

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Deep Learning Across Healthcare Spectrums: Genomic Insights, Social Determinants Analysis, and Imaging Diagnostics in Complex Diseases

Permalink

<https://escholarship.org/uc/item/54v9m94m>

Author

Sun, Shenghuan

Publication Date

2024

Peer reviewed|Thesis/dissertation

Deep Learning Across Healthcare Spectrums: Imaging Diagnostics, Social Determinants Analysis, and Genomic Insights in Complex Diseases

by
Shenghuan Sun

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Atul Butte Atul Butte
51E283C7897F40E... Chair

DocuSigned by:
Marina Sirota Marina Sirota

DocuSigned by:
Ashish Raj Ashish Raj

DocuSigned by:
Gregory Goldgof Gregory Goldgof
64BD269C9EDB417...

Committee Members

Copyright 2024

by

Shenghuan Sun

Dedication and Acknowledgements

Expressing my gratitude in words alone falls short of capturing the depth of my appreciation for the numerous individuals whose paths I've crossed in both my academic and professional journeys. Their influence has been pivotal in my growth, and I am eager to give back in any way I can in the future.

Prof. Atul Butte stands out as a beacon of inspiration for me. Under Atul's mentorship, I've gained invaluable insights into becoming a more impactful scientist, with a broad vision aimed at tackling tangible challenges. His guidance has steered me toward creating meaningful solutions that benefit healthcare professionals, patients, and the wider medical community. Atul's unwavering belief in my potential and his generous support through various hurdles have been nothing short of remarkable. My deepest thanks go to Prof. Butte for his relentless advocacy and mentorship over the last four years. Joining your team introduced me to many extraordinary people, and together, you have all directed me towards a fulfilling path that I am excited to follow for many years to come.

I am deeply grateful to Prof. Greg Goldgof, whose invaluable contribution has been essential to the success of my work. Reflecting on our first encounter, it was Greg who unveiled the fascinating realm of hematopathology to me, inspiring me to explore the application of computer vision in pathology. Our collaboration has since flourished, with Greg providing mentorship that delves into intricate details, elevating my understanding and approach to our shared projects. I fondly recall the numerous occasions when Greg patiently clarified complex medical concepts, engaging in thorough discussions about the most suitable machine learning frameworks for our endeavors. He has been an extraordinary mentor, a trusted friend, and an exemplary figure in my life.

I consider myself incredibly fortunate to have been under the mentorship of Prof. Ahmed Alaa, despite our acquaintance spanning only a year. The opportunity to share and refine my research ideas and experimental plans with him has been a privilege. His guidance has markedly enhanced my computer science capabilities, for which I am profoundly grateful. Prof. Alaa's unwavering support and assistance have been invaluable. Moreover, I am continually impressed by his exceptional writing prowess and the clarity and elegance of the figures he produces.

My understanding of clinical NLP has been profoundly shaped by Dr. Madhumita Sushil, whose intelligence and talent as a postdoctoral researcher in Atul's lab are nothing short of remarkable. I am confident in her future success within the academic sphere. With her assistance, I was able to contribute to two manuscripts, benefiting greatly from her comprehensive support and hands-on expertise. As she transitions to a faculty position, I am certain she will excel and make significant contributions to her field.

Collaboration and learning alongside my colleagues in Atul's lab have also been pivotal to my growth. I extend my gratitude to Zicheng, Travis, Chris for their invaluable assistance with my manuscripts, and to Michelle, Brenda, and Jayson for the enriching discussions we've shared. Special thanks to Boris for his technical support with the GPU servers. The collective support and encouragement from the team have been instrumental in the completion of numerous projects over the years.

In addition to the support from Atul's lab, I have been fortunate to receive guidance from Prof. Iain Carmichael in computational pathology. My rotations with Prof. Marina Sirota, Prof. Ashish Raj, and Prof. Jingjing Li were highly rewarding. Jingjing provided an exceptional introduction to my UCSF experience, while Marina's support was crucial for my first co-authored paper. Ashish introduced me to computational neuroscience, and although the rebuttal process for our paper was challenging, it was through these rigorous experiences that I gained invaluable learning and achieved my first paper as the lead author.

I am not who I am right now without my family's support. Though as the first generation who went to graduate school in my family, I did not receive a lot of career guidance. However, all my family have been working so hard for their own and each other's lives. I learned that attitude and use it in my career trajectory. My family are PhDs for their own lives, and I cannot thank them enough for what they taught me along the way. My partner might be the only one who help with my PhD study, there are many up and down along my PhD, she is always there to make me relaxed and realized there is always something more important in life than the PhD work. Now she is embarking for her journey of getting a PhD, hopefully I will be at least half helpful as she did to me. I also want to wish her good luck.

Originating from a quaint town in China, my decision to pursue studies in the USA was significantly influenced by my avid interest in the NBA. This journey, filled with transformative experiences, was made possible by the unwavering support of many individuals who have been pillars in my life. I vividly recall my initial days in the USA, marked by a distinct English accent that posed communication challenges, making even simple tasks like applying for a debit card at the Bank of America a prolonged endeavor of two hours. Reflecting on this journey, from those humble beginnings to where I stand today, fills me with amazement and gratitude. The journey has indeed been a remarkable adventure, punctuated with learning and growth.

My journey, academic and professional alike, would not have been the same without the myriad of individuals who have enriched my life with their wisdom, guidance, and opportunities. The culmination of my PhD is a testament to the collective wisdom and support bestowed upon me, particularly under the astute leadership and mentorship of Prof. Butte. This achievement is not solely my own but a mosaic of contributions from every person who has been part of my story.

Deep Learning Across Healthcare Spectrums: Genomic Insights, Social Determinants Analysis, and Imaging Diagnostics in Complex Diseases

Shenghuan Sun

Abstract

The burgeoning interest in leveraging deep learning within the medical field heralds a promising frontier for enhancing disease understanding and patient care. Yet, this technological advance is not without its challenges. One significant issue is the underutilization of diverse data types; medical records and biological factors, while crucial, do not encompass the entirety of necessary information. Social Determinants of Health (SDoH), for instance, play a pivotal role in disease comprehension but are often neglected in research. Furthermore, while deep learning holds potential for diagnosis and aiding clinical decisions, the absence of rigorous external validation undermines its reliability. Many models, despite performing well in initial settings, falter under broader, real-world scrutiny. Additionally, the tendency to harness large datasets and maximize feature inclusion for disease analysis sometimes overshadows the value of engineered features. These more targeted, hypothesis-driven attributes can sometimes offer clearer insights into disease mechanisms, a nuance that is frequently overlooked in the rush towards big data approaches.

These challenges manifest distinctly across different data modalities in medical research. In the realm of Electronic Health Records (EHR), the exploration of disease mechanisms often prioritizes medical data, inadvertently sidelining non-medical but equally vital Social Determinants of Health (SDoH) such as financial stability, mental health, and physical activity. This oversight can skew our understanding of disease etiology and patient outcomes. In medical imaging, the rapid development and deployment of deep learning models boast of enhanced diagnostic accuracy. Yet, this domain is particularly susceptible to the

pitfalls of insufficient external validation. Minor perturbations or "noise" within the imaging data can dramatically compromise the predictive reliability of these models, emphasizing the need for robust validation processes. Genomic studies, on the other hand, face the challenge of signal dilution amidst the vast array of genomic features. The pursuit of correlations across tens of thousands of genes often overlooks the critical influence of covariates and noise, potentially obscuring the true biological signals vital for understanding disease processes. Each of these issues highlights the complexity of medical data analysis and the need for nuanced approaches that consider the full spectrum of relevant factors.

This dissertation is dedicated to the development and application of innovative computational strategies, employing practical deep learning techniques to address these prevailing challenges. Firstly, it underscores the necessity of integrating comprehensive and meaningful features in deep learning research, with a particular emphasis on the inclusion of Social Determinants of Health (SDoH) factors, to present a more holistic view of disease mechanisms. Secondly, it demonstrates the imperative role of high-quality data, coupled with human feedback and rigorous external validation, in enhancing the reliability and applicability of deep learning frameworks within the medical domain. Thirdly, the dissertation advocates for the strategic use of high-level feature engineering, as opposed to relying on an overwhelming volume of features, to decipher complex biological systems.

Table of Contents

| | |
|--|----------|
| 1 Chapter 1: Introduction | 1 |
| 1.1 Dissertation Overview | 1 |
| 1.2 The following chapters | 2 |
| 2 Chapter 2: Topic Modeling on Clinical Social Work Notes for Exploring Social Determinants of Health Factors | 4 |
| 2.1 ABSTRACT | 5 |
| 2.2 INTRODUCTION..... | 7 |
| 2.3 MATERIALS AND METHODS..... | 9 |
| 2.4 RESULTS..... | 14 |
| 2.5 DISCUSSION | 17 |
| 2.6 CONCLUSION | 20 |
| 2.7 ACKNOWLEDGEMENTS | 21 |
| 2.8 DATA AVAILABILITY..... | 21 |
| 2.9 COMPETING INTERESTS STATEMENT | 21 |
| 2.10 FUNDING STATEMENT..... | 22 |
| 2.11 REFERENCES | 32 |

3 Chapter 3: Revealing the impact of social circumstances on the selection of cancer

| | |
|---|-----------|
| therapy through natural language processing of social work notes | 37 |
| 3.1 ABSTRACT | 38 |
| 3.2 INTRODUCTION..... | 39 |
| 3.3 MATERIALS AND METHODS..... | 42 |
| 3.4 RESULTS | 45 |
| 3.5 DISCUSSION | 48 |
| 3.6 CONCLUSION | 50 |
| 3.7 CONTRIBUTORSHIP STATEMENT | 51 |
| 3.8 ACKNOWLEDGEMENTS | 51 |
| 3.9 COMPETING INTERESTS STATEMENT | 51 |
| 3.10 FUNDNG STATEMENT | 52 |
| 3.11 DATA AVAILBILTY STATEMENT | 52 |
| 3.12 REFERENCES..... | 74 |

4 Chapter 4: Aligning Synthetic Medical Images with Clinical Knowledge using

| | |
|-----------------------------|-----------|
| Human Feedback | 78 |
| 4.1 ABSTRACT | 78 |
| 4.2 INTRODUCTION..... | 79 |
| 4.3 RELATED WORK..... | 87 |
| 4.5 RESULTS | 90 |
| 4.6 CONCLUSIONS..... | 93 |

| | |
|---|------------|
| 4.7 REFERENCES..... | 111 |
| 5 Chapter 5: Spatial cell-type enrichment predicts mouse brain connectivity | 117 |
| 5.1 ABSTRACT | 117 |
| 5.2 INTRODUCTION..... | 118 |
| 5.3 RESULTS..... | 120 |
| 5.4 DISCUSSION | 130 |
| 5.5 FUTURE DIRECTIONS..... | 137 |
| 5.6 CONCLUSIONS..... | 139 |
| 5.7 ACKNOWLEDGMENTS | 139 |
| 5.8 DECLARATION OF INTERESTS..... | 139 |
| 5.10 REFERENCES..... | 172 |
| 6 Chapter 6: DeepHeme: A High-Performance, Generalizable, Deep Ensemble for Bone Marrow Morphometry and Hematologic Diagnosis..... | 178 |
| 6.1 ABSTRACT | 179 |
| 6.2 INTRODUCTION..... | 179 |
| 6.3 RESULTS..... | 181 |
| 6.4 DISCUSSION | 188 |
| 6.5 METHODS..... | 190 |
| 6.6 REFERENCES..... | 216 |

| | |
|--|------------|
| 7 Chapter 7: Conclusion: Summary and Future Work..... | 222 |
| 7.1 SUMMARY | 222 |
| 7.2 FUTURE WORK | 226 |

List of Figures

| | |
|--|-----|
| Figure 2.1 Retrieval of clinical social work notes for the study | 23 |
| Figure 2.2 Topic proportion comparison for different categories | 24 |
| Figure 2.3 Word frequency calculation..... | 25 |
| Figure 3.1 The overall workflow | 53 |
| Figure 3.2 Data exploration on social work notes | 54 |
| Figure 3.3 Illustration of BERT-MS-n model..... | 55 |
| Figure 3.4 Feature importance analysis for SDOH factors in ablation study | 56 |
| Figure 3.5 Pie chart showing the different proportions | 57 |
| Figure 3.6 Example deidentified social work notes contain abusive history information..... | 59 |
| Figure 4.1 Overview of our pathologist-in-the-loop synthetic data generation framework | 94 |
| Figure 4.2 Samples for biologically implausible synthetic images | 95 |
| Figure 4.3 Generation of refined cell sub-types. | 96 |
| Figure 4.4 Quantitative and qualitative impact of pathologist feedback on synthetic images | 97 |
| Figure 4.5 Representative samples from different baseline models | 98 |
| Figure 4.6 Representative samples from the conditional diffusion model before (left) and after (right) incorporating pathologist feedback. | 99 |
| Figure 4.7 Representative samples from the conditional diffusion model. | 100 |
| Figure 4.8 Representative samples from the finetuned model with 10% of the pathologist feedback points. | 101 |
| Figure 4.9 Representative samples from the finetuned model with 50% of the pathologist feedback points. | 102 |
| Figure 4.10 Representative samples from the finetuned model with 100% of the pathologist feedback points | 103 |

| | |
|---|-----|
| Figure 5.1 Study design | 144 |
| Figure 5.2 Machine learning applied to regional cell type distributions predicts both the existence of connectivity and connectivity density | 145 |
| Figure 5.3 Interrogating the individual contributions of cell types | 146 |
| Figure 5.4 Most important cell type contributors vary depending on inter-regional distance..... | 147 |
| Figure 5.5 Distribution of top contributors to long-range connectivity (from Zeisel, et al. data) | 149 |
| Figure 5.6 AMBCA connectivity matrix properties | 150 |
| Figure 5.7 Confusion matrices for binary connectome prediction. Performance is shown for both the Zeisel, <i>et al.</i> | 151 |
| Figure 5.8 Zeisel, <i>et al.</i> similarity and connectome prediction, ipsilateral and contralateral. | 152 |
| Figure 5.9 Random forest predictions with and without zero-filtering. | 153 |
| Figure 5.10 MISS in-sample and out-of-sample error for the Tasic, <i>et al.</i> , dataset. | 154 |
| Figure 5.11 Connectivity prediction using the Tasic, <i>et al.</i> dataset | 155 |
| Figure 5.12 Tasic, <i>et al.</i> similarity and connectome prediction, ipsilateral and contralateral | 156 |
| Figure 5.13 Correspondence between Zeisel, <i>et al.</i> and Tasic, <i>et al.</i> similarity matrices, ipsilateral and contralateral..... | 157 |
| Figure 5.14 Feature importance, Tasic, <i>et al.</i> dataset | 158 |
| Figure 5.15 Feature importance, binary connectivity prediction..... | 159 |
| Figure 5.16 Feature importance for neocortical-to-other and other-to-neocortical connectivity density prediction, separated source and target cell-type features, Zeisel, <i>et al.</i> | 160 |
| Figure 5.17 Feature importance, short- and long-range connectivity, Tasic, <i>et al.</i> | 161 |
| Figure 5.18 Feature importance for short- and long-range connectivity prediction, separated source and target cell-type features, Zeisel, <i>et al.</i> | 162 |
| Figure 5.19 Feature importance for short- and long-range connectivity prediction, separated source and target cell-type features, Tasic, <i>et al.</i> | 163 |
| Figure 5.20 Taxonomic distance..... | 164 |

| | |
|--|-----|
| Figure 6.1 Workflow of DeepHeme-SE..... | 196 |
| Figure 6.2 External validation and expert evaluation..... | 197 |
| Figure 6.3 Exploring Advanced Snapshot Ensemble in Bone Marrow Cell Classification | 198 |
| Figure 6.4 Model Learns Underlying Hematopoietic Developmental Relationships..... | 199 |
| Figure 6.5 Application of DeepHeme-SE in diagnosing different leukemia types. | 200 |
| Figure 6.6 Software for Supporting DeepHeme-SE | 201 |
| Figure 6.7 Workflow of DeepHeme..... | 202 |
| Figure 6.8 Snapshot ensemble concept..... | 203 |
| Figure 6.9 Venn Diagram..... | 204 |
| Figure 6.10 UpSet Plot | 205 |
| Figure 6.11 UMAP embedding for cells with high nucleus:cytoplasm (N:C) ratio | 206 |
| Figure 6.12 One-vs-One AUC Analysis | 207 |
| Figure 6.13 Confusion Matrix on UCSF dataset | 208 |
| Figure 6.14 Confusion Matrix on MSF dataset | 209 |
| Figure 6.15 Saliency Maps | 210 |

List of Tables

| | |
|--|-----|
| Table 3.1 Model performance of different classifiers..... | 60 |
| Table 3.2 BERT MS model achieved superior performance in AUC, MACRO F1, as well as MACRO RECALL. | 61 |
| Table 3.3 Demographic characteristics for breast cancer patients in our cohort..... | 62 |
| Table 3.4 Summary characteristics of social factors (smoking and marital status) for breast cancer patients extracted from structured data | 63 |
| Table 3.5 Model performances of common machine learning classifiers using SDOH related structured tabular data on targeted therapy administration. | 64 |
| Table 3.6 The properties of notes for breast cancer patient’s cohort. (Measure the tokens length and compare with 512 tokens). | 65 |
| Table 3.7 The words in the Keywords column are the representative words used to define the topics .. | 66 |
| Table 3.8 Model performance of different classifiers..... | 67 |
| Table 3.9 The removal of notes drug mentioning in the prediction pipeline..... | 68 |
| Table 3.10 Model performance of leveraging SDOH topics appearance on regimen prediction, without semantic meanings. | 69 |
| Table 3.11 Demographic characteristics for all breast cancer patients..... | 70 |
| Table 3.12 Summary characteristics of social factors (smoking and marital status) for all breast cancer patients | 71 |
| Table 3.14 | 73 |
| Table 4.1 Pathologist Evaluation Criteria..... | 104 |
| Table 4.2 Breakdown of bone marrow image patches by morphological cell type and pathologist feedback. | 105 |
| Table 4.3 Expert evaluation of synthetic data..... | 106 |

| | |
|---|-----|
| Table 4.4 Evaluation of synthetic data using fidelity and diversity metrics. | 107 |
| Table 4.5 Accuracy of classifiers trained on real and synthetic data. | 108 |
| Table 4.6 Performance comparison for baseline generative models in the cell classification task. | 109 |
| Table 4.7 Model finetuned with the new subtypes of Neutrophil cells. | 110 |
| Table 5.1 Random forest model performance..... | 165 |
| Table 5.2 Dataset information for each of the three source datasets used..... | 166 |
| Table 5.3 Model comparison for classification, Zeisel, et al..... | 167 |
| Table 5.4 Model comparison for regression, Zeisel, et al..... | 168 |
| Table 5.5 Model performance of training and testing the random forest model using two different methods | 169 |
| Table 5.6 Model comparison for classification, Tasic, et al..... | 170 |
| Table 5.7 Model comparison for regression, Tasic, et al. | 171 |
| Table 6.1 Multi-institutional Datasets..... | 211 |
| Table 6.2 Precision score from three experts..... | 212 |
| Table 6.3 Recall score from three experts..... | 213 |
| Table 6.4 Comparison to other deep-learning-based bone marrow cell classifiers | 214 |
| Table 6.5 Comparison between Single Snapshot Model, Standard Learning Rate Scheduler, and Snapshot Ensemble Model Performances. | 215 |

1 Chapter 1: Introduction

This chapter provides an overview of the dissertation. The primary motivation and research need of this dissertation will be summarized followed by concise description of each following chapter.

1.1 Dissertation Overview

The integration of deep learning into healthcare represents a significant leap forward in our ability to diagnose, understand, and treat complex diseases. Yet, the transition from theoretical models to practical applications is fraught with challenges that span the collection and analysis of diverse data types, from electronic health records (EHR) and medical imaging to genomic data. Central to these challenges is the underutilization of critical data types such as Social Determinants of Health (SDoH), which are often overlooked in favor of more traditional medical data, despite their proven impact on health outcomes. Moreover, the reliance on large datasets and the drive to include as many features as possible can sometimes obscure rather than illuminate the underlying mechanisms of diseases.

This dissertation contends with these obstacles by proposing innovative computational strategies that harness deep learning to improve our understanding and management of complex diseases.

- 1) Comprehensive Data Utilization: It stresses the importance of integrating diverse data types, including Social Determinants of Health (SDoH), to provide a more complete picture of disease mechanisms and patient care.
- 2) Human Expertise: The work highlights the invaluable role of human feedback and expertise in refining deep learning models, ensuring their relevance and reliability in clinical applications.

- 3) Strategic Feature Engineering: The dissertation advocates for careful feature selection, such as prioritizing cell type over gene enrichment, to improve model interpretability and efficacy in uncovering disease insights.

1.2 The following chapters

Chapter 2: "Topic Modeling on Clinical Social Work Notes for Exploring Social Determinants of Health Factors" - This chapter explores the use of topic modeling to analyze social work notes, highlighting the significance of Social Determinants of Health (SDoH) in understanding patient health and treatment outcomes.

Chapter 3: "Revealing the impact of social circumstances on the selection of cancer therapy through natural language processing of social work notes" - It focuses on the practical application of SDoH information, demonstrating how social circumstances influence cancer therapy selection, underscoring the importance of a holistic view in treatment decisions.

Chapter 4: "Aligning Synthetic Medical Images with Clinical Knowledge using Human Feedback" - This chapter illustrates the integration of human expertise with deep learning, particularly in enhancing the accuracy and clinical relevance of synthetic medical images through feedback.

Chapter 5: "Spatial Cell Type Enrichment Predicts Mouse Brain Connectivity" - Highlights strategic feature engineering by employing spatial cell type enrichment to improve the understanding of mouse brain connectivity, showcasing the importance of focused and relevant feature selection in model development.

Chapter 6: "DeepHeme: A High-Performance, Generalizable, Deep Ensemble for Bone Marrow Morphometry and Hematologic Diagnosis" - Demonstrates the value of extensive human annotation and

external validation in developing a deep learning ensemble that advances bone marrow analysis and hematologic diagnosis, ensuring model reliability and applicability.

Chapter 7: “Conclusion: Summary And Future Work” provides a summary for the dissertation and put forward the exciting future work.

2 Chapter 2: Topic Modeling on Clinical Social Work Notes for Exploring Social Determinants of Health Factors

Shenghuan Sun, B.S. 1, Travis Zack, MD, PhD^{1,4}, Christopher Y.K. Williams, MD 1, Madhumita Sushil, PhD^{1†}, Atul J. Butte, MD, PhD* 1, 2, 3†

1. Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA
2. Center for Data-driven Insights and Innovation, University of California, Office of the President, Oakland, CA, USA
3. Department of Pediatrics, University of California, San Francisco, CA, 94158, USA
4. Division of Hematology/Oncology, Department of Medicine, UCSF, San Francisco, California, USA.

*Author to whom correspondence should be addressed.

† Equal Contribution

2.1 ABSTRACT

OBJECTIVE

Existing research on social determinants of health (SDoH) predominantly focuses on physician notes and structured data within Electronic Medical Records (EMRs). This study posits that social work notes are an untapped, potentially rich source for SDoH information. We hypothesize that clinical notes recorded by social workers, whose role is to ameliorate social and economic factors, might provide a complementary information source of data on SDoH compared to physician notes, which primarily concentrate on medical diagnoses and treatments. We aimed to use word frequency analysis and topic modeling to identify prevalent terms and robust topics of discussion within a large cohort of social work notes including both outpatient and in-patient consultations.

MATERIALS AND METHODS

We retrieved a diverse, deidentified corpus of 0.95 million clinical social work notes from 181,644 patients at the University of California, San Francisco. We conducted word frequency analysis related to ICD-10 chapters to identify prevalent terms within the notes. We then applied Latent Dirichlet Allocation (LDA) topic modeling analysis to characterize this corpus and identify potential topics of discussion, which was further stratified by note types and disease groups.

RESULTS

Word frequency analysis primarily identified medical-related terms associated with specific ICD10 chapters, though it also detected some subtle SDoH terms. In contrast, the LDA topic modeling analysis extracted 11 topics explicitly related to social determinants of health risk factors, such as financial status, abuse history, social support, risk of death, and mental health. The topic modeling approach effectively demonstrated variations between different types of social work notes and across patients with different types of diseases or conditions.

DISCUSSION

Our findings highlight LDA topic modeling's effectiveness in extracting SDoH-related themes and capturing variations in social work notes, demonstrating its potential for informing targeted interventions for at-risk populations.

CONCLUSION

Social work notes offer a wealth of unique and valuable information on an individual's SDoH. These notes present consistent and meaningful topics of discussion that can be effectively analyzed and utilized to improve patient care and inform targeted interventions for at-risk populations.

LAY SUMMARY

This study explored the untapped potential of social work notes to understand health-related factors shaped by our social and economic backgrounds. While past research often turned to doctor's notes or specific sections of medical records, the insights within social worker notes, which detail individuals' social challenges, remained largely uncharted. Analyzing close to a million such notes from the University of California, San Francisco, using standard and rigorously measured methods, we found 11 main discussion themes related to social and economic health risks. These themes covered areas like financial challenges, history of abuse, and mental well-being. Our findings suggest that social work notes provide valuable context about patients' life situations. Utilizing this information could be instrumental in creating more personalized care strategies for individuals navigating challenges stemming from their social and economic circumstances.

2.2 INTRODUCTION

Social determinants of health (SDoH) are non-medical factors that influence health outcomes, including the conditions in which people are born, grow, work, live, and age, as well as the wider set of forces and systems shaping daily life, such as economic policies, development agendas, social norms, and political systems [1,2,3,4]. These factors contribute significantly to health disparities due to systemic disadvantages and biases [5,6]. Systemic disadvantages refer to unequal distribution of resources and opportunities, while bias refers to unfair treatment based on social, economic, or demographic characteristics. Health inequities, which are unfair and avoidable differences in health among population groups, can arise from these determinants and warrant ethical consideration [7]. For example, mental health during pregnancy plays a pivotal role in both the mother's and the unborn child's well-being [8]. In a similar vein, lifestyle choices and living environments are intricately linked to the health outcomes of diabetes patients, with significant correlations observed [9]. These examples illustrate how systemic disadvantages and biases contribute to health inequities, underlining the importance of addressing SDoH in medical treatments for these conditions [5,10,11,12].

Social work notes written by social workers contain comprehensive information on social determinants of health (SDoH) compared to other common clinical note types documented by clinicians or medical professionals. Examples of social aspects covered in social work notes include living conditions, family support, access to transportation, employment status, and education level. While other types of notes such as nursing notes, discharge summaries and hospital progress notes may include some SDoH-related information such as insurance status, and health related aspect such as food and physical environment, they typically focus on specific aspects of patient care and may not provide as extensive information on SDoH as social work notes, which are written to provide a more complete view of these factors [13,14,15]. However, our capacity to research sociodemographic and socioeconomic health outcomes is still quite constrained. Most assessments of SDoH are not present in structured data [16]. Instead, much of this

information is collected in unstructured notes, making the information largely inaccessible without advanced technical processing. The inability to easily extract this information limits research into the effects of SDoH on care delivery and success.

To understand the information embedded in the social work notes and to characterize specific SDoH factors covered across nearly one million notes, we explored the use of unsupervised methods for topic modeling. Topic modeling methods based on Latent Dirichlet Allocation (LDA) have been previously successful in finding hidden structures (topics) from large corpora [17,18], the utility of which we further explored in this study. The large collection of social work notes analyzed in this study spanned a diverse cohort of patient demographics and disease groups. This allowed us to develop a comprehensive understanding of the underlying SDoH topics from different note types for a variety of disease chapters. We explored several methods to circumvent the inherent limitations of topic modeling approaches, such as pre-determining a fixed number of clusters, intrinsic randomness, and need for human-based interpretation.

BACKGROUND AND SIGNIFICANCE

Computational understanding of the free text in clinical notes is well known to be an open challenge, including the extraction of structured information from these documents [19]. Some progress has been made in extracting SDoH factors from clinical text using named entity recognition (NER), an NLP method of extracting pre-defined concepts from text [20,21]. Both machine learning-based and traditional rule-based NER have been developed and tested [21-23]. While NER approaches have been shown to be effective, they can be time-consuming [24].

Topic modeling methods have been widely applied for unbiased topic discovery from large collections of documents [25-27] and have been used in the fields of social science[28], environmental science[29], political science[30], and in biological and medical contexts[13]. Recent studies, such as work by Meaney C et al[13], have begun to explore latent topics in clinical notes. However, to our knowledge, LDA topic

modeling has not been heavily used to assess corpora of social work notes for SDoH factors, likely due to the general availability of sufficiently large corpora.

Clinical social workers are licensed professionals who specialize in identifying and addressing social and environmental barriers experienced by patients. In particular, text notes documented by clinical social workers are an invaluable data resource for understanding SDoH information in patients. As such, the clinical notes written by social workers often include specific text capturing an individual's SDoH. Yet, to date, social work notes have been a relatively under-utilized data source and have not been extensively investigated for understanding SDoH[31].

This study aims to explore the potential of social work notes as a rich source of data on social determinants of health (SDoH) by analyzing the most meaningful social work terminology across different disease chapters and applying Latent Dirichlet Allocation (LDA) topic modeling to identify robust topics of discussion within a large cohort of social work notes. By doing so, we seek to uncover clinically relevant SDoH information contained in these notes and their potential impact on patient and public health, demonstrating the value of social work notes in understanding SDoH factors.

2.3 MATERIALS AND METHODS

DATA SOURCES AND PATIENT DEMOGRAPHICS

This study uses the deidentified clinical notes at UCSF recorded between 2012 and 2021[32]. The study was approved by the Institutional Review Board (IRB) of the University of California, San Francisco (UCSF; IRB #18-25163). Our cohort consists of the following demographic distribution: Gender - Male: 95,387 (52.5%), Female: 85,635 (47.1%); Race - White: 22,839 (12.6%), Black: 21,120 (11.6%), Asian: 47,723 (26.3%), Native American: 14,813 (8.2%), Other: 75,149 (41.4%); Age - Median: 33 years (Range: 12-58); Ethnicity - Hispanic: 41,386 (22.8%), Non-Hispanic: 128,018 (70.5%).

DATA PREPROCESSING

We initiated our research by collecting clinical notes from a de-identified dataset, specifically selecting those entries where the metadata contained the term 'social'—case-insensitive—within the encounter type, department name, specialty, or provider type, thus designating these as ‘social work notes’. From the extensive corpus of 106 million notes representing 1.2 million patients, this focused query yielded 2.5 million social work notes attributed to 181,644 unique patients. To ensure the quality and relevance of our data, we excluded notes under 30 characters, anticipating they would not provide substantial content. Duplicate notes were also removed to eliminate redundancy and decrease computational demands. Following this stringent quality control process, we distilled the dataset down to 1 million notes corresponding to the same 181,644 patients, which formed the basis for our downstream topic modeling analysis, as depicted in Figure 2.1.

TOPIC MODELING WITH LATENT DIRICHLET ALLOCATION (LDA) ANALYSIS

While word frequency calculations can provide preliminary insights about term relevance, this view is too limited to understand what broader topics may be contained within social work notes. In contrast, topic modeling is a field of unsupervised learning that learns statistical associations between words or groups of words to identify “topics”: clusters of words that tend to co-occur within the same document. Latent Dirichlet Allocation (LDA) is a generative probabilistic model, that assumes that each document is a combination of a few different topics, and that each word's presence can be attributed to particular topics in the document. The result is a list of clusters, each of which contains a collection of distinct words. The combination of words in a cluster can be used for topic model interpretation. Python package *gensim* was used for the implementation [33]. We used *gensim.models.ldamodel.LdaModel* for the actual analysis. The core estimation code is based on Hoffman et al[34]. Python package *nltk* was used. As a preprocessing step, English language stop words and special characters including ‘\t’, ‘\n’, ‘\s’ were removed from note text. The resulting text from all social work notes were vectorized and topics were inferred with the LDA algorithm. In addition to the analysis on the complete cohort of social work notes, in order to investigate

the topic distribution across specific social work note categories, we additionally analyzed the four largest categories of social work notes: Progress Notes, Interdisciplinary, Telephone encounters, and Group Notes. We also extended the investigation to social work note subsets across 10 ICD-10 disease chapters. These subsets were determined by investigating encounter-specific ICD-10 diagnostic codes. The common stop words were also excluded, using *stopwords.words('english')* from nltk package[35]. To overcome the inherent stochasticity of topic modeling approaches and ensure the reliability of our findings, we ran five independent modeling analyses for each category of notes. This allowed us to capture consistent patterns and topics across different iterations, increasing our confidence in the identified topics and their relevance to the respective disease groups. Another critical step in LDA topic modeling was determining the optimal cluster number, which is further discussed in the next sub-section. Furthermore, when extending the analysis to different note types, we labeled the inferred topics using heuristics described further.

DETERMINING THE OPTIMAL NUMBER OF TOPICS FOR NOTES

One of the most important hyper-parameters for LDA analysis is the number of topics K . Generally, if K is chosen to be too small, the model will lack the capacity to provide a holistic summary of complex document collections; and returned topical vectors may combine semantically unrelated words/tokens[36]. Conversely, if K is chosen to be too large, the returned topical vectors may be redundant, and a parsimonious explanation of a complex phenomenon may not be achieved. We used two evaluation metrics, topic coherence [37,38] and topic similarity [39], to systematically determine the optimal number of clusters. Topic Coherence (C) quantifies the score of a single topic by measuring the degree of semantic similarity between high-scoring words in the topic[40]. The measure helps distinguish between topics that are semantically interpretable and those that are artifacts of statistical inference. The coherence metric we compute is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity[38]. Similarly, topic similarity (S) measures how similar two clusters are considering the words contained in the topics. The lower the values are, the less redundant the topic distribution is. For quantifying topic similarity,

we use Jaccard similarity[39]. Furthermore, there are alternative ways to evaluate the quality of topic discovery, such as assessing 'topic diversity'[41]. Considering these evaluation metrics in future work may provide further insights into the performance of our methods. An ideal solution would have a high topic coherence and low similarity metric. To decide the optimal number of clusters, for each analysis, we ran the LDA analysis with the number of clusters K ranging from 10 to 50, simultaneously computing C and S scores. The number having the i^{th} highest C value, j^{th} smallest S value, and the minimum $i + j$ among all runs was selected as the final number of topics (Figure 2.3). We found that the best cluster number for analyzing the entire notes repository was 17.

TOPIC MODELING PER NOTES WITH CERTAIN TYPE

In order to investigate the topic distribution across specific note categories, we applied topic modeling on the four largest categories of social work notes: Progress Notes, Interdisciplinary, Telephone encounters, and Group Notes. This approach allowed us to gain insights into the prevalence of certain topics within these major categories and assess their potential impact on the overall topic modeling results. We used the same pipeline for identifying the optimal number of clusters as described earlier in the Methods section. To ensure robustness in our results, given the inherent randomness of the LDA method, we conducted each analysis across five different iterations for every category. This approach allowed us to capture a broader range of variability, thereby increasing the reliability of our findings. The results from these five iterations were then pooled together. This pooling strategy was instrumental in developing a well-grounded heuristic for labeling the topic clusters, ensuring our results were reflective of consistent patterns observed across all iterations, rather than being influenced by any single run's anomalies. In our analysis, we determined that the optimal number of clusters for most of the analyses we conducted is approximately 20. This balances the trade-off between coherence and similarity metrics, ensuring that we obtain semantically interpretable and non-redundant topic clusters, which provide meaningful insights into the underlying document collection. Consequently, we used 20 clusters for the majority of our note analyses, including those focused on note subtypes or disease chapters. However, we found that the best cluster number for analyzing the

entire notes repository was 17, so we utilized 17 clusters for the topic modeling of all notes combined (See previous session).

Topic labelling heuristics. Apart from labeling topics determined from the entire cohort of social work notes, our analysis screened 20 topic clusters (determined experimentally; see Results) for all 14 categories of notes (10 disease chapters and 4 social work note types) for 5 independent runs (to reduce stochasticity), thereby resulting in 1400 topic clusters that required further labeling. To assign labels to all 1400 topics, we developed a heuristic to automatically assign topic labels for subsequent analyses, the details of which are discussed next. We first constructed the dictionary of topic names and the corresponding words by manually analyzing the topic modeling results for one run on the complete corpus of 0.95 million social work notes at UCSF. Then we expanded the individual topic clusters by first retrieving 20 most similar words to the words comprising topic clusters based on the cosine similarity of their word embeddings[42]. Any words that were not relevant to the topic label, as determined through manual review, were not considered further. The final dictionary of topic labels and the set of words used to label the topics is shown in Table 2. In our approach, we automatically assigned topic labels to individual word clusters by calculating the intersection over union (IOU) ratio for the words in a cluster. This enabled us to assign labels to all 1,400 topic clusters from our analysis. The details can be found in the pseudo-code below. To address your professor's concerns, we used the IOU of word frequencies within each cluster. We assigned the label with the maximum IOU, but only if there was an overlap of at least two words. If none of the topics met this criterion, we did not assign a topic to the word cluster.

Code for the paper is available on https://github.com/ShenghuanSun/LDA_TM

WORD FREQUENCY CALCULATION

To perform a preliminary investigation of disease-specific features in the social work notes, 10 disease chapters were identified with ICD-10 codes: (1) Diseases of the nervous system (G00-G99), (2) Diseases

of the circulatory system (I00-I99), (3) Diseases of the respiratory system (J00-J99), (4) Diseases of the digestive system (K00-K95), (5) Diseases of the musculoskeletal system and connective tissue (M00-M99), (6) Diseases of the genitourinary system (N00-N99), (7) Pregnancy, childbirth and the puerperium (O00-O9A), (8) Congenital malformations, deformations and chromosomal abnormalities (Q00-Q99) (9) Neoplasms (C00-D49), (10) Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (D50-D89). Chi-squared statistics was used to compare the frequency of words across different note categories (chi2 function from sklearn.feature selection was used to this end). After ranking the P values and removing stop words, the top five potential meaningful words were visualized by the word frequency calculation. Python package scikit-learn was used to conduct the analysis[43]. To embed and tokenize the unstructured notes, text.CountVectorizer function from sklearn.feature extraction package was used.

2.4 RESULTS

We retrieved a total of 0.95 million de-identified clinical social work notes generated between 2012 and 2021 (see Methods) from our UCSF Information Commons[32] (Figure 2.1). The majority of notes were classified as Progress Notes, Interdisciplinary Notes, or Telephone Encounter Notes; other note categories included Patient Instructions, Group Note, Letter, which comprised fewer than 5 percent each. These notes covered 181,644 patients of which 95387 (52.5%) were female. The median age of these patients was 33 years. Among them, 69,211 patients had only one note; 65,100 patients had between 2 and 5 notes, and 47,333 patients had more than 5 notes (S. Table 1, S. Figure 2.2B). The demographics distribution is presented in Table 1. No demographic feature was statistically associated with the number of notes for each patient (S. Table 1).

In addition to analyzing the number of notes, we were also interested in exploring the medical conditions associated with patients who received social work notes. This aspect can provide valuable insights into the

factors contributing to the need for social work intervention. To investigate this, we collected the ICD-10 codes for the encounters during which social work notes were recorded for the patients. These ICD-10 codes were then mapped at the chapter level[34]. The three most frequent ICD-10 chapters found to be associated with a social work note were "Mental, Behavioral and Neurodevelopmental disorders", "Factors influencing health status and contact with health services", and "Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified" (S. Table 2).

USING LDA TO EXTRACT TOPICS IN SOCIAL WORK NOTES

Looking at the word components of each topic (Table 1), we discovered a few diverse clusters that cover many different social aspects of patients including social service (Topic 11), abuse history (Topic 14), phone call/ online communications (Topic 12), living condition/ lifestyle (Topic 16), risk of death (Topic 8), group session (Topic 7), consultation/ appointment (Topic 5), family (Topic 4, 6), and mental health (Topic 1). Many of these topics are consistent with topics covering social determinants of health; most importantly, most of the information potentially conveyed through these topics are absent in the structured data. Of note, in our parameter exploration, we found that increasing the number of clusters can lead to additional recognizable topics, such as food availability (data not shown), although we also obtain redundant topics.

TOPIC MODELING ON SPECIFIC NOTE CATEGORIES

Analyzing the topics appearance in each note subtype, we found that social work notes in the Progress Notes category contained a higher percentage of clinically related topics, such as Mental Health (4.32%) and Clinician/Hospital/Medication-related information (8.40%), along with a smaller proportion of SDoH-related topics like Insurance/Income, Abuse history, Social support (10.46%), and Family (6.29%). Compared to Progress Notes, Telephone Encounter notes contained a larger proportion of topics related to Insurance/Income (3.93%), Phone call/Online (7.47%), Social support (11.56%), and Family (8.08%). Interestingly, telephone encounter notes lacked information about the Risk of death (0%), which may be

because the discussions on this topic are not appropriate for telephone encounters. Furthermore, Group Notes, which are the notes taken during group therapy, describe the group's progress and dynamics. As expected, Group Notes have a more uneven topic category distribution, with a higher percentage of Group session (24.69%) and Phone call/Online (12.71%) -related topics(Figure 2.2A).

We also applied LDA analysis to the social work notes associated with 10 ICD-10 chapters described earlier (Figure 2.2B). We observed that most diseases have a similar topic proportion distribution, for example, most of them are enriched for Social support and Family topics. In particular, Social support is highly represented in notes related to Neoplasms (21.51%) and Diseases of the digestive system (22.47%). Family topics are also frequently mentioned in notes associated with Diseases of the nervous system (23.31%), Pregnancy, childbirth, and the puerperium (20.1%), and Congenital malformations, deformations, and chromosomal abnormalities (21.43%). However, some differences were identified between the ICD-10 chapters. Notes associated with disorders of mental health and pregnancy contain a higher percentage of SDoH topics on mental health, as would be expected. Mental health topics are more frequently mentioned in clinical notes around pregnancy than even in nervous system disorders. Interestingly, the Family topic area was often mentioned in notes associated with congenital malformation abnormalities. In summary, the analysis demonstrated both the commonness and uniqueness of topics around social determinants of health covered across the various diseases and conditions which afflict patients.

WORD FREQUENCY ON INDIVIDUAL DISEASE

In addition to performing topic modeling on social work notes associated with 10 ICD-10 chapters, we also conducted a word frequency analysis. This analysis highlighted that note from each ICD-10 chapter contained both disease-specific terms and a limited number of disease-specific SDoH topics. For instance, notes from patients with neoplasms frequently mentioned terms like 'oncology', 'chemotherapy', and 'tumor', while those associated with musculoskeletal disorders often included words such as 'arthritis' and 'rheumatology'. In addition to these disease-specific words, there were observable patterns in the prevalence

of certain SDoH-related terms. Words like ‘mindfulness’ appeared predominantly in chapters on Pregnancy and the Nervous System, and ‘wheelchair’ was a recurrent term in Musculoskeletal disorders. Notably, conditions related to pregnancy showed a significant presence of mental health topics, indicating a frequent assessment of this aspect in social work notes for pregnancy care (Figure 2.3).

Overall, the word frequency analysis serves as a complementary tool to topic modeling. While topic modeling is adept at uncovering general patterns, predominantly SDoH topics, in social work notes, word frequency analysis, with its focused approach, tends to reveal features specific to particular diseases, especially when comparing different ICD-10 chapters.

2.5 DISCUSSION

We used an unsupervised topic modeling method called LDA modeling on our corpus of 0.95 million de-identified clinical social work notes. We showed that topic modeling can be used to (1) extract the hidden themes from this huge corpus of clinical notes and identify the critical information embedded in the notes, namely social determinants of health (SDoH) factors; and (2) calculate the proportion of each theme across different subsets of the note corpus and systemically characterize notes of different types. Using simple term frequency methods on this large corpus, we found that specific SDoH terms tend to be enriched in notes from patients within different disease categories, including wheelchair for patients with musculoskeletal disorders and depression for patients with pregnancy diagnoses, suggesting that these populations may be more at risk for these SDoH features.

We extracted several concrete SDoH-related topics, thus providing insight into the information that may be extracted from these corpora for facilitating future work around understanding how these topics correlate with health outcomes. During our comparison of notes of different subtypes, we found that the topic distribution of notes for specific types of diseases contains similar information but showed different levels

of enrichment, representing the unique features of each disease set. As one of many examples, our analysis shows how mental health issues are frequently documented around pregnancy (Figure 2.2B). This type of information can help us better understand the social determinants of most concern to patients when interacting with the health system.

The specific topics identified in our study were in line with findings from a previous publication [13]. This recent research extracted information on physical, mental, and social health by applying the non-negative matrix factorization (NMF) topic modeling method to 382,666 primary care clinical notes. However, that study exclusively examined physician-generated notes, whereas our focus was on social work notes, enabling us to uncover a broader range of SDoH topics. In our paper, we identified several additional topics, including but not limited to Living Condition/Lifestyle, Family, Risk of Death, and Abuse History.

Our research has several potential use cases. First, it aids computational sociology and epidemiology studies by identifying key factors that influence health outcomes. This extraction process lays the groundwork for in-depth analysis within these fields. Second, the findings from computational analyses can substantiate policy decisions. By providing empirical evidence, these findings can guide regulations and interventions aimed at health equity. Lastly, for participating healthcare providers, these extracted SDoH factors offer insights for effective resource allocation, particularly in supporting vulnerable groups. Overall, understanding the distribution of SDoH topics in patient records is crucial for developing targeted interventions and preventive strategies, aimed at addressing the root causes of health disparities.

Our study has several strengths. We performed analysis on a large corpus of notes, which to our knowledge is the largest social work notes data set to be used in a similar study. Instead of focusing on a single disease category or specific medical topic, we aimed at comprehensively finding the potential SDoH topics in all types of clinical social notes for a variety of diseases. Furthermore, to obtain a thorough understanding of the information embedded in social worker notes and capture the richness and complexity of the rhetoric in

these notes, we conducted complementary analyses: a word frequency enrichment analysis allowed us to identify specific terms more frequently associated with particular ICD-10 chapters, which demonstrated the prevalence of disease-related terms in social work notes, providing a more granular view of the data. Second, the use of Latent Dirichlet Allocation (LDA) allowed us to identify broader topics of increased relevance in these disease groups. It helped us uncover patterns related to social determinants of health, offering a higher-level perspective on the data.

Recognizing the intrinsic instability of LDA topic modeling methods, we enhanced the robustness of our results by independently searching for optimal hyperparameters to predefine topic numbers. Additionally, we ensured reliability by conducting each analysis across five iterations for every category (See methods). However, it is possible to still obtain different topic clusters with a different set of hyperparameters. Moreover, other topic modeling algorithms, such as NMF[13, 44] and BERTopic[45], could be explored to compare their performance and suitability for our specific task. In addition, we developed topic labeling heuristics that allow us to assign topics to the individual clusters. However, the heuristics may not cover all topic-related keywords, and in the future, it may be interesting to revisit our heuristic to expand upon the topic clusters further to make them more generalizable. State-of-the-art large language models like ChatGPT offer significant potential for improving our pipeline, particularly in the nuanced task of assigning topic labels[46-48]. With effective prompt engineering, these models could systematically extract patterns from social work notes, enhancing the depth and accuracy of our statistical analyses, and potentially uncovering new insights in social determinants of health. We also exclusively utilized ICD-10 codes, acknowledging the prospective merit of incorporating ICD-9 in future research. Another limitation of our study is the lack of structured EHR data for recording comorbidities, insurance, and living status. These factors are relevant to SDoH and could provide valuable insights into the relationships between health outcomes and social determinants. The absence of such data may limit our ability to fully capture the complex interplay of these factors and their effects on health. Finally, we did not explicitly exclude negations or the lengthy expression, as they still contribute to the overall discussion of certain topics.

However, we acknowledge that the consideration of negation is crucial for a more nuanced understanding of the information contained in clinical notes, and for more accurate analysis of the semantic meaning of the identified topics.

Our study opens pathways for several key areas of future research. For data scientists and computational researchers, future research should focus on combining these identified themes with predictive modeling techniques to assess their correlation with future health outcomes. This integration would not only validate the relevance of the identified SDoH themes but also provide a more holistic understanding of patient care dynamics and health outcomes. For healthcare practitioners, the challenge lies in integrating SDoH insights into patient care and public health policies. This demands not only an understanding of clinical informatics but also an insight into health policy and administration. Collaborating with experts in these fields could lead to developing actionable strategies that utilize our findings to improve healthcare delivery and policy decisions.

2.6 CONCLUSION

Social work notes contain rich and unique information about social determinants of health factors, frequently only recorded in text notes. SDoH factors are critical for analyzing health outcomes, and this study identified detailed categories of SDoH information covered by social work notes. Furthermore, the study demonstrated that different categories of notes emphasize different aspects of social determinants of health, despite belonging to social work consultations. The findings from this study would form a basis of potential future research questions around this utilizing SDoH to uncover health disparities and SDoH-associated disease trajectories, as well as methods to extract comprehensive SDoH-related information from clinical notes.

2.7 ACKNOWLEDGEMENTS

We thank all researchers, clinicians, and social workers who help collect clinical notes data. We thank everyone in Dr. Atul J. Butte's lab for helpful discussion and feedback. We thank staff members in the Bakar Computational Health Sciences Institute and UCSF IT Services who build and maintain the UCSF Information Commons. We thank the Wynton High-Performance Computing (HPC) cluster for making available the needed computation capacity.

2.8 DATA AVAILABILITY

The data that support the findings of this study are available from the Information Commons platform at UCSF, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of UCSF.

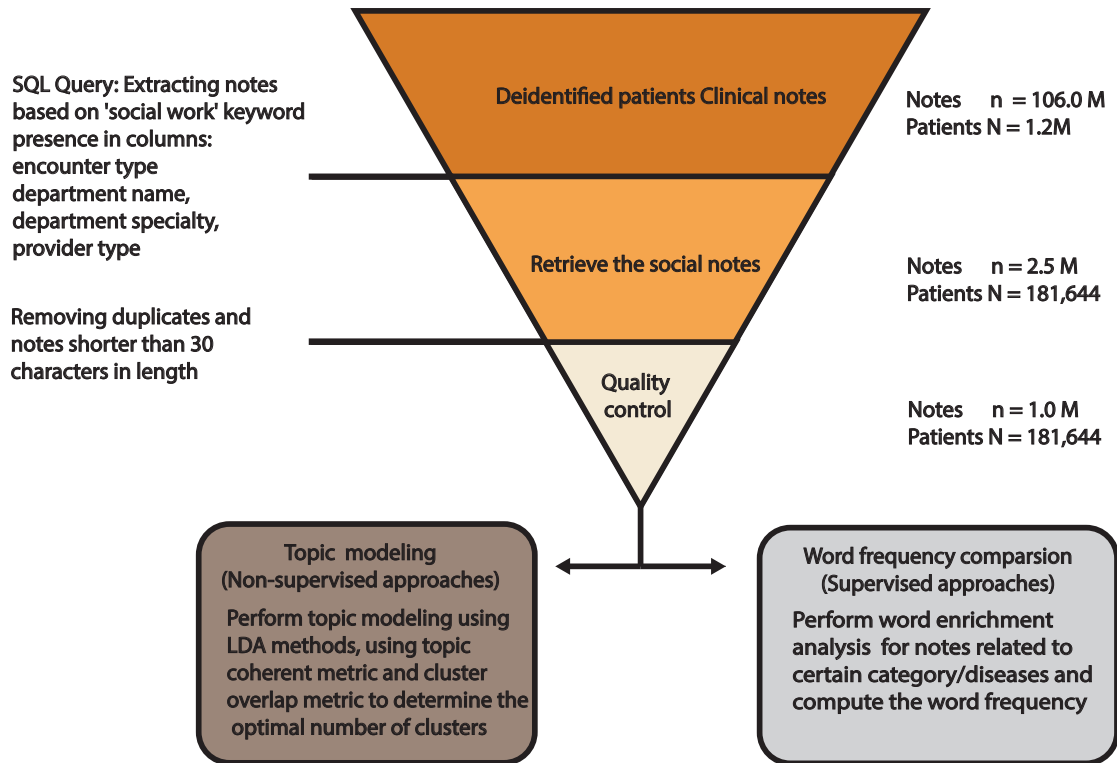
2.9 COMPETING INTERESTS STATEMENT

AJB is a co-founder and consultant to Personalis and NuMedii; consultant to Mango Tree Corporation, and in the recent past, Samsung, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, and Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Meta (Facebook), Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty Pharma, Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, Nuna Health, Assay Depot, and Vet24seven, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars,

Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical or disease specific foundations and associations, and health systems. AJB receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. AJB's research has been funded by NIH, Peraton (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity. The authors have declared that no competing interests exist.

2.10 FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors



Cohort selection and notes analysis workflow

Figure 2.1 Retrieval of clinical social work notes for the study

The social work notes from the UCSF Information Commons between 2012 and 2021 were initially retrieved. Notes that were duplicated or extremely short were excluded, which resulted in a corpus of 0.95 million notes. Later, the notes were analyzed using two methods: word frequency calculation (Bottom Left) and topic modeling (Bottom Right). Later, the word frequency was compared between different disease chapters. For topic modeling, Latent Dirichlet Allocation was used to identify the topics in individual social work notes. Topic coherence metric and Jaccard distance were implemented to decide the optimal clustering results.

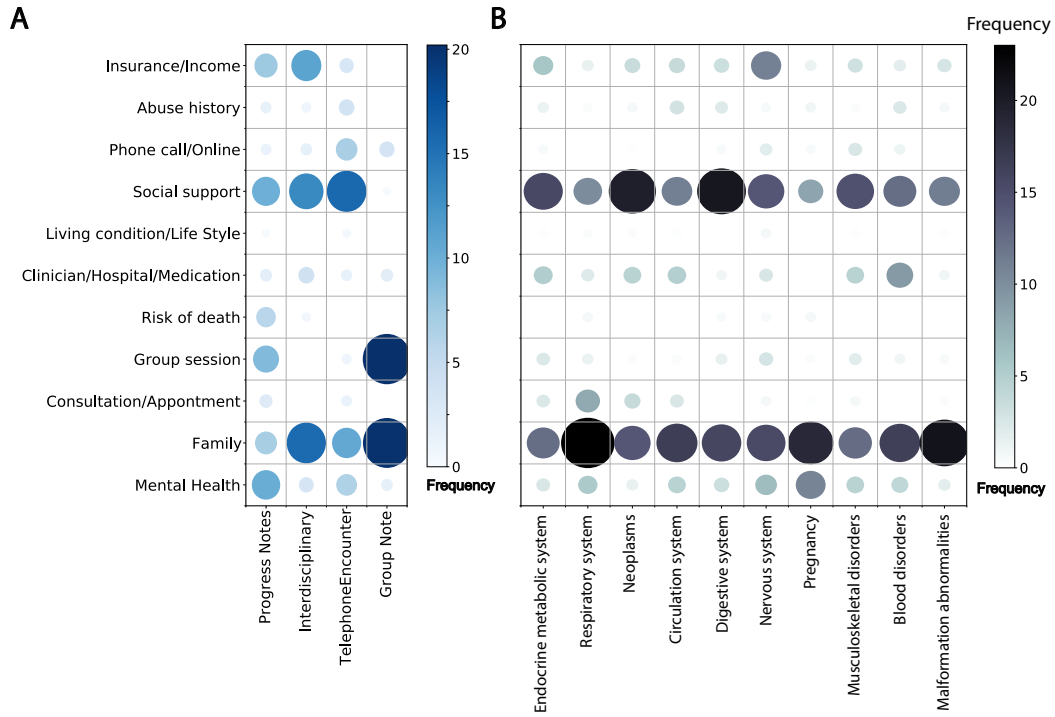


Figure 2.2 Topic proportion comparison for different categories

A. Topic proportion comparison for different note types. B. Topic proportion comparison for different disease chapters. Size and color of the circle represent proportion of each topic.

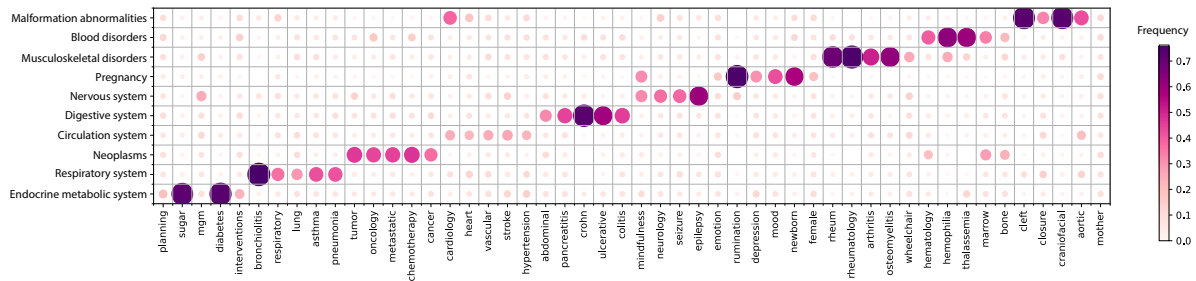


Figure 2.3 Word frequency calculation

Word frequency calculation for social work notes associated with each ICD-10 chapter. The proportion of the words in social work notes associated with each ICD-10 chapter is shown by the heatmap.

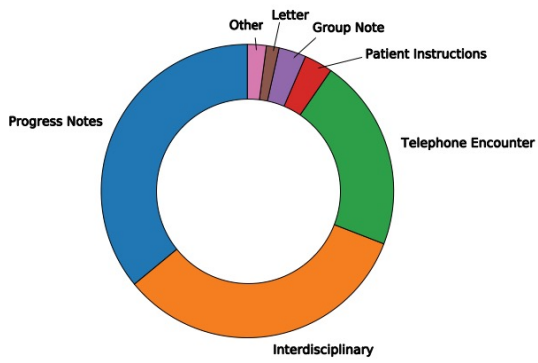
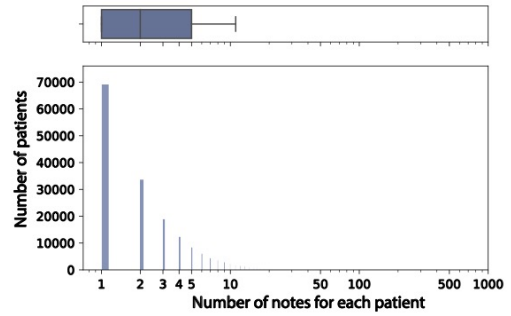
A**B**

Figure 2.4. Data exploration on social work notes

A. Pie chart showing the proportions of patients in different categories. B. Boxplot and histogram showing the number of notes for the individual patients. The scale of x-axis is log10-transformed. The mode, mean, and median are 1, 5.8, and 2.

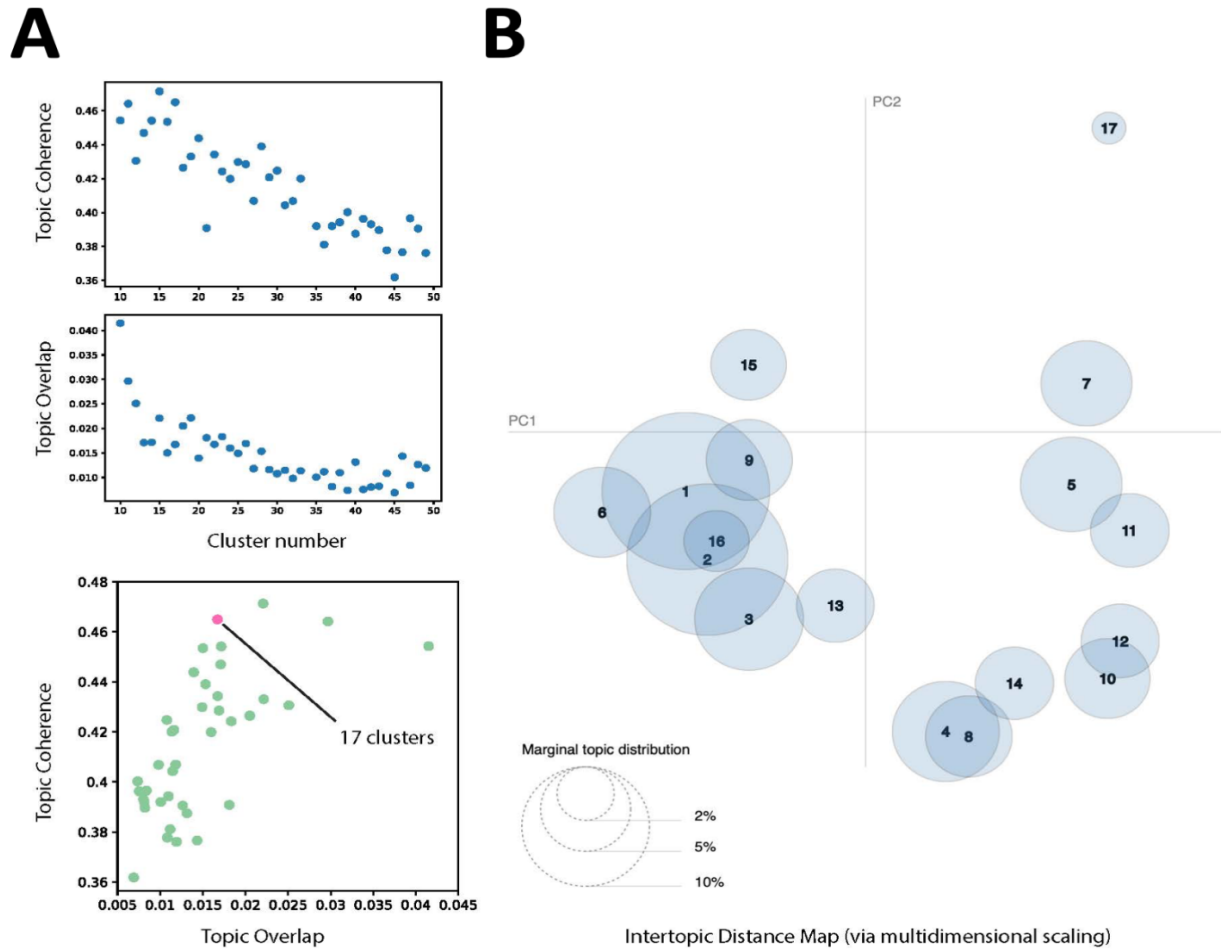


Figure 2.5 Topic modeling clustering on whole social work notes.

A. Pipeline for determining the optimal number of clusters for the LDA method. The top: the plot of cluster number versus the topic coherence metric; The middle: the plot of the cluster number versus the topic overlap metric (measured by jaccard similarity metric); The bottom: the plot of the topic overlap metric versus the topic coherence metric. The number of clusters is chosen as 17 because it has the lowest topic overlap metric value while having the highest topic coherence metric value (see **Methods**)

B. Inter-topic distance mapping for the individual cluster. Each circle represents an inferred topic. The coordinates for each circle correspond to the first two Principal components. The radius size indicates the frequency of topic existence on each note.

Table 2.1 Topic modeling results for all social work notes

Each row is an inferred topic, which is composed of 10 words.

| Clusters | Key Words |
|----------|---|
| 1 | goal, anxiety, problem, term, depression, mood, therapy, symptom, long, treatment |
| 2 | recommendation, wife, education, treatment, patient, form, appearance, ongoing, advocate, trauma |
| 3 | hospital, self, day, pain, other, connection, recent, feeling, side, number |
| 4 | mother, father, family, room, information, nurse, source, concrete, control, instruction |
| 5 | session, consultation, telehealth, location, time, tool, objective, parking, other, treatment |
| 6 | parent, family, school, child, sister, support, place, year, well, initial |
| 7 | group, intervention, patient, discussion, response, time, summary, progress, participant, skill |
| 8 | risk, chronic, thought, normal, imminent, status, testing, intervention, speech, suicide |
| 9 | client, health, service, caregiver, mental, therapist, therapy, behavioral, individual, group |
| 10 | well, when, time, week, also, able, state, more, friend, very |
| 11 | social, service, support, family, assessment, medical, time, note, concern, ongoing |
| 12 | care, home, plan, phone, contact, work, information, resource, call, support |
| 13 | time, clinician, name, date, code, behavior, risk, number, plan, provider |
| 14 | history, child, other, factor, current, none, substance, abuse, psychiatric, year |
| 15 | donor, donation, potential, employment, understanding, risk, decision, independent, process, care |
| 16 | night, morning, hour, sleep, house, already, less, past, aggressive, evening |
| 17 | transplant, medication, post, support, health, insurance, husband, psychosocial, message, history |

Table 2.2 Keywords for topic assignment

The words in the Keywords column are the representative words used to define the topics.

| Topics | Key Words |
|--|--|
| Mental health | mental, depression, anxiety, mood, psychological, physical, cognitive, emotional, mind, psychiatric |
| Family | family, parent, father, mother, child, children, sister, parents, relatives, clan, childhood, friends |
| Consultation/Appointment | appointment, consultation, consult, questionnaire, question, advice, biographical, Wikipedia, relevant, questions, know, documentation |
| Group session | group, intervention, session, interprets, community, class, organization, together, part, organization |
| Risk of death | suicide, suicidal, risk, crisis, homicide, murder, commit, bombing, murdered, murders, bomber, killing, convicted, victims |
| Clinician/Hospital/Medication | patient, medication, hospital, medical, clinic, clinician, treatment, therapy, surgery, symptoms, patients, drugs, diagnosis, treatments, prescribed |
| Living condition/Lifestyle | shelter, housing, house, living, sleep, bedtime, building, buildings, urban, employment, suburban, campus, acres |
| Social support | social, service, support, referral, recommendation, recommend, worker, resource, supports, provide, supporting, supported, allow, providing, assistance, benefit, help |
| Telephone/Encounter/Online communication | telehealth, phone, call, video, telephone, mobile, wireless, msg, cellular, dial, email, calling, networks, calls, messages, telephones, internet |
| Abuse history | abuse, history, addiction, alcohol, drugs, allegations, victim, violence, sexual, rape, dependence |
| Insurance/Income | insurance, income, coverage, financial, contracts, banking, finance, liability, private, pay |

Table 2.3 Descriptive statistics for clinical social work notes corpus and contributing patient samples

Few Notes: Number of notes ≤ 1 ; Several Notes: $2 \leq$ Number of notes < 5 ; Many Notes: Number of notes ≥ 5 .

| | Few Notes (N=69211) | Several Notes (N=65100) | Many Notes (N=47333) | Overall (N=181644) |
|------------------------|--------------------------------|--|-------------------------------------|-------------------------------|
| Sex | | | | |
| Female | 36372 (52.6%) | 34285 (52.7%) | 24730 (52.2%) | 95387 (52.5%) |
| Male | 32608 (47.1%) | 30626 (47.0%) | 22401 (47.3%) | 85635 (47.1%) |
| Unknown | 231 (0.3%) | 189 (0.3%) | 202 (0.4%) | 622 (0.3%) |
| Ethnicity | | | | |
| Hispanic/Latino | 14891 (21.5%) | 14451 (22.2%) | 12044 (25.4%) | 41386 (22.8%) |
| Not Hispanic/Latino | 48758 (70.4%) | 46011 (70.7%) | 33249 (70.2%) | 128018 (70.5%) |
| Unknown | 5562 (8.0%) | 4638 (7.1%) | 2040 (4.3%) | 12240 (6.7%) |
| Race | | | | |
| Asian | 8651 (12.5%) | 8578 (13.2%) | 5610 (11.9%) | 22839 (12.6%) |
| Black/African | 7153 (10.3%) | 7148 (11.0%) | 6819 (14.4%) | 21120 (11.6%) |
| Other | 17594 (25.4%) | 16922 (26.0%) | 13207 (27.9%) | 47723 (26.3%) |
| Unknown | 6683 (9.7%) | 5493 (8.4%) | 2637 (5.6%) | 14813 (8.2%) |
| White | 29130 (42.1%) | 26959 (41.4%) | 19060 (40.3%) | 75149 (41.4%) |
| Age | | | | |
| median (Q1- Q3) | 32 (11 - 58) | 35 (13 - 59) | 30 (10 - 57) | 33 (12 - 58) |
| Missing | 111 (0.2%) | 135 (0.2%) | 175 (0.4%) | 421 (0.2%) |

Table 2.4 Most frequent ICD-10 codes for patients with social work notes

| ICD10 code | Diagnosis | Notes | Patients |
|------------|---|--------|----------|
| F01-F99 | Mental, Behavioral and Neurodevelopmental disorders | 106348 | 15388 |
| Z00-Z99 | Factors influencing health status and contact with health services | 36399 | 25995 |
| R00-R99 | Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified | 25011 | 18820 |
| E00-E89 | Endocrine, nutritional and metabolic diseases | 13307 | 8488 |
| S00-T88 | Injury, poisoning and certain other consequences of external causes | 10508 | 5092 |
| I00-I99 | Diseases of the circulatory system | 9832 | 8450 |
| K00-K95 | Diseases of the digestive system | 7711 | 6267 |
| G00-G99 | Diseases of the nervous system | 7589 | 6187 |
| O00-O9A | Pregnancy, childbirth and the puerperium | 7370 | 5628 |
| D50-D89 | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism | 6289 | 4598 |
| Q00-Q99 | Congenital malformations, deformations and chromosomal abnormalities | 6041 | 4507 |
| C00-D49 | Neoplasms | 6025 | 4515 |
| N00-N99 | Diseases of the genitourinary system | 5570 | 4926 |
| J00-J99 | Diseases of the respiratory system | 5294 | 4690 |
| M00-M99 | Diseases of the musculoskeletal system and connective tissue | 5018 | 4228 |
| P00-P96 | Certain conditions originating in the perinatal period | 4700 | 4426 |
| A00-B99 | Certain infectious and parasitic diseases | 2867 | 2536 |
| V00-Y99 | External causes of morbidity | 2104 | 2053 |
| L00-L99 | Diseases of the skin and subcutaneous tissue | 2023 | 1853 |
| H00-H59 | Diseases of the eye and adnexa | 889 | 841 |
| H60-H95 | Diseases of the ear and mastoid process | 562 | 516 |
| U00-U85 | Codes for special purposes | 85 | 84 |

2.11 REFERENCES

- [1] Marmot M. Social determinants of health inequalities. *The lancet* 2005;365:1099–104.
- [2] Organization WH. Social determinants of health. WHO Regional Office for South-East Asia 2008.
- [3] World Health Organization. A Conceptual Framework for Action on the Social Determinants of Health. Social Determinants of Health Discussion Paper 2, 2010
- [4] Sun S, Zack T, Williams CY, Butte AJ, Sushil M. Revealing the impact of social circumstances on the selection of cancer therapy through natural language processing of social work notes. arXiv preprint arXiv:2306.09877. 2023 Jun 16.
- [5] Hill-Briggs F, Adler NE, Berkowitz SA, et al. Social determinants of health and diabetes: a scientific review. *Diabetes care* 2021;44:258–79.
- [6] White-Williams C, Rossi LP, Bittner VA, et al. Addressing social determinants of health in the care of patients with heart failure: a scientific statement from the American Heart Association. *Circulation* 2020;141:e841–63.
- [7] Marmot M, Friel S, Bell R, Houweling TA, Taylor S. Closing the gap in a generation: health equity through action on the social determinants of health. *The lancet*. 2008 Nov 8;372(9650):1661-9.
- [8] Federenko IS, Wadhwa PD. Women's mental health during pregnancy influences fetal and infant developmental and health outcomes. *CNS spectrums*. 2004 Mar;9(3):198-206.
- [9] Hill-Briggs F, Adler NE, Berkowitz SA, Chin MH, Gary-Webb TL, Navas-Acien A, Thornton PL, Haire-Joshu D. Social determinants of health and diabetes: a scientific review. *Diabetes care*. 2021 Jan;44(1):258.
- [10] Coffey PM, Ralph AP, Krause VL. The role of social determinants of health in the risk and prevention of group A streptococcal infection, acute rheumatic fever and rheumatic heart disease: a systematic review. *PLoS neglected tropical diseases* 2018;12:e0006577.
- [11] Calixto O-J, Anaya J-M. Socioeconomic status. The relationship with health and autoimmune diseases. *Autoimmunity reviews* 2014;13:641–54.

- [12] Singu S, Acharya A, Challagundla K, et al. Impact of social determinants of health on the emerging COVID-19 pandemic in the United States. *Frontiers in public health* 2020;:406.
- [13] Meaney C, Escobar M, Moineddin R, Stukel TA, Kalia S, Aliarzadeh B, Chen T, O'Neill B, Greiver M. Non-negative matrix factorization temporal topic models and clinical text data identify COVID-19 pandemic effects on primary healthcare and community health in Toronto, Canada. *Journal of Biomedical Informatics*. 2022 Apr 1;128:104034
- [14] Hobensack M, Song J, Oh S, Evans L, Davoudi A, Bowles KH, McDonald MV, Barrón Y, Sridharan S, Wallace AS, Topaz M. Social Risk Factors are Associated with Risk for Hospitalization in Home Health Care: A Natural Language Processing Study. *J Am Med Dir Assoc*. 2023 Aug 5:S1525-8610(23)00621-7. doi: 10.1016/j.jamda.2023.06.031. Epub ahead of print. PMID: 37553081.
- [15] Keyhani S, Mowery DL, Chapman BE, Conway M, South BR, Madden E, Chapman WW. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: an information content analysis.
- [16] Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc* 2021;28:2716–27. doi:10.1093/jamia/ocab170
- [17] Pyo S, Kim E, Kim M. LDA-Based Unified Topic Modeling for Similar TV User Grouping and TV Program Recommendation. *IEEE Trans Cybern* 2015;45:1476–90. doi:10.1109/TCYB.2014.2353577
- [18] Min K-B, Song S-H, Min J-Y. Topic Modeling of Social Networking Service Data on Occupational Accidents in Korea: Latent Dirichlet Allocation Analysis. *J Med Internet Res* 2020;22:e19222. doi:10.2196/19222
- [19] Rosenbloom ST, Denny JC, Xu H, et al. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association* 2011;18:181–6.
- [20] Conway M, Keyhani S, Christensen L, et al. Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. *Journal of biomedical semantics* 2019;10:1–10.

- [21] Lituiev D, Lacar B, Pak S, et al. Automatic Extraction of Social Determinants of Health from Medical Notes of Chronic Lower Back Pain Patients. 2022;:2022.03.04.22271541. doi:10.1101/2022.03.04.22271541
- [22] Chen ES, Carter EW, Sarkar IN, et al. Examining the use, contents, and quality of free-text tobacco use documentation in the electronic health record. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association 2014. 366.
- [23] Bejan CA, Angiolillo J, Conway D, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *Journal of the American Medical Informatics Association* 2018;25:61–71.
- [24] Alghamdi R, Alfalqi K. A survey of topic modeling in text mining. *Int J Adv Comput Sci Appl(IJACSA)* 2015;6.
- [25] Hofmann T. Probabilistic Latent Semantic Analysis. 2013. doi:10.48550/arXiv.1301.6705
- [26] Deerwester S, Dumais ST, Furnas GW, et al. Indexing by latent semantic analysis. *Journal of the American society for information science* 1990;41:391–407.
- [27] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *the Journal of machine Learning research* 2003;3:993–1022.
- [28] Hong L, Davison BD. Empirical study of topic modeling in twitter. In: *Proceedings of the first workshop on social media analytics*. 2010. 80–8.
- [29] Girdhar Y, Giguere P, Dudek G. Autonomous adaptive underwater exploration using online topic modeling. In: *Experimental Robotics*. Springer 2013. 789–802.
- [30] Chen B, Zhu L, Kifer D, et al. What is an opinion about? exploring political standpoints using opinion scoring model. In: *Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010.
- [31] Wang M, Pantell MS, Gottlieb LM, et al. Documentation and review of social determinants of health data in the EHR: measures and associated insights. *Journal of the American Medical Informatics Association* 2021;28:2608–16. doi:10.1093/jamia/ocab194

- [32] University of California, San Francisco, ARS. UCSF DeID CDW. R20220207. UCSF Data Resources. <https://data.ucsf.edu/research> (accessed 12 Sep 2022).
- [33] Řeh\uu00f1ek R, Sojka P. Gensim—statistical semantics in python. Retrieved from genism.org 2011.
- [34] Hoffman M, Bach F, Blei D. Online learning for latent dirichlet allocation. *advances in neural information processing systems* 2010;23.
- [35] Bird S. NLTK: the natural language toolkit. In: *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. 2006. 69–72.
- [36] Rieger J, Rahnenführer J, Jentsch C. Improving Latent Dirichlet Allocation: On Reliability of the Novel Method LDAPrototype. In: *International Conference on Applications of Natural Language to Information Systems*. Springer 2020. 118–25.
- [37] Rosner F, Hinneburg A, Röder M, et al. Evaluating topic coherence measures. *arXiv preprint arXiv:14036397* 2014.
- [38] Syed S, Spruit M. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2017. 165–74. doi:10.1109/DSAA.2017.61
- [39] Jaccard P. The distribution of the flora in the alpine zone. 1. *New phytologist* 1912;11:37–50.
- [40] Organization WH. *International Statistical Classification of Diseases and related health problems: Alphabetical index*. World Health Organization 2004.
- [41] Dieng AB, Ruiz FJ, Blei DM. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*. 2020 Jul 1;8:439-53.
- [42] Picture DS Nikhil Thorat, Charles Nicholson, Big. Embedding projector - visualization of high-dimensional data. <http://projector.tensorflow.org> (accessed 10 Nov 2022).
- [43] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 2011;12:2825–30.
- [44] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999 Oct 21;401(6755):788-91.

- [45] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. 2022 Mar 11.
- [46] Rijcken E, Scheepers F, Zervanou K, Spruit M, Mosteiro P, Kaymak U. Towards Interpreting Topic Models with ChatGPT. In The 20th World Congress of the International Fuzzy Systems Association 2023.
- [47] OpenAI. ChatGPT: Optimizing language models for dialogue [Internet]. OpenAI; 2022 [cited [Your Access Date]]. Available from: <https://openai.com/blog/chatgpt>
- [48] OpenAI. GPT-4 technical report [Internet]. arXiv; 2023 [cited [Your Access Date]]. Available from: <https://arxiv.org/abs/2303.08774>

3 Chapter 3: Revealing the impact of social circumstances on the selection of cancer therapy through natural language processing of social work notes

Shenghuan Sun¹, Travis Zack^{1,4}, Christopher Y.K. Williams¹, Atul J. Butte^{1,2,3}, Madhumita Sushil¹

1. Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA

2. Center for Data-driven Insights and Innovation, University of California, Office of the President, Oakland, CA, USA

3. Department of Pediatrics, University of California, San Francisco, CA, 94158, USA

4. Division of Hematology/Oncology, Department of Medicine, University of California, San Francisco, San Francisco, California, USA.

3.1 ABSTRACT

OBJECTIVE

We aimed to investigate the impact of social circumstances on cancer therapy selection using natural language processing to derive insights from social worker documentation.

MATERIALS AND METHODS

We developed and employed a Bidirectional Encoder Representations from Transformers (BERT) based approach, using a hierarchical multi-step BERT model (BERT-MS), to predict the prescription of targeted cancer therapy to patients based solely on documentation by clinical social workers. Our corpus included free-text clinical social work notes, combined with medication prescription information, for all patients treated for breast cancer at UCSF between 2012 and 2021. We conducted a feature importance analysis to identify the specific social circumstances that impact cancer therapy regimen.

RESULTS

Using only social work notes, we consistently predicted the administration of targeted therapies, suggesting systematic differences in treatment selection exist due to non-clinical factors. The findings were confirmed by several language models, with GatorTron achieving the best performance with an AUROC of 0.721 and a Macro F1 score of 0.616. The UCSF BERT-MS model, capable of leveraging multiple pieces of notes, surpassed the UCSF-BERT model in both AUROC and Macro-F1. Our feature importance analysis identified several clinically intuitive social determinants of health (SDOH) that potentially contribute to disparities in treatment.

DISCUSSION

Leveraging social work notes can be instrumental in identifying disparities in clinical decision-making. Hypotheses generated in an automated way could be used to guide patient-specific quality improvement interventions. Further validation with diverse clinical outcomes and prospective studies is essential.

CONCLUSIONS

Our findings indicate that significant disparities exist among breast cancer patients receiving different types of therapies based on social determinants of health. Social work reports play a crucial role in understanding these disparities in clinical decision-making.

3.2 INTRODUCTION

Clinical decisions biased by social disparities lead to significant discrepancies in outcome and pose significant public health concerns [1–3]. Clinical decisions are influenced not only by clinical criteria but also by non-clinical factors such as race, gender, perceived financial stability, and more, which are collectively referred to as social determinants of health (SDOH) [4–6]. There is growing evidence that many minority groups are less likely to receive standard of care [6,9,10]. One pressing example is the decision to initiate anti-neoplastic treatments, which are becoming increasingly expensive and associated with financial toxicities [7]. While new, targeted agents often are better tolerated and more effective than previous treatments, they can come with a high price tag not always fully covered by insurance, leaving clinicians with a moral decision when balancing efficacy and cost. Financial constraints are but one example of factors that can potentially influence the treatment decision [8].

In this work, we demonstrated a strong association between specific features within social work (SW) clinical documentation and the choice of expensive, targeted therapy prescription for patients with breast cancer. Using a pretrained Bidirectional Encoder Representations from Transformers (BERT) model, we

showed that the unstructured SW notes, without detailed diagnostic or therapeutic information, can predict whether targeted therapy was prescribed for a given patient. Moreover, we developed a hierarchical language model for prediction over long sequences of clinical notes and successfully increased the predictability of the outcome. To understand which SDOH factors are used by the model for prediction, we measured the importance of SDOH factors by deleting words belonging to specific SDOH topics. Several critical contributors emerged, including socio-economic factors, abuse history, and risk of death. Our findings demonstrate that SW notes can reveal the impact of a patient's social environment on medical treatment prescription without requiring expensive and time-consuming manual annotation. Our hierarchical modeling approach will inform the development of models capable of leveraging multiple clinical notes for prediction.

BACKGROUND AND SIGNIFICANCE

A growing body of evidence indicates that SDOH factors significantly impact patient health and behaviors[5,6,11,12]. However, SDOH factors not only affect patients but also influence the clinical decision-making process recommended by physicians[4]. Ideally, clinical decision-making should be rooted in evidence-based practices, cognizant of the complex interaction between a patient's background and SDOH that could affect both their trust in the medical system and their overall disease trajectory. In reality, though, physicians are inevitably influenced by a wide range of non-clinical factors, with many of these non-clinical factors rooted in unconscious bias[13,14]. Previous research showed that clinical management decisions can be influenced by socioeconomic status[8], race[15], gender[16], adherence to treatment[17], patient behavior[18], attitude[19], and even physician personal characteristics[20].

Although it is well-known that SDOH-related, non-medical factors are crucial contributors to health and clinical outcomes, extracting non-medical and social factors from electronic medical records remains challenging. While information such as smoking, alcohol, and primary insurance status is increasingly accessible in structured fields, many social factors that are increasingly recognized as being important to

successful treatment are either not captured or are not a focus of structured physician documentation. Various aspects that physicians consider, including patient personalities, preferences, faith, concerns, professional interactions, family support, and living situations, can often be missing or improperly addressed within physician notes [4]. Due to this, our capacity to understand the relationship between these critical aspects of SDOH is constrained by the data we choose to focus on, as well as the accessibility of the information within.

Compared to general clinical documentation, notes written by social workers (SW notes) contain comprehensive social information[21,22]. Social workers are professionals who specialize in navigating a patient through the barriers that may interfere with receiving adequate medical care[23,24]. They can evaluate the many aspects of patients' life outside of medicine that can impact their ability to receive treatment. These include insurance concerns, financial concerns, social and daily living support, and ancillary support such as transportation, mental health, and housing. Because of this focus on the non-medical barriers that may affect medical care, SW notes could be invaluable in understanding the non-medical factors that influence medical decision-making.

Demonstrating that social work notes, considered in isolation, can be predictive of complex clinical decisions would highlight the power that can be derived from understanding how SDOH affects clinical decision-making. Doing this requires the development of new methods in natural language processing (NLP) to transform the nuances within SW documentation of complex social topics into predictive features around the clinical decision-making for costly drugs.

3.3 MATERIALS AND METHODS

STUDY DESIGN AND COHORT SELECTION

This study used a deidentified clinical note corpus at UCSF available within the UCSF Information Commons. The research was conducted under the IRB #18-25163. Our corpus included the deidentified social work notes of all patients treated at UCSF for breast cancer between 2012 and 2021 (Figure 3.1). Breast cancer diagnosis was identified using the ICD9 code 174 and the ICD10 code C50 through the UCSF Clinical Data Warehouse. We obtained 2496 patients matching these codes, with available social work reports. We then retrieved the medications ordered or prescribed for these patients, then categorized these as “targeted therapy” medications or not based on the definitions in the Targeted Cancer Therapies Fact Sheet from National Cancer Institute[25]. Patients in the cohort who received targeted therapy at least once were categorized into the 'Targeted therapy administered' group (TT-Yes); patients who did not receive any targeted therapy were categorized into the 'Targeted therapy not administered' group (TT-No). Though we are working with social work notes, we still found that drug information was mentioned in less than 10% of the overall social work notes. To prevent information leakage, we masked the drug information prior to any further processing. Specifically, in expressions like "Tamoxifen was administered to the patient", we replaced the drug name, here Tamoxifen, with the word "drug". The full list of drug names we masked are included as S. Table 12.

DEEP LEARNING MODELS FOR SENTENCE CLASSIFICATION

We used the latest and the longest social work note per patient to predict cancer therapy selection. Patient notes were randomly split in an 8:1:1 ratio into training, validation and test sets. We trained our algorithm on the training set, using the early stopping approach to help with parameter tuning on the validation set. We ran our algorithm 5 times for each model and evaluated the model performance using the validation set. The cross-entropy loss function was used for optimization. After training and hyperparameter tuning, the

model was tested on the held-out test set to compute model performance. Median scores over 5 runs are reported here.

We compared several biomedical BERT models in this research, including: GatorTron-OG [41], a Megatron BERT model pre-trained on de-identified clinical notes from the University of Florida, the UCSF-BERT model[26], which is a cased BERT model pretrained on the UCSF clinical notes publicly , SciBERT[28], ClinicalBERT[29], BioLM[30], and Biomed-Roberta[35]. All of these models have been pre-trained on a large corpus of scientific texts, PubMed, PMC, and/or clinical notes from the MIMIC-III corpus[36]. We fine-tuned each of these models for the classification task.

To rule out the possibility of finding results at random, we implemented three distinct dummy classifiers as a control. Dummy (Prior): This strategy always predicts the most frequent class in the training set. Dummy (Stratified): This strategy generates predictions by respecting the class distribution of the training set. It randomly predicts class labels based on the distribution of the training set. Dummy (Uniform): This strategy generates predictions uniformly at random.

EVALUATION METRICS

Model evaluation results were reported for the testing dataset only. For the classification task, Area Under the Receiver Operating Characteristic curve (AUROC), F1 score, precision, and recall metrics are reported. In order to address the issue of data imbalance, which can impede the interpretation of model performance, we used macro-averaged format for F1, precision, and recall score. F1 score is the harmonic mean of precision and recall.

Notably, macro-averaged computation uses the arithmetic mean of all the per-class scores, which provides equal weight to all the classes. We used `sklearn.metrics` from the `scikit-learn` python package for programming[32].

CONSTRUCTING THE BERT-MS MODEL

Although most patients in our dataset have several relevant SW notes (median = 11, Figure 3.2B), the BERT models used for classification are unable to accept more than a maximum of 512 tokens, which cannot handle more than one social work note piece. We were interested in knowing whether integrating more notes and thus more information about a patient's social history would improve the prediction. However, retraining a language model with an input length several times longer would take considerable time and computation resources and is impractical in an academic environment[27]. Consequently, we developed a multistep, hierarchical BERT model that can integrate several notes named MS-n, where n refers to the maximum number of notes allowed by the model (Figure 3.3).

The MS-n model was trained in two steps (Figure 3.3, and Supplementary Algorithm 1). First, all clinical notes for a single patient were treated as independent instances for phase 1 fine-tuning. Each note for a single patient was assigned the same binary patient-level label indicating whether targeted therapy was administered to the patient. The BERT model was fine-tuned in this setup and the validation loss was computed for backpropagation. Consequently, in the second phase, intermediate note-level representations were extracted from the resulting model of phase 1 finetuning and concatenated for phase 2 finetuning. The phase 2 BERT model was initialized with these concatenated note-level representations, the intermediate layer weights were frozen, and the classification layer of the model was fine-tuned further. Phase 1 fine-tuning was critical because it could extract the lower-dimensional hidden representations of each note. In this manner, we were able to train a hierarchical language model that can integrate n-fold information without expending the model parameter n-folds. We built several UCSF BERT-MS-n models including MS-3, MS-5, MS-8, MS-10, that correspond to the use of at most 3, 5, 8 and 10 notes.

FEATURE IMPORTANCE ANALYSIS

To understand which SDOH factors are used by the model for prediction, we used feature ablation methods to measure the importance of different SDOH factors. We examined the effect on model performance of

removing keywords associated with the following topics: Mental health, Family, Consultation/Appointment, Group session, Risk of death, Clinician/Hospital/Medication, Living condition/Lifestyle/Social support, Telephone encounter/Online communication, Abuse history (all forms), and Insurance/Income. These categories, and keywords associated with each category, were selected following the LDA topic modeling analysis as described by Sun et al [33] (Supplementary Table 5). Specifically, we removed a set of words belonging to each SDOH topic iteratively from the test set only and compared the decrease in model performance represented by the decrease in F1 score. We conducted these experiments on MS-5 model which has the best predictive performance.

To account for differences in the prevalence of various topics mentioned across patients (e.g 96% of notes contained keywords in the ‘Social support’ topic whereas only 10% of notes relate to the ‘Risk of death’ category), we normalize the importance of each topic by their frequency. We present both the raw feature important score and the important score normalized by topic frequency in Supplementary Figure 3.2.

3.4 RESULTS

PATIENTS STRUCTURED CHARACTERISTICS AND THEIR SOCIAL WORK NOTES

We identified 2496 patients with breast cancer with available deidentified social work notes (Figure 3.1); 97.9% of patients were female and 2.1% were male. There were 59.7% White/Caucasian patients, 18.1% Asian, 10.1% Hispanic/Latino, 6.5% Black/African, and 15.7% Other (S. Table 1). No obvious difference was observed when comparing the demographic information between patient with and without social work notes except an increase proportion of Asian population (S. Table 9). Overall, 70% of patients in the cohort received targeted therapy at least once [‘Targeted therapy administered’ group] (TT-Yes), compared to 30% of patients who did not receive any targeted therapy [‘Targeted therapy not administered’ group] (TT-No).

First, we explored whether SDOH information within structured data alone could stratify these patients. For the 2496 patients identified, we found information regarding demographics, marriage status, and smoking history was present, but data on patient financial status, education level, and other important SDOH were absent from the structured data. Machine learning-based approaches leveraging all available demographic information, marriage status and smoking history failed to predict the administration of targeted therapy in patients (S. Table 3), which is not surprising given the sparsity of the available data as well as the difficulty of the task.

In contrast, our prior research has demonstrated that social work notes possess a wealth of information relating to SDOH, including details on frequently discussed topics such as mental health, insurance status, and family support (Figure 3.2C, D)[33]. This qualitative observation suggested that social work notes encompass a wealth of SDOH factors, which may be captured by pre-trained language models when predicting the administration of therapy regimens to patients.

GATORTRON-OG OUTPERFORMS OTHER LANGUAGE MODELS IN PREDICTING THERAPY

We fine-tuned several pretrained biomedical BERT models to predict the targeted therapy administration directly from the social work notes of breast cancer patients [26-31, 41]. Given that the maximum sequence length supported by a regular BERT model is 512 tokens, we used the longest note for each patient to maximize the amount of information available for classification. Table 1 shows the prediction performance of different deep-learning classification models. To further ensure that related clues or other explicit medical information were not present in these notes, we additionally quantified model performance on a subset of notes that do not mention any drugs. This approach achieved similar performance, demonstrating the reliability of masking the drug names (Supplementary Table 7). GatorTron-OG achieved the best result with a Macro F1 of 0.616 and AUROC score of 0.721. UCSF-BERT also held good classification performance with a Macro F1 of 0.599 and AUROC score of 0.675, although it did not outperform the GatorTron-OG model. This can be attributed to the fact that GatorTron model is larger in size and is trained

on a larger cohort of clinical data. RoBERTa models (BioLM and Biomed-Roberta) performed generally better than BERT-base models (SciBERT, ClinicalBERT) potentially because of their dynamic masking strategy during pretraining such that the masked token changes during each training epoch[31]. This suggests that pretraining BERT-based models with clinical data can be helpful for achieving superior performance on domain-specific tasks. We also ran our tasks on three random baseline models, each of which ruled out the random performance from different perspectives (See methods). Our model significantly outperformed the random baselines (Table 1).

INTEGRATING MULTIPLE CLINICAL NOTES FOR PREDICTION

Given that the median number of clinical social notes per patient in our cohort is 11, we built several multi-step (MS-n) models including MS-3, MS-5, MS-8, MS-10, allowing the analysis of up to 3, 5, 8, and 10 notes respectively. We used UCSF-BERT for this because it is smaller in size, and hence has lower training complexity than the GatorTron-OG model, while having comparable performance. Table 2 compares the prediction performance of UCSF BERT_MS-n models with the UCSF BERT model using a single social work note. Generally, the UCSF BERT_MS-n models achieved better results, demonstrating the advantage of incorporating more clinical notes.

IDENTIFYING THE SDOH FACTORS THAT INFLUENCE MODEL DECISIONS

To explore the role different SDOH factors may have in predicting utilization of targeted therapy, we assessed the importance of SDOH factors by feature ablation methods (See Methods). The 11 topics that we tested were mentioned with varying frequency in the social work notes (Supplementary Figure 3.1). The notes belonging in each topic have similar class proportions: 70% “TT-Yes” group and 30% “TT-No” group. Of note, simple machine learning frameworks leveraging the presence of SDOH topics as binary features were not sufficient to predict the administration of targeted therapy (Supplementary Table 8).

We identified several SDOH topics, including Abuse History, Risk of Death, and Social Support, as the most significant influencers that the model leveraged in the prediction task (Figure 3.4). Other SDOH topics such as Family, Living Condition also had obvious impact in model decision making. However, besides the broad topic area “medical factors”, the common topics relating to medical aspects, Mental Health and Group Session, had a lower influence on the model prediction. As the neutral control, topic Consultation and TelephoneEncounter played a less important role in the prediction task. Interestingly, Finance, which represents the socioeconomic factor that likely influences patients’ decisions in therapy regimen, did not come up as an important regulator in the process. Overall, we successfully used model interpretability methods to analyze the trained language model to discover the SDOH factors that are not frequently considered to be influencers of the prescription of more financially toxic oncology medications.

3.5 DISCUSSION

This study demonstrated that clinical social work documentation, which focuses on social determinants of health rather than treatment plans, can be predictive of whether targeted therapies are administered to patients with breast cancer and highlights a potential SDOH-dependent disparity in therapy administration. Additionally, we developed a hierarchical modeling technique to incorporate the large volume of note data within any given chart, which often exceeds the processing capacity of the state-of-the-art NLP models. This technique can leverage multiple notes for prediction without adding a significant amount of computation burden. Finally, we performed a feature importance analysis by ablation of SDOH-related keywords to better understand which topics within social work notes have the greatest contribution to model performance.

We found that pretraining a language model on similar data sources is important for better prediction performance in specialized domains, particularly from small datasets that are common in clinical studies.

Among all the transformer-based models we explored (Table 1), Gatortron-OG achieved the best prediction performance on our task. Moreover, with our hierarchical BERT model, we showed that integrating multiple notes, and consequently more information about a patient, improves model performance. It is generally accepted that including more comprehensive patient information, either from clinical notes written at different times during a health encounter or for a different purpose, will lead to better performance for prediction tasks. Although alternate methods that allow longer input text exist, such as the Longformer technique[38], these approaches usually require retraining a large language model, which can be time-consuming and computationally expensive.

In our feature importance analysis, we found that financial factors are not the sole SDOH factors influencing therapy regimen decisions, as initially hypothesized. Our feature importance analysis revealed other significant factors, including "risk of death" and "abuse history", led to decreases in model performance when removed from social work note text. Notably, simply extracted the appearance, mentioning of these topics mentioning in the social work notes as the feature to perform an prediction failed to achieved ideal performance. This demonstrated it is important to consider the context of these mentions within sentences further emphasizes their importance (S. Table 8, S. Table 11.). This study broadens our understanding of the various factors affecting therapy regimen choices, suggesting that a more comprehensive approach is needed when considering SDOH factors in clinical informatics. Future research should explore additional factors and their potential impact on therapy decisions to ensure a more holistic understanding of patient care.

There are several limitations to this study. While our research showed that social work reports that encompass SDOH information are predictive of the administered breast cancer therapy regimen, integration of structured data and other types of text reports may both highlight other aspects driving the disparity in treatment choice and improve overall predictive performance. While the aim of our paper is to demonstrate the utility of social work notes, comprehensively predicting therapy regimen decisions is complex and

beyond the scope of the current paper. Systematically extracting and converting all SDOH factors from clinical notes to structured data may create additional opportunities for further analysis. In addition, our data was limited to cancer therapy treatment decisions at a single academic medical center. The driving forces behind treatment decisions for patients at other centers may differ, as may the overall distribution of SDOH factors themselves. Patients may already be preselected in unrecorded ways to receive a social worker consultation. Future work should seek to integrate data across institutions with differing practices to further validate our findings. Regarding our BERT-MS-n method, our model used a MLP as the classification layer after integrating note-specific representations. The reasons for selecting MLP include its efficiency and simplicity, making it easy to manipulate and understand. However, we acknowledge that advanced layers like transformers might be better suited than MLP to aggregate note information. We leave the investigation of this possibility for future research. Finally, while our methods for model interpretability are able to uncover important social topics that are associated with the observed disparity, the methods for post-hoc interpretability may not be entirely faithful to the originally trained model, which is an inherent limitation of the current state-of-the-art methods in NLP[39].

3.6 CONCLUSION

In conclusion, our study demonstrates the potential of utilizing transformer-based deep learning approaches for predicting clinical outcomes using social work reports. Specifically, our findings indicate the presence of notable disparities in treatment regimens, which can be attributed to social determinants of health. By creating a hierarchical model that can incorporate additional notes, we observed an enhancement in overall model performance. Through the use of ablation methods to better understand model interpretability, we highlighted the variety of SDOH factors that can influence therapy regimen selection for patients with breast cancer. Future research should extend this analysis to explore the impact of SDOH on treatment selection at other institutions and for different types of cancer.

3.7 CONTRIBUTORSHIP STATEMENT

AJB put forward the research idea. SS, MS and AJB designed the study. SS developed the methods, analyzed the data, and drafted the manuscript. AJB and MS supervised the study. All authors contributed to manuscript review and editing.

3.8 ACKNOWLEDGEMENTS

We thank all researchers, clinicians, and social workers who help collect clinical notes data in our UCSF Information Commons. We thank everyone in Dr. Atul J. Butte's lab for helpful discussion and feedback. We thank staff members in the Bakar Computational Health Sciences Institute and UCSF IT Services who build and maintain the UCSF Information Commons. We thank the Wynton High-Performance Computing (HPC) cluster support team for making available the needed computation capacity.

3.9 COMPETING INTERESTS STATEMENT

AJB is a co-founder and consultant to Personalis and NuMedii; consultant to Mango Tree Corporation, and in the recent past, Samsung, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, and Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Meta (Facebook), Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty Pharma, Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, Nuna Health, Assay Depot, and Vet24seven, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical

or disease specific foundations and associations, and health systems. AJB receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. AJB's research has been funded by NIH, Peraton (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity. None of these entities had any bearing on the design of this study or the writing of the manuscript.

3.10 FUNDNG STATEMENT

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

3.11 DATA AVAILBILTY STATEMENT

The data that support the findings of this study are available from the Information Commons platform at UCSF, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of UCSF.

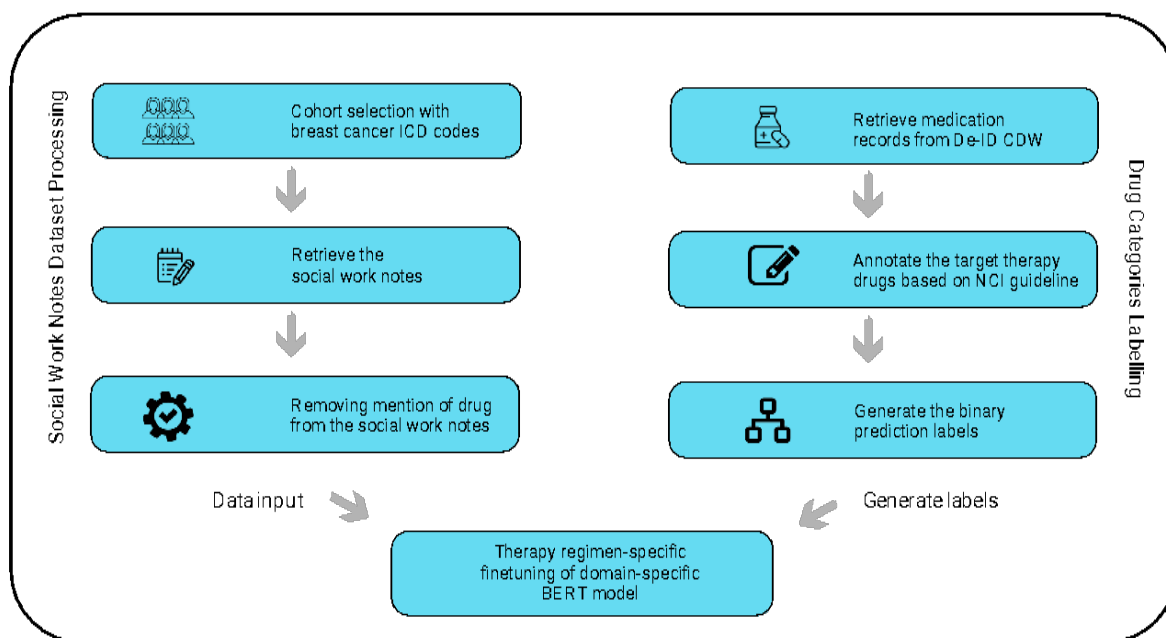
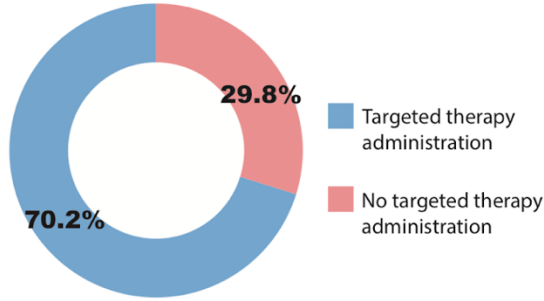


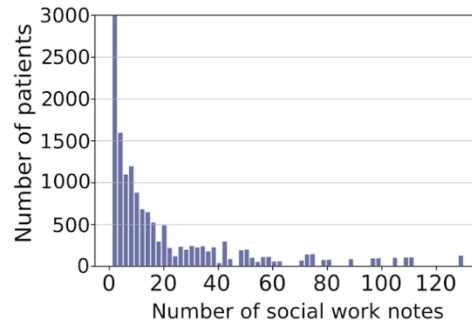
Figure 3.1 The overall workflow

We implement an end-to-end BERT-base classification model to predict the category of treatment administration for breast cancer patients at UCSF. We first retrieved the patients’ social work notes from UCSF de-identified Caboodle Data Warehouse (DeID-CDW) between 2012 and 2021. We then annotated whether an individual patient has ever received targeted therapy based on the Targeted Cancer Therapies Fact Sheet from National Cancer Institute. In this manner, we obtained 2496 patients, of which 70% received targeted therapy. The dataset was further split into 8:1:1 ratio, corresponding to training, validation, and test sets.

A.



B.



C.

Data: 62 year old female with a new diagnosis of left breast DCIS; SW referral received from clinic RN for emotional support.
 Assessment: SW spoke with patient on the phone in Spanish. She lives in ***** with a family that she ***** for. She recently found out that she **has to leave that job because her treatment schedule does not allow enough time for the nanny job.** She can live with the family for the next month but may have to leave after that and will have to find new housing. **She does not have family or friends in *****.** She states that she would like to apply for SSDI when she terminates this job as she wants to retire now. **She pays out of pocket for her ***** insurance plan.** She stated that she has felt

Losing job

Finanacial difficulty

Lack of familiy support

D.

Data: Pt is 45yo female being followed in ***** clinic for follow up of recurrent NET. SW referred to case by ***** RN in context of Financial Navigation pilot program re: pt's recent issues with her Medi-Cal coverage. SW called pt for planned phone consult.
 Assessment: Per pt: she recently received a letter from Medi-Cal saying that **her income is too high and she'll lose coverage.** Pt relayed her recent history of coverage as **having become eligible for Medi-Cal in 2015 due** and ***** in Medi-Cal Managed Care (*****) at that time with Health Plan of ***** (*****), as did her husband and 11yo dtr. Pt reported that she's never had any issues with her Medi-Cal before now, however she only just recently ***** about annual re-determination processes

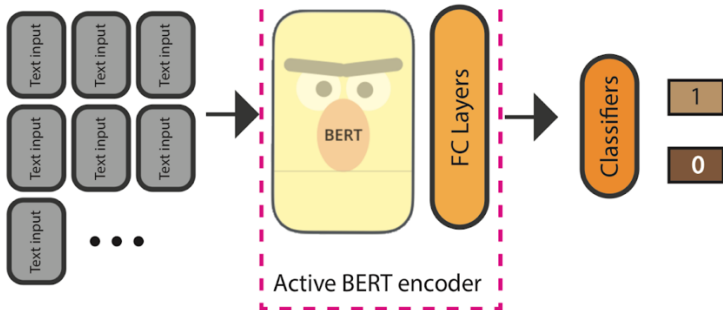
Insurance related information

High incomes

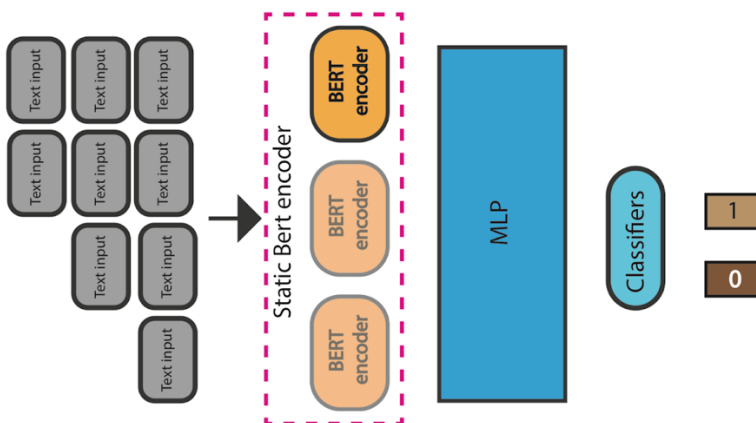
Figure 3.2 Data exploration on social work notes

A. Pie chart showing the different proportions of patients in the two categories. B. Histogram showing the number of notes for the individual patients (mode = 2, mean = 22, median = 11). C. Example deidentified social work notes. Top: Example patient who did not have any targeted therapy administration. Bottom: Example patient who received at least one dose of targeted therapy.

Step 1 Trained the Bert encoder using individual notes



A B C



Step 2 Trained the MLP model for individual patients

Figure 3.3 Illustration of BERT-MS-n model

To use long sequences of clinical notes for prediction, we built a hierarchical BERT model (BERT-MS), where the first step divides a long sequence of notes into multiple independent instances and then trains the single BERT classifier on the individual chunks in the training set. In the second step, we concatenate the BERT representations of all notes of the same patient and further fit them into a multilayer perceptron for the training. FC: Fully connected Layer; MLP: Multi-layer perceptron.

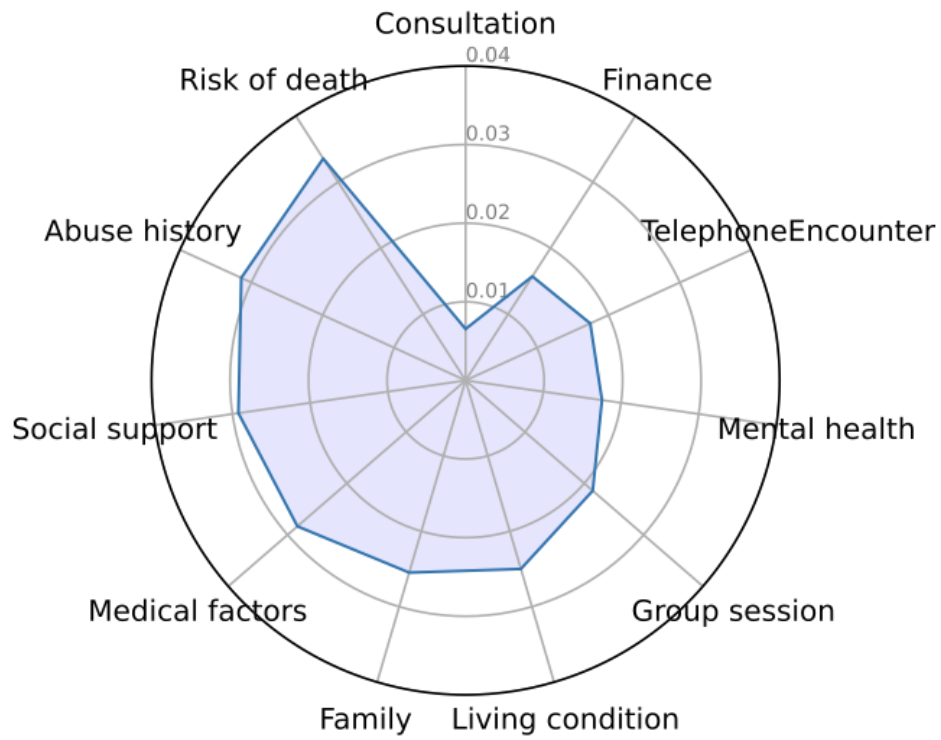


Figure 3.4 Feature importance analysis for SDOH factors in ablation study

The radar chart shows the feature importance of SDOH topics, represented by the decrease of F1 score.

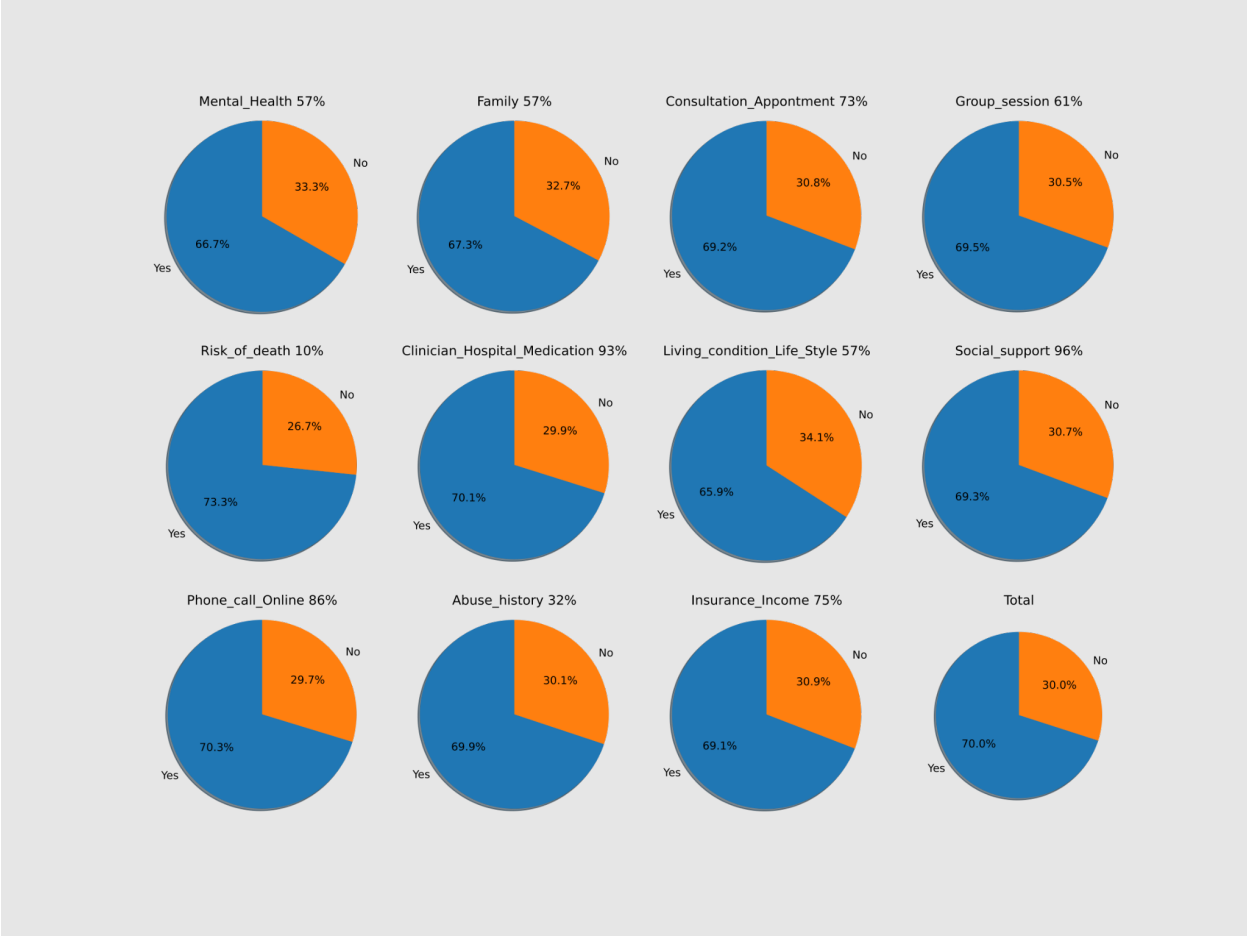


Figure 3.5 Pie chart showing the different proportions

The percentage on the right of each topic indicates the frequency of whether words in the topics existed in individual social work notes. Orange: Patients who did not receive any targeted therapy. Blue: Patients who received at least one dose of targeted therapy.

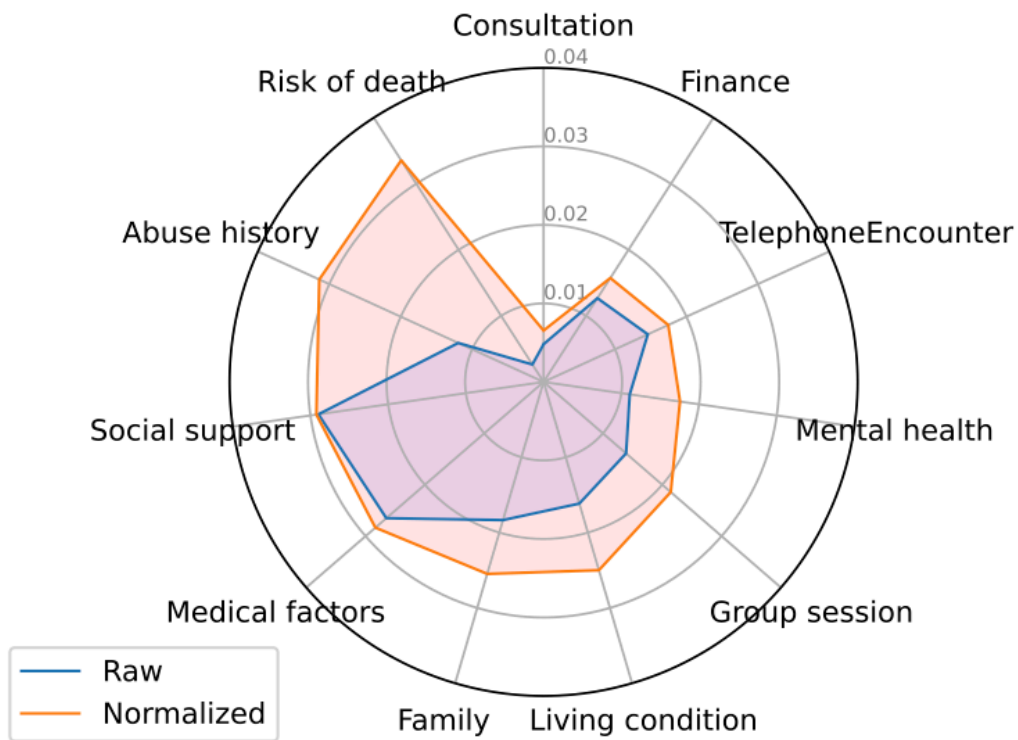


Figure 3.6 The radar chart shows the feature importance of SDOH topics

Feature importance is defined to be the decrease in the F1 score of *Targeted therapy not administered* class across the entire test set overall (Raw: Blue), and across the notes that actually contain these words (Normalized: Orange).

'Social Work Note 08/22/2015 PT previously seen by {...} Family composition/living situation: PT is originally from China, but has lived in the ***** for over ten years. Pt lives w/ her in-***** in ***** and husband and 6yo daughter. Legal: No legal issues identified. Emotional status/coping: PT endorsed stable mood during *****. Support: PT describes FOB as supportive, though they do not have much time together d/t his work schedule Risk Factors: hx of depression; PT describes in-***** as verbally abusive (denies physical abuse); Limited social support Finances/Employment: PT is receptionist at a dental clinic. {...}''

'Data: Ms. ***** was referred to ***** Work Services for assistance with resources related to Domestic Violence. (See note from 05/11/15) Her husband pushed her last week and also pushed her in October. She states that she has not "hit" her (hx of pushing 3x) SW provided NP referral info last week during the medical appointment.

'Social Work Data: SW referred to this pt by team pharmacist {...} Pt confirmed that she used the words "mean" and "controlling" to describe her husband when she was speaking to the team pharmacist. She elaborated to provide me with numerous examples of instances when she felt that her husband was "*****" and "putting [me] down". Pt said her husband has been like this for the entirety of their marriage and that he has current health issues ("memory impairment" and "is deaf") that have caused him to be more abusive as of late. {...}''

Figure 3.6 Example deidentified social work notes contain abusive history information

Table 3.1 Model performance of different classifiers

GatorTron-OG achieved supervisor performance in AUC, MACRO F1, as well as MACRO RECALL.

| Model | AUC | MACRO F1 | MACRO PRECISION | MACRO RECALL |
|--------------------|--------------|-----------------|------------------------|---------------------|
| Gatortron-OG | 0.721 | 0.616 | 0.624 | 0.611 |
| UCSF BERT | 0.675 | 0.599 | 0.604 | 0.596 |
| ClinicalBERT | 0.627 | 0.578 | 0.584 | 0.576 |
| SciBERT | 0.616 | 0.532 | 0.606 | 0.533 |
| BioLM | 0.671 | 0.583 | 0.615 | 0.580 |
| Biomed-RoBERTa | 0.667 | 0.584 | 0.592 | 0.581 |
| Dummy (Prior) | 0.500 | 0.412 | 0.350 | 0.491 |
| Dummy (stratified) | 0.504 | 0.525 | 0.529 | 0.603 |
| Dummy (Uniform) | 0.500 | 0.509 | 0.522 | 0.602 |

Table 3.2 BERT MS model achieved superior performance in AUC, MACRO F1, as well as MACRO RECALL.

| | AUC | MACRO F1 | MACRO PRECISION | MACRO RECALL |
|-----------------|--------------|-----------------|------------------------|---------------------|
| UCSF BERT | 0.675 | 0.599 | 0.604 | 0.596 |
| UCSF BERT MS-3 | 0.707 | 0.620 | 0.660 | 0.612 |
| UCSF BERT MS-5 | 0.702 | 0.624 | 0.637 | 0.615 |
| UCSF BERT MS-8 | 0.718 | 0.623 | 0.645 | 0.616 |
| UCSF BERT MS-10 | 0.706 | 0.596 | 0.665 | 0.594 |

Table 3.3 Demographic characteristics for breast cancer patients in our cohort

| | Targeted therapy not administered (TT-No, N=597) | Targeted therapy administered (TT-Yes, N=1899) | Overall (N=2496) |
|------------------------|---|---|---------------------|
| Sex | | | |
| Female | 572 (95.8%) | 1871 (98.5%) | 2443 (97.9%) |
| Male | 25 (4.2%) | 28 (1.5%) | 53 (2.1%) |
| Ethnicity | | | |
| Hispanic/Latino | 72 (12.1%) | 180 (9.5%) | 252 (10.1%) |
| Not Hispanic or Latino | 497 (83.2%) | 1646 (86.7%) | 2143 (85.9%) |
| Other | 28 (4.6%) | 73 (2.0%) | 101 (2.2%) |
| Race | | | |
| Asian | 110 (18.4%) | 341 (18.0%) | 451 (18.1%) |
| Black/African | 33 (5.5%) | 129 (6.8%) | 162 (6.5%) |
| White or Caucasian | 355 (59.5%) | 1136 (59.8%) | 1491 (59.7%) |
| Other | 99 (17.6%) | 283 (15.4%) | 382(15.7%) |

Table 3.4 Summary characteristics of social factors (smoking and marital status) for breast cancer patients extracted from structured data

| | Targeted therapy not administered (TT-No, N=597) | Targeted therapy administered (TT-Yes, N=1899) | Overall (N=2496) |
|---------------------------------------|--|--|------------------|
| Smoking status | | | |
| Current Everyday Smoker | 11 (1.8%) | 39 (2.1%) | 50 (2.0%) |
| Current Some Day Smoker | 6 (1.0%) | 21 (1.1%) | 27 (1.1%) |
| Former Smoker | 177 (29.6%) | 569 (30.0%) | 746 (29.9%) |
| Never Assessed | 2 (0.3%) | 9 (0.5%) | 11 (0.4%) |
| Never Smoker | 395 (66.2%) | 1232 (64.9%) | 1627 (65.2%) |
| Passive Smoke Exposure - Never Smoker | 5 (0.8%) | 17 (0.9%) | 22 (0.9%) |
| Smoker, Current Status Unknown | 1 (0.2%) | 1 (0.1%) | 2 (0.1%) |
| *Unknown | 0 (0%) | 6 (0.3%) | 6 (0.2%) |
| Light Tobacco Smoker | 0 (0%) | 3 (0.2%) | 3 (0.1%) |
| Unknown If Ever Smoked | 0 (0%) | 2 (0.1%) | 2 (0.1%) |
| Marital status | | | |
| *Unspecified | 1 (0.2%) | 0 (0%) | 1 (0.0%) |
| Divorced | 65 (10.9%) | 212 (11.2%) | 277 (11.1%) |
| Legally Separated | 4 (0.7%) | 15 (0.8%) | 19 (0.8%) |
| Married | 286 (47.9%) | 909 (47.9%) | 1195 (47.9%) |
| Registered Domestic Partner | 7 (1.2%) | 9 (0.5%) | 16 (0.6%) |
| Significant Other | 14 (2.3%) | 37 (1.9%) | 51 (2.0%) |
| Single | 164 (27.5%) | 480 (25.3%) | 644 (25.8%) |
| Unknown/Declined | 18 (3.0%) | 43 (2.3%) | 61 (2.4%) |
| Widowed | 38 (6.4%) | 193 (10.2%) | 231 (9.3%) |

Table 3.5 Model performances of common machine learning classifiers using SDOH related structured tabular data on targeted therapy administration.

| | AUC | MACRO F1 | MACRO PRECISION | MACRO RECALL |
|----------------------------|------------|-----------------|------------------------|---------------------|
| KNeighborsClassifier | 0.497 | 0.491 | 0.496 | 0.497 |
| SVM Classifier | 0.500 | 0.434 | 0.383 | 0.500 |
| RandomForestClassifier | 0.519 | 0.483 | 0.592 | 0.517 |
| GradientBoostingClassifier | 0.509 | 0.458 | 0.635 | 0.509 |

Table 3.6 The properties of notes for breast cancer patient’s cohort. (Measure the tokens length and compare with 512 tokens).

| Percentage of notes longer than >300 words | Percentage of notes longer than >400 words | Percentage of notes longer than >2000 characters | Percentage of notes longer than >2500 characters |
|--|--|--|--|
| 34.0% | 22.2% | 26.1% | 17.9% |

Table 3.7 The words in the Keywords column are the representative words used to define the topics.

| Topics | Keywords |
|--|--|
| Family | family, parent, father, mother, child, children, sister, parents, relatives, clan, childhood, friends |
| Consultation/Appointment | appointment, consultation, consult, questionnaire, question, advice, biographical, Wikipedia, relevant, questions, know, documentation |
| Group session | group, intervention, session, interpers, community, class, organization, together, part, organization |
| Risk of death | suicide, suicidal, risk, crisis, homicide, murder, commit, bombing, murdered, murders, bomber, killing, convicted, victims |
| Medical factors | patient, medication, hospital, medical, clinic, clinician, treatment, therapy, surgery, symptoms, patients, drugs, diagnosis, treatments, prescribed |
| Living condition/Lifestyle | shelter, housing, house, living, sleep, bedtime, building, buildings, urban, employment, suburban, campus, acres |
| Social support | social, service, support, referral, recommendation, recommend, worker, resource, supports, provide, supporting, supported, allow, providing, assistance, benefit, help |
| TelephoneEcounter/Online communication | telehealth, phone, call, video, telephone, mobile, wireless, gsm, cellular, dial, email, calling, networks, calls, messages, telephones, internet |
| Abuse history | abuse, history, addiction, alcohol, drugs, allegations, victim, violence, sexual, rape, dependence |
| Insurance/Income | insurance, income, coverage, financial, contracts, banking, finance, liability, private, pay |

Table 3.8 Model performance of different classifiers

External big language model Gatortron achieved the state-of-the-art performance, demonstrating the reliability of our discovery.

| Model | AUC | MACRO F1 | MACRO PRECISION | MACRO RECALL |
|--------------------|--------------|-----------------|------------------------|---------------------|
| UCSF BERT | 0.675 | 0.599 | 0.604 | 0.596 |
| Gatortron-OG | 0.721 | 0.616 | 0.624 | 0.611 |
| ClinicalBERT | 0.627 | 0.578 | 0.584 | 0.576 |
| SciBERT | 0.616 | 0.532 | 0.606 | 0.533 |
| BioLM | 0.671 | 0.583 | 0.615 | 0.580 |
| Biomed-RoBERTa | 0.667 | 0.584 | 0.592 | 0.581 |
| Dummy (Prior) | 0.500 | 0.412 | 0.350 | 0.491 |
| Dummy (stratified) | 0.504 | 0.525 | 0.529 | 0.603 |
| Dummy (Uniform) | 0.500 | 0.509 | 0.522 | 0.602 |

Table 3.9 The removal of notes drug mentioning in the prediction pipeline.

| Model | AUC | MACRO F1 | MACRO PRECISION | MACRO RECALL |
|---|------------|-----------------|------------------------|---------------------|
| UCSF BERT (with Drug info masked) | 0.675 | 0.599 | 0.604 | 0.596 |
| UCSF BERT excluding notes mentioning Drug | 0.696 | 0.585 | 0.622 | 0.562 |

Table 3.10 Model performance of leveraging SDOH topics appearance on regimen prediction, without semantic meanings.

| | F1 | Precision | Recall | Accuracy |
|------------------------------|-----------|------------------|---------------|-----------------|
| SVM | 0.408 | 0.345 | 0.500 | 0.690 |
| Logistic Regression | 0.408 | 0.345 | 0.500 | 0.690 |
| Random Forest | 0.513 | 0.516 | 0.514 | 0.603 |
| Multilayer perceptron | 0.456 | 0.598 | 0.51 | 0.690 |
| Naive Bayes | 0.401 | 0.406 | 0.399 | 0.466 |

Table 3.11 Demographic characteristics for all breast cancer patients

| | Overall (N=30631) |
|------------------------|-------------------|
| Sex | |
| Female | 30026 (98.0%) |
| Male | 589 (1.9%) |
| Ethnicity | |
| Hispanic/Latino | 1913 (6.2%) |
| Not Hispanic or Latino | 23407 (76.4%) |
| Other | 5311 (17.4%) |
| Race | |
| Asian | 2809 (9.2%) |
| Black/African | 1411 (4.6%) |
| White or Caucasian | 18744 (61.2%) |
| Other | 7667 (25.0%) |

Table 3.12 Summary characteristics of social factors (smoking and marital status) for all breast cancer patients

| Smoking Status | Overall (N=30631) |
|---------------------------------------|--------------------------|
| *Not Applicable | 1 (0.0%) |
| *Unknown | 10416 (34.0%) |
| *Unspecified | 1425 (4.7%) |
| Current Every Day Smoker | 135 (0.4%) |
| Current Some Day Smoker | 43 (0.1%) |
| Every Day | 210 (0.7%) |
| Former | 3336 (10.9%) |
| Former Smoker | 1724 (5.6%) |
| Heavy Smoker | 5 (0.0%) |
| Heavy Tobacco Smoker | 1 (0.0%) |
| Light Smoker | 24 (0.1%) |
| Light Tobacco Smoker | 5 (0.0%) |
| Never | 8053 (26.3%) |
| Never Assessed | 925 (3.0%) |
| Never Smoker | 3533 (11.5%) |
| Passive Smoke Exposure - Never Smoker | 129 (0.4%) |
| Smoker, Current Status Unknown | 36 (0.1%) |
| Some Days | 78 (0.3%) |
| Unknown | 70 (0.2%) |
| Unknown If Ever Smoked | 38 (0.1%) |
| Maritalstatus | |
| | 3 (0.0%) |
| *Not Applicable | 1 (0.0%) |
| *Unknown | 1 (0.0%) |
| *Unspecified | 35 (0.1%) |
| Divorced | 2506 (8.2%) |
| Legally Separated | 197 (0.6%) |
| Married | 16229 (53.0%) |
| RDP-Dissolved | 3 (0.0%) |

| Smoking Status | Overall (N=30631) |
|-----------------------------|--------------------------|
| RDP-LG SEP | 1 (0.0%) |
| RDP-Widowed | 38 (0.1%) |
| Registered Domestic Partner | 74 (0.2%) |
| Significant Other | 225 (0.7%) |
| Single | 6284 (20.5%) |
| Unknown/Declined | 2069 (6.8%) |
| Widowed | 2965 (9.7%) |

Table 3.13

The drug names that we masked.

| Carboplatin | Intravenous | RiTUXimab | Carfilzomib | CARBOplatin | Trastuzumab- anns |
|------------------------|--------------------|------------------|--------------------|--------------------|------------------------------|
| Etoposide | Brigatinib | Hydroxyurea | Cisplatin | Ruxolitinib | Inotuzumab |
| Ibrutinib | Irinotecan | Ciloleucel | Anastrozole | Copanlisib | Dabrafenib |
| Mechlorethamine | Venetoclax | Acalabrutinib | Procarbazine | Larotrectinib | Triptorelin |
| Flutamide | Mitotane | Melphalan | Methoxsalen | Midostaurin | InFLIXimab |
| Abiraterone | Avelumab | Cytarabine | VinCRISTine | Lapatinib | Ifosfamide |
| FluorouraciL | Erlotinib | DAUNOrubicin | DOCEtaxeL | Ramucirumab | VinORELBine |
| Thioguanine | Alpelisib | Decitabine | IDArubicin | Encorafenib | Capecitabine |
| Trastuzumab | Goserelin | Pemetrexed | Ivosidenib | Cladribine | MetHOTREXate |
| SORAFenib | Interferon | Alemtuzumab | Daratumumab | Elotuzumab | Tamoxifen |
| MitoMYcin | Certolizumab | Intrathecal | Vemurafenib | Bicalutamide | MegestroL |
| Axitinib | Gemtuzumab | Durvalumab | Enasidenib | Ipilimumab | CabazitaxeL |
| Emtansine | Olaparib | PACLitaxeL | Temozolomide | Hyaluronid | Bevacizumab |
| Doxorubicin | Lenvatinib | PEMEtrexed | DOXOrubicin | Cabozantinib | Osimertinib |
| Mercaptopurine | Cetuximab | Etanercept | Bleomycin | Rituximab | Hyaluronidase |
| Intravesical | Deruxtecan | EpiRUBicin | Siltuximab | Epirubicin | Everolimus |
| Neratinib | Trastuzumab | Bevacizumab | Enzalutamide | Vedotin | Gilteritinib |
| Eribulin | Panitumumab | Fludarabine | Vandetanib | Topotecan | Megestrol |
| Exemestane | Niraparib | Adalimumab | Trifluridine | Tipiracil | Regorafenib |
| PAZOPanib | Fulvestrant | Ceritinib | Lenalidomide | Afatinib | CISplatin |
| Pomalidomide | Laherparepvec | Tucatinib | Mebutate | Sorafenib | Clofarabine |
| Thiotepa | Bortezomib | Ixazomib | Fluorouracil | Dasatinib | Trastuzumab |
| Talimogene | Nilotinib | Obinutuzumab | Cyclophosphamide | Sipuleucel-T | Panobinostat |
| Carmustine | EriBULin | Toremifene | Trioxide | Govitecan- hziy | Diclofenac |
| Dacarbazine | Leuprolide | Binimetinib | Betadex | Methotrexate | Letrozole |
| ChlorambuciL | Rucaparib | Pazopanib | Atezolizumab | Paclitaxel | Crizotinib |
| PACLitaxel- protein | Aminolevulinic | Pembrolizumab | Infliximab | Ribociclib | Hyaluronidase |
| SUNItinib | Abemaciclib | Trametinib | Sunitinib | Ixabepilone | Lorlatinib |
| Sacituzumab | Vorinostat | | | | |

3.12 REFERENCES

- [1] Dehon E, Weiss N, Jones J, et al. A systematic review of the impact of physician implicit racial bias on clinical decision making. *Academic Emergency Medicine* 2017;24:895–904.
- [2] Murray E, Pollack L, White M, et al. Clinical decision-making: Patients’ preferences and experiences. *Patient education and counseling* 2007;65:189–96.
- [3] Stipelman CH, Kukhareva PV, Trepman E, et al. Electronic Health Record-Integrated Clinical Decision Support for Clinicians Serving Populations Facing Health Care Disparities: Literature Review. *Yearbook of Medical Informatics* 2022;31:184–98.
- [4] Hajjaj F, Salek M, Basra M, et al. Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice. *J R Soc Med* 2010;103:178–87. doi:10.1258/jrsm.2010.100104
- [5] Marmot M, Wilkinson R. *Social Determinants of Health*. OUP Oxford 2005.
- [6] Davidson J, Vashisht R, Butte AJ. From Genes to Geography, from Cells to Community, from Biomolecules to Behaviors: The Importance of Social Determinants of Health. *Biomolecules* 2022;12:1449. doi:10.3390/biom12101449
- [7] Tsimberidou A-M. Targeted therapy in cancer. *Cancer Chemother Pharmacol* 2015;76:1113–32. doi:10.1007/s00280-015-2861-1
- [8] Bernheim: Influence of patients’ socioeconomic... - Google Scholar. https://scholar.google.com/scholar_lookup?journal=Ann+Fam+Med&title=Influence+of+patients%27+socioeconomic+status+on+clinical+management+decisions:+a+qualitative+study&author=SM+Bernheim&author=JS+Ross&author=HM+Krumholz&author=EH+Bradley&volume=6&publication_year=2008&pages=53-9&pmid=18195315& (accessed 16 Nov 2022).
- [9] Maynard C, Fisher LD, Passamani ER, et al. Blacks in the coronary artery surgery study (CASS): race and clinical decision making. *Am J Public Health* 1986;76:1446–8. doi:10.2105/AJPH.76.12.1446

- [10] Kressin NR, Petersen LA. Racial Differences in the Use of Invasive Cardiovascular Procedures: Review of the Literature and Prescription for Future Research. *Ann Intern Med* 2001;135:352–66. doi:10.7326/0003-4819-135-5-200109040-00012
- [11] Wilder ME, Kulie P, Jensen C, et al. The Impact of Social Determinants of Health on Medication Adherence: a Systematic Review and Meta-analysis. *J Gen Intern Med* 2021;36:1359–70. doi:10.1007/s11606-020-06447-0
- [12] Braveman P, Egerter S, Williams DR. The Social Determinants of Health: Coming of Age. *Annual Review of Public Health* 2011;32:381–98. doi:10.1146/annurev-publhealth-031210-101218
- [13] Philpot LM, Ebbert JO, Hurt RT. A survey of the attitudes, beliefs and knowledge about medical cannabis among primary care providers. *BMC Family practice* 2019;20:1–7.
- [14] Glatzer M, Panje CM, Sirén C, et al. Decision making criteria in oncology. *Oncology* 2020;98:370–8.
- [15] Differences in access to zidovudine (AZT) among symptomatic HIV-infected persons | SpringerLink. <https://link.springer.com/article/10.1007/BF02599388> (accessed 16 Nov 2022).
- [16] Verbrugge LM, Steiner RP. Physician Treatment of Men and Women Patients: Sex Bias or Appropriate Care? *Medical Care* 1981;19:609–32.
- [17] Who is targeted for lifestyle advice? A cross-sectional survey in two general practices. | *British Journal of General Practice*. <https://bjgp.org/content/49/447/806.short> (accessed 16 Nov 2022).
- [18] Impact of medical and nonmedical factors on physician decision making for HIV/AIDS antiretroviral treatment. - Abstract - Europe PMC. <https://europepmc.org/article/med/10866232> (accessed 16 Nov 2022).
- [19] ‘difficult patient’ as perceived by family physicians | *Family Practice* | Oxford Academic. <https://academic.oup.com/fampra/article/18/5/495/664879> (accessed 16 Nov 2022).
- [20] Sociologic Influences on Decision-Making by Clinicians | *Annals of Internal Medicine*. <https://www.acpjournals.org/doi/abs/10.7326/0003-4819-90-6->

957?casa_token=aZcuQs1XsGgAAAAA:ITQxqv9wCTdXzU2WamGwnSfdVosnoNkuHqHgcT81
JDrDsp5zpcitCelahF7V3RaxCruw4RBo1wyTQ (accessed 16 Nov 2022).

- [21] Social Work and the Social Determinants of Health Perspective: A Good Fit | Health & Social Work | Oxford Academic. <https://academic.oup.com/hsw/article/35/4/310/609930> (accessed 4 Apr 2023).
- [22] Rine CM. Social Determinants of Health: Grand Challenges in Social Work's Future. *Health & Social Work* 2016;41:143–5. doi:10.1093/hsw/hlw028
- [23] Riessman CK, Quinney L. Narrative in social work: A critical review. *Qualitative social work* 2005;4:391–412.
- [24] Ross H, Dritz R, Morano B, et al. The unique role of the social worker within the Hospital at Home care delivery team. *Soc Work Health Care* 2021;60:354–68. doi:10.1080/00981389.2021.1894308
- [25] Targeted Therapy for Cancer - NCI. 2014. <https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies> (accessed 2 Jan 2023).
- [26] Sushil M, Ludwig D, Butte AJ, et al. Developing a general-purpose clinical language inference model from a large corpus of clinical notes. arXiv preprint arXiv:221006566 2022.
- [27] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs] Published Online First: 24 May 2019. <http://arxiv.org/abs/1810.04805> (accessed 2 May 2022).
- [28] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:190310676 2019.
- [29] Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. 2020. doi:10.48550/arXiv.1904.05342
- [30] Lewis P, Ott M, Du J, et al. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: : Association for Computational Linguistics 2020. 146–57. doi:10.18653/v1/2020.clinicalnlp-1.17

- [31] Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. doi:10.48550/arXiv.1907.11692
- [32] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 2011;12:2825–30.
- [33] Sun S, Zack T, Sushil M, et al. Topic Modeling on Clinical Social Work Notes for Exploring Social Determinants of Health Factors. arXiv preprint arXiv:221201462 2022.
- [34] University of California, San Francisco, ARS. UCSF DeID CDW. R20220207. UCSF Data Resources. <https://data.ucsf.edu/research> (accessed 12 Sep 2022).
- [35] Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:200410964 2020.
- [36] Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Scientific data 2016;3:1–9.
- [37] Sushil M, Butte AJ, Schuit E, et al. Cross-institution text mining to uncover clinical associations: a case study relating social factors and code status in intensive care medicine. arXiv preprint arXiv:230106570 2023.
- [38] Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. 2020. doi:10.48550/arXiv.2004.05150
- [39] Lipton ZC. The Mythos of Model Interpretability. 2017. doi:10.48550/arXiv.1606.03490
- [40] Norgeot, B., Muenzen, K., Peterson, T.A. et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. npj Digit. Med. 3, 57 (2020). <https://doi.org/10.1038/s41746-020-0258-y>

4 Chapter 4: Aligning Synthetic Medical Images with Clinical Knowledge using Human Feedback

Shenghuan Sun* University of California, San Francisco shenghuan.sun@ucsf.edu Gregory M. Goldgof*
Memorial Sloan Kettering Cancer Center goldgofg@mskcc.org Atul Butte University of California, San
Francisco atul.butte@ucsf.edu Ahmed M. Alaa UC Berkeley and UCSF amalaa@berkeley.edu

4.1 ABSTRACT

Generative models capable of capturing nuanced clinical features in medical images hold great promise for facilitating clinical data sharing, enhancing rare disease datasets, and efficiently synthesizing annotated medical images at scale. Despite their potential, assessing the quality of synthetic medical images remains a challenge. While modern generative models can synthesize visually-realistic medical images, the clinical validity of these images may be called into question. Domainagnostic scores, such as FID score, precision, and recall, cannot incorporate clinical knowledge and are, therefore, not suitable for assessing clinical sensibility. Additionally, there are numerous unpredictable ways in which generative models may fail to synthesize clinically plausible images, making it challenging to anticipate potential failures and manually design scores for their detection. To address these challenges, this paper introduces a pathologist-in-the-loop framework for generating clinically-plausible synthetic medical images. Starting with a diffusion model pretrained using real images, our framework comprises three steps: (1) evaluating the generated images by expert pathologists to assess whether they satisfy clinical desiderata, (2) training a reward model that predicts the pathologist feedback on new samples, and (3) incorporating expert knowledge into the diffusion model by using the reward model to inform a finetuning objective. We show that human feedback significantly improves the quality of synthetic images in terms of fidelity, diversity, utility in downstream applications, and plausibility as evaluated by experts. We also show that human feedback can teach the

model new clinical concepts not annotated in the original training data. Our results demonstrate the value of incorporating human feedback in clinical applications where generative models may struggle to capture extensive domain knowledge from raw data alone.

4.2 INTRODUCTION

Diffusion models have recently shown incredible success in the conditional generation of high-fidelity natural, stylized and artistic images [1–6]. The generative capabilities of these models can be leveraged to create synthetic data in application domains where obtaining large-scale annotated datasets is challenging. The medical imaging field is one such domain, where there is often a difficulty in obtaining high-quality labeled datasets [7]. This difficulty may stem from the regulatory hurdles that impede data sharing [8], the costs involved in getting experts to manually annotate images [9], or the natural scarcity of data in rare diseases [10]. Generative (diffusion) models may provide a partial solution to these problems by synthesizing high-fidelity medical images that can be easily shared among researchers to replace or augment real data in downstream modeling applications [11, 12].

WHAT SETS MEDICAL IMAGE SYNTHESIS APART FROM IMAGE GENERATION IN OTHER FIELDS?

Unlike mainstream generative modeling applications that prioritize visually realistic or artistically expressive images, synthetic medical images require a different approach. They must be grounded in objective clinical and biological knowledge, and as such, they leave no room for creative or unrestricted generation. Given that the ultimate goal of synthetic medical images is to be used in downstream modeling and analysis, they must faithfully reflect nuanced features that represent various clinical concepts, such as cell types [13], disease subtypes [7], and anatomies [14]. Off-the-shelf image generation models are not capable of recognizing or generating clinical concepts, rendering them unsuitable for generating plausible medical images without further adaptation [15, 16]. Therefore, our aim is to develop a framework for

generating synthetic medical images that not only exhibit visual realism but also demonstrate biological plausibility and alignment with clinical expertise.

One way to generate synthetic medical images is to finetune a pretrained “foundation” vision model, such as Stable Diffusion, that has been trained on billions of natural images (such as the LAION-5B dataset [17]), using real medical images. With a sufficiently large set of medical images, we can expect the finetuned model to capture the clinical knowledge encoded in medical images. However, the sample sizes of annotated medical images are typically limited to a few thousand. When a large vision model is finetuned on such a dataset using generic objective functions (such as the likelihood function), the model may capture only the generic features that make the medical images appear visually realistic, but it may miss nuanced features that make them biologically plausible and compliant with clinical domain knowledge (see examples in the next Section). Designing domain-specific objective functions for finetuning that ensure a generative model adheres to clinical knowledge is challenging. The difficulty arises from the numerous unpredictable ways in which these models can generate images that lack clinical plausibility. As a result, it is impractical to anticipate every possible failure scenario and manually construct a loss function that penalizes such instances.

SUMMARY OF CONTRIBUTIONS

In this paper, we develop a pathologist-in-the-loop framework for synthesizing medical images that align with clinical knowledge. Our framework is motivated by the success of reinforcement learning with human feedback (RLHF) in aligning the outputs of large language models (LLMs) with human preference [18, 19], and is directly inspired by [20], where human feedback was used to align the visual outputs of a generative model with input text prompts. To generate clinically-plausible medical images, our framework (outlined in Figure 4.1) comprises 3 steps:

Step 1: We train a (conditional) diffusion model using real medical images. We then sample a synthetic dataset from the model to be evaluated by a pathologist. Each image is carefully examined, and the pathologist provides feedback on whether it meets the necessary criteria for clinical plausibility.

Step 2: We collate a dataset of synthetic images paired with pathologist feedback and train a reward model to predict the pathologist feedback, i.e., clinical plausibility, on new images.

Step 3: Finally, the reward model in Step 2 is utilized to incorporate expert knowledge into the generative model. This is achieved by finetuning the diffusion model using a reward-weighted loss function, which penalizes the generation of images that the pathologist considers clinically implausible.

Throughout this paper, we apply the steps above to the synthetic generation of bone marrow image patches, but the same conceptual framework can generalize to any medical imaging modality. We gathered pathologist feedback on thousands of synthetic images of various cell types generated by a conditional diffusion model. Then, we analyzed the impact of this feedback on the quality of the finetuned synthetic images. Our findings suggest that incorporating pathologist feedback significantly enhances the quality of synthetic images in terms of all existing quality metrics such as fidelity, accuracy of downstream predictive models, and clinical plausibility as evaluated by experts. Additionally, it also improves qualities that are not directly addressed in the pathologist evaluation, such as the diversity of synthetic samples. Furthermore, we show that human feedback can teach the generative model new clinical concepts, such as more refined identification of cell types, that are not annotated in the original training data. These results demonstrate the value of incorporating human feedback in clinical applications where generative models may not be readily suited to capture intricate and extensive clinical domain knowledge from raw data alone.

PATHOLOGIST-IN-THE-LOOP GENERATION OF SYNTHETIC MEDICAL IMAGES

In this Section, we provide a detailed description of our synthetic medical image generation framework. We will use a running example pertaining to single-cell images extracted from bone marrow aspirate whole slide images. Details of this setup and the dataset used in our study are provided in Section 4.

Step 1: Pathologist feedback collection. The first step in our framework starts with training a generative model to synthesize medical images through the standard training procedure. As we discuss in more detail in Section 4, we utilized a dataset of 2,048 bone marrow image patches to train a conditional diffusion model [6]. The model was trained to generate class-conditional images, where an image class corresponds to a cell type. We conducted an exploratory analysis where we found that neither latent diffusion nor Stable Diffusion models yielded superior results compared to a customized diffusion model that we employed in this study (See Appendix A). We opted for using a class-conditional model rather than a text-conditional model as we found that existing pretrained vision-language models were not fit for capturing the scientific jargon related to bone marrow cell types.

Given a dataset of real images $D_r = \{(x_i, c_i)\}_{i=1}^n$, where x is a medical image and $c \in C$ is an image class, we train a diffusion model to generate class-conditional images through the forward process:

$$x_{t+1} = x_t - \alpha_t \cdot \nabla_x \log(p_\theta(x_t|x, c)) + \epsilon_t, \quad (1)$$

where $\epsilon_t \sim N(0, \rho^2)$ is the noise term at time-step t , x_t is the data point at time-step t , α_t is the step size at t , θ is the model parameters and $\nabla_x \log(p_\theta(x_t|x, c))$ is the gradient of the log probability distribution with respect to x , conditioned on the original data x and class c . Once the model is pretrained, we sample a synthetic dataset $D_s = \{(x_e^j, c_e^j)\}_{j=1}^n$ by first sampling a class c_e from C , and then sampling a medical image x_e conditioned on the class c_e through the reverse diffusion process. Pathologist evaluation. Each image in the synthetic dataset $D_s = \{(x_e^j, c_e^j)\}_{j=1}^n$ generated by the pretrained model is inspected by an expert pathologist to assess its clinical plausibility. The objective of

this evaluation is to identify the specific inaccuracies in the synthetic data that can only be identified by an expert, and provide feedback for the model to refine its synthesized images in the finetuning step. When a model is trained with only a modest number of real image samples, it may generate bone marrow image patches that look visually appealing but are not biologically plausible. In Figure 4.2, we present 8 synthetic images sampled from the conditional diffusion model, which correspond to four different cell types. Each of these images achieves high precision and fidelity scores individually, but they also have biological implausibilities such as inaccurate cell coloring or nucleus shapes. Therefore, models that prioritize visual features without considering biological knowledge may miss important clinical features required for synthetic images to be useful for downstream analysis. Generic evaluation scores (e.g., [21, 22]) cannot diagnose these failures because they also lack biological domain knowledge. By incorporating feedback from pathologists, we can refine the generative model by identifying biological information that is missed by the pretrained model.

The expert pathologist examined each synthetic image and provided a feedback score on its biological plausibility. The evaluation typically involved inspecting the image and checking 7 aspects that contribute to its perceived plausibility (Table 1). These aspects pertain to the consistency of the shapes, sizes, patterns and colors of the contents of a synthetic bone marrow image with the cell type etc. Among the determinants of plausibility is the cell size—different cells have different sizes, e.g., Lymphocytes are generally smaller than Monocytes or Neutrophils. Nucleus shape and size also depend on the cell type, e.g., Band Neutrophils have a horseshoe-shaped nucleus, whereas Segmented Neutrophils have a multi-lobed nucleus. Chromatin patterns within the nucleus are dense and clumped in Lymphocytes, while in Myeloid cells they are diffuse and fine. The number, size and color of granules also contribute to plausibility. Detailed explanation of all criteria is provided in the Appendix.

Among the determinants of plausibility is the cell size—different cells have different sizes, e.g., Lymphocytes are generally smaller than Monocytes or Neutrophils. Nucleus shape and size also depend on

the cell type, e.g., Band Neutrophils have a horseshoe-shaped nucleus, whereas Segmented Neutrophils have a multi-lobed nucleus. Chromatin patterns within the nucleus are dense and clumped in Lymphocytes, while in Myeloid cells they are diffuse and fine. The number, size and color of granules also contribute to plausibility. Detailed explanation of all criteria is provided in the Appendix. Note that we have full control over the number of synthetic images n_s , i.e., we can sample an arbitrary number of synthetic images from the conditional diffusion model. The key limiting factor on n_s is the time-consuming nature of the feedback collection process. To enable scalable feedback collection, we limited our study to binary feedback, i.e., the pathologist flagged a synthetic image as “implausible” if they found a violation of any of the criteria in Table 1 upon visual inspection. We collected these binary signals and did not pursue a full checklist on all plausibility criteria for each synthetic image. The output of Step 1 is an annotated dataset $D_s = \{(x_{e_j}, e_{c_j}, y_{e_j})\}_{j=1}^{n_s}$, where $y_{e_j} \in \{0, 1\}$ is the pathologist feedback on the j -th synthetic image, where $y_{e_j} = 1$ means that the image is implausible.

Step 2: Clinical Plausibility Reward Modeling. We conceptualize the pathologist as a "labeling function" $\Gamma: X \times C \rightarrow \{0, 1\}$ that maps the observed synthetic image x_e and declared cell type (class) e_c to a binary plausibility score. In Step 2, we model the "pathologist" by learning their labeling function Γ based on their feedback annotations.

To train a model Γ that estimates the pathologist labeling function, we construct a training dataset comprising a mixture of real and synthetic images as follows:

- Synthetic images: We construct a dataset $D_s^\Gamma = \{(x_{e_j}, e_{c_j}, y_{e_j})\}_{j=1}^{n_s}$ comprising the synthetic images and corresponding pathologist feedback collected in Step 1.
- Real images: We build a dataset $D_r^\Gamma = \{(x_i, e_{c_i}, y_i)\}_{i=1}^{n_r}$ comprising the real images and pseudo-labels defined as $y_i = 1 \{c_i \neq e_{c_i}\}$, where $e_{c_i} \sim \text{Uniform}(1, 2, \dots, |C|)$.

We combine both datasets $D\Gamma = D\Gamma_r \cup D\Gamma_s$ to construct a training dataset for the model Γ . The real dataset $D\Gamma_r$ is built by randomly permuting the image class and assigning an implausibility label of 1 if the permuted class does not coincide with the true class. We use the real dataset to augment the synthetic dataset with the annotated pathologist feedback. By augmenting the datasets $D\Gamma_r$ and $D\Gamma_s$, we teach the model Γ to recognize two forms of implausibility in image generation:

Instances where the synthetic image looks clinically plausible but belongs to a wrong cell type (i.e., training examples in $D\Gamma_r$). Instances where the synthetic image is visually consistent with the correct cell type but fails to meet some of the plausibility criteria in Table 1 (i.e., subset of the training examples in $D\Gamma_s$). We call the resulting model Γ a clinical plausibility reward model. Using the augmented feedback dataset $D\Gamma$, we train the reward model by minimizing the mean square error as follows:

$$L\Gamma(\phi) = \sum_{j \in D\Gamma_s} (y_{ej} - \Gamma\phi(x_{ej}, ec_j))^2 + \lambda_r \sum_{i \in D\Gamma_r} (y_i - \Gamma\phi(x_i, ec_i))^2, \quad (2)$$

where λ_r is a hyper-parameter that controls the contribution of real images in training the reward functions, and ϕ is the parameter of the reward model.

Step 3: Clinically-informed Finetuning. In the final step, we refine the diffusion model by leveraging the pathologist feedback. Specifically, we incorporate domain knowledge into the model by utilizing the reward model Γ in the finetuning objective. Following [20], we use a reward-weighted negative log-likelihood (NLL) objective, i.e.,

$$L(\theta, \phi_b) = E_{(x, ec) \sim D_s} [-\Gamma\phi_b(x, ec) \cdot \log(p_\theta(x|ec))] + \beta_r \cdot E_{(x,c) \sim D_r} [-\log(p_\theta(x|c))], \quad (3)$$

to finetune the conditional diffusion model, where β_r is a hyper-parameter and ϕ_b is the reward model parameters obtained by minimizing (2) in Step 2. The finetuning objective in (3) incorporates the

pathologist knowledge through the reward model, which predicts the pathologist evaluation of the synthetic images that the model generates as it updates its parameters θ . The reward-weighted objective penalizes the generation of images that do not align with the pathologist preferences; hence we expect that the finetuned model will be less likely to generate clinically implausible synthetic images.

BONUS STEP: FEEDBACK-DRIVEN GENERATION OF NEW CLINICAL CONCEPTS

Besides refining generative models for clinical plausibility, pathologist feedback can be used to incorporate novel clinical concepts into the generative process that were not initially labeled in the real dataset. This could allow generative models to continuously adapt in changing clinical environments. For instance, in our bone marrow image generation setup, pathologist feedback can refine image generation by introducing new sub-types of the original cell types in C , as illustrated in Figure 4.3. Instead of collecting pathologist feedback that is limited to clinical plausibility, we also collect their annotation of new cell sub-types (e.g., segmented and band variants of Neutrophil cell types). Next, we train an auxiliary model $\Gamma_{\rho}(x)$ with parameter ρ to classify the new sub-types based on the pathologist annotations. Finally, we finetune the conditional diffusion model through a combined loss function, that incorporates the two forms of pathologist feedback, i.e., annotations of new cell types and clinical plausibility.

PATHOLOGIST FEEDBACK VS. AUTOMATED FEEDBACK

To assess the added value of human feedback, we consider a baseline where the generative model is supplemented with automatically generated feedback on clinical plausibility. To this end, we implement a baseline based on classifier-guided diffusion, where a classifier serves as an automatic feedback signal that deems a synthetic image implausible if it does not match the corresponding cell type. For this baseline, we train an auxiliary classifier to predict the cell type c of an image x , and then we incorporate the gradient of the log-likelihood of this classifier in the training objective as described in the referenced literature (See the Supplementary material for implementation details).

4.3 RELATED WORK

Before delving into the experimental results, we discuss three strands of literature that are related to our synthetic data generation framework. These include generative modeling for synthetic clinical data, evaluation of generative models and learning from human feedback.

GENERATIVE MODELING OF SYNTHETIC MEDICAL IMAGES

The dominant approach for synthesizing medical images is to train or finetune a generative model, such as a Variational Autoencoder (VAE) [24], a Generative Adversarial Network (GAN) [25], or a diffusion model [6, 23], using a sufficiently large sample of images from the desired modality. Owing to their recent success in achieving state-of-the-art results in high-fidelity image synthesis [23], diffusion probabilistic models have become the model of choice for medical image synthesis applications [12, 26–29]. In [12], the Stable Diffusion model—an open-source pretrained diffusion model—was used to generate synthetic X-ray images, and in [28] it was shown that diffusion models can synthesize high-quality Magnetic Resonance Images (MRI) and Computed Tomography (CT) images. [27] used latent diffusion models to generate synthetic images from high-resolution 3D brain images. All of these models are trained with the standard likelihood objective and the synthetic images are typically evaluated through downstream classification tasks or generic, domain-agnostic metrics for image fidelity. To the best of our knowledge, none of the previous studies have explored a human-AI collaboration approach to synthetic image generation or incorporated clinical knowledge into generative models of medical images.

EVALUATION OF SYNTHETIC IMAGES IN THE MEDICAL DOMAIN AND BEYOND

Unlike discriminative modeling (i.e., predictive modeling) where model accuracy can be straightforwardly evaluated by comparing the model predictions with ground-truth labels in a testing set, evaluating the quality of generative models can be quite challenging since we do not have a “ground-truth” for defining what makes a synthetic sample is of high or low quality. Devising a generic score to evaluate a generative

model can be tricky since there are many potential modes of failure [22]. Consequently, it is essential to design robust multidimensional scores that capture the most relevant failure modes for a given application. Recently, there have been various attempts at defining domain-specific scores [30–32] as well as generic scores for evaluating the quality of synthetic images. Examples include the FID score which is based on a distributional distance between real and synthetic images [33]. Other examples for sample level evaluation metrics include the precision and recall metrics [21] which check if synthetic data resides in the support of the real data distribution. However, these scores do not encode clinical domain knowledge, which is critical for identifying failures in generating clinically meaningful images. Traditional scores of medical image quality include signal- and contrast-to-noise ratio [34–36], mean structural similarity [37]. These scores are typically applied to real images and cannot be repurposed to judge the generative capacity of a synthetic data model in a meaningful way. The lack of an automated score for detecting clinically implausible synthetic medical images is a key motivation for our work. We believe that the most reliable way to assess the quality of a synthetic image is to have it evaluated by an expert pathologist. From this perspective, the reward model Γ in Section 2.2 can be thought of as a data-driven score for image quality trained using pathologist evaluations.

LEARNING FROM HUMAN FEEDBACK

The success of many modern generative models can be attributed in part to finetuning using feedback solicited from human annotators. The utilization of human feedback in model finetuning is very common in natural language processing applications, particularly in finetuning of large language models (LLMs). Examples for applications where human feedback was applied include translation [38], web question-answering [39] and instruction tuning [40–42]. The key idea in these applications is that by asking a human annotator to rate different responses from the same model, one can use such annotations to finetune the model to align with human preference. Similar ideas have been applied to align computer vision models with human preferences [20, 43–46]. In the context of our application, the goal is to align the outputs of a generative model with the preferences of pathologists, which are naturally aligned with clinical domain

knowledge. Our finetuning objective builds on the recent work in [20] and [45], which use human feedback to align text-prompts with generated images using a reward-weighted likelihood score.

4.4 EXPERIMENTS

In this Section, we conduct a series of experiments to evaluate the utility of pathologist feedback in improving the quality of synthetic medical images. In the next Subsection, we start by providing a detailed description of the single-cell bone marrow image dataset used in our experiments.

BONE MARROW CELLS DATASET

In all experiments, we used a dataset of hematopathologist consensus-annotated single-cell images extracted from bone marrow aspirate (BMA) whole slide images. The images were obtained from the clinical archives of an academic medical center. The dataset comprised 2,048 images, with the images evenly distributed across 16 morphological classes (cell types), with 160 images per class (Table 2). These classes encompass varied cell types found in a standard bone marrow differential. The dataset covers the complete maturation spectrum of Erythroid and Neutrophil cells, from Proerythroblast to mature Erythrocyte and from Myeloid blast to mature Neutrophils, respectively. The dataset also differentiates mature Eosinophils with segmented nuclei from immature Eosinophils and features Monocytes, Basophils, and Mast cells. Bone marrow cell counting and differentiating between various cell types is a complex task that poses challenges even for experienced hematologists. Hence, we expect pathologist feedback to significantly improve the quality of bone marrow image synthesis.

SYNTHETIC DATA GENERATION AND PATHOLOGIST FEEDBACK COLLECTION

We used a conditional diffusion model trained on real images (64×64 pixels in size) to generate synthetic image patches. Training was conducted using 128 images per cell type, with 32 images per cell type held out for testing and evaluating all performance metrics. For the reward model $\Gamma(x, c)$, we used a ResNeXt-

50 model [47] pre-trained on a cell type classification task to obtain embeddings for individual images, and then concatenated the embeddings with one-hot encoded identifiers of image class (cell type) as inputs to a feed-forward neural network that predicts clinical plausibility. Further details on model architectures and selected hyper-parameters are provided in the Appendix.

We collected feedback from an expert pathologist on 3,936 synthetic images generated from the diffusion model. The pathologist identified most of these images as implausible—the rate of implausible images was as high as 85% for some cell types (e.g., Basophil cells, see Table 2). After training the reward model using pathologist feedback, we finetune the diffusion model as described in Section 2.

4.5 RESULTS

EXPERT EVALUATION OF SYNTHETIC DATA QUALITY

To evaluate the impact of pathologist feedback on the generated synthetic data, we created two synthetic datasets: a sample from the diffusion model (before finetuning with pathologist feedback) and a sample from the finetuned version of the model after incorporating the pathologist feedback. Each dataset comprised 400 images (25 images per cell type). An expert pathologist was asked to evaluate the two samples and label each image as plausible or implausible (in a manner similar to the feedback collection process). Table 3 lists the fraction of clinically plausible images per cell type for the two synthetic datasets (before and after finetuning using the pathologist feedback) as evaluated by an expert hematopathologist. As we can see, the pathologist feedback leads to a significant boost in the quality of synthetic images across all cell types, increasing the average rate of clinical plausibility from 0.21 to 0.75. Note that in this experiment, the human evaluator emulates the reward function $\Gamma(x, c)$. Hence, the improved performance of the finetuned model indicates success in learning the pathologist preferences. Evaluating synthetic data using fidelity & diversity scores. In addition to expert evaluation, we also evaluated the two synthetic datasets generated in the previous experiment using standard metrics for evaluating generative models. We

considered the Precision, Recall and Coverage metrics [21, 22]. Precision measures the fraction of synthetic samples that resides in the support of the real data distribution and is used to measure fidelity. Recall and Coverage measure the “diversity” of synthetic samples, i.e., the fraction of real images that are represented in the output of a generative model. In addition to the two synthetic datasets (with and without feedback) generated in the previous experiment, we also evaluated a third synthetic dataset finetuned using the automatic feedback (classifier-guidance) approach described in Section 2.5. The results in Table 4 show that pathologist feedback improves the quality of synthetic data compared to the two baselines across all metrics under consideration. Interestingly, we see that feedback not improves fidelity of synthetic images, but it also improves the diversity of samples, which was not one the criteria considered in the pathologist feedback.

DOWNSTREAM MODELING WITH SYNTHETIC DATA

Morphology-based classification of cells is a key step in the diagnosis of hematologic malignancies. We evaluated the utility of synthetic medical images in training a cell-type classification model. In this experiment, we train a ResNext-50 model to classify the 16 cell types using real data, synthetic data from the pretrained model with no feedback, and synthetic data from the model finetuned with pathologist feedback. To ensure a fair comparison, we created synthetic datasets consisting of 128 images per cell type, matching the size of the real data. The classification accuracy of all models was then tested on the held-out real dataset, containing 32 images per cell type. Results are shown in Table 5. Unsurprisingly, the model trained on real data demonstrated superior performance across all accuracy metrics, exhibiting a significant gap compared to the model trained on synthetic data without human feedback. However, incorporating pathologist feedback helped narrow this gap and improved the quality of synthetic data to the point where the resulting classifiers only slightly underperformed compared to the one trained on real data. To evaluate the marginal value of human feedback, we also considered two ablated versions of our synthetic data generation process. These included a synthetic dataset generated using automatic feedback (Section 2.5) as well as a reward model trained using the pathologist feedback on synthetic data only (i.e., dataset D Γ s)

without real data augmentation (Section 2.2). The results in Table 5 show that automatic feedback only marginally improves performance, which aligns with the results in Table 4. Real data augmentation (i.e., finetuning on $D \Gamma_s + D \Gamma_r$) slightly improves classification accuracy, but most of the performance gains are achieved by finetuning on the pathologist-labeled dataset $D \Gamma_s$.

IMPACT OF THE NUMBER OF FEEDBACK POINTS

How much human feedback is necessary to align the generative model with the preferences of pathologists? In Figure 4.4, we analyze the effect of different amounts of pathologist-labeled synthetic images on training the reward model. We explore four scenarios: 0%, 10%, 50%, and 100% of the 3,936 synthetic images labeled by the pathologist. For each scenario, we fine-tune the pretrained model using a reward model trained with the corresponding fraction of synthetic images. We then repeat the cell classification experiment to evaluate the accuracy of the classifiers trained using synthetic data generated in these four scenarios. Figure 4.4(a) shows that as the number of pathologist-labeled synthetic images increases, all accuracy metrics increase to improve over the pretrained model performance and become closer to the accuracy of training on real data. Additionally, we see that even a modest amount of feedback (e.g., 10% of pathologist-labeled synthetic images) can have a significant impact on performance. Figure 4.4(b) also show the qualitative improvement in the quality of synthetic images as the amount of human feedback increases.

INCORPORATING NEW CLINICAL CONCEPTS USING PATHOLOGIST FEEDBACK

Finally, we evaluate the feedback driven generation approach outlined in Section 2.4. Here, our goal is not only to leverage pathologist feedback for rating the plausibility of synthetic images, but also to harness their expertise in providing additional annotations that can enable the conditional diffusion model to generate more refined categories of bone marrow image patches, e.g., abnormal cell types that develop from preexisting normal cell types. Hence, refining the generative model to synthesize new cell subtypes can help continuously update the model to capture new pathological cells and build downstream diagnostic

models. In this experiment, we focus on finetuning the model to distinguish between band Neutrophils and segmented Neutrophils (Figure 4.5(c)). These two sub-categories are often lumped together in bone marrow cell typing, as was the case in our dataset (see Table 2). However, in many clinical settings, differentiating between the two subtypes is essential. For instance, a high percentage of band Neutrophils can indicate an acute infection, inflammation, or other pathological conditions. We collected pathologist annotations of band and segmented Neutrophils and trained a subtype classifier to augment the plausibility reward as described in Section 2.4. The finetuned model was able to condition on the new classes and generate plausible samples of the two Neutrophil subtypes (Figure 4.5(c)), while retaining the classification accuracy with respect to the new classes (See Appendix for detailed results).

4.6 CONCLUSIONS

Synthetic data generation holds great potential for facilitating the sharing of clinical data and enriching rare diseases datasets. However, existing generative models and evaluation metrics lack the ability to incorporate clinical knowledge. Consequently, they often fall short in producing clinically plausible and valuable images. This paper introduces a pathologist-in-the-loop framework for generating clinically plausible synthetic medical images. Our framework involves finetuning a pretrained generative model through feedback provided by pathologists, thereby aligning the synthetic data generation process with clinical expertise. Through the evaluation of synthetic bone marrow patches by expert hematopathologists, leveraging thousands of feedback points, we demonstrate that human input significantly enhances the quality of synthetic images. These results underscore the importance of incorporating human feedback in clinical applications, particularly when generative models encounter challenges in capturing nuanced domain knowledge solely from raw data.

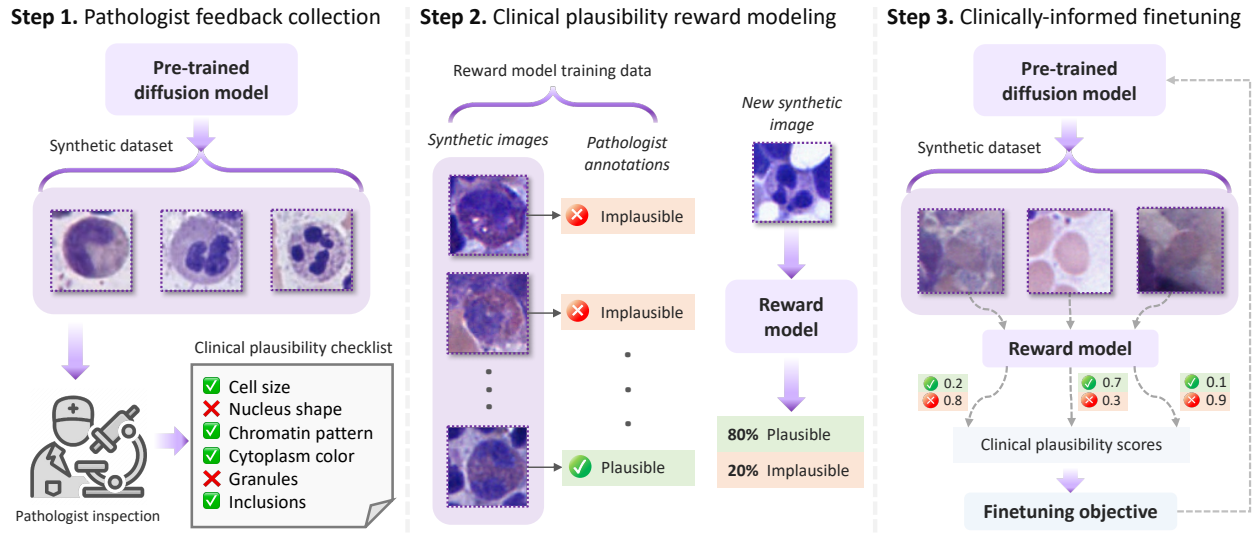


Figure 4.1 Overview of our pathologist-in-the-loop synthetic data generation framework

(1) Step 1, a synthetic dataset is sampled from a generative model pretrained using a dataset of real medical images. The dataset is then inspected by a pathologist who examines each image to determine its plausibility based on a set of criteria. For each image, the pathologist provides binary feedback, labeling a synthetic image as "1" if it fails to meet all the plausibility criteria. (2) In Step 2, the synthetic images and pathologist feedback obtained in Step 1 are used to train a reward model that predicts human feedback on new images. (3) Finally, the generative model is finetuned via an objective function that uses the reward model to incentivize the generation of clinically plausible images.

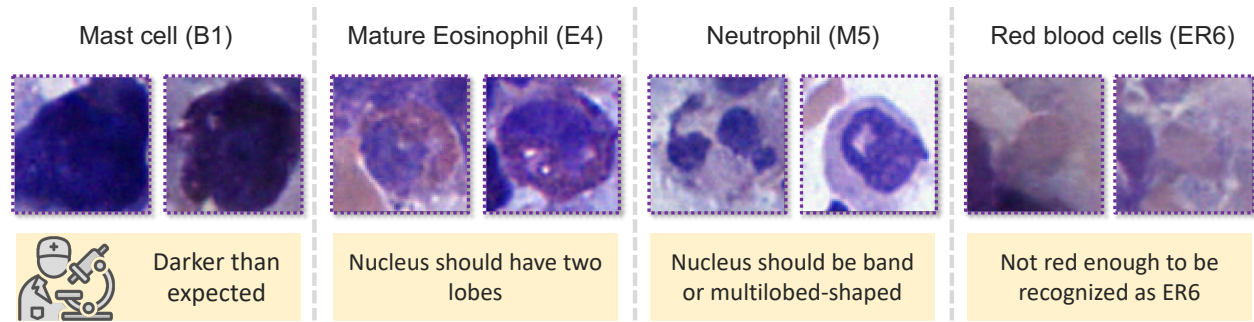


Figure 4.2 Samples for biologically implausible synthetic images

On the bottom panels, we show pathologist evaluations detailing the reasons for their implausibility.

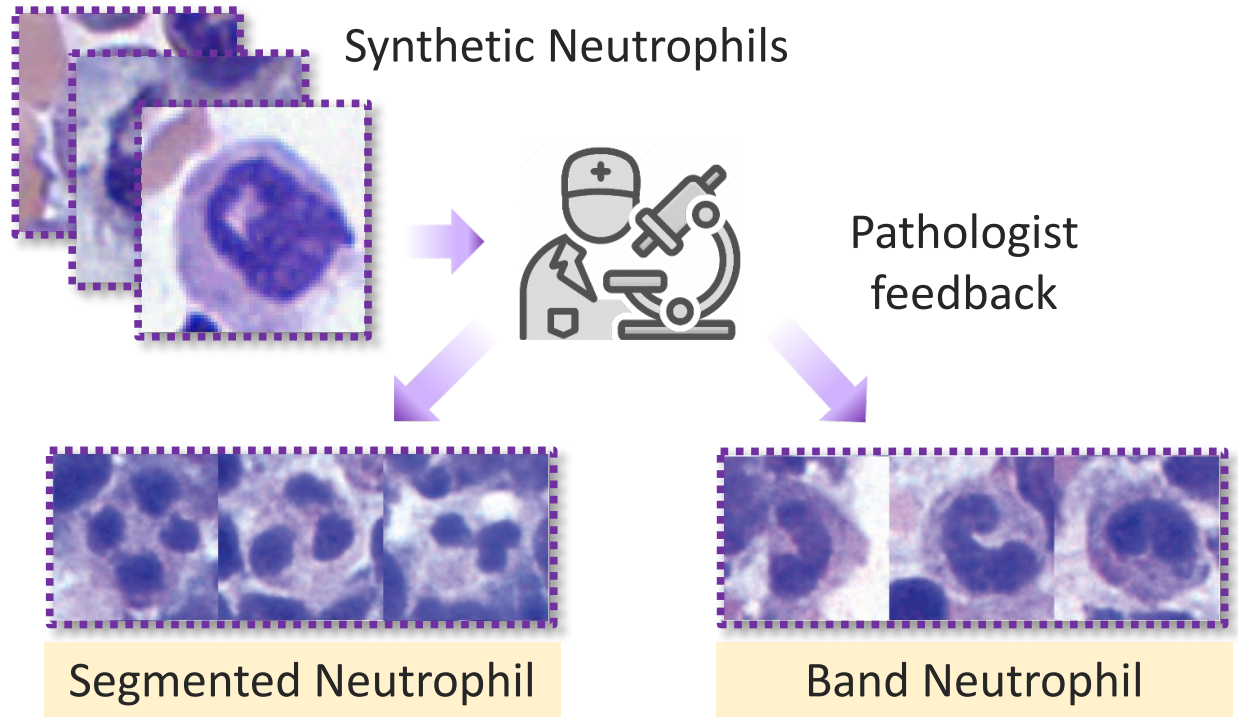


Figure 4.3 Generation of refined cell sub-types.

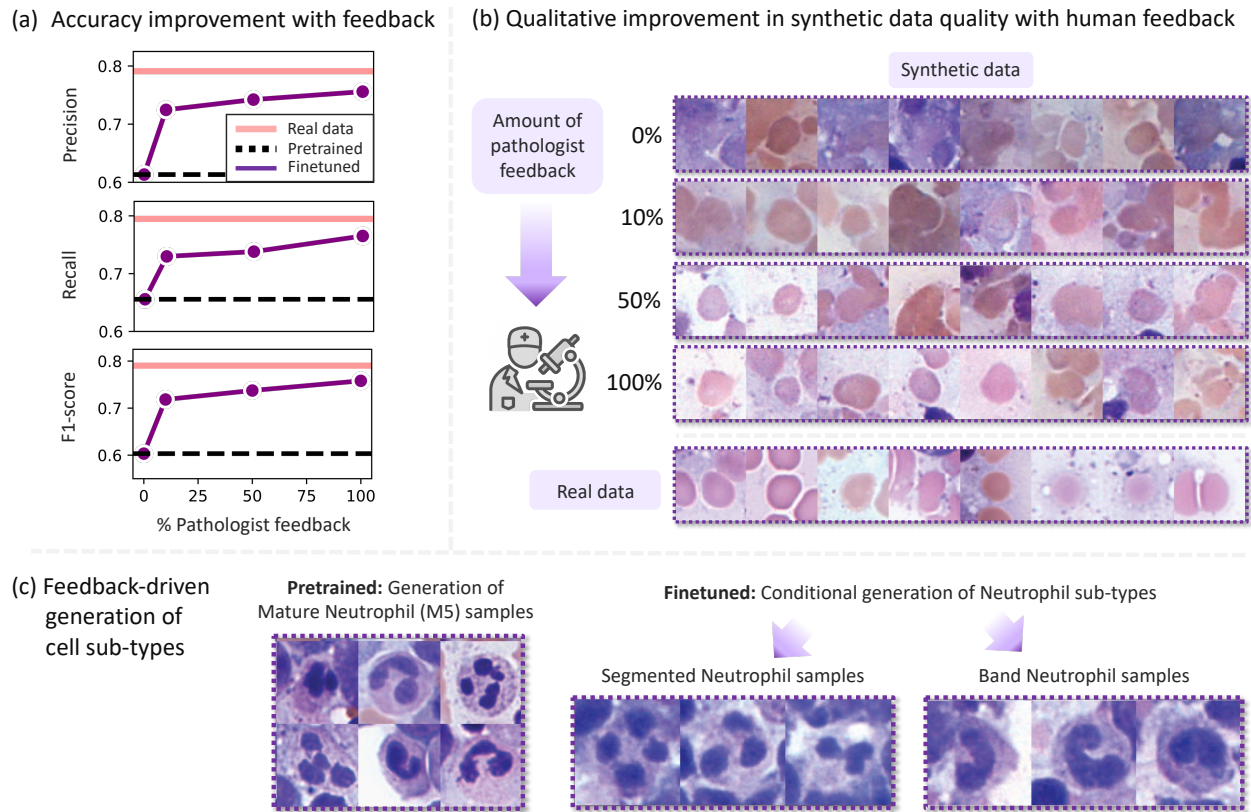


Figure 4.4 Quantitative and qualitative impact of pathologist feedback on synthetic images

(a) Accuracy of cell-types classifiers trained on synthetic data with varying amounts of pathologist feedback. (b) Visual inspection of random synthetic samples from diffusion models finetuned with varying amounts of pathologist feedback. (c) Feedback-driven conditional generation of new (segmented and band) subtypes of Neutrophil cells.

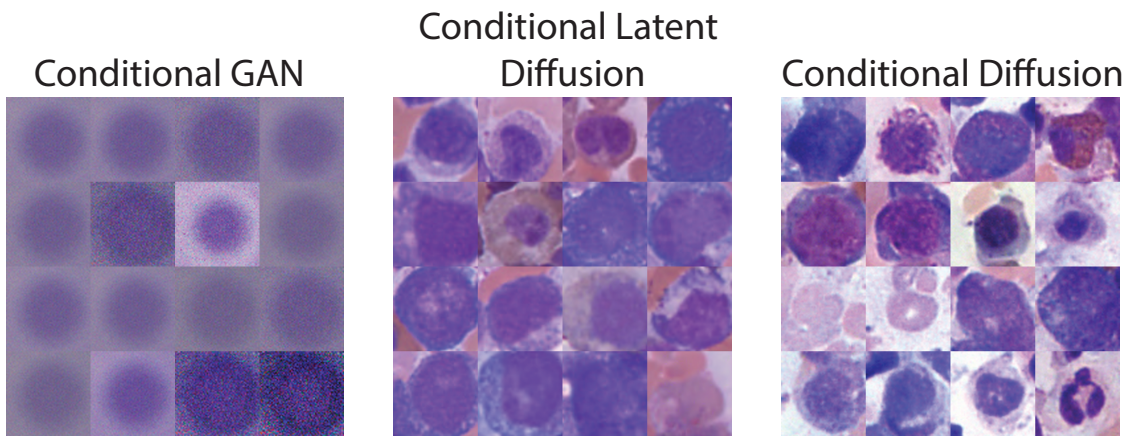


Figure 4.5 Representative samples from different baseline models

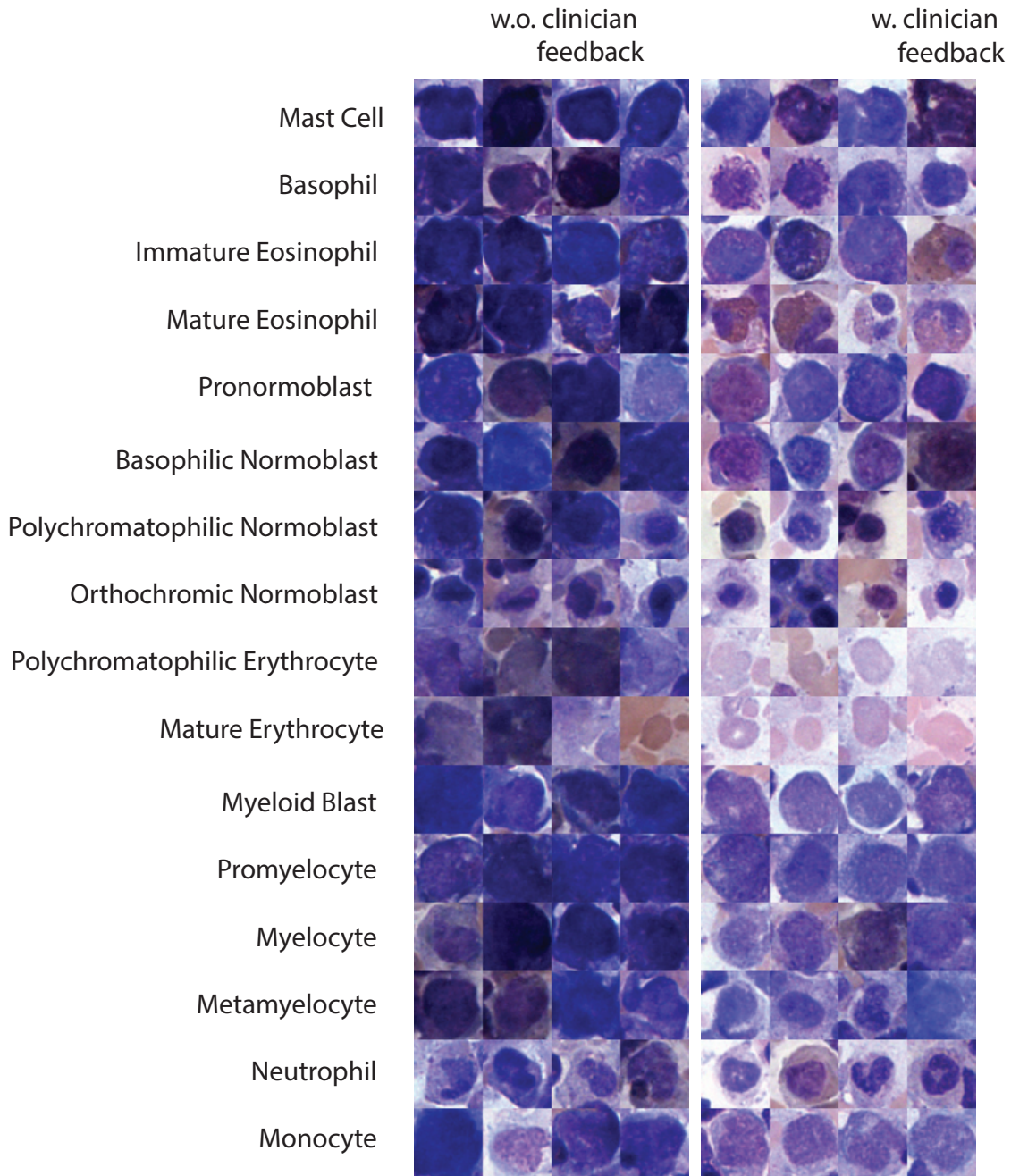


Figure 4.6 Representative samples from the conditional diffusion model before (left) and after (right) incorporating pathologist feedback.

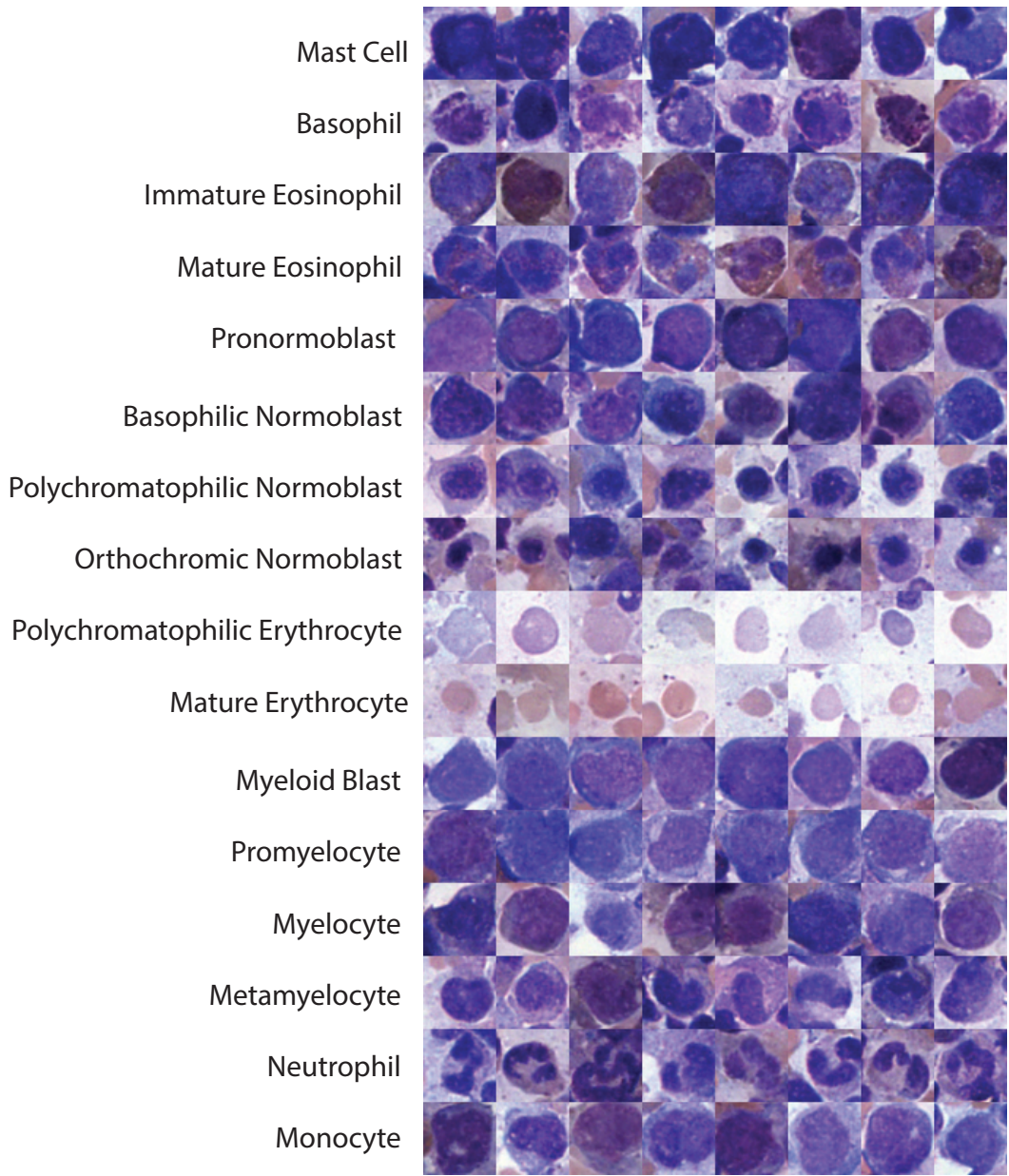


Figure 4.7 Representative samples from the conditional diffusion model.

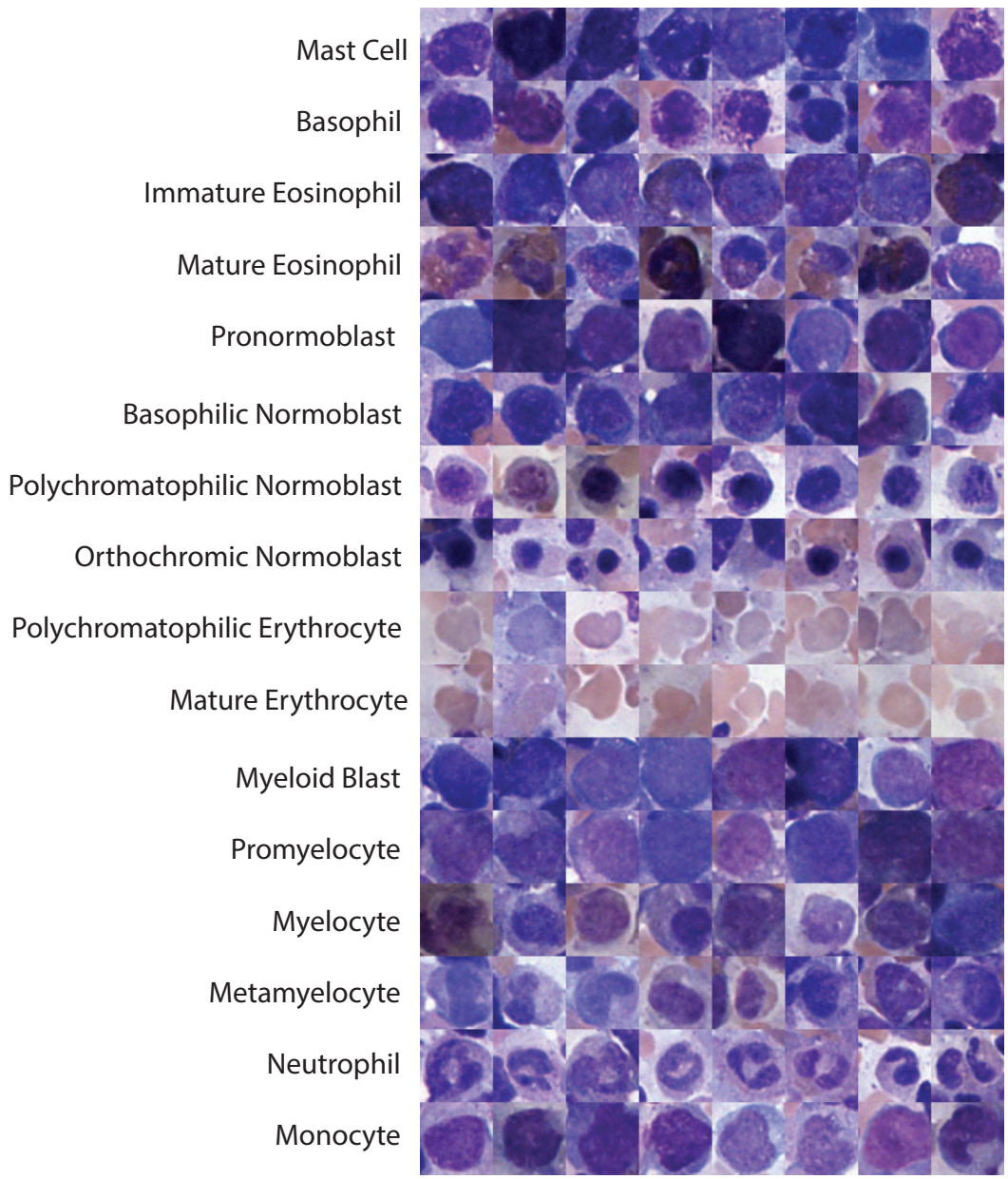


Figure 4.8 Representative samples from the finetuned model with 10% of the pathologist feedback points.

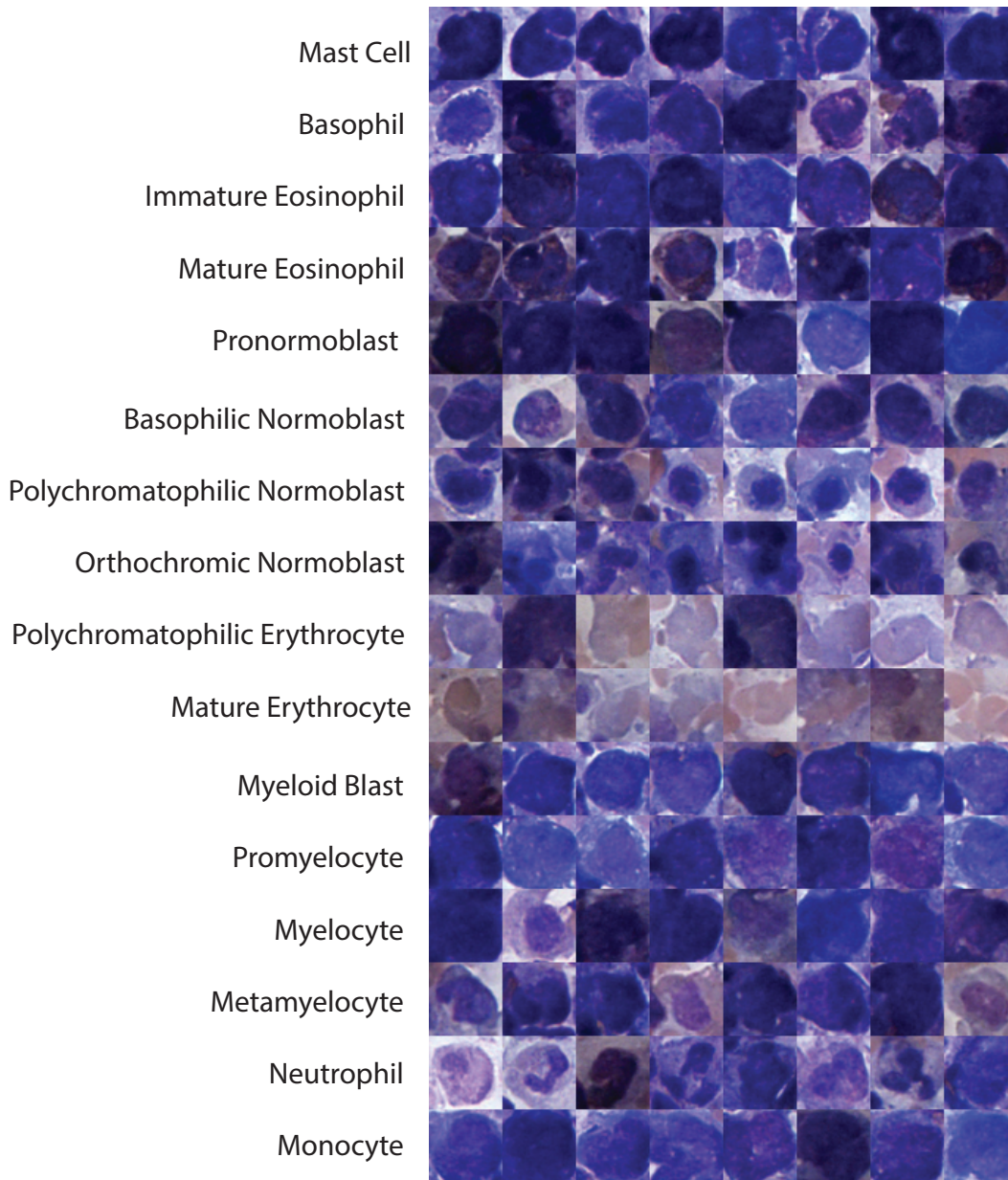


Figure 4.9 Representative samples from the finetuned model with 50% of the pathologist feedback points.

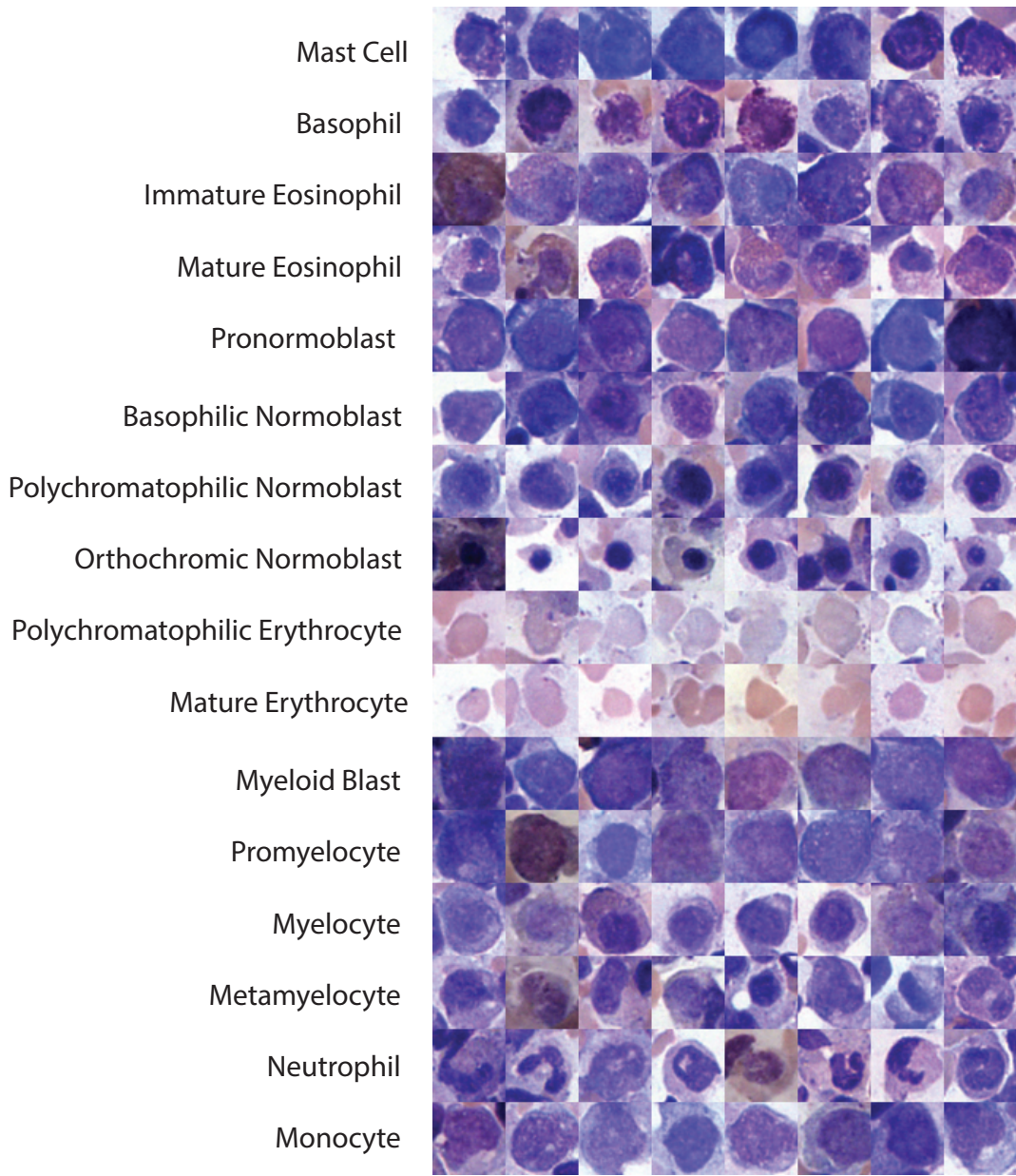


Figure 4.10 Representative samples from the finetuned model with 100% of the pathologist feedback points

Table 4.1 Pathologist Evaluation Criteria

| Criteria | |
|------------------------------------|----------------------|
| 1. Cell size | 5. Chromatin pattern |
| 2. Nucleus shape and size | 6. Inclusions |
| 3. Nucleus-to-cytoplasm ratio | 7. Granules |
| 4. Cytoplasm color and consistency | |

Table 4.2 Breakdown of bone marrow image patches by morphological cell type and pathologist feedback.

| Morphological cell type | Code | Training | Testing | Synthetic data | Plausible | Implausible |
|--------------------------------|-------------|-----------------|----------------|-----------------------|------------------|--------------------|
| Mast Cell | B1 | 128 | 32 | 213 | 72 | 141 |
| Basophil | B2 | 128 | 32 | 214 | 29 | 185 |
| Immature Eosinophil | E1 | 128 | 32 | 213 | 49 | 164 |
| Mature Eosinophil | E4 | 128 | 32 | 224 | 53 | 171 |
| Pronormoblast | ER1 | 128 | 32 | 256 | 97 | 159 |
| Basophilic Normoblast | ER2 | 128 | 32 | 256 | 49 | 207 |
| Polychromatophilic Normoblast | ER3 | 128 | 32 | 256 | 129 | 127 |
| Orthochromic Normoblast | ER4 | 128 | 32 | 256 | 118 | 138 |
| Polychromatophilic Erythrocyte | ER5 | 128 | 32 | 256 | 141 | 115 |
| Mature Erythrocyte | ER6 | 128 | 32 | 256 | 120 | 136 |
| Myeloid Blast | M1 | 128 | 32 | 256 | 138 | 118 |
| Promyelocyte | M2 | 128 | 32 | 256 | 107 | 149 |
| Myelocyte | M3 | 128 | 32 | 256 | 131 | 125 |
| Metamyelocyte | M4 | 128 | 32 | 256 | 88 | 168 |
| Mature Neutrophil | M5 | 128 | 32 | 256 | 80 | 176 |
| Monocyte | MO2 | 128 | 32 | 256 | 83 | 173 |

Table 4.3 Expert evaluation of synthetic data.

| | Clinically Plausible | |
|-----|----------------------|----------------|
| | No feedback | Path. feedback |
| B1 | 0.40 | 0.92 |
| B2 | 0.16 | 1.00 |
| E1 | 0.12 | 0.80 |
| E4 | 0.24 | 0.52 |
| ER1 | 0.44 | 0.96 |
| ER2 | 0.28 | 0.64 |
| ER3 | 0.24 | 0.68 |
| ER4 | 0.20 | 0.84 |
| ER5 | 0.20 | 0.76 |
| ER6 | 0.20 | 0.96 |
| M1 | 0.20 | 0.84 |
| M2 | 0.20 | 0.64 |
| M3 | 0.08 | 0.56 |
| M4 | 0.20 | 0.84 |
| M5 | 0.08 | 0.68 |
| MO2 | 0.12 | 0.40 |

Table 4.4 Evaluation of synthetic data using fidelity and diversity metrics.

| Training data | Precision | Recall | Coverage |
|----------------------------|------------------|---------------|-----------------|
| Synthetic (no feedback) | 68.06 | 52.00 | 56.98 |
| Synthetic (auto. feedback) | 74.80 | 43.90 | 61.33 |
| Synthetic (with feedback) | 81.01 | 56.74 | 84.57 |

Table 4.5 Accuracy of classifiers trained on real and synthetic data.

| Training data | F1 | Accuracy | Precision | Recall |
|--|--------------|-----------------|------------------|---------------|
| Synthetic (no feedback) | 60.33 | 95.17 | 61.33 | 65.56 |
| Synthetic (auto. feedback) | 63.47 | 95.58 | 64.64 | 65.68 |
| Synthetic (with feedback, \mathcal{D}^{Γ_s}) | 71.80 | 96.41 | 71.29 | 74.51 |
| Synthetic (with feedback, \mathcal{D}^{Γ}) | 75.80 | 96.95 | 75.59 | 76.51 |
| Real | 79.03 | 97.39 | 79.10 | 79.47 |

Table 4.6 Performance comparison for baseline generative models in the cell classification task.

| Training Data | Test Data | AUC | F1 | Precision | Recall |
|------------------------------------|------------------|------------|-----------|------------------|---------------|
| Conditional diffusion model | Real | 0.929482 | 0.567672 | 0.604237 | 0.604146 |
| Conditional GAN model | Real | 0.413551 | 0.003225 | 0.002186 | 0.015804 |
| Conditional Latent diffusion model | Real | 0.566388 | 0.008565 | 0.048372 | 0.051456 |

Table 4.7 Model finetuned with the new subtypes of Neutrophil cells.

| AUC score | Accuracy | Precision | Recall |
|------------------|-----------------|------------------|---------------|
| 81.90 | 71.88 | 67.95 | 82.81 |

4.7 REFERENCES

- [1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Highresolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022.
- [2] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In International Conference on Machine Learning, pages 8821–8831. PMLR, 2021.
- [3] Aditya Ramesh, Prfulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.
- [4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33:6840–6851, 2020.
- [5] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 32, 2019.
- [6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.
- [7] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5(6):493–497, 2021.
- [8] Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, Gorka Epelde, et al. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. JMIR medical informatics, 8(7):e18910, 2020.
- [9] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis, 71:102062, 2021.
- [10] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep learning for medical image

- processing: Overview, challenges and the future. *Classification in BioApps: Automation of Decision Making*, pages 323–350, 2018.
- [17] Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. arXiv preprint arXiv:2211.01323, 2022.
- [18] Hazrat Ali, Shafaq Murad, and Zubair Shah. Spot the fake lungs: Generating synthetic medical images using neural diffusion models. In *Artificial Intelligence and Cognitive Science: 30th Irish Conference, AICS 2022, Munster, Ireland, December 8–9, 2022, Revised Selected Papers*, pages 32–39. Springer, 2023.
- [19] Faisal Mahmood, Daniel Borders, Richard J Chen, Gregory N McKay, Kevan J Salimian,
- [20] Alexander Baras, and Nicholas J Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE transactions on medical imaging*, 39(11):3257–3267, 2019.
- [21] Hajar Emami, Ming Dong, and Carri K Glide-Hurst. Attention-guided generative adversarial network to address atypical anatomy in synthetic ct generation. In *2020 IEEE 21st international conference on information reuse and integration for data science (IRI)*, pages 188–193. IEEE, 2020.
- [22] Lisa C Adams, Felix Busch, Daniel Truhn, Marcus R Makowski, Hugo JWL Aerts, and Keno K Bressemer. What does dall-e 2 know about radiology? *Journal of Medical Internet Research*, 25:e43110, 2023.
- [23] Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. arXiv preprint arXiv:2210.04133, 2022.
- [24] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman,
- [25] Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402, 2022. 10

- [27] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593, 2019.
- [28] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- [29] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192, 2023.
- [30] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.
- [31] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.
- [32] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.
- [34] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [35] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarbürger, Christiane Kuhl, Tianci Wang, Tianyu Han, Sven Nebelung, Jakob Nikolas Kather, et al. Diffusion probabilistic models beat gans on medical images. arXiv preprint arXiv:2212.07501, 2022.

- [36] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 117–126. Springer, 2022.
- [37] Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarbuerger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baessler, Sebastian Foersch, et al. Medical diffusion–denoising diffusion probabilistic models for 3d medical image generation. *arXiv preprint arXiv:2211.03364*, 2022.
- [38] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*, 2022.
- [39] Tonghe Wang, Yang Lei, Zhen Tian, Xue Dong, Yingzi Liu, Xiaojun Jiang, Walter J Curran, Tian Liu, Hui-Kuo Shu, and Xiaofeng Yang. Deep learning-based image quality improvement for low-dose computed tomography simulation in radiation therapy. *Journal of Medical Imaging*, 6(4):043504–043504, 2019.
- [40] Ilona Urbaniak and Marcin Wolter. Quality assessment of compressed and resized medical images based on pattern recognition using a convolutional neural network. *Communications in Nonlinear Science and Numerical Simulation*, 95:105582, 2021.
- [41] Daniel Lévy and Arzav Jain. Breast mass classification from mammograms using deep convolutional neural networks. *arXiv preprint arXiv:1612.00542*, 2016.
- [42] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [43] Harry Nyquist. Certain factors affecting telegraph speed. *Transactions of the American Institute of Electrical Engineers*, 43:412–422, 1924.

- [44] Stefan Winkler and Praveen Mohandas. The evolution of video quality measurement: From psnr to hybrid metrics. *IEEE transactions on Broadcasting*, 54(3):660–668, 2008. 11
- [45] WA Edelstein, PA Bottomley, HR Hart, and LS Smith. Signal, noise, and contrast in nuclear magnetic resonance (nmr) imaging. *J Comput Assist Tomogr*, 7(3):391–401, 1983.
- [46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [47] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [48] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [49] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [50] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- [51] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [52] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.
- [53] Cheng-Kang Ted Chao and Yotam Gingold. Text-guided image-and-shape editing and generation: A short survey. *arXiv preprint arXiv:2304.09244*, 2023.

- [54] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. arXiv preprint arXiv:2303.14420, 2023.
- [55] Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767, 2023.
- [56] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017.
- [57] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In International Conference on Machine Learning, pages 8162–8171. PMLR, 2021.
- [58] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
- [59] Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood, The Journal of the American Society of Hematology*, 138(20):1917–1927, 2021.

5 Chapter 5: Spatial cell-type enrichment predicts mouse brain connectivity

Shenghuan Sun^{1,*}, Justin Torok^{1,*}, Christopher Mezas², Daren Ma¹, and Ashish Raj^{1,**}

¹ Department of Radiology, University of California, San Francisco – San Francisco, CA, United States

² Cold Spring Harbor Laboratory – Cold Spring Harbor, NY, United States

* These authors contributed equally to this work

** Lead Contact: Ashish Raj (ashish.raj@ucsf.edu)

5.1 ABSTRACT

A fundamental neuroscience topic is the link between the brain's molecular, cellular and cytoarchitectonic properties and structural connectivity (SC). Recent studies relate inter-regional connectivity to gene expression, but the relationship to regional cell-type distributions remains understudied. Here, we utilize whole-brain mapping of neuronal and non-neuronal subtypes via the Matrix Inversion and Subset Selection (MISS) algorithm to model inter-regional connectivity as a function of regional cell-type composition with machine learning. We deployed random forest algorithms for predicting connectivity from cell type densities, demonstrating surprisingly strong prediction accuracy of cell types in general, and particular non-neuronal cells such as oligodendrocytes. We found evidence of a strong distance-dependency in the cell-connectivity relationship, with layer-specific excitatory neurons contributing the most for long-range connectivity, while vascular and astroglia were salient for short-range connections. Our results demonstrate a link between cell types and connectivity, providing a roadmap for examining this relationship in other species, including humans.

5.2 INTRODUCTION

The structural connectome, which represents the density of physical projections between brain regions and is measured by such techniques as viral tracing and diffusion tensor imaging, is a coarse wiring diagram of the central nervous system (CNS)[1–4]. Complex molecular processes during embryonic development encourage the formation of connections between brain regions, and later postnatal pruning results in structural connectomes with a remarkable degree of conservation between healthy individuals. There is a strong interest in gaining a rigorous measure of how gene expression and cell type composition of brain regions relate to connectivity[5, 6], which can deepen our understanding of how brain circuits mature during the development of the CNS and how they are disrupted in neurodegenerative diseases, among other areas of inquiry. While the correlation between regional gene expression and connectivity is well established in mice[5], [7–9] and humans[10–12], the methods used to determine this association are mainly correlative or analytic. Correlation or regression with high-dimensional input feature spaces carries a risk of overfitting, and, as a result, often fails to generalize to unseen data[13]. As an alternative approach, Ji, et al.[14] applied random forest methods to predict the presence or absence of brain connectivity from gene expression with high accuracy, but did not attempt to predict the amount of connectivity density. Other groups[5], [14] report that connected regions tend to have higher correlated gene expression patterns than regions that are not, which naturally raises the question of whether the connected brain regions share common cell types. A step in this direction was taken by Huang et al., who demonstrated BRICseq, a powerful technique capable of mapping individual axonal projections along with the neuronal subtypes to which they belong[15]. However, their methodology has not yet been scaled up to produce a dataset of comparable spatial coverage to the Allen Mouse Brain Connectivity Atlas (AMBCA)[2], which is perhaps the most thorough mesoscale connectome currently available. Therefore, it is not yet clear how distributions of different types of cells - the fundamental units of connectivity - relate to the whole-brain connectome, nor have any unbiased, data-driven methods been applied to attempt to reconstruct the mouse connectome from regional cell type densities. Although the success of prior studies in using gene expression-based markers to predict

connectivity suggests that cell type distributions will also be predictive of the connectome, the paucity of available whole-brain cell type distributions has made it difficult to test the hypothesis. Indeed, before the advent of spatial transcriptomics and single cell gene profiling the question would have been impossible to answer quantitatively on the whole brain level.

Here, we take advantage of these emerging technologies to develop a comprehensive data-driven computational machinery needed to address this question. We first implemented an algorithm to produce regional cell type enrichment from spatially resolved gene expression data, following a specialized method we have recently developed called Matrix Inversion and Subset Selection (MISS)[16]. This method is essentially a cell type deconvolution algorithm that was shown to faithfully reproduce cell type distributions in the mouse brain using Allen Gene Expression Atlas (AGEA)[17] and publicly available single-cell RNA sequencing data[18], [19]. Then, using inferred cell type enrichment distributions as input features, we applied a number of machine learning methods to reconstruct the mesoscale mouse structural connectome from AMBCA2. Among all the models tested, the random forest (RF) algorithm outperformed other approaches at predicting both the presence or absence of a connection between any given region pair as well as the actual connectivity density values.

We were able to predict the structural connectome with a surprisingly high level of accuracy, despite that the fact that the construction of fiber connectivity is a highly complex and iterative biological process with many determinants not strictly captured by constituent cell types. We replicated our findings with a second, different set of cell type distributions inferred by MISS. Despite the two datasets having a widely different number of individual cell types, both achieved almost identical performance on the connectivity prediction task, indicating that our approach is not an artifact of a particular input feature set. Our results quantitatively demonstrate that regional cell type distributions can explain most of the variance in inter-regional connectivity.

To uncover the individual actors in this process, we undertook a thorough feature importance (F.I.) analysis, with both confirmatory and surprising outcomes. Strikingly, oligodendrocytes were implicated as the most important cell type feature for recreating connectivity. Oligodendrocytes are the brain's myelin and fiber maintenance cells; their role in predicting connectivity is not unexpected, but their prominence in this role has not received adequate attention. A deeper dive also uncovered that non-neuronal cells generally dominate neuronal cells as predictors of connectivity, another surprising finding. Additionally, we identified a strong distance-dependency in the cell-connectivity relationship, with layer-specific excitatory and medium spiny neurons contributing most for predicting long-range connectivity, while non-neuronal cells were more salient for short-range connections. Indeed, the cell types necessary for reconstructing long-range connections are generally different from those most useful for predicting local connectivity, suggesting that these may be maintained by distinct biological pathways. Together, our findings suggest a hitherto under-explored role of specific cell types that play outsize roles in forming and/or maintaining connections.

5.3 RESULTS

OVERVIEW OF THE STUDY PIPELINE

A schematic of the analytic pipeline is displayed in Figure 5.1. We used previously computed regional densities for 200 neuronal and non-neuronal cell types from publicly available single-cell RNA-sequencing (scRNAseq) data from Zeisel, et al.¹⁹ and in situ hybridization (ISH) data from the Allen Institute for Brain Science (AIBS)¹⁷ using the Matrix Inversion and Subset Selection (MISS) algorithm¹⁶ (Figure 1-i). For confirmatory analyses, we also utilized the densities of 25 cell types from the Tasic, et al. scRNAseq dataset^{[16], [18], [20]}. We normalized these raw MISS-inferred densities to create enrichment scores to prevent the scale of these features from artificially influencing the machine learning algorithms' outputs (see Methods). The

connectivity data we attempted to reconstruct was derived from the AMBCA (<http://connectivity.brain-map.org>)[2], which we normalized by volume of the source region, resulting in a 424×424 matrix of normalized connection strengths (Figure 5.1-ii; see also Methods). Our choice of normalization is motivated by the observation by Abdelnour, et al. and others that connectome degree is correlated with region volume[21]; therefore, we marginalized out the effect of source-region volume prior to all analyses. As we were only interested in connectivity between disparate regions and not self-connectivity, we set all diagonal entries of the connectivity matrix to zero. Finally, several machine learning methods were implemented to infer the whole-brain connectome from the regional cell type enrichment scores, which we evaluated quantitatively (Figure 5.1-iii). We also note that we considered the enrichment scores within regions sending out connections (“source”) and within regions receiving connections (“target”) as separate features, resulting in models with 400 total features for the Zeisel, et al. dataset and 50 total features for the Tasic, et al. dataset.

PREDICTING THE EXISTENCE OR ABSENCE OF CONNECTIVITY

We first addressed whether regional cell-type enrichment features can be used to predict the existence or absence of connectivity between any given pair of regions, because the underlying biological difference between zero connectivity and non-zero connectivity is qualitatively different from any differences in degree of connectivity between region pairs (see Methods). Figure 5.1A shows the proportions of zero and non-zero values within the ABMCA, indicating that the mouse brain connectome is approximately 64% sparse. To perform this binary classification task, we began with common unsupervised clustering methods Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). Neither approach could distinguish region pairs that form connections from those that do not (Figures 5.2A and B, respectively). However, the random forest (RF) algorithm produced excellent classification results (Figure 5.2C; Table 5.1)22. The confusion matrix in S. Figure 2A shows that the RF model predicted the existence of connectivity between pairs of regions with an accuracy of 0.80 for the Zeisel, et al. dataset (see also S. Data Table 2). AUROC (Area Under the Receiving Operator Characteristic) and AUPR (Area Under

Precision-Recall curve) values for RF were 0.87 and 0.80, respectively (Figure 5.2C). Thus, regional cell type enrichment profiles can predict the presence of connectivity, paralleling prior findings based on gene expression [14].

PREDICTING CONNECTIVITY DENSITY

We next turned to the task of predicting the connectivity density[2], [15], [23], [24], which we define to be a measure proportional to the number of axonal tracts per unit of source region volume between any region pair. We first examined whether region pairs with similar cell type compositions were likely to be more densely connected. Figures 5.2D and E (left panel) depict heat maps of the ipsilateral regional cross-correlation matrix with respect to cell-type enrichment scores and the mouse connectome, respectively (see also S. Figure 3). While there is a degree of visual similarity, the two measures are only weakly correlated (Pearson's $R = 0.32$; Figure 5.2F). This agrees with previous work suggesting that coupled regions tended to have higher levels of gene expression similarity[5], [14]. We conclude that inter-regional similarity in cell type enrichment profile is related to, but insufficiently predictive of, the whole brain connectome. Given that the connectivity density distribution is mostly comprised of very small values with a number of prominent outliers (Figure 5.6B), we hypothesized that nonlinear predictive models would be more appropriate. Similar to the binary classification task, we found that the RF model recreated connectivity from cell-type enrichment with a high degree of accuracy (Adjusted $R^2 = 0.60$, Root-mean-square deviation = 0.60, 10-fold cross-validation; Table 1; S. Data Table 3). Excellent visual similarity between the connectivity predicted by RF using cell types and the ground truth can be observed in matrix heatmaps (Figure 5.2E, right) and scatter plots (Figure 5.2G), with a Pearson's correlation of 0.79.

To more thoroughly explore the significance of these results, we constructed five collections of randomly generated null models, each of which had the same number of input features as the Zeisel, et al. dataset (i.e. 200 each for source and target region). Figure 5.2H displays distributions of R^2 values from each type of null distribution representing 500 random model instances, and the red vertical line indicates the

performance of the cell-type-based model (Table 1; see also Methods). As expected, the least informative models incorporate no gene-expression information. The purely random models (purple curve), which involved assigning regional cell-type-enrichment scores from a uniform random distribution, were completely uninformative. When these regional values were randomly assigned using distributions whose means depended on the anatomical parcel to which each region belonged (green curve; see also Methods), the predictions improve markedly, reflecting key biology of the anatomical relationships between regions, but remain poor. Performance further improves when scrambling the values of the AGEA before applying MISS on the highly informative MRx3 gene subset (yellow curve; see Methods and Meziyas, et al.¹⁶ for details), but it is much lower than the true cell-type distributions. We also explored the performance of gene expression directly with two different sampling methods: 1) randomly selecting 200 genes within the 4083-gene AGEA (red curve), and 2) randomly selecting 200 genes within the 1360-gene MRx3 subset (blue curve). The model using cell-type features significantly outperforms those using fully random gene sampling, indicating that cell types contain key information for predicting connectivity that is not uniformly reflected in the expression of individual genes. We achieved comparable prediction accuracy using subsets of informative MRx3 genes and cell types; given that MRx3 specifically selects genes based on how well they discriminate between cell types transcriptomically^[16], the agreement between these two types of input features is expected.

In the above analyses, we separated the tasks of predicting the presence or absence of connectivity (binary classification) and predicting the density of connections among connected region pairs (regression). From both biological and machine learning perspectives, these are distinct questions and therefore we chose to address them individually. Nevertheless, we also implemented a RF algorithm to predict connectivity density in the AMBCA without first removing unconnected region pairs (Figure 5.9). As expected, we found that agreement was not as strong between ground truth and predicted connectivity when the zeroes were not first filtered out; however, the adjusted R² of 0.42. We also achieved strong performance (R² = 0.50) when

we split our training and test sets by source region rather than purely randomly (S. Data Table 4). Several other common machine learning algorithms were implemented to reconstruct both the binary connectome and predict connectivity density, which, however, fail to achieve superior performance over random forest (S. Data Tables 2 and 3).

CONFIRMATION WITH AN INDEPENDENT CELL TYPE DATASET

We tested whether the random forest algorithm could also recreate whole-brain connectivity using an independently curated collection of cell types to form the input feature space. For this purpose we used MISS-inferred distributions of the scRNAseq dataset from Tasic, et al., which sampled 25 cell types within the mouse neocortex and thalamus[18],[20]. A natural question to ask is whether the lack of sampling outside of the neocortex and thalamus may bias the whole-brain predictions of cell-type density from this dataset. To address this concern, we have previously shown that the prediction error within unsampled regions is comparable to that within sampled regions (reproduced from Meziaris, et al. in Figure 5.10)[16]. t-SNE and PCA were also unable to separate region pairs that share a connection from those that do not using the Tasic, et al. dataset (5.11A and B). But when we used this less expansive set of cell types, we were still able to produce an accurate recreation of the binarized connectome (AUROC = 0.85, AUPR = 0.78; Table 1; Figure 5.11C – only a modest decrease from the 200-type Zeisel, et al.-derived results (Figure 2C; Table 1). The matrix of cell-type similarity is, again, only weakly correlated to the connectome (Pearson's $R = 0.21$, $p\text{-value} = 0.0$; S. Figures 5.6D–F; Figure 5.12). Notably, the two cell-type similarity matrices created with 25 and 200 features, respectively, are strongly correlated with each other (Pearson's $R = 0.79$, $p\text{-value} = 0.0$; S. Figure 8), which we expected given the reliability of the MISS algorithm. The machine learning models were similarly successful in predicting the connectome with the Tasic, et al. dataset, only modestly underperforming relative to the 200-type Zeisel, et al. dataset (Table 1; Figure 5.11H; Data Tables 5 and 6). Notably, only RF was able to perform both the classification and regression tasks successfully (Tables 5 and 6), reinforcing that RF is uniquely suited to this problem.

FEATURE IMPORTANCE ANALYSIS TO IDENTIFY KEY CELLULAR MEDIATORS OF CONNECTIVITY.

We next asked which cell types contribute the most to predictions of inter-regional connectivity. Unlike other machine learning models that give outputs whose dependencies are difficult to discern, RF models are amenable to feature importance (F.I.) analysis^{22, 25}(see also Methods). F.I. can be thought of as a measure of how much information is contributed by a given feature relative to all other features. Therefore, for each RF model we determined the importance of each cell-type feature, and grouped them by “supertype” as determined by their scRNAseq-based taxonomies. Please refer to S. Data Tables 7–10 for the list of cell-type names and the supertypes to which they belong. We show these as box plots for the Zeisel, et al. connectivity density RF model in Figures 3A–C, where each data point represents the average F.I. score for each cell type across the 10 cross-validation test sets. We considered the salience of each cell type in terms of its source-region (Figure 3A) and target-region (Figure 3B) F.I., as well as its overall salience as an average of source-region and target-region F.I. scores (Figure 3C). The corresponding results for the Tasic, et al. connectivity density RF model (S. Figures 6E and G) and the classification RF models (Figure 2C; S. Figure 6C) are shown in S. Figures 9A–C and S. Figure 10, respectively. Overall, we found that that oligodendrocytes (Oligo) were the most important contributors to both binary connectivity and connectivity density prediction at the whole-connectome level for both the Zeisel, et al. and Tasic, et al. datasets (Figure 3C; S. Figures 9C and 10). On a more granular level, the source-region cell-type F.I. scores strongly resembled the averaged values, with oligodendrocytes again having the highest scores in both the Zeisel et al. and Tasic, et al. datasets (Figure 3A; S. Figure 9A). However, when considering only the target regions’ cell-type compositions, we found that a number of neuronal cell types had higher F.I. scores than oligodendrocytes, with medium spiny neurons (MSN) being a notable outlier for Zeisel, et al. (Figure 3B). We found qualitatively similar results when we retrained the RF model to predict the connectivity densities from neocortical to non-neocortical regions and vice versa (S. Figure 11). We elaborate upon the implications of the divergence between source and target cell-type compositions in Discussion.

More generally, the non-neuronal supertypes were more salient in the RF models than neuronal supertypes. We show the voxel-wise distributions of these non-neuronal Zeisel, et al. and Tasic, et al. supertypes in Figure 5.3D and Figure 5.14D, respectively. Overall, the apparent consistency of these feature importance results between the two independently curated scRNAseq datasets suggests a true biological connection between these non-neuronal support cells and connectivity at a whole-brain level.

THE EFFECT OF INTER-REGIONAL DISTANCE ON PREDICTING CONNECTIVITY DENSITY

Although adult cell-type distributions are highly informative for reconstructing the mouse connectome, the unexplained variance in the data likely comes from other biological factors. For instance, we found that there is a strong inverse relationship between inter-regional center-to-center distance and connectivity density (Pearson's $R = -0.33$, $p\text{-value} = 0.0$; Figure 5.4A), indicating that there is a bias towards short-range connections in the mouse brain. Using spatial distance as a sole predictor of connectivity density produced an RF model with an average R^2 of 0.12, indicating that distance contributes modest but significant information (Figure 5.4B). Further, including it along with the cell-type distributions produced RF models with higher R^2 values ($\Delta R^2 = 0.09$; Figure 5.4B). By contrast, using the taxonomic distance matrix as a predictor, where distance is defined in terms of how early each region-pair separated anatomically during development²⁶, contributed less information than spatial distance and did not provide an improvement over cell-type enrichment scores (Figure 5.4B; see also Methods). These results indicate that inter-regional spatial distance contributes information that is at least partly independent of that contributed by regional cell-type composition, while the information from taxonomic distance is fully captured by differences in regional cell-type composition. When we looked at the distance dependence of connectivity density within major anatomical region groups, we found that each set of regions generally has a broad distribution of connection lengths (overall interquartile range = [2.2 mm, 4.6 mm]; Figure 5.4C). However, while each distribution is left-skewed, indicating that shorter-ranged connections predominate, we found that neocortical regions mediate a disproportionate number of the long-range connections in the brain. Consequently, we were interested in whether there was a distance dependence to cell-type F.I., as has been

suggested previously[27], [28]. We therefore trained the RF algorithm on the upper and lower quartiles of connections by distance separately and determined the F.I. scores per cell supertype as above (Figures 5.4D and E). The RF models achieved similar fits regardless of distance bin ($R^2 = 0.61$ and 0.58 for short-range and long-range connectivity, respectively) and performed comparably well to the model of whole-brain connectivity (Table 5.1). However, clear differences emerged at the level of F.I. between short-range and long-range connectivity. Although oligodendrocyte distributions from the Zeisel, et al. and Tasic, et al. datasets were not the strongest contributors to the RF model of short-range connectivity as they were for whole-brain connectivity, they remained among the top features, and generally non-neuronal cells had stronger source-and-target-averaged F.I. scores than neurons, as above (Figure 5.4D; Figure 5.17A). In particular, immune cells (Immune) and vascular cells (Vasc) exhibited the strongest contributions to short-range connectivity for the Zeisel et al. and Tasic, et al. datasets, respectively. Of the neuronal supertypes, forebrain glutamatergic neurons (Neo Glu, Thal Glu, Hip Neo Glu) had particularly weak F.I. scores. Interestingly, this trend is reversed for reconstructing long-range connectivity: for both datasets, we found that these three neuronal cell-type distributions were consistently among the most salient features (Figure 5.4E; Figure 5.17B). As with the target-region cell-type F.I. analysis for Zeisel, et al., the supertype with the highest F.I. score was striatal medium spiny neurons (MSN), which are unique to that dataset (Figure 5.4E). We summarize these results in Figure 4F, which shows that, for both the Zeisel, et al. and Tasic, et al. datasets: 1) non-neuronal cell types, and in particular vascular and immune cells, contribute predominantly to predicting short-range connectivity as opposed to long-range connectivity; and 2) telencephalic glutamatergic neurons contribute little to models of short-range connectivity, but they are over-represented among types that predict long-range connectivity. In short, while cell-type-based RF models can reconstruct short-range and long-range connectivity with a similar degree of accuracy as the whole-brain connectome, the saliency of the cell-type features markedly differs between these models. A more nuanced picture emerged when we considered the source- and target-region cell-type contributions to short- and long-range connectivity prediction separately (Figures 5.18 and 5.19). The contributions of individual source- and target-region non-neuronal cells were variable; as a class, they generally exceeded

neuronal supertypes when considering only source-region supertypes in predicting short-range connectivity. In other words, consistent with the above analyses, while non-neuronal contributions predominated when considering overall connectivity prediction (Figure 5.3C; Figure 5.14C), this was driven predominantly by their source-region F.I. scores and the prediction of shorter-distance connectivity densities. The medium spiny neuron and telencephalic glutamatergic supertypes also exhibited interesting trends when separating source- and target- region features. As mentioned above, MSN was the strongest contributor among target-region features to overall connectivity prediction (Figure 5.3B) and among source-and-target-averaged features to long-range connectivity prediction (Figure 5.4E). However, we found that, among only target-region features, MSN was in fact was the strongest contributor to both short- and long-range connectivity prediction (Figures 5.18B and D) and did not strongly contribute as a source-region feature to long-range connectivity prediction (Figure 5.18C). For both the Zeisel, et al. Hip Neo Glu and Tasic, et al. Neo Glu supertypes, there was no effect of separating out source-region from target-region supertype features, providing similarly weak contributions to short-range connectivity prediction (S. Figures 5.18A–B and 5.19A–B) and similarly strong contributions to long-range connectivity prediction (S. Figures 5.18C–D and 5.19C–D). In summary, while medium spiny neurons and telencephalic glutamatergic neurons both disproportionately contributed to predicting long-range connectivity, the contributions between source- and target-region enrichment scores markedly differed between them. To further examine the underpinnings of the discrepancy between the supertypes most critical for predicting short-range and long-range F.I., we examined whether there was a relationship between how variably distributed the Zeisel, et al. supertypes were across the brain and F.I. We hypothesized that more spatially homogeneous cell types would contribute less to the RF model’s predictiveness. As shown in Figure 4G, we indeed found that, for the RF models predicting all connectivity (left panel) and short-range connectivity (center panel), there was a statistically significant negative association between each supertype’s average F.I. score and its spatial coefficient of variation (CoV), with Pearson’s R values of -0.67 (p-value = 3.5×10^{-3}) and -0.69 (p-value = 2.0×10^{-3}), respectively. However, while the association trended negative for the long-range RF model (Figure 4G, right panel), it was weak and not statistically significant (Pearson’s R = -0.10, p-value = 0.72).

The two outliers with especially high long-range F.I. scores, MSN and Hip Neo Glu, have intermediate CoV values (Figure 5.4G, right panel), which agrees with their distributions being highly specific to a relatively large set of regions (Figure 5.5A and B). Therefore, we conclude that the contributions of supertypes to long-range connectivity density predictions in particular cannot be simply explained by spatial heterogeneity.

NEURONAL CONTRIBUTIONS TO LONG-RANGE CONNECTIVITY

To explore some of the relationships between cell-type distributions and connectivity qualitatively, we show the distributions of Hip Neo Glu and MSN, the two supertypes from the Zeisel, et al. dataset with the highest average F.I. for predicting long-range connectivity (Figure 5.5A and B). The Hip Neo Glu supertype comprises twenty-four individual cell types, all of which are excitatory and located within neocortical and hippocampal regions, and the MSN supertype comprises six types of striatal medium spiny neuron¹⁹. As expected, based on their taxonomy, Hip Neo Glu cells are confined to the neocortex and hippocampus, while MSN cells are entirely within the striatum. Given the high degree of regional specificity of these cell-type supertypes, we also show the strongest long-range connections to and from the neocortex (Figure 5.5C) and the striatum (Figure 5.5D). More specifically, for the neocortex, these include projections to hindbrain nuclei and contralateral neocortical-neocortical connections (Figure 5.5C). The main long-range projections from the striatum originate in the olfactory tubercle and terminate in the periaqueductal gray of the midbrain, while it receives its strongest long-range inputs primarily from contralateral neocortical regions (Figure 5.5D). In this way, we can link the anatomical distributions of cell types to specific subsets of inter-regional connections.

5.4 DISCUSSION

SUMMARY OF KEY RESULTS

Our results constitute practical applications of data-driven machine learning models for reconstructing whole brain inter-regional connectivity using spatial cell type enrichment distributions. We split this reconstruction into two tasks: a classification task to predict the existence and absence of connections between each region pair, and a regression task to predict the values of connectivity density between all connected region pairs. We find that using the comprehensively sampled Zeisel, et al. cell-type distributions[16], [19], random forest models are able to perform both tasks with a high degree of accuracy (Figure 2; Table 1), which we replicate using the smaller Tasic, et al. dataset (Figure 5.11; Table 5.1)[16], [18], [20]. Post hoc feature importance analyses implicate oligodendrocytes as especially critical in correctly recreating the whole brain connectome (Figure 5.3; Figure 5.14). We further consider inter-regional distance as an important predictor of the density of brain connectivity (Figure 5.4). When feature importance is evaluated separately for short-range versus long-range connections, we find that medium spiny neurons and telencephalic glutamatergic neurons appear to be far more important for recreating long-range connectivity than for short-range connectivity, while non-neuronal cell types are more important for recreating short-range connectivity. We discuss below the implications of our findings, some confirmatory and some unexpected, in the context of current literature.

PREDICTING BINARY AS WELL AS WEIGHTED CONNECTOMES

We divided our ML prediction tasks by separately predicting the absence or presence of a connection and the connectivity density between any given region pair two reasons. First, the connectivity data are quite sparse (36% nonzero region pairs), which can significantly impact the the ability of the model to generalize. Second, a zero connectivity density value might not necessarily mean there is no connectivity between two regions at all; rather, it might only mean the intensity was not able to pass the threshold of observability imposed by the mesoscale connectome methodology[2]. Nevertheless, when we attempted to predict

connectivity density for the whole connectome (including region pairs with zero connectivity density), the RF model exhibited strong agreement with the ground-truth connectome, although not nearly as high as that with the zero values removed (Figure 5.8).

MODEL PERFORMANCE IS REPLICATED ACROSS TWO DIFFERENT SCRNASEQ DATASETS

We were able to replicate the results of our primary dataset - the 200-type Zeisel, et al. dataset[19] - using a separate, 25-type dataset from Tasic, et al.[18, 20] (Figure 5.2; Figure 5.11; see also Methods). Interestingly, we found that the Zeisel, et al. dataset performed only modestly better despite containing a far more diverse array of cell types sampled from a more comprehensive set of brain regions. One possibility is that, because training accuracy is close to 1 even for the Tasic, et al. dataset, there is a limit to how well cell type features in the test set can reconstruct connectivity using machine learning. This observation is supported by the results from the RF models using subsets of genes (Figure 5.2H, red and blue curves), whose performance also did not exceed that of either cell-type model. It is possible that a subset of the Zeisel, et al. cell types might outperform the 25 cell types from Tasic, et al., but the current study design is not well-suited for exploring all combinatorial possibilities. Alternatively, it may be that the 25 cell types inferred from the Tasic, et al. dataset, despite representing only a subset of mouse neuronal diversity, provide close to maximal information content for reconstructing brain connectivity. For example, the four non-neuronal supertypes (astrocytes, oligodendrocytes, immune cells, and vascular cells) from the two datasets are qualitatively very similar in spatial distribution (Figure 5.3D; Figure 5.13D) and consistently have higher F.I. scores than most neuronal supertypes (Figure 5.3C; Figure 5.14C). Further, for the more regionally specific long-range connections (Figure 5.4C), both datasets have robust supertypes of the telencephalic glutamatergic neurons that were especially important in reconstructing the long-range connectome (Figure 5.4E; Figure 5.18B). Nevertheless, that we were able to create models with high predictive accuracy with two sets of cell type enrichment scores coming from independently sampled scRNAseq datasets reinforces the central claim that adult cell type distributions strongly reflect the brain connectome.

COMPARISON WITH PREVIOUS WORK

Our work is preceded by several previous attempts to model the wiring diagram of the brain. Henriksen, et al. modeled the mouse mesoscale connectome with graph-theory-based approaches[7], and Reimann, et al. built a null model for the micro-connectome integrating the macro- and mesoscale connectomics⁹. Although these studies are not directly related to our current effort, they highlight the importance of graph-theoretic features and generative models in studying the mesoscale mouse connectome. In the present study, we have focused almost exclusively on molecular or cellular signatures of connectivity, but these studies indicate that future work incorporating additional graph theoretic contributors for predicting brain wiring diagram could be fruitful. An approach much closer to ours was taken by French, et al., who built statistical models correlating the gene expression signatures of 17,530 genes in 142 anatomical regions from the Allen Brain Atlas, and identified a subset of genes that are statistically correlated with the brain's wiring diagram⁵. They found a strong association between transcriptomic data and the connectome, which motivated us to create a predictive model of whole-brain connectivity from spatially distributed biological features. Ji, et al. went a step further by performing machine learning to predict the existence or absence of brain connectivity

from gene expression, using a previous version of the AMBCA as a target[14]. Their approach yielded a very similar predictive accuracy and AUC as the classification results we present here, and their results underscore that random forest appears to be an excellent approach whether the features are based on regional gene expression or cell type distributions. However, in addition to predicting the existence of connectivity, here we also demonstrate that cell-type densities can be used to recreate the actual connectivity density values with high accuracy.

An alternative, experimental approach linking cell types to connectivity is BRICseq, which allows for the high-throughput mapping of axonal tracts alongside the transcriptomic profiling of the projecting

neurons[15]. However, BRICseq has not yet been scaled up to produce a connectivity map of comparable spatial resolution and coverage as the AMBCA[2, 15]. Therefore, to our knowledge, no prior approach has been able to computationally link regional cell-type composition and whole-brain connectivity.

Cell-type density versus gene expression as predictors of connectivity. We propose here that cell type features are a valuable alternative to gene expression for recreating the brain connectome for the following reasons:

- 1) Cells are the most fundamental unit responsible for inter-regional connectivity.
- 2) Most neural cell types have roughly fixed functions and spatial locations in the adult brain, whereas expression for many genes is highly temporally variable.
- 3) Using gene expression requires informed feature selection given the sheer number of mammalian genes and gene variants. While previous authors have reported such feature selection procedures, they necessarily rely on prior assumptions or knowledge.
- 4) The larger the feature set (e.g., using the entire mouse transcriptome), the higher the risk of overfitting and non-generalizability.

Throughout our study, we have taken care to address these challenges, and the use of a small number of cell type features, particularly for the confirmatory Tasic, et al. dataset, was considered a means of avoiding these pitfalls. Compared with the thousands of gene features used in prior studies[5], [14], the sets of 25 and 200 cell types should form a more parsimonious input feature space.

That being said, tremendous effort has been invested in obtaining gene profiles of cell-type-specific marker genes as well as genes involved in processes related to the formation and maintenance of projections between brain regions. Our work both complements those efforts and also shows that we can obtain cellular signatures from genes using the MISS algorithm which are not necessarily single-cell markers but nevertheless contribute significant information for predicting connectivity.

Oligodendrocytes are disproportionately associated with whole-brain connectivity patterns. Our analysis demonstrated the importance of oligodendrocyte cell types in recreating the whole-brain connectome. Oligodendrocytes are the most predictive feature in the random forest model for both datasets (Figure 5.3A–C; Figure 5.14A–C), and are also among the highly predictive features when analyzing the feature importance for the classification task (Figure 5.15).

Biologically, oligodendrocytes produce the myelin sheath insulating neuronal axons[29,30]. They help protect the vulnerable axons from parenchymal chemokines and cytokines, and ensure the fast and efficient movement of action potentials[29–31]. Dysfunction of oligodendrocytes can interfere with normal microstructure and functional connectivity in the mouse brain[32].

Oligodendrocyte myelination was also shown in previous work to be able to regulate the loss of synapses³³. Moreover, recent work from Buchanan, et al. showed that oligodendrocyte precursor cells can prune axons in the mouse neocortex[34]. When we modeled short-range and long-range connectivity separately, we found that while oligodendrocytes contributed strongly to short-range connectivity, they were somewhat less informative for reconstructing long-range connectivity for the Zeisel, et al. dataset (Figure 5.4D–E). Overall, our results underscore the critical role this cell type plays in maintaining white matter integrity.

Non-neuronal cells contribute to whole-brain and short-range connectivity. Non-neuronal cell types also had high feature importance and we highlight them below. Brain vascular cells compose the blood-brain barrier, which protects the vulnerable central nervous system (CNS), and they interact with the CNS for supporting neuronal cells with nutrients, energy, and oxygen[35–39]. Their breakdown is strongly correlated with brain connectivity disruption and cognitive defects[35,39]. Brain endothelial cells are involved in the process of neurovascular coupling[40,41], whereby local neural activity stimulates subsequent blood flow changes in the corresponding downstream locations[41,42]. That endothelial cells are more important for short- and medium-range connections but not for long-range ones supports a role in

local circuit maintenance rather than long projections. We also found that immune cell and astrocytes play an outsize role in predicting connectivity compared to neuronal cell types. Previous studies have indicated that there is an association between inflammation and functional brain connectivity[43, 44]. Similarly, astrocytes, the most abundant glial cells in the CNS, have critical impact in maintaining many physiological functions of neurons. Germane to this investigation, previous experimental work has shown the existence of bidirectional interactions between astrocytes and synapses[45].

Further, we found that non-neuronal cell types contribute disproportionately to predicting short-range connectivity. Of these, immune cells were the most important supertype for the Zeisel, et al. dataset and vascular cells were the most important supertype for the Tasic, et al. dataset, although all non-neuronal superotypes tended to have higher F.I. scores than most of the neuronal superotypes (Figure 5.4D; Figure 5.17A). There are multiple reasons why these non-neuronal cells have higher F.I. scores for predicting short-range as opposed to long-range connectivity. Generally, many non-neuronal cell types are thought to impact and interact with neighboring neuronal cell bodies in the gray matter, which may result in the mediation of more local, short-range connectivity. Alternatively, it is possible that non-neuronal cells, in their various roles supporting neuronal function, are important in the formation and maintenance of all connections in the CNS (Figure 5.3A–C; Figure 5.14A–C). However, given that F.I. is a relative measure of the model information provided by a given feature, non-neuronal cell types contribute at most moderately to the long-range models of connectivity because certain neuronal cell types have an outsize distance-dependent effect (see below; Figure 5.4E; Figure 5.17B). The distance dependence of cell-type contributions to connectivity is an important line of inquiry for future studies.

NEURONAL SUBTYPES DIFFERENTIALLY MEDIATE LONG-RANGE CONNECTIVITY

In addition to oligodendrocytes, we found that telencephalic glutamatergic neurons and striatal medium spiny neurons were among the most salient classes of cell types, but only for predicting long-range connections (Figure 5.4E; Figure 5.17B). The former are well known to project to remote locations within

and outside of the neocortex (Figure 5A,C), and therefore their prominence in long-range but not shorter connections is consistent with their neurobiology. It is particularly striking that the telencephalic glutamatergic cell supertypes in both the blue, et al. and Tasic, et al. datasets (Neo Glu and Hip Neo Glu, respectively) are also among the least important features for predicting short-range connectivity (Figure 5.4D; Figure 5.17A), suggesting that these neurons predominantly engage in long-range connections. Similarly, the high F.I. of medium spiny neurons is concordant with their function, as these are long-range-projecting, inhibitory neurons. Medium spiny neurons comprise a significant fraction of neurons in the striatum and are involved in dopamine signaling; notably, these neurons selectively exhibit altered behavior in several psychiatric disorders⁴⁶. When we look at the F.I. of individual cell types between the two datasets, we see a similar pattern as we do at the supertype level (Figure 5.4G). In particular, telencephalic glutamatergic neurons contribute weakly to predicting short-range connectivity and are overrepresented among types with high F.I. scores for predicting long-range connectivity. Since telencephalic glutamatergic neurons comprise many of the long-range, inter-regional connections of the brain, the distance dependence we observed is biologically plausible. One interesting difference in the ways in which these two classes of cell types contribute to connectivity density prediction emerges when examining the contributions of source-region and target-region cell-type features separately. The Neo Glu and Hip Neo Glu supertypes were disproportionately informative for predicting long-range connectivity when considering either source or target (Figures 5.18C–D and 19C–D), whereas the target-region MSN supertype had more relative importance for predicting both short- and long-range connectivity and was only moderately informative as a source-region feature (Figure 5.18). As shown in Figure 5A, the distribution of Hip Neo Glu is entirely telencephalic; these regions are involved in a disproportionate fraction of long-range connections (Figure 5.4C), the strongest of which tended to be contralateral and intra-neocortical (Figure 5.5C). That source- and target-region F.I. were both high for Hip Neo Glu reflects the intra-cortical nature of these connections. By contrast, the striatum, and caudoputamen in particular, have many more incoming long-range connections than outgoing long-range connections (Figure 5.5D), and therefore there should be a large difference between source- and target-region F.I. for MSN, which we observed (Figure 5.17B, D). Taken

together, these results suggest that the formation and maintenance of brain connections requires a wide array of cell types. However, we caution that this kind of feature importance analysis will benefit from further experimental work to elucidate in more detail the biological roles of the identified cell types with respect to connectivity.

5.5 FUTURE DIRECTIONS

One extension of the current method would be to apply feature selection on either of the cell type datasets used here, which may facilitate the development of more predictive models. Additionally, machine learning models that integrate both cellular features and anatomic/morphological features can be expected to improve current predictions. Creating cell-to-cell or even voxel-to-voxel level connectivity and benchmarking against known neuronal cell-type-specific signaling pathways would be beneficial for future research but will require higher-resolution data. Given the conservation of central nervous system properties in mammals, we may also be able to apply these data-driven methods to the human brain.

LIMITATIONS OF THE STUDY

The primary limitation of the current work is that cell type enrichment does not accommodate other factors critical for determining brain connectivity, including Neural polarity, cell maturation and migration. Further, despite their ability to produce F.I., random forest models are less interpretable than generalized linear models. RF models, an ensemble of decision trees, can also suffer from overfitting, since any constituent decision tree may be sensitive to data variations and noise. However, we note that the issue of overfitting cuts across almost all machine learning methods and is not specific to RF. In this study we have taken great care at various steps to minimize this risk, starting from the basic design of using only the cell-type features from the two connecting regions, and eschewing full brain or neighboring regional features. Also, as mentioned above, we have not explored feature selection to produce a minimal set of informative cell types, either for the 25-type Tasic, et al. or the 200-type Zeisel, et al. dataset, and therefore it is possible that the

model performance demonstrated here could be further enhanced. Finally, we were still limited by the resolution of both the mouse brain connectome and cell type density maps, and therefore did not attempt to separately predict additional features of keen interest, such as cell polarity.

Several caveats are worth mentioning in regards to the input features used here. First, we used the coronal series of the AGEA, which contains far fewer unique genes (4083) than the sagittal series and has a neuron and hippocampal bias¹⁷ for the MISS pipeline and the null models of Figure 5.2H. The coronal series, however, has a superior spatial resolution of 200- μm ; ultimately, we decided that higher accuracy in regional quantification of gene expression was more important than the limitations inherent to the gene set. Similarly, our choice to use cell densities inferred using the MISS algorithm was motivated by its comprehensive spatial coverage. Within the MISS pipeline, we apply a gene selection algorithm called MRx3 to filter out thousands of uninformative genes for the purpose of reconstructing cell-type densities¹⁶, so having a more expansive gene set may not necessarily lead to significantly better predictions. However, we note that several promising technologies are emerging that have demonstrated single-cell-level resolution of brain tissue, such as STARmap^[47], osmFISH^[48], and merFISH^[49], ^[50]. While the spatial resolution and direct transcriptomic mapping of cell types using these methods is impressive, they have not yet been scaled up beyond single regions. More recent work using BARseq mapped approximately 1.2 million individual cells within the mouse forebrain and labelled them using 107 marker genes⁵¹; however, the authors biased their sampling towards neocortical glutamatergic neurons and so this dataset lacks the breadth of transcriptomic diversity captured within the Zeisel, et al. dataset used here. Therefore, for exploring the architecture of the whole-brain connectome at a mesoscopic scale as it relates to cell-type distributions, we chose to use MISS for its breadth of spatial coverage and amount of cell-type diversity. Many questions about whole-brain microarchitecture, which would require mapping cell types and projections at a single-cell level to answer, remain the subject of future work in this area.

5.6 CONCLUSIONS

We report a data-driven approach that successfully predicts whole-brain connectivity from regional cell type information in the mouse brain. We report quantitative evidence of the vital importance of interareal distance and non-neural cell types in recreating connectivity, especially of oligodendrocytes and other non-neuronal cell types. Our results may provide guidelines for future experimental analysis, and can be extended to other mammals, including humans.

Author Contributions AR and SS conceived of the presented idea of the study. AR developed the theory and main experiment design. Sun performed the early machine learning experiments and generated early results with the help of DM. CM and JT verified the analytical methods. JT further introduced MISS as an important analytical tool that expanded the scope of the current research. SS performed data collection and built the machine learning pipeline and JT implemented the statistical analysis and contributed in feature selection. All authors discussed the results and contributed to the final manuscript. JT and SS took lead in the manuscript writing and figure generating. DM, CM, and AR polished the manuscript; each of them has provided important advice based on their domain knowledge. AR supervised the project's progress and provided necessary guidance on general writing and submission.

5.7 ACKNOWLEDGMENTS

This work was supported by the following NIH grants: R01NS092802, RF1AG062196, R01AG072753.

5.8 DECLARATION OF INTERESTS

The authors declare no competing interests.

5.9 METHOD DETAILS

We use two primary sources of data: MISS-derived cell type enrichment scores, which are themselves a function of gene expression data and serve as our models' input features, and the Allen Mouse Brain Connectivity Atlas (AMBCA), which serves as our empirical ground truth for training and testing our models. These data are available at the DOI listed in the Key Resources Table above.

NOTE ON MISS

We developed the Matrix Inversion and Subset Selection (MISS) algorithm to deconvolve spatially resolved gene expression data, such as those provided by the AGEA, into cell type densities using cell-type-specific gene expression signatures from scRNAseq. The fundamental problem can be stated as: $E=C \times D$, where E is the genes-by-voxels matrix from the AGEA, C is the genes-by-cell-types matrix from scRNAseq, and D is the unknown cell-types-by-voxels matrix to be determined.

MISS-DERIVED CELL TYPE FEATURES

Although the Allen Gene Expression Atlas (AGEA) contains spatially resolved gene expression information for thousands of genes, a similar dataset directly mapping a comprehensive set of cell types in the mouse brain has not been produced. The MISS pipeline is capable of deconvolving the spatial gene expression data from the AGEA into cell type densities with cell-type-specific single-cell RNA-seq (scRNAseq) data.

MOUSE CONNECTIVITY

We use the AMBCA as the source of the mouse connectome we reconstruct from cell type features, which was assembled using viral tracing. The resulting mesoscale connectome, C , is represented as a matrix, with $C(i, j)$ representing the total connectivity from region i to region j .

MACHINE LEARNING METHODS

We implemented several machine learning methods for predicting brain connectivity, dividing our ML prediction tasks by separately predicting the absence or presence of a connection and the connectivity density between any given region pair.

CELL TYPE INPUT FEATURES

For both the prediction tasks, we used the cell-type enrichment vectors from both the source and target regions, resulting in a comprehensive set of features for our models.

NULL MODEL INPUT FEATURES

To benchmark the performance of our model, we used several types of "null" input features, including purely random, region-coupled, scrambled MISS, random genes, and random MRx3 features.

RANDOM FOREST

The main findings were obtained from random forest models, known for generating a number of decision trees on various sub-samples of the dataset and using averaging to improve predictive accuracy and control overfitting.

OTHER ML MODELS

In addition to random forest, we tested several common machine learning algorithms including linear models like ridge and LASSO, support vector machines (SVMs) with a Radial (RBF) kernel, and other models implemented by Scikit-learn.

NEURAL NETWORK MODELS

We explored the performance of modern neural-network-based models, including shallow and deep learning models like the multilayer perceptron (MLP), and more advanced neural network models using a Pytorch-based multi-layer perceptron.

MODEL PERFORMANCE EVALUATION

All model evaluation results are reported for the testing dataset only, after 10-fold cross-validation, with metrics such as precision, recall, Root Mean Square Error (RMSE), and R-squared score used for evaluation.

3D BRAIN VISUALIZATION

We used Brainframe, an in-house MATLAB package, to generate the 3D mouse brains, the distribution of gene expression and cell-type patterns within, and the brain connectome.

INTER-REGIONAL DISTANCE MATRIX CALCULATION

To calculate the distance between each region-pair in the mouse CCF, we determined the center of mass of each region and used the pairwise Euclidean distance between these regional centers of mass as a proxy for the lengths of the white matter tracts connecting them.

FEATURE INTERPRETATION FROM RANDOM FOREST MODELS

To decompose the random forest model and calculate the importance of each input feature, we used the Scikit-learn Python package, calculating a node's importance for each decision tree and averaging the importance across all trees.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed using Python and MATLAB programming languages. Machine learning approaches were assessed and averaged over ten-fold cross-validation, with the Standard Error of

the Mean (SEM) computed for precision and dispersion. Statistical significance was defined based on p-values, and appropriate techniques were used for randomization, stratification, and sample size estimation. Accuracy and AUROC for the classification tasks were obtained directly from the Python implementation of random forest. R2 and Pearson's R values were obtained using standard linear regression. Two-sample t-tests following Fisher's R-to-Z transformation were used to compare model performance. Preliminary analyses were conducted to ensure data met the assumptions of the chosen statistical methods, addressing any deviations through data transformation or non-parametric alternatives.

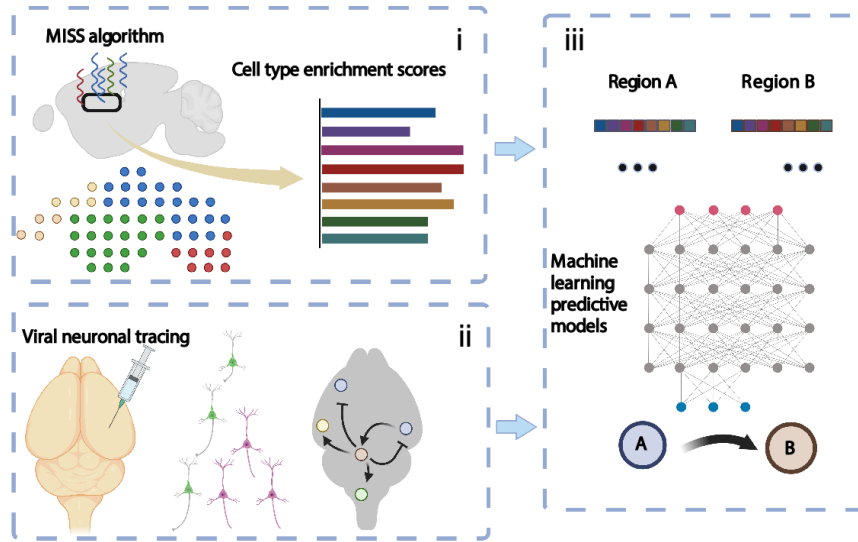


Figure 5.1 Study design

Top left: The spatial quantification of cell type enrichment was computed with the computational pipeline MISS16 from publicly available gene expression data. Bottom Left: The brain connectivity graph was measured by Allen Mouse Brain Connectivity Atlas (AMBCA) using viral neuronal tracing techniques. Right: Machine learning algorithms were then implemented to predict the connectivity between each two regions.

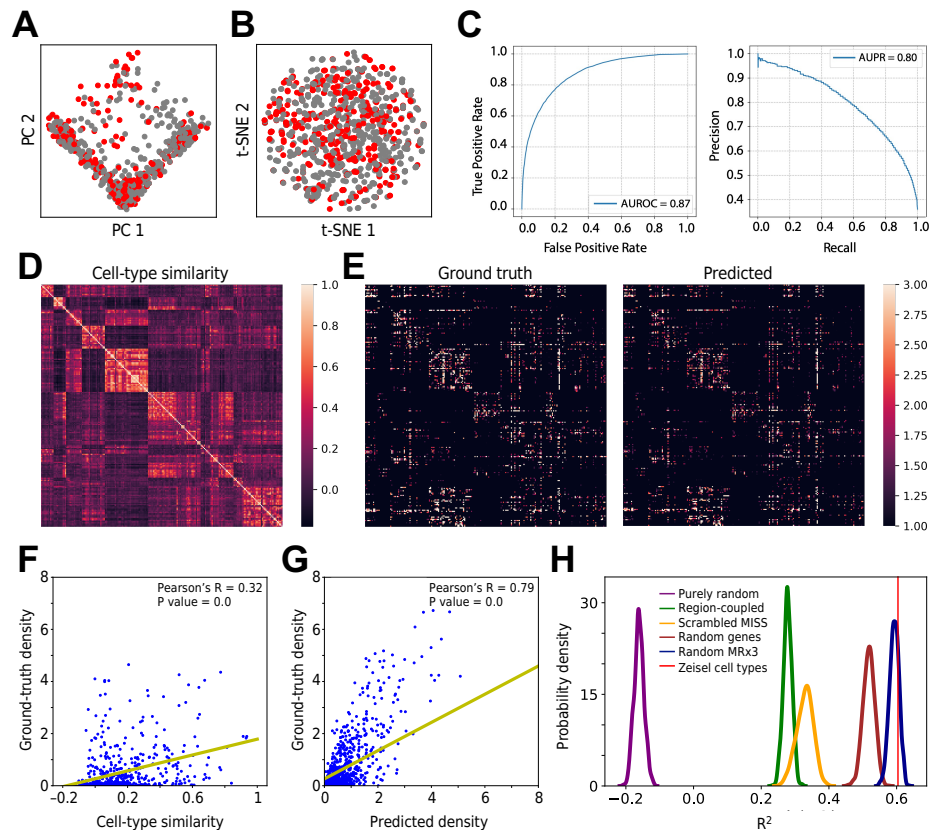


Figure 5.2 Machine learning applied to regional cell type distributions predicts both the existence of connectivity and connectivity density

A. Principal component analysis (PCA) of the cell type spatial quantification array. B. t-distributed stochastic neighbor embedding (t-SNE) of the cell type spatial quantification array. Neither method shows distinct clusters based on the presence or absence of connectivity. C. Performance evaluation of the classifier model using ten-fold cross-validation. Left: The receiver operating characteristic curve (AUROC = 0.87). Right: The precision recall curve (AUPR = 0.80). D. Cellular similarity matrix (quantified using Pearson correlation) of spatial cell type enrichment quantification across brain regions, ipsilateral only. E. Left: Brain connectivity matrix (log₂-transformed). Right: RF prediction without splitting the training and test set. The depicted matrices' rows and columns represent individual regions, and the connectivity between regions is denoted by the matrix entries. The random forest model was able to qualitatively reconstruct the whole brain connectome. F. Scatter plot of pairwise cellular similarity (as depicted in D) between two regions' cell type distribution vectors versus the log-transformed connectivity strength between the two regions (as depicted in E, left), and the fitted linear regression curve (Pearson's R = 0.32, p-value = 0.0). G. Scatter plot showing the correlation between the ground truth connectivity strength between all regions pairs with non-zero connectivity and their predicted values for connectivity using cell types as predictors in the RF model (test set only), along with the fitted linear regression curve (Pearson's R = 0.79, p-value = 0.0). H. Distributions of R² values from null models using five types of inputs in the figure below, each with the same number of features as the Zeisel, et al. dataset (i.e. 200): purely random white noise (purple); region-coupled white noise (green); cell-type "distributions" obtained from MISS after scrambling the regional gene expression values in the AGEA (yellow); randomly selected genes from the 4083-gene AGEA (red); and randomly selected genes from the 1360-gene "high-information" subset used to infer the Zeisel, et al. cell-type distributions in MISS (blue)¹⁶. The red vertical line indicates the performance of the cell-type-based model presented in the manuscript. Each distribution represents 500 random model instances.

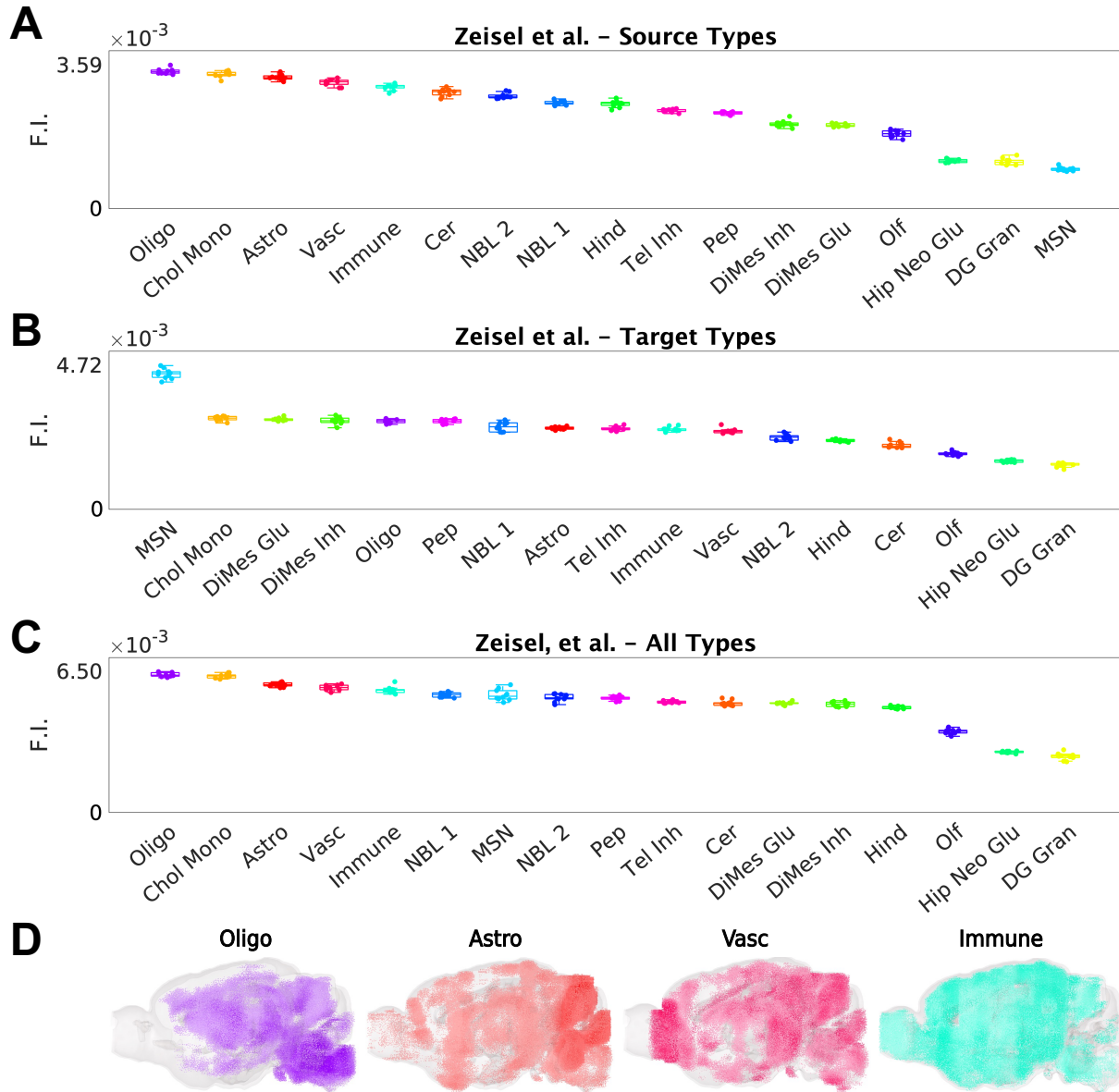


Figure 5.3 Interrogating the individual contributions of cell types

A.Box plots showing the feature importance values of all source-region cell-type features in the random forest model for the Zeisel et al, with the Standard Error of the Mean (SEM) computed as the average across ten-fold cross-validation. cell types, grouped by supertype. B. F.I. values for target region cell-type features, with the SEM computed as the average across ten-fold cross-validation. C. F.I. values for all cell-type features, with the SEM computed as the average across ten-fold cross-validation. D. Sagittal views of cell type densities at the voxel level as inferred by MISS for the corresponding Zeisel, et al. cell-type classes. Please refer to S. Data Tables 7–10 for the full cell type names and description.

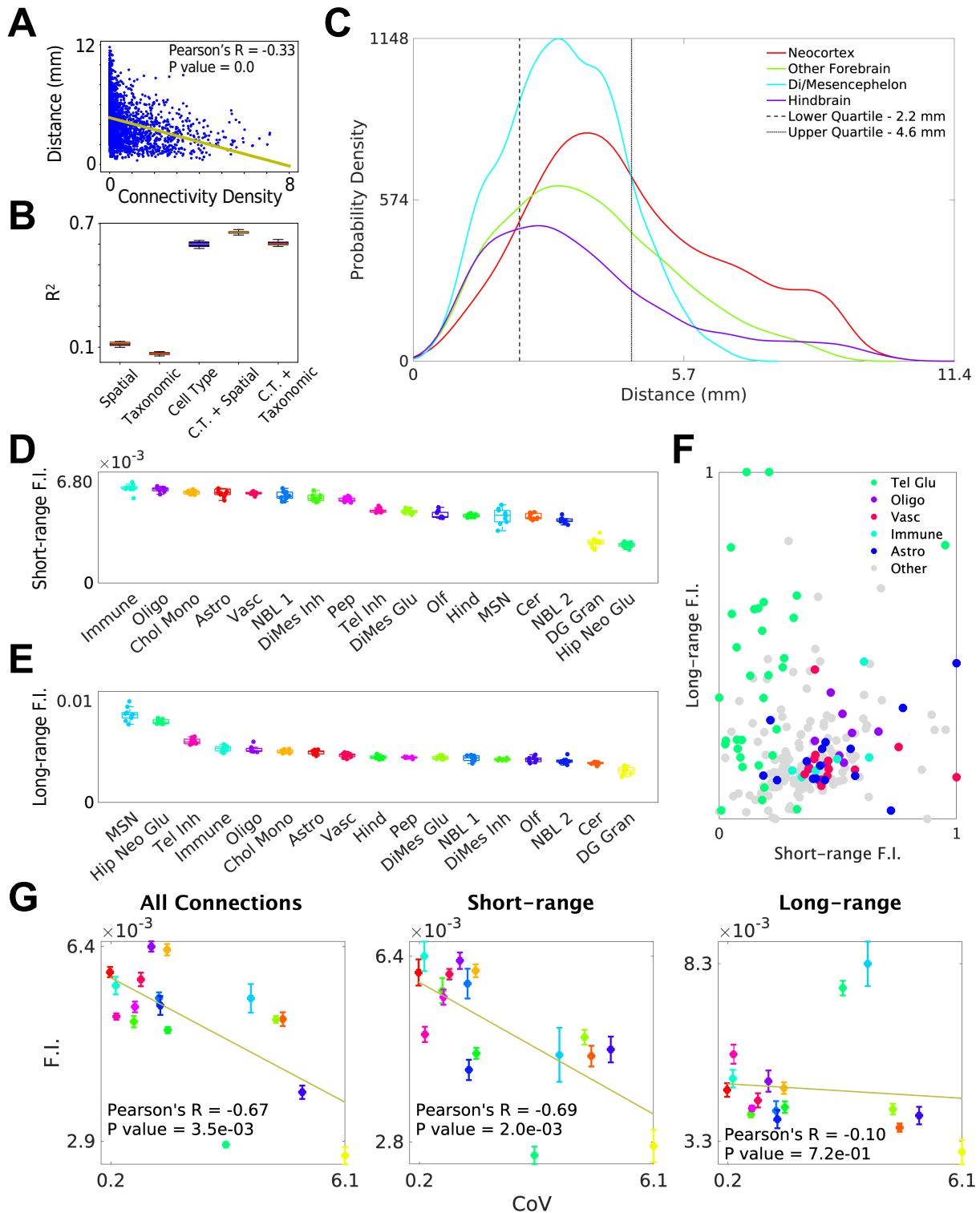


Figure 5.4 Most important cell type contributors vary depending on inter-regional distance
 A. Scatter plot of inter-regional distance and connectivity, showing that distance has a weak correlation with connection strength. B. Box plots of R^2 values following ten-fold cross-validation using different combinations of input features. From left to right: spatial distance matrix, taxonomic distance matrix, cell-type enrichment scores, cell-type enrichment scores with spatial (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) distance, cell-type enrichment scores with taxonomic distance. C. Kernel density estimate (KDE) plot of the probability of two regions being connected as a function of inter-regional distance. The individual lines represent the subregions comprising the neocortex (red), the combination of subregions within the amygdala, cortical subplate, hippocampal formation, olfactory bulb, pallidum, and striatum (green), the combination of subregions within the hypothalamus, thalamus, and midbrain (cyan), and the combination of subregions within the cerebellum, pons, and medulla (purple). Interquartile range is shown with the black dashed and dotted lines. D. Box plots showing the importance of cell-type classes in the random forest model for the lower 25th quartile of connections by distance for the Zeisel, et al. dataset, with the SEM computed as the average across ten-fold cross-validation. E. Box plots showing the importance of cell-type features in the random forest model for the upper 75th quartile of connections by distance for the Zeisel, et al. dataset, with the SEM computed as the average across ten-fold cross-validation. F. Scatter plot of long-range versus short-range for all of the individual cell types within both datasets. We highlight in color the most important cell types: telencephalic glutamatergic neurons (Tel Glu; a combination of the Tasic, et al. Neo Glu and Zeisel, et al. Hip Neo Glu cell supertypes), oligodendrocyte subtypes (Oligo), vascular cell types (Vasc), immune cell subtypes (Immune), and astrocyte subtypes (Astro). G. Scatter plots of the Zeisel et al. supertype for all connections (left), short-range connections only (center), and long-range connections only (right), regressed against the regional coefficient of variation (CoV) of the cell-supertype densities. There is a strongly negative and statistically significant negative relationship between F.I. and CoV for all connections and short-range connections, but not long-range connections. Please refer to S. Data Tables 7–10 for the full cell type names and descriptions.

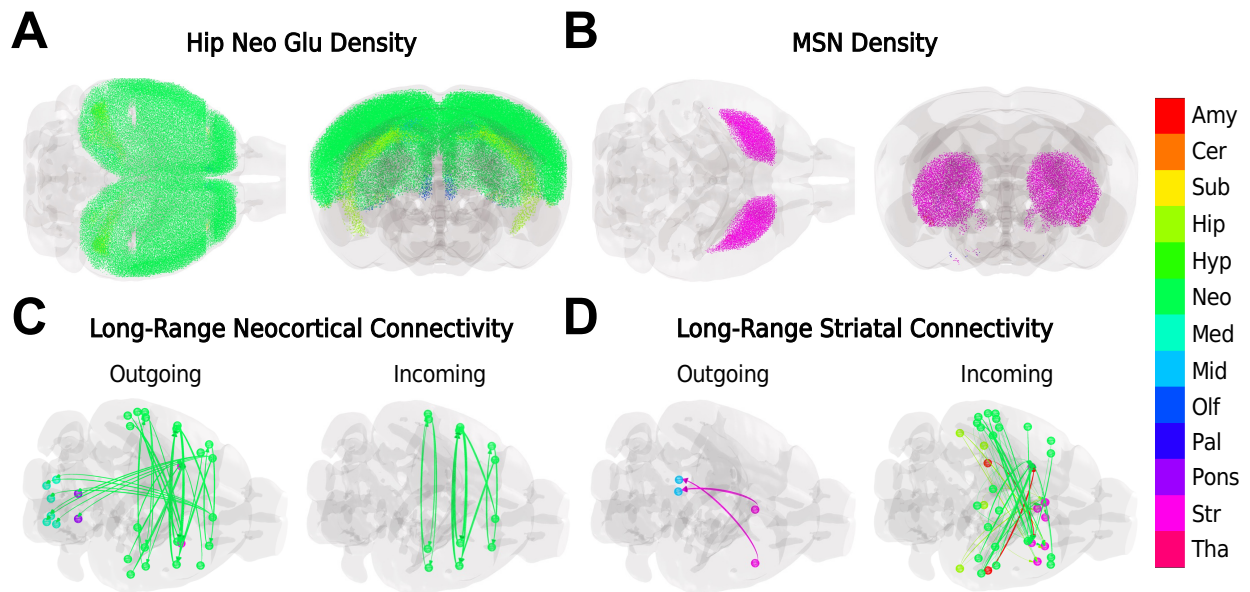
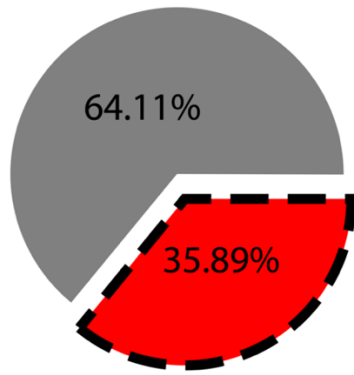


Figure 5.5 Distribution of top contributors to long-range connectivity (from Zeisel, et al. data)

A. Glass-brain representations of the first principal component of Hipp Neo Glu neuronal distributions (number of types = 24). B. Glass-brain representations of the first principal component of MSN neuronal distributions (number of types = 6). C. Glass-brain representations of the long-range connectivity from (Left) and to (Right) neocortical regions. For clarity, only the upper 95th percentile of connections by connectivity density are depicted. D. Glass-brain representations of the long-range connectivity from (Left) and to (Right) striatal regions. For clarity, only the upper 50th percentile of connections by connectivity density are depicted. The colors correspond to the following major region groups: Amy – amygdala; Cer – cerebellum; Sub – cortical subplate; Hip – hippocampus; Hyp – hypothalamus; Neo – neocortex; Med – medulla; Mid – midbrain; Olf – olfactory; Pal – pallidum; Pons – pons; Str – striatum; Tha – thalamus.

A

● Connection ● No connection

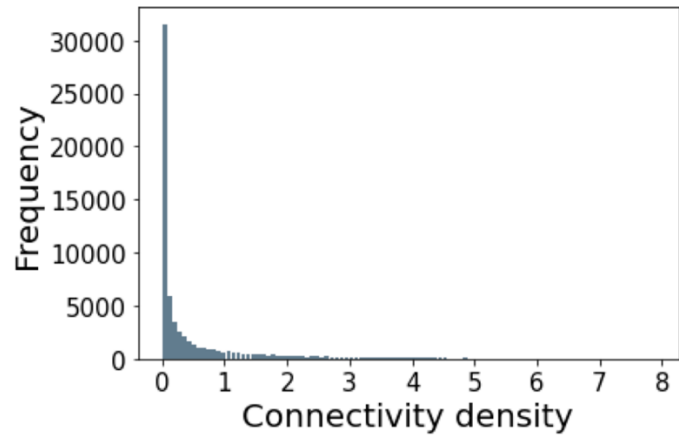
B

Figure 5.6 AMBCA connectivity matrix properties

A. Pie chart showing the sparsity of the connectome. **B.** Histogram showing of the log-transformed connectivity distribution.

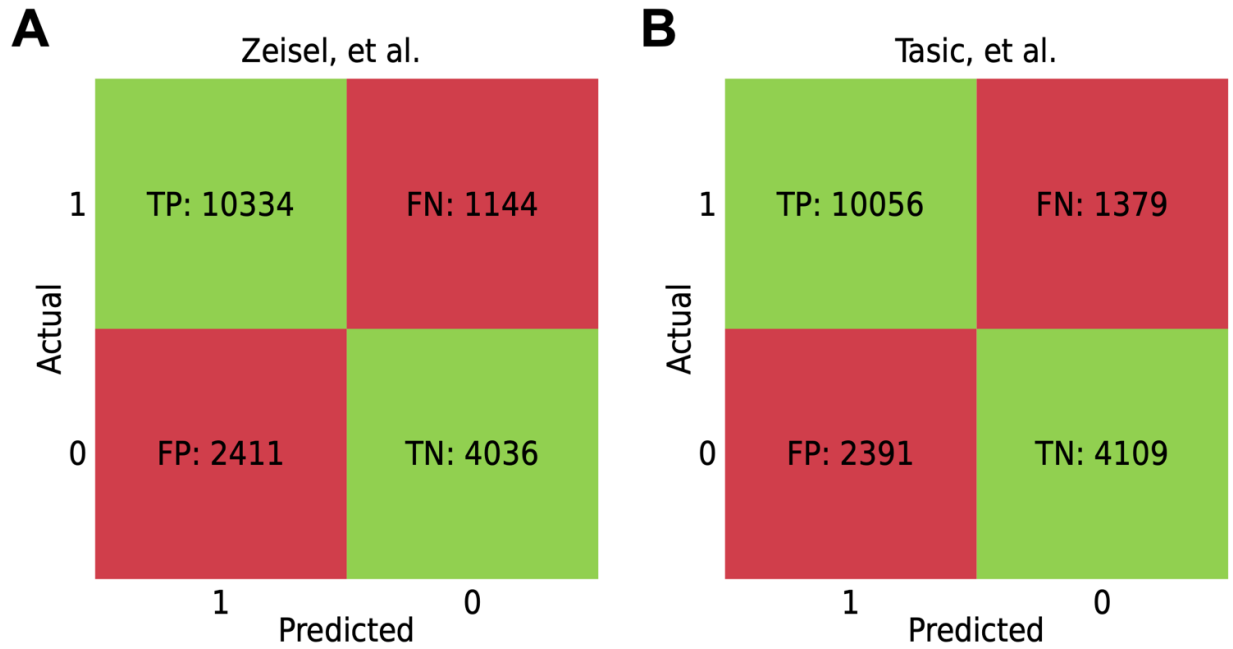


Figure 5.7 Confusion matrices for binary connectome prediction. Performance is shown for both the Zeisel, *et al.*

(A.) and Tasic, *et al.* (B.) datasets.

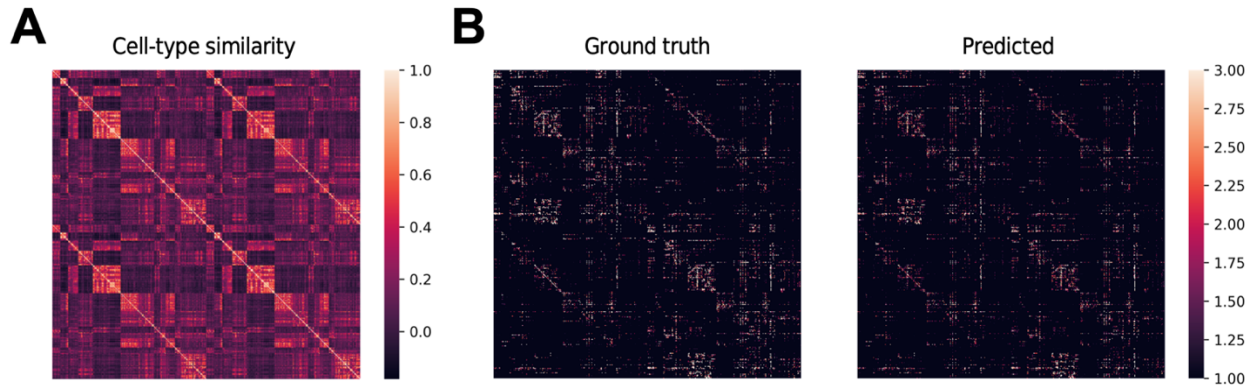


Figure 5.8 Zeisel, *et al.* similarity and connectome prediction, ipsilateral and contralateral.

A. Heatmap of the Zeisel, *et al.* regional cell-type similarity matrix, where both ipsilateral and contralateral hemispheres are shown. **B.** Full ground-truth connectivity matrix (*left*) and the RF model prediction using the Zeisel, *et al.* dataset (*right*). These panels correspond to the ipsilateral-only views in **Figure 2D** and **E**, respectively.

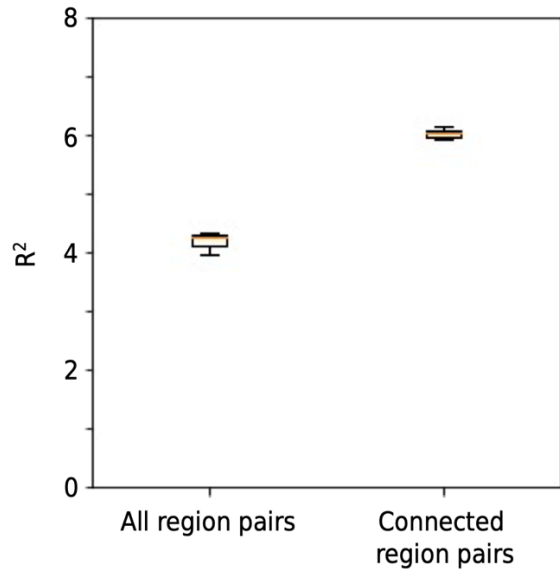


Figure 5.9 Random forest predictions with and without zero-filtering.

Box plots of the test-set performance on predicting connectivity density across 10 cross-validation iterations using all region pairs in the mouse brain (*left*) and only region-pairs with nonzero connectivity density (*right*).

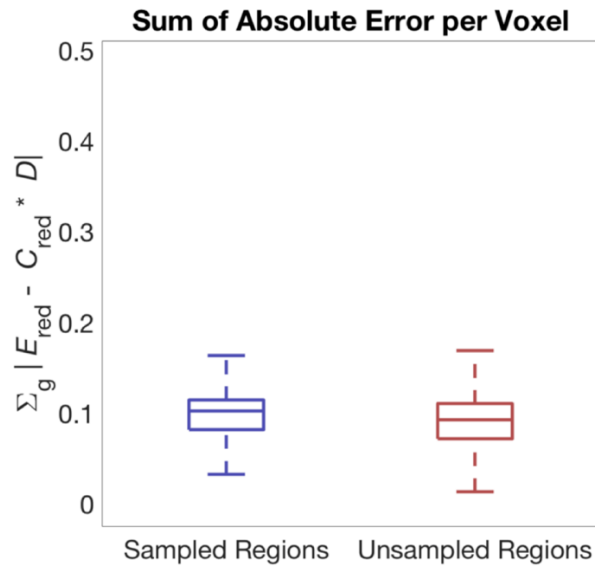


Figure 5.10 MISS in-sample and out-of-sample error for the Tasic, *et al.*, dataset.

Box plots of the sum of squared error (SSE) for the MISS predictions for the Tasic, *et al.* datasets in voxels within regions that were sampled for cell types (*left*) and all others (*right*). Reproduced from Mezas, *et al.*²

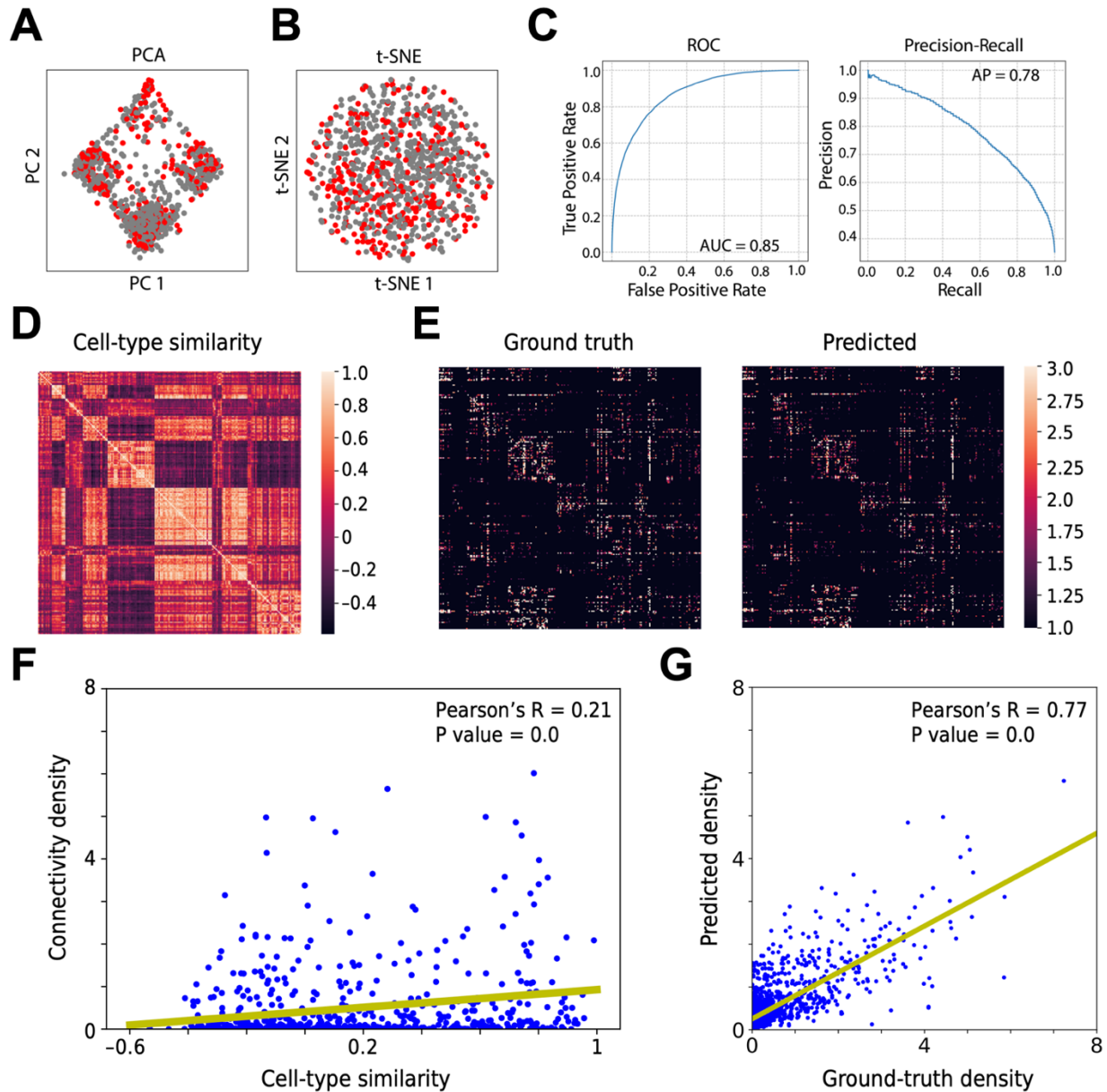


Figure 5.11 Connectivity prediction using the Tasic, *et al.* dataset

A. Principal component analysis (PCA) of the cell type spatial quantification array. **B.** t-distributed stochastic neighbor embedding (t-SNE) of the cell type spatial quantification array. **C.** Performance evaluation of the classifier model using the receiver operating characteristic (ROC) curve (AUROC = 0.85, *left*) and the precision recall curve (AUPR = 0.78, *right*). **D.** Cellular similarity matrix (quantified using Pearson correlation) of spatial cell type enrichment quantification across brain regions. **E.** Brain connectivity matrix (log₂-transformed, ipsilateral only, *left*) and the RF prediction without splitting the training and test set (*right*). The depicted matrices' rows and columns represent individual regions, and the connectivity between regions is denoted by the matrix entries. **F.** Scatter plot of pairwise cellular similarity (**D**) between two regions' cell type distribution vectors versus the log-transformed connectivity strength between the two regions (**E**, *left*), and the fitted linear regression curve (Pearson's R = 0.21, $p = 0.0$). **G.** Scatter plot showing the correlation between the ground truth connectivity strength between all regions pairs with non-zero connectivity and their predicted values for connectivity using cell types as predictors in the RF model (test set only), along with the fitted linear regression curve (Pearson's R = 0.77, $p = 0.0$).

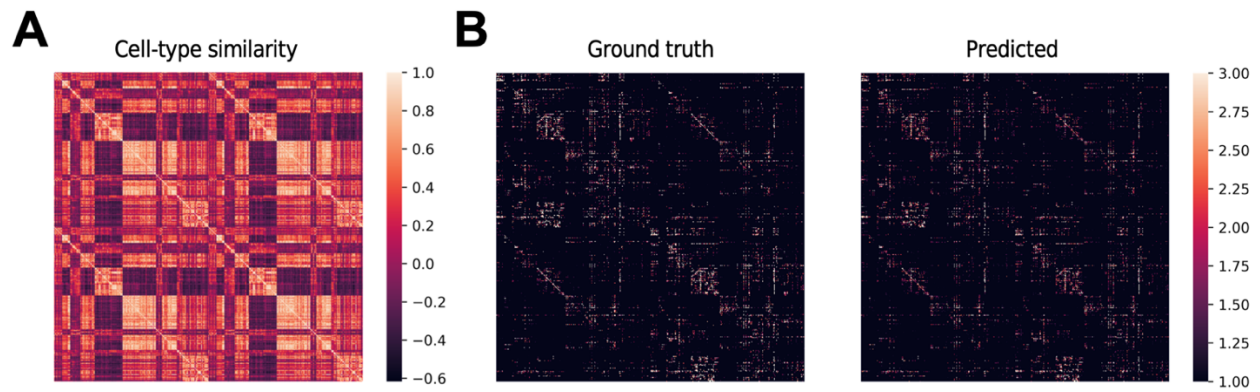


Figure 5.12 Tasic, *et al.* similarity and connectome prediction, ipsilateral and contralateral

A. Heatmap of the Tasic, *et al.* regional cell-type similarity matrix, where both ipsilateral and contralateral hemispheres are shown. B. Full ground-truth connectivity matrix (*left*) and the RF model prediction using the Tasic, *et al.* dataset (*right*). These panels correspond to the ipsilateral-only views in S. Figure 6D and E, respectively.

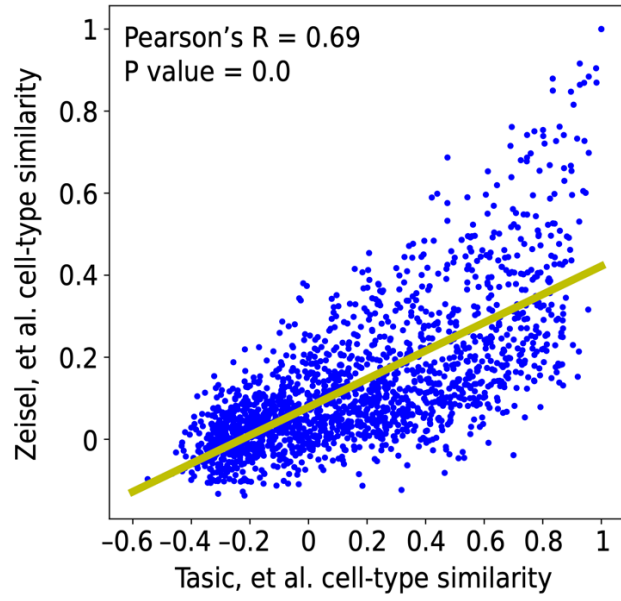


Figure 5.13 Correspondence between *Zeisel, et al.* and *Tasic, et al.* similarity matrices, ipsilateral and contralateral

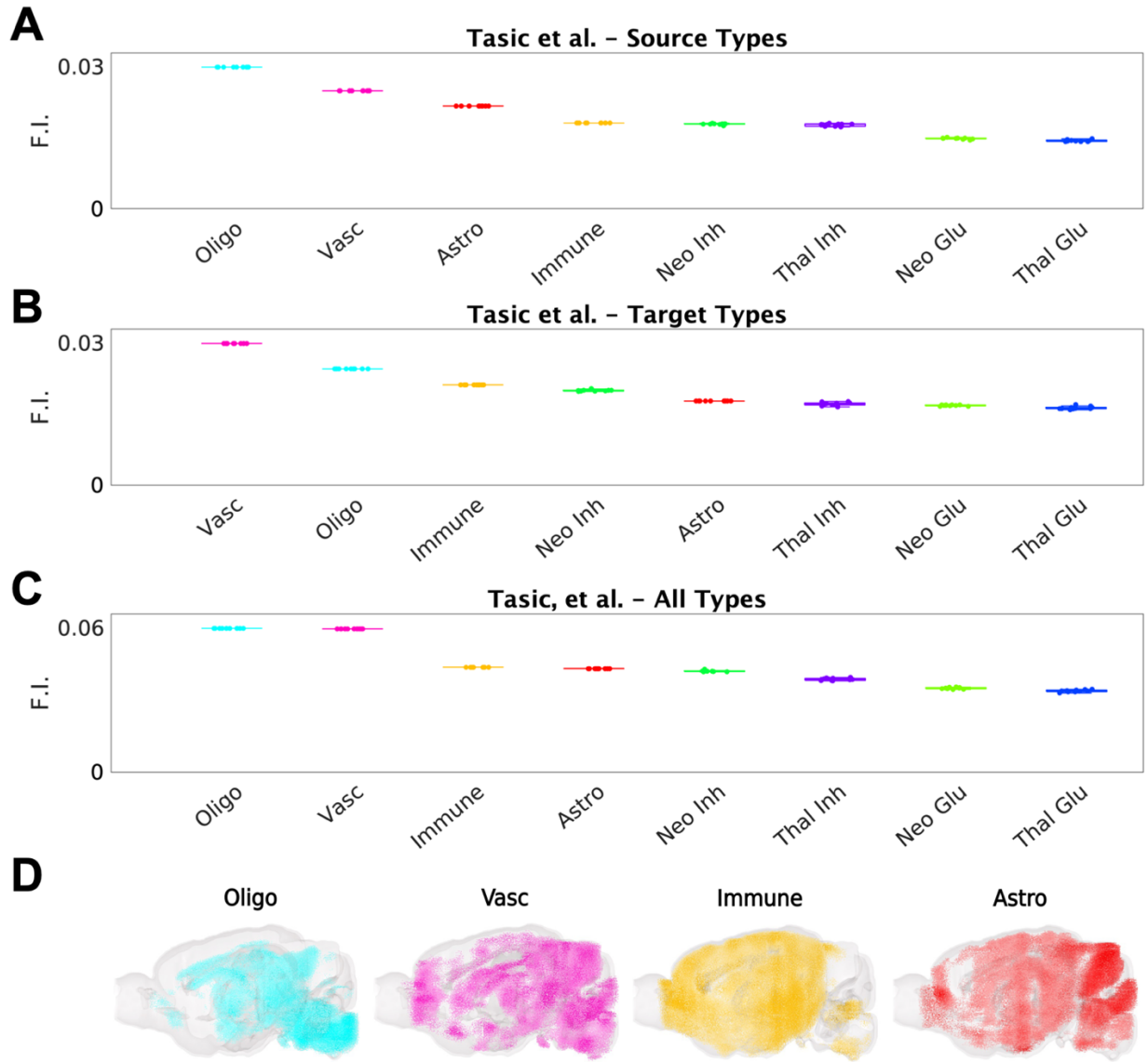


Figure 5.14 Feature importance, Tasic, *et al.* dataset

A. Box plots showing the feature importance values of source-region cell type features in the random forest model for the Tasic, *et al.* cell types, grouped by supertype. B. Feature importance values for target region cell-type features. C. Feature importance values for all cell-type features. D. Sagittal views of cell type densities at the voxel level as inferred by MISS for the corresponding Tasic, *et al.* cell-type classes. The error bars represent the standard error, calculated across ten-fold cross-validation. Please refer to S. Data Tables 7-10 for the full cell type names and description.

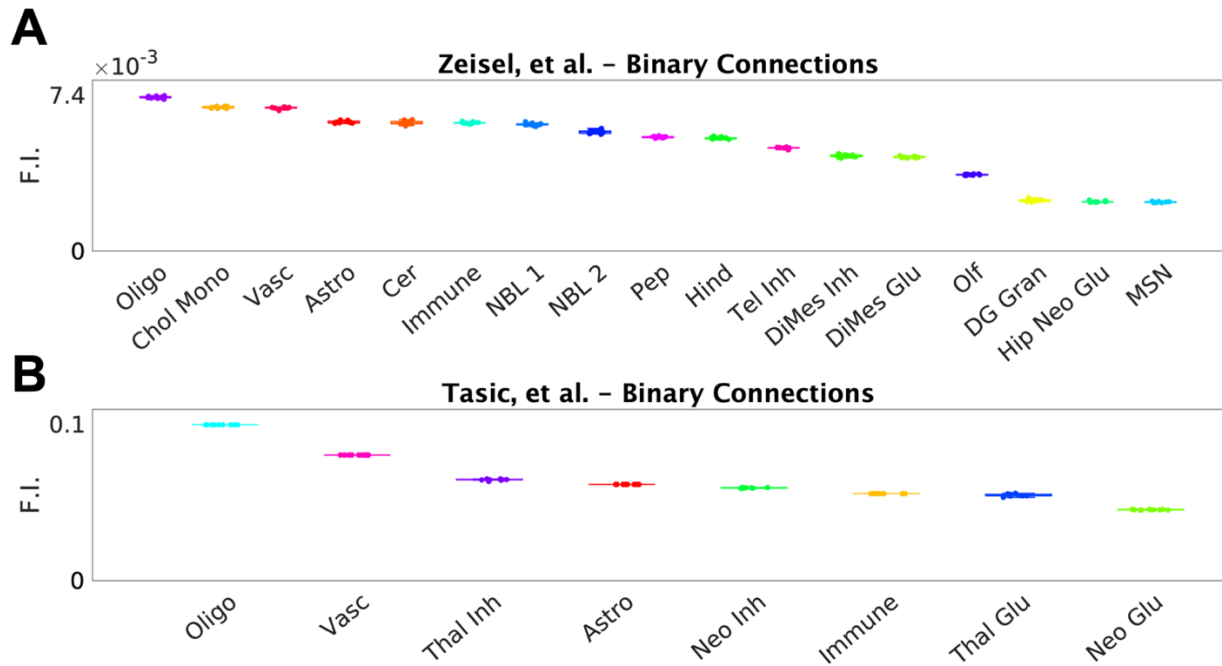


Figure 5.15 Feature importance, binary connectivity prediction

A. Box plots showing the feature importance values of all cell-type features in the random forest model for the classification of region-pairs into connected and unconnected bins, grouped by supertype. **B.** Classification feature importance values for the Tasic, *et al.* cell types. The error bars represent the standard error, calculated across ten-fold cross-validation.

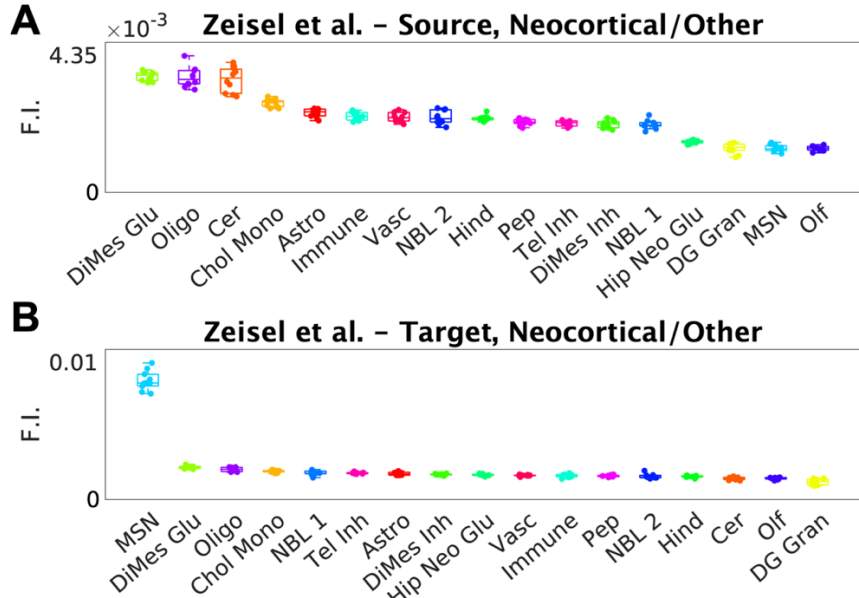


Figure 5.16 Feature importance for neocortical-to-other and other-to-neocortical connectivity density prediction, separated source and target cell-type features, Zeisel, *et al*

A. Box plots showing the feature importance values of all source-region cell-type features in the random forest model for the prediction of the connectivity density values for all neocortical-to-other and other-to-neocortical connections, grouped by supertype. **B.** Feature importance values for target region cell-type features and neocortical-to-other and other-to-neocortical connections only. The error bars represent the standard error, calculated across ten-fold cross-validation.

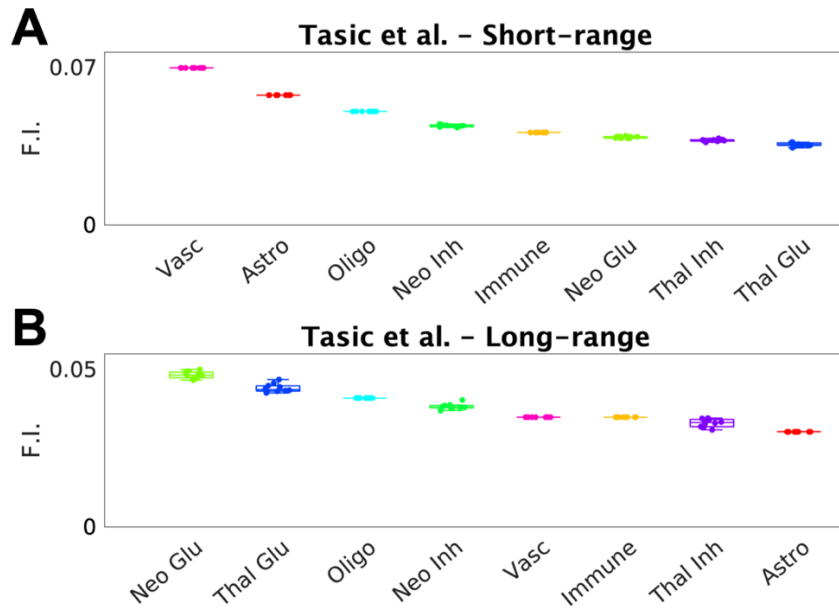


Figure 5.17 Feature importance, short- and long-range connectivity, Tasic, *et al.*

A. Box plots showing the feature importance values of all cell-type features in the random forest model for predicting the lower quartile of connections by inter-regional distance, grouped by supertype. **B.** Feature importance values for predicting the upper quartile of connections by inter-regional distance, grouped by supertype.

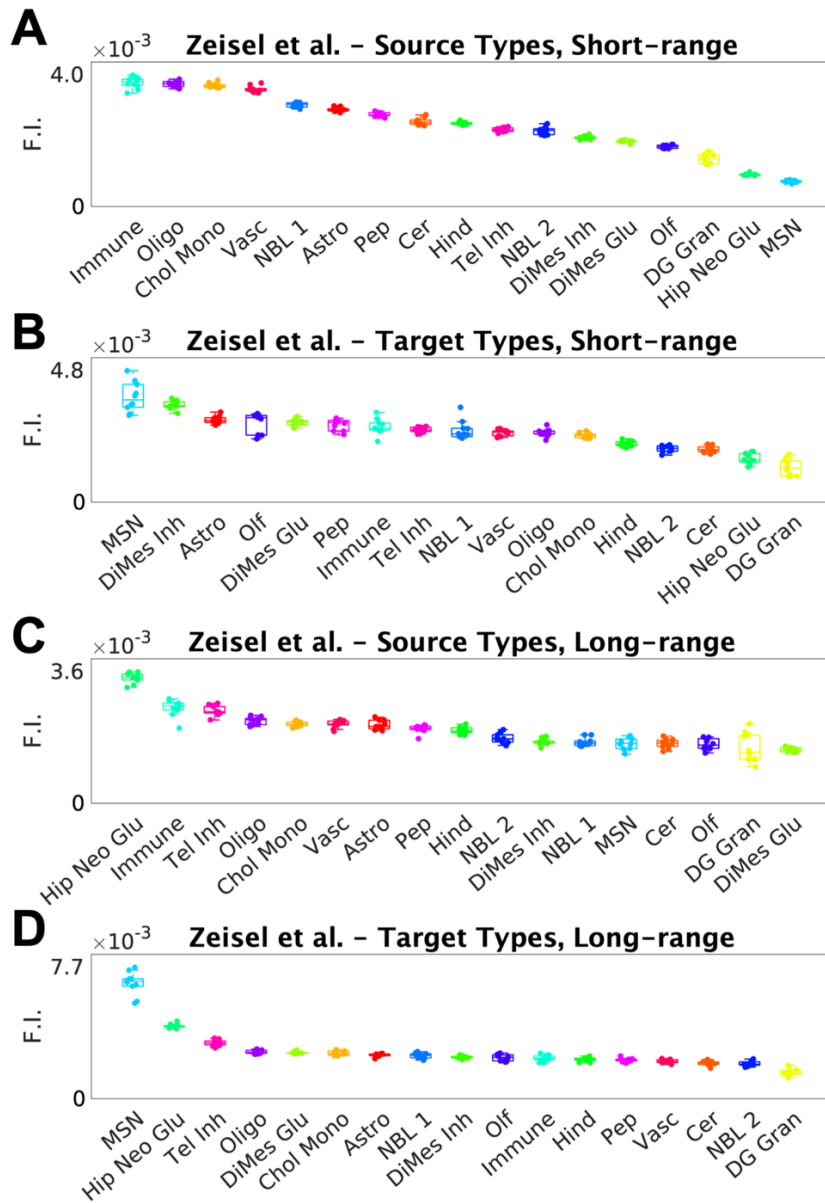


Figure 5.18 Feature importance for short- and long-range connectivity prediction, separated source and target cell-type features, Zeisel, *et al.*

A. Box plots showing the feature importance values of all source-region cell-type features in the random forest model for the prediction of the lower quartile of connectivity density values by inter-regional distance, grouped by supertype. **B.** Feature importance values for target region cell-type features and lower quartile of connectivity values by distance only. **C.** Feature importance values for source region cell-type features and upper quartile of connectivity values by distance only. **D.** Feature importance values for target region cell-type features and upper quartile of connectivity values by distance only.

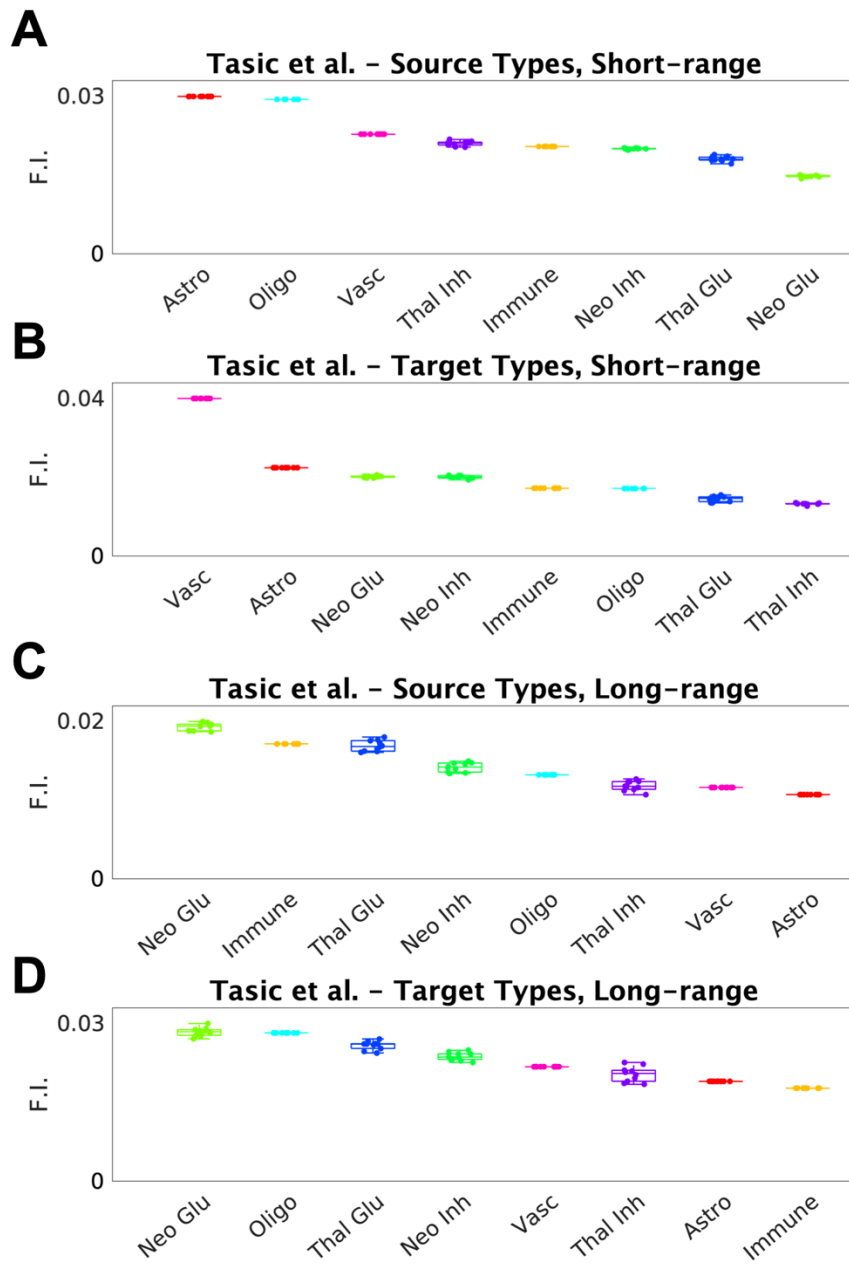


Figure 5.19 Feature importance for short- and long-range connectivity prediction, separated source and target cell-type features, Tasic, *et al*

A. Box plots showing the feature importance values of all source-region cell-type features in the random forest model for the prediction of the lower quartile of connectivity density values by inter-regional distance, grouped by supertype. **B.** Feature importance values for target region cell-type features and lower quartile of connectivity values by distance only. **C.** Feature importance values for source region cell-type features and upper quartile of connectivity values by distance only. **D.** Feature importance values for target region cell-type features and upper quartile of connectivity values by distance only. The error bars represent the standard error, calculated across ten-fold cross-validation.

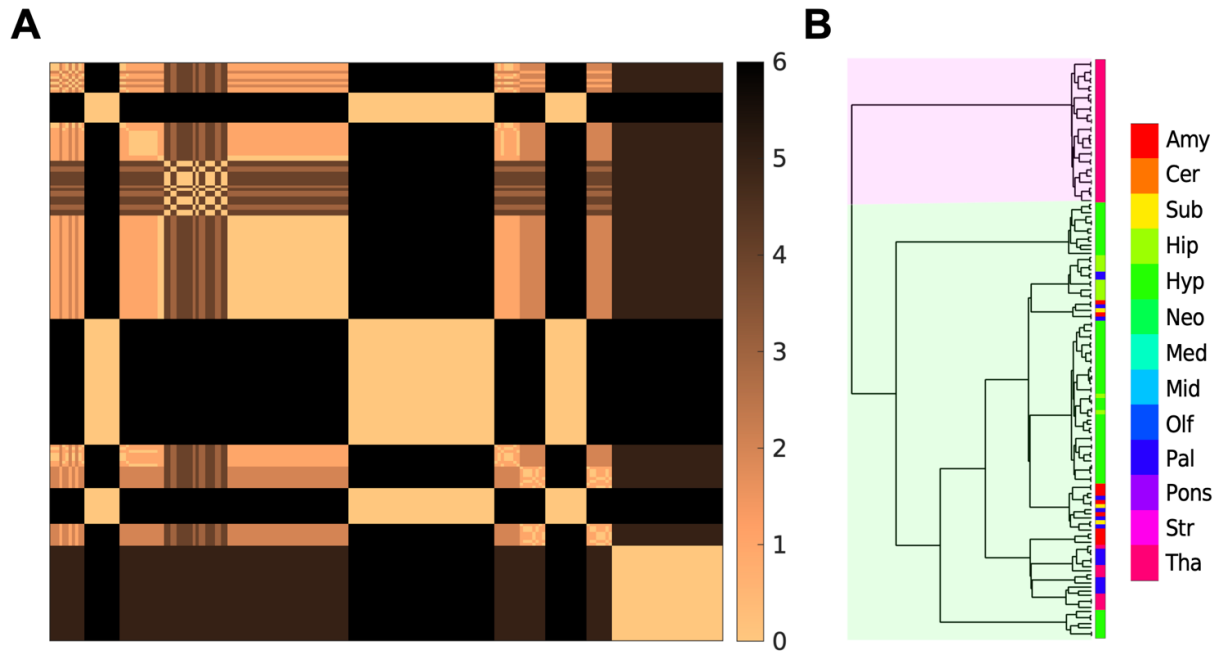


Figure 5.20 Taxonomic distance

A. Heatmap of taxonomic (hierarchical clustering tree distance) between the 212 AIBS regions. Each region-pair was assigned an integer value based on the number of branch points separating them in the hierarchical clustering tree, with 6 being the maximum tree distance. **B.** Hierarchical clustering tree for forebrain regions only (splits 2-6).

Table 5.1 Random forest model performance

Summary of results for the different random forest models explored, broken up by cell-type dataset (Zeisel et al. or Tasic, et al.), model task (classification to predict binary connectivity or regression to predict connectivity density), and connectome subset (portion of the connectivity matrix being predicted). We report accuracy for classification models and mean R² values for the regression models. See also S. Data Tables 2–6.

| Dataset | Model task | Connectome subset | Accuracy | R ² |
|----------------|----------------|---------------------------|----------|----------------|
| Zeisel, et al. | Classification | All | 0.864 | -- |
| Zeisel, et al. | Regression | All | -- | 0.604 |
| Zeisel, et al. | Regression | Short-range | -- | 0.614 |
| Zeisel, et al. | Regression | Long-range | -- | 0.577 |
| Zeisel, et al. | Regression | Neocortical to/from other | -- | 0.585 |
| Tasic, et al. | Classification | All | 0.856 | -- |
| Tasic, et al. | Regression | All | -- | 0.587 |
| Tasic, et al. | Regression | Short-range | -- | 0.608 |
| Tasic, et al. | Regression | Long-range | -- | 0.581 |

Table 5.2 Dataset information for each of the three source datasets used

| Dataset Name | Description | Reprocessing Methods |
|---|--|---|
| AMBCA² | The Allen Mouse Brain Connectivity Atlas is a mesoscale connectome representing the whole-brain wiring diagram. Brain-wide, region-specific axonal projections were mapped into a common 3D space using viral tracing. | See Methods |
| Tasic, <i>et al.</i> scRNAseq dataset^{18, 20} | Single-cell RNA sequencing data obtained by the AIBS from three distinct brain regions (primary visual cortex, secondary motor cortex, and lateral geniculate complex), representing data from 25557 individual cells. | Grouped into 25 distinct cell types and mapped using MISS ¹⁶ |
| Zeisel, <i>et al.</i> scRNAseq dataset¹⁹ | A single-cell RNA sequencing data consisting of 144147 individual cells sampled from twelve regions throughout the mouse brain. | 200 cell types mapped using MISS ¹⁶ |

Table 5.3 Model comparison for classification, Zeisel, *et al*

| Methods | AUROC score | Accuracy |
|------------------------|--------------------|-----------------|
| Random Forest | 0.864 | 0.798 |
| Ridge | 0.557 | 0.658 |
| Lasso | 0.579 | 0.653 |
| Support Vector Machine | 0.685 | 0.766 |
| MLP Classifier | 0.702 | 0.744 |

Table 5.4 Model comparison for regression, Zeisel, *et al*

| Methods | Adjusted R² score | RMSE |
|------------------------|-------------------------------------|-------------|
| Random Forest | 0.604 | 0.575 |
| Ridge | 0.043 | 0.919 |
| Lasso | 0.088 | 0.926 |
| Support Vector Machine | 0.377 | 0.791 |
| MLP Regressor | 0.550 | 0.654 |

Table 5.5 Model performance of training and testing the random forest model using two different methods

Model performance of training and testing the random forest model using two different methods of creating training and test sets: purely random (*left*, see also **Table 1**) and separating by source region (*right*).

| Dataset | Purely Random (R^2) | Source-Region Separated (R^2) |
|-----------------------|---|---|
| <i>Zeisel, et al.</i> | 0.604 | 0.502 |
| <i>Tasic, et al.</i> | 0.587 | 0.465 |

Table 5.6 Model comparison for classification, Tasic, *et al.*

| Methods | AUROC score | Accuracy |
|------------------------------|--------------------|-----------------|
| Random Forest | 0.856 | 0.790 |
| Ridge | 0.530 | 0.647 |
| Lasso | 0.578 | 0.641 |
| Support Vector Machine | 0.673 | 0.738 |
| MLP Classifier | 0.712 | 0.759 |
| Decision Tree Classifier | 0.704 | 0.727 |
| Gradient Boosting Classifier | 0.605 | 0.694 |
| Extra Trees Classifier | 0.687 | 0.711 |
| K-Neighbors Classifier | 0.744 | 0.774 |

Table 5.7 Model comparison for regression, Tasic, *et al.*

| Methods | Adjusted R² score | RMSE |
|-----------------------------|-------------------------------------|-------------|
| Random Forest | 0.587 | 0.605 |
| Ridge | 0.060 | 0.911 |
| Lasso | 0.061 | 0.910 |
| Support Vector Machine | 0.363 | 0.750 |
| MLP Regressor | 0.515 | 0.654 |
| Decision Tree Regressor | 0.121 | 0.881 |
| Gradient Boosting Regressor | 0.328 | 0.770 |
| Extra Trees Regressor | 0.599 | 0.595 |
| K-Neighbors Regressor | 0.501 | 0.664 |

5.10 REFERENCES

- [1] Sporns, O., Tononi, G., and Kotter, R. (2005). PLoS Comput Biol The human connectome: A structural description of the human brain. PLoS Comput. Biol. 1, e42.
- [2] Oh, S.W., Harris, J.A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A.M., et al. (2014). A mesoscale connectome of the mouse brain. Nature 508, 207–214. <https://doi.org/10.1038/nature13186>.
- [3] Bullmore, E.T., and Bassett, D.S. (2011). Annu Rev Clin Psychol Brain graphs: graphical models of the human brain connectome. Annu. Rev. Clin. Psychol. 7, 113–140.
- [4] Zeng, H. (2018). Mesoscale connectomics. Curr Opin Neurobiol Mesoscale connectomics. Curr Opin Neurobiol 50, 154–162.
- [5] French, L., and Pavlidis, P. (2011). Relationships between gene expression and brain wiring in the adult rodent brain. PLoS Comput. Biol. 7, e1001049. <https://doi.org/10.1371/journal.pcbi.1001049>.
- [6] Tan, P.P.C., French, L., and Pavlidis, P. (2013). Neuron-enriched gene expression patterns are regionally anti-correlated with oligodendrocyte-enriched patterns in the adult mouse and human brain. Front. Neurosci.
- [7] Henriksen, S., Pang, R., and Wronkiewicz, M. (2016). A simple generative model of the mouse mesoscale connectome. Elife 5, e12366. <https://doi.org/10.7554/elife.12366>.
- [8] Fulcher, B.D., and Fornito, A. (2016). A transcriptional signature of hub connectivity in the mouse connectome. Proc. Natl. Acad. Sci. USA 113, 1435–1440.
- [9] Reimann, M.W., Gevaert, M., Shi, Y., Lu, H., Markram, H., and Muller, E. (2019). A null model of the mouse whole-neocortex micro-connectome. Nat. Commun. 10, 3903. <https://doi.org/10.1038/s41467-019-11630-x>.
- [10] Goel, P., Kuceyeski, A., Locastro, E., and Raj, A. (2014). Spatial patterns of genome-wide expression profiles reflect anatomic and fiber connectivity architecture of healthy human brain. Human Brain Mapping 35, 4204–4218. <https://doi.org/10.1002/hbm.22471>.

- [11] Ve'rtes, P.E., Rittman, T., Whitaker, K.J., Romero-Garcia, R., Va' sa, F., Kitzbichler, M.G., Wagstyl, K., Fonagy, P., Dolan, R.J., Jones, P.B., et al. (2016). Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371, 20150362.
- [12] Diez, I., and Sepulcre, J. (2018). Neurogenetic profiles delineate large-scale connectivity dynamics of the human brain. *Nature communications* 9.
- [13] Shen, X., Finn, E.S., Scheinost, D., Rosenberg, M.D., Chun, M.M., Papademetris, X., and Constable, R.T. (2017). Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* 12, 506–518. <https://doi.org/10.1038/nprot.2016.178>.
- [14] Ji, S., Fakhry, A., and Deng, H. (2014). Integrative analysis of the connectivity and gene expression atlases in the mouse brain. *Neuroimage* 84, 245–253.
- [15] Huang, L., Kechschull, J.M., Furth, D., Musall, S., Kaufman, M.T., Church- € land, A.K., and Zador, A.M. (2020). BRICseq bridges brain-wide interregional connectivity to neural activity and gene expression in single animals. *Cell* 182, 177–188.e27. <https://doi.org/10.1016/j.cell.2020.05.029>.
- [16] Mezas, C., Torok, J., Maia, P.D., Markley, E., and Raj, A. (2022). Matrix inversion and subset selection (miss): A pipeline for mapping of diverse cell types across the murine brain. *Proc. Natl. Acad. Sci. USA* 119, e2111786119.
- [17] Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168–176. <https://doi.org/10.1038/nature05453>.
- [18] Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. *Nature* 563, 72–78. <https://doi.org/10.1038/s41586-018-0654-5>.
- [19] Zeisel, A., Hochgerner, H., Lo" nnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Ha" rring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular architecture of the mouse nervous system. *Cell* 174, 999–1014.e22. <https://doi.org/10.1016/j.cell.2018.06.021>.

- [20] Allen, A.I.B.S. (2018). Cell Types Database - Technical White Paper: Transcriptomics.
- [21] Abdelnour, F., Voss, H.U., and Raj, A. (2014). Network diffusion accurately models the relationship between structural and functional brain connectivity networks. *Neuroimage* 90, 335–347. <https://doi.org/10.1016/j.neuroimage.2013.12.039>.
- [22] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- [23] Fakhry, A., and Ji, S. (2015). High-resolution prediction of mouse brain connectivity using gene expression patterns. *Methods* 73, 71–78. <https://doi.org/10.1016/j.ymeth.2014.07.011>.
- [24] Fornito, A., Arnatkeviciute, A., and Fulcher, B.D. (2019). Bridging the gap between connectome and transcriptome. *Trends Cognit. Sci.* 23, 34–50. <https://doi.org/10.1016/j.tics.2018.10.005>.
- [25] Anaissi, A., Goyal, M., Catchpoole, D.R., Braytee, A., and Kennedy, P.J. (2016). Ensemble feature learning of genomic data using support vector machine. *PLoS One* 11, e0157330. <https://doi.org/10.1371/journal.pone.0157330>.
- [26] Allen, A.I.B.S. (2013). Developing Mouse Brain Atlas - Technical White Paper: Reference Atlases for the Allen Developing Mouse Brain Atlas.
- [27] Ouyang, M., Kang, H., Detre, J.A., Roberts, T.P.L., and Huang, H. (2017). Short-range connections in the developmental connectome during typical and atypical brain maturation. *Neurosci. Biobehav. Rev.* 83, 109–122. <https://doi.org/10.1016/j.neubiorev.2017.10.007>.
- [28] Naze, S., Proix, T., Atasoy, S., and Kozloski, J.R. (2021). Robustness of connectome harmonics to local gray matter and long-range white matter connectivity changes. *Neuroimage* 224, 117364. <https://doi.org/10.1016/j.neuroimage.2020.117364>.
- [29] Pfeiffer, S.E., Warrington, A.E., and BANSAL, R. (1993). The oligodendrocyte and its many cellular processes. *Trends Cell Biol.* 3, 191–197. [https://doi.org/10.1016/0962-8924\(93\)90213-k](https://doi.org/10.1016/0962-8924(93)90213-k).
- [30] Emery, B. (2010). Regulation of oligodendrocyte differentiation and myelination. *Science* 330, 779–782. <https://doi.org/10.1126/science.1190927>.

- [31] Eroglu, C., and Barres, B.A. (2010). Regulation of synaptic connectivity by glia. *Nature* 468, 223–231. <https://doi.org/10.1038/nature09612>. 3
- [32] Kawamura, A., Abe, Y., Seki, F., Katayama, Y., Nishiyama, M., Takata, N., Tanaka, K.F., Okano, H., and Nakayama, K.I. (2020). Chd8 mutation in oligodendrocytes alters microstructure and functional connectivity in the mouse brain. *Mol. Brain* 13, 160. <https://doi.org/10.1186/s13041-020-00699-x>.
- [33] Wang, F., Yang, Y.J., Yang, N., Chen, X.J., Huang, N.X., Zhang, J., Wu, Y., Liu, Z., Gao, X., Li, T., et al. (2018). Enhancing oligodendrocyte myelination rescues synaptic loss and improves functional recovery after chronic hypoxia. *Neuron* 99, 689–701.e5. <https://doi.org/10.1016/j.neuron.2018.07.017>.
- [34] Buchanan, J., Elabbady, L., Collman, F., Jorstad, N.L., Bakken, T.E., Ott, C., Glatzer, J., Bleckert, A.A., Bodor, A.L., Brittan, D., et al. (2021). Oligodendrocyte precursor cells prune axons in the mouse neocortex. Preprint at bioRxiv. <https://doi.org/10.1101/2021.05.29.446047>.
- [35] Abbott, N.J., Patabendige, A.A.K., Dolman, D.E.M., Yusof, S.R., and Begley, D.J. (2010). Structure and function of the blood–brain barrier. *Neurobiol. Dis.* 37, 13–25. <https://doi.org/10.1016/j.nbd.2009.07.030>.
- [36] Langen, U.H., Ayloo, S., and Gu, C. (2019). Development and cell biology of the blood-brain barrier. *Annu. Rev. Cell Dev. Biol.* 35, 591–613. <https://doi.org/10.1146/annurev-cellbio-100617-062608>.
- [37] Chow, B.W., and Gu, C. (2015). The molecular constituents of the blood– brain barrier. *Trends Neurosci.* 598–608. <https://doi.org/10.1016/j.tins.2015.08.003>.
- [38] Daneman, R., and Prat, A. (2015). The blood–brain barrier. *Cold Spring Harbor Perspect. Biol.* 7, a020412. <https://doi.org/10.1101/cshperspect.a020412>.
- [39] Ballabh, P., Braun, A., and Nedergaard, M. (2004). The blood–brain barrier: an overview. *Neurobiol. Dis.* 16, 1–13. <https://doi.org/10.1016/j.nbd.2003.12.016>.
- [40] Cauli, B., and Hamel, E. (2010). Revisiting the role of neurons in neurovascular coupling. *Front. Neuroenergetics* 2, 9. <https://doi.org/10.3389/fnene.2010.00009>.

- [41] Chow, B.W., Nunez, V., Kaplan, L., Granger, A.J., Bistrong, K., Zucker, H.L., Kumar, P., Sabatini, B.L., and Gu, C. (2020). Caveolae in CNS arterioles mediate neurovascular coupling. *Nature* 579, 106–110. <https://doi.org/10.1038/s41586-020-2026-1>.
- [42] Kaplan, L., Chow, B.W., and Gu, C. (2020). Neuronal regulation of the blood–brain barrier and neurovascular coupling. *Nat. Rev. Neurosci.* 21, 416–432. <https://doi.org/10.1038/s41583-020-0322-2>.
- [43] Jafari, A., de Lima Xavier, L., Bernstein, J.D., Simonyan, K., and Bleier, B.S. (2021). Association of sinonasal inflammation with functional brain connectivity. *JAMA Otolaryngol. Head Neck Surg.* 147, 534–543.
- [44] Morimoto, K., and Nakajima, K. (2019). Role of the Immune System in the Development of the Central Nervous System. *Front. Neurosci.* 13, 916. <https://doi.org/10.3389/fnins.2019.00916>.
- [45] Allen, N.J., and Eroglu, C. (2017). Cell biology of astrocyte-synapse interactions. *Neuron* 96, 697–708.
- [46] Chuhma, N., Tanaka, K.F., Hen, R., and Rayport, S. (2011). Functional Connectome of the Striatal Medium Spiny Neuron. *J. Neurosci.* 31, 1183–1192. <https://doi.org/10.1523/JNEUROSCI.3833-10.2011>.
- [47] Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361, eaat5691. <https://doi.org/10.1126/science.aat5691>.
- [48] Codeluppi, S., Borm, L.E., Zeisel, A., La Manno, G., van Lunteren, J.A., Svensson, C.I., and Linnarsson, S. (2018). Spatial organization of the somatosensory cortex revealed by osmfish. *Nat. Methods* 15, 932–935. <https://doi.org/10.1038/s41592-018-0175-z>.
- [49] Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., Rubinstein, N.D., Hao, J., Regev, A., Dulac, C., and Zhuang, X. (2018). Molecular, spatial and functional single-cell profiling of the hypothalamic preoptic region. *Science* 362, eaau5324. <https://doi.org/10.1126/science.aau5324>.

- [50] Zhang, M., Eichhorn, S.W., Zingg, B., Yao, Z., Cotter, K., Zeng, H., Dong, H., and Zhuang, X. (2021). Spatially resolved cell atlas of the mouse primary motor cortex by merfish. *Nature* 598, 137–143. <https://doi.org/10.1038/s41586-021-03705-x>.
- [51] Chen, X., Fischer, S., Zhang, A., Gillis, J., and Zador, A.M. (2022). Modular cell type organization of cortical areas revealed by in situ sequencing. Preprint at bioRxiv 598. <https://doi.org/10.1101/2022.11.06.515380>.
- [52] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. <https://doi.org/10.1109/TPAMI.2005.159>.
- [53] Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 832–844.
- [54] Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.
- [55] Qi, Y. (2012). Random forest for bioinformatics. *Ensemble Machine Learning*, 307–323 (Springer).
- [56] Segal, M.R. (2004). *Machine Learning Benchmarks and Random Forest Regression* (UCSF: Center for Bioinformatics and Molecular Biostatistics).
- [57] Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.
- [58] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* 58, 267–288.
- [59] Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152.
- [60] Chang, C.-C., Yu, S.C., McQuoid, D.R., Messer, D.F., Taylor, W.D., Singh, K., Boyd, B.D., Krishnan, K.R.R., MacFall, J.R., Steffens, D.C., and Payne, M.E. (2011). Libsvm: a library for support vector machines. *Psychiatr. Res.* 193, 1–6.

6 Chapter 6: DeepHeme: A High-Performance, Generalizable, Deep Ensemble for Bone Marrow Morphometry and Hematologic Diagnosis

Shenghuan Sun¹, Jacob Van Cleave², Linlin Wang³, Brenda Fried², Zhanghan Yin⁸, Fabienne Lucas⁴, Leonardo Boiocchi², Hasan Bilal², Laura Brown³, Jacob D. Spector⁵, Orly Ardon², Leonardo Boiocchi², Irem Sahver Isgor², Rohan Sardana², Jeeyeon Baik², Menglei Zhu², Mikhail Roshal², Chuanyi M. Lu^{3, 6}, Aijaz Syed², Dmitry B. Goldgof⁷, Iain, Ahmet Dogan², Sonam Prakash³, Atul J. Butte¹, and Gregory M. Goldgof²

¹Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA

²Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA

³Department of Laboratory Medicine, University of California, San Francisco, CA, USA

⁴Department of Pathology, Brigham and Women's Hospital/Harvard Medical School, Boston, MA, USA

⁵Department of Laboratory Medicine, Boston Children's Hospital/Harvard Medical School, Boston, MA, USA

⁶Department of Laboratory Medicine, Veterans Affairs Medical Center, San Francisco, CA, USA

⁷Department of Computer Science, University of South Florida, Tampa, FL, USA

⁸Department of Statistics, University of California, Berkeley, CA, USA

6.1 ABSTRACT

Histomorphology analysis of the bone marrow aspirate (BMA) is pivotal for the diagnostic workup of a broad range of hematological disorders. However, this skill is time consuming, error prone, and highly complex, requiring years to master, as it is among the most technically challenging in the field of pathology due to the number of cell types involved. Convolutional neural networks-based models for the automatic classification of bone marrow cell morphology demonstrate significant potential in improving diagnostic efficiency and accuracy, however, existing deep learning approaches in this field fall short in expert-level performance and generalizability outside of a single dataset. Working with multiple hematologists, we curated an extensive dataset of 41,595 consensus-annotated single-cell images spanning 23 morphological classes from BMA whole slide images (WSIs) from two academic centers and trained DeepHeme-SE, a bone marrow classification tool based on a snapshot ensemble deep learning framework. DeepHeme-SE outperforms previous models across in accuracy while expanding the total number of differentiable cell classes. In addition, DeepHeme SE is the first to show generalizability across different medical centers. Finally, we conducted systematic comparisons with three medical experts from differing academic hospitals, demonstrating that DeepHeme-SE can match or surpass the diagnostic accuracy of human experts at cell classification while improving reproducibility. Accurate and generalizable cell classification represents a significant step towards automated analysis of hematopathology slides and the development of quantitative morphology predictive biomarkers.

6.2 INTRODUCTION

Bone marrow aspirates (BMAs) are pivotal in diagnosing a spectrum of hematologic diseases, accounting for approximately 10% of global cancer cases and deaths[1], [2]. The accurate categorization of morphologic cell types is essential for diagnostic accuracy, prognosis determination, and treatment planning[3], [4]. However, BMA analysis is technical complex, time consuming, and suffers from interobserver variability. This underscores an urgent need for advanced, automated approaches^{5, 6}. Recent studies have demonstrated the potential of automated bone marrow cell classification using both whole

slide images (WSIs)[7–9] and microscope camera images[10], [11]. Specifically, Chandradevan et al. employed a combination of Faster R-CNN and CNN networks for the detection and classification of cells within manually-annotated BMA regions[7]. Similarly, Matek et al. and Tayebi et al. have contributed significantly by developing extensive expert-annotated datasets to train CNN-based models for cell classification[9], [11]. In their approach, Matek et al. collaborated with clinical laboratory staff to assemble their dataset, whereas Tayebi et al. worked closely with hematopathologists for dataset collection. More recently, Lewis et al. have advanced the field by introducing a fully automated pipeline for bone marrow cell counts⁸. However, a notable gap in the existing literature is the lack of external validation to demonstrate the generalizability of these models. While Matek et al. attempted such analysis, the results were not optimal[11]. External validation is essential to ascertain the reliability and robustness of these models, especially when faced with domain shifts. Moreover, to the best of our knowledge, few studies have reported an average F1 performance exceeding 0.8 in the context of automated bone marrow cell classification[8], [9], [11–13]. Two primary challenges impede progress in this area: the limited quality of datasets and the underutilization of advanced deep learning techniques. A significant limitation of most current datasets is their annotation, often not verified by a consensus among hematologists. For instance, Chandradevan et al.[7] utilized a dataset annotated through hematologist consensus, yet it comprised only 9,269 images. Another critical issue pertains to the predictive performance for rare classes, which is predominantly hindered by insufficient data. For example, the smallest class in the datasets of Chandradevan et al.[7], Matek et al.[11], and Tayebi et al.[9] contained only 62, 8, and 7 images, respectively. In such scenarios, regardless of the sophistication of the deep learning models employed, the performance tends to be suboptimal. Building on the theme of enhancing model performance, one effective strategy is the application of ensemble methods, which often outperform single models[14], [15] and have proven effective in multiple medical applications[16–18]. In the multi-class setting of bone marrow cell typing, a single network might not achieve optimal accuracy for all classes simultaneously. Standard learning rate schedulers typically focus on minimizing overall loss. However, even with strategies like

weighted loss functions and up/down sampling, a model cannot guarantee optimal performance for each individual class[19], [20].

As the learning rate decreases, the model may improve predictions for certain classes, yet at the expense of others, leading to imbalanced performance. To date, ensemble methods have not been explored in the realm of bone marrow cell typing, potentially due to the increased computational demands of training multiple models. Ensemble methods are also analogous to the consensus conferences held by real world pathologists, who each day discuss the most difficult cases with their colleagues, thereby creating a diagnosis that is an ensemble of experts. It is methodologically intriguing to investigate whether ensemble techniques could offer a practical solution for improving bone marrow cell typing, while concurrently minimizing the need for additional computational resources. In this study, we sought to investigate how ensemble and consensus approaches may improve AI-based bone marrow classification. In collaboration with hematopathologists from various institutions, we have curated a dataset comprising 41,595 pathologist consensus annotated single-cell images, spanning 23 morphological classes. The method we developed, DeepHeme-SE, uses the snapshot ensemble technique to enhance the bone marrow cell classification without compromising on computational efficiency in terms of memory usage and processing time[21].

6.3 RESULTS

DEEPHEME-SE: AI-POWERED BONE MARROW MORPHOMETRY

To address the challenges of developing a reliable bone marrow cell classifier, we introduce the DeepHeme-SE framework (Figure 6.1). DeepHeme-SE aims to achieve state-of-the-art performance by focusing on two key aspects: dataset quality and methodological innovation. One novelty we have is we curated a comprehensive library of 41,595 images, categorized into 23 distinct classes through a consensus among a panel of three expert hematopathologists (Figure 6.1A, Figure 6.7). This dataset is, to our knowledge, one

of the first in bone marrow cell typing to be annotated by a consensus of hematopathology specialists. To thoroughly test model adaptability, we included an independent cohort from an unrelated academic medical center. This external test set was purposefully designed to capture the broad variability typical across domains, reflecting differences in slide preparation, staining protocols, scanning equipment, populations and disease phenotype. We released a dataset for benchmark evaluation that will be available on the DeepHeme web application. Our approach diverges from traditional methods, which typically depend on a single, optimized deep learning model. By incorporating snapshot ensemble techniques, we enable the use of ensemble methods in deep learning-based predictions. This innovation allows us to bypass the significant computational expenses usually associated with such processes. (Figure 6.1 B). We performed comprehensive downstream analysis to show the capability of DeepHeme-SE including cell typing/counting, feature embedding and clustering as well as interpretability analysis (Figure 6.1 C). Last but not least, we built a web application to allow scientists to interact with DeepHeme algorithms (Figure 6.1 D).

SNAPSHOT ENSEMBLE ADVANCES BONE MARROW CELL CLASSIFICATION

DeepHeme-SE, leveraging the snapshot ensemble technique, enhances model performance by harnessing the benefits of ensemble methods. This approach simultaneously avoids the high computational costs typically incurred, thanks to the efficient snapshot nature employed during training. The traditional learning rate scheduler generally decrease as the validation error decrease and the increase of iteration. Snapshot ensemble, on the other hand, has a cyclic learning rate scheduler (Figure 6.8). By increasing the learning rate dramatically in a periodic manner, we can obtain several models with different weights . The training loss increased every time we boosted the learning rate but the model will converged to a different local minimal when we decreased the learning rate again (Figure 6.8). After the training, we perform a set of analysis including Venn Diagram(Figure 6.9) and UpSet plot(Figure 6.10) to show that each model have their similarity but also share a noticeable level of diversity. This demonstrated snapshot ensemble, even

with the same architectures, is able to produce a set of models that are distinguishing from each other(Figure 6.11). Given the diversity of insights from DeepHeme-SE's models, the next step involves determining the most effective strategy to integrate these perspectives into one conclusive prediction. We proposed three different methods as shown in Figure 6.2A (See Methods for the details). The first method is weighted voting, where we weighted the model prediction by the SoftMax score for each cell type class. The second method is plurality voting, the final prediction is the prediction results with the most counts, regardless the actual probability scores. The third method is intended to the confidence of each model, overall prediction option as the one with the highest probability likelihood. Consequently, weighted voting methods have shown to be superior to other techniques in bone marrow cell typing, demonstrating a significant yet not dramatic edge(Figure 6.1D,E).

When we compared the weighted voting methods performance with the individual snapshot models on all the cell types, we found that different individual has a varied performance and the pooled model achieved superior performance in almost all of them (Figure 6.11). One specific example is the performance for myeloblast, an critical cell type for determining the pathology conditions for most leukemia diseases. The best individual model only achieved F1 score of 0.71 but the snapshot ensemble model increase the model performance into 0.76 (Figure 6.11). Besides showing that snapshot ensemble improved the performance of ResNeXt50 model significantly. We also performed this analysis across several different popular conventional neural network model including Inception V3, EfficientNetV2, GoogLeNet and VGG19(Figure 2). As a more proper control, we trained all deep learning models with standard learning rate schedule for 50 epochs. We found snapshot ensemble outperformed the individual model with standard learning rate schedule even they were trained about the same epoch number and learning rate for the standard control is kept being optimized. When reviewing existing research on bone marrow cell classification, it's important to note that certain studies have provided limited evaluation metrics, typically focusing on AUC. This context, combined with the fact that each study often employs different datasets

and a varied number of cell classes, makes direct comparison a nuanced task. However, DeepHeme-SE encompasses the most extensive array of cell classes compared to other published works and outperforms other models in performance based on mean F1-score, mean precision, mean recall and number of classes with F1-score about 0.8 (Figure 6.2F).

MODEL LEARNS UNDERLYING HEMATOPOIETIC DEVELOPMENTAL RELATIONSHIPS

Interested in whether our classifier has learned relevant and consistent information, we embedded the extracted features represented in the flattened final convolutional layer of the network with 1000 dimensions into 2 dimensions using the Uniform Manifold Approximation and Projection (UMAP) algorithm[25]. This was done to visualize and explore how the different classes are being grouped together or separated from each other (Figure 6.3A). The UMAP has recapitulated much of the hematopoietic structure known to biologists. Notably, myeloblasts and proerythroblasts are linked at the UMAP's top left, suggesting a shared origin from hematopoietic stem cells. This is followed by distinct pathways representing neutrophil and erythroid development. A significant morphological division is observed only between orthochromic erythroblasts and polychromatic erythrocytes, marked by the presence or absence of a nucleus. The transitions between other cell classes in these lineages are more subtle and subject to interpretation. Almost all related hematopoietic cell types are connected, except for lymphocytes and plasma cells, which mature outside the marrow. The algorithm also discerns morphological similarities across lineages, grouping cells with similar nuclear-to-cytoplasmic ratios and showing connectivity only among directly related cell clusters (Figure 6.14) The UMAP's ability to mirror biological relationships indicates that the algorithm is effectively learning pertinent morphological features, rather than relying on confounders or shortcuts which would not reflect these relationships.

Our findings show that classes further apart are more easily distinguishable, while those in closer proximity pose more separation challenges, as evidenced by the one versus one AUC (Figure 6.12). This observation

suggests potential for uncovering new biological relationships and cell classes through the integration of single-cell image datasets with dimension reduction techniques.

MODEL GENERALIZABILITY

Multi-site generalization has proven difficult to achieve in many areas of medical computer vision [26], [27]. In bone marrow classification, to our knowledge, no algorithm has demonstrated reliable multi-site generalization for this problem. To evaluate DeepHeme-SE's ability to generalize to an external dataset, we next tested the classifier on images from a completely independent hospital system, Memorial Sloan Kettering Cancer Center (MSK). Images were scanned using either a Hamamatsu S360 or a Leica Aperio AT2, and then annotated using the same annotation strategy as the original dataset. Figure 6.2A summarizes the performance of DeepHeme-SE on the external dataset. We see a mild decrease in the small of the cell types but overall, the DeepHeme-SE generalizes well on external data set. Notably, the performance disparity between the external data and our in-house test set is significantly mitigated by the snapshot ensemble, improving average F1 score by 0.03 (Figure 6.2B). This suggests that the Snapshot ensemble approach holds great potential for enhancing model generalization on external datasets, making it a valuable technique for advancing the performance and applicability of deep learning models in various domains.

COMPARISON WITH CLINICAL EXPERTS

To assess whether DeepHeme-SE achieves clinical-level accuracy, we compared its performance with that of three subspecialty hematopathologists from well-established cancer centers (MSKCC, UCSF, and Brigham and Women's Hospital). These experts performed the same classification task as the algorithm, blinded to each other's assessments and the gold-standard labels, which were determined by consensus. A random selection of 25 images from each of the 23 classes (575 images) in the UCSF test set was chosen for review. DeepHeme-SE achieved hematopathologist-level or better performance across all 23 classes, with a mean precision and recall of $(0.91 \pm 0.00, 0.91 \pm 0.00)$ compared to $(0.78 \pm 0.05, 0.76 \pm 0.06)$ for

hematopathologists. The mean standard deviation for precision and recall across the classes was (0.03 ± 0.02 , 0.04 ± 0.02) for the AI, versus (0.08 ± 0.06 , $0.10 \pm 0.00.05$) for hematologists, demonstrating a major improvement in reproducibility. In terms of speed, the three hematopathologists took an average of approximately 3 hours to label 575 images, while DeepHeme-SE completed the same task in 0.36 seconds—nearly 30,000 times faster than the experts. It is important to note that we do not claim our model to be superior to clinician experts, who in a real-world setting can use image context, clinical history, multimodal clinical data and expert judgement to improve their performance on cell classification. However, given the image alone, DeepHeme-SE demonstrates comparable or better performance. Given the fact that our dataset encompasses the most abundant cell type classes to date, we believe that we have made significant strides in optimizing the model's performance.

EVALUATING DEEPHEME'S DIAGNOSTIC UTILITY

The ability to discern the distribution of cell types is crucial for the diagnosis of various hematological conditions. Building upon the robust cell-typing capabilities of DeepHeme-SE, we sought to investigate its application in the clinical diagnosis realm. To approximate the manual cell counting procedure employed by pathologists, who select regions of interest and classify cells, we used DeepHeme-SE to automate the prediction of cell types. Our evaluation engaged samples from three de-identified patients, each with classic hematological diagnoses. Patient 1 exhibited a normal marrow profile, typified by a blast count under 5%. Patient 2 was identified with chronic myelomonocytic leukemia, marked by a moderately progressive blast count between 5% and 20%. Patient 3's diagnosis was acute myeloid leukemia, an aggressive leukemia requiring immediate intervention, with a blast count greater than 20%.

Figures 6.4 A-C underscore DeepHeme-SE's ability to distinguish between these diagnosis based on morphometric quantification and in particular, myeloid blast count. By combining this approach with

automated region selection and cell detection, DeepHeme-SE could provide fully automated graphical and text reports to pathologists to improve their workflows.

CLOUD-BASED WEB APPLICATION

To demonstrate the performance of the DeepHeme-SE algorithm and to encourage further collaboration and development, we have built a cloud deployment where users can test it (<https://hemepath.ai/deepheme.html>). The application allows users to test the algorithm on images from either the UCSF or MSK test sets. They can also upload their own bone marrow aspirate images. Images may be cropped from 400x-equivalent WSIs or images captured from microscope cameras at 400x. A full featured web application for analyzing WSIs is also available for those with user credentials (Figure 6.6B). The full version includes a WSI image viewer, region selection tools, cell selection tools and tools for automated cell detection. It also includes tools for rapid visualization of the results at the slide or region level, as well as tools for rapid annotation (Figure 6.6B). This interface can be used to build clinical validation datasets for any hospital interested in deploying the DeepHeme system.

CLINICAL TESTING AND DEPLOYMENT ARCHITECTURE

We have implemented a framework for clinical validation at MSK summarized (Figure 6.6C). WSI scanners are connected to a shared network Isolon server on which all WSI images are stored. A newly scanned image spawns a call to the electronic health record (EHR) to determine if the slide is a bone marrow aspirate or not and collect necessary clinical information. The slides are then tiled and accessioned in a PostgreSQL relational database. The tiled images undergo region selection, cell detection, cell classification, as well as a series of statistical quality assessments. The final differential count is communicated back to the EHR along with a hyperlink to the web application where the user can review the slide, review the results, and view visualizations that provide evidence and explainability for the results (Figure 6.6C). This

system will be used for clinical testing and validation prior to its clinical deployment. Containerizing each element allows for easier deployment across different hospitals and research environments including on-prem, cloud, and hybrid architectures, allowing compliance with institutional infrastructure policies regarding protective health information.

6.4 DISCUSSION

In this study, we presented DeepHeme-SE, a method tailored for the classification of cells in bone marrow aspirate. Our findings underscore the significant impact of using an expert consensus-annotated dataset combined with the robust capabilities of ensemble deep neural networks. This synergistic approach outperforms previous published work, and yields results comparable to those of hematopathologists. DeepHeme-SE also demonstrates strong adaptability across various medical centers, WSI scanners, and a broad range of hematologic diseases. Notably, our results also illustrate the capacity of neural networks to decipher and represent the underlying biological relationships inherent in labeled image datasets. Additionally, we have developed platforms for both scientific collaboration and clinical testing, aiming to facilitate further research and practical applications of our findings. Our research affirms the critical role of ensemble methods in deep learning, particularly for medical imaging applications, as corroborated by various studies[35–38]. Unlike many medical diagnostic tasks such as drug response, cancer grading, and COVID detection, which typically involve binary or a small number of classes, bone marrow cell typing presents a uniquely challenging scenario with its requirement for classification across an extensive array of over 20 distinct classes. This complexity significantly elevates the challenge, exemplifying situations where the advantages of ensemble methods become especially valuable[21], [39], [40]. Among ensemble methods, snapshot ensemble methods stand out due to their distinct advantages. One key benefit of snapshot ensembles is their efficiency in training time. Unlike traditional methods that require extensive training for each model in the ensemble, snapshot ensembles streamline this process, leading to significant time savings[41]. Additionally, these methods alleviate the often cumbersome task of selecting different architectures and fine-tuning hyperparameters for each model in the ensemble. Instead, snapshot ensembles allow for a unified approach where a single architecture can be optimized and then leveraged to generate multiple models at different training phases. This not only simplifies the model

development process but also ensures consistency across the ensemble, making it a particularly effective strategy in medical imaging where precision and reliability are paramount. Last but not least, our attempt for improving saliency map with snapshot ensemble could also be innovative. A novel aspect of this work, compared with other bone marrow classifier, is our efforts in external validation. Through a series of analyses (Figure 6.15, and Figure 6.16), we demonstrate that Deep-SE can maintain remarkably consistent performance across both in-domain and external datasets. However, it is important to recognize that our results do not signify a complete resolution of the bone marrow classification challenge. For instance, the model's predictive accuracy may falter when tasked with identifying specific cell types not included in our current dataset. Nonetheless, our findings offer substantial evidence that a combination of high-quality dataset configuration and adequately trained models can effectively overcome challenges associated with color variation, staining differences, and other domain-specific shifts in image characteristics. Overall, our external analysis, augmented by expert comparison, represents a significant stride towards automating the evaluation of bone marrow (BM) cell morphology using automated image classification algorithms.

One current limitation of DeepHeme-SE is its dependence on manual steps in the diagnostic pipeline. Future developments will focus on integrating DeepHeme-SE with automated region selection and cell detection algorithms. This integration aims to establish a fully autonomous diagnostic pipeline. The efficacy of this pipeline will be rigorously evaluated by comparing its outputs with bone marrow differentials and diagnoses extracted from pathology reports in clinical archives and prospective trials. Additionally, entities for specific disease morphologies, such as hairy cell leukemia cells, could be added. Such advancements are not only expected to streamline diagnostic processes, but also to enhance the accuracy and reliability of medical assessments.

Our study paves the way for several promising avenues in enhancing bone marrow classification. Firstly, including a wider array of cell types for AI-based bone marrow classification. This is particularly significant as our findings demonstrate that high performance can be maintained even with an extensive number of cell classes, given the availability of high-quality data. For instance, incorporating abnormal cells associated with specific subtypes of acute myeloid leukemia could greatly enhance the model's

diagnostic capabilities for more nuanced disease categories. Such an expansion would not only deepen our understanding but also improve the identification of a diverse range of cellular abnormalities. Secondly, integrating DeepHeme-SE, or other bone marrow classifiers into a fully automated diagnostic pipeline, as proposed by existing literature, is a critical advancement. This integration would not only streamline the diagnostic process but also facilitate more complex analyses, such as genotype determination, drug response prediction, and survival rate estimation. Lastly, a collaborative effort between clinicians and AI scientists is essential for the clinical deployment of software utilizing the bone marrow classifier. This partnership will be crucial in ensuring that the technology is effectively translated into a clinical setting, benefiting real-world patient diagnosis and treatment. In summary, the potential applications of DeepHeme-SE are extensive. Its ability to maintain high performance across a large number of cell classes is just the beginning. Future research should aim to validate these diagnostic tools clinically, ensuring the efficacy and reliability in real patient scenarios. This progression from theoretical model to practical application represents an exciting frontier in the field of medical AI (Figure 6.6).

6.5 METHODS

CASE IDENTIFICATION, WHOLE SLIDE IMAGING, AND IMAGE ANNOTATION

50 aspirate slides from 50 unique patients with normal BMA morphology were selected from the UCSF Parnassus adult hospital and UCSF Benioff Children's Hospital between 2017 and 2020. The UCSF slide set was separated at this stage into 40 slides used for training and validation. 10 slides were kept as a hold-out test set to ensure accurate reporting of the model's performance on unseen patient cases. A library of 41,595 images was assigned into one of 23 classes by consensus decision of an expert panel of three hematopathologists (Table 6.1). 30,394 images in the training set and 8,507 images in the test set.

The 23 image classes represent all cell types included in a standard bone marrow differential, as well as differentiation stages of trilineage hematopoietic cells (Figure 6.1b). The full spectrum of erythroid and neutrophil maturation was included, from proerythroblast to mature erythrocyte and from myeloid blast to segmented neutrophil, respectively. Along the megakaryocytic lineage, megakaryocytes and platelet clumps were assessed. The lymphoid lineage included lymphocytes and plasma cells. Eosinophils were

separated into mature eosinophils with segmented nuclei and immature eosinophils. In addition, the set included monocytes, basophils and mast cells. Additional classes include artifacts and mitotic bodies to probe for cellular states. Artifacts are cells that are broken as a result of the biopsy procedure or slide preparation process. These cells cannot be used for classification. The presence of mitotic bodies is a proxy for the mitotic rate of the sample, which is itself a clinically prognostic biomarker[42–45]. Because of the differences in the relative distribution of bone marrow cell types, special efforts were made to identify additional examples of rare classes including myeloid blasts, basophils, mast cells, and mitotic bodies[46]. Slides from MSK were scanned using a Hamamatsu S360 scanner and were taken from the archives of the clinical service. They include morphologically normal samples and a range of abnormal samples.

CASE IDENTIFICATION AND WHOLE SLIDE IMAGING

Two new datasets were created to develop and test the performance of our deep learning algorithm, one from UCSF and one from MSK. All UCSF slides were randomly selected from the adult and pediatric hematopathology clinical service based on normal morphology and adequate specimen. WSIs were scanned at 400x-equivalent magnification using either a Leica Aperio AT Turbo, Leica Aperio AT2, or GT450 and saved as .svs files. MSK slides come from the clinical hematopathology service and represent the range of normal and abnormal cases seen there. Slides were scanned using a Hamamatsu S360 and saved as .ndpi files. All slides were scanned using a high density of focus points and a single z-plane. Slides include a range of quality reflecting variations in stain intensity, slide preparation, and slide age common to clinical archives.

IMAGE LIBRARY ANNOTATION

Images were annotated using annotation software developed in-house. To compensate for variations in slide preparation and stain intensity, as well as to replicate features of a manual microscope, the software's viewer permits modification of brightness, contrast, and zoom. To compensate for variations in slide preparation and stain intensity, as well as to replicate features of a manual microscope, the software's viewer permits modification of brightness, contrast, and zoom. Images from both the UCSF and MSK datasets were annotated using a 3-step process. Initial image classification was performed by a single

pathologist. A second audit was performed by a single pathologist to correct any errors made with the first round of classification. Finally, a panel of three hematopathologists reviewed the final sorted cell lists to provide a consensus label that was used as the gold-standard for each image. For each image, the annotated cell is in the center of the image, based on the whole cell, not the nucleus, except for the following classes. Since megakaryocytes are larger than the field of view, image centers were placed in multiple non-overlapping locations within the megakaryocyte to capture different fields of view. For cells undergoing mitosis, the centers may have been placed in either the center of the mitotic figure, the center of the cell, or both. For platelet clumps, the center of the object was placed in the middle of the clump. Images were exported as 96x96 pixel PNGs with a resolution of 72px/inch.

SNAPSHOT ENSEMBLE

Snapshot Ensemble is a training technique for neural networks that enables the capture of several different learned models at different epochs during training, which can then be combined into an ensemble with minimal extra computational cost. It leverages the cyclical learning rate policy to identify the optimal stopping points (snapshots) of the model during training, allowing us to store these snapshots and later aggregate them to improve performance and robustness. In our implementation of Snapshot Ensemble, we adopt the following steps: 1. For different neural network we chose for the bone marrow cell typing task, we initialize a deep neural network with pretrained weights from ImageNet. Our network architecture is designed to be conducive to the learning task, taking into account the complexity of the input data and the desired output. 2. We employ a cyclic learning rate policy where the learning rate cyclically varies between reasonable boundary values. This allows the model to converge to several local minima along its training path. The length of a cycle is predetermined and is set such that the learning rate will have made a complete cycle back to its initial value by the time we take a snapshot. At the end of each learning rate cycle, when the rate is at its lowest and the model is presumed to be at a local minimum, we take a snapshot of the model weights. This does not require the training to stop, and we continue the training process by increasing the learning rate once again as per the cyclic learning rate policy. After training is completed over several cycles, we combine the snapshots by averaging their predictions. This can be done in different ways which we will elaborate later. For our experiment, we trained the neural network model over a total of 50 epochs. These epochs were divided into 5 snapshot

periods, each consisting of 10 epochs, during which the learning rate decreases progressively. This decrement is in accordance with a predefined schedule, adhering to the cyclical learning rate policy discussed earlier. At the conclusion of each 10-epoch snapshot period, rather than allowing the learning rate to reach its minimum value before taking a snapshot, we introduce an early stopping criterion based on the model's performance on a held-out validation set. This early stopping mechanism is crucial in identifying the most performant model within each period. It's important to note that the 'best' model — the one we take a snapshot of — does not necessarily coincide with the last epoch of the snapshot period before the learning rate is reset. It is the model that exhibits the highest validation performance during the snapshot period.

METHODOLOGY FOR ENSEMBLE PREDICTION SYNTHESIS

In the realm of predictive modeling with deep learning, synthesizing the output from multiple models—each potentially offering a unique viewpoint—can significantly enhance the robustness and accuracy of the final decision. Given that our models maintain architectural consistency and provide predictions along with corresponding likelihoods on a comparable scale, we have formulated three distinct strategies to integrate their individual inferences.

Pooling Method: Our first approach, termed 'Pooling', involves aggregating the probability distributions provided by each model. By averaging the predicted probabilities for each class across all snapshots, we achieve a consensus prediction that embodies the collective wisdom of the ensemble.

Voting Method: The second approach relies on a democratic 'Voting' system. Each model casts a vote for its predicted class, and the class with the majority of votes is selected as the final prediction. This method capitalizes on the strength of the most frequent outcome, effectively harnessing the power of numbers.

Confidence-Based Selection: The third method prioritizes the confidence level of individual models. Here, the final prediction is chosen based on the highest probability likelihood among all the predictions. This

technique assumes that the model's confidence in its prediction—a reflection of the maximum likelihood—is a direct indicator of its accuracy.

The implementation of these techniques aims to enhance decision-making by leveraging the unique strengths of each model in the ensemble. By considering both the collective agreement and individual confidence levels, our methods strive to distill a more precise and reliable final prediction from the snapshot ensemble.

NEURAL NETWORK STRUCTURE, TRAINING AND TESTING

In this study, a diverse array of deep learning architectures was utilized, including ResNeXt-50, EfficientNetV2, VGG19, GoogLeNet, and Inception V3, to classify bone marrow cells. Each architecture was selected based on its reported efficacy in image recognition tasks. We initiated training with the ResNeXt-50 architecture[47], following its documented success in similar classification tasks by Matek et al[1]. Furthermore, we adapted other models to fit the specific requirements of our image data and classification objectives.

Our dataset presented an imbalanced distribution of cell types. To address this, we applied up-sampling to balance the classes before proceeding with data augmentation. We leveraged the Albumentations Python library to implement 20 different augmentation transformations[48], enhancing both the shape and color characteristics of the images. This extensive augmentation process resulted in a robust dataset of approximately 50,000 images per class, totaling 1.15 million images. We included shape augmentations such as rotations, flips, shears, and resizing. For color, we incorporated adjustments in contrast, brightness, and added Gaussian noise, alongside stain-color augmentation to enhance the dataset's variability and representativeness. For our experiments, we conducted 50 iterations overall. Each model was initially loaded with ImageNet-pretrained weights and then fine-tuned on our augmented bone marrow cell images. We adapted the input size to accept 96x96 pixel images and modified the output layer to classify the 23 cell types defined in our annotation scheme. To optimize our models, we employed the Adam optimizer with an initial learning rate of 0.001 and a large batch size of 1024. The loss function used was binary cross-entropy, suitable for our one-hot encoded targets. All training was performed on

NVIDIA TITAN RTX graphics processing units, where training of the ResNeXt model took approximately 12 hours of computing time. For training and validation, we used 40 images from slides from the UCSF dataset, whereas the rest 10 slides were used as the UCSF test set. 5-fold cross validation was performed on the training/validation set. All model tuning and parameters adjustment were performed during training and validation. All the numbers reported in the paper come from analysis of the unseen test sets. Results were then averaged across the 5 different cross-validation networks.

UMAP INTERPRETATION

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) was used to represent the information that the deep learning classifier learned²⁵. We embedded the extracted features represented in the flattened final convolutional layer of the network into 2 dimensions for each member of the data set using the UMAP algorithm. UMAP works by using nearest-neighbor-descent technique to identify the closest neighbors. The nearest neighbors that were previously found are then connected to create a graph[53]. The next stage for UMAP is to map the approximation manifold to a lower-dimensional space, in our case two dimensions, after learning it from the higher-dimensional environment. To perform these calculations, we used the umap-learn package in Python[54].

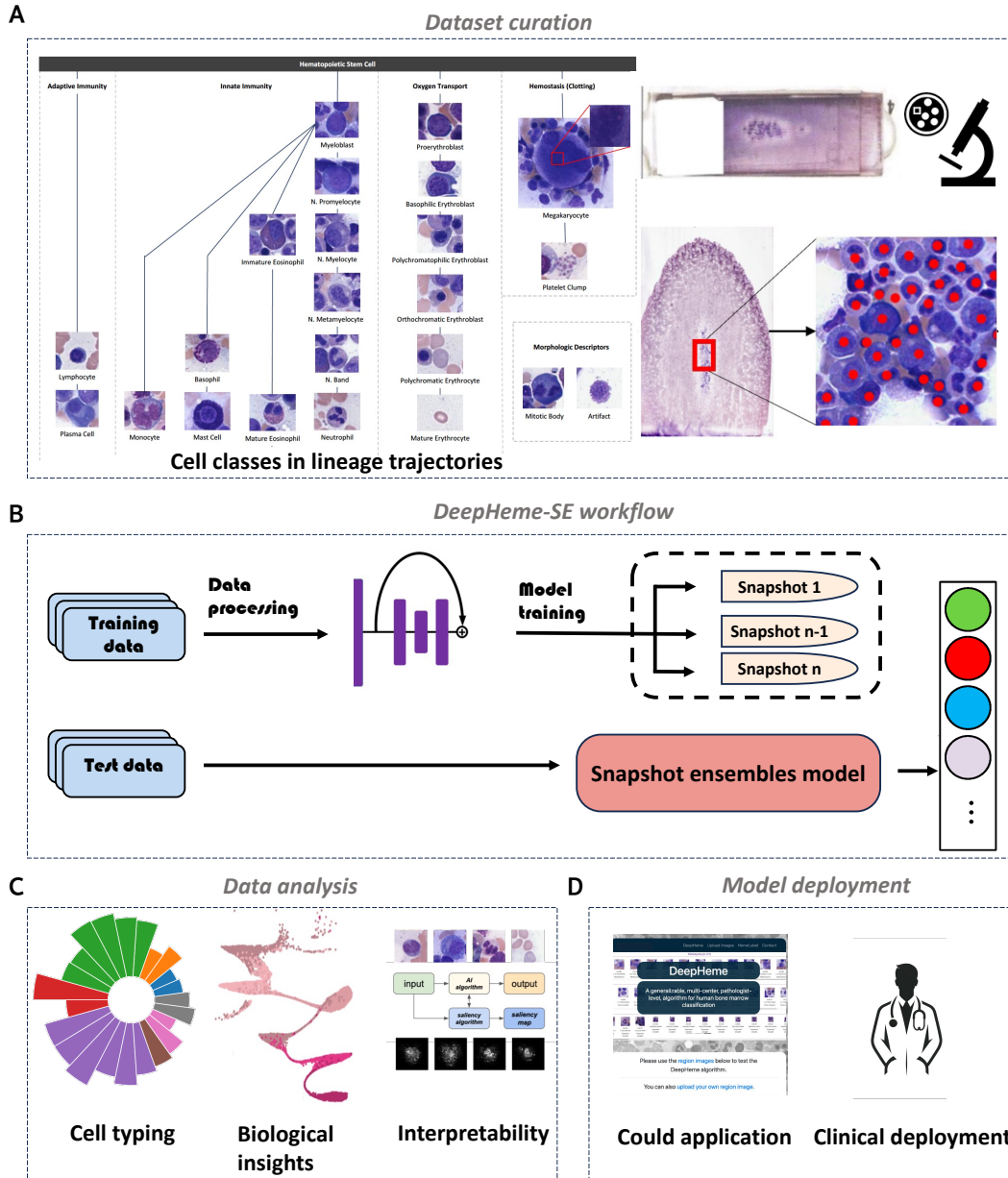


Figure 6.1 Workflow of DeepHeme-SE

(A) Whole slide images of bone marrow aspirates were digitized using whole slide scanners. Regions of interest were selected by hematopathologists and the location and classification of cells was labeled by a consensus of three hematopathologists. A diagram of cells and morphologic labels included in the study, as well as their relationship to each other in the hematopoietic tree. (B) The overall framework of group convolution snapshot ensemble method. (C) The illustration of the downstream analysis for DeepHeme. From left to right: Cell typing, Feature embedding and clustering, interpretability analysis. (D) A web application has been built for scientists to interact with the DeepHeme algorithm, as well as a clinical deployment framework that interfaces with the digital slide scanning laboratory and the electronic health record.

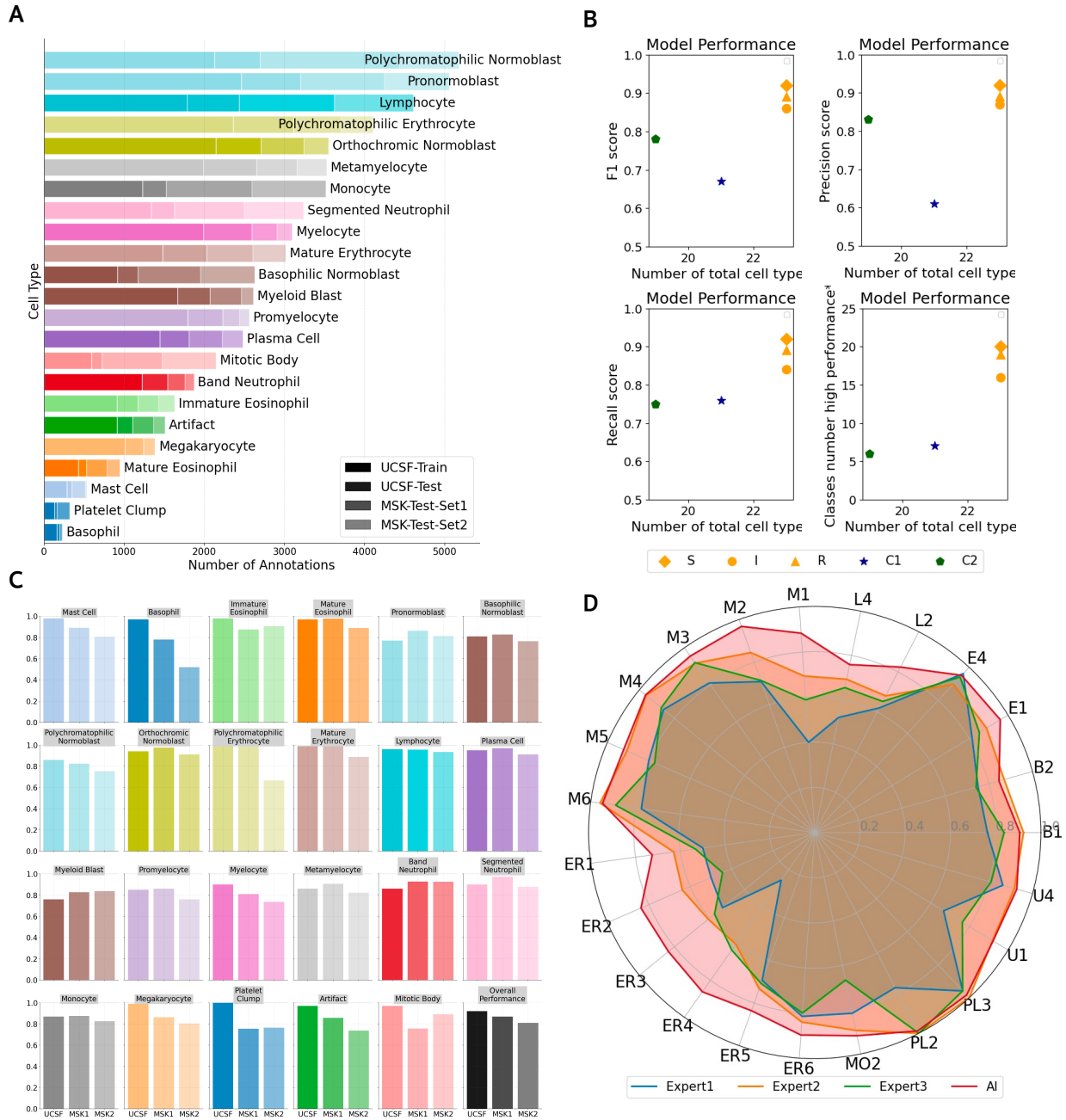


Figure 6.2 External validation and expert evaluation

(A) Data distribution for individual cell types. (B) Comparing with the previous work. C1 and C1 are the bone marrow classifier whose mean model performance we can found^{9, 11}. I refer to the mean performance of individual snapshot ensemble classifier, R is the bone marrow classifier regularly trained for the same epoch as the whole snapshot ensemble framework, S refers to snapshot ensembles. (C) Boxplot demonstrated the model performance on UCSF-Test-set, MSK-Test-set-1 and MSK-Test-set-2. (D) Radar plot comparing the Deep-SE performance compared with three clinical experts.



Figure 6.3 Exploring Advanced Snapshot Ensemble in Bone Marrow Cell Classification

(A) Visual representation of three ensemble summarization strategies: Pooling, Voting, and Confidence. (B) Bar chart comparing F1 scores for 23 cell types using different methods—blue for Pooling, red for Voting, and green for Confidence. (C) Boxplot illustrating performance variations among the three summarization methods. (D) F1 Score Comparison of Five Snapshot Modules and Snapshot Ensemble. (E) Line graph depicting the performance of pooling method (continuous line) against individual snapshot modules (dashed lines). (F) Dumbbell plots contrasting model performance across snapshot ensemble, standard learning rate scheduler, and individual snapshot models, highlighting the efficacy and robustness of the ensemble approach.

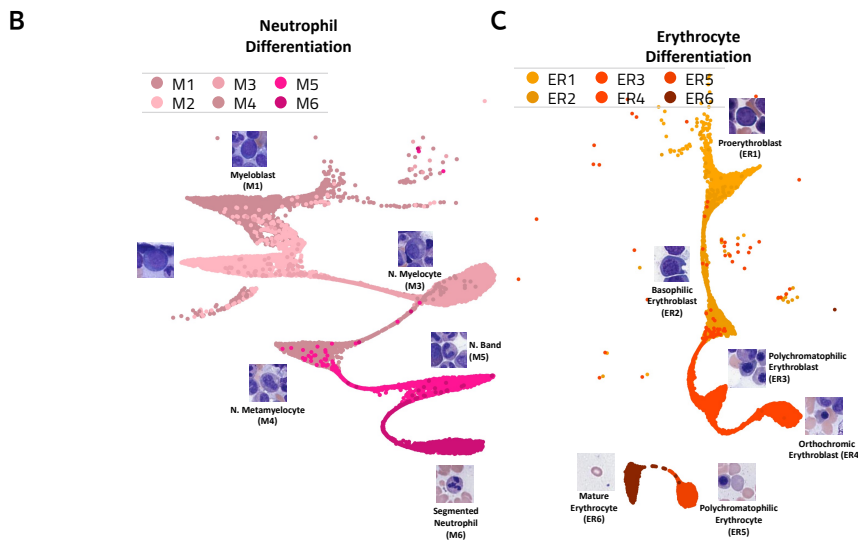
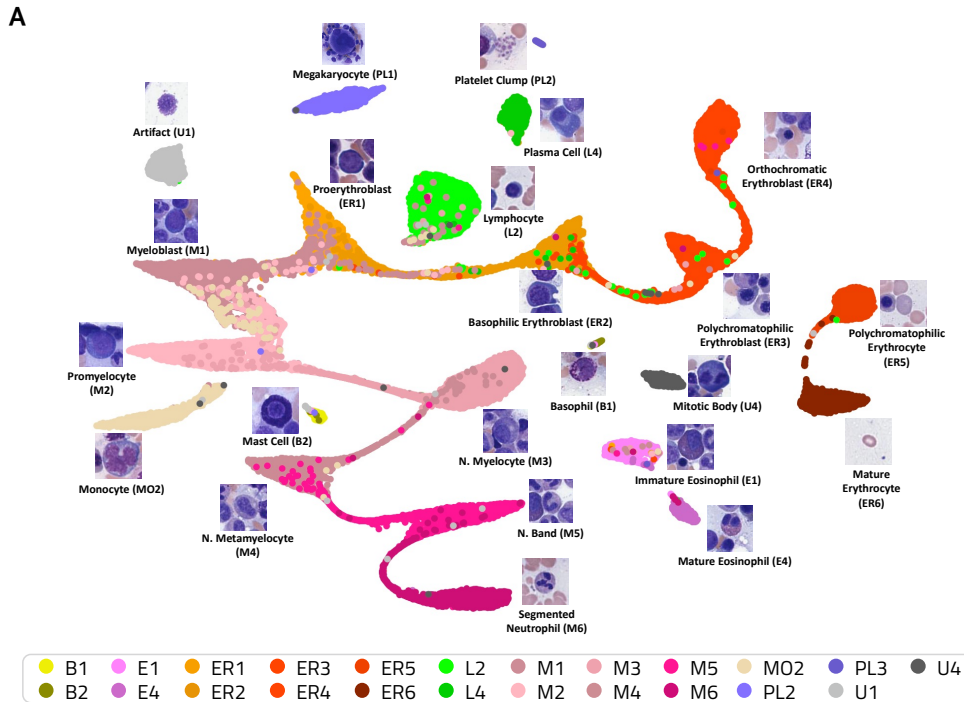


Figure 6.4 Model Learns Underlying Hematopoietic Developmental Relationships

(A) UMAP embedding of extracted features recapitulates biological relationships. The shape of the UMAP recapitulates major aspects of hematopoiesis, of which the untrained neural network has no prior knowledge, suggesting it has been learned from the training images. Bridges between clusters reflect the natural continuum and lineage trajectories between adjacent cell types. (B) Neutrophil differentiation. The complete spectrum of neutrophil development from myeloblast to segmented neutrophil has been learned by the algorithm. (C) Erythrocyte differentiation. Similarly, the full spectrum of erythroid development has also been learned by the algorithm. The break between orthochromatic erythroblasts (ER5) and polychromatic erythrocytes (ER6) likely reflects their clear morphologic boundary (the presence of a nucleus). Such clear morphologic boundaries do not exist between other cell categories, which are defined based on multiple, subjective features.

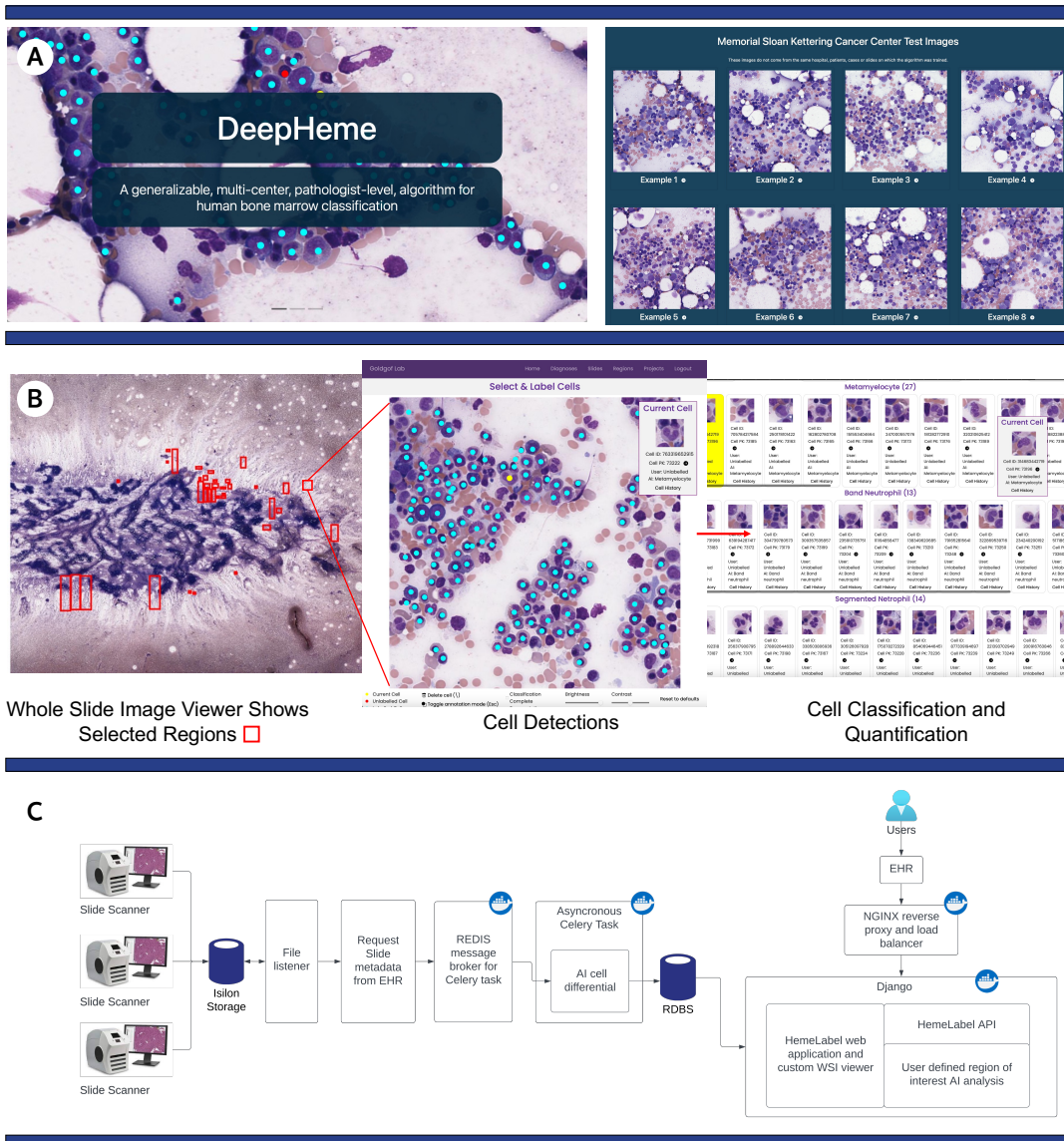


Figure 6.6 Software for Supporting DeepHeme-SE

A) A cloud-based web application has been created to encourage collaboration and further development including a web application where the user has the option to upload their own image, or select one of several sample images to test the performance of the algorithm. B) The software includes a whole slide image viewer, region view with cell detection algorithm, and views for rapidly reviewing cell classification and quantification. This can be used to annotate datasets when building validation sets for new hospitals, or for providing explanations and verifiable results to clinicians using DeepHeme SE for rapid review. C) This describes the clinical testing and deployment architecture at MSKCC for the DeepHeme system. Multiple WSI scanners place images in a shared network drive. The creation of these images spawn a process that automates communication to the EHR to confirm the specimen type, and then

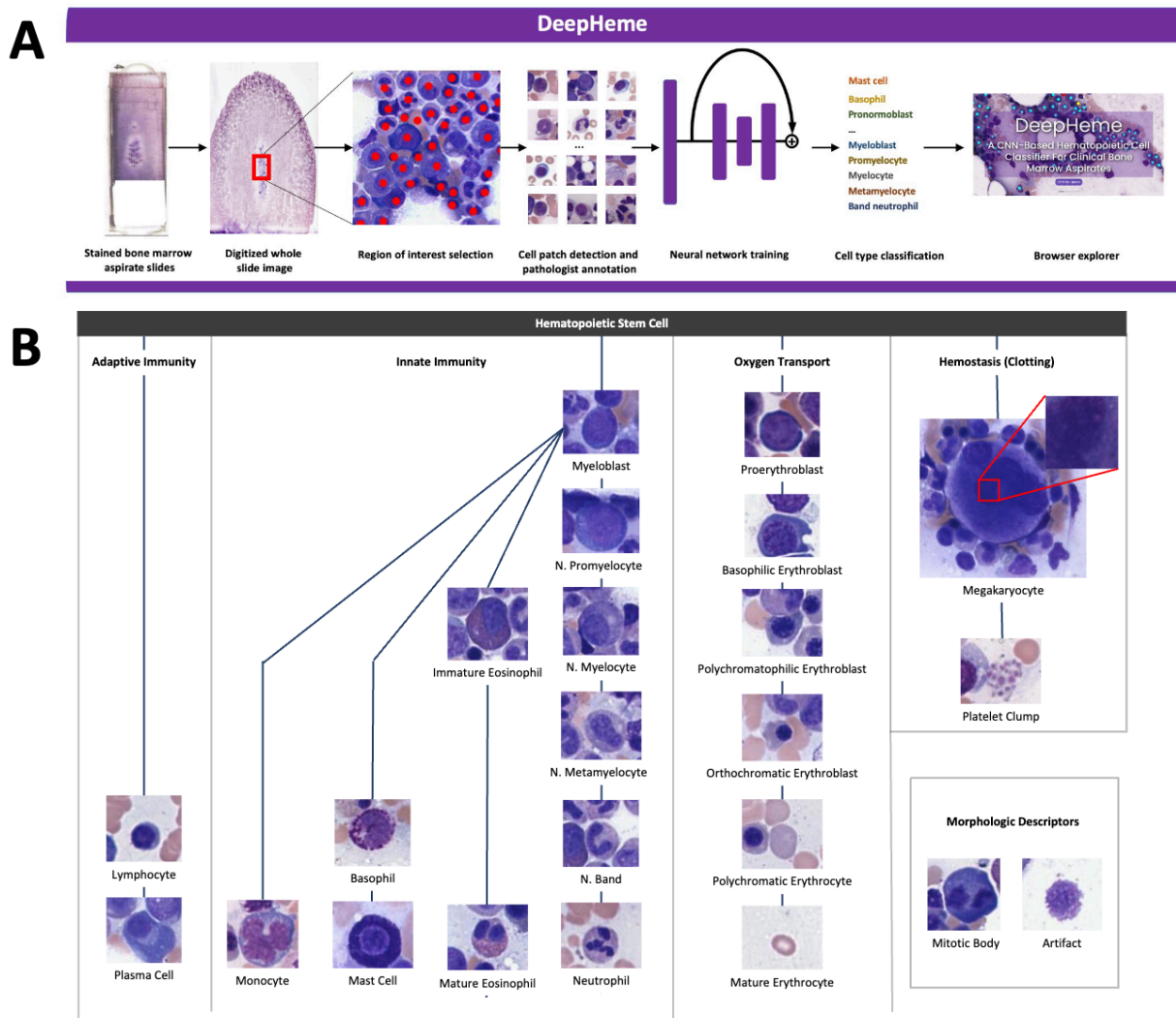


Figure 6.7 Workflow of DeepHeme

A) Experimental workflow. Whole slide images of bone marrow aspirates were digitized using whole slide scanners. Regions of interest were selected by hematopathologists and the location and classification of cells was labeled by a consensus of three hematopathologists. The single cell images were used to train and test convolutional neural networks with ResNext-50 architecture to produce DeepHeme, an algorithm that classifies single cell images into 23 different cell classes. A web application was built, where scientists can interact with the DeepHeme algorithm (<https://www.hemepath.ai/deepheme.html>). B) Cell classes, lineage trajectories, and physiologic functions. A diagram of cells and morphologic labels included in the study, as well as their relationship to each other in the hematopoietic tree. In addition to classes of cells, two important morphologic categories were included: mitotic body and artifact.

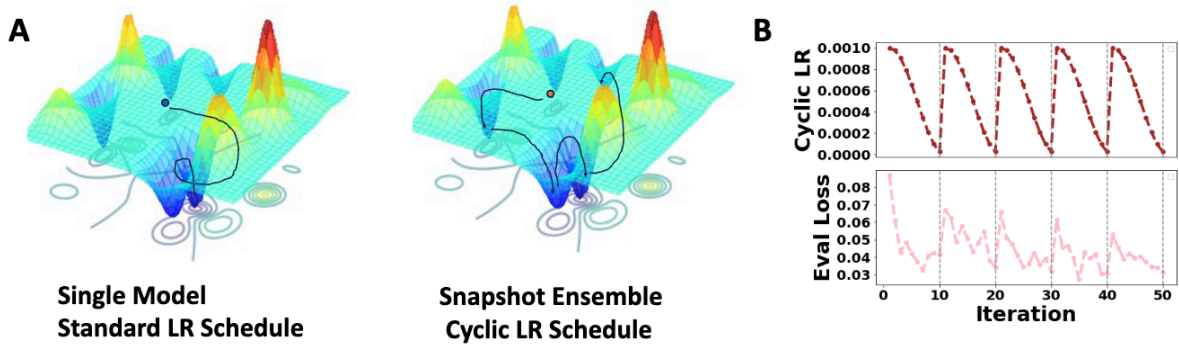


Figure 6.8 Snapshot ensemble concept

A). Comparison of traditional model and snapshot ensemble optimization process. B). The changing process of cyclic cosine learning rate and corresponding loss in the evaluation set.

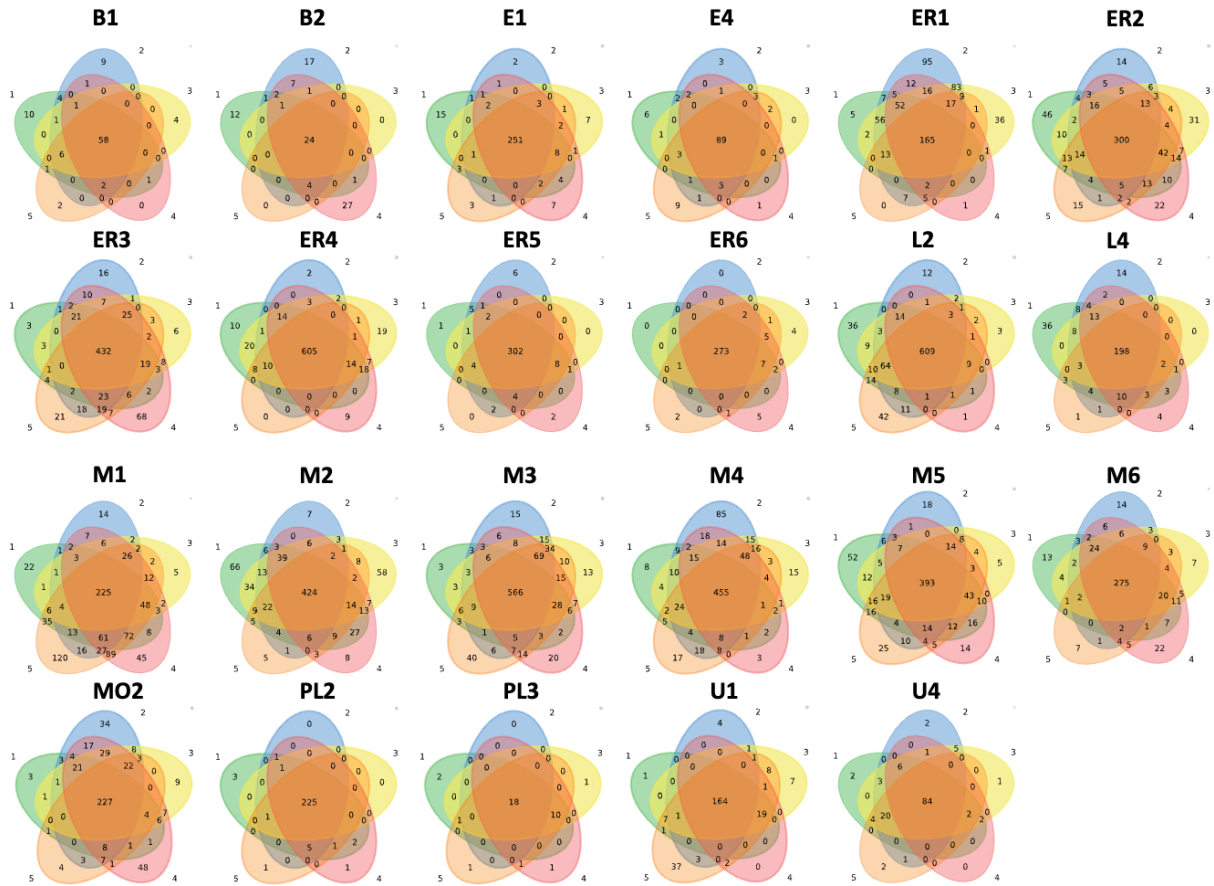


Figure 6.9 Venn Diagram

Venn Diagram of Prediction Results for 23 Cell Types Across Five Snapshot Models This diagram illustrates the intersection and unique aspects of prediction results for different cell types across five distinct snapshot individual modules.

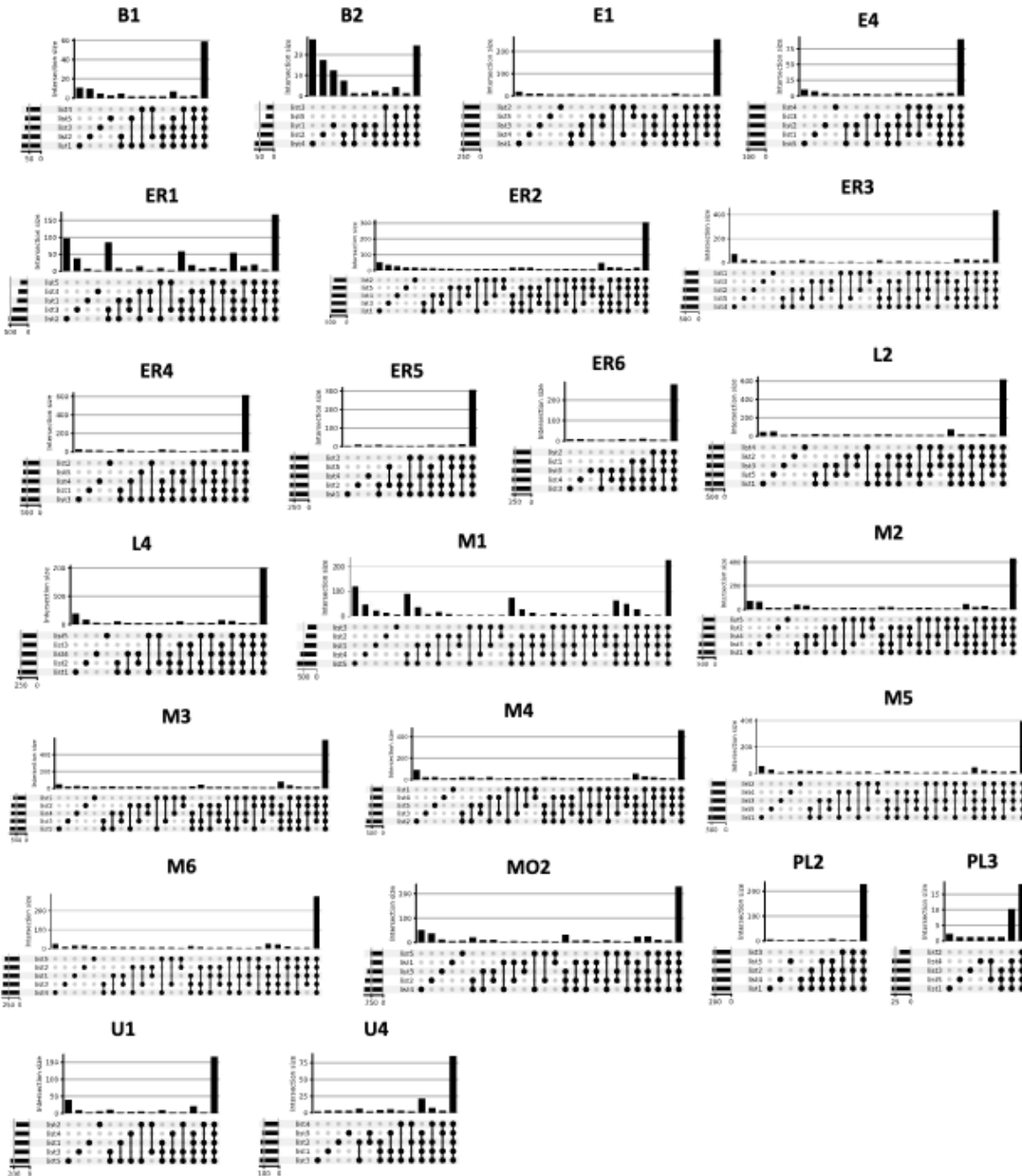


Figure 6.10 UpSet Plot

UpSet Plot of Prediction Results for 23 Cell Types Across Five Snapshot Modules This UpSet plot provides an alternative visualization to the previously presented Venn diagram, illustrating the prediction results cell types across the five Snapshot Modules.

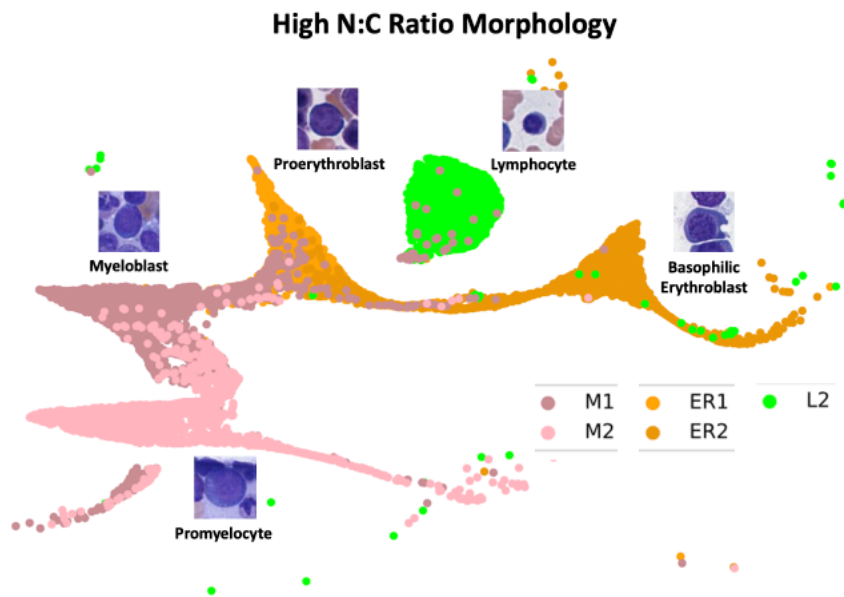


Figure 6.11 UMAP embedding for cells with high nucleus:cytoplasm (N:C) ratio

All five cell classes in our dataset with high N:C ratio co-localize, while only cell clusters that are directly related to each other are attached by bridges. Of note, the bridge between myeloblasts (M1) and proerythroblasts (ER1) reflects the location of the theoretical hematopoietic stem cell that exists as a precursor between them.

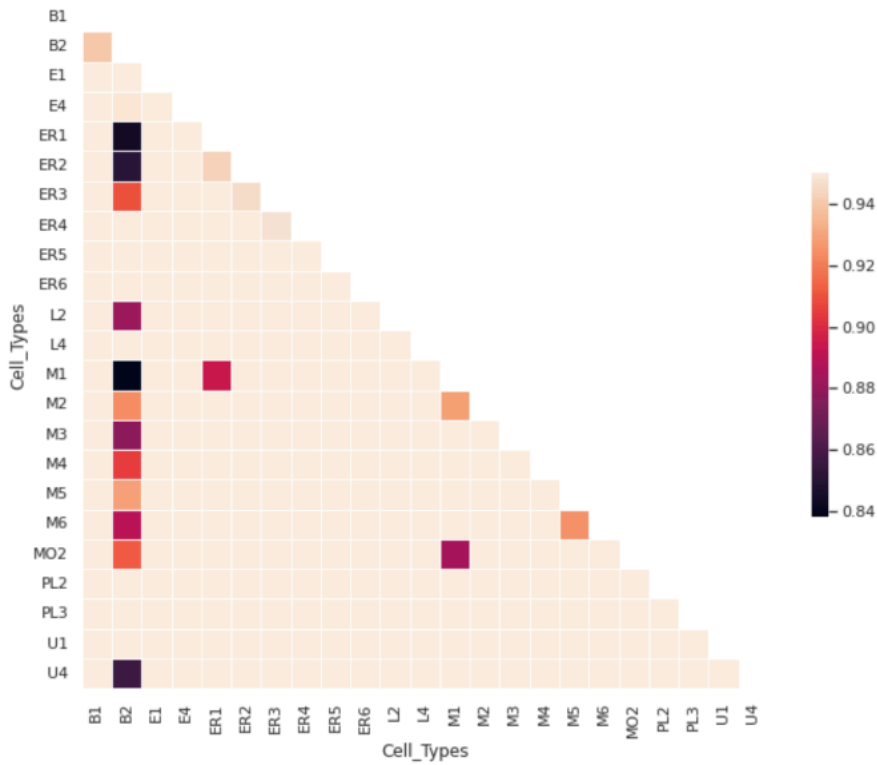


Figure 6.12 One-vs-One AUC Analysis

Here, the analysis focuses on the Area Under the Curve (AUC) for one-versus-one comparisons among the classes. This measure provides insights into the separability of each class pair, with higher AUC values indicating greater ease of distinguishing between the two classes.

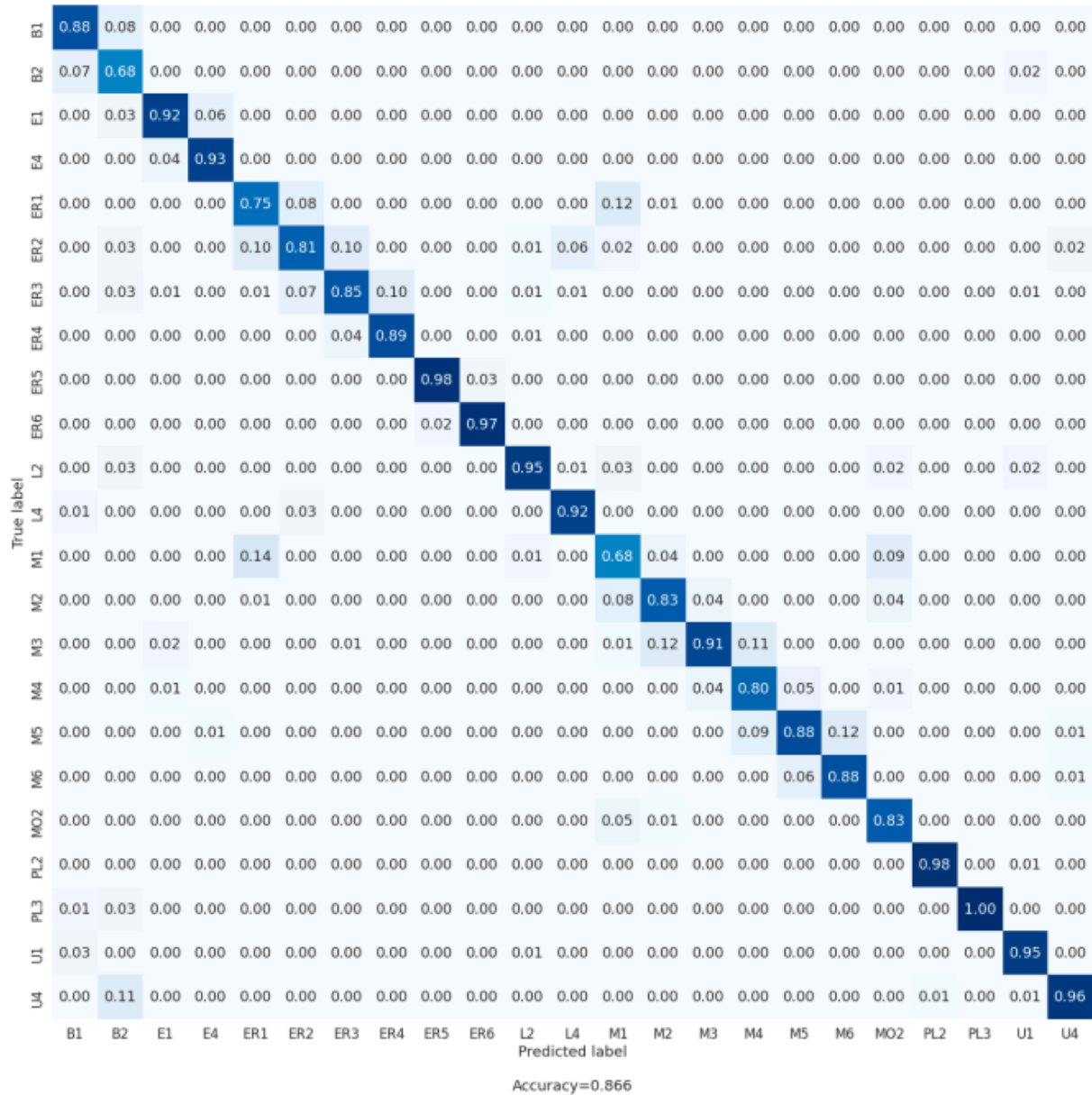


Figure 6.13 Confusion Matrix on UCSF dataset

This figure shows the confusion matrix of prediction on the test set of UCSF images. Most misclassifications are between biologically adjacent cell classes, reflecting the true ambiguity between edge cases. Notable examples include myeloid blast (M1) vs erythroid blast (ER1). These are developmentally adjacent cell types and as a result have some morphologic overlap.

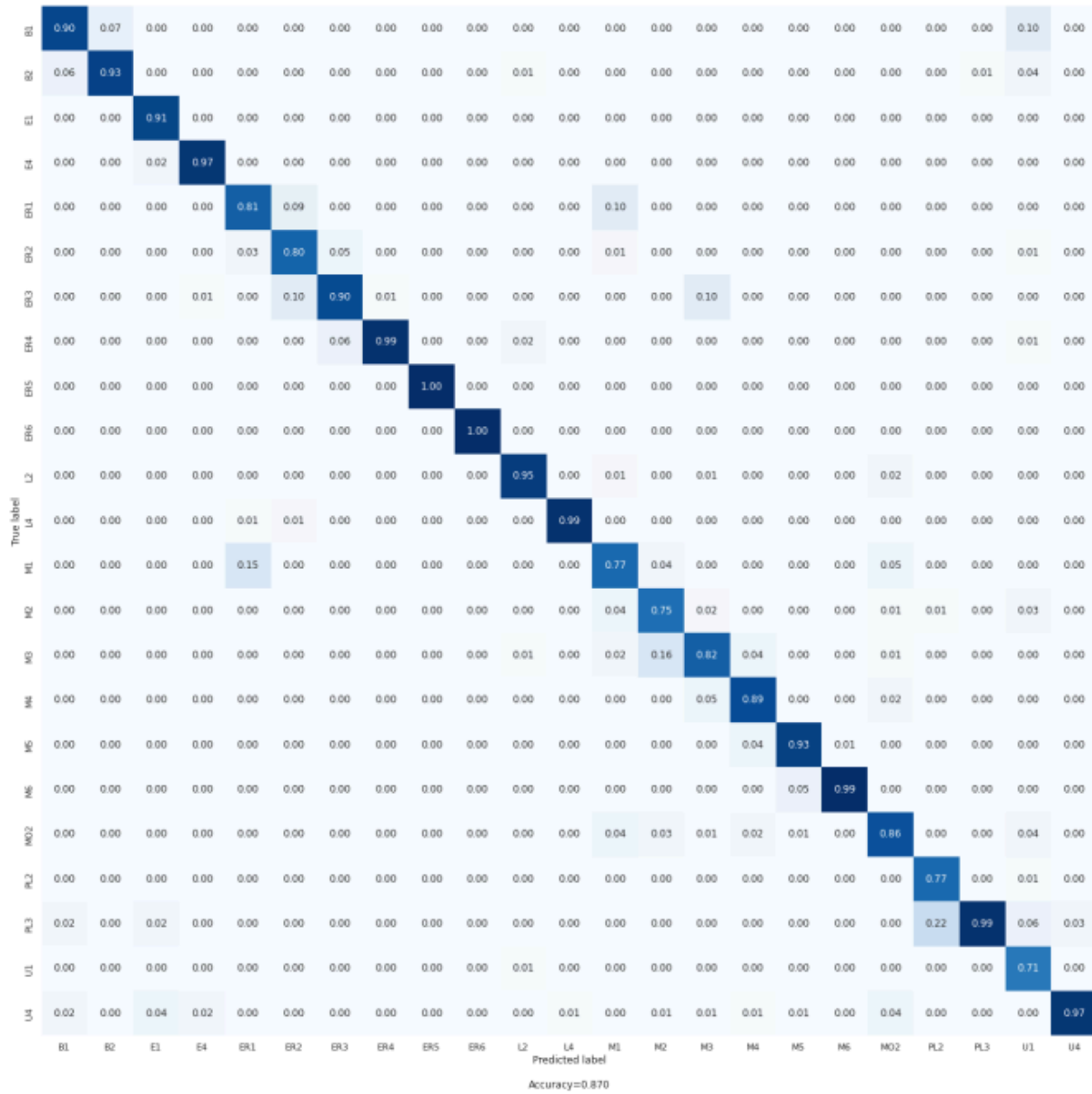


Figure 6.14 Confusion Matrix on MSF dataset

This figure shows the confusion matrix of prediction on the test set of MSK images.

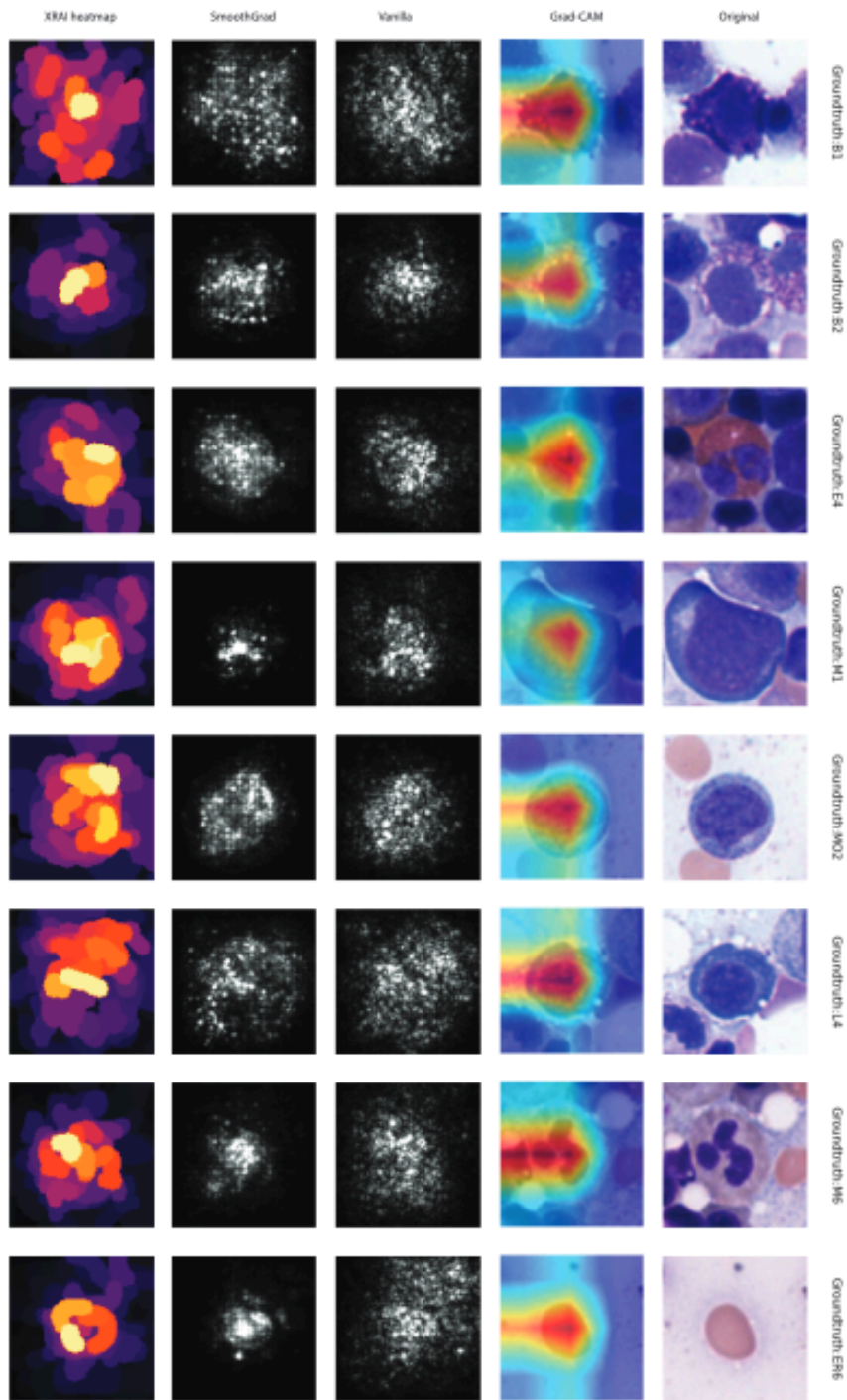


Figure 6.15 Saliency Maps

This figure shows randomly selected saliency maps from 8 classes using different mapping algorithms. image processing through the AI algorithm and image registration in a relational database (RDBS). Processed that are containerized are denoted by the Docker icon. Containerizing each element allows for easy deployment across different hospitals and research environments including en-prem, cloud, and hybrid architectures, allowing compliance with varying institutional policies regarding protective health information.

Table 6.1 Multi-institutional Datasets.

This table shows the Number of hematopathologist-labeled, single cell images, per cell category, evaluated in the training, test, and external validation sets. Training and test sets were separated at the slide level to avoid testing on images from slides on which training had been performed. Each slide was obtained from a unique patient undergoing bone marrow evaluation, with results showing normal hematopoiesis.

| Number of Patients/Slides | Code | UCSF Train | UCSF Test | MSK Test1 | MSK Test2 | Total |
|--------------------------------|------|------------|-----------|-----------|-----------|--------|
| Cell Class | | | | | | |
| Mast Cell | B1 | 284 | 68 | 58 | 18 | 428 |
| Basophil | B2 | 164 | 33 | 22 | 4 | 223 |
| Immature Eosinophil | E1 | 910 | 266 | 43 | 202 | 1421 |
| Mature Eosinophil | E4 | 431 | 98 | 89 | 163 | 781 |
| Pronormoblast | ER1 | 1667 | 404 | 229 | 149 | 2449 |
| Basophilic Normoblast | ER2 | 1795 | 437 | 80 | 119 | 2431 |
| Polychromatophilic Normoblast | ER3 | 1990 | 605 | 100 | 191 | 2886 |
| Orthochromic Normoblast | ER4 | 1991 | 658 | 105 | 368 | 3122 |
| Polychromatophilic Erythrocyte | ER5 | 1223 | 322 | 99 | 113 | 1757 |
| Mature Erythrocyte | ER6 | 1338 | 290 | 117 | 738 | 2483 |
| Lymphocyte | L2 | 2464 | 739 | 93 | 810 | 4106 |
| Plasma Cell | L4 | 916 | 255 | 96 | 675 | 1942 |
| Myeloid Blast | M1 | 2131 | 567 | 229 | 737 | 3664 |
| Promyelocyte | M2 | 2146 | 560 | 210 | 305 | 3221 |
| Myelocyte | M3 | 2359 | 758 | 158 | 399 | 3674 |
| Metamyelocyte | M4 | 1481 | 549 | 109 | 410 | 2549 |
| Band Neutrophil | M5 | 1785 | 652 | 128 | 985 | 3550 |
| Segmented Neutrophil | M6 | 1446 | 363 | 145 | 260 | 2214 |
| Monocyte | MO2 | 1231 | 296 | 131 | 920 | 2578 |
| Megakaryocyte | PL2 | 1007 | 232 | 145 | 0 | 1384 |
| Platelet Clump | PL3 | 131 | 30 | 160 | 0 | 321 |
| Artifact | UI | 910 | 195 | 75 | 141 | 1321 |
| Mitotic Body | U4 | 594 | 130 | 73 | 675 | 1472 |
| Number of Annotated Images | | 30,394 | 8,507 | 2,694 | 8,382 | 49,977 |

Table 6.2 Precision score from three experts

| Model | Expert Scores (Precision) | | |
|-------|---------------------------|---------|---------|
| | Expert1 | Expert2 | Expert3 |
| B1 | 0.70 | 0.89 | 1.00 |
| B2 | 0.78 | 0.95 | 0.69 |
| E1 | 0.71 | 0.83 | 0.79 |
| E4 | 0.93 | 0.92 | 0.92 |
| ER1 | 0.55 | 0.73 | 0.67 |
| ER2 | 0.52 | 0.67 | 0.60 |
| ER3 | 0.53 | 0.71 | 0.58 |
| ER4 | 0.74 | 0.95 | 0.65 |
| ER5 | 1.00 | 0.96 | 0.92 |
| ER6 | 0.76 | 0.93 | 0.91 |
| L2 | 0.90 | 0.86 | 0.69 |
| L4 | 0.65 | 1.00 | 1.00 |
| M1 | 0.67 | 0.59 | 0.60 |
| M2 | 0.52 | 0.68 | 0.50 |
| M3 | 0.47 | 0.55 | 0.47 |
| M4 | 0.67 | 0.72 | 0.74 |
| M5 | 0.71 | 0.63 | 0.74 |
| M6 | 0.76 | 0.84 | 0.73 |
| MO2 | 0.83 | 0.95 | 0.82 |
| PL2 | 0.89 | 1.00 | 1.00 |
| PL3 | 1.00 | 1.00 | 1.00 |
| U1 | 0.52 | 0.92 | 0.70 |
| U4 | 1.00 | 0.96 | 1.00 |

Table 6.3 Recall score from three experts.

| Model | Expert Scores | | |
|-------|---------------|---------|---------|
| | Expert1 | Expert2 | Expert3 |
| B1 | 0.84 | 0.96 | 0.72 |
| B2 | 0.72 | 0.8 | 0.8 |
| E1 | 0.96 | 0.96 | 0.92 |
| E4 | 1.0 | 0.88 | 0.96 |
| ER1 | 0.72 | 0.64 | 0.64 |
| ER2 | 0.52 | 0.72 | 0.72 |
| ER3 | 0.32 | 0.68 | 0.6 |
| ER4 | 0.68 | 0.76 | 0.8 |
| ER5 | 0.68 | 0.88 | 0.92 |
| ER6 | 1.0 | 1.0 | 0.84 |
| L2 | 0.72 | 0.96 | 0.88 |
| L4 | 0.96 | 0.92 | 0.8 |
| M1 | 0.4 | 0.68 | 0.48 |
| M2 | 0.48 | 0.6 | 0.4 |
| M3 | 0.6 | 0.68 | 0.72 |
| M4 | 0.16 | 0.52 | 0.56 |
| M5 | 0.68 | 0.88 | 0.68 |
| M6 | 0.88 | 0.84 | 0.88 |
| MO2 | 0.8 | 0.84 | 0.56 |
| PL2 | 0.68 | 1.0 | 1.0 |
| PL3 | 0.92 | 1.0 | 0.92 |
| U1 | 0.92 | 0.92 | 0.84 |
| U4 | 0.76 | 0.88 | 0.68 |

Table 6.4 Comparison to other deep-learning-based bone marrow cell classifiers

This table compares DeepHeme to other works that use CNNs to classify single cell images from 400x-equivalent images. All image totals are for the entire study, including training and test sets. We define high performance is defined as >0.8 precision and recall. Images in smallest class is a measure of dataset quality. DeepHeme matches or outperforms currently published algorithms across multiple metrics.

| | Chandradevan et al. | Lewis et al. | Matek et al. | Tayebi et al. | DeepHeme |
|--|-----------------------------|---------------------|-----------------------------|----------------------|-----------------------------|
| Annotation Standard | Hematopathologist consensus | Not reported | Clinical Laboratory Staff | Hematopathologist | Hematopathologist consensus |
| Image Source | WSI (light) | WSI (light) | Microscope Camera under Oil | WSI (light) | WSI (light) |
| Images | 9,269 | 23,609 | 171,374 | 26,782 | 41,595 |
| Image Classes | 12 | 16 | 21 | 19 | 23 |
| Images in smallest class | 62 | 155 | 8 | 7 | 219 |
| Classes with high performance | Not reported | Not reported | 7 | 6 | 19 |
| Mean F1-Score | Not reported | Not reported | 0.67 | 0.78 | 0.89 |
| Mean Precision | Not reported | Not reported | 0.61 | 0.83 | 0.89 |
| Mean Recall | Not reported | Not reported | 0.76 | 0.75 | 0.89 |
| Slide Level External Validation | No | Yes | No | Yes | Yes |
| Expert Comparison | No | No | No | No | Yes |
| Institution Level External Validation | No | No | No | No | Yes |

Table 6.5 Comparison between Single Snapshot Model, Standard Learning Rate Scheduler, and Snapshot Ensemble Model Performances.

| Model Names | Single snapshot model | Standard LR scheduler | Snapshot ensemble |
|----------------|-----------------------|-----------------------|-------------------|
| ResNext50 | 0.861 | 0.885 | 0.915 |
| Inception V3 | 0.841 | 0.881 | 0.898 |
| EfficientNetV2 | 0.849 | 0.875 | 0.891 |
| GoogLeNet | 0.845 | 0.861 | 0.877 |
| Vgg19 | 0.825 | 0.845 | 0.863 |

6.6 REFERENCES

- [1] Foucar, K. et al. Concordance among hematopathologists in classifying blasts plus promonocytes: A bone marrow pathology group study. 42, 418–422, DOI: <https://doi.org/10.1111/ijlh.13212> eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijlh.13212>.
- [2] Global Burden of Disease Cancer Collaboration et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: A systematic analysis for the global burden of disease study. 4, 1553–1568, DOI: 10.1001/jamaoncol.2018.2706.
- [3] Alaggio, R. et al. The 5th edition of the world health organization classification of haematolymphoid tumours: Lymphoid neoplasms. 36, 1720–1748, DOI: 10.1038/s41375-022-01620-2. Number: 7 Publisher: Nature Publishing Group.
- [4] Campo, E. et al. The international consensus classification of mature lymphoid neoplasms: a report from the clinical advisory committee. 140, 1229–1253, DOI: 10.1182/blood.2022015851.
- [5] Zhang, Y. et al. Comparison of the revised 4th (2016) and 5th (2022) editions of the world health organization classification of myelodysplastic neoplasms. 36, 2875–2882, DOI: 10.1038/s41375-022-01718-7. Number: 12 Publisher: Nature Publishing Group.
- [6] Cree, I. A. The WHO classification of haematolymphoid tumours. 36, 1701–1702, DOI: 10.1038/s41375-022-01625-x. Number: 7 Publisher: Nature Publishing Group.
- [7] Chandradevan, R. et al. Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. 100, 98–109, DOI: 10.1038/s41374-019-03257. Number:1 Publisher: Nature Publishing Group.
- [8] Lewis, J. E. et al. An automated pipeline for differential cell counts on whole-slide bone marrow aspirate smears. 36, 100003, DOI: 10.1016/j.modpat.2022.100003.

- [9] Tayebi, R. M. et al. Automated bone marrow cytology using deep learning to generate a histogram of cell types. 2, 1–14, DOI: 10.1038/s43856-022-00107-6. Number: 1 Publisher: Nature Publishing Group.
- [10] Choi, J. W. et al. White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. 12, e0189259, DOI: 10.1371/journal.pone.0189259. Publisher: Public Library of Science.
- [11] Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T. & Marr, C. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image dataset. DOI: 10.1182/blood.2020010568.
- [12] Fazeli, S., Samiei, A., Lee, T. D. & Sarrafzadeh, M. Beyond labels: Visual representations for bone marrow
- [13] cell morphology recognition. In 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), 111–117 (IEEE, 2023).
- [14] Manescu, P. et al. Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning. *Scientific Reports* 13, 2562 (2023).
- [15] Opitz, D. & Maclin, R. Popular ensemble methods: An empirical study. *Journal artificial intelligence research* 11, 169–198 (1999).
- [16] Sagi, O. & Rokach, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery* 8, e1249 (2018).
- [17] Yang, Y., Lv, H. & Chen, N. A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review* 56, 5545–5589 (2023).
- [18] Hameed, Z., Zahia, S., Garcia-Zapirain, B., Javier Aguirre, J. & Maria Vanegas, A. Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors* 20, 4373 (2020).

- [19] Kundu, R., Das, R., Geem, Z. W., Han, G.-T. & Sarkar, R. Pneumonia detection in chest x-ray images using an ensemble of deep learning models. *PloS one* 16, e0256630 (2021).
- [20] Wang, Q., Ma, Y., Zhao, K. & Tian, Y. A comprehensive survey of loss functions in machine learning. *Annals Data Science* 1–26 (2020).
- [21] Batista, G. E., Prati, R. C. & Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter* 6, 20–29 (2004).
- [22] Huang, G. et al. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109* (2017).
- [23] Szegedy, C. et al. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9 (2015).
- [24] Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114 (PMLR, 2019).
- [25] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [26] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
- [27] Ahmed, K. B., Goldgof, G. M., Paul, R., Goldgof, D. B. & Hall, L. O. Discovery of a generalization gap of convolutional neural networks on COVID-19 x-rays classification. 9, 72970–72979, DOI: 10.1109/ACCESS.2021.3079716. Conference Name: IEEE Access.
- [28] Ben Ahmed, K., Hall, L. O., Goldgof, D. B. & Fogarty, R. Achieving multisite generalization for CNN-based disease diagnosis models by mitigating shortcut learning. 10, 78726–78738, DOI: 10.1109/ACCESS.2022.3193700. Conference Name: IEEE Access.
- [29] Adebayo, J. et al. Sanity checks for saliency maps. In *Neural Information Processing Systems* (2018).
- [30] Yona, G. & Greenfeld, D. Revisiting sanity checks for saliency maps. *arXiv preprint arXiv:2110.14297* (2021).

- [31] Kim, B. et al. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 4149–4157 (IEEE, 2019).
- [32] Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis machine intelligence* 20, 1254–1259 (1998).
- [33] Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).
- [34] Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328 (PMLR, 2017).
- [35] Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
- [36] Kumar, A., Kim, J., Lyndon, D., Fulham, M. & Feng, D. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE journal biomedical health informatics* 21, 31–40 (2016).
- [37] Dong, X., Yu, Z., Cao, W., Shi, Y. & Ma, Q. A survey on ensemble learning. *Frontiers Computer Science* 14, 241–258 (2020).
- [38] Mazurowski, M. A., Buda, M., Saha, A. & Bashir, M. R. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri. *Journal magnetic resonance imaging* 49, 939–954 (2019).
- [39] Zhang, J., Xie, Y., Wu, Q. & Xia, Y. Medical image classification using synergic deep learning. *Medical image analysis* 54, 10–19 (2019).
- [40] Chattopadhyay, S., Singh, P. K., Ijaz, M. F., Kim, S. & Sarkar, R. Snapensemfs: a snapshot ensembling-based deep feature selection model for colorectal cancer histological analysis. *Scientific Reports* 13, 9937 (2023).

- [41] Annavarapu, C. S. R. et al. Deep learning-based improved snapshot ensemble technique for covid-19 chest x-ray classification. *Applied Intelligence* 51, 3104–3120 (2021).
- [42] Dietterich, T. G. et al. Ensemble learning. *The handbook brain theory neural networks* 2, 110–125 (2002).
- [43] Mestrum, S. G. C. et al. Integration of the ki-67 proliferation index into the ogata score improves its diagnostic sensitivity for low-grade myelodysplastic syndromes. 113, 106789, DOI: 10.1016/j.leukres.2022.106789.
- [44] Mestrum, S. G. C. et al. The proliferation index of erythroid cells predicts the development of transfusion-dependence in myelodysplastic syndrome patients with mildly reduced hemoglobin levels at initial diagnosis. 6, e804, DOI: 10.1097/HS9.0000000000000804.
- [45] Cree, I. A. et al. Counting mitoses: SI(ze) matters! 34, 1651–1657, DOI: 10.1038/s41379-021-00825-7. Number: 9 Publisher: Nature Publishing Group.
- [46] Song, M.-K. et al. High ki-67 expression in involved bone marrow predicts worse clinical outcome in diffuse large b cell lymphoma patients treated with r-CHOP therapy. 101, 140–147, DOI: 10.1007/s12185-014-1719-3.
- [47] Kaushansky, K. et al. *Williams Hematology*, 10th Edition (McGraw Hill / Medical), 10th edition edn.
- [48] Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. 5987–5995, DOI: 10.1109/CVPR.2017.634. ISSN: 1063-6919.
- [49] Buslaev, A. et al. Albuumentations: Fast and flexible image augmentations. *Information* 11, DOI: 10.3390/info11020125 (2020).
- [50] Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. 128, 336–359, DOI: 10.1007/s11263-019-01228-7. 1610.02391[cs].
- [51] Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise, DOI: 10.48550/arXiv.1706.03825. 1706.03825[cs,stat].

- [52] Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps, DOI: 10.48550/arXiv.1312.6034. 1312.6034[cs].
- [53] Kapishnikov, A., Bolukbasi, T., Viégas, F. & Terry, M. XRAI: Better attributions through regions, DOI: 10.48550/arXiv.1906.02825. 1906.02825[cs,stat].
- [54] Dong, W., Moses, C. & Li, K. Efficient k-nearest neighbor graph construction for generic similarity measures. In Proceedings of the 20th international conference on World wide web, 577–586 (2011).
- [55] McInnes, L., Healy, J., Saul, N. & Grossberger, L. Umap: Uniform manifold approximation and projection. The Journal Open Source Software 3, 861 (2018).

7 Chapter 7: Conclusion: Summary and Future Work

7.1 SUMMARY

This dissertation outlines my endeavor to tailor machine learning techniques for the intricate task of deciphering complex diseases. Through a combination of innovative methods and practical applications, this work showcases the significant potential of leveraging medical data through novel computational approaches to support clinical practices and inform regulatory decisions. The ultimate goal of this research is to improve patient care and contribute positively to public health on a broader scale.

Machine learning and deep learning represent revolutionary techniques with the power to transform healthcare, offering unprecedented capabilities in data analysis, pattern recognition, and predictive modeling. However, it's crucial to acknowledge that medicine is a uniquely complex field, distinct from many other areas where these technologies have been applied. The intricacies of medical science, combined with the ethical, legal, and personal nuances of patient care, present a landscape replete with pitfalls and challenges that can be easily overlooked or oversimplified by researchers from either the machine learning or medical domains. These challenges range from the risk of algorithmic bias to the oversimplification of complex biological interactions, and from data privacy concerns to the interpretability of model outputs in a clinical context. In the work presented in this dissertation, I have endeavored to address some of these issues, or at least to propose viable solutions.

EQUALLY VITAL: THE INDISPENSABLE ROLE OF SOCIAL DETERMINANTS IN HEALTH OUTCOMES

In the pursuit of a more comprehensive understanding of patient health, my work underscores the critical importance of Social Determinants of Health (SDoH), revealing that medical factors alone do not paint the full picture.

Chapter 2 delves into "Topic Modeling on Clinical Social Work Notes for Exploring Social Determinants of Health Factors," employing advanced topic modeling techniques to sift through clinical social work notes. This innovative approach sheds light on the nuanced ways in which SDoH factors, such as economic stability, education, and social context, play a pivotal role in patient health and treatment outcomes. By analyzing these often-overlooked aspects of patient records, this chapter demonstrates the profound impact of social circumstances on health, advocating for their integration into holistic patient care strategies.

Building on this foundation, Chapter 3, "Revealing the impact of social circumstances on the selection of cancer therapy through natural language processing of social work notes," takes a more focused look at how these social determinants influence critical medical decisions, specifically in the context of cancer therapy selection. Through the lens of natural language processing, this work highlights the intricate relationship between a patient's social environment and their treatment pathway, emphasizing that factors such as access to healthy food or supportive family structures can significantly sway therapeutic choices. This chapter not only reinforces the indispensability of considering SDoH in medical decision-making but also illustrates the practical implications of such considerations in tailoring treatment to individual patient needs.

HUMAN EXPERTISE IN AI: ENHANCING DEEP LEARNING THROUGH ANNOTATION AND FEEDBACK

The integration of human expertise into the development and refinement of AI, particularly in deep learning models, stands as a pivotal advancement in the quest to harness the full potential of these technologies in healthcare. This segment of the dissertation delves into the critical role of detailed, high-quality human annotations and the strategic use of human feedback to substantially elevate model performance, especially when extensive annotations are not feasible.

Chapter 4, "Aligning Synthetic Medical Images with Clinical Knowledge using Human Feedback," highlights this interplay, demonstrating how the integration of clinical insights through feedback loops can refine the accuracy and relevance of synthetic medical images. This process not only tailors the AI's learning trajectory but also ensures the generated images are clinically plausible, thereby enhancing the model's utility and trustworthiness in medical applications. Such feedback mechanisms prove invaluable, especially in scenarios where direct and detailed annotations may be limited or infeasible, allowing for continuous model improvement and alignment with evolving clinical knowledge.

In contrast, Chapter 6, "DeepHeme: A High-Performance, Generalizable, Deep Ensemble for Bone Marrow Morphometry and Hematologic Diagnosis," showcases the transformative power of detailed human annotation. This meticulous process involves the comprehensive labeling of medical images or data by experts, providing a rich, nuanced dataset that serves as the foundation for training highly accurate and reliable deep learning models. DeepHeme leverages this extensive annotated dataset to achieve superior performance in bone marrow analysis and hematologic diagnosis, setting new benchmarks for model reliability and applicability in clinical practice. The chapter underscores the indispensable role of human-annotated data in developing AI tools that not only perform with high precision but also align closely with the complexities of real-world medical diagnostics.

STRATEGIC FEATURE ENGINEERING: ILLUMINATING DISEASE MECHANISMS WITH PRECISION

Chapter 5, "Spatial Cell Type Enrichment Predicts Mouse Brain Connectivity," serves as a testament to the power of strategic feature engineering in machine learning for biomedical research. This section of the dissertation spotlights the pivotal role of selecting biologically relevant features, such as spatial cell type enrichment, to improve the interpretability and accuracy of models that study complex biological systems. By prioritizing cell type information over gene enrichment, this work showcases the ability to uncover deeper insights into mouse brain connectivity, emphasizing the value of focused feature selection. This approach not only enhances the predictive performance of the model but also ensures that the features employed are directly aligned with biological reality, offering clearer insights into the underlying mechanisms of disease.

In summary, this dissertation has presented a multifaceted exploration of real-world electronic health records (EHR) through computational lenses, ranging from classical machine learning techniques to cutting-edge natural language processing models based on transformers. We have ventured beyond the conventional focus on purely medical factors, incorporating the often-overlooked social determinants of health to provide a more comprehensive understanding of patient well-being. Our methodologies have harnessed the power of detailed human annotations and feedback to refine deep learning models, ensuring their practical applicability and enhancing their performance in clinical settings. Additionally, we have underscored the significance of strategic feature engineering, advocating for the prioritization of relevant biological features to advance model interpretability and the discovery of disease insights.

These diverse approaches, from the granular examination of social factors to the incorporation of expert human input and the deliberate selection of model features, collectively aim to generate high-quality, real-world evidence. This evidence is crafted not only to address research gaps and medical needs but also to

inform and guide improved clinical care practices. As we strive for methodological innovation, the ultimate goal of this research remains steadfast: to contribute meaningfully to patient outcomes and public health, harnessing the transformative power of machine learning to illuminate the complex tapestry of human health.

7.2 FUTURE WORK

I believe this dissertation has carved out several pathways for enhancing machine learning applications in healthcare, each with its unique potential for future expansion. As we look ahead, integrating Social Determinants of Health (SDoH) with clinical data emerges as a pivotal step towards enriching the characterization of complex diseases. This comprehensive approach promises to refine diagnostic accuracy and tailor preventive and therapeutic interventions more closely to patient-specific contexts. The intricate tapestry of a patient's life, woven with threads of environmental, socioeconomic, and lifestyle patterns, alongside their clinical picture, could offer unprecedented insights into addressing and preempting health disparities.


Progressing further, the pursuit of blending engineered genomic and imaging features stands out as a frontier with immense potential to predict clinical outcomes and understand patient heterogeneity. This integration aims to reveal the complex interplay between phenotypic expressions and genetic underpinnings, potentially illuminating new subtypes of diseases and leading to more targeted therapeutic strategies. Moreover, the correlation between imaging biomarkers and genomic data could unlock novel understandings of disease mechanisms, fostering advancements in the realms of personalized and precision medicine.

Lastly, the fusion of natural language processing with imaging data to develop comprehensive vision-language models represents an exciting horizon for medical diagnostics. The envisioned models would not only automate the interpretation of medical images but also contextualize them within the rich narrative of clinical notes, offering a multifaceted diagnostic tool. Such advancements could streamline the diagnostic process, providing clinicians with synthesized, actionable insights. The journey to realize this vision will necessitate a concerted effort to refine sophisticated algorithms capable of discerning the nuanced interplay between textual and visual information, ensuring clinical utility and accuracy. The road ahead is paved with both challenges and opportunities, beckoning a future where machine learning not only complements but significantly enhances clinical decision-making and patient care.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

000A457907724D1... Author Signature

2/26/2024
Date