAL AUTOENCODER BASED IMAGE COMPRESSION

Mohammad Sadegh Nasr[⋆,1,2] *Amir Hajighasemi*[⋆,1,2] *Paul Koomey*[1,2]
Parisa Boodaghi Malidarreh[1,2] *Michael Robben*[2,4] *Jillur Rahman Saurav*[1,2]
Helen H Shang[1,3] *Manfred Huber*[1] *Jacob M. Luber*[†,1,2,4]

[1] Department of Computer Science and Engineering, University of Texas at Arlington
[2] **Multi-Interprofessional Center for Health Informatics, University of Texas at Arlington**
[3]Division of Internal Medicine, Ronald Reagan University of California Los Angeles Medical Center
[4] Department of Bioengineering, University of Texas at Arlington

ABSTRACT

In this paper, we introduce a Variational Autoencoder (VAE) based training approach that can compress and decompress cancer pathology slides at a compression ratio of 1:512, which is better than the previously reported state of the art (SOTA) in the literature, while still maintaining accuracy in clinical validation tasks. The compression approach was tested on more common computer vision datasets such as CIFAR10, and we explore which image characteristics enable this compression ratio on cancer imaging data but not generic images. We generate and visualize embeddings from the compressed latent space and demonstrate how they are useful for clinical interpretation of data, and how in the future such latent embeddings can be used to accelerate search of clinical imaging data.

Index Terms— Histopathology cancer slides, autoencoder, image compression, latent space, clinical image search

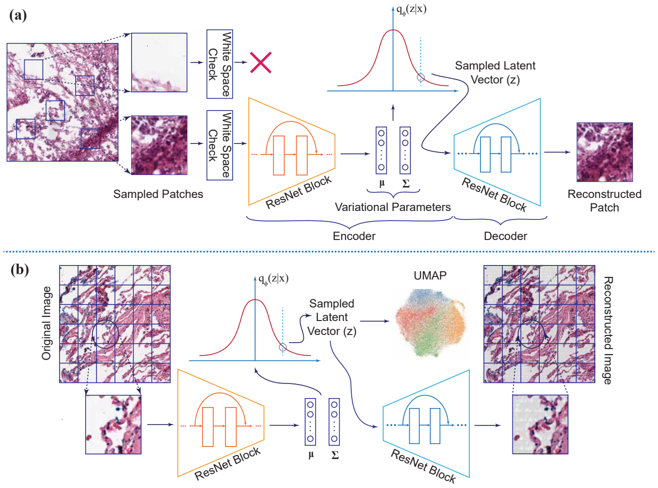


Fig. 1. **(a)** Overview of the VAE training pipeline. **(b)** Overview of the pipeline at inference. For generating UMAP plots, a similar patch sampling as training is used.

1. INTRODUCTION

Histopathological images derived from cross sectional tissue microscopy are used in the clinical setting for diagnosis of various diseases and conditions [1]. Hemotoxylin and Eosin (H&E) staining, which introduce a contrast dye for the discernment of nuclear and cytoplasmic structures, has long been used to determine carcinomal regions of excised tissue from cancer patients [2]. For this reason, databases of tumor patient slides, such as the NIH Genomic Data Commons (GDC), have been compiled for researchers to access tens of thousands of cancer patients' histopathological data. The GDC itself contains more than 30,000 Whole Slide Images (WSIs) which, with each slide representing over a billion pixels each, is stored on over 20 TB of data. Most purposes, from retrieval to transmission, local storage, and data analysis would benefit from efficient, indexable storage structures of this WSI data [3]. This is especially applicable to image search algorithms for large whole slide image databases [4].

Several solutions have been proposed for the efficient storage and indexing of cancer tissue image data. Classic compression formulas such as JPEG2000 can successfully reduce image size at a compression ratio of 32:1 before becoming unusable for histopathological classification of malignancy [5]. Compression and scaling has also been found to adversely effect tissue segmentation up to ratios of 50:1 [6]. In contrast to discrete cosine transformation models, neural networks have been proven to retain high efficiency and fidelity in the lossy compression of image data [7]. While neural networks seek to store image data in latent space representations, not every network does this at equivalent efficiency or accuracy [8]. Several studies have demonstrated that Variational Auto En-

[⋆]These authors contributed equally to this work.
[†]Responsible author. Email: jacob.luber@uta.edu

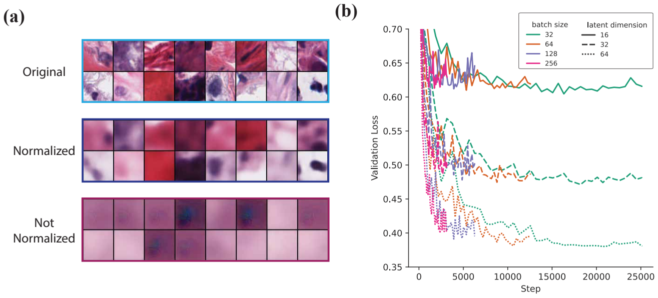


Fig. 2. **(a)** Example of how normalization affects the performance of our pipeline. Both models are trained using the exact same hyper-parameters (`latent_dim` = 64). **(b)** The effect of batch size and latent dimension of validation loss. For better visualization, early stopping is not used for these experiments.

coders (VAEs) retain higher image quality and lower noise ratios at extreme compression ratios [9, 10, 11]. Tellez et al., [12] showed in a benchmark study that VAE compression of medical tissue images to a latent space of 128 (>5000:1 compression ratio) retained the most details of the original whole slide image compared to 4 other encoders. In the current study, we develop a VAE to compress and index images in latent space for fast complex search of whole slide H&E cancer images.

2. METHODS

2.1. Dataset

The dataset we used for this study is publicly available at the NCI GDC data portal (Sec. 5). These are real samples from cancer patients in the US, and all samples contain cancerous cells. For this study, We downloaded 20% of the available `.svs` samples for primary sites: ”Brain”, ”Breast”, ”Bronchus and Lung”, and ”Colon” (647, 551, 580, and 267 images, respectively).

2.2. Latent Variables and VAE

For an observation $\boldsymbol{x}^{(i)}$, its latent vector of variables is assumed to be an unobserved random variable $\boldsymbol{z}^{(i)}$ sampled from a lower dimension space (latent space) that is involved in producing $\boldsymbol{x}^{i}$ in a random process [13]. For a particular task, it is assumed that using latent variables removes non-informative dimensionality and is suitable for downstream machine learning tasks. Since the latent space is unobserved, latent variables should be somehow inferred. Autoencoders and Variational Auto Encoders are two very effective methods for inferring these latent variables and encoding very high-dimensional data into highly compressed latent space with minimal loss of information. VAEs, as opposed to regular Auto Encoders, try to find a distribution for each latent variable, rather than single point estimate, resulting in a regularized latent space with generative capability.

VAEs are comprised of two parts: an encoder and a decoder (Fig.1-a). If the latent variable $\boldsymbol{z}^{(i)}$ and data point $\boldsymbol{x}^{(i)}$ are sampled from parametric probability distributions $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ for some parameter $\boldsymbol{\theta}$, then the encoder will try to estimate the approximate posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ with variational parameter $\boldsymbol{\phi}$. The decoder tries to find the likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. The model can be trained by minimizing the loss introduced in Eq.1 over all observations ([13]). The first term of the loss is called the Kullback–Leibler (KL) divergence term, which is introduced to ensure that the variational approximation is as informative as the generative true posterior. The second term is reconstruction loss, which makes sure the generated output from the learned latent distribution is close to the original input. In our experiments, we used a weighted loss with a KL term coefficient of 0.1.

$$
\begin{aligned}
\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)}\right) = & - D_{KL}\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}||p_{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\right)\right) \\
& + \mathbb{E}_{q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right)}\left[\log p_{\theta}\left(\boldsymbol{x}^{(i)}|\boldsymbol{z}\right)\right]
\end{aligned} \tag{1}
$$

2.3. Training and Inference Pipelines

As illustrated in Fig.1, we use two pipelines for training and inference. For the training phase (Fig.1-a), a selected number of patches from whole slide images (WSIs) in the training and validation set are randomly sampled. A white space filter is utilized to ensure that these patches are not blank, and that they do not overlap. The mean and standard deviation of all patches sampled from the training set is calculated and all patches are normalized using the standard score method with these values (not shown in Fig.1). The inverse transformations are also stored to be applied to the outputs of the model.

Our model assumes a Gaussian prior and a Gaussian approximate posterior. The encoder learns the parameters of the Gaussian prior and the decoder uses a re-parameterized sample from this prior and tries to reconstruct the input. Both encoder and decoder use ResNet18 ([14]) architectures. A ResNet50 archiecture (not-shown) provides similar performance; ResNet18 was selected to keep the number of model parameters as small as possible for future downstream deployment in the clinic.

During inference (Fig.1-b), to perform a a compression/decompression task, the test image is fully tiled. Each patch is then fed into the trained networks and stitched together once all patches are reconstructed. However, for the UMAP experiment, the same patch sampling algorithm used for training is used to generate random patches to be fed to the model. For the reconstruction task, we want a whole image, but for the UMAP plot, sample latent variables are enough.

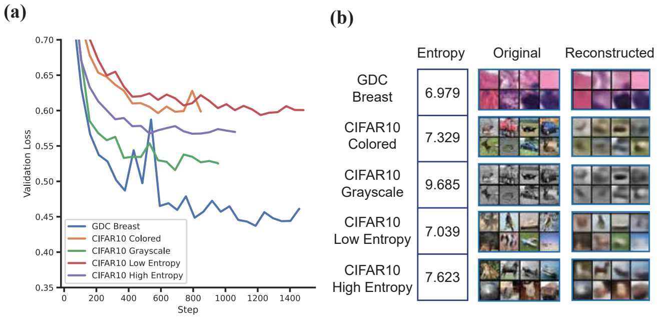


Fig. 3. Effect of dataset entropy and color content on peroformance. All hyper-parameters are the same for all 5 models.

2.4. Dimension Reduction and UMAP

Uniform Manifold Approximation and Projection ([15]) is a manifold based dimensionality reduction algorithm used for visualizing and clustering high dimensional datasets. This algorithm tries to reduce the points in a manner that the distance between resulting points would be still meaningful. UMAP is utilized to visualize and demonstrate that not only do the latent vectors learned by our pipeline provide visually accurate decompressed images, but also they contain relevant clinical information from different cancer types (Fig.4). UMAP can use many metrics for distance calculation; ”cosine similarity” was selected for its ability to capture correlation features.

3. SETTINGS AND EXPERIMENTS

In this section, we summarize different scenarios and their experimental settings used for training and validating the compression and latent space approximation of histopathology images, and establish that the compression ratio our pipeline achieves is state of the art.

3.1. Training Settings

For hyperparamter tuning, the effect of normalization of data on the quality of outcome was tested (based on visual inspection), and it was concluded that normalization is necessary for acceptable results (Fig. 2-a). Since all datasets are normalized using the same procedure, the validation can be perceived as a metric to compare the performance of different models on different datasets. All experiments are conducted using and early stopping on validation loss with `patience` = 5 unless mentioned otherwise.

As illustrated in Fig. 2-b, higher batch sizes result into faster objective minimization, but lower batch sizes eventually results in better validation loss due to a higher regularization effect ([16]). To take the middle ground, all experiments were conducted using a batch size of 128 unless mentioned otherwise. Also, as expected, higher latent dimensions resulted into a better performance.

The model is developed with PyTorch Lightning API. All experiments were conducted using the DDP parallelization strategy on an NVIDIA DGX A100 with 8, 80 GB A100 GPUs, and a learning rate of 10^{-4}.

3.2. Compression Experiments

Experimental results demonstrate a better performance of our compression model on histopathology slides than is achieved on images of every day objects datasets such as in CIFAR10 ([17]). We first hypothesised that this diffeence is rooted in the difference of entropy between the average image in these two datasets. Entropy is a way of calculating the context information of a datapoint. We reasoned that low entropy images are more compressible han high entropy ones. Therefore, we divided the CIFAR10 dataset by entropy with a high entropy fold (average entropy = 7.623) and a low entropy fold (average entropy = 7.039), each containing 30,000 images, and ran two experiments to see which one is more compressible when fed through our model. For both experiments, batch size was set to 256, latent dimension was set to 16, and the input images were of dimension $32 \times 32 \times 3$. The results are shown in Fig. 3. The final validation loss for low entropy and high entropy datasets are 0.601 and 0.570, respectively contradicted our original hypothesis. We ran the same experiment on the same number of patches sampled for the breast cancer slides, and although having lower entropy, it showed a better performance (numbers are reported in Fig3). Hence, we concluded that entropy is not a reliable factor to explain the SOTA performance of our VAE compression pipeline on cancer imaging data.

We then hypothesized that color distribution may be a contributing factor. H&E slides are limited to the colors present in tissue, while CIFAR10 images have a more diverse color distribution. For this hypothesis, we randomly chose 30,000 images from CIFAR10 dataset. Using the same settings, we ran one experiment on the sampled images and another on the same images but with grayscale transformation to eliminate olor diversity. The final validation loss for colored dataset is 0.599 and for the grayscale dataset is 0.525 (Fig. 2-a). The lower validation loss indicates that less color content can be attributed to a better comprehensibility.

3.3. Validation Experiments

In order to examine whether the latent space preserves necessary information for downstream clinical tasks, we tested the accuracy of original slide images against regenerated slide images on CLAM ([18]), the state-of-the-art model in lung cancer classification from H&E slides. We first used CLAM on the original test set for the two classification tasks, i.e. ”tumor vs. normal” and ”sub-typing” between Lung Adenocarcinomas (LUAD) and Squamous Cell Carcinomas (LUSC). Then, we created a reconstructed (post compression) version of the test set using our inference pipeline (Fig. 1-b). This

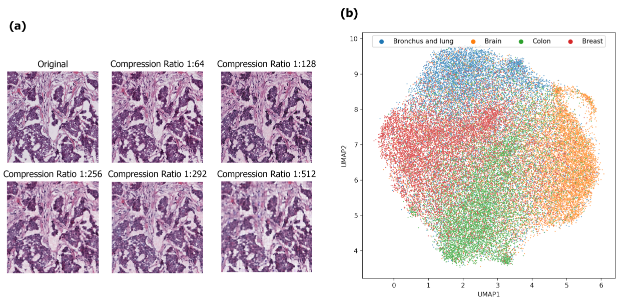


Fig. 4. **(a)** The reconstruction results for breast cancer tissues at 5 different compression ratios. **(b)** UMAP plot generated on 4 different tissue types with a compression ratio of 1:64.

reconstructed test set was then run through the same classification problems as the original images. We then calculated the percentage of the images that had the same label for both original and reconstructed images over all test images as a measure of performance and observed that our compression did not decrease performance on clinical application tasks.

To test the clinical information preservation of the latent space, we chose a model trained on lung tissues with the highest compression ratio (1:512) in our pipeline (Sec. 3.3). This compression ratio is twice as high as the best models introduced in the literature ([12, 19]). For the ”tumor vs. normal” task, the reconstructed images did not show loss of performance, however, this level of compression made it difficult for lung cancer sub-typing model to perform as before.

We used 900 images from the GDC TCGA (Sec. 5) including 450 samples for each LUAD and LUSC sub-types for training and 100 images (50 LUAD, 50 LUSC) for testing. The CLAM model has 10-fold validated pre-trained weights; thus, we calculated the performance in a 10-fold setting, too.

3.4. UMAP Experiments

To show that the latent space preserves important and clinically relevant information, 4 models were trained with a latent space of size 64 on different tissue types (brain, breast, bronchus and lung, and colon) on 20,000 patches of size 64×64 pixels, and tested them on 10,000 patches from their respective tissue type. We then ran the latent vectors of the test patches through the UMAP algorithm using the ”cosine” distance as the similarity metric. The results are shown in Fig.4.

4. RESULTS AND CONCLUSION

Fig. 4-a shows the impact of various compression ratios on VAE output images. At lower compression ratios, reconstructed images more closely resemble original input images. Importantly, we see a marked improvement in histologic features that are critical for interpretability such as refined cell-to-cell borders and sharper demarcation of cytoplasmic vs. nuclei compartments. Moreover, in Fig. 4-b, we use UMAP to visualize the latent space vectors learned by our pipeline. The UMAP captures intra-tumor and across-tumor relationships, separating all four tissue types into distinct clusters. Interestingly, clusters of brain and colon cancers share overlapping boundaries whereas the breast cancer cluster is uniquely separated from the brain cancer cluster. Also, the UMAP identifies a distinct sub-cluster of brain tumor samples that does not overlap with any other cancer types.

We envision our pipeline being useful to clinicians and researchers across multiple domains. One potential application is more accurate sub-typing and diagnoses of poorly understood cancers. A notable example of this is brain cancer, which contains over 150 different histologic subtypes, many of which are so rare that a pathologist may only encounter a handful of cases in his or her career ([20]). In our UMAP visualization of the latent space, there is an unexpected but distinct sub-cluster of brain tumor samples that does not overlap with other cancer types (Fig. 4-b). Further characterization of this sub-cluster and its unique attributes could provide novel insights into intra-tumor relationships in brain cancer.

Our pipeline also facilitates experiments across different tumor types. The latent space separates breast, colon, lung/bronchus, and brain tissue into unique clusters, demonstrating the preservation of important histological features. Interestingly, we see a closer clustering between brain and colon cancer versus brain and breast or lung (Fig. 4-b). More investigation into these relationships is warranted – one possible explanation of this phenomenon could be due to both brain and colon tissue containing ganglion nerve cells whereas breast and lung tissue do not. In the future, our embedding approach could be deployed to a hospital system and linked to the electronic health record (EHR) to help clinicians diagnose patients with rare disorders: the images closest to that of the input patient in UMAP embedding have the most similarities, and their records could be retrieved to better contextualize a differential diagnosis for the query patient.

However, our pipeline carries several limitations. To start, we will need to further explore acceptable thresholds of reconstruction loss introduced via our VAE-based architecture. Additionally, our model architecture lacks human interpretable features, which may lead to higher levels of end-user distrust as “peeking under the hood” to audit our model for biases or errors may be more limited. Along these lines, any insights or novel conclusions will still require manual review and interpretation by human pathologists. In future iterations of this work, we intend to improve upon these areas.

5. DATA AND CODE AVAILABILITY

All dataset used in this study are accessible from NCI GDC portal at portal.gdc.cancer.gov/repository. The code is also accessible at github.com/jacobluber/uta_cancer_search.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by the Genomic Data Commons (GDC) provided by the National Cancer Institute of the National Institues of Health (NIH/NCI). Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

This work was supported by the Cancer Prevention and Research Institute of Texas (CPRIT) Recruitment of First-Time, Tenure-Track Faculty Members Grant (RR220015) (JML) and University of Texas System Rising STARs award (JML). The authors wish to thank Abhishek Dubey, Peng Jiang, , and Eytan Ruppin for fruitful discussions. The authors want to thank Thomas Shipman and Edward Gonzalez for their support of computing infrastructure.

8. REFERENCES

[1] Metin N Gurcan, Laura E Boucheron, et al., "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.

[2] Lei He, L Rodney Long, et al., "Histology image analysis for carcinoma detection and grading," *Computer methods and programs in biomedicine*, vol. 107, no. 3, pp. 538–556, 2012.

[3] M Khalid Khan Niazi, Yuzhang Lin, et al., "Pathological image compression for big data image analysis: Application to hotspot detection in breast cancer," *Artificial intelligence in medicine*, vol. 95, pp. 82–87, 2019.

[4] Shivam Kalra, Hamid R Tizhoosh, et al., "Yottixel–an image search engine for large archives of histopathology whole slide images," *Medical Image Analysis*, vol. 65, pp. 101757, 2020.

[5] Elizabeth A Krupinski, Jeffrey P Johnson, et al., "Compressing pathology whole-slide images using a human and model observer evaluation," *Journal of pathology informatics*, vol. 3, no. 1, pp. 17, 2012.

[6] Juho Konsti, Mikael Lundin, et al., "Effect of image compression and scaling on automated scoring of immunohistochemical stainings and segmentation of tumor epithelium," *Diagnostic Pathology*, vol. 7, no. 1, pp. 1–9, 2012.

[7] Hamdy S Soliman and Mohammed Omari, "A neural networks approach to image data compression," *Applied Soft Computing*, vol. 6, no. 3, pp. 258–271, 2006.

[8] Sonain Jamil, Md Piran, et al., "Learning-driven lossy image compression; a comprehensive survey," *arXiv preprint arXiv:2201.09240*, 2022.

[9] Yueyu Hu, Wenhan Yang, et al., "Learning end-to-end lossy image compression: A benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[10] Salvator Lombardo, Jun Han, et al., "Deep generative video compression," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] M Akín Yílmaz, Onur Keleş, et al., "Self-organized variational autoencoders (self-vae) for learned image compression," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3732–3736.

[12] David Tellez, Geert Litjens, et al., "Neural image compression for gigapixel histopathology image analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 567–578, 2019.

[13] Diederik P. Kingma and Max Welling, "Auto-Encoding Variational Bayes," May 2014, arXiv:1312.6114 [cs, stat].

[14] Kaiming He, Xiangyu Zhang, et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.

[17] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[18] Ming Y Lu, Drew FK Williamson, et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[19] Chengkuan Chen, Ming Y. Lu, et al., "Fast and scalable search of whole-slide images via self-supervised deep learning," *Nature Biomedical Engineering*, pp. 1–15, Oct. 2022, Publisher: Nature Publishing Group.

[20] Thomas Roetzer-Pejrimovsky, Anna-Christina Moser, et al., "The Digital Brain Tumour Atlas, an open histopathology resource," *Scientific Data*, vol. 9, no. 1, pp. 55, Feb. 2022.

4

Paper 3: Readiness of Histopathology Slide Search for Clinic

4.1 Introductory Comments

This chapter presents the paper ”Histopathology Slide Indexing and Search: Are We There Yet?” which delves into the critical aspect of developing efficient and accurate search engines for histopathology slides. The focus of this study is on enhancing the capabilities of digital pathology by facilitating the retrieval of relevant cases, thereby significantly improving diagnostic processes and patient care.

4.1.1 Benefits of an Efficient Histopathology Search Engine

1. **Enhanced Diagnostic Accuracy**: Having a robust search engine in histopathology can dramatically improve the accuracy of diagnoses. By retrieving the most similar cases to the patient at hand, pathologists can compare and contrast current cases with previously diagnosed ones, leading to more informed and precise diagnoses. This comparative analysis is especially beneficial in complex or rare cases where historical reference can provide crucial insights.

2. **Speed and Efficiency in Clinical Practice**: An efficient search system significantly reduces the time spent by pathologists in finding relevant cases or reference materials. This efficiency is not only beneficial in terms of workflow optimization but also crucial in time-sensitive situations where quick diagnosis can lead to faster treatment decisions.

3. **Educational and Research Benefits**: Such a search engine also serves as a valuable educational tool for medical students and trainees, allowing them to access a vast repository of cases for study and comparison. For researchers, it facilitates the exploration of pathological data, enabling them to identify patterns and correlations that may not be immediately apparent.

4. **Personalized Patient Care**: With the ability to quickly retrieve similar cases, pathol-

ogists and clinicians can offer more personalized care to patients. Understanding how similar cases have progressed and responded to treatments can inform more tailored and effective treatment plans.

5. **Advancing Digital Pathology**: Implementing a sophisticated search engine is a significant step forward in the digitization of pathology. It not only enhances the current capabilities of digital pathology systems but also opens up new possibilities for integrating AI and machine learning tools to further refine search and analysis processes.

The following pages will comprehensively detail the manuscript of ”Histopathology Slide Indexing and Search: Are We There Yet?” including the title page, abstract, main body, references, and any additional material. The chapter is structured to align with the publisher’s requirements and is renumbered for consistency within the overall thesis/dissertation.

HISTOPATHOLOGY SLIDE INDEXING AND SEARCH: ARE WE THERE YET?

Helen H. Shang, MD, MS*, **Mohammad Sadegh Nasr***, Jai Prakash Veerla, Jillur Rahman Saurav, Amir Hajighasemi, Parisa Boodaghi Malidarreh, Manfred Huber, PhD, Chace Moleta, MD, Jitin Makker, MD, and Jacob M. Luber, PhD

Submitted to the New England Journal of Medicine AI (NEJM AI)

The first page of the article begins on the next page.

* Indicates co-first authors.

Histopathology Slide Indexing and Search: Are We There Yet?

Helen H. Shang, MD, MS*[,1,3] Mohammad Sadegh Nasr[*,3,4] Jai Prakash Veerla[3,4]
Jillur Rahman Saurav[3,4] Amir Hajighasemi[3,4] Parisa Boodaghi Malidarreh[3,4]
Manfred Huber, PhD [3] Chace Moleta, MD[2] Jitin Makker, MD[2]
Jacob M. Luber, PhD†[,3,4,5,6]

[1] *Department of Internal Medicine, Ronald Reagan University of California Los Angeles Medical Center*
[2] *Department of Pathology & Laboratory Medicine, Ronald Reagan University of California Los Angeles Medical Center*
[3] *Department of Computer Science and Engineering, The University of Texas at Arlington*
[4] *Multi-Interprofessional Center for Health Informatics, The University of Texas at Arlington*
[5] *Department of Bioengineering, The University of Texas at Arlington*
[5] *Department of Biology, The University of Texas at Arlington*

Abstract

The search and retrieval of digital histopathology slides is an important task that has yet to be solved. In this case study, we investigate the clinical readiness of four state-of-the-art histopathology slide search engines, Yottixel, SISH, RetCCL, and HSHR on both unseen datasets and several patient cases. We provide a qualitative and quantitative assessment of each model's performance in providing retrieval results that are reliable and useful to pathologists. We found high levels of performance across all models using conventional metrics for tissue and subtyping search. Upon testing the models on real patient cases, we found the results were still less than ideal for clinical use. Based on our findings, we propose a minimal set of requirements to further advance the development of accurate and reliable histopathology image search engines for successful clinical adoption.

1 Introduction

As histopathology slides become increasingly digitized, the process of manually searching and retrieving slides has become increasingly more time-consuming for pathologists (Hegde et al.; Z. Li et al.). Recently, there has been growing interest in the development of automated search and retrieval systems for digital histopathology slides (Kalra et al.; C. Chen et al.; Wang et al.; Hegde et al.; Kalra et al.; S. Li et al.), which can help pathologists identify similar cases in both developing and narrowing a differential diagnosis. These systems leverage advances in artificial intelligence and machine learning to analyze large volumes of slides efficiently and accurately.

The exploration of medical image databases predominantly relies on content-based image retrieval (CBIR) (Kalra et al.; Z. Li et al.; Lew et al.). CBIR systems initially transform images into a feature-based database accompanied by corresponding indices. Subsequently, by utilizing a similarity metric, the retrieval process simplifies into a k-nearest neighbors problem. Extracting features from extensive whole slide images (WSI) is typically achieved through either the sub-setting method, which focuses on a small section of a large pathology image to significantly reduce processing time, or the tiling method, which segments images into manageable patches (i.e., tiles) for intra-patch processing (Kalra et al.; Gutman et al.).

Among the recent end-to-end systems proposed for histopathology image search, Yottixel (Kalra et al.), SISH (C. Chen et al.), RetCCL (Wang et al.), and HSHR (S. Li et al.) have emerged as influential contenders, showcasing promising outcomes. Yottixel pioneered the processing of large-scale WSIs by introducing the concept of mosaics. Instead of extracting features from the entire WSI, Yottixel's approach involves extracting features from mosaic tiles using a DenseNet-based feature extractor. Additionally, Yottixel incorporates the notion of barcoding (Tizhoosh; Tizhoosh et al.) to facilitate expedited retrieval by binarizing the extracted

[*] These authors contributed equally to this work.
[†] Corresponding author. Email: jacob.luber@uta.edu

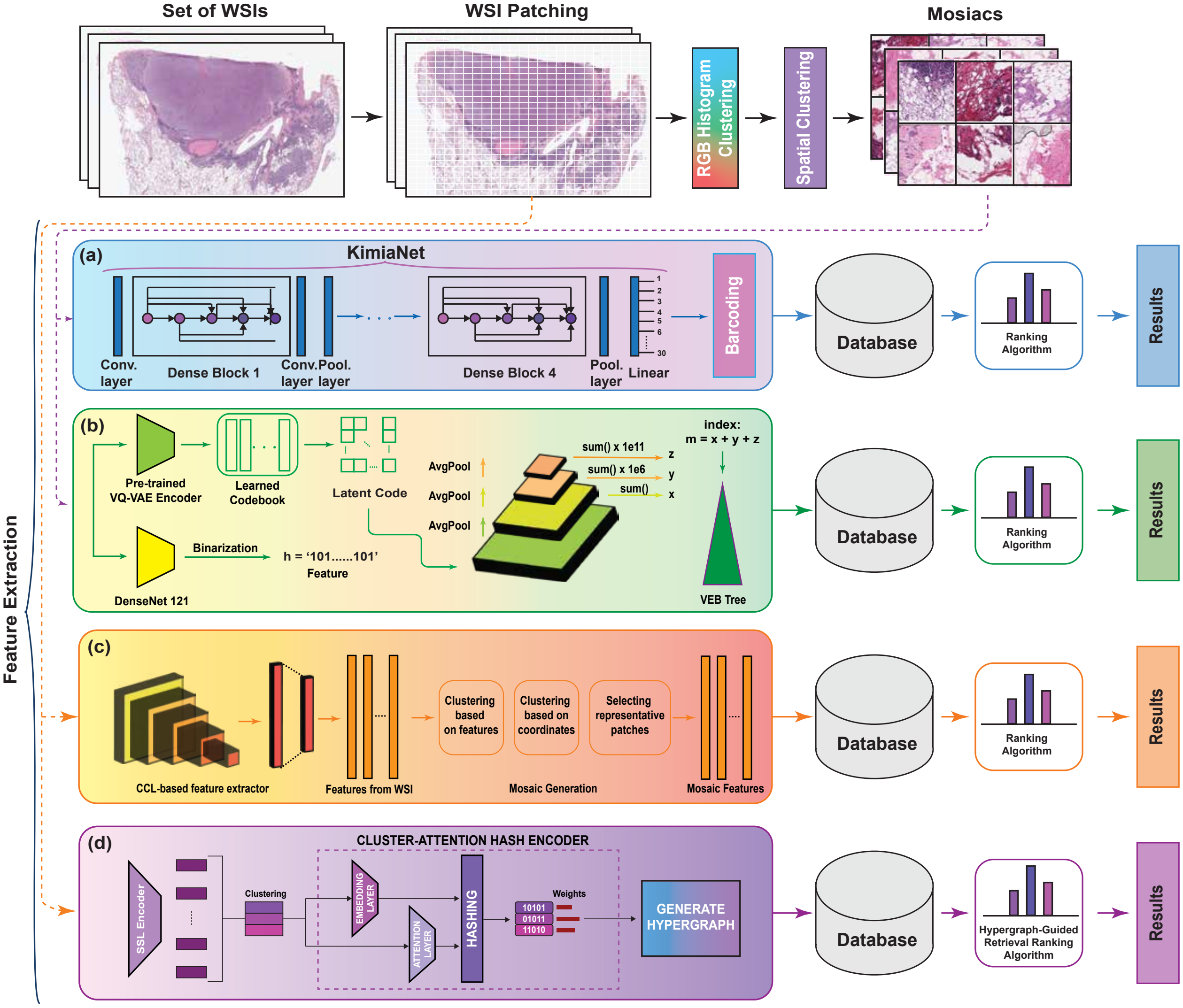


Figure 1: A summary of feature extraction and database creation processes proposed by **(a)** Yottixel (Kalra et al.), **(b)** SISH (C. Chen et al.), **(c)** RetCCL (Wang et al.), and **(d)** HSHR (S. Li et al.). The feature extractor of Yottixel is switched with KimiaNet (Riasatian et al.).

features. Similarly, SISH adopts a framework very similar to Yottixel but incorporates an additional VQ-VAE-based (Oord et al.) feature extractor. SISH also introduces advanced VEB tree-based (van Emde Boas) indexing and ranking algorithms to enhance the quality of the retrieved samples. In contrast, RetCCL employs the mosaic concept as well, but uniquely converts WSIs to mosaics after extracting features from tiled WSIs. Moreover, RetCCL introduces an effective contrastive-based feature extractor to improve feature quality. Finally, HSHR expands the idea of using Self-Supervised Learning (SSL) to both extract the mosaics and creating hash codes from them. It uses SimCLR (T. Chen et al.) to train the feature extractor and uses MOCO (He et al.) to train the Cluster-Attention Hash Encoder (CaEncoder). The results are then processed to create a hypergraph which leads to similarity-based WSI retrieval.

The introduction of successive systems claiming to have achieved state-of-the-art performance in the search and retrieval of histopathology slides, often supported by statistical metrics demonstrating agreement with trained pathologists' judgments, raises the fundamental question of whether this problem has been satisfactorily addressed. Specifically, it prompts an inquiry into the readiness of these systems for deployment in clinical settings, where they can provide genuinely valuable information to pathologists, especially in challenging cases where even the most experienced group of pathologists struggle to reach a consensus.

In this case study, we evaluate these models on patient cases from our health system and several external datasets. Our objective is to provide a quantitative analysis of these models' performances on unseen slides while offering a qualitative assessment of the usefulness of these models in the clinical setting and potential areas of improvement. To ensure fairness, we constructed each model's database using a fixed number of slides from The Cancer Genome Atlas (TCGA) (Weinstein et al.), while employing the same feature extractors as published by the original authors.

In subsequent sections, we provide an overview of our methods and approach towards implementation (Section 2). We then report our quantitative and and qualitative analysis of model performance (Section 3). Finally, we discuss the current state of histopathology slide search engines and propose a set of minimal requirements for real-world deployment based on our findings (Section 4).

2 Methods

2.1 Search Engines

Search engines commonly comprise two fundamental components: indexing and database generation, as well as ranking and retrieval. Given the large-scale nature of the images involved in this study, feature extraction becomes imperative for effective indexing. In terms of ranking, a suitable similarity measure is crucial, followed by post-processing steps to ensure result quality. In the supplementary methods section, we provide an overview of the feature extraction techniques, database indexing approaches, employed similarity measures, and result ranking methodologies utilized by the four primary methods under investigation (Fig. 1). Please be advised that all the hyper-parameters employed here are the parameters recommended by the authors of the models. We additionally share our code that we used to re implement methods that were not made available by the authors.

To summarize, The **Yottixel** method creates a mosaic of patches from whole slide images (WSIs), applies a feature extractor, KimiaNet (Riasatian et al.), and generates binary codes, or barcodes, from the extracted features. These barcodes represent each WSI, form a database, and enable the retrieval of related slides or patches based on the median of the minimum Hamming distances (Hamming).

The **SISH** method also generates a mosaic and uses DenseNet for feature extraction similar to Yottixel, but it further employs a pretrained VQ-VAE for index creation. Querying in SISH involves converting a slide into a mosaic, generating indices and features, and utilizing the ”guided VEB search” algorithm to retrieve top slides based on Hamming distance.

RetCCL takes inspiration from both Yottixel and SISH but applies a unique approach by obtaining contrastive-based feature vectors for each patch within the segmented foreground tiles. The method employs a clustering-guided contrastive learning method with two InfoNCE losses to capture irregular regions in patches, which is particularly important given the prevalence of normal cells in WSIs (Oord et al.).

Finally, **HSHR** first trains a encoder in a self-supervised manner on a small subset of patches extracted from database slides. By clustering the features from this encoder, it creates mosaics for each slide, and then passes the the features of mosaics to the CaEncoder. By following teh guidelines of MOCO (He et al.), they train the desired CaEncoders and this way, they are able to create hashings and weights for each slides. Hashing and weights are then incorporated into building a hypergraph for the database. Every query slide is then considered a new node and hyperedge in this hypergraph database and similarity scores can be calculated for it. Per the authors' emphasize on global perception of WSIs, HSHR is not designed to be used with patch retrieval tasks.

A more detailed explanation of these models also can be found in Supplementary Materials (Supplementary Section 4). Specifically, Supplementary Algorithms 1 to 4 would summarize the process a query slide would undergo in all the discussed methods. Moreover, time complexity of various stages of ranking and retrieval of these methods are also juxtaposed in Supplementary Table 4.

2.2 Database Slides

To ensure a fair comparison among all models, it was necessary to have consistent slides in the databases of each model. We constructed the database using slides available in TCGA (Weinstein et al.). Given our focus on lung, brain, and liver as primary sites for testing (see Section 2.4), it was essential to include slides from these sites in the databases. Additionally, to introduce a challenging aspect, slides from breast and colon were added to ensure that site retrieval experiments were not trivial. For each site, we randomly selected between 50 to 75 slides from subtypes containing at least 75 slides. The varying number of slides aimed to introduce class imbalance, mirroring real-world scenarios where some subtypes have more samples than others. Importantly, none of the slides in the database shared the same patient ID. The resulting database comprised 508 slides from 5 different sites and 8 different subtypes (Supplementary Table 1).

It is worth noting that in each experiment, we utilized the pre-trained feature extractors provided by the respective authors (except for the backbone of HSHR, see Supplementary Section 4). These feature extractors were trained on different datasets of varying sizes. The relatively small size of our database does not affect the performance of these models, as the only aspect influenced by data size is the feature extractor. As long as we have samples of the same class as the query within the database, a correctly functioning model should be capable of retrieving them.

Due to preprocessing criteria, we were not able to include 6 slides for Yottixel, 1 slide for SISH,. 4 slides for HSHR in the database. These slides are listed in Supplementary Table 3.

2.3 Test Datasets

In order to conduct fair quantitative experiments and avoid data leakage, we needed to acquire test slides that were not seen by the encoders of these models. Table 1 summarizes the all the datasets used in validation experiments. Except from the in house UCLA dataset, all other datasets are downloaded from the publicly available Cancer Imaging Archive database (Clark et al.) (Supplementary Table 2). For all the datasets, we made sure not to have samples from the same patient using the patient identifiers provided.

All UCLA slides were sourced from real clinical cases at our institution to best approximate real-world scenarios. Team members who were responsible for algorithmic implementation were blinded from the ground truth to reduce the likelihood of bias.

Table 1: Summary of test slides used for experiments. Abbreviations are based on (Kalra et al.).

Experiment	Slides	Dataset	Site	Diagnosis
UCLA	slide1	In House	lung	LUAD
	slide2	In House	brain	LGG
	slide3	In House	liver	LIHC
Reader Study	MSB-09151-01-11	CMB-CRC	colon	COAD
	MSB-09977-01-22	CMB-LCA	lung	LUSC
	Her2Pos_Case_66	Yale Her2+ Cohort	breast	BRCA
Microscope Study	34 slides	CPTAC-GBM (Leica)	brain	GBM
	34 slides	UPENN-GBM (Hamamatsu)	brain	GBM
HER2+ prediction	93 slides	Yale Her2+ Cohort	breast	BRCA
	97 slides	Yale Her2- Cohort	breast	BRCA
Ablation	85 slides	Yale Trastuzumab Cohort	breast	BRCA

2.4 Experiments

In general, we have three types of experiments: site (tissue) retrieval, subtype retrieval, and patch retrieval. We define "site" as the tissue of cancer origin and "subtype" as the final diagnosis, which is specific to the tissue type. For patch retrieval tasks, the models should return the closest patches to a query patch as opposed to WSIs to WSIs. For subtype and patch retrieval experimetns, we limited the search database to the slides with the same tissue type as the query.

Consistent with prior work on WSI search algorithms, we chose majority top-k accuracy (mMV@K) and mean average precision (mAP@K) as our quantitative metric of performance, which returns the predicted label by majority vote amongst the top K slides. For tissue search, K has commonly been 10 for mMV and 5 for mAP. As with prior studies, we choose mMV@1,3,5 and mAP@3,5 for subtype search (See Supplementary Algorithms 5 and 6).

We have designed 5 experiments to validate the methods for different purposes. The UCLA experiment aims to bring qualitative evaluation to the 3 in house slides by evaluating them for all three tasks. Reader study is designed to bring pathologists' point of view to the quality retrieved patches by the models. The microscope study is intended to measure the robustness of the models' performances with respect to different microscope brands. The Her2+ prediction tries to answer the question whether there is evidence that the models would perform differently given different sub-subtype. And finally, since the authors' of Yottixel mentioned their model would perform better using KimiaNet instead of DenseNet pretrained on ImageNet, we designed the ablation study to measure this change in performance.

3 Results

Table 2: Evaluation results of different methods on UCLA slides for primary site retrieval task.

Method	**UCLA Slides**	**MV@1**	**MV@3**	**MV@5**	**MV@10**	**AP@3**	**AP@5**
YOTTIXEL + KimiaNet	**Slide1**	0	0	0	0	0	0
	Slide2	1	1	1	1	1	1
	Slide3	0	0	0	0	0	0
SISH + DenseNet	**Slide1**	0	-	-	-	0	0
	Slide2	1	1	1	1	1	0.888
	Slide3	0	0	-	-	0	0
RetCCL	**Slide1**	1	0	1	1	1	0.750
	Slide2	0	1	1	1	0.583	0.679
	Slide3	1	1	1	1	1	1
HSHR	**Slide1**	1	1	1	1	1	1
	Slide2	1	1	1	1	1	1
	Slide3	0	0	0	0	0.500	0.500

Detailed results are included in the supplement, and a concise summary is provided here.

3.1 Tissue and subtype retrieval

A quantitative analysis of model performance on our patient cases is reported in Table 2 and Table 3. While we cannot arrive at definitive conclusions on algorithmic performance due to the few number of cases here, we observed higher mMV and mAP metrics for RetCCL on both tissue and subtype retrieval relative to the other models.

Supplementary Table 9 and Supplementary Table 10 show the results of our quantitative evaluations on GBM and BRCA datasets encompassing 344 unseen slides. Of note, UPENN GBM was not feasible for RetCCL due to extensive computational burdens. When testing the models on brain tissue retrieval, we found that RetCCL had the highest performance at mMV @ 10 although there is almost a 15 point drop in performance versus the mMV @ 10 score reported by the authors at 90.21. For subtyping on GBM versus LGG, we see that HSHR and SISH are the top performers at mMV@5 for GBM versus LBB subtyping, which are lower than prior work by the authors of HSHR showing a mMV@5 of 0.937 and 0.916 for HSHR and SISH, respectively, when tested on 3580 TGCA GBM and LGG slides.

For breast tissue retrieval on the Yale Trastuzumab dataset, Yotixxel performed the best on both metrics with a mMV@10 of 0.588 and mAP@5 of 0.650. Of note, the authors of RetCCL previously showed that Yotixxel achieved a mMV@10 of 0.663 relative to RetCCL's score of 0.914 on a set of frozen WSIs. Subtype search on the BRCA dataset was not performed as all of the cases were of the same diagnosis.

3.2 Visual review of query results

To better investigate discrepancies in performance, we reviewed the top five ranked results on subtyping and tissue search for three different WSI slides from our own patient cases as illustrated in Figure 2. Additional details on patient cases and slide preparation methods can be found in the Supplementary Section on Patient Cases.

In Figure 2, we see several errors made on tissue and subtype search. We also review of patch-level results for Yotixxel, SISH, and RetCCL on two patches from our LUAD case, one showcasing tumoral tissue and the

Table 3: Evaluation results of different methods on UCLA slides for subtype retrieval task.

Method	**UCLA Slides**	**MV@1**	**MV@3**	**MV@5**	**AP@3**	**AP@5**
YOTTIXEL + KimiaNet	**Slide1**	0	0	1	0.500	0.533
	Slide2	0	0	0	0	0
	Slide3	1	1	1	1	1
SISH + DenseNet	**Slide1**	1	-	-	1	1
	Slide2	0	0	0	0	0
	Slide3	1	1	1	1	1
RetCCL	**Slide1**	1	1	1	1	1
	Slide2	1	1	1	1	1
	Slide3	1	0	0	1	0.700
HSHR	**Slide1**	1	1	1	1	0.867
	Slide2	1	0	0	1	1
	Slide3	0	0	0	0.500	0.500

other with normal alveolar tissue (Figure 3). HSHR was excluded due to its inability for patch-level search. We found that all three algorithms are capable of retrieving patches containing tumoral and alveolar tissue but there were visual discrepancies in some of more granular features on slides. For example, while all models retrieve patches corresponding to alveoli, all three models return patches with varying degrees of necrosis, inflammation, and hyperplasia, which may point pathologists towards different diagnoses and treatments.

3.3 Reader study

We next compared model performance on patch-level retrieval results qualitatively. Supplementary Figure 1 shows the Mean Opinion Score (MOS) of seven pathologists on the top three ranked results when querying a patch containing tumor from three WSI H&E slides. We include Yottixel, RetCCL, and SISH but not HSHR given the latter's inability to perform patch-level search. For consistency, we used MOS as an evaluation metric based on prior studies on the quality of WSI search results. We found SISH had overlapping performance with Yotixxel and RetCCL due to higher variance but when comparing RetCCL versus Yotixxel, we see a statistically significant improvement in RetCCL performance. However, the Fleiss' Kappa for all algorithms combined was 0.131 suggesting low rates of agreement amongst pathologists. Quantitative performance metrics are provided in Supplementary Tables 5 and 6 on our reader study slides.

3.4 HER2+ prediction

To test the richness of feature representations, we compared the ability of the four models in distinguishing between HER2+ and HER2- BRCA (Supplementary Figures 3,4,5, & 6). At present, immunohistochemistry is required for the interpretation of HER2 status although recent work using a CNN-based architecture was capable of predicting HER2 positivity with an AUC of 0.81 on untested datasets (Farahmand et al.). All models tended to predict HER2- slides with greater precision despite the nearly 50-50 distribution of HER2- and HER2+ cases in the dataset, suggesting an ability to learn subtle features specific to HER2 status.

For this experiment, we calculated MV@10 and AP@5 for tissue retrieval task in both Her2+ and Her2- cohorts. Then using the Shapiro-Wilk normality test, and Levene's homogeneity of variances test, we checked the distribution of these two metrics. Since they did not pass the Independent T test requirements, we conducted a non-parametric Mann-Whitney U Test. We concluded that all 4 models would have a better performance in terms of AP@5 on Her2- cohort for the site retrieval task. However, for MV@10, the evidence was only significant for Yottixel and HSHR (Supplementary Table 8).

3.5 Ablation study

Figure 6 shows comparative analysis of Yottixel and SISH algorithms using two distinct networks, Kima Net and Densenet, on the Yale Trastuzumab dataset is presented. The figure is subdivided into two subplots,

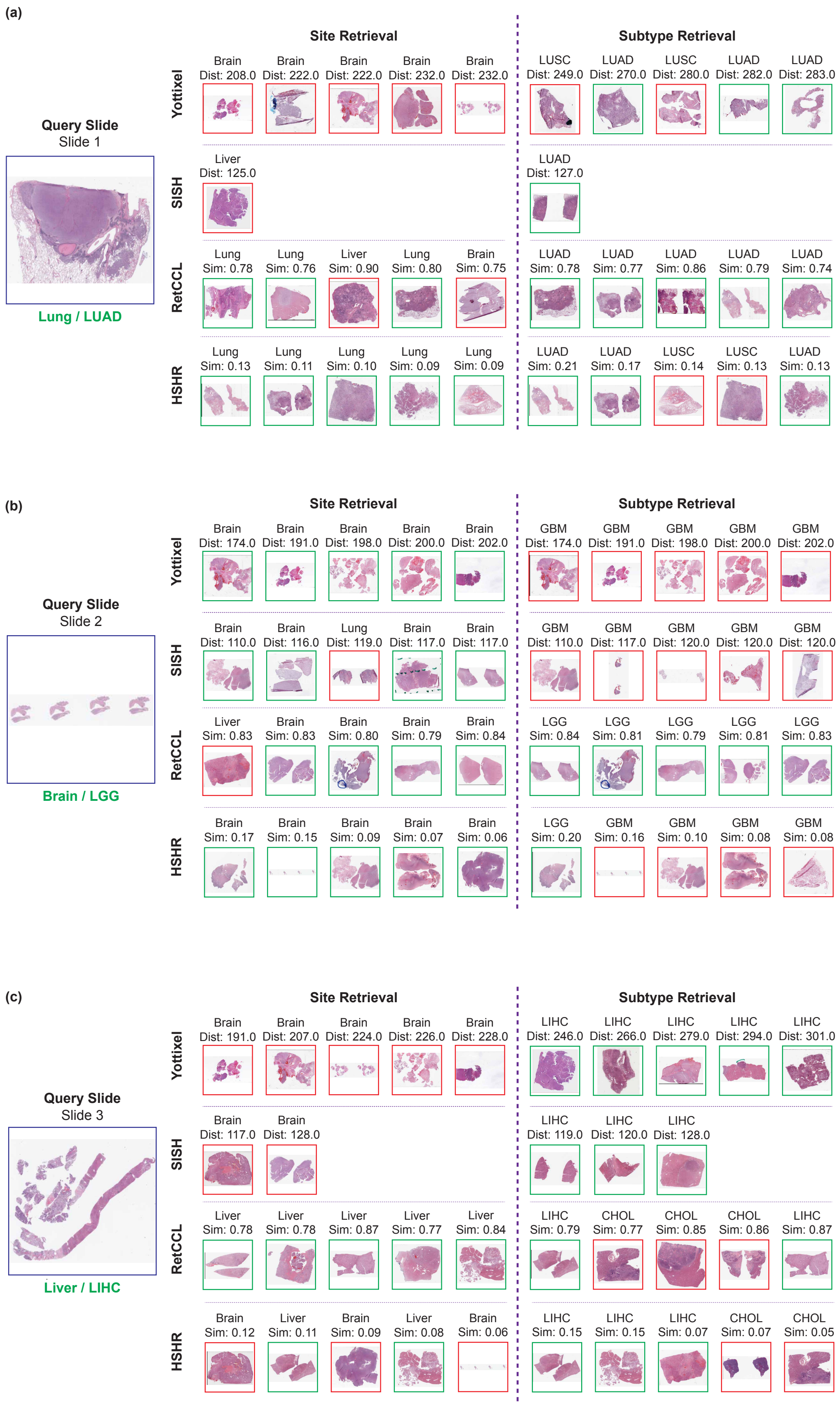
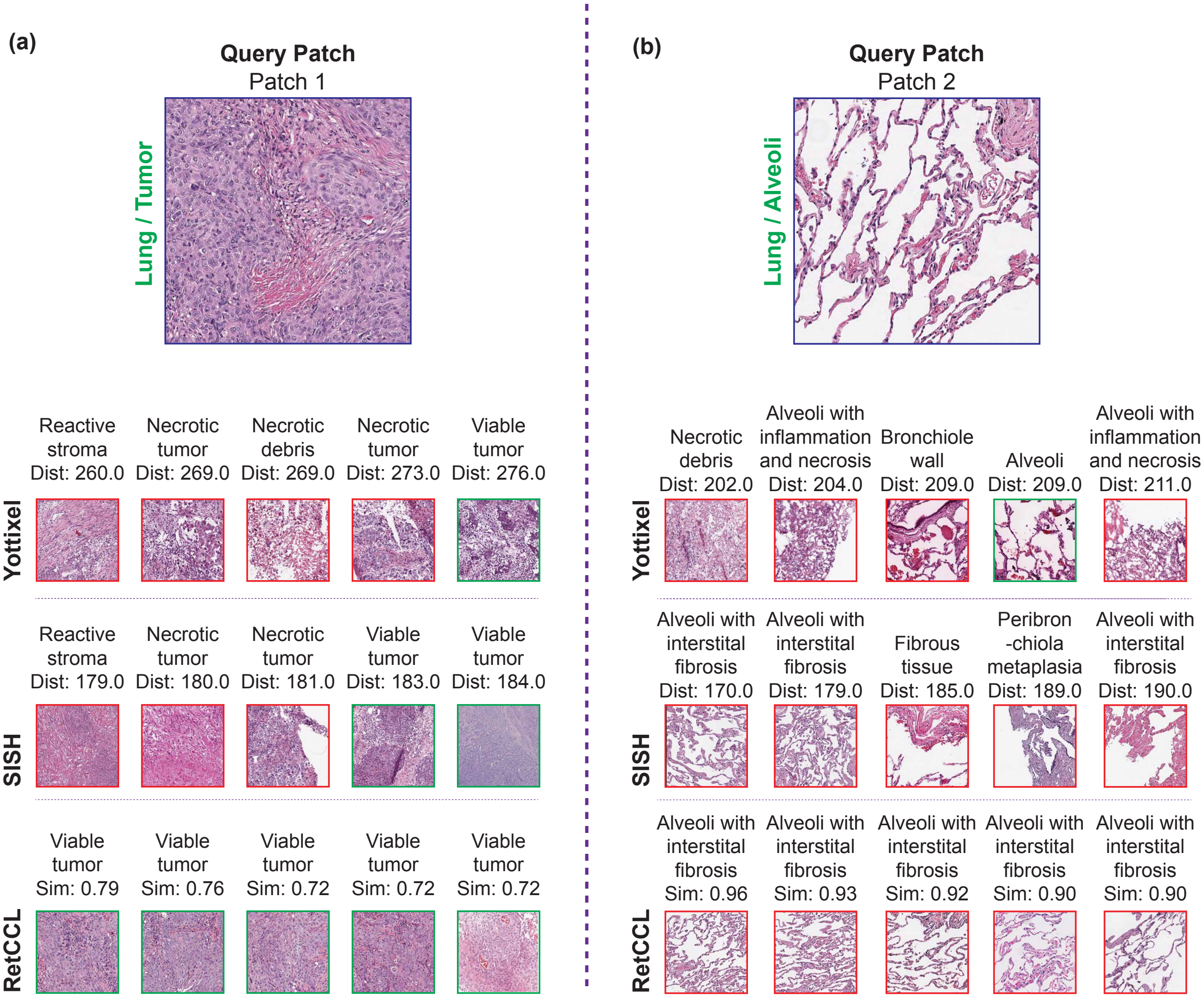


Figure 2: Results of site retrieval (left) and sub-type retrieval (right) at slide level for all three test slides. Correct labels are printed in green under query slides. Green border means correct label; red border means wrong label. For details about distances and similarities, see Section 2.1.

each illustrating the performance of a base algorithm, Yottixel or SISH, evaluated across six different metrics: mmV @ 1, 3, 5, 10 and mAP @ 3 and 5. Configurations employing Kima Net generally outperform those using Densenet across our evaluation metrics. Disparities in performance underscore the impact of the network choice on the efficacy of the algorithms, highlighting the importance of optimal network-algorithm pairing.

3.6 Microscope study

We investigated differences in performance for Yotixxel, SISH, and HSHR on two GBM datasets utilizing different microscopes, UPENN and CPTAC GBM, to test model generalizability. As RetCCL was not compatible with the UPENN dataset, it was excluded from this experiment. We found that all three algorithms had overlapping results with p-values of > 0.05 across both subtype and tissue retrieval at mMV @ 5 and 10 on both GBM datasets, indicating that differences in microscopes did not affect performance. Full results can be found in the Supplementary Table 7.



(a)
Query Patch
Patch 1
Lung / Tumor
Reactive stroma Dist: 260.0
Necrotic tumor Dist: 269.0
Necrotic debris Dist: 269.0
Necrotic tumor Dist: 273.0
Viable tumor Dist: 276.0
Yottixel
Reactive stroma Dist: 179.0
Necrotic tumor Dist: 180.0
Necrotic tumor Dist: 181.0
Viable tumor Dist: 183.0
Viable tumor Dist: 184.0
SISH
Viable tumor Sim: 0.79
Viable tumor Sim: 0.76
Viable tumor Sim: 0.72
Viable tumor Sim: 0.72
Viable tumor Sim: 0.72
RetCCL
(b)
Query Patch
Patch 2
Lung / Alveoli
Necrotic debris Dist: 202.0
Alveoli with inflammation and necrosis Dist: 204.0
Bronchiole wall Dist: 209.0
Alveoli Dist: 209.0
Alveoli with inflammation and necrosis Dist: 211.0
Yottixel
Alveoli with interstital fibrosis Dist: 170.0
Alveoli with interstital fibrosis Dist: 179.0
Fibrous tissue Dist: 185.0
Peribron -chiola metaplasia Dist: 189.0
Alveoli with interstital fibrosis Dist: 190.0
SISH
Alveoli with interstital fibrosis Sim: 0.96
Alveoli with interstital fibrosis Sim: 0.93
Alveoli with interstital fibrosis Sim: 0.92
Alveoli with interstital fibrosis Sim: 0.90
Alveoli with interstital fibrosis Sim: 0.90
RetCCL

Figure 3: Results of patch retrieval for two patches from Slide 1. Correct labels are printed in green to the left of query patches. Green border means correct label; red border means wrong label. For details about distances and similarities, see Section 2.1.

4 Discussion

In this case study, we evaluated the performance and clinical utility of state-of-the-art histopathology slide search engines, Yotixxel, SISH, and RetCCL and HSHR. To our knowledge, this is the first independent, external validation of these four models. While all models demonstrate significant advancements in the field, we also noticed clinically significant shortcomings.

Based on our findings, we propose a framework of minimal requirements to facilitate the development of these systems both fairly and transparently while minimizing patient harm and maximizing clinical utility:

1. **Richness of feature representations:** We observed several clinically relevant inconsistencies across our queried and retrieved patches, suggesting the need for improvements in feature extraction. Our ablation study also highlights the importance of encoder architecture on model performance. Large pre-trained models such as Virchow (Vorontsov et al.), with extensive training on 1.5 million images, are a potential solution, similar to natural image processing models such as VGG16 (Simonyan and Zisserman) and ResNet-50 (He et al., Deep Residual Learning for Image Recognition). This strategy can also facilitate the prioritization of downstream tasks while alleviating extensive computational burdens in the pre-training stage.

2. **Systematic and rigorous evaluations:** We found it challenging to evaluate relative model performance due to variable results across the four algorithms in our experiments. For example, RetCCL had the highest performance on brain tissue retrieval but Yottixel had better performance on breast tissue. Likewise, while we found that pathologists tended to rank patches retrieved by RetCCL highly, due to large inter-observer variation on quality, relative performance was still uncertain. Our findings highlight the need for systematic and rigorous methodologies for model evaluation.

3. **Robustness:** A vital measure of clinical applicability is model performance across diverse clinical populations and environments, as in the case of real-world health systems. This is motivated by our findings that all models experienced a notable decrease in performance when tested against unseen datasets. However, we were pleased to find that batch level effects from external factors such as microscope type did not result in statistically significant differences on tissue and subtyping performance. We propose that all future models be validated for their generalizability in addition to precision.

4. **Replicability:** We found variable degrees of replicability across the models evaluated in this work. Some were constructed for specific slide ratios and resolutions, while others were capable of handling variable slide formats. Models also differed in terms of ease of applicability and transparency due to differences in the availability of model development code and indexed databases. Model transparency fosters scientific integrity and accelerates the pace of innovation through collective efforts and unbiased evaluations.

5. **Clinical benefit:** For clinical adoption, the unique aspects of each model and their value need to be clear for end-users. To date, the ongoing development of WSI search algorithms has focused on achieving stepwise improvements in performance on tissue and subtyping search. While these are worthy goals, we propose the inclusion of new tasks based on existing challenges in the field of pathology, such as helping to determine the tissue of cancerous origin in the metastatic setting.

6. **Computational Efficiency:** Clinical applicability not only requires theoretical efficiency, but also hinges on the performance of these systems in a real-world, high-demand environment. Thus, discussions on computational efficiency should reflect the realities of clinical implementation. Especially, we need to make sure the processes of querying an expanding the database are really efficient. These two processes are of utmost importance for a sustainable search engine. We recommend the database indexing algorithms to be online (i.e. adding a new data point does not require compiling the whole database from scratch). Also, one shortcoming of models like RetCCL, SISH, and Yottixel is that they generate mosaics as a percentage of number of patches in a cluster. Now, if like GBM UPENN, slides become very large, number of patches becomes unnecessarily large and query time becomes very long. We recommend models such as HSHR at use a fixed size for patches for this matter.

As the field of digital pathology continues to evolve, we anticipate exciting developments in the near future. These will likely include more efficient and reliable systems for indexing and searching of histopathology slides, increasingly robust algorithms for feature extraction, and potentially transformative diagnostic tools. Given the high-stakes nature of patient care, we anticipate a significant amount of work ahead to ensure the validity of these models prior to clinical adoption. As we continue to make strides in the development of histopathology slide search engines, our proposed criteria ensures that these systems are not only theoretically sound but also ready for meaningful clinical adoption.

Code and Data Availability

The test slides along with the updated source codes for all three methods used to generate the results can be found at github.com/jacobluber/PathologySearchComparison. All other data used in databases are publicly available at portal.gdc.cancer.gov and www.cancerimagingarchive.net/. The list of data included in the database can also be found in the github repository above.

Funding

This work was supported by a University of Texas System Rising STARs Award (J.M.L) and the CPRIT First Time Faculty Award (J.M.L).

Competing Interests

No competing interests are disclosed by the authors.

Acknowledgements

We thank the following for their contributions: David Fernandez-Hazoury, M.D., Department of Pathology, Harbor-UCLA Medical Center; Kenechukwu Ojukwu, M.D., M.P.P.,Bone/Soft Tissue Fellow, Department of Pathology and Laboratory Medicine, National Clinician Scholars Program (NCSP) Fellow; Alex Oliveira-Kowaleski, M.D., UCLA Department of Pathology and Laboratory Medicine; Julie Y. Kim, D.O., M.S., Harbor-UCLA Medical Center; Daniel P. Stefanko, M.D. Ph.D., UCLA Department of Pathology and Laboratory Medicine.

References

Chen, Chengkuan, et al. “Fast and scalable search of whole-slide images via self-supervised deep learning”. *Nature Biomedical Engineering*, volume 6, number 12, Dec. 2022, Number: 12 Publisher: Nature Publishing Group, Pages 1420–34. https://doi.org/10.1038/s41551-022-00929-8.

Chen, Ting, et al. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709, June 30, 2020. *arXiv*, arxiv.org/abs/2002.05709[cs,stat], https://doi.org/10.48550/arXiv.2002.05709.

Clark, Kenneth, et al. “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository”. *Journal of Digital Imaging*, volume 26, number 6, Dec. 1, 2013, Pages 1045–57. https://doi.org/10.1007/s10278-013-9622-7.

Farahmand, Saman, et al. “Deep learning trained on hematoxylin and eosin tumor region of Interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer”. *Modern Pathology*, volume 35, number 1, 2022, Pages 44–51.

Gutman, David A., et al. “Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data”. *Journal of the American Medical Informatics Association: JAMIA*, volume 20, number 6, 2013, Pages 1091–98. https://doi.org/10.1136/amiajnl-2012-001469.

Hamming, Richard W. “Error detecting and error correcting codes”. *The Bell system technical journal*, volume 29, number 2, 1950, Pages 147–60.

He, Kaiming, et al. Deep Residual Learning for Image Recognition. 2015. *arXiv*, arxiv.org/abs/1512.03385.

———. Deep Residual Learning for Image Recognition. arXiv:1512.03385, Dec. 10, 2015. *arXiv*, arxiv.org/abs/1512.03385[cs], https://doi.org/10.48550/arXiv.1512.03385.

He, Kaiming, et al. Momentum Contrast for Unsupervised Visual Representation Learning. arXiv:1911.05722, Mar. 23, 2020. *arXiv*, arxiv.org/abs/1911.05722[cs], https://doi.org/10.48550/arXiv.1911.05722.

Hegde, Narayan, et al. “Similar image search for histopathology: SMILY”. *npj Digital Medicine*, volume 2, number 1, June 2019, Number: 1 Publisher: Nature Publishing Group, Pages 1–9. https://doi.org/10.1038/s41746-019-0131-z.

Kalra, Shivam, et al. “Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence”. *NPJ Digital Medicine*, volume 3, Mar. 2020, Page 31. https://doi.org/10.1038/s41746-020-0238-2.

Kalra, Shivam, et al. “Yottixel - An Image Search Engine for Large Archives of Histopathology Whole Slide Images”. *Medical Image Analysis*, volume 65, Oct. 2020, Page 101757. https://doi.org/10.1016/j.media.2020.101757.

Lew, Michael S., et al. “Content-based multimedia information retrieval: State of the art and challenges”. *ACM Transactions on Multimedia Computing, Communications, and Applications*, volume 2, number 1, Feb. 2006, Pages 1–19. https://doi.org/10.1145/1126004.1126005.

Li, Shengrui, et al. “High-Order Correlation-Guided Slide-Level Histology Retrieval With Self-Supervised Hashing”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 45, number 9, Sept. 2023, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, Pages 11008–23. https://doi.org/10.1109/TPAMI.2023.3269810.

Li, Zhongyu, et al. “Large-scale retrieval for medical image analytics: A comprehensive review”. *Medical Image Analysis*, volume 43, Jan. 2018, Pages 66–84. https://doi.org/10.1016/j.media.2017.09.007.

Oord, Aaron van den, et al. Neural Discrete Representation Learning. May 2018, arXiv:1711.00937 [cs]. https://doi.org/10.48550/arXiv.1711.00937.

Oord, Aaron van den, et al. Representation Learning with Contrastive Predictive Coding. Jan. 2019, arXiv:1807.03748 [cs, stat]. https://doi.org/10.48550/arXiv.1807.03748.

Riasatian, Abtin, et al. Fine-Tuning and Training of DenseNet for Histopathology Image Representation Using TCGA Diagnostic Slides. Jan. 2021, arXiv:2101.07903 [eess]. https://doi.org/10.48550/arXiv.2101.07903.

Simonyan, Karen, and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015. *arXiv*, arxiv.org/abs/1409.1556.

Tizhoosh, H. R. “Barcode annotations for medical image retrieval: A preliminary investigation”. *2015 IEEE International Conference on Image Processing (ICIP)*. Sept. 2015, Pages 818–22, https://doi.org/10.1109/ICIP.2015.7350913.

Tizhoosh, H. R., et al. MinMax Radon Barcodes for Medical Image Retrieval. Oct. 2016, arXiv:1610.00318 [cs]. https://doi.org/10.48550/arXiv.1610.00318.

Van Emde Boas, P. “Preserving order in a forest in less than logarithmic time and linear space”. *Information Processing Letters*, volume 6, number 3, June 1977, Pages 80–82. https://doi.org/10.1016/0020-0190(77)90031-X.

Vorontsov, Eugene, et al. Virchow: A Million-Slide Digital Pathology Foundation Model. 2023. *arXiv*, arxiv.org/abs/2309.07778.

Wang, Xiyue, et al. “RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval”. *Medical Image Analysis*, volume 83, Jan. 2023, Page 102645. https://doi.org/10.1016/j.media.2022.102645.

Weinstein, John N., et al. “The Cancer Genome Atlas Pan-Cancer analysis project”. *Nature Genetics*, volume 45, number 10, Oct. 2013, Number: 10 Publisher: Nature Publishing Group, Pages 1113–20. https://doi.org/10.1038/ng.2764.

Supplementary Material

Datasets

Supplementary Table 1: Summary of database used for comparison experiments. Abbreviations are based on (Kalra et al.).

Primary Site	Project Name (Subtype)	Abbr.	Num. Slides	Num. Selected Slides
Brain	Glioblastoma Multiforme	GBM	2040	61
	Brain Lower Grade Glioma	LGG	1543	69
	Lymphoid Neasm Diffuse Large B-cell Lymphoma	DLBC	4	0
Breast	Breast Invasive Carcinoma	BRCA	2704	72
	Lymphoid Neasm Diffuse Large B-cell Lymphoma	DLBC	2	0
Bronchus and lung	Lung Adenocarcinoma	LUAD	1359	68
	Lung Squamous Cell Carcinoma	LUSC	1265	57
	Mesothelioma	MESO	2	0
Colon	Colon Adenocarcinoma/Rectum Adenocarcinoma[a]	COAD/READ	1307+18	59 (COAD)
	Lymphoid Neasm Diffuse Large B-cell Lymphoma	DLBC	6	0
	Sarcoma	SARC	4	0
Liver and intrahepatic bile ducts	Liver Hepatocellular Carcinoma	LIHC	778	72
	Cholangiocarcinoma	CHOL	80	50

[a] Although from pathologist point of view Colon Adenocarcinoma and Rectum Adenocarcinoma are genetically and morphologically the same entity, TCGA considers them different projects.

Supplementary Table 2: Access links to the test datasets used in experiments.

Dataset	link
CMB-CRC	wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=93257955
CMB-LCA	wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=93258420
Yale Her2	wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=119702524
Yale Trastuzumab	wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=119702524
CPTAC-GBM	wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=30671232
UPENN-GBM	wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70225642

H&E Staining and Preparation

Tissues were stained with Harris' hematoxylin solution for 6 h at a temperature of 60 °C–70 °C and were then rinsed in tap water until the water was colorless. Next, 10% acetic acid and 85% ethanol in water were used to differentiate the tissue 2 times for 2 h and 10 h, and the tissues were rinsed with tap water. In the bluing step, we soaked the tissue in saturated lithium carbonate solution for 12 h and then rinsed it with tap water. Finally, staining was performed with eosin Y ethanol solution for 48 h. Tissues were dehydrated with 95% ethanol twice for 0.5 h, and then soaked in xylene for 1 h at 60 °C–70 °C followed by paraffin for 12 h.

Supplementary Table 3: Unprocessed slides in the database for each model.

Model	**Slide IDs**
Yottixel	9ecf91d4-0d9e-4400-bf38-99420acd14cc f18b6fc0-6f40-4f0d-82ef-0b092a21b6bf 846087b8-f70c-4970-a1b7-24d403229801 c95681f3-53d4-4b15-833d-ff68f171965e 71dc7ba0-a623-4aaf-9502-f2fe9d188401 2dc5d0b4-04ff-4731-bda1-8ad7cd0fa345
SISH	2dc5d0b4-04ff-4731-bda1-8ad7cd0fa345
RetCCL	All slides in the database processed.
HSHR	f2d5aa37-d9ce-4264-a447-fc69dd0d7d85 a2658e39-e476-44b2-99ee-118056cf6201 f84130fe-4853-4252-a292-9372aeea4a5d 22904f9d-0788-463c-9961-02629cf9a85f

The stained tissues were cut into 7 µm slices, dewaxed, mounted with neutral balsam and then imaged using Nikon NIS-Elements microscopy.

Search Engines Methods

In the **Yottixel** method, the initial preprocessing step involves segmenting the foreground from the background in large whole slide images (WSIs). The segmented foreground is then divided into patches of size 1000×1000 for 20× slides and 2000×2000 for 40× slides. The 2000×2000 patches are resized to 1000×1000 before being input to the feature extractor. These patches undergo clustering using the K-means algorithm, resulting in 9 clusters based on the RGB histogram of each patch. A further selection process is applied, retaining 15% of the patches in each cluster using another K-means clustering method based on the spatial coordinates. This final collection of patches forms a "mosaic." The Yottixel model, as recommended by its authors (Kalra et al.), employs KimiaNet (Riasatian et al.), a fine-tuned version of DenseNet specifically designed for histopathology slides, as the primary feature extractor (Fig. 1a). The outputs of the feature extractor undergo barcoding, where binary codes are generated from the extracted features. Thus, each WSI is represented by a set of barcodes (BoBs). The database comprises BoBs for each slide in the dataset. The distance between two BoBs is calculated as the median of the minimum Hamming distances (Hamming) between each barcode in the first BoB and all barcodes in the second BoB. When a query slide is introduced, it is converted into a BoB. The distance between the query BoB and all BoBs in the database is computed, and the top 5 slides with the lowest distances are returned. For patch retrieval, the query BoB is not required, and instead, the top 5 patches from all BoBs with the minimum distances to the query patch are retrieved.

The **SISH** method uses a similar approach to Yottixel for mosaic generation, with patch sizes of 1024×1024 for 20× slides and 2048×2048 for 40× slides. After mosaic generation, artifacts such as pure white patches are filtered out. The feature extraction in SISH consists of two parts: feature and index (Fig. 1b). The feature extraction process is the same as Yottixel, where each patch in the mosaic is fed into a pretrained DenseNet, and the resulting features are binarized. The index, however, is obtained from a pretrained VQ-VAE. The patch is encoded, resulting in a latent code, which is then subjected to three layers of average pooling. The output of these layers is multiplied by scaling factors, and the sum of these results represents the index in the VEB tree. This creates the database. When querying a slide, it is converted into a mosaic, and indices and features are generated from the patches in the query mosaic. The "guided VEB search" algorithm is utilized, leveraging the properties of VEB trees, forward and backward searches, and entropy-based uncertainty calculations to retrieve the top slides based on hamming distance. The ranking algorithm accounts for class imbalance when returning the results. For patch retrieval, an index and feature are created for the query patch using a similar approach, and the best matches are found among the patches in the mosaics of the database. They also use a hamming distance threshold of 128 to make sure they only keep high quality results. That is why sometimes they return only a few matches.

RetCCL, drawing inspiration from both Yottixel and SISH, adopts a distinct approach. Instead of clustering patches based on RGB histogram values, RetCCL first obtains contrastive-based feature vectors for each patch within the segmented foreground tiles. These features serve as inputs for a 9-class K-means clustering. Within each cluster, an additional K-means process based on spatial coordinates is performed to select 20% of the patches. These selected patches form the mosaics, which constitute the database. The proposed feature extraction algorithm in RetCCL utilizes a clustering-guided contrastive learning method, employing the InfoNCE loss introduced in (Oord et al.) (Fig. 1c). Given the prevalence of normal cells in WSIs, learning irregularities from a limited number of patches becomes crucial. The self-supervised feature extractor employs two InfoNCE losses to capture irregular regions in patches. It worth mentioning that RetCCL was not able to perform the indexing on UPENN GBM dataset within a reasonable run time and was not utilized for certain experiments involving this dataset.

For retrieving similar slides, a query slide is first transformed into a mosaic, generating a set of features for each patch in the mosaic. Similarity between two patches is measured using cosine similarity between their feature vectors. The retrieval process involves returning a set of patches in the database that exhibit a similarity score of at least 70% to the query patch. Each query patch and its corresponding results form a "bag." To account for class imbalance, an entropy-based uncertainty measure is calculated based on the occurrence of each label within the bag. Patch members in the bag are sorted according to this entropy measure. A threshold is then determined to remove lower quality results. Ultimately, the top 5 samples within each bag are returned as the final results for slide retrieval. For patch retrieval, only the top 5 patches with the highest cosine similarity scores are returned.

In **HSHR**, the first step is to train the ResNet18 (He et al., Deep Residual Learning for Image Recognition) backbone encoder using the SimCLR (T. Chen et al.) approach. Unlike the other methods, authors had not provided their backbone pre-trained weights, so we trained it from the scratch. The training data for this backbone was approximately $508 \times 100 = 50,800$ randomly patches of size 224×224. We trained it for 200 epochs using the same hyperparameters as the authors on 2 Nvidia A100 GPUs. Once the backbone is trained, it is used to extract the features for all densely-patched patches for each WSI. These features are used to train a 20-class k-means clustering algorithm. The features of the centroids of these clusters create the mosaic for each WSI. These features are then passed to CaEncoder for generating the hashes and attention weights. Unlike the backbone, the authors had provided the weights for their CaEncoder, and we used the same weights in our experiments. The outputs of CaEncoder is further used to create a hypergraph for the database using Eq. 15 in (S. Li et al.). We used $K = 10$ in this equation.

Once the hypergraph for database is constructed, each query slide would go through the same pipeline and becomes and turns into hash codes and attention weights. Using these values, the query can be appended to the hypergraph as a new vertex and a new hyperedge. Updating the hypergraph, the similarity score between this vertex and all vertices in the database can be calculate using Eq. 19 in (S. Li et al.). Then the top-k results are returned. We used $\alpha = \beta = 1$ in this equation.

Methodological Breakdown of Query and Retrieval Processes

Algorithm 1: Yottixel Algorithm

```
Input: Image $I$
Output: Top 5 retrieved slides
Patch the image densely to get patches $p_1, p_2, \ldots$;
Perform RGB histogram clustering;
Perform spatial clustering and calculate mosaic patches;
Feed mosaic patches to KimiaNet for feature extraction and calculate barcode for each patch;
foreach patch in input slide do
    Calculate hamming distance between barcode of input patch with barcodes of patches from all slides in database;
Choose the median of the list of minimum hamming distances for each slide in database;
Retrieve the slides with the top five smallest medians;
```

Algorithm 2: SISH Algorithm

```
Input: Image $I$
Output: Similar slides to query slide $I$
Patch the image densely to get patches $p_1, p_2, \ldots$;
Perform RGB histogram clustering;
Perform spatial clustering and create mosaic patches;
Feed mosaic patches to DenseNet and VQ-VAE encoder to calculate parameters $h$ and $m$
  respectively, and create the VEB tree;
Apply guided-search algorithm to tuples of $m$ and $h$ to calculate corresponding set of tuples $r$;
Create a set of candidate indices $mi, c+$ and $mi, c-$ along with the original $mi$;
Call helper functions forward-search and backward-search on $mi, c+$ and $mi, c-$ respectively;
Take the results $RI = \{r_1, r_2, \ldots, r_k\}$ from Guided-Search as input by Results Ranking Algorithm;
Return similar slides to query slide $I$;
```

Algorithm 3: RetCCL Algorithm

```
Input: Image $I$
Output: Top $k$ similar WSIs
Patch the image densely to get patches $p_1, p_2, \ldots$;
Feed patches to feature extraction algorithm;
Cluster based on extracted features, then on coordinates to create mosaic patches;
foreach patch in query WSI do
    Perform Knn search to retrieve a bag of most similar patches in database to each patch, using
      cosine similarity in pretrained SSL encoder's learned embedding space;
Calculate entropy within each bag, reorder bags by entropy;
Remove bags with low quality based on mean of cosine similarity scores in top-5;
foreach bag do
    Perform voting for each diagnosis within the bag, get the top-5 samples, then do majority vote to
      get associated WSI;
Retrieve top-$k$ similar WSIs;
```

Algorithm 4: HSHR Algorithm

```
Input: Image $I$
Output: Top $k$ similar WSIs
Patch the image densely to get patches $p_1, p_2, \ldots$;
Feed patches to feature extraction algorithm;
Cluster based on extracted features, then on coordinates to create mosaic patches;
foreach patch in query WSI do
    Create a bag containing the query patch and its retrieved patches;
Calculate entropy within each bag, reorder bags by entropy;
Remove bags with low quality based on mean of cosine similarity scores in top-5;
foreach bag do
    Perform voting for each diagnosis within the bag, get the top-5 samples, then do majority vote to
      get associated WSI;
Retrieve top-$k$ similar WSIs;
```

Complexity Analysis

The time complexities of the key operations in different algorithms are summarized in Table Supplementary Table 4. In the table, n (Yottixel) denotes the length of the hamming vector, m (Yottixel) denotes the maximum number of patches for each WSI, T (common across Yottixel, HSHR, and RetCCL) denotes the total number of slides in the database, B (SISH) denotes the number of patches in a WSI, K (RetCCL, HSHR) denotes the number of patches of the query slide, M (RetCCL, HSHR) denotes the total number of diagnoses in the database, and H (HSHR) denotes the hash vector size.

Supplementary Table 4: Time complexity analysis for different components of the compared methods.

Algorithm	Operation and Time Complexity
Yottixel	Hamming distance calculation: $O(n.T.m^2)$ Minimum: $O(T.m^2)$ Median: $O(T.\log(T))$
SISH	Search Performance: $O(1)$ Ranking: $O(B')$, $B' = 0.05 \cdot B$
RetCCL	Cosine similarity calculation: $O(K.B)$ Probability calculation: $O(M.K.B)$ Entropy calculation: $O(K.M)$ Sorting bags based on entropies: $O(K^2)$ Mean of cosine similarity scores: $O(T.K)$ Removing bags with low quality: $O(K)$
HSHR	Hamming distance: $O(H.T.M^2)$ Sorting: $O(M)$ Incidence Matrix calculation: $O(K.M.T)$ Cross similarity: $O(T^2)$ Vertex similarity: $O(T^3)$ Hyperedge similarity: $O(T^3)$ Final sorting: $O(T)$

Studies

Performance Metrics

As discussed, we use majority voting and average precision at k as main performance metrics in different experiments. Supplementary Algorithms line 5 and line 6 summarize they way we defined these metrics. The important point about majority voting at k is that if it returns `None`, that sample would not be counted in the average, while 0 outputs are counted towards average.

Algorithm 5: Majority Voting at k

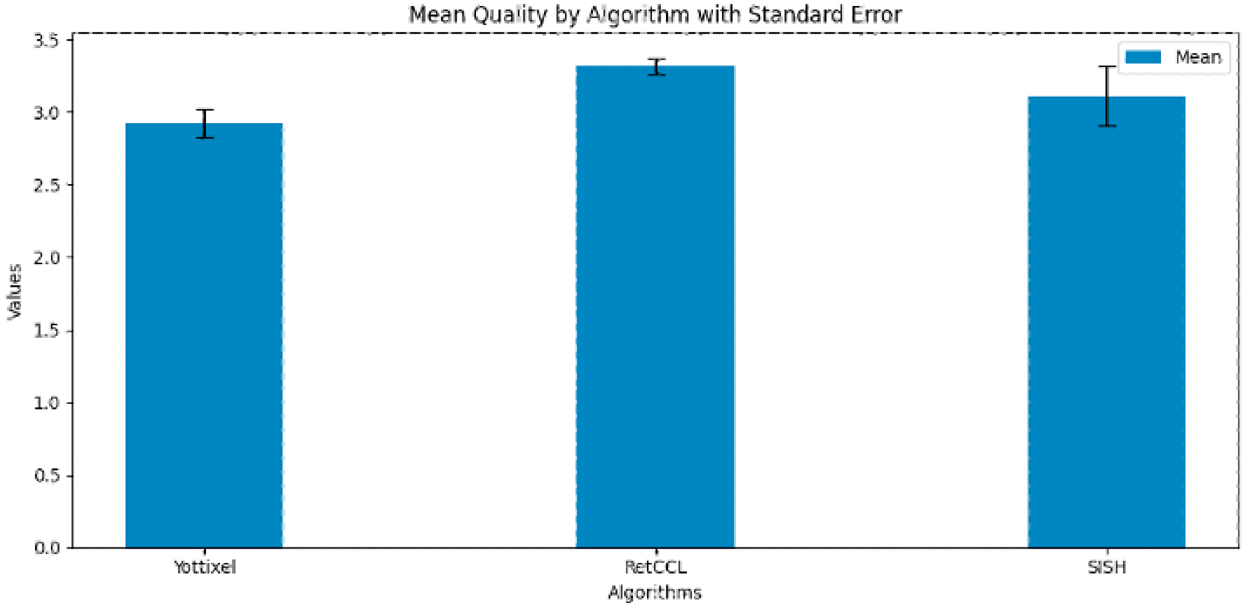
```
   Data: $row$, $k$
   Result: Result of the majority vote
1  $votes \leftarrow$ empty list
2  for $i = 1$ to $k$ do
3      $ret\_site \leftarrow row$[f'ret_i_site']
4      if $ret\_site$ is null then
5          append $-1$ to $votes$
6      else
7          append $ret\_site$ to $votes$
8  if not $votes$ then
9      return 0
10 $counter \leftarrow$ Counter($votes$)
11 $most\_common \leftarrow counter$.most_common(1)
12 if $most\_common[0][0] = row['query\_site']$ then
13     return 1
14 else
15     if $most\_common[0][0] = -1$ then
16         return None
17     else
18         return 0
```

Algorithm 6: Average Precision at k

```
Data: $row$, $k$
Result: Average precision at k
$relevant\_count \leftarrow 0$
$precision\_sum \leftarrow 0$
for $i = 1$ to $k$ do
    $ret\_site \leftarrow row$[f'ret_i_site']
    if $ret\_site$ is $None$ then
        continue
    if $ret\_site = row['query\_site']$ then
        $relevant\_count \leftarrow relevant\_count + 1$
        $precision\_sum \leftarrow precision\_sum + \frac{relevant\_count}{i}$
if $relevant\_count = 0$ then
    return 0
return $\frac{precision\_sum}{\min(k, relevant\_count)}$
```

Reader study

Seven pathologists were shown the top three ranked results on patch-level retrieval for one patch across three different H&E slides. All pathologists were shown the original queried patch of interest but were blinded to the algorithms involved, ranked order of the retrieved patches, and diagnoses of both the queried and retrieved patched. Pathologists were then asked for their Mean Opinion Score (MOS) based on their perspective on the quality of the results, ranked from one to five with higher scores indicating higher quality.



Supplementary Figure 1: Reader Study: Average quality ratings and standard deviation for each algorithm as evaluated by seven pathologists.

Extended Results

We choose three internal patient cases for visual review of tissue, subtype, and patch-level search results. Slide 1 is a Lung Adenocarcinoma (LUAD) case from a patient's partial lobectomy. Slide 2 is a Low Grade Glioma (LGG) that was retrieved by surgical biopsy. Slide 3 is from a patient with Hepatocellular Carcinoma (LIHC) who underwent a liver biopsy by fine needle aspiration.

Supplementary Table 5: Evaluation results of different methods on Reader Study slides for primary site retrieval task.

Method	**Reader Study**	**MV@1**	**MV@3**	**MV@5**	**MV@10**	**AP@3**	**AP@5**
YOTTIXEL + KimiaNet	**MSB-09151-01-11**	0	0	0	0	0	0.250
	MSB-09977-01-22	1	1	1	0	1	1
	Her2Pos_Case_66	1	1	1	1	0.833	0.806
SISH + DenseNet	**MSB-09151-01-11**	0	0	0	0	0	0
	MSB-09977-01-22	0	0	0	0	0	0
	Her2Pos_Case_66	0	0	0	0	0	0
RetCCL	**MSB-09151-01-11**	0	0	0	1	0	0.250
	MSB-09977-01-22	1	1	0	0	0	1
	Her2Pos_Case_66	0	0	0	1	0	0.325
HSHR	**MSB-09151-01-11**	1	1	1	1	0.833	0.833
	MSB-09977-01-22	1	1	1	1	1	0.700
	Her2Pos_Case_66	1	1	1	0	1	0.867

Supplementary Table 6: Evaluation results of different methods on Reader Study slides for subtype retrieval task.

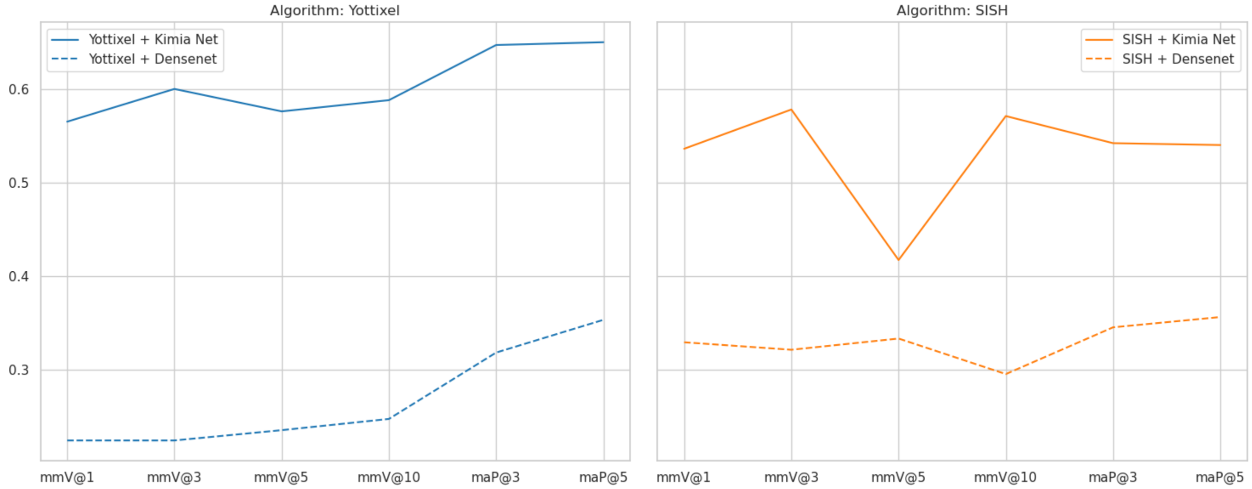
Method	**Reader Study**	**MV@1**	**MV@3**	**MV@5**	**AP@3**	**AP@5**
YOTTIXEL + KimiaNet	**MSB-09151-01-11**	-	-	-	-	-
	MSB-09977-01-22	0	0	0	0	0
	Her2Pos_Case_66	-	-	-	-	-
SISH + DenseNet	**MSB-09151-01-11**	-	-	-	-	-
	MSB-09977-01-22	0	0	0	0	0
	Her2Pos_Case_66	-	-	-	-	-
RetCCL	**MSB-09151-01-11**	-	-	-	-	-
	MSB-09977-01-22	0	1	1	0.583	0.589
	Her2Pos_Case_66	-	-	-	-	-
HSHR	**MSB-09151-01-11**	-	-	-	-	-
	MSB-09977-01-22	1	1	1	0.833	0.806
	Her2Pos_Case_66	-	-	-	-	-

Supplementary Table 7: Mann-Whitney U Test Results for Microscope

Microscope	**Metric**	**U-statistic**	**P-value**
SISH	MV_at_10_site	nan	nan
	AP_at_5_site	578.000	1.0000000000000000
	MV_at_5_subtype	nan	nan
	AP_at_5_subtype	543.000	0.6156890923465373
HSHR	MV_at_10_site	658.500	0.0670269674521155
	AP_at_5_site	523.500	0.6135535212442216
	MV_at_5_subtype	672.500	0.0905435676918205
	AP_at_5_subtype	636.000	0.3353506099579945
Yottixel	MV_at_10_site	523.500	0.5298120557113710
	AP_at_5_site	455.500	0.1566946582170631
	MV_at_5_subtype	532.000	0.6415761864226559
	AP_at_5_subtype	534.500	0.7372510705087723

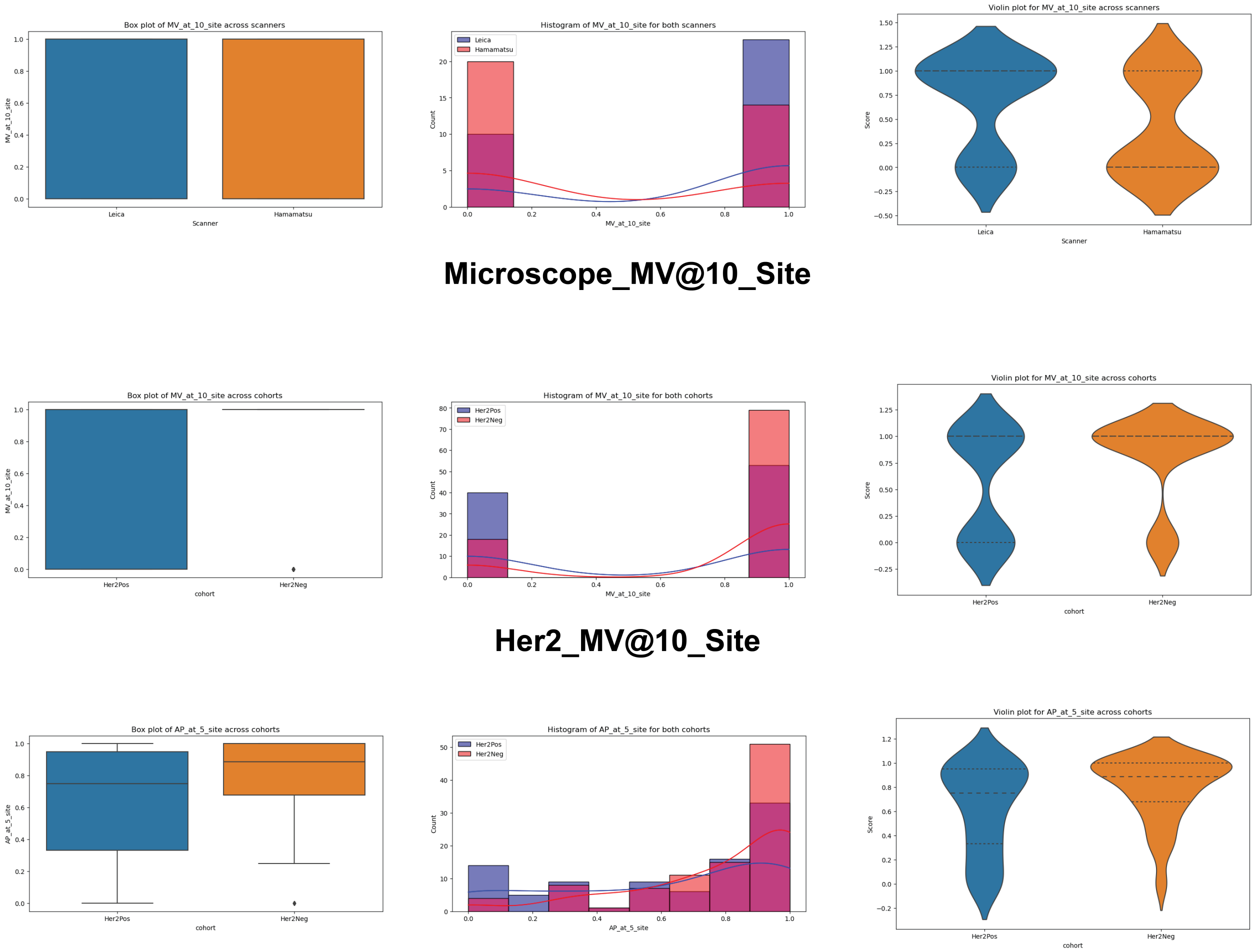
Supplementary Table 8: Mann-Whitney U For Her2+

Model	Metric	U-statistic	P-value
Yottixel	MV_at_10_site	2906.500	0.0000006008842402
	AP_at_5_site	3287.500	0.0008328526912513
SISH	MV_at_10_site	nan	nan
	AP_at_5_site	3651.500	0.0130448102685896
RetCCL	MV_at_10_site	nan	nan
	AP_at_5_site	2483.500	0.0000000486711422
HSHR	MV_at_10_site	3407.500	0.0002646668324846
	AP_at_5_site	3274.500	0.0008856874714843



Supplementary Figure 2: Performance comparison of Yottixel and SISH algorithms using Kima Net and Densenet on the Yale Trastuzumab dataset across various metrics.

HSHR



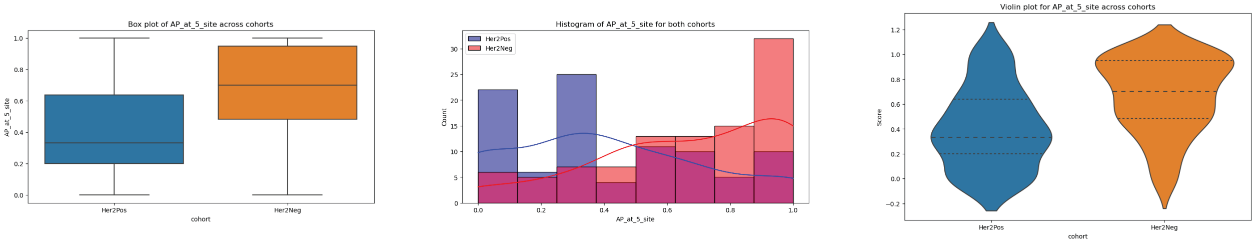
Microscope_MV@10_Site

Her2_MV@10_Site

Her2_AP@5_Site

Supplementary Figure 3: Metrics for HSHR Microscope comparison analysis and Her2 analysis.

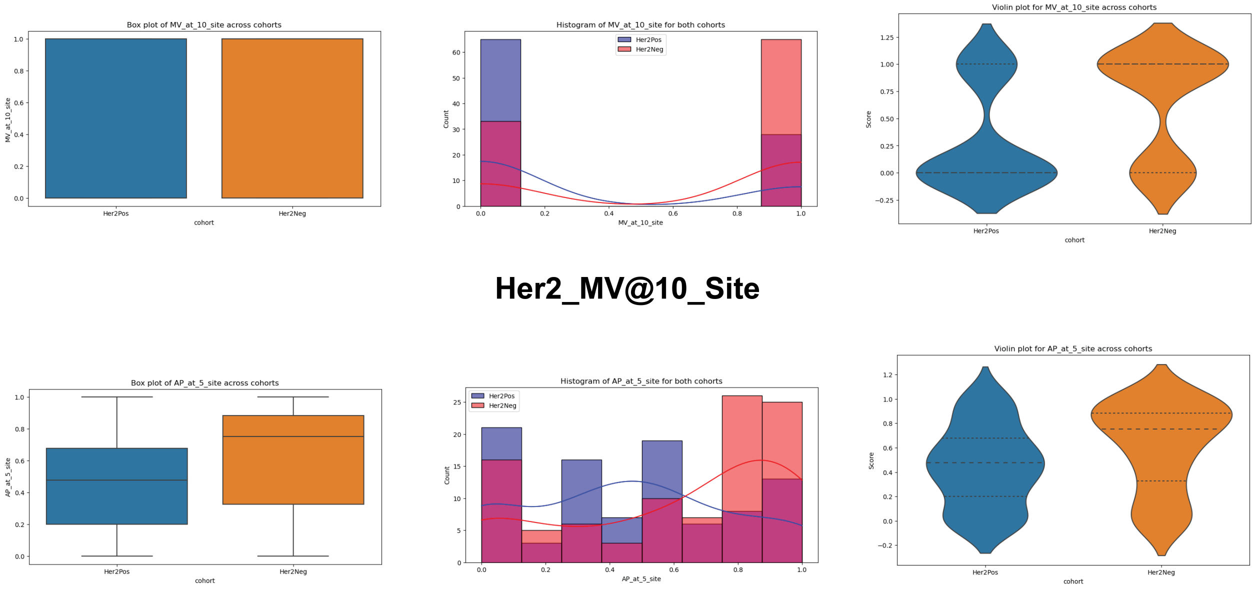
RetCCL



Her2_AP@5_Site

Supplementary Figure 4: Metrics for RetCCL Her2 analysis.

YOTTIXEL

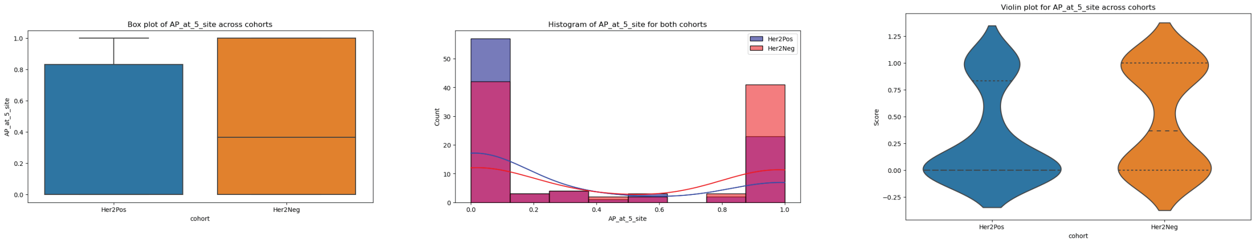


Her2_MV@10_Site

Her2_AP@5_Site

Supplementary Figure 5: Metrics for YOTTIXEL Her2 analysis.

SISH



Her2_AP@5_Site

Supplementary Figure 6: Metrics for SISH Her2 analysis.

Supplementary Table 9: Evaluation results of different methods on large datasets for primary site retrieval task. *mMV* means mean Majority Voting score and mAP means mean Average Precision.

Method	Dataset	mMV@1	mMV@3	mMV@5	mMV@10	mAP@3	mAP@5
YOTTIXEL + KimiaNet	**UPENN GBM**	0.706	0.735	0.735	0.794	0.752	0.765
	CPTAC GBM	0.606	0.606	0.667	0.727	0.646	0.644
	Yale Her2 Pos	0.548	0.559	0.538	0.570	0.627	0.624
	Yale Her2 Neg	0.742	0.742	0.784	0.814	0.809	0.788
	Yale Trastuzumab	0.565	0.600	0.576	0.588	0.647	0.650
SISH + DenseNet	**UPENN GBM**	0.735	0.706	0.706	0.758	0.730	0.720
	CPTAC GBM	0.706	0.697	0.697	0.710	0.725	0.729
	Yale Her2 Pos	0.269	0.247	0.221	0.167	0.294	0.304
	Yale Her2 Neg	0.454	0.442	0.446	0.375	0.476	0.480
	Yale Trastuzumab	0.329	0.321	0.333	0.295	0.345	0.356
RetCCL	**UPENN GBM**	-	-	-	-	-	-
	CPTAC GBM	0.588	0.594	0.750	0.846	0.728	0.736
	Yale Her2 Pos	0.172	0.301	0.355	0.481	0.345	0.404
	Yale Her2 Neg	0.510	0.602	0.742	0.864	0.656	0.671
	Yale Trastuzumab	0.212	0.247	0.353	0.420	0.382	0.428
HSHR	**UPENN GBM**	0.794	0.824	0.765	0.735	0.816	0.809
	CPTAC GBM	0.758	0.727	0.818	0.909	0.813	0.810
	Yale Her2 Pos	0.247	0.323	0.323	0.301	0.429	0.449
	Yale Her2 Neg	0.571	0.592	0.622	0.663	0.632	0.612
		0.447	0.447	0.494	0.529	0.581	0.569
YOTTIXEL + DenseNet	**Yale Trastuzumab**	0.224	0.224	0.235	0.247	0.318	0.353
SISH + KimiaNet	**Yale Trastuzumab**	0.536	0.578	0.417	0.571	0.542	0.540

Supplementary Table 10: Evaluation results of different methods on large datasets for subtype retrieval task.

Method	Dataset	mMV@1	mMV@3	mMV@5	mAP@3	mAP@5
YOTTIXEL + KimiaNet	**UPENN GBM**	0.382	0.382	0.294	0.468	0.454
	CPTAC GBM	0.303	0.303	0.242	0.399	0.418
SISH + DenseNet	**UPENN GBM**	0.706	0.656	0.677	0.725	0.725
	CPTAC GBM	0.676	0.667	0.656	0.686	0.681
RetCCL	**UPENN GBM**	-	-	-	-	-
	CPTAC GBM	0.529	0.375	0.406	0.667	0.648
HSHR	**UPENN GBM**	0.618	0.559	0.559	0.721	0.704
	CPTAC GBM	0.758	0.727	0.758	0.806	0.785

5

General Conclusions

This dissertation embarked on a journey to enhance the capabilities of artificial intelligence in biomedical imaging, a field that stands at the intersection of medical research, computational science, and technology. The central objective was to address pressing challenges in the analysis and prediction of spatial transcriptomics, cancer pathology, and histopathology slide analysis. The research questions poised at the beginning of this dissertation focused on how advanced AI methodologies could be leveraged to improve the precision, efficiency, and predictive power in these domains.

Each of the included studies adopted a unique methodological approach, tailored to the specific challenges and nuances of the respective research areas. In the first study, we explored the use of Random Forest Regression combined with spatial point processes to predict the future distribution of cells in spatial transcriptomics. This approach represented a fusion of computational modeling with intricate biological data, offering a novel perspective on gene expression analysis at the single-molecule level.

The second paper shifted focus to the realm of cancer histopathology, where a Variational Autoencoder (VAE) was used for the compression of large-scale histopathological images. This study tackled the significant challenge of managing voluminous biomedical image data without compromising the clinical relevancy and integrity of the images.

Finally, in the third study, we examined the current landscape of histopathology slide indexing and search systems. The emphasis was on evaluating and enhancing the efficiency and accuracy of these systems using state-of-the-art AI techniques. This included a comprehensive analysis of existing systems and the development of an improved methodology for slide retrieval.

Collectively, these studies underscore the transformative potential of AI in biomedical imaging, each contributing to a facet of the broader goal of advancing the field. They represent an amalgamation of innovative computational techniques and deep domain knowledge, pushing the boundaries of how AI can be utilized to extract meaningful insights from complex biological data.

The methodologies adopted across these studies not only address the specific challenges within each research area but also contribute to the overarching theme of enhancing AI's role in biomedical imaging. The integration of these diverse yet complementary approaches highlights the multidisciplinary nature of the field and sets the stage for the conclusions and future directions discussed in the subsequent sections of this chapter.

5.1 Summary of Key Findings

5.1.1 Paper 1: Predicting the Future States of Gene Expression

The first study in this dissertation made significant strides in the field of spatial transcriptomics, particularly in predicting cellular behavior and gene expression patterns during the embryogenesis of Drosophila. This research utilized Random Forest Regression in conjunction with spatial point processes, marking a novel approach in the analysis of super-resolution whole embryo spatial transcriptomics imaging.

Key outcomes of this study include:

1. **Development of a Predictive Model**: The implementation of Random Forest Regression, combined with Ripley's K-function, enabled the accurate prediction of the future distribution of cells expressing the Sog-D gene. This approach was pivotal in understanding the dynamic nature of gene expression during the embryogenesis process.

2. **Enhanced Resolution and Predictive Accuracy**: The study achieved a significant breakthrough in analyzing gene expression at a sub-cellular, single-molecule resolution. This high-resolution analysis allowed for a more nuanced understanding of cellular dynamics and gene expression patterns.

3. **Novel Methodological Integration**: By leveraging temporally resolved spatial point processes, the study introduced a methodological innovation that bridged the gap between static and dynamic analyses of gene expression. This integration was instrumental in moving beyond traditional static snapshots of gene expression to a more dynamic and predictive model.

4. **Practical Implications**: The findings of this study have profound implications for understanding the complex mechanisms of gene expression regulation during development. The predictive model provides insights that are crucial for further research in developmental biology, potentially influencing studies in disease progression and therapeutic interventions.

The outcomes from this study not only contribute significantly to the field of spatial transcriptomics but also exemplify the potential of combining computational models with biological data to yield deeper insights into complex biological processes.

This section of the final chapter offers a concise yet comprehensive summary of the critical findings from the first paper, showcasing the innovative approaches and significant contributions to the field of spatial transcriptomics.

5.1.2 Paper 2: Clinically Relevant Histopathology Slide Compression

The second paper in this dissertation focused on addressing the challenge of efficiently managing and analyzing large cancer histopathology slide datasets through advanced image compression techniques. This study employed a Variational Autoencoder (VAE) based approach, offering innovative solutions to the field of digital pathology.

Key outcomes of this study include:

1. **State-of-the-Art Image Compression**: The development and implementation of a VAE model achieved a remarkable compression ratio of 1:512, surpassing previous benchmarks in the field. This high compression ratio was attained while maintaining the accuracy and integrity of the histopathological images, crucial for clinical validations.

2. **Preservation of Clinically Relevant Information**: One of the most significant achievements of this approach was the ability to compress images without losing critical histological features necessary for accurate medical diagnosis and research. This balance between compression efficiency and data integrity marks a significant advancement in medical image processing.

3. **Enhancement of Image Retrieval and Analysis**: The study demonstrated that the compressed images, through their latent space embeddings, retained essential clinical information. This finding is vital for the development of efficient image search algorithms in large whole slide image databases, greatly facilitating the retrieval and analysis process.

4. **Methodological Innovations**: The use of a DenseNet-based architecture within the VAE model, along with methodological improvements for handling large-scale image data, showcased the potential of deep learning techniques in revolutionizing medical image analysis.

5. **Implications for Digital Pathology**: The research presented in this paper has broad implications for the field of digital pathology. By significantly reducing the storage and computational requirements for large datasets without compromising the

quality of analysis, this approach paves the way for more scalable and efficient digital pathology practices.

The findings from this study contribute significantly to the ongoing efforts to integrate AI and machine learning into the realm of medical imaging, particularly in optimizing the storage, retrieval, and analysis of large-scale histopathological datasets.

5.1.3 Paper 3: Readiness of Histopathology Slide Search for Clinic

The third paper in the dissertation delves into the critical aspect of indexing and searching histopathology slides, a key component in the efficient handling and analysis of digital pathology data. This study provided a comprehensive evaluation of the current state of histopathology slide indexing and search systems, assessing their effectiveness and identifying areas for improvement.

Key outcomes of this study include:

1. **Evaluation of Existing Systems**: The paper conducted a thorough analysis of existing histopathology slide search engines, such as Yottixel, SISH, RetCCL, and HSHR. This evaluation provided valuable insights into the strengths and limitations of current methodologies, highlighting the advancements made in the field and the challenges that still persist.

2. **Innovations in Feature Extraction and Retrieval**: A significant focus of the study was on the methods of feature extraction from Whole Slide Images (WSIs). The paper discussed various techniques, including the innovative approach of Yottixel,

which uses a DenseNet-based feature extractor on mosaic tiles, offering a more efficient processing method for large-scale WSIs.

3. **Identification of Challenges and Gaps**: The research identified key challenges in the field, such as the need for higher precision in search results, handling variability in slide preparation and imaging, and integrating search systems into clinical workflows. These challenges are critical in understanding the current limitations and guiding future developments.

4. **Future Directions for Histopathology Search Engines**: The study proposed potential improvements and future directions for histopathology slide indexing and search systems. It emphasized the importance of developing more sophisticated and user-friendly systems that can handle the increasing scale of digital pathology and provide clinically relevant search results.

5. **Implications for Digital Pathology**: The findings of this study have broad implications for digital pathology, particularly in enhancing the diagnostic process and supporting medical research. Efficient and accurate search systems are essential for pathologists and researchers to navigate through large volumes of digital slides, improving both the speed and quality of medical diagnoses and research.

The insights gained from this paper contribute significantly to the field of digital pathology, particularly in the context of AI and machine learning. It highlights the need for ongoing innovation in histopathology slide indexing and search systems, aiming to meet the growing demands of digital pathology and enhance the efficiency and accuracy of medical imaging analysis.

With the completion of this summary of key findings from the third paper, the dissertation concludes its exploration of advancements in AI applications in biomedical imaging, setting the stage for the final conclusions and future research directions in the subsequent sections.

5.2 Discussion on the Integration of Results

This dissertation, through its individual studies, presents a cohesive narrative that showcases the advancement of AI in biomedical imaging. Each paper, while focusing on a unique aspect of this broad field, contributes to a collective understanding and enhancement of the capabilities in analysis and prediction.

5.2.1 Interrelation of Findings Across Studies

The first paper on spatial transcriptomics introduced a novel method to predict cell distribution using Random Forest Regression, advancing our ability to interpret complex gene expression patterns at a single-molecule resolution. This study set the tone for precision and predictive accuracy, essential themes that resonate through the subsequent papers.

The second paper, focusing on the application of a Variational Autoencoder for image compression in cancer histopathology, further underscored the theme of precision but added a layer of efficiency in handling large-scale datasets. By ensuring that critical histological information is preserved even after significant data compression, this study complemented the first by enhancing the manageability of large-scale biomedical data without losing analytical accuracy.

The third paper ventured into the realm of histopathology slide indexing and search systems, addressing the practical challenges of managing and retrieving vast amounts of digital pathology data. This study highlighted the importance of sophisticated AI-driven systems that can navigate through the complexities of histopathology slides, ensuring that the valuable data generated by the methodologies in the previous papers are accessible and usable.

5.2.2 Advancements in Analysis and Prediction in Biomedical Imaging

Collectively, these studies have significantly advanced the analysis and prediction capabilities in biomedical imaging. They have done so by:

1. **Enhancing Predictive Modeling**: The first paper's predictive modeling approach in spatial transcriptomics represents a significant leap in understanding dynamic biological processes, a critical aspect of precision medicine.

2. **Improving Data Management and Efficiency**: The second paper's contribution to efficient data management through innovative image compression techniques has enabled the handling of large-scale histopathology data, a bottleneck in digital pathology.

3. **Optimizing Data Retrieval and Usage**: The third paper's focus on improving slide indexing and search systems ensures that the vast amounts of data generated and managed are effectively utilized, thereby supporting both clinical decision-making and research.

5.2.3 Contributions to the Central Research Questions

In addressing the central research questions posed at the outset of this dissertation, these studies collectively demonstrate how AI can be leveraged to not only analyze complex biomedical data but also predict future states and trends within these data. They illustrate the potential of AI to transform biomedical imaging from a field that is traditionally reactive (focused on diagnosis) to one that is increasingly predictive, aiding in prognosis and personalized medicine.

In conclusion, the integration of results from these individual papers presents a comprehensive picture of how AI can revolutionize various aspects of biomedical imaging. From enhancing the precision of predictive models and improving the efficiency of data handling to optimizing the retrieval and application of vast datasets, these studies collectively push the boundaries of what is possible in the realm of biomedical research and clinical practice.

5.3 Impact of the Research

The research presented in this dissertation has made significant contributions to the fields of biomedical imaging and artificial intelligence, both from a theoretical standpoint and in terms of practical applications. The advancements achieved in these studies have implications that extend far beyond the scope of the individual papers, influencing clinical practice, medical research, and the broader field of AI in healthcare.

5.3.1 Contributions to the Field of Biomedical Imaging and AI

1. **Advancing Predictive Analytics in Biomedical Imaging**: The dissertation's first paper introduces advanced predictive modeling in the realm of spatial transcriptomics, representing a major leap in our ability to forecast cellular behavior and gene expression. This advance not only contributes to the field of biomedical imaging but also exemplifies the application of AI in predicting complex biological processes.

2. **Innovating in Efficient Data Management and Compression**: The implementation of a Variational Autoencoder for image compression, as presented in the second paper, addresses a critical need in digital pathology – managing large-scale histopathological data efficiently. This innovation stands at the intersection of biomedical imaging and AI, demonstrating how deep learning can be applied to solve practical data management challenges in medicine.

3. **Enhancing Accessibility and Usability of Biomedical Data**: The third paper's focus on optimizing histopathology slide indexing and search systems has direct implications for the accessibility and usability of vast amounts of digital pathology data. This contribution is vital for the effective application of AI in biomedical imaging, ensuring that data are not only well-managed but also readily accessible for analysis and decision-making.

5.3.2 Practical Implications for Clinical Practice and Medical Research

1. **Enhancing Diagnostic Accuracy and Efficiency**: The techniques developed in these studies have the potential to significantly enhance the accuracy and efficiency

of disease diagnosis. By providing more precise predictive models and efficient data management systems, clinicians can access and interpret relevant medical imaging data more quickly and accurately, leading to improved patient outcomes.

2. **Facilitating Personalized Medicine**: The ability to predict cellular behavior and analyze large-scale histopathological data more effectively paves the way for personalized medicine. These advances allow for a more nuanced understanding of individual patient conditions, enabling tailored treatment strategies based on specific disease characteristics and patient profiles.

3. **Accelerating Medical Research**: The methodologies and technologies developed in this dissertation will also accelerate medical research by providing more efficient tools for data analysis. Researchers can leverage these AI-driven techniques to uncover new insights into disease mechanisms, treatment responses, and epidemiological trends.

In summary, the research presented in this dissertation has substantial implications for the future of biomedical imaging and AI. It contributes to the advancement of the field by introducing innovative methodologies and technologies that enhance our ability to analyze, predict, and utilize medical imaging data. These contributions not only represent significant academic achievements but also have the potential to transform clinical practice and medical research, ultimately improving patient care and health outcomes.

5.4 Reflection on Methodological Approaches

The methodologies employed in this dissertation represent a significant advancement in the integration of AI with biomedical imaging. Each study utilized distinct approaches tailored to specific challenges, offering a broad perspective on the potential and limitations of current AI technologies in this field.

5.4.1 Critical Analysis of the Methodologies

1. **Spatial Transcriptomics (Paper 1)**: The use of Random Forest Regression combined with spatial point processes was a novel approach in predicting cell distribution in spatial transcriptomics. This methodology allowed for high-resolution analysis and predictive modeling, which is critical in understanding dynamic gene expression patterns. However, the complexity of the model and the specificity of the data might limit its applicability to different types of transcriptomic data or other biological processes.

2. **Image Compression in Histopathology (Paper 2)**: The Variational Autoencoder approach for compressing cancer histopathology slides marked a significant advancement in digital pathology. Its ability to maintain image integrity at high compression rates is a key strength. Nevertheless, the reliance on deep learning models necessitates substantial computational resources, and the model's performance may vary with different types of histopathological data.

3. **Histopathology Slide Indexing and Search (Paper 3)**: The exploration of current histopathology slide indexing and search systems highlighted the need for more so-

phisticated AI-driven methods. While the study provided valuable insights into the current state of these systems, it was more of an evaluative approach rather than a development of new technology. The findings point towards the potential for future innovations in this area.

5.4.2 Strengths, Limitations, and Generalizability

1. **Strengths**: Across all studies, the methodologies demonstrated the power of AI to provide deeper insights and more efficient solutions in biomedical imaging. They showcased the potential for AI to transform data analysis from a largely manual and time-consuming process to an automated, efficient, and more accurate one.

2. **Limitations**: A common limitation across these methodologies is the need for large datasets and computational resources, which can be a barrier to widespread adoption. Additionally, the specificity of some methods to certain data types or conditions can limit their applicability in broader contexts.

3. **Generalizability**: While the methods employed in each study showed promising results in their specific applications, the generalizability of these approaches to other areas of biomedical imaging or different diseases remains to be thoroughly tested. Future research should focus on adapting and testing these methodologies in varied contexts to fully realize their potential.

In conclusion, the methodological approaches used in this dissertation represent a significant stride in the application of AI in biomedical imaging. They offer a balance between innovation and practicality, providing solutions to some of the field's most pressing challenges while also highlighting areas for future improvement. The insights gained from these approaches form a solid foundation for further research and development in this rapidly evolving field.

5.5 Future Research Directions

The findings from this dissertation open several avenues for future research in the field of AI and biomedical imaging. These directions not only extend the work presented but also propose methodological enhancements and explore new areas of investigation.

5.5.1 Paper 1: Enhancing Spatial Transcriptomics Predictions

Future research stemming from the first paper could focus on integrating image and vision deep learning methods to predict the actual active cells, rather than just their distribution. This advancement would hinge on the availability of larger datasets that provide more comprehensive spatial and temporal information. By employing advanced image processing and pattern recognition techniques, researchers could achieve a more granular and accurate prediction of cell behavior, enhancing our understanding of complex biological processes.

5.5.2 Paper 2: Advancing Autoencoder Models for Image Compression

For the second paper, an exciting area of future research involves training an autoencoder on the latent variable while conditioning it on clinical features. This approach would allow the model to incorporate relevant clinical information, potentially leading to more clinically pertinent image compression. Additionally, improving the model to a Vector Quantized-Variational AutoEncoder (VQ-VAE) could offer better control over the latent space and enhance the quality of the reconstructed images. These improvements would make the model more robust and applicable in diverse clinical settings.

5.5.3 Paper 3: Foundation Models for Search in Histopathology

In relation to the third paper, future work should include testing foundation models for search and retrieval in histopathology. This exploration should aim to identify additional requirements for effective search mechanisms in digital pathology. By extending the evaluation criteria and incorporating more sophisticated AI models, researchers can develop more advanced search systems that address the nuanced needs of pathologists and researchers. This would significantly contribute to making digital pathology more efficient and effective.

5.5.4 General Suggestions for Methodological Improvements

Across all areas, there is a continuous need for methodological improvements. Enhancing computational efficiency, reducing the reliance on large datasets, and improving the generalizability of models are crucial goals. Furthermore, exploring new applications of AI in areas such as real-time diagnostics, prognostic modeling, and personalized treatment planning could significantly impact patient care. The integration of AI with emerging technologies like augmented reality and robotics in surgery, for instance, presents a fascinating area for exploration.

In conclusion, the future research directions proposed here aim to build upon the foundational work of this dissertation, seeking to push the boundaries of AI in biomedical imaging further. By addressing these potential areas for research and methodological improvements, the field can continue to evolve and make substantial contributions to medical science and patient care.

5.6 General Conclusions

This dissertation has traversed a significant journey in the realm of AI-enhanced biomedical imaging, demonstrating how advanced computational techniques can revolutionize our understanding and capabilities in this field. The collective findings from the individual studies present a comprehensive narrative of innovation, challenge, and progress.

1. **Advancement in Predictive Analytics**: The dissertation has significantly contributed to predictive analytics in biomedical imaging, particularly in spatial transcriptomics and histopathology. It has shown how AI can be harnessed to predict complex biological behaviors and improve the accuracy of medical diagnoses.

2. **Efficiency in Data Management**: The research underscored the importance of efficient data management in digital pathology. By introducing advanced image compression techniques, it has paved the way for more effective handling and analysis of large-scale histopathological data.

3. **Enhancement of Data Accessibility and Analysis**: The studies emphasized the need for sophisticated systems for indexing and searching histopathology slides. Improving these systems is key to making the vast amounts of digital pathology data more accessible and analyzable, thereby enhancing both clinical practice and medical research.

4. **Implications for Clinical Practice and Research**: The methodologies and findings of this dissertation hold significant implications for clinical practice and medical research. They demonstrate the potential of AI to enhance the precision and efficiency of medical imaging analysis, which is crucial for both patient care and scientific discovery.

5.7 Final Remarks

The journey of this research has been one of exploration, discovery, and innovation. It has highlighted the immense potential of AI in transforming biomedical imaging, opening new doors for diagnosis, prognosis, and treatment in medicine. This work has laid a foundation for future research, proposing new directions and methodologies that could further advance the field.

The future of AI in biomedical imaging is bright, with endless possibilities for enhancing patient care and medical research. As technology continues to evolve, so too will the methods and applications of AI in this field, promising a future where AI and medicine are inextricably linked for the betterment of human health.

A

Copyright Permission for Paper 2

CCC RightsLink

Sign in/Register

IEEE Requesting permission to reuse content from an IEEE publication

Clinically Relevant Latent Space Embedding of Cancer Histopathology Slides Through Variational Autoencoder based Image Compression

Conference Proceedings:
2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)

Author: Mohammad Sadegh Nasr

Publisher: IEEE

Date: 18 April 2023

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

B

Multiple Author Release Forms

UNIVERSITY OF
TEXAS
ARLINGTON
GRADUATE SCHOOL

Multiple Author Release for Master's Thesis or Doctoral Dissertation

Thesis / Dissertation Writer Name: Mohammad Sadegh Nasr

Co-Author Name: Parisa Boodaghi Malidarreh

Title(s) of Co-Authored Work(s):

"Predicting Future States with Spatial ...

Point Processes in Single Molecule ...

Resolution Spatial Transcriptomics"

Co-Author Statement of Consent:

As the co-author of the above named work(s), I acknowledge the above-named thesis / dissertation writer contributed substantially to the content of the work(s) listed above. I authorize the thesis / dissertation writer named above to use the listed work(s) in their thesis / dissertation.

Co-Author Signature:

Date: 12/5/2023

Please maintain a copy of this completed form for your records.

You may be entitled to know what information The University of Texas at Arlington (UT Arlington) collects concerning you. You may review and have UT Arlington correct this information according to procedures set forth in UTS 139. The law is found in sections 552.021, 552.023 and 559.004 of the Texas Government Code.

UNIVERSITY OF TEXAS ARLINGTON | GRADUATE SCHOOL

Multiple Author Release for Master's Thesis or Doctoral Dissertation

Thesis / Dissertation Writer Name: Mohammad Sadegh Nasr

Co-Author Name: Biraaj Rout

Title(s) of Co-Authored Work(s):

"Predicting Future States with Spatial ...

Point Processes in Single Molecule ...

Resolution Spatial Transcriptomics"

Co-Author Statement of Consent:

As the co-author of the above named work(s), I acknowledge the above-named thesis / dissertation writer contributed substantially to the content of the work(s) listed above. I authorize the thesis / dissertation writer named above to use the listed work(s) in their thesis / dissertation.

Co-Author Signature: Brout

Date: 12/05/2023

Please maintain a copy of this completed form for your records.

You may be entitled to know what information The University of Texas at Arlington (UT Arlington) collects concerning you. You may review and have UT Arlington correct this information according to procedures set forth in UTS 139. The law is found in sections 552.021, 552.023 and 559.004 of the Texas Government Code.

UNIVERSITY OF TEXAS ARLINGTON | GRADUATE SCHOOL

Multiple Author Release for Master's Thesis or Doctoral Dissertation

Thesis / Dissertation Writer Name: Mohammad Sadegh Nasr

Co-Author Name: Priyanshi Borad

Title(s) of Co-Authored Work(s):

"Predicting Future States with Spatial ...
Point Processes in Single Molecule ...
Resolution Spatial Transcriptomics"

Co-Author Statement of Consent:

As the co-author of the above named work(s), I acknowledge the above-named thesis / dissertation writer contributed substantially to the content of the work(s) listed above. I authorize the thesis / dissertation writer named above to use the listed work(s) in their thesis / dissertation.

Co-Author Signature: P.H.Borad

Date: 12/5/2023

Please maintain a copy of this completed form for your records.

UNIVERSITY OF TEXAS ARLINGTON | GRADUATE SCHOOL

Multiple Author Release for Master's Thesis or Doctoral Dissertation

Thesis / Dissertation Writer Name: Mohammad Sadegh Nasr

Co-Author Name: Jillur Rahman Saurav

Title(s) of Co-Authored Work(s):

"Predicting Future States with Spatial ...

Point Processes in Single Molecule ...

Resolution Spatial Transcriptomics"

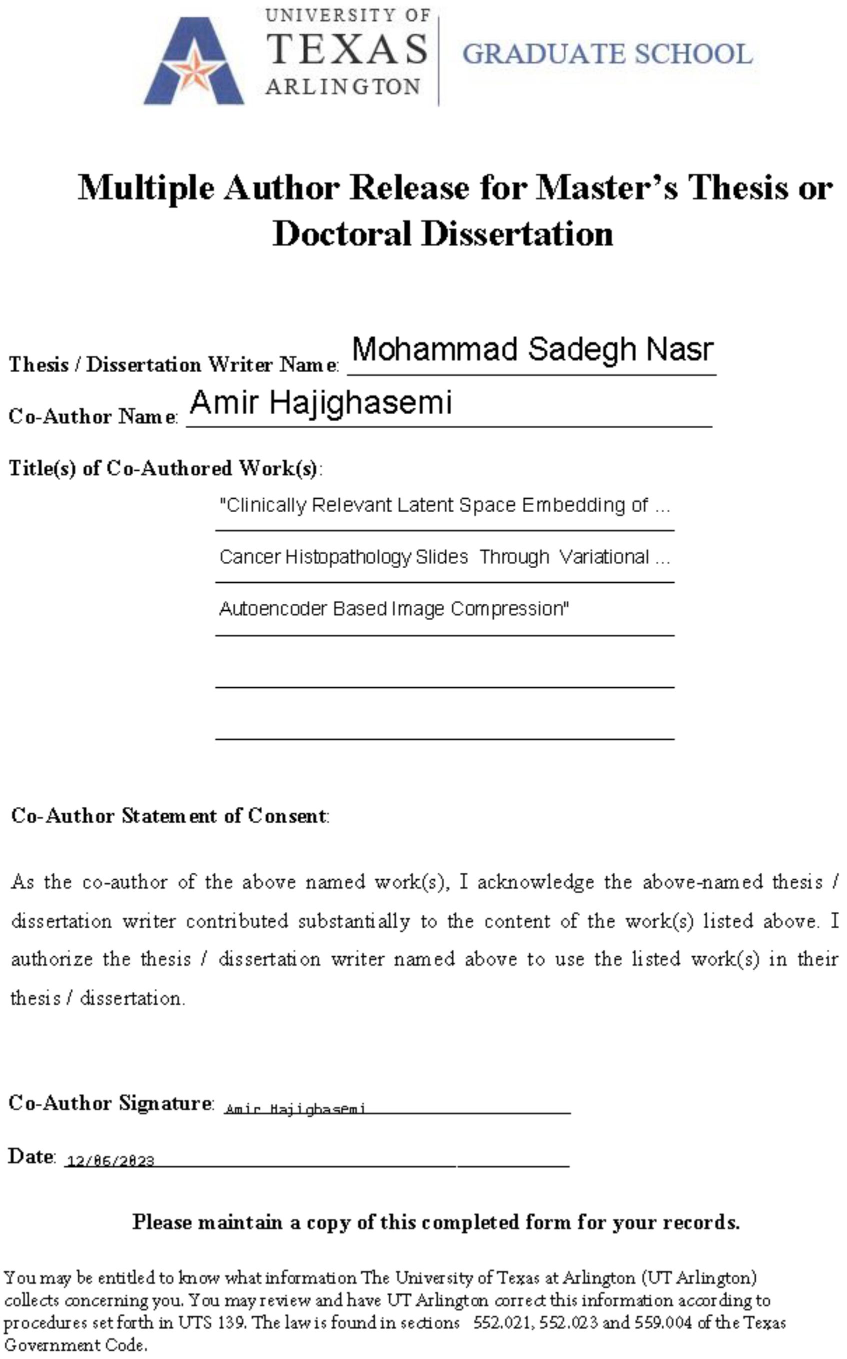
Co-Author Statement of Consent:

As the co-author of the above named work(s), I acknowledge the above-named thesis / dissertation writer contributed substantially to the content of the work(s) listed above. I authorize the thesis / dissertation writer named above to use the listed work(s) in their thesis / dissertation.

Co-Author Signature:

Date: 12/05/2023

Please maintain a copy of this completed form for your records.



UNIVERSITY OF TEXAS ARLINGTON | GRADUATE SCHOOL

Multiple Author Release for Master's Thesis or Doctoral Dissertation

Thesis / Dissertation Writer Name: Mohammad Sadegh Nasr

Co-Author Name: Amir Hajighasemi

Title(s) of Co-Authored Work(s):

"Clinically Relevant Latent Space Embedding of ...

Cancer Histopathology Slides Through Variational ...

Autoencoder Based Image Compression"

Co-Author Statement of Consent:

As the co-author of the above named work(s), I acknowledge the above-named thesis / dissertation writer contributed substantially to the content of the work(s) listed above. I authorize the thesis / dissertation writer named above to use the listed work(s) in their thesis / dissertation.

Co-Author Signature: Amir Hajighasemi

Date: 12/06/2023

Please maintain a copy of this completed form for your records.

You may be entitled to know what information The University of Texas at Arlington (UT Arlington) collects concerning you. You may review and have UT Arlington correct this information according to procedures set forth in UTS 139. The law is found in sections 552.021, 552.023 and 559.004 of the Texas Government Code.

UNIVERSITY OF TEXAS ARLINGTON | GRADUATE SCHOOL

Multiple Author Release for Master's Thesis or Doctoral Dissertation

Thesis / Dissertation Writer Name: Mohammad Sadegh Nasr

Co-Author Name: Helen H. Shang

Title(s) of Co-Authored Work(s):

"Histopathology Slide Indexing ...

and Search: Are We There Yet?"

Co-Author Statement of Consent:

As the co-author of the above named work(s), I acknowledge the above-named thesis / dissertation writer contributed substantially to the content of the work(s) listed above. I authorize the thesis / dissertation writer named above to use the listed work(s) in their thesis / dissertation.

Co-Author Signature:

Date: 12/5/23

Please maintain a copy of this completed form for your records.

References

[1] Amisha, Malik, P., Pathania, M., & Rathaur, V. K. (2019). Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care*, 8(7), 2328–2331.

[2] Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M., & Hamilton, P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1), 16878. Number: 1 Publisher: Nature Publishing Group.

[3] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8), 1301–1309. Number: 8 Publisher: Nature Publishing Group.

[4] Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, 538(7623), 20. Cg_type: Nature News Section: News Feature.

[5] Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine*, 378(11), 981–983. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMp1714229.

[6] Chen, C., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Schaumberg, A. J., & Mahmood, F. (2022). Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, 6(12), 1420–1434. Number: 12 Publisher: Nature Publishing Group.

[7] Chen, J., Suo, S., Tam, P. P., Han, J.-D. J., Peng, G., & Jing, N. (2017). Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nature Protocols*, 12(3), 566–580. Number: 3 Publisher: Nature Publishing Group.

[8] Dayao, M. T., Trevino, A., Kim, H., Ruffalo, M., D'Angio, H. B., Preska, R., Duvvuri, U., Mayer, A. T., & Bar-Joseph, Z. (2023). Deriving spatial features from in situ proteomics imaging to enhance cancer survival analysis. *Bioinformatics*, 39(Supplement_1), i140–i148.

[9] Dunipace, L., Saunders, A., Ashe, H. L., & Stathopoulos, A. (2013). Autoregulatory Feedback Controls Sequential Action of cis-Regulatory Modules at the brinker Locus. *Developmental Cell*, 26(5), 536–543.

[10] Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., & and the CAMELYON16 Consortium (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22), 2199–2210.

[11] Gerke, S., Minssen, T., & Cohen, G. (2020). Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare. In A. Bohr & K. Memarzadeh (Eds.), *Artificial Intelligence in Healthcare* (pp. 295–336). Academic Press.

[12] Giger, M. L. (2018). Machine Learning in Medical Imaging. *Journal of the American College of Radiology: JACR*, 15(3 Pt B), 512–520.

[13] Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological Image Analysis: A Review. *IEEE Reviews in Biomedical Engineering*, 2, 147–171. Conference Name: IEEE Reviews in Biomedical Engineering.

[14] He, L., Long, L. R., Antani, S., & Thoma, G. R. (2012). Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine*, 107(3), 538–556.

[15] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4), e1312. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1312.

[16] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature reviews. Cancer*, 18(8), 500–510.

[17] Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., & Saltz, J. H. (2016). Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification.

In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2424–2433). ISSN: 1063-6919.

[18] Hu, Y., Yang, W., Ma, Z., & Liu, J. (2022). Learning End-to-End Lossy Image Compression: A Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8), 4194–4211. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[19] Jamil, S., Piran, M. J., Rahman, M., & Kwon, O.-J. (2023). Learning-driven lossy image compression: A comprehensive survey. *Engineering Applications of Artificial Intelligence*, 123, 106361.

[20] Kalra, S., Tizhoosh, H. R., Choi, C., Shah, S., Diamandis, P., Campbell, C. J. V., & Pantanowitz, L. (2020). Yottixel – An Image Search Engine for Large Archives of Histopathology Whole Slide Images. *Medical Image Analysis*, 65, 101757.

[21] Kingma, D. P. & Welling, M. (2014). Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

[22] Komura, D. & Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, 16, 34–42.

[23] Konsti, J., Lundin, M., Linder, N., Haglund, C., Blomqvist, C., Nevanlinna, H., Aaltonen, K., Nordling, S., & Lundin, J. (2012). Effect of image compression and scaling on automated scoring of immunohistochemical stainings and segmentation of tumor epithelium. *Diagnostic Pathology*, 7(1), 29.

[24] Koromila, T. & Stathopoulos, A. (2019). Distinct Roles of Broadly Expressed Repressors Support Dynamic Enhancer Action and Change in Time. *Cell Reports*, 28(4), 855–863.e5.

[25] Krupinski, E. A., Johnson, J. P., Jaw, S., Graham, A. R., & Weinstein, R. S. (2012). Compressing pathology whole-slide images using a human and model observer evaluation. *Journal of Pathology Informatics*, 3(1), 17.

[26] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., & Kharchenko, P. V. (2018).

RNA velocity of single cells. *Nature*, 560(7719), 494–498. Number: 7719 Publisher: Nature Publishing Group.

[27] Langlotz, C. P. (2019). Will Artificial Intelligence Replace Radiologists? *Radiology: Artificial Intelligence*, 1(3), e190058. Publisher: Radiological Society of North America.

[28] Li, S., Zhao, Y., Zhang, J., Yu, T., Zhang, J., & Gao, Y. (2023). High-Order Correlation-Guided Slide-Level Histology Retrieval With Self-Supervised Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(09), 11008–11023. Publisher: IEEE Computer Society.

[29] Lim, B., Heist, T., Levine, M., & Fukaya, T. (2018). Visualization of Transvection in Living Drosophila Embryos. *Molecular Cell*, 70(2), 287–296.e6. Publisher: Elsevier.

[30] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.

[31] Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen van de Kaa, C., Bult, P., van Ginneken, B., & van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, 6(1), 26286. Number: 1 Publisher: Nature Publishing Group.

[32] Lombardo, S., HAN, J., Schroers, C., & Mandt, S. (2019). Deep Generative Video Compression. In *Advances in Neural Information Processing Systems*, volume 32: Curran Associates, Inc.

[33] Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell*, 167(5), 1170–1187.

[34] Lucas, T., Ferraro, T., Roelens, B., De Las Heras Chanes, J., Walczak, A. M., Coppey, M., & Dostatni, N. (2013). Live Imaging of Bicoid-Dependent Transcription in Drosophila Embryos. *Current Biology*, 23(21), 2135–2139.

[35] Madabhushi, A. & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33, 170–175.

[36] Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2019). Deep learning in radiology: An overview of the concepts and a survey of the state of the art with

focus on MRI. *Journal of Magnetic Resonance Imaging*, 49(4), 939–954. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.26534.

[37] Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., & Zhuang, X. (2016). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39), 11046–11051. Publisher: Proceedings of the National Academy of Sciences.

[38] Moor, A. E. & Itzkovitz, S. (2017). Spatial transcriptomics: paving the way for tissue-level systems biology. *Current Opinion in Biotechnology*, 46, 126–133.

[39] Naik, N., Madani, A., Esteva, A., Keskar, N. S., Press, M. F., Ruderman, D., Agus, D. B., & Socher, R. (2020). Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. *Nature Communications*, 11(1), 5727. Number: 1 Publisher: Nature Publishing Group.

[40] Niazi, M. K. K., Lin, Y., Liu, F., Ashok, A., Marcellin, M. W., Tozbikian, G., Gurcan, M. N., & Bilgin, A. (2019). Pathological image compression for big data image analysis: Application to hotspot detection in breast cancer. *Artificial Intelligence in Medicine*, 95, 82–87.

[41] Obermeyer, Z. & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375(13), 1216–1219. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMp1606181.

[42] Perry, M. W., Bothma, J. P., Luu, R. D., & Levine, M. (2012). Precision of Hunchback expression in the Drosophila embryo. *Current biology : CB*, 22(23), 2247–2252.

[43] Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., & Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434), 1463–1467. Publisher: American Association for the Advancement of Science.

[44] Saramago, P., Yang, H., Llewellyn, A., Walker, R., Harden, M., Palmer, S., Griffin, S., Simmonds, M., Saramago, P., Yang, H., Llewellyn, A., Walker, R., Harden, M., Palmer, S., Griffin, S., & Simmonds, M. (2018). *High-throughput non-invasive prenatal testing for fetal rhesus D status in RhD-negative women not known to be*

sensitised to the RhD antigen: a systematic review and economic evaluation. NIHR Journals Library.

[45] Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual review of biomedical engineering*, 19, 221–248.

[46] Soliman, H. S. & Omari, M. (2006). A neural networks approach to image data compression. *Applied Soft Computing*, 6(3), 258–271.

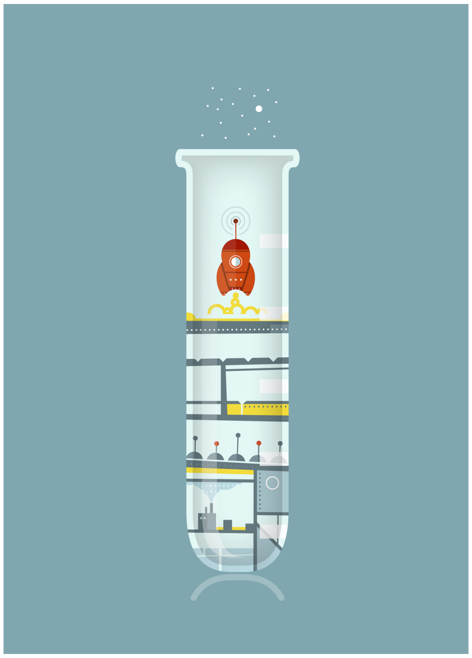
[47] Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, �., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., & Frisén, J. (2016). Visualization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

[48] Tellez, D., Litjens, G., van der Laak, J., & Ciompi, F. (2021). Neural Image Compression for Gigapixel Histopathology Image Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2), 567–578. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[49] Titano, J. J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., Swinburne, N., Zech, J., Kim, J., Bederson, J., Mocco, J., Drayer, B., Lehar, J., Cho, S., Costa, A., & Oermann, E. K. (2018). Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature Medicine*, 24(9), 1337–1341. Number: 9 Publisher: Nature Publishing Group.

[50] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. Number: 1 Publisher: Nature Publishing Group.

[51] Wang, X., Du, Y., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2023). RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis*, 83, 102645.



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.