

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Harnessing Emerging Technologies in Medicine: Practical Applications from the Micro to the Macro

Permalink

<https://escholarship.org/uc/item/5mk8n0jj>

Author

Wong, Daniel

Publication Date

2021

Peer reviewed|Thesis/dissertation

Harnessing Emerging Technologies in Medicine: Practical Applications from the Micro
to the Macro

by
Daniel Wong

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:

Michael J Keiser

Michael J Keiser

4DF1BD06D670465...

Chair

DocuSigned by:

Atul Butte

Atul Butte

DocuSigned by:

Sourav Bandyopadhyay

Sourav Bandyopadhyay

F1E602E27884493...

Committee Members

Copyright 2021

by

Daniel Wong

ACKNOWLEDGEMENTS

I thank my advisers Michael Keiser and Atul Butte for shaping me into the scientist that I am today. Despite both being incredibly busy, I appreciate the many both brief and extensive moments of discussing science with them, probing new hypotheses, and brainstorming how to advance scientific knowledge to be used for the benefit of many. I thank Michael Keiser for actively leading my day-to-day mentorship, and for getting into the weeds of scientific challenges with me and synergistically helping work towards solutions. I thank Atul Butte for his high-level guidance of my broader scientific journey, and for his invaluable insights into the future of medicine. I thank them both for the contagious zeal they have towards their craft. Much of my own excitement and optimism I have towards scientific progress and its utility to human betterment can be attributed to their mentorship.

I also thank my qualifying exam committee for helping to focus and sharpen my science. I have Sourav Bandyopadhyay, Marina Sirota, Martin Kampmann, and Rima Arnaout, to thank for providing positive pressure to test my scientific approaches, challenging my hypotheses, and preparing me to think on-the-fly to properly defend and present my science to an audience.

I thank my thesis committee consisting of Michael Keiser, Atul Butte, and Sourav Bandyopadhyay for helping me achieve major milestones in my academic career. They provided life insight into broader areas outside of graduate school, and prepared me to be an active contributor to the scientific community, wherever I will be in the future.

I dedicate this thesis to my family, especially my two parents Liu and Tony, without whom none of this would be possible. They made so many sacrifices to ensure that I could devote my energies to academia and working with my mind, as opposed to laboring with my hands and body in hard labor. I am truly privileged, and this position comes largely from no merit or works of my own, but from the circumstances and sacrifices of others. I would also like to thank my larger church family, and all my brothers and sisters in Christ. I am so grateful that I do not need to run this race alone (1 Corinthians 9:24-27), but have so many others to count on for life's greatest challenges that only God, and not science even as much as I love it, can address.

The text of Chapter 3 is a reprint of material as it appears in:

Wong, D.R. et al. "Prototype of running clinical trials in an untrustworthy environment using blockchain." *Nature Communications* **10**, 917 (2019).

ABSTRACT

Harnessing Emerging Technologies in Medicine: Practical Applications from the Micro to the Macro

Daniel Wong

The field of computing has made incredible strides in how we process and analyze information. Here I consider how to take these advances in theory and apply them in novel ways to different scopes of the medical pipeline: from the cellular level, to the clinic, and finally to the broadest level of healthcare and data management between different institutions. In particular, I sought to explore how the emergent technologies of deep learning and blockchain technologies can provide practical improvements to biomedical research, with a focus in mostly the neurodegenerative space.

I investigated how to improve high-content high-throughput drug screens through deep learning—moving experimentation and analysis off the bench and into the servers as a first pass to generate hypotheses. There's a plethora of information found in these screens. Through deep learning, such information can be refined to actionable insights, and complex algorithms can make sense of data that is uninterpretable to humans. So much time and money is lost in pharmaceutical endeavors due to following up on poor drug candidates. I hypothesized that if we had better predictions to begin with, then the search space for new drugs would be optimized.

Hence, I developed an *in-silico* approach to phenotypic drug screens, by providing a deep learning method to glean related biological signals between markers and project them into images used for hypothesis generation. This resulted in better triaging efforts for finding active compounds in an Alzheimer's-centered drug screen. More generally, I showed that this method can apply to another phenotypic screen in a completely different biological system.

I showed that deep learning can not only assist at the bench, but also be useful in the clinical research space. Again focusing on neurodegenerative efforts, I developed deep learning methods to help neuropathologists identify amyloid beta pathologies. Following the idea that there is wisdom to be found in the crowd, I showed how consensus learning and ensembling over many experts can provide a robust framework for accurate detection despite being in a space where there may be expert disagreement. These efforts provided not only automation, but also potential for standardization in the field as I provided a method that is not only accurate but also completely consistent. I showed the method's potential for practical adoption through a prospective validation study.

Finally, at the broad level of data exchange between multiple medical institutions and data centers, I introduced a blockchain-based system into the clinical network, showing that blockchain technology has relevant applications outside of merely financial problems and questions. In particular, I made advances in the clinical trial process, and performed a proof-of-concept in which I simulated a real clinical trial on a more secure, traceable, and trustworthy data management system. We showed that such a system is immune to data tampering, and also provides a useful audit trail for regulators.

TABLE OF CONTENTS

INTRODUCTION	1
CHAPTER 1: TRANS-CHANNEL FLUORESCENCE LEARNING IMPROVES HIGH-CONTENT SCREENING FOR ALZHEIMER'S DISEASE THERAPEUTICS	5
1.1 ABSTRACT	6
1.2 INTRODUCTION	7
1.3 RESULTS	10
1.3.1 We Constructed a Drug Screen Dataset of Tau Prion Propagation for ML Training	10
1.3.2 We Trained a Trans-Channel ML Model to Predict Hidden Tau Phosphorylation	14
1.3.3 Trans-Channel ML Improved Hit Rate and Compound Triaging in Tauopathy HCS	20
1.3.4 We Tested Trans-Channel ML Method on an HCS with Different Biological Conditions	26
1.4 DISCUSSION	30
1.5 METHODS	35
1.5.1 Tg2541 Mouse Line	35
1.5.2 Experimental Setup for the Cellular Tau(P301S)-YFP Aggregation Assay Training Dataset	36
1.5.3 Constructing the Archival HCS Dataset	39
1.5.4 Training and Evaluation of the ML Models from the Tauopathy Dataset	39
1.5.5 Performance Metrics	41
1.5.6 Construction and Evaluation of PQML and PQC	42
1.5.7 Experimental Setup for the Secondary Dose-Response Experiments	43

	viii
1.5.8 Assessing Activity from Dose Response Curves	44
1.5.9 Experimental Setup for Validation Assay of Functional Genomics Screen in Osteosarcoma Line	45
1.5.10 Training and Evaluation of the ML Models Used for the Osteosarcoma Dataset	46
1.5.11 Quantification and Statistical Analyses	47
1.6 SUPPLEMENTAL FIGURES	48
CHAPTER 2: DEEP LEARNING FROM MULTIPLE EXPERTS IMPROVES IDENTIFICATION OF AMYLOID NEUROPATHOLOGIES	61
2.1 ABSTRACT	62
2.2 INTRODUCTION	63
2.3 RESULTS	67
2.3.1 We curated a multi-annotator dataset and found annotation differences among experts and consensus schemes	67
2.3.2 CNNs mimic human annotators, and consensus CNNs mimic a consensus of experts	73
2.3.3 Consensus provided superior models and annotation benchmarks	75
2.3.4 Model interpretation reflected differences in human expertise	79
2.3.5 Ensemble learners bolstered performance and were robust to intentionally noisy “annotators”	82
2.3.6 Annotators performed a second phase of annotation and were self-consistent	86
2.3.7 Models prospectively enriched for minority A β classes and favored consensus learning	88
2.4 DISCUSSION	91
2.5 METHODS	97
2.5.1 Slide Preparation	97

	ix
2.5.2 Data Collection and Annotation (Phase-one)	97
2.5.3 Training and Evaluation of DL Models	99
2.5.4 Assessing Consensus Model and Benchmark Superiority	101
2.5.5 Model Interpretability	103
2.5.6 Ensemble Training and Evaluation	103
2.5.7 Setup of Phase-Two Prospective Validation and Analysis	106
2.6 SUPPLEMENTAL FIGURES	110
CHAPTER 3: PROTOTYPE OF RUNNING CLINICAL TRIALS IN AN UNTRUSTWORTHY ENVIRONMENT USING BLOCKCHAIN	128
3.1 ABSTRACT	128
3.2 INTRODUCTION	129
3.3 RESULTS	131
3.3.1 Network Protocol	131
3.3.2 Data Transaction Details	134
3.3.3 Version Controls	139
3.3.4 Simulation of a Previously Completed Clinical Trial	139
3.3.5 Patient and Clinical Investigator Encounters	140
3.3.6 User Mediated Corruption	141
3.3.7 Storage Corruption	142
3.3.8 Zero-knowledge Proof of Purity	143
3.3.9 Expedited Adverse Event Reporting	144
3.4 DISCUSSION	144
3.5 METHODS	150
3.5.1 Portal Implementation and Simulated Trial Data	150
3.5.2 Code Availability	150

	x
3.5.3 CRF Extraction	150
3.6 SUPPLEMENTAL MATERIALS	151
3.6.1 Supplemental Methods	151
3.6.2 Supplemental Discussion	152
3.6.3 Supplemental Tables	154
3.6.4 Supplemental Notes	155
CONCLUSION	205
REFERENCES	208

LIST OF FIGURES

Figure 1.1: We constructed a cellular Tau(P301S)-YFP aggregation assay that modeled prion propagation in vitro with an additional AT8-pTau channel for ML training.	13
Figure 1.2: The ML model generated a phosphorylated AT8-pTau channel, solely from DAPI and YFP-tau channels.	16
Figure 1.3: The ML model achieved good performance.	18
Figure 1.4: Transforming the archival HCS data with ML-predicted AT8-pTau images revealed previously unknown active compounds.	24
Figure 1.5: ML triaged active compounds better than the conventional method.	26
Figure 1.6: The trans-fluorescence learning method generated accurate cyclin-B1 signal from Hoechst signal, on an independent and biologically unrelated dataset.	30
Supplemental Figure 1.1: Additional image examples showing trans-channel predictions.	49
Supplemental Figure 1.2: Transchannel-learning from single channel input to objective AT8.	51
Supplemental Figure 1.3: Overlap of AT8-pTau signal and YFP-tau signal.	53
Supplemental Figure 1.4: Pearson performance by drug condition.	54
Supplemental Figure 1.5: Excitation and emission spectra of fluorophores used in the two biologically distinct datasets.	55
Supplemental Figure 1.6: Ablation testing on the tau dataset.	57
Supplemental Figure 1.7: Additional triaging analysis of missed but active compounds, and also enrichment by a different ML Standard.	59
Supplemental Figure 1.8: Binary thresholding of the label image in constructing the AUC.	60
Figure 2.1: We curated annotations of A β neuropathologies from multiple experts, and found differing degrees of consensus.	71

	xii
Figure 2.2: Inter-rater agreement varies by class and annotator.	72
Figure 2.3: We trained models to learn human annotation behavior and consensus strategies.	74
Figure 2.4: Consensus models performed better than individual-expert models across all benchmarks.	78
Figure 2.5: Class activation maps (CAM) of DL models indicate progression of human expertise.	82
Figure 2.6: Ensembles improve performance and are robust to false information.	86
Figure 2.7: Models prospectively predict human annotation, with consensus models performing the most consistently.	91
Supplemental Figure 2.1. Demographics of the 43 patients from all institutions.	111
Supplemental Figure 2.2: Negative, flag, and not sure annotations for both phase one and phase two.	112
Supplemental Figure 2.3: Models capture interrater agreement patterns among experts.	114
Supplemental Figure 2.4: Model performance by stain type for the three A β classes.	116
Supplemental Figure 2.5: Model evaluation differences on color-normalized versus unnormalized test images.	117
Supplemental Figure 2.6: Consensus benchmarks facilitate higher model performance.	119
Supplemental Figure 2.7: CAM analysis of UG2.	121
Supplemental Figure 2.8: Intra-rater agreement for self-enrichment and consensus-enrichment.	123
Supplemental Figure 2.9: Ensemble weights.	125
Supplemental Figure 2.10: CAMs for each class.	126
Supplemental Figure 2.11: Distribution of positive annotations in the test set.	127
Figure 3.1: The idealized clinical trial network in the context of a blockchain-based record system.	132

Figure 3.2: The growing blockchain.

135

Figure 3.3: Portal functionality.

138

LIST OF TABLES

Table 1.1: Comparison of dose-response success rates using different methods	22
Supplemental Table 3.1: Subject arm and codes	154
Supplemental Table 3.2: Arm description	154

LIST OF ABBREVIATIONS

AD: Alzheimer's disease

AUPRC: area under the precision recall curve

AUROC: area under the receiver operating characteristic

A β : amyloid beta

CAA: cerebral amyloid angiopathy

CAMs: class activation maps

CNN: convolutional neural network

CRF: case report form

CRO: Clinical Research Organization

DAPI: 4',6-diamidino-2-phenylindole

DI: deionized

DL: deep learning

DPBS: Dulbecco's Phosphate

DSMB: Data Safety Management Board

EHR: electronic health record

FDA: Food and Drug Administration

GFP: green fluorescent protein

HCS: high-content high-throughput screening

HSV: hue saturation value

ICH: International Council for Harmonization of Requirements for Pharmaceuticals for Human Use

ICMJE: International Committee of Medical Journal Editors

IRB: Institutional Review Board

IVRS: Interactive Voice Response System

ML: machine learning

MSE: mean squared error

NFTs: neurofibrillary tangles

NIAID: National Institute of Allergy and Infectious Disease

NP#: neuropathologist #

PFA: paraformaldehyde

PQC: conventional method's priority queue

PQML: ML-derived priority queue

PTA: Phosphotungstic acid

UG#: undergraduate novice #

WSIs: whole slide images

YFP-tau: 0N4R isoform of tau fused to the yellow fluorescent protein

INTRODUCTION

Coming into graduate school with a background in both molecular biology and computer science, I knew that the two fields had many opportunities for integration. The last few years has seen something of a data revolution, with so much data being generated every day, yet so much of it being under-utilized. The biomedical field is no exception, and is exploding with new yet untapped data every day. Why is data not being used to its full potential? Either because there's too much and it becomes unmanageable, or because we don't have the proper methods and technologies to probe these complex datasets. These problems exist at all levels of the medical and biological research pipeline. They span from 1) micro-level challenges like cell-based drug discovery campaigns for combatting disease, to 2) higher-scope problems like equipping doctors with the right tools to analyze patient data, and finally to 3) the most macro-level problems like data management and exchange between health institutions. During my PhD work, I traversed this pipeline spanning from the micro to the macro, all with the theme of adapting and developing new technologies to make key insights and improvements from what is often unwieldy and complex data.

Starting from the lowest, most granular level of the pipeline, I proposed a practical improvement to cellular drug screening campaigns. At the level of cellular research, the data is arguably the most complex of any other levels due to the intricacy of biological systems that far surpasses any man-made system. One common way to generate data is for drug developers to

model a disease system and selectively perturb them to reduce the disease phenotype. Similar to other levels of the pipeline, there are antiquated methods that require much manual labor and time, such as non-generalizable image analysis software that must be tuned with each new experiment. Also, when drug developers search for new drugs, they often need to perform costly and time-intensive laboratory experiments. One can imagine that previously performed experiments might contain useful information to aid answering these new questions. Probing and mining old data for new insights can both save time and effort, and provide cheap, actionable hypotheses to then follow up via new experiments. I hypothesized that there is much more useful phenotypic information within existing data repositories. Although difficult to harness and interpret by humans, this information can actually be extracted and learned by machines. This can aid drug developers by avoiding costly new experiments and relying on already-existing large datasets. This is covered in Chapter 1.

The next section of this thesis dissertation is a level up in scope, and focuses on improvement of pathological diagnoses at the individual pathologist level. The same problems present at the micro level are present here as well, with complex data, cumbersome existing methods, and uncertainty in how to best interpret data. However, these problems take on slightly different forms. Human disease phenotypes are complex, and oftentimes diagnoses are not agreed upon even among experts in the field. How are we to cure diseases if we cannot even arrive at a consensus of what that disease is, or what cases can be classified as such? Hence, accurate and standardized pathology is of the utmost importance. Yet it faces problems like

lacking consensus, and requiring much time and labor to perform high-skilled tasks that require years of medical training to perfect. Moreover, common data sources like immunohistochemical images are information-rich and hence difficult to analyze. Just like the cellular data at the micro-level of the biomedical research pipeline, gleaning information is difficult due to the presence of complex information. I address these limitations, and provide a standardized ground truth for challenging pathology tasks where unified agreement among experts may not exist. I also train deep learning algorithms to learn a high-skilled task in a matter of hours that a human pathologist would require years to master. Chapter 2 focuses on this standardization and automation, which are often lacking in pathology, yet are critical to accurate disease diagnosis, understanding, and treatment. Although here I present a case study specifically in computer vision applied to neuropathology, the methodology and insight applies more generally to pathology.

Finally, the last part of this thesis dissertation moves up in scope to the bird's eye view of the clinical space and the connections between different research institutions, hospitals, and regulators. The sheer amount and complexity of patient and clinical trial data did not seem to be appropriately managed by current methods, many of which are manual processes done by hand with old software systems. In fact, around 70% of medical devices still run on antiquated operating systems like Windows XP¹, and important medical data is often shared with old hardware tools like fax machines². The clinical space suffers from antiquity and reliance on old methods. Although this problem is far too ambitious to solve for one PhD thesis, I saw an

opportunity in the clinical trial space to bridge part of this gap in where we should be with all the amazing technologies available to us today, and where we currently are. Change is hard to initiate, but my hope is that with one PhD thesis dissertation at a time and one paper at a time, practice will adopt theory, and technological innovation in academia will make inroads into the clinical space. My proof of concept and solution to one of the many problems in clinical trial management is presented in Chapter 3, which was published in Nature Communications. Via blockchains, I demonstrate how clinical trial management can be completely transparent, secure, and traceable, all with gains in efficiency.

Chapter 1: Trans-channel fluorescence learning improves high-content screening for Alzheimer's disease therapeutics

Daniel R. Wong^{1,2,3,4,5,6,7}, Jay Conrad^{**1}, Noah Johnson^{**1}, Jacob Ayers¹, Annelies
Laeremans^{4,8}, Joanne C. Lee¹, Jisoo Lee¹, Stanley B. Prusiner¹, Sourav Bandyopadhyay^{4,8}, Atul
J. Butte^{2,6,7}, Nick Paras¹, Michael J. Keiser^{1,2,3,4,5*}

1. Institute for Neurodegenerative Diseases, University of California, San Francisco, CA, 94158, USA
2. Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, 94158, USA
3. Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, 94158, USA
4. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, 94158, USA
5. Kavli Institute for Fundamental Neuroscience, University of California, San Francisco, San Francisco, CA, 94158, USA

6. Department of Pediatrics, University of California, San Francisco, CA, 94158, USA
7. Center for Data-Driven Insights and Innovation, University of California, Office of the President, Oakland, CA, 94607, USA
8. Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, 94158, USA

****These authors contributed equally**

1.1 ABSTRACT

In microscopy-based drug screens, fluorescent markers carry critical information on how compounds affect different biological processes. However, practical considerations may hinder the use of certain fluorescent markers. Here, we present a deep learning method for overcoming this limitation. We accurately generated predicted fluorescent signals from other related markers and validated this new machine learning (ML) method on two biologically distinct datasets. We used the ML method to improve the selection of potentially efficacious therapeutic compounds for Alzheimer's disease (AD) from high-content high-throughput screening (HCS). The ML method identified novel compounds that effectively blocked tau aggregation, which would have been missed by traditional screening approaches unguided by ML. The ML method also improved triaging efficiency of compound rankings over a current in-house screening approach. We reproduced this ML method on a biologically independent cancer-based dataset,

demonstrating its generalizability. The approach is disease-agnostic and applicable across fluorescence microscopy datasets.

1.2 INTRODUCTION

HCS efforts generate a wealth of complex phenotypic information that is pivotal to the drug discovery process. Extracting biological insight from these observations, however, often requires performing multiple experiments, which can demand extensive time and resources³. We posit that ML can extract actionable information that is otherwise encoded within HCS and learn the phenotypic relationships between related biological processes. As a corollary, we can potentially glean information on related but different biological processes—increasing the power of computationally mining large archival HCS datasets to gain new information.

The traditional way to gather more information within HCS is simply to add biological markers to track different processes. However, we cannot introduce new markers for archival datasets because the experiments are finished. Furthermore, capturing additional markers may be impractical due to expensive or cumbersome visualization procedures, like the optimization of multi-channel fluorescent immunohistochemistry or interference across detection wavelengths^{4,5}. We overcame these marker limitations computationally by directly *learning* the phenotypic relationships between a highly-informative but cumbersome marker and other similar yet more easily-accessible markers. These hidden relationships were then projected into *de novo* images that displayed the desired fluorescent signal of the cumbersome marker. In this study, we focused

on an archival HCS dataset that tracked the phenotypic effects of small molecules for the treatment of AD.

Despite extensive drug discovery efforts, no effective treatments exist that prevent or even slow the progression of AD⁶. Of the many therapeutic targets under investigation⁷, the pathogenic^{8–15} misfolding and accumulation of tau protein into neurofibrillary tangles (NFTs) has emerged as a target mechanism¹². Younger decedents tend to have more aggressively propagating tau prions than older decedents¹⁴. Furthermore, studies have linked this tau prion propagation^{16,17} and aggregation¹⁸ to increased tau hyperphosphorylation. Although hyperphosphorylation is not required to form ordered assemblies of tau^{19,20}, NFTs in the brains of deceased AD patients are highly enriched for hyperphosphorylated tau (pTau)^{18,21}. Hyperphosphorylation inhibits the binding of tau to microtubules^{22,23}, thus preventing normal microtubule assembly^{23,24} and axonal transport²⁵. It may also increase the propensity of tau to be recruited into NFTs^{8,25,26}. These factors promote disease phenotypes including synaptic dysfunction, enhanced neuroinflammation, and neuronal cell loss^{23,25,26}.

Accordingly, discovering compounds that inhibit this tau prion propagation may improve our understanding of the disease and provide avenues for novel therapeutics. Our study focused on inhibiting the propagation and aggregation of tau in cells exposed to tau prions. To screen for compounds, we developed a HCS procedure that utilizes a biosensor cell-line overexpressing a 0N4R isoform of tau fused to the yellow fluorescent protein (YFP-tau)²⁷. Due to the correlation of tau oligomerization with increased hyperphosphorylation, we posited that tracking the level of

pTau in response to drug treatment would provide an additional point of validation to identify molecules that inhibit prion propagation. Furthermore, we hoped that synthesizing pTau signal would eliminate spurious signal such as contamination of the YFP channel by autofluorescent cell debris. The AT8 antibody allows us to quantify these correlated hyperphosphorylation events and is traditionally used to label disease-associated paired helical filaments in human tissue samples²⁸. This antibody is specific to the Ser202/Thr205 epitope^{29,30}, which is one of the most hyperphosphorylated residues in AD-afflicted brains^{18,31}. Furthermore, hyperphosphorylation of this region has been linked to increased prion propagation^{16,17} and aggregation¹⁸. Although AT8 immunoreactivity is a useful surrogate to identify a disease-relevant phenotype and recognize clinically relevant pTau²⁹, it was not used in our HCS because immunostaining would preclude some advantages of live-cell imaging^{32,33}. Therefore, we turned to a different solution based on ML. We computationally synthesized a representation of the AT8 channel from existing data, rather than repeating the vast HCS with this useful antibody, which would have been laborious and cost-prohibitive.

Two seminal papers outlined by Christiansen³⁴ and Ounkomol³⁵ describe a ML method for predicting fluorescence images from bright-field images. Augmented microscopy approaches like these extract latent information from images post hoc³⁶ and in a goal-driven way^{37,38}. We reasoned that if two fluorescent channels were sufficiently related, we could train models to extract hidden information and generate “trans-channel” images depicting AT8-pTau from related YFP-tau images. Our logic hinged on the expectation that a ML method could both

identify subtle aggregate morphology and exploit the phenotypic correlations between tau hyperphosphorylation and aggregation.

We collected a three-channel dataset—4',6-diamidino-2-phenylindole (DAPI) nuclear marker, YFP-tau, and AT8-pTau—one year after the original HCS experiment that had only used YFP-tau and DAPI. After training our generative model on this new three-channel dataset, we applied this model to the archival HCS dataset to synthesize predicted AT8-pTau channels. We then evaluated whether the computed AT8-pTau images would better guide discovery efforts and improve the screen. To our knowledge, this is the first prospective demonstration of reliably constructing new fluorescent images *in silico* from pre-existing fluorescent markers—providing a new way to improve the drug discovery process and leverage complex biological features otherwise latent within historical HCS archives. Moreover, we assessed the generalizability of this trans-channel method by applying it to an entirely unrelated biological context of a genome-wide functional screen in osteosarcoma cells.

1.3 RESULTS

1.3.1 We Constructed a Drug Screen Dataset of Tau Prion Propagation for ML Training

We sought to improve in vitro screening by constructing a new image channel from abundant, common, and relatively inexpensive channels—even after a screen's completion. The archival HCS data contained extensive information on many compounds, but it only contained YFP-tau and DAPI images and was missing the valuable but cumbersome AT8-pTau channel.

Therefore, our first task was to generate a completely independent dataset to train a ML model to predict the AT8-pTau channel solely from the YFP-tau and nuclear DAPI channels.

We constructed the new training dataset using a fixed cell system and collected all three channels (YFP-tau, DAPI, and AT8-pTau) across six 384-well plates representing a range of conditions. Ideally, ML models leverage large and diverse datasets during their training. Hence, we collected varied image data so that the model could learn to capture information about different conditions and perturbation scenarios that may be found in the HCS dataset, which itself has many different compounds present. In the training dataset, we curated a range of small molecule perturbations, each resulting in different tau aggregation and cell viability phenotypes. We used six different compounds to capture three different phenotypic spaces: reducing intracellular tau aggregation without affecting cell viability; reducing tau aggregation and reducing cell viability; and increasing tau aggregation and reducing cell viability. See Methods for a detailed description of the drug perturbation scenarios.

We constructed the three-channel training dataset to model infectious tau propagation similar to the archival HCS dataset (Methods). HEK293T cells were transfected to overexpress a YFP-tau mutant containing a familial mutation (P301S) which increases the propensity of tau to form prions that aggregate^{27,39}. We then seeded the cells with an infectious mouse brain lysate passaged once through HEK293T cells expressing the microtubule-binding repeat domain of human tau with the P301L and V337M mutations. We immediately treated the cells with small molecule compounds, fixed the cells, and initially introduced the monoclonal AT8 antibody to

measure phosphorylation events. We imaged three channels—DAPI, YFP-tau, and AT8-pTau (Figure 1.1)—to obtain 57,600 non-overlapping images that were 3 x 2048 x 2048 pixels in size. The archival HCS dataset was similar, differing predominantly in its use of live-cell imaging without AT8, and also containing thousands of compounds. Otherwise, the protocols were identical (Methods).

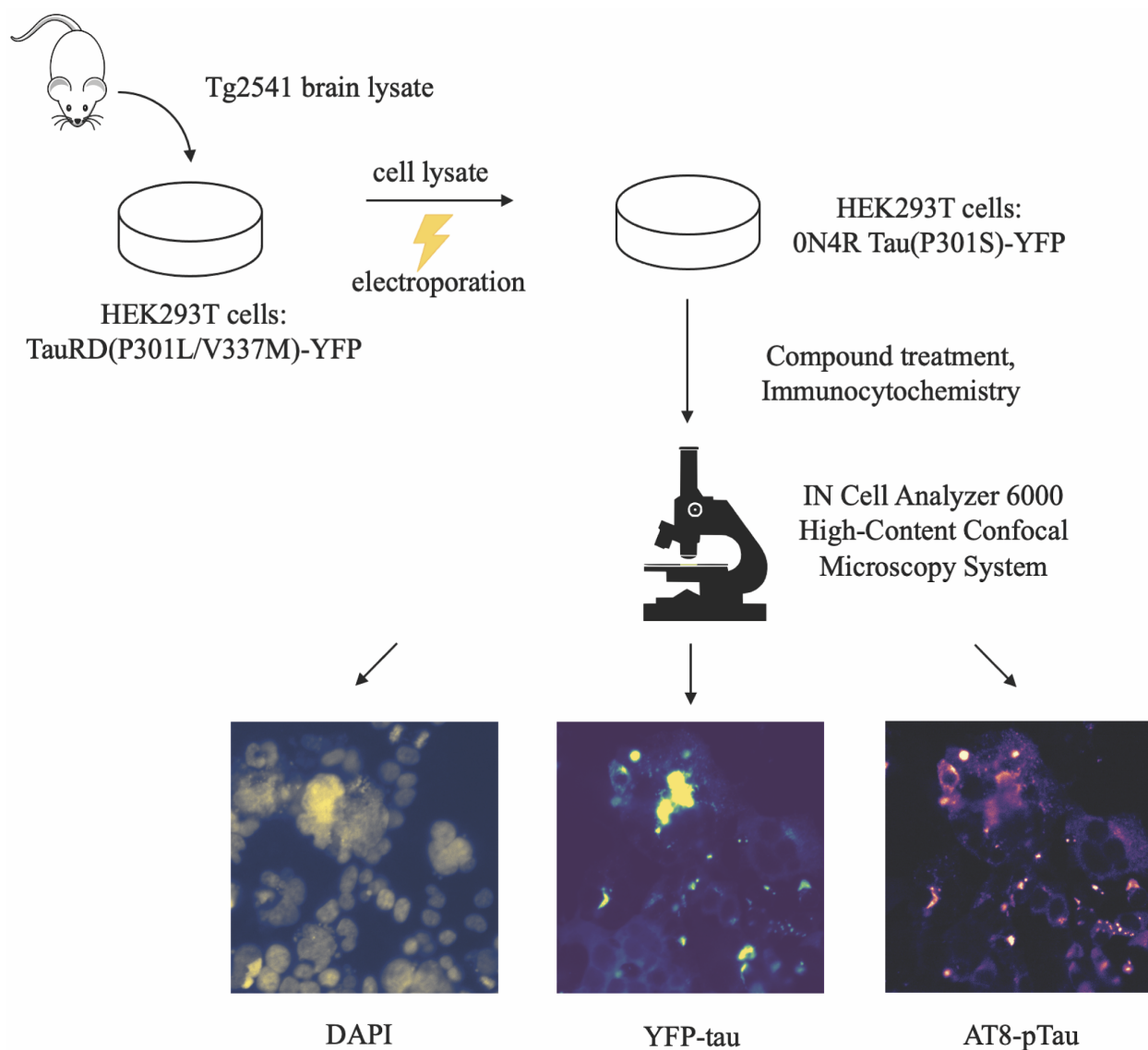


Figure 1.1: We constructed a cellular Tau(P301S)-YFP aggregation assay that modeled prion propagation *in vitro* with an additional AT8-pTau channel for ML training.

An overview of the experiment used to model the adverse propagation and aggregation of tau. We used brain lysate from a tauopathy mouse model to infect a human cellular model of mutant tau that has a higher propensity to aggregate. The experiment yielded three channels used for training the ML model: DAPI, YFP-tau, and AT8-pTau. Images were enhanced with ImageJ's auto-enhance feature solely for visualization purposes.

1.3.2 We Trained a Trans-Channel ML Model to Predict Hidden Tau Phosphorylation

To learn a mapping from the input DAPI and YFP-tau channels to the output AT8-pTau channel, we developed a custom convolutional neural network (CNN) model motivated by the U-Net architecture first proposed by Ronneberger³⁷. Our design employed “skip connections”³⁷ that preserved lower complexity features from earlier layers and combined them with higher complexity features from deeper layers. We concatenated these layers and took a convolution operation over this combination (Figure 1.2; Methods). We made our PyTorch code and fully trained models openly available (Methods). We randomly stratified the three-channel images into a 70% train and 30% held-out test split (Methods). On visual inspection, the ML model removed aberrant signal and accurately mapped the phenotypes of the related channels to construct realistic AT8-pTau images with a high resemblance to the actual objectives (Figure 1.2).

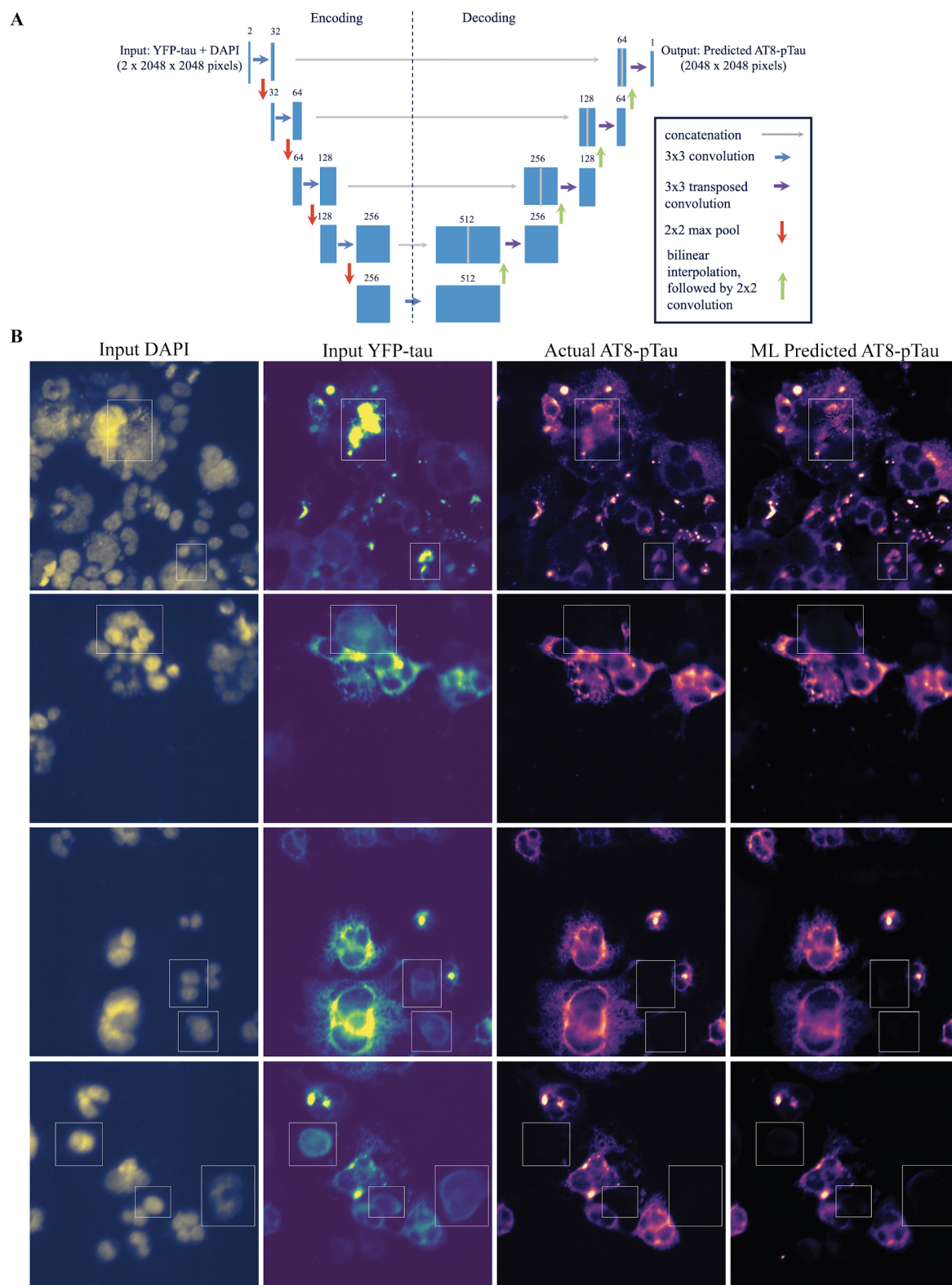


Figure 1.2: The ML model generated a phosphorylated AT8-pTau channel, solely from DAPI and YFP-tau channels.

(a) Given two image channels (DAPI and YFP-tau) as input, the neural network predicted the desired AT8-pTau image. The architecture, motivated by U-Net³⁷, comprised two phases: an “encoding” half which increased the cardinality of convolutional filters while reducing the image dimensions, and a “decoding” half which decreased the cardinality of filters while upsampling the image dimensions with bilinear interpolation and learnable transposed convolutions. The numbers above each box indicate the number of filters used. (b) From the input DAPI image (left column) and the input YFP-tau image (second column), the *in silico* predicted AT8-pTau image (rightmost column) is shown versus the actual, hidden AT8-pTau image (third column). White boxes mark where the AT8-pTau images most diverged from YFP-tau on visual inspection. These held-out test images were not seen during model training. For visualization purposes only, images were cropped to 512 x 512 pixels, auto-enhanced using ImageJ, and colored using the python library Matplotlib. Pearson correlation values between the predicted and actual images displayed are 0.88, 0.92, 0.93, and 0.86 from the top row to the bottom.

The ML model learned to identify and enhance regions of interest in the YFP-tau channel, namely tau phosphorylated at the Ser202/Thr205 epitope. The model learned to effectively generate the AT8-pTau images by gleaning visual information about the relationships between a hidden phosphorylation readout and the YFP and DAPI channels. It removed YFP-tau signal that was not present in the AT8-pTau image and selectively retained pertinent signal (interesting regions are boxed in Figure 1.2). On visual inspection, there did not appear to be an intuitive set of visual criteria or easily predefined attributes like pixel intensity by which a human could accomplish the same task (e.g, see Supplemental Figure 1.1). The model’s decision-making appeared to be non-trivial.

We evaluated the model quantitatively on the held-out test set of YFP-tau and DAPI inputs and AT8-pTau objectives. To assess image similarity between the observed and predicted AT8-pTau images, we calculated the Pearson correlation and also the mean squared error (MSE)

metric, which is more commonly used in regression tasks. We favored Pearson because it is invariant to pixel intensity scaling: a Pearson correlation coefficient of 1.0 indicates exactly identical images (with a possible constant scaling factor), whereas 0.0 indicates a complete pixel-wise disagreement. We obtained an average Pearson correlation of 0.74 ± 0.19 over the test set of $n=17,280$ images (Figure 1.3). This corresponded to a MSE of 0.52 ± 0.39 after normalizing to zero-mean and unit-variance (Figure 1.3). To quantify its pixel-wise performance, we measured how well the model balances true positive pTau pixels with false positives, assessing the error rate using the area under the curve of the receiver operating characteristic (AUROC). The model achieved an $\text{AUROC} = 0.96$ (Figure 1.3). Importantly, we found that learning from either single channel alone was insufficient to glean the AT8 signal (Supplemental Figure 1.2).

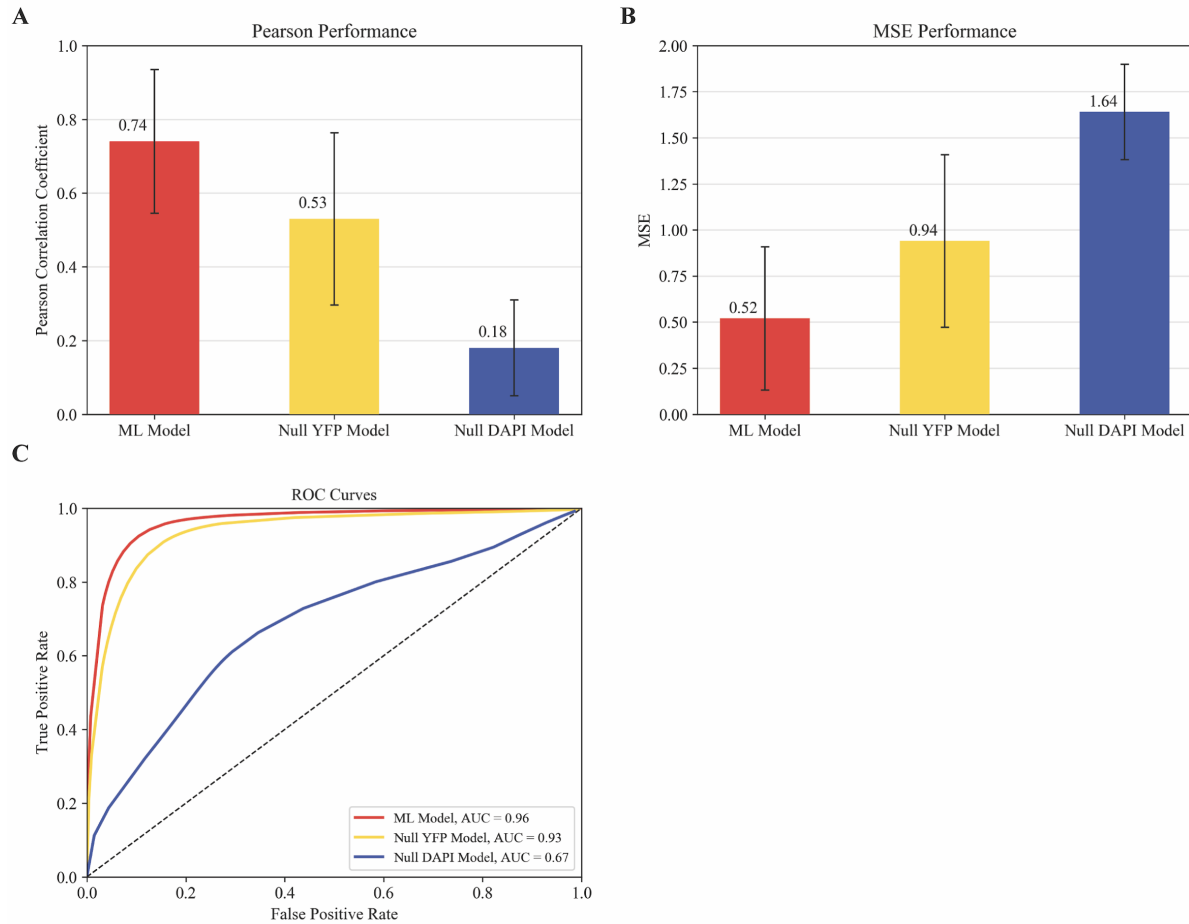


Figure 1.3: The ML model achieved good performance.

(a) The average Pearson performance over the held-out test set for three models: the ML model, versus two null models derived from each of the input channels respectively. Error bars show one standard deviation in each direction centered at the mean. Assessing the ML model's superior performance over each of the null models resulted in statistical significance of $p < 0.00001$. (b) Equivalent MSE results after normalizing all images to zero-mean and unit-variance (Methods). (c) The receiver operating characteristic (ROC) curves for the ML model (red) exceed the Null YFP Model (yellow) and the Null DAPI Model (blue). A binary pixel threshold of 1.0 was used to binarize the label image, while the predicted images were assessed across a range of pixel thresholds (Methods).

The YFP-tau and AT8-pTau channels contained substantial overlapping information. If they did not, the prediction task would be difficult or impossible. In theory, signal from the YFP-

tau is a superset of the AT8-pTau signal and includes all of the introduced tau and aggregated protein (Supplemental Figure 1.3). Thus, given the similarity between the YFP-tau channel and AT8-pTau channel, we might theoretically achieve good “prediction” by simply extracting the input YFP-tau *as-is* and stipulating this as the model’s predicted output. This trivial Null YFP Model required no ML, and yielded a Pearson correlation of 0.53 ± 0.23 (corresponding to an MSE of 0.94 ± 0.47) over the entire dataset (Figure 1.3). The ML model exceeded this baseline with an increase in average Pearson correlation of 0.21 ($p < 0.00001$; corresponding to a decrease in average MSE of 0.42), which is consistent with the model learning to generate an output that more closely approximated the phosphorylation state than was already provided by the YFP-tau input alone. We also created a Null DAPI Model that simply returned the input DAPI image as the output; this model obtained an average Pearson correlation of 0.18 ± 0.13 (corresponding to a MSE of 1.64 ± 0.26). The ML model’s test performance was consistent across all six drug perturbations (Supplemental Figure 1.4).

As with most multichannel imaging studies, we were wary of image bleed-through that the model might exploit as a hidden crutch. Excitation and emission plots of the different fluorophores are shown in Supplemental Figure 1.5. To test the hypothesis that the ML model leveraged undetectable but pernicious hidden AT8 signal within the input channels, we performed pixel-intensity ablations on the input images to eliminate potential low-intensity bleed-through signal and then assessed performance (Supplemental Figure 1.6). We did not detect significant image bleed-through as a confounder. Accordingly, we note that any

confounding signal augmenting a model's performance would be of no help when applied subsequently to the archival HCS dataset, since the archival HCS did not undergo immunohistochemistry preparation and had neither AT8 antibody nor fluorophore.

1.3.3 Trans-Channel ML Improved Hit Rate and Compound Triaging in Tauopathy HCS

Turning to its prospective use in drug discovery, we applied the trans-channel ML model to the archival HCS dataset and found practical improvements. For most HCS pipelines, immunostaining is not normally employed due to batch variability and increased labor and cost, and must be balanced against the advantages of live-cell formats such as time-course data collection. If ML could inexpensively infer these immunohistochemistry channels, then it could enhance HCS efforts by providing previously unavailable channels and inferences—thereby improving drug discovery pipelines. Similarly, one might imagine a wealth of data could be mined in a hypothesis-guided way from substantial existing datasets of completed screens in the public and private sectors, thus advancing biological knowledge and improving medical therapies.

Compounds are conventionally ranked by priority, favoring those that lower tau aggregation for secondary dose-response testing. Medicinal chemists also consider chemical structure when prioritizing. Instead, we took a naive approach by having the ML model make decisions based solely on cellular phenotype—unguided by human intuition.

To assess the ML method for practical use, we directly compared it to the conventional method of drug discovery. Using data from the archival HCS dataset that was devoid of immunohistochemistry procedures, we predicted an AT8-pTau channel for each compound using the trained model. Due to the large dataset size, we randomly selected one run of the HCS (consisting of 1,600 unique compounds) for our analysis. We constructed a ML-derived priority queue (PQML), ranking compounds based on aggregation scores using the predicted, machine-learned AT8-pTau images. We calculated aggregation scores using cellular image software provided by General Electric (Methods), as we did in the original archival HCS evaluation pipeline a year before. The conventional method's priority queue (PQC) ranked each compound solely by the aggregation scores of pre-existing YFP-tau images. Hence each compound received rankings in both queues and these rankings could disagree substantially.

We prospectively collected complete dose-response profiles for the top 40 compounds from each queue. We chose 40 compounds (per Methods) for each queue due to cost and labor limitations. Despite operating solely on a computationally-generated channel, the ML queue's hit rate was comparable with the conventional method and effectively proposed overlooked compounds that passed dose-response testing. Of the 40 compounds tested from PQML, 11 passed secondary testing (27.5%). Of the 40 compounds tested from PQC, 12 were active (30%). Importantly, compounds that were highly prioritized by both the ML and conventional methods obtained a much higher hit rate than either method alone. Ten compounds ranked within the top 40 of both queues and 6 out of these 10 were confirmed by dose-response. Taken together,

progressing compounds by this combined criterion yielded a success rate of 60%—an improvement of 30% over the conventional method and of 32.5% over the ML method (see Table 1.1). However, our sample size was smaller for these overlap compounds, which must be considered when assessing generalizability.

Table 1.1: Comparison of dose-response success rates using different methods.

The Random Baseline is approximated from a previous screen, in which 26 active compounds were discovered in a screen consisting of 60,851 compounds.

Method	Dose-Response Success Rate
Random Baseline	0.043% (26 / 60,851)
Prospectively test top 40 compounds in PQC	30% (12/40)
Prospectively test top 40 compounds in PQML	27.5% (11/40)
Test compounds that are shared by both PQC's top 40 and PQML's top 40	60% (6/10)

Interestingly, 5 of the 11 active compounds in PQML's top 40 were effectively missed in the PQC, which ranked them 539th, 545th, 560th, 582nd, and 1,596th out of 1,600 possible compound ranks. The activity of one such compound is shown in Figure 1.4. The ML model eliminated YFP-tau aggregate signal that had falsely lowered these compounds' priority in the PQC (Figure 1.4).

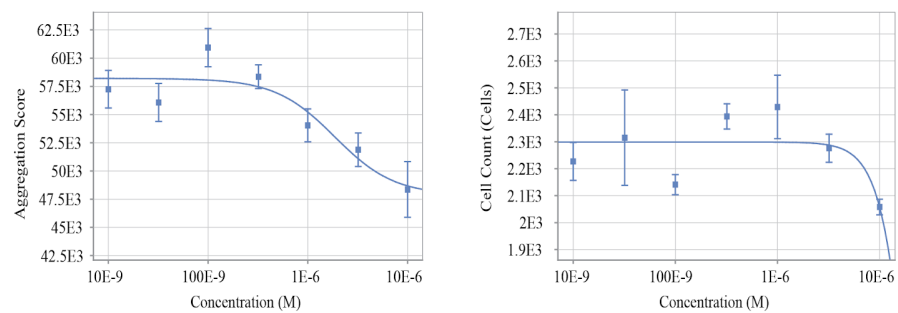
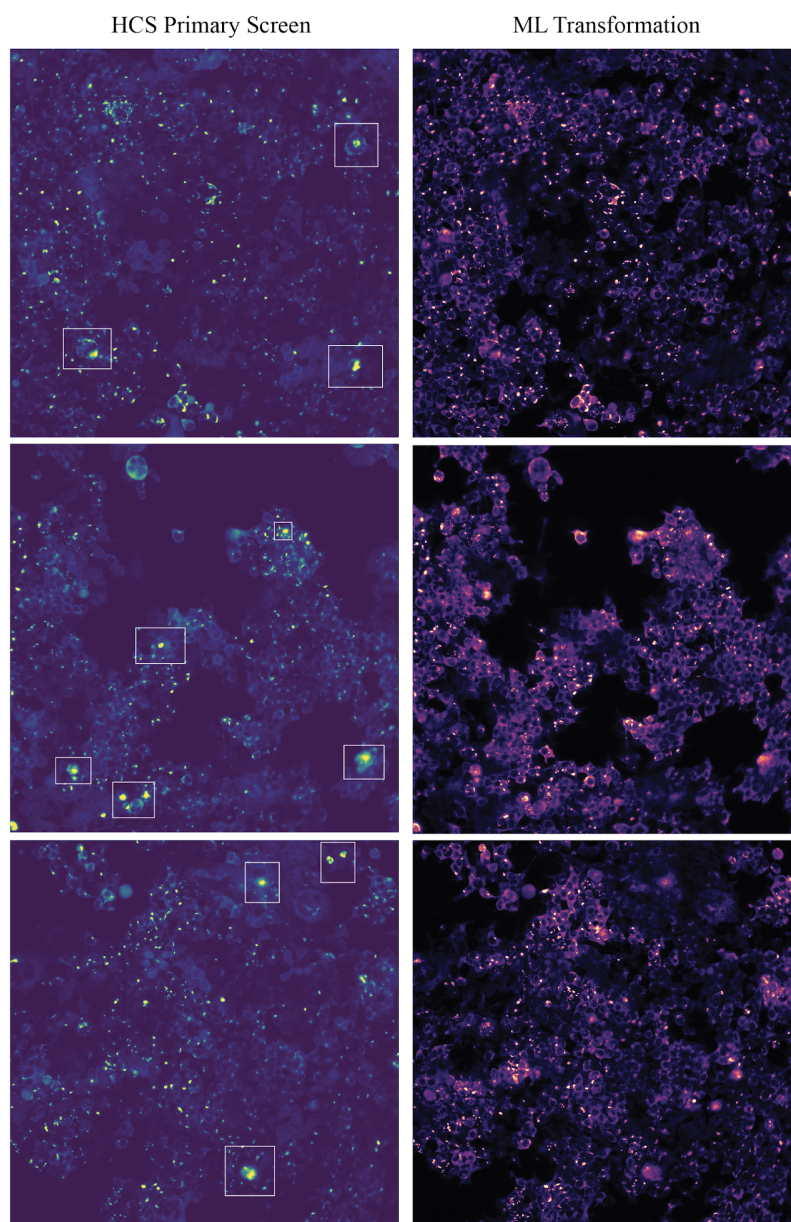
A**B**

Figure 1.4: Transforming the archival HCS data with ML-predicted AT8-pTau images revealed previously unknown active compounds.

(a) We rescored a plate consisting of 1,600 compounds from the archival HCS, and prospectively collected dose-responses for compounds that resulted in the lowest tau aggregation according to either PQML or PQC. Example dose-response (left) and cell count assay (right) of an active compound. Although an indicator of cell viability, cell count is not a perfect metric for toxicity, since it is possible for the compounds to inhibit the innately prolific nature of HEK293T cells. These curves are drawn from the compound DRW545, which was ranked well by the ML method (22nd), but poorly (545th) by the conventional method. (b) Example images for active compounds that the ML-based rescoring rescued. Top row: Compound DRW1596 (ranked 1596th in PQC versus 14th in PQML), Middle: compound DRW560 (ranked 560th in PQC versus 15th in PQML), Bottom: compound DRW545 (ranked 545th in PQC versus 22nd in PQML). YFP-tau images from the archival HCS dataset (left column) differed from the ML-predicted AT8-pTau images used for rescoring the HCS (right column), with white boxes highlighting example regions. In these regions of interest, the ML predicted the aggregate signal as not being phosphorylated at the residue of interest. We auto-enhanced and colored images with ImageJ and Matplotlib solely for visualization.

Additionally, the ML method achieved higher enrichment⁴⁰ for active compounds. We discovered a total of 17 unique dose-response confirmed compounds from both lists. For these compounds, PQML had an area under the enrichment curve of 0.93 versus PQC's area under the enrichment curve of 0.85 (Figure 1.5). These active compounds achieved an average rank of 119 in PQML, which was better than nearly half their average rank of 235 in PQC (Figure 1.5). Therefore, fewer compounds from PQML would need to be tested to find the same number of active compounds from PQC.

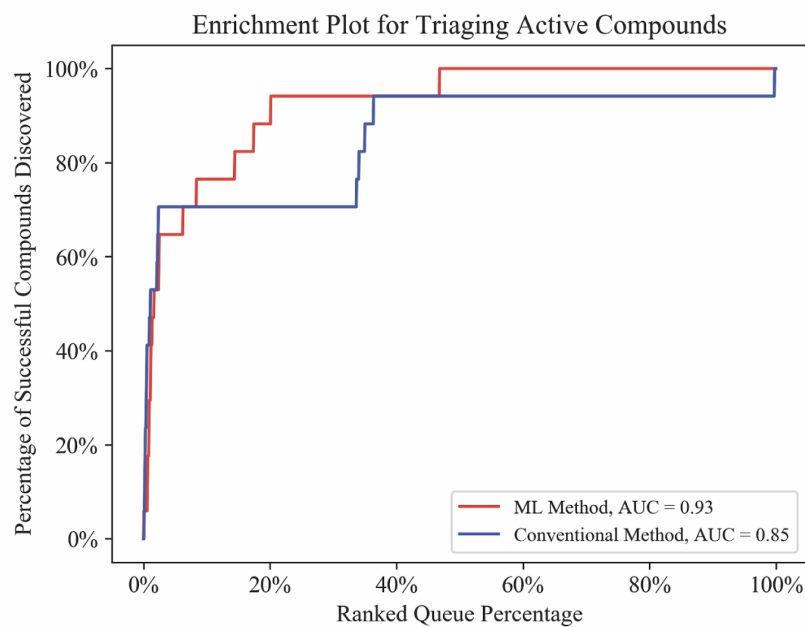
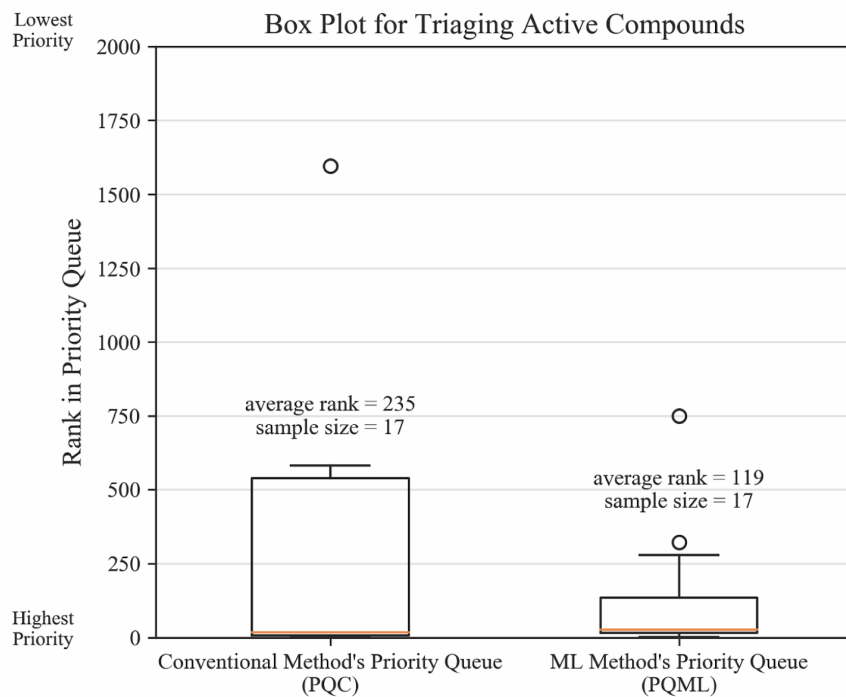
A**B**

Figure 1.5: ML triaged active compounds better than the conventional method.

(a) Enrichment plots comparing the conventional compound testing method with prospective, ML-guided dose-response assays. We considered the set of all active compounds ($n=17$). The y-axis shows what fraction of active compounds were discovered in the top $x\%$ of the ranked priority queue (x-axis) (100% corresponds to rank 1,600). The ML method's larger area under the curve (AUC) indicates that its compound ranking achieves a higher success rate on prospective dose-response testing. (b) Box plots of rank accounting for all active compounds. The median line is in orange, whiskers are set at 1.5x interquartile range, and outliers are anything outside the 1.5x interquartile range (plotted as dots). PQML on average prioritizes active compounds about twice as well as PQC.

Of the 17 total active compounds from both lists, 5 fell outside of PQC's top 40, and 6 fell outside of PQML's top 40. However, to recover the active compounds missed in each case, we would need to test more than twice as many candidate compounds on average ($n=764$) by the PQC rankings as would be necessary by the PQML rankings ($n=302$). Hence, the ML method ranked missed-but-active compounds more than twice as well as the conventional method—including them in a smaller, more tractable search space (Supplemental Figure 1.7). Furthermore, when we derived a compound's dose-response activity from the ML-derived AT8-pTau images instead of scoring all queues solely by the conventional YFP-tau images, the ML method exceeded the conventional method with larger margins (Supplemental Figure 1.7).

1.3.4 We Tested Trans-Channel ML Method on an HCS with Different Biological Conditions

To assess whether the success of our trans-channel ML approach was specific merely to the tauopathy compound HCS, we investigated whether the method of learning fluorescent signal from related markers applied to a HCS unrelated in its biology and perturbation. We performed

the same trans-channel fluorescence learning task. However, compared to the prion tauopathy screen, this cancer-based HCS used an entirely different cell line, microscope, perturbation method, and fluorescent markers. ML models can exploit dataset-specific patterns in subtle ways; thus, our goal was to assess the approach's generalizability by applying it where many of the data parameters differed from the tauopathy screen.

We performed a functional genomics HCS in a cancer cell model. In this fluorescence-based arrayed whole-genome screen (details in Methods), we plated an osteosarcoma cell line (U2OS) with a perturbation scheme that systematically knocked down all coding genes with siRNA. We marked cellular DNA with Hoechst fluorescent dye and tracked the cyclin-B1 protein with a green fluorescent protein (GFP) fusion. These two markers are biologically related, with cyclin-B1 specific to the G2/M phase of mitosis and also involved in DNA damage repair⁴¹. In theory, the two markers contain shared information, so we investigated whether the ML model could learn this signal relationship like it did in the tauopathy dataset. We collected a total of 324,989 images.

Unlike the tauopathy dataset, we observed subtle bleed-through signal in the raw Hoechst channel (Figure 1.6; see Supplemental Figure 1.5 for excitation and emission spectra). Hence, as a preprocessing step to mitigate bleed-through, we ablated all pixels below the 95th percentile in our Hoechst images, which inevitably removed some Hoechst signal as well (Figure 1.6). We performed threefold cross-validation to predict cyclin-B1 signal solely from the ablated Hoechst channel. These models achieved an average pixel-wise Pearson correlation for channel prediction

of 0.75 ± 0.13 (corresponding to an MSE of 0.50 ± 0.26) on $n=108,330$ test images (Figure 1.6). This was 87.5% ($p < 0.00001$) higher than the Null Model that used the ablated input as its prediction (Pearson correlation of 0.40 ± 0.14 ; MSE of 1.21 ± 0.28). Despite significant ablations to the models' input, the models were able to learn the non-trivial cyclin-B1 phenotype. Upon visual inspection, predicting where high-intensity cyclin-B1 signal resided solely from the Hoechst signal did not appear obvious. As an unexpected benefit, the ablation procedure resulted in the model predicting images that were mostly free of signal artifacts in the cyclin-B1 channel (Figure 1.6).

When we omitted the ablation procedure and trained a threefold cross-validation on raw images, the ML model accurately predicted cyclin-B1 signal and produced detailed and realistic images that matched the observed images to which it had been blinded (Figure 1.6). This resulted in an average Pearson correlation of 0.85 ± 0.08 (corresponding to an MSE of 0.30 ± 0.15) on $n=108,330$ test images (Figure 1.6). This was 93.2% higher ($p < 0.00001$) than the Null Model that stipulated the raw Hoechst image as its prediction (Pearson correlation of 0.44 ± 0.12 ; MSE of 1.11 ± 0.23). Training on raw, unablated images resulted in a 13.3% ($p < 0.00001$) increase in average Pearson performance versus the ablated training procedure.

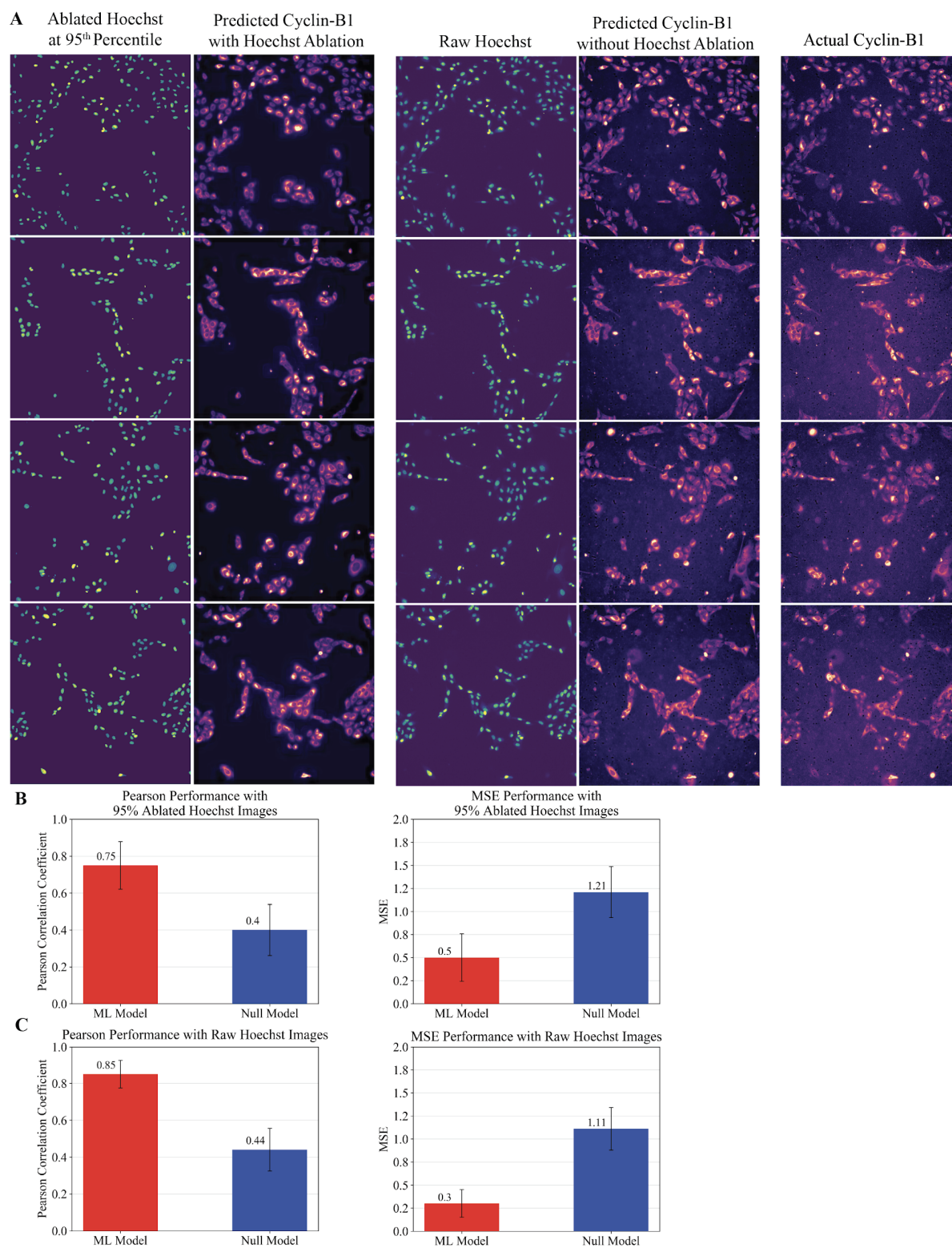


Figure 1.6: The trans-fluorescence learning method generated accurate cyclin-B1 signal from Hoechst signal, on an independent and biologically unrelated dataset.

(a) Given ablated Hoechst images (leftmost column) a ML model trained to output cyclin-B1 images (rightmost column) achieved high *in silico* prediction of this channel (second column). Given unmodified Hoechst stain images (third column), a separate ML model constructed high-fidelity and detailed predictions (fourth column) of the actual cyclin-B1 channel. This panel contains images from the test set ($n=108,330$ images) never seen during model training. For visualization, all images were auto-enhanced and colored as in prior Figures. (b) Assessing the Hoechst-ablated model's performance. Left: average Pearson correlation coefficient when assessing pixel-wise similarity between actual and predicted Cyclin-B1 images in threefold cross-validation of a dataset of $n=324,989$ total image pairs, across $g=16,194$ unique functional genomic perturbations; Right: MSE from the same analysis. Assessing the ML Model's superior average performance over the Null Model resulted in statistical significance of $p < 0.00001$. (c) Assessing the model trained using raw, unablated Hoechst images, performing the same analysis as in (b). The ML Model's superior average performance over the Null Model likewise resulted in $p < 0.00001$. Error bars show one standard deviation in each direction centered at the mean.

1.4 DISCUSSION

We developed and assessed a method to computationally augment fluorescence microscopy and HCS. This method could be especially useful for tapping into otherwise hidden information in large and information-rich archival datasets. Three observations merit emphasis: (1) trans-fluorescence models generated biologically informative images that were drop-in replacements for existing HCS workflows; (2) the model operated on an archival HCS dataset that was independent from the newly-created training dataset and facilitated discovery of new compounds; (3) the method was generalizable to a different biological environment (a functional genomics screen in an osteosarcoma line).

We readily integrated the ML trans-channel predictions into a conventional compound evaluation workflow because all of the *in silico* channel images were generated as drop-in replacements for the equivalent *in vitro* experiment. Generating a predicted AT8-pTau image from existing YFP-tau and DAPI fusion images required only fractions of a second. With the proper hardware or cloud resources, this method scales to large, high-volume HCS datasets and is parallelizable. Its ability to accurately predict trans-channel fluorescence can potentially serve as a biologically informed and economically practical means of hypothesis generation, especially if the desired marker in question has barriers to widespread utilization. Furthermore, accurate *in silico* labeling decreases variability from possible experimental labeling artifacts.

As measured by dose-response confirmed compounds, the ML method performed similarly to the conventional method in the drug discovery task. We were intrigued, however, that the ML method found a comparable number of active compounds, many of which were missed in the original screening campaign. A strong comparison could not be made between the approaches and their propensity to find active compounds at the small scale of 17 unique active compounds, when the methods' hit rates differed by one hit. Rather, the demonstrated utility of the ML method was to triage compounds more efficiently and to rescue efficacious compounds that were missed due to poor rank in the primary screen. Encapsulating information on AT8-marked phosphorylation and projecting it onto the HCS dataset facilitated compound ranking and subsequent validation.

We acquired the three-channel fixed-cell HCS training dataset using a slightly different protocol than the archival live-cell HCS dataset. Separated by differences in protocol, cell format, experimenters, and a year, the ML model relied solely on extracting hidden signal of phenotypic phosphorylation from YFP-tau images—with no AT8 labeling in the HCS, and thus no possibility of fluorescent bleed-through. This implies that the model has value outside of the strict data regime on which it was trained. This is not always the case in ML applications, which often do not generalize⁴². The archival HCS test dataset consisted completely of compounds that were not part of the training dataset, yet the model was able to operate on new images and a new and larger compound space of greater than a thousand compounds to facilitate discovery.

By evaluating the method on a biologically unrelated osteosarcoma dataset, we showed that the method of trans-channel fluorescent learning generalized across disparate data domains. Intended only as a test in a different disease and biological domain, we did not seek to address nuanced biological questions related to accurately constructing cyclin-B1 signal solely from Hoechst signal. We hope that researchers can tailor the method to their own use cases in order to augment their phenotypic screening endeavors and expand to use cases we have not envisioned. Furthermore, the architecture is fully convolutional, invariant to image size, and adaptable to any input shape. Hence, we expect this method to aid development efforts across a wide range of image applications.

Several caveats limit and focus the scope of this study. The cell line for the tau experiment was not neuronal, and HEK cells were chosen for practicality—HEK cell lines are

frequently used in drug screens since they are easy to grow and transfect to express a protein of interest. Additionally, the method, as with most ML applications, was data intensive and required many image examples for learning. This could be a pragmatic hurdle, but we hope to leverage transfer learning⁴³ as openly shared cellular image data become increasingly common⁴⁴.

Accordingly, we have made all of the tauopathy training dataset and one full plate of the osteosarcoma functional genomics HCS dataset publicly available (Methods). Furthermore, the AT8-pTau channel images collected experimentally could be missing information on tau that is phosphorylated at other epitopes of interest. Thus, what we currently interpret as YFP off-target signal—i.e., the “false positive” pixels that show fluorescence in the YFP-tau but *not* in the AT8-pTau channels—could represent disease-relevant phosphorylation at sites other than the Ser202/Thr205 epitope. A future exercise would be to use antibodies for other commonly hyperphosphorylated epitopes. Consequently, we did not seek with this study to implicate tau hyperphosphorylation as a disease-inducing mechanism, but rather used the AT8 marker as an approximate surrogate for a relevant disease phenotype.

Trans-channel learning is a tool for hypothesis-guided biological discovery. When a new biologically-informative channel can be reliably predicted on an archival dataset, it decodes actionable high-content signal that can guide compound prioritization and rescue missed opportunities hidden in completed HCSs for drug discovery. Conversely, when employed instead as a means to *falsify* hypotheses about the relatedness of biological processes, the failure of a best-effort trans-channel approach to learn one channel when it succeeds in learning other trans-

channel relationships may be informative when assessing the biological and visual independence of a pair of biological markers (Supplemental Figure 1.2). We hope that the techniques and datasets we describe and make available here, which may be attempted on any archived high-content screen, will be of broad use to the microscopy, screening, and drug discovery communities.

1.5 METHODS

1.5.1 Tg2541 Mouse Line

The Tg2541 transgenic mouse line expressed the 0N4R isoform of human tau, under control of the murine neuron-specific Thy1.2 genetic promoter⁴⁵. Tg2541 mice were originally generated on a mixed C57BL/6J $\times$ CBA/Ca background⁴⁵ and were then bred onto a C57BL/6J background using marker-assisted backcrossing for eight generations before intercrossing to generate Tg2541 homozygous congenic mice. These mice were kept at room temperature. Homozygous Tg2541 mice on a congenic C57BL/6J background were kindly provided by Dr. Michel Goedert (Medical Research Council, Cambridge, UK). The mice were maintained in a facility accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International in accordance with the Guide for the Care and Use of Laboratory Animals. All procedures for animal use were approved by the University of California, San Francisco's Institutional Animal Care and Use Committee.

Phosphotungstic acid precipitation of tau protein in Tg2541 mouse brain samples

Pooled brain homogenate from 6–7 month old Tg2541 mice (both male and female) was prepared as previously described²⁷. Briefly, a 10% (weight/volume) homogenate in DPBS was prepared using a rotor-stator type tissue homogenizer (Omni International). Phosphotungstic acid (PTA) precipitation of the brain homogenate was then performed as previously described^{14,46}. Briefly, 10% brain homogenate was incubated with 2% sarkosyl (Sigma Aldrich) and 0.5%

benzonase (Sigma Aldrich) at 37°C with constant agitation at 1,200 rotations per minute for 2 hours on an orbital shaker. Sodium PTA was dissolved in water and the pH was adjusted to 7.0. A final concentration of 2% sodium PTA was added to the brain homogenate and incubated overnight under the same agitation conditions. The brain homogenate was then centrifuged at $16,100 \times g$ for 30 minutes at room temperature, and the supernatant was removed. The pellet was resuspended in 2% sarkosyl and 2% sodium PTA in DPBS and incubated for 1 hour at room temperature. The sample was centrifuged again under the same conditions, the supernatant was removed, and the pellet was resuspended in DPBS, using 10% of the initial starting volume, and stored at -80°C until further use.

1.5.2 Experimental Setup for the Cellular Tau(P301S)-YFP Aggregation Assay Training Dataset

Tau(P301S)-YFP cells were developed by transfecting human embryonic kidney cells (HEK293T female; ATCC) using electroporation to overexpress the full-length 0N4R isoform of human tau containing the familial disease-linked P301S missense mutation and the yellow fluorescent protein (YFP) fused to the C-terminus for visualization. A stable monoclonal line was maintained in Dulbecco's modified enriched medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. Tau(P301S)-YFP cells in confluent flasks were collected using trypsin and resuspended in electroporation buffer at a concentration of 8.4×10^7 cells/mL. HEK293T cells that overexpress the microtubule-binding repeat domain (RD) of 4R human tau with the P301L and V337M mutations¹⁵ [TauRD(P301L/V337M)-YFP

cells] were maintained in DMEM supplemented with 10% FBS and 1% penicillin/streptomycin. Tau seeds were prepared by passaging PTA-precipitated tau protein from Tg2541 mouse brain samples once through TauRD(P301L/V337M)-YFP cells and then collecting the cell lysate, as described previously²⁷. All cells were cultured at 37°C and 5% CO₂ in a humidified incubator until analysis.

Cell lysate containing tau seeds was combined with the Tau(P301S)-YFP cells at 1.8 µg/mL and electroporation was performed on an Amaxa Nucleofactor instrument (Lonza Bioscience) using the program #Q-001. Seeded Tau(P301S)-YFP cells were then plated at 5 x 10³ cells/well in a 384-well plate (Greiner) coated with poly-L-ornithine (PLO, 30-70kDa; Sigma Aldrich). To coat plates, PLO was dissolved at 1 mg/mL in 0.15 M sodium borate buffer pH 8.4 and filter-sterilized, and then further diluted to 0.1 mg/ml in sterile deionized (DI) water. The PLO coating solution was added to the culture plate wells and incubated in culture plate wells for six hours at room temperature, then removed. The wells were washed three times with sterile DI water and air-dried in a biosafety cabinet.

Immediately following cell plating, small-molecule drug compounds prepared in pure DMSO were added at a fixed concentration of 10µM. The small molecules were selected to represent the range of effects on tau aggregation and cell viability by compounds in drug libraries of interest, developed by Daiichi Sankyo (Tokyo, Japan). Actual compound identifiers are proprietary and are obfuscated to identifiers in the form “DRW#”. Two compounds were selected for reducing tau aggregation and also reducing cell viability (DRW1 and DRW2); and

two compounds were selected for increasing tau aggregation and reducing cell viability (DRW3 and DRW4); two compounds were selected for reducing intracellular tau aggregation without affecting cell viability (DRW5 and DRW6). Control wells were also prepared with cells electroporated with or without tau seeds and the same volume of DMSO but no drug compound.

At four days post-seeding, cells were washed three times with sterile Dulbecco's Phosphate (DPBS) and then fixed with 4% paraformaldehyde (PFA) in DPBS for 20 minutes at room temperature. Cells were washed with DBPS and then permeabilized with 0.1% Triton X-100 in DPBS for 20 minutes at room temperature. Cells were washed and then blocked with 3% bovine serum albumin (BSA; Millipore) in DPBS for one hour at room temperature. Cells were then incubated with mouse monoclonal antibody AT8 (1:500; Thermo Fisher Scientific) and rabbit polyclonal antibody MCM2 (1:500; Abcam) in 3% BSA in DBPS overnight at 4°C. Cells were washed with DBPS and then incubated with Alexa Fluor 594 Plus-conjugated goat anti-mouse and anti-rabbit secondary antibodies (1:500; Thermo Fisher Scientific) in 3% BSA in DBPS for 90 min at room temperature. Cells were washed with DBPS and then covered with DPBS and imaged on an InCell Analyzer 6000 High-Content Confocal Microscopy System (GE Healthcare). Twenty five non-overlapping, 2048 x 2048 pixel fields were captured per well. We chose markers with minimal overlap of the excitation spectra to minimize potential image bleed-through (Supplemental Figure 1.5). This full training dataset was uploaded here for open access (DOI: 10.17605/OSF.IO/XNTD6).

1.5.3 Constructing the Archival HCS Dataset

The experimental setup for the archival HCS data was the same setup used to construct the training set (the cellular Tau(P301S)-YFP aggregation assay), with the exceptions that the HCS data did not undergo any of the immunocytochemistry steps, and live-cell imaging was used instead. We constructed this HCS dataset more than one year prior to the three-channel training dataset. Only two imaging channels were captured, a nuclear DAPI channel and a YFP-fused tau channel. No AT8 antibody was plated. Thus no AT8-pTau channel was captured, and there was no possibility of bleed-through. Furthermore, the compound library in the HCS data was much larger and more diverse than that used for model training. The original six compounds used in the training set with AT8 were not included in the random HCS subset of 1,600 unique compounds evaluated by the ML model.

1.5.4 Training and Evaluation of the ML Models from the Tauopathy Dataset

The full source code and fully trained models are available at <https://github.com/keiserlab/trans-channel-paper>. The necessary python packages are PyTorch, pandas, cv2, cupy, numpy, and sklearn. Training and testing of the model was performed on one Nvidia Geforce GTX 1080 graphics card, and completed in 30 epochs totaling 111 hours of training.

The full dataset consisted of six different drug perturbations: two reduced aggregation while maintaining cell count; two reduced aggregation while reducing cell count; and two

increased aggregation while reducing cell count. We imaged a total of six 384-well plates, one plate for each drug. Each well was divided into four fields and each field was imaged, resulting in a full image dataset consisting of 57,600 TIFF images each of size 2048 x 2048 pixels for each of the three channels. As a preprocessing step, we linearly scaled all images from the original TIFF range of 0 to 65,535 to the more tractable range of 0 to 255 by dividing each pixel by 65,535 and then multiplying by 255. We converted the resulting images to 32-bit floating point. The images were randomly shuffled, and then split into a 70% training set and 30% test set.

We used the fully convolutional architecture in Figure 1.2 for training. As inputs to the model, we concatenated the transformed YFP-tau image and the transformed DAPI image into a tensor of size 2 x 2048 x 2048 pixels. The model generated a predicted AT8-pTau image of size 2048 x 2048 pixels.

We used a stochastic gradient descent optimizer with momentum = 0.90, a learning rate = 0.001, and a batch size of one. We trained the model for 30 epochs. We used a negative Pearson correlation loss function for training, in which we minimized the following objective function:

$$\text{NPCC}(A, B) = \frac{-\sum_i (A_i - A_\mu)(B_i - B_\mu)}{\sqrt{\sum_i (A_i - A_\mu)^2 \sum_i (B_i - B_\mu)^2}}$$

such that A and B were the images being compared, i was summed over all pixels,

A_μ was the average of image A , and B_μ was the average of image B

1.5.5 Performance Metrics

When constructing the Pearson performance metric, we iterated over the test set, flattened all of the images to one dimensional vectors, and then found the Pearson correlation of the flattened label image with the flattened ML-predicted image using the Numpy library's `corrcoef` function.

When constructing the MSE metric, we normalized all AT8-pTau and predicted AT8-pTau images to have zero-mean and unit-variance in order to account for differences in underlying pixel value distributions between the label images and the predicted images. Normalizing both distributions placed them in a more comparable regime for calculating MSE.

When calculating MSE, we normalized each image by subtracting the average of that image, and then dividing by the standard deviation. The MSE was determined as follows:

$$\text{MSE}(A, B) = \frac{\sum_j (A_j - B_j)^2}{n}$$

Such that A_j and B_j were the normalized images being compared, and j was the index of the image out of a total of n test images.

To construct the ROC curve, we first normalized the images to have zero-mean and unit-variance. We then compared the predicted image to the actual, and chose a pixel threshold t such that any pixel greater than or equal to t was considered as positive for signal, and anything lower

than t was negative for signal. We chose an intensity threshold of 1.0 to binarize the image, which retained most of the features of interest and aggregation signal (Supplemental Figure 1.8).

Hence, we performed pixel wise classification, and derived an ROC curve for each image. The ROC curve in Figure 1.3 is the average ROC curve across the full test set of images. For the thresholds applied to the predicted image, we chose the pixel value range 0 (permissive) to 1000 (non-permissive) with different increments. In the range -1.0 to -0.1 and the range 0.1 to 1.0, we chose incrementing values of 0.1. From the range -0.1 to 0.1 we chose a more fine grained increment with a step size of 0.01. In the range 1.0 to 4.0 we chose a step size of 1.0 because of the infrequency of these values. Finally, we evaluated the pixel intensity threshold of 1,000, for which no pixels assumed this value.

1.5.6 Construction and Evaluation of PQML and PQC

From the available YFP-tau fusion images and DAPI images of the HCS subset of 1,600 compounds, we constructed predicted AT8-pTau images for each pair of YFP-tau and DAPI images. We took the YFP and DAPI images and linearly scaled the pixel values to the range 0 to 255 inclusive. We then concatenated the two images together to form a tensor of dimensions 2 x 2048 x 2048, and finally inputted this into the trained model. Next, we scaled the output of the model back to the original space by first dividing by 255 and then multiplying by the maximal possible pixel value of 65,535. We then scored these images for tau aggregation using proprietary software from General Electric, called the “InCell Analyzer,” which yielded a score based on puncta count and area. This same algorithm was used for previous HCS efforts, and it

was applied consistently to all images. We obtained each small molecule's aggregation score by averaging over all of that compound's images' aggregation scores—which consisted of averaging over four non-overlapping fields within a single well treated with the compound. We ranked all of the compounds by their aggregation score, with compounds inducing lower aggregation being higher priority than compounds with a higher aggregation score. A lower rank in the queue indicated higher priority (e.g. compounds with ranks first, second, and third were the three highest priority compounds). For constructing PQC, we used the same method as the one to construct PQML, except that we obtained aggregation scores from the original YFP-tau images instead of their corresponding ML-predicted AT8-pTau images.

To calculate enrichment curve AUCs (Figure 1.5), we analyzed the set of known active compounds ($n=17$) discovered in the study. For each queue, we generated an enrichment curve using the rankings of the active compounds. We calculated AUC by integrating with Numpy's composite trapezoidal function (`numpy.trapz`). We calculated average active compound ranks by averaging over each active compound's index in the queue (Figure 1.5).

1.5.7 Experimental Setup for the Secondary Dose-Response Experiments

For the dose-response experiments, we performed the same protocol as the cellular Tau(P301S)-YFP aggregation assay, with the exceptions being the drug doses and the compounds that we tested. We plated drug concentrations in half logs from 10nM to 10 μ M. We tested the top 40-ranked compounds of PQML and the top 40-ranked compounds of PQC. In accordance with the cellular Tau(P301S)-YFP aggregation assay, we did not use AT8 because it

would demand significant additional time and labor. For each compound, we independently replicated each concentration in four separate wells. We calculated the average aggregation scores at each concentration, and used these averages to construct a dose-response plot (e.g., Figure 1.4).

1.5.8 Assessing Activity from Dose Response Curves

A compound that succeeded in a secondary dose-response test demonstrated a dose-dependent decrease in aggregation as a function of concentration, while maintaining a therapeutic window that preceded a noticeable drop in cell count. A compound did not need to induce a perfect sigmoidal shape to be considered active. A compound was considered active if and only if all of the following four conditions were satisfied.

1. The best fit curve strictly had an aggregation score that decreased with increasing concentration of compound.
2. The effective response was at least 6,000 aggregation units. We required this so that only compounds were chosen that induced an effective decrease in aggregation. Medicinal chemists who worked closely on the screen chose 6,000 subjectively as a minimal threshold before we began the analyses. If the curve was monotonically increasing with increasing concentration, this condition was automatically not satisfied, and the compound was considered inactive.
3. The half maximal effective concentration (EC₅₀) was lower than 10 μ M.

4. The concentration at which aggregation began to be ameliorated at half maximal response was lower than the concentration at which the compound decreased cell count by half. If the cell count trend increased with concentration, or was unaffected, then this condition was automatically satisfied. Otherwise, the logarithm of the EC50 of the aggregation curve minus the logarithm of the EC50 of the cell count curve must be less than -0.10. We chose -0.10 as a more stringent threshold, as opposed to the difference being anything less than 0.0. This ensured that the compound's concentration at which it exhibited half of its maximum response occurred at a lower concentration than when the compound began to decrease cell count by half, and that this difference was not effectively zero.

1.5.9 Experimental Setup for Validation Assay of Functional Genomics Screen in Osteosarcoma Line

We generated a completely different HCS experiment than the tauopathy study in order to assess the ML method's generalizability. This osteosarcoma dataset was subjected to completely different biological conditions. U2OS cells (female) expressing a stable CCNB1-GFP construct were plated into 384 well plates with 500 cells per well and reverse transfected with an esiRNA library (10ng, Sigma Aldrich) using Hiperfect transfection reagent. The U2OS cells were cultured in DMEM medium containing 10% foetal bovine serum. The library can be found at <https://iccb.med.harvard.edu/sigma-esirna-human-1> and also <https://iccb.med.harvard.edu/sigma-esirna-human-2>. Our method was inspired by the assay presented in Bray et al.⁴⁷ We performed 16,194 unique functional genomic perturbations.

After 72 hours, cells were stained with 5 μ g/mL of Hoechst 33342 dye (ThermoFisher), incubated for 60 minutes at 37°C, washed with PBS, fixed in 4% PF, and scanned on a Thermo Cell Insight NXT high content microscopy system using a 10X objective. We captured the two channels, resulting in 324,989 Hoechst and cyclin-B1 image pairs, each of dimension 1104 x 1104 pixels.

1.5.10 Training and Evaluation of the ML Models Used for the Osteosarcoma Dataset

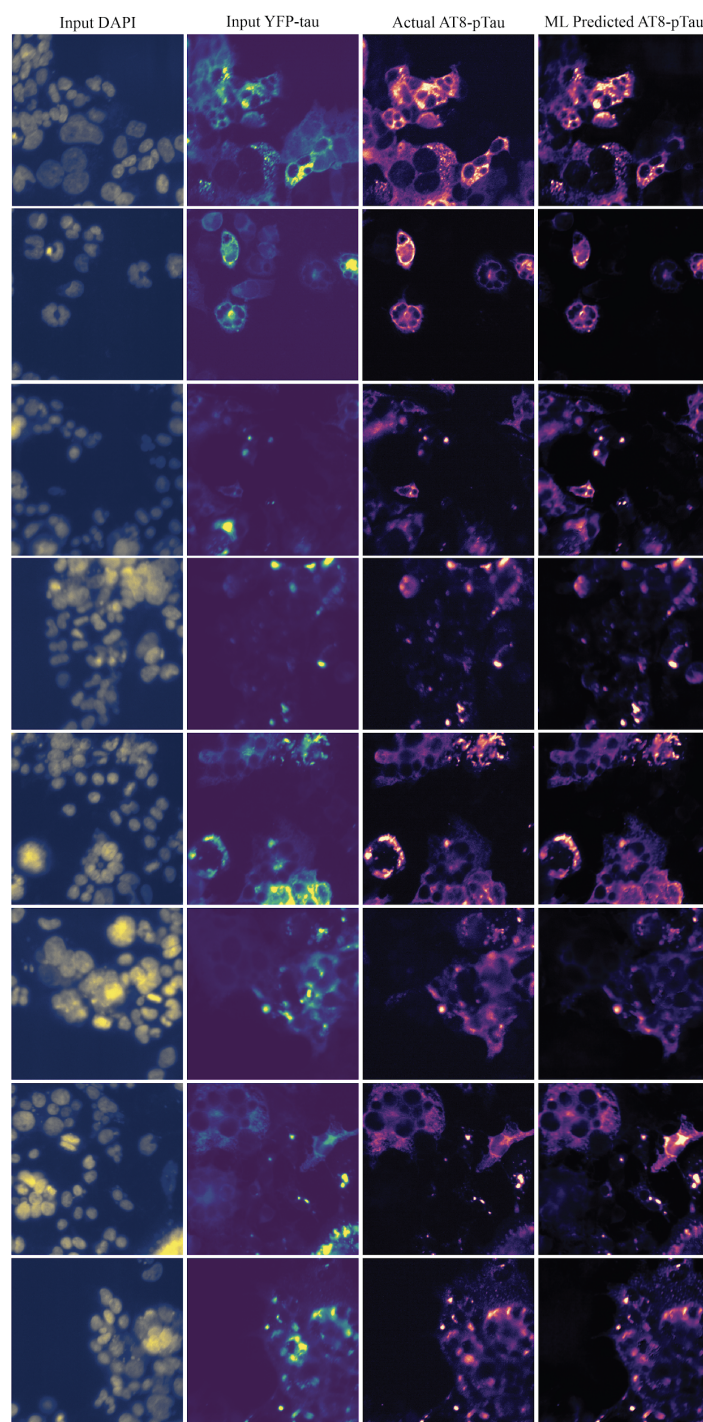
We performed two training experiments, one using ablated Hoechst images, and one using the raw Hoechst images. For the training using ablated Hoechst images, we ablated our Hoechst images at the 95th percentile pixel-intensity threshold. Afterwards, we performed the same exact training procedure as the one we used for the tauopathy experiment, except for the following exceptions: 1) the input to the model was one dimensional (the Hoechst channel) instead of two; 2) We trained for 10 epochs instead of 30 epochs; 3) We trained a threefold cross-validation instead of a single random training-test split.

For training using the unablated, raw images, we applied the same preprocessing and training procedure as the one used for training with ablated images, except for the following exceptions: 1) we left the images intact and did not perform any ablations; 2) we trained for 5 epochs instead of 10. We trained for a shorter time because the model converged faster, likely because the task of learning from raw unablated images was easier.

1.5.11 Quantification and Statistical Analyses

For all statistical significance tests, we used a two sample, one-sided z test. The null hypothesis stated that performance means were equal. The alternative hypothesis stated that the average performance from ML is greater than the average performance from the non-ML approach. Significance was set at $p < 0.05$. The values of n correspond to the number of instances in the test set (see Method Details for values of n). Sample sizes were sufficiently large to perform the test. A sample is one comparison between a predicted channel and the actual channel. We assume normality due to the large sample size.

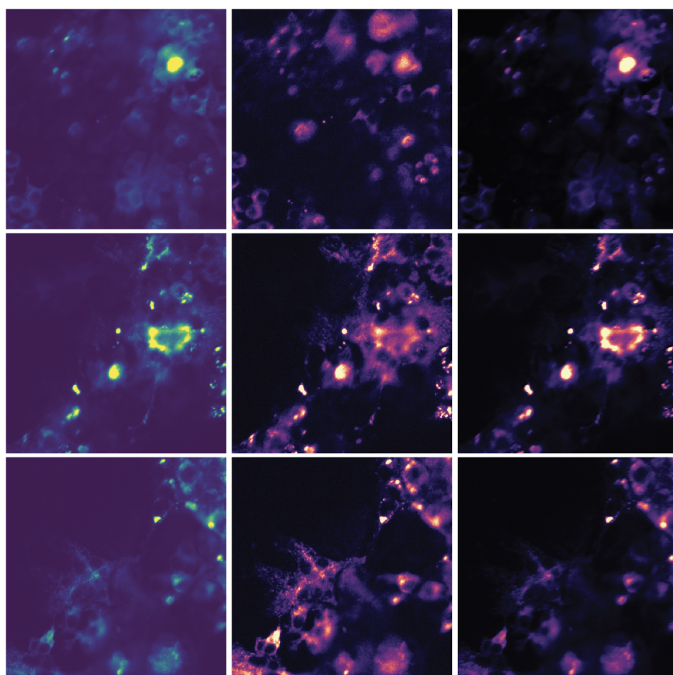
1.6 SUPPLEMENTAL FIGURES



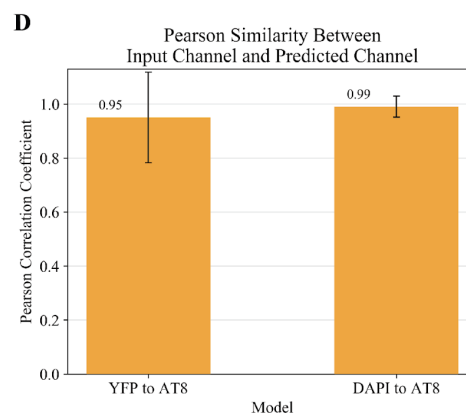
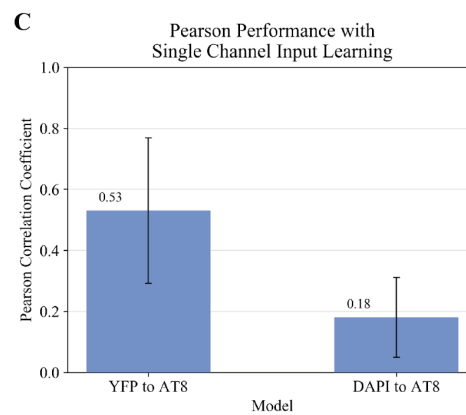
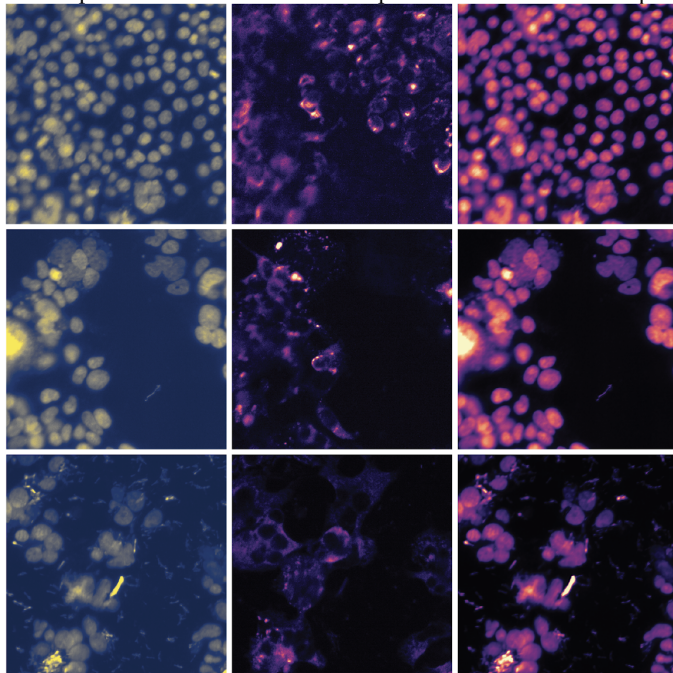
Supplemental Figure 1.1: Additional image examples showing trans-channel predictions.

The DAPI channel (leftmost) is concatenated with the YFP-tau channel (second column) and inputted to the trained ML Model, which produces a prediction of the AT8-pTau image channel (rightmost). This prediction has high similarity to the actually imaged AT8-pTau channel (third column). All images are pulled from the test set never seen during training.

A Input YFP-tau Actual AT8-pTau ML Predicted AT8-pTau

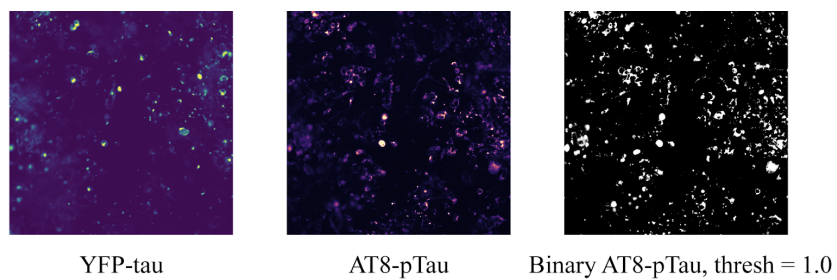


B Input DAPI Actual AT8-pTau ML Predicted AT8-pTau

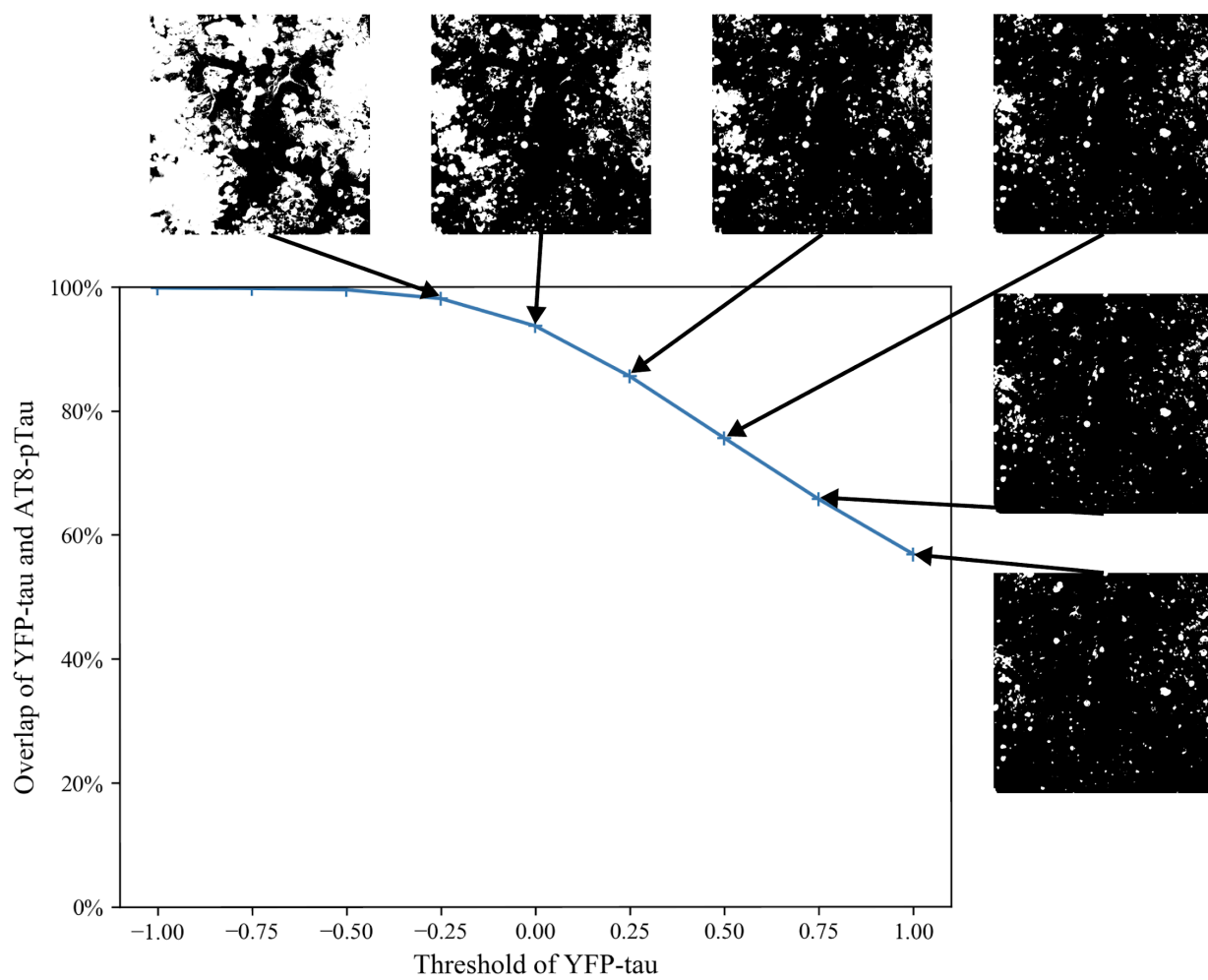


Supplemental Figure 1.2: Transchannel-learning from single channel input to objective AT8.

(a) Example input, label, and predicted images from a model trained to map just the YFP-tau channel (without the DAPI channel) to the AT8-pTau channel. (b) Example input, label, and predicted images from a model trained to map just the DAPI channel (without the YFP-tau channel) to the AT8-pTau channel. (c) Pearson performance of models trained with just a single channel, left: model trained to map just YFP-tau to AT8-pTau, right: model trained to map just DAPI to AT8-pTau. The YFP-tau to AT8-pTau model's performance was nearly identical to the Null Model that simply returned the input YFP-tau channel. The DAPI to AT8-pTau model failed to learn the AT8-pTau channel, as we would expect if the DAPI and AT8-pTau channels are not sufficiently biologically related. (d) Pearson correlation between the single channel model's input and the model's prediction. For both the YFP-tau to AT8-pTau model and the DAPI to AT8-pTau model, the input image is almost identical to the predicted image, indicating that the model learned to simply return its input as the final prediction to maximize the Pearson correlation coefficient. Error bars for (c) and (d) span one standard deviation in each direction, centered at the mean.

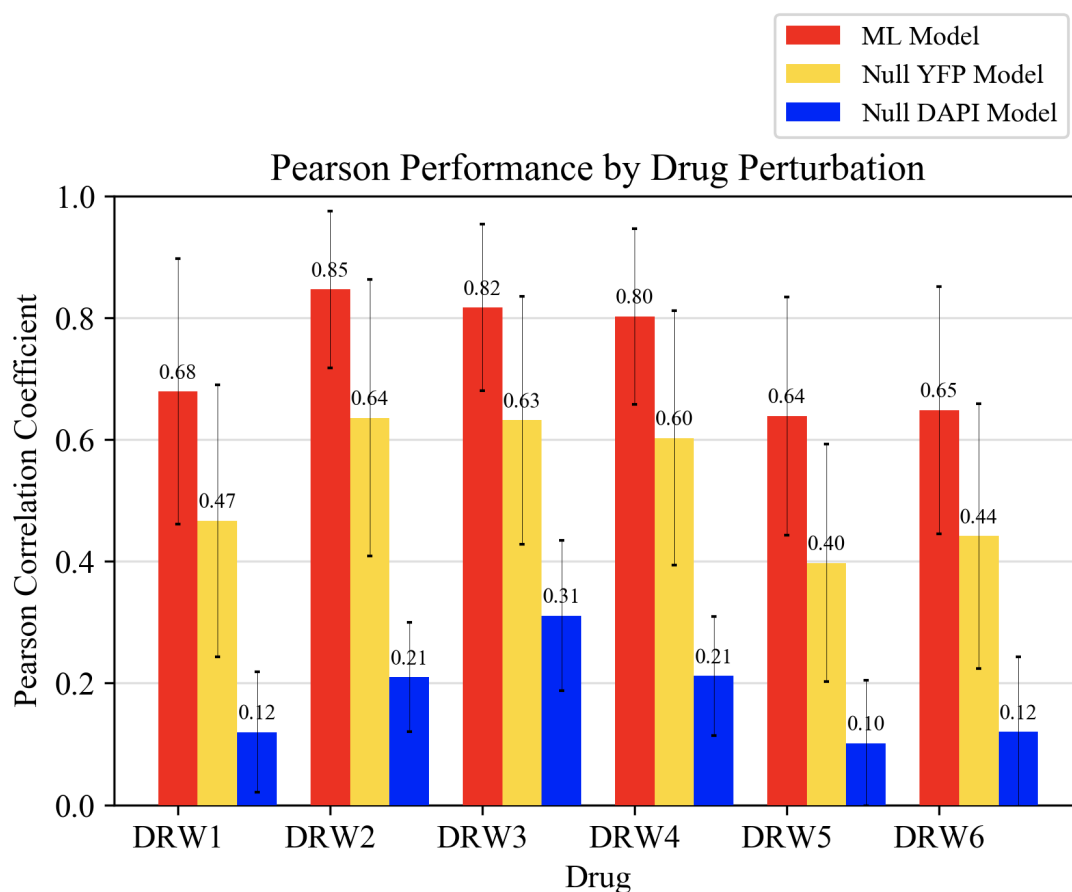


Binary YFP-tau images at different thresholds:



Supplemental Figure 1.3: Overlap of AT8-pTau signal and YFP-tau signal.

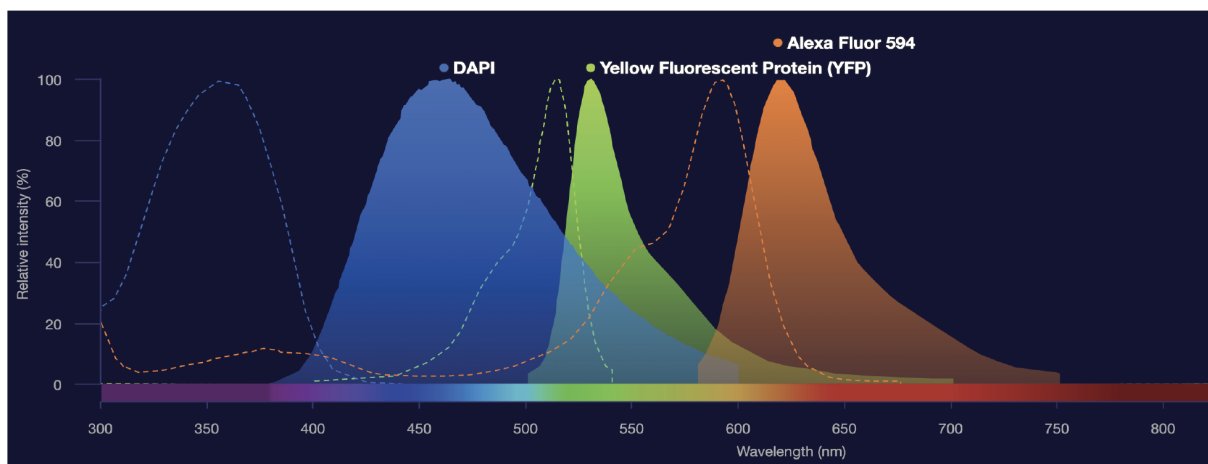
The x-axis shows the pixel value threshold that binarized each YFP-tau image, such that any pixel greater than or equal to the threshold was considered positive for signal, and anything below this threshold was considered negative for signal. The y-axis shows the average overlap between YFP-tau and AT8-pTau signal (at a threshold of 1.0) as a fraction of AT8-pTau positive signal that was also present in YFP-tau at the same pixel location. We calculated overlaps over the entire image dataset. Example binarized YFP-tau are shown at each threshold. Images at threshold = -0.50 and lower are not shown and consist of mostly signal-positive white images.



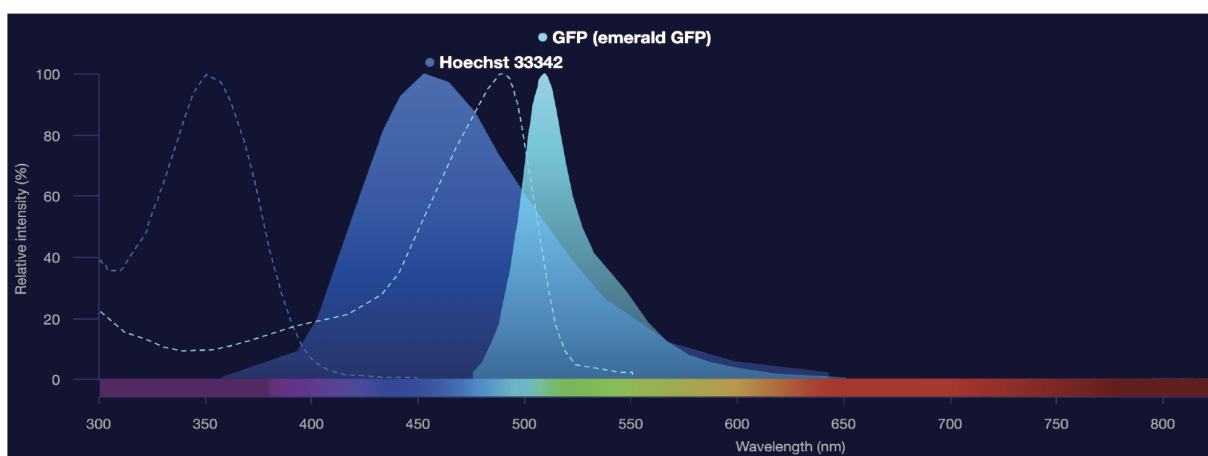
Supplemental Figure 1.4: Pearson performance by drug condition.

Performance by drug across the 17,280 held-out test images. Each test image comes from one of six drug perturbations. Error bars show one standard deviation in each direction centered at the mean.

A



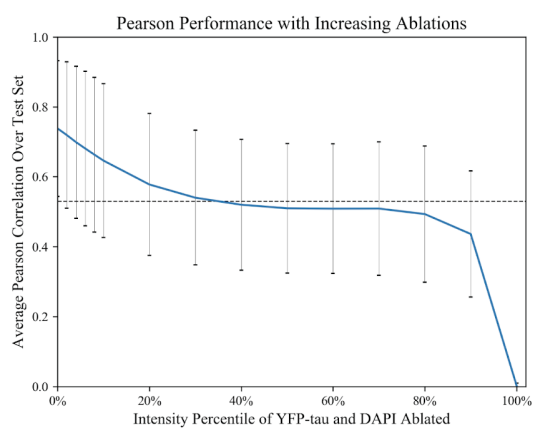
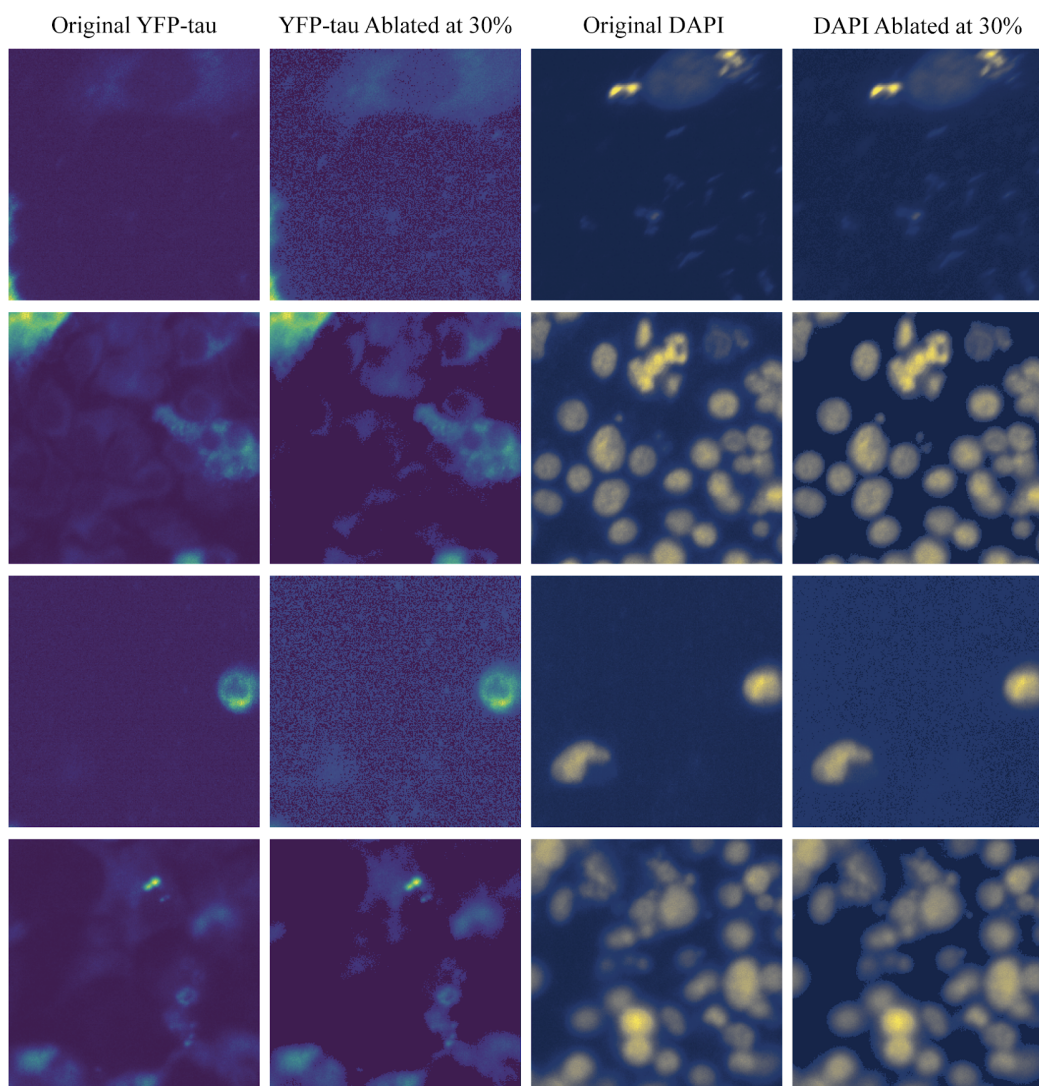
B



Supplemental Figure 1.5: Excitation and emission spectra of fluorophores used in the two biologically distinct datasets.

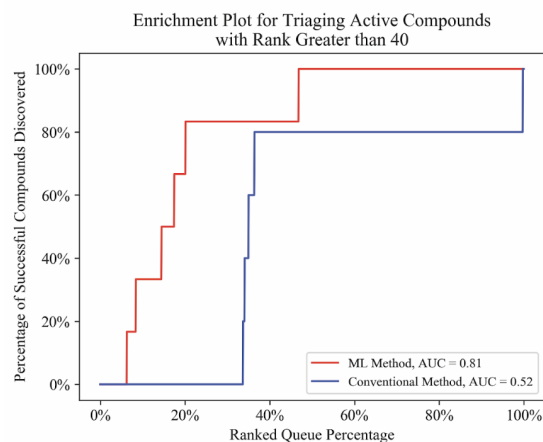
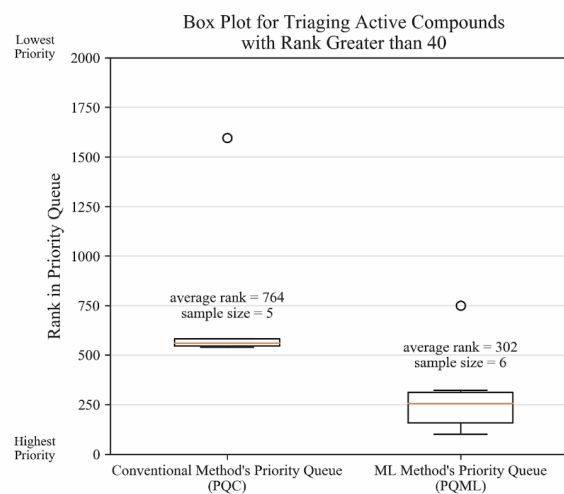
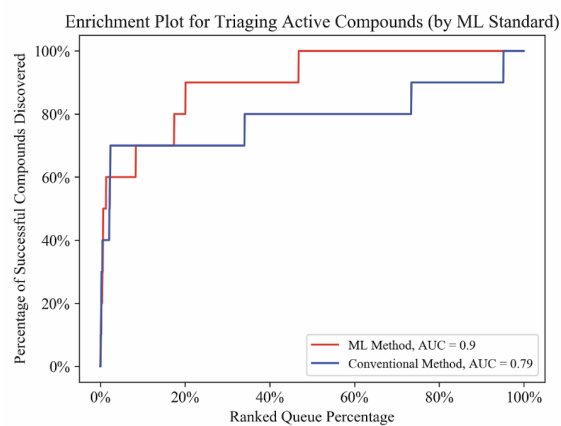
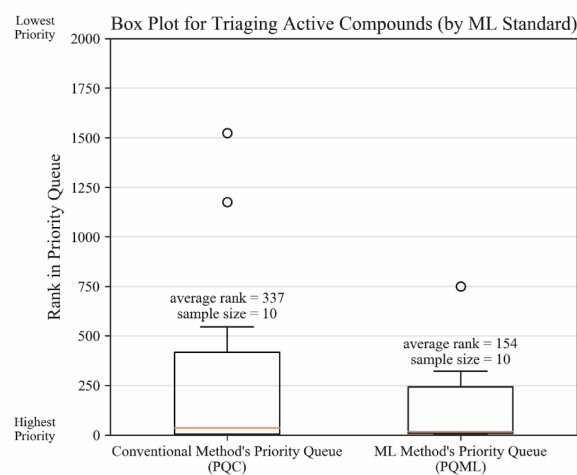
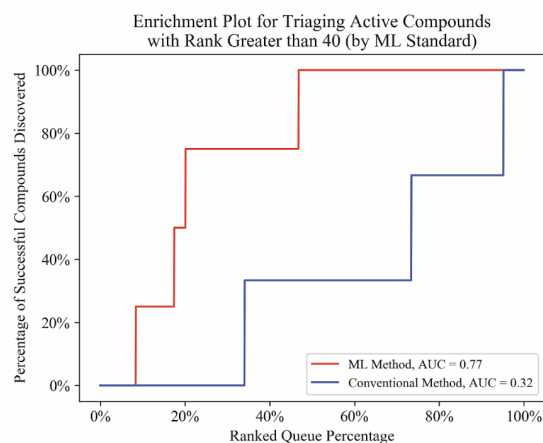
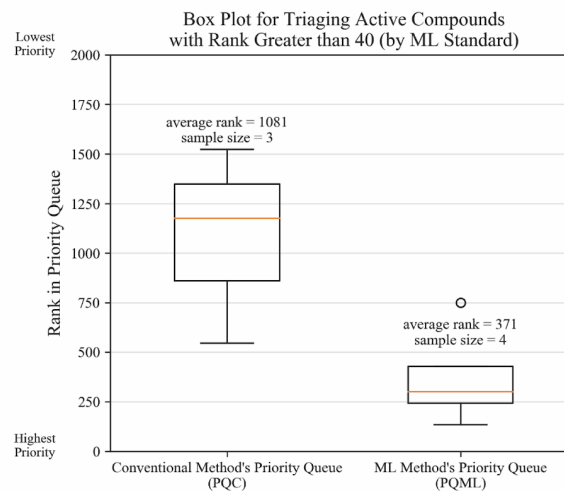
(a) Excitation and emission of the three-channel tauopathy training dataset. The x-axis plot the wavelength, and the y-axis plots the relative intensity. The dashed lines indicate the excitation plots, and the solid lines indicate the emission plots. Alexa Fluor 594 refers to the secondary antibody used for the AT8-pTau channel (b) Excitation and emission of the functional genomics screen in osteosarcoma cells. These graphs were constructed with ThermoFisher's Fluorescence SpectraViewer:

<https://www.thermofisher.com/us/en/home/life-science/cell-analysis/labeling-chemistry/fluorescence-spectraviewer.html#!/>

A**B**

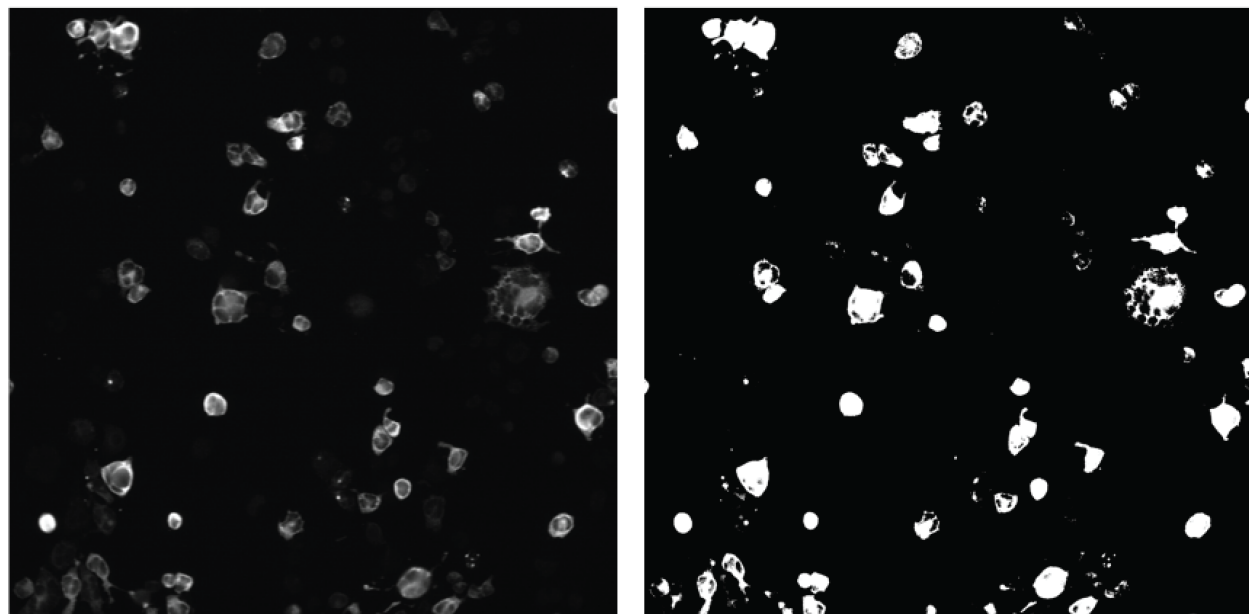
Supplemental Figure 1.6: Ablation testing on the tau dataset.

(a) We systematically ablated the input channels (both DAPI and YFP-tau) and assessed its effect on performance. We incrementally ablated our input images by reassigning pixels under a specific pixel-value threshold t to the value of 0. The threshold t was increased from no ablation to a complete erasure of the image. The y-axis plots the average Pearson correlation over the full test set at different percentile ablations (x-axis). We performed a more fine grained search at low ablations where faint bleed-through signal would occur if it was present. We took a step size of 2% from the 0th percentile to the 10th percentile, and a step size of 10% from the 10th percentile on to the 100th percentile. The dashed horizontal line indicates the Null YFP Model's average performance with no ablations. The ML model's performance under ablation approaches the Null YFP Model's performance with no ablation at an intensity percentile of 30%. Error bars span one standard deviation in each direction above and below the average. (b) Example images are shown for both original YFP-tau images (leftmost column), YFP-tau images ablated at 30% (second column), original DAPI images (third column), and DAPI images ablated at 30% (rightmost column). We can see noticeable visual abnormalities at a high ablation of 30%. Images were auto-enhanced and colored with ImageJ and Matplotlib for visualization purposes only.

A**B****C****D****E****F**

Supplemental Figure 1.7: Additional triaging analysis of missed but active compounds, and also enrichment by a different ML Standard.

(a) An enrichment plot for active compounds that had rank greater than 40. (b) A box plot is depicted for active compounds residing outside of a queue's top 40 list. The ML achieved an average rank of 302, which was more than twice as low as the average rank of 764 for missed but active compounds in PQC. (c) - (f) correspond to enrichment analyses when assessing a compound's dose-response activity from the ML-derived AT8-pTau images instead of the YFP-tau images (denoted here as the "ML Standard"). Other than this difference in how images are scored for aggregation, the ML Standard is identical to the standard presented in the main text. (c) Enrichment curve for all active compounds by the ML Standard. By this standard, there were 10 active compounds. (d) Box plot for active compounds by the ML Standard. (e) Enrichment curve by the ML Standard for active compounds residing outside of the top 40 list. (f) Box plot by the ML Standard for active compounds residing outside of the top 40 list.



Supplemental Figure 1.8: Binary thresholding of the label image in constructing the AUC.

The image on the left is the actual AT8-pTau label image (enhanced with ImageJ's auto-enhance function for visualization purposes only). The image on the right is the result of normalizing the raw AT8-pTau image, and then thresholding with a pixel value = 1.0, such that any pixels greater than or equal to 1.0 are shown in white as positive signal, and any pixel with intensity less than 1.0 is shown in black as negative signal.

Chapter 2: Deep learning from multiple experts improves identification of amyloid neuropathologies

Daniel R. Wong^{1,2,3,4,5}, Ziqi Tang^{2,3,4}, Nicholas C. Mew^{2,3,4}, Sakshi Das⁶, Justin Athey⁶, Kirsty E. McAleese⁷, Julia K. Kofler⁸, Margaret E. Flanagan^{9,10}, Ewa Borys¹¹, Charles L. White III¹², Atul J. Butte^{1,5,13}, Brittany N. Dugger^{6*}, Michael J. Keiser^{1,2,3,4*}

1. Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, 94158, USA

2. Institute for Neurodegenerative Diseases, University of California, San Francisco, CA, 94158, USA

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

4. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, 94158, USA

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

6. Department of Pathology and Laboratory Medicine, School of Medicine, University of California, Davis, Sacramento, CA 95817, USA

7. Translation and Clinical Research Institute, Newcastle University, Newcastle, United Kingdom
8. Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA, 15260, USA
9. Department of Pathology, Northwestern University, Evanston, IL, 60208, USA
10. Mesulam Center for Cognitive Neurology and Alzheimer’s Disease, Northwestern Medicine, Chicago, IL, 60611, USA
11. Department of Pathology, Loyola University Medical Center, Maywood, IL, 60153, USA
12. Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA
13. Center for Data-Driven Insights and Innovation, University of California, Office of the President, Oakland, CA, 94607, USA

*These authors contributed equally

2.1 ABSTRACT

Pathologists can label pathologies differently, making it challenging to design deep learning (DL) algorithms in the absence of one ground truth. We present a DL approach that draws on an expert cohort, weighs each contribution, and is robust to noisy labels. We collected

100,495 annotations on 20,099 candidate amyloid beta neuropathologies (cerebral amyloid angiopathy (CAA), and cored and diffuse plaques) from three institutions, independently annotated by five experts. DL methods trained on a consensus-of-two strategy yielded 12.6-26% improvements by area under the precision recall curve (AUPRC) when compared to those that learned individualized annotations. This strategy surpassed individual-expert models, even when unfairly assessed on benchmarks favoring them. Moreover, ensembling over individual models was robust to hidden random annotators. Saliency mapping on neuropathologies illustrated how human expertise may progress from novice to expert. In blind prospective tests of 52,555 subsequent expert-annotated images, the models labeled pathologies like their human counterparts (consensus model AUPRC=0.74 cored; 0.69 CAA). This study demonstrates a means to combine multiple ground truths into a common-ground DL model.

2.2 INTRODUCTION

There are inherent differences and areas of complementary expertise in evaluation within the medical disciplines. This becomes nuanced when differences in classification could lead to high-stakes differences in prognosis, treatment, and prevention measures⁴⁸⁻⁵⁰. Pathology assessment can benefit from more standardized approaches that improve quality and reliability, especially for imaging data⁵¹⁻⁵⁷. However, in cases of diagnostic discrepancy, it can be difficult to determine what is objective truth^{58,59}. Furthermore, medical professions such as neuropathology are also facing dwindling workforces despite high demand for their services^{60,61}. Hence, medical centers may not all have pathologists to render interpretations or may lack

bandwidth to collect multiple opinions on each case. Here, we present an automated framework to learn from multiple neuropathology experts and provide a robust labeling co-pilot tool that is resilient to discrepancies among different expertises.

Alzheimer’s disease is a neurodegenerative disease where specific proteins accumulate in select neuroanatomic regions⁶². Studying these accumulations and accurately characterizing their presence, localization, and distribution improves understanding of its pathophysiology⁶³. Extracellular accumulations of amyloid beta ($A\beta$) into plaques are a pathological hallmark of Alzheimer’s disease. $A\beta$ aggregates form diverse plaque morphologies, as well as deposits within blood vessels, termed cerebral amyloid angiopathy (CAA)^{64,65}. These morphologies may change with disease severity and correlate to select clinical features^{66,67}. Historically, however, the Consortium to Establish a Registry for Alzheimer’s Disease uses a semiquantitative measure of neuropathological phenotypes in its criteria⁵¹. Having more consistent, quantitative, and precise anatomic measures of $A\beta$ aggregates would not only standardize research endeavors to better understand pathophysiology across institutions, but would likewise allow for detection of more subtle pathophysiological differences.

Pathological annotation can be a laborious and time-intensive process that leverages the unique training of the expert^{52,63,68}. In previous work, we automated a single expert’s annotations using DL on images of Alzheimer’s disease pathologies⁶⁹. This approach was later validated by an independent study with a different cohort⁷⁰. However, individual bias of the expert remained, and it was not yet clear if these methods could scale across multiple experts, institutions, and

data modalities—all of which are critical for assessing generalizability⁷¹. In the current study, we draw on the contributions of five experts across a methodically and geographically diverse dataset.

Even if individual experts could be perfectly augmented by models trained to reflect their annotation expertise, a common ground truth is difficult to ascertain for challenging pathology tasks where experts make different assessments. Guan and Hinton⁷² trained neural network ensembles to mimic individual doctors, and created a benchmark of ground-truth examples of retinal neuropathy diagnosis by adjudicating decisions among experts. Although an active discussion may be a judicial way to handle more difficult labeling, a consortium of experts is not always available, requires much labor, and presents confounding social factors that may bias labeling. Avoiding human adjudication and performing the task accurately and automatically at scale—while still utilizing diverse expert proficiency—may be optimal and time-efficient as a first-pass or a second-opinion assessment. Other approaches have assigned reliability scores to each expert, and used statistical procedures to estimate ground truth^{73–77}. We assessed complementary approaches, by both modeling individual learners and also augmenting a consensus-voting scheme among experts. The consensus models learned the final assessment using various voting thresholds, resulting in different levels of sensitivity and precision. We evaluated which of these two approaches was the most robust and performant—learning from individual assessments or learning from some consensus of experts. Finally, we built on the

approach presented by Guan and Hinton⁷², and created ensembles that were robust to intentionally poor and noisy information.

Although DL can potentially provide more standardized and quantitative assessments for pathology^{78,79}, the complex mathematical functions learned can lack interpretability^{80,81}. There is a justifiable hesitancy to adopt such technology in high-stakes scenarios without knowing how a model's decisions are made^{82–84}. Recent advances in interpreting complex DL models, primarily through the use of saliency mapping, begin to elucidate this decision-making process⁸⁵. Employing such techniques, we found that DL models trained on individuals and on consensus annotations consistently learned relevant pathological features not explicitly given during training. The models' feature detection was granular to the pixel level, and may help better understand features considered by human experts with different levels of expertise.

In addition to interpretability, we evaluated these models in a prospective research study. The DL models accurately adopted annotation patterns of their expert counterparts. We found that models trained from a consensus of experts were able to (1) reproduce consensus annotations, (2) prospectively filter and enrich for neuropathologies in new whole slide images (WSIs), (3) and exceed the performance of models trained from individual experts at the same tasks. We found that modeling a consensus of experts was performant even when evaluated on different benchmarks that were expected to favor individual models, advocating for the robustness of consensus learning. This methodology may be a powerful means to leverage diverse and complementary human expertise—among pathologists and more generally.

2.3 RESULTS

2.3.1 We curated a multi-annotator dataset and found annotation differences among experts and consensus schemes

In prior work⁶⁹, we developed a convolutional neural network (CNN) pipeline to automatically identify three different A β neuropathologies for a single expert. In this current study, we generalized this method to five experts (NP1-NP5) and two novice annotators (UG1 and UG2). Oftentimes, pragmatic DL applications are trained on data with limited diversity, resulting in inability to effectively generalize to data outside their training corpus⁴². To counter this, we validated our method using a dataset of 43 WSIs obtained from three different institutions, where each used different histological staining procedures (Methods). Supplemental Figure 2.1 contains demographic information on these 43 research patients.

We organized the study into two phases of data collection. In phase-one, we collected independent annotations from the seven annotators on the same 20,099 images derived from 29 WSIs. As in previous work, we color-normalized the WSIs⁸⁶, identified candidate A β aggregates with conventional computer vision techniques, and center-cropped these candidates to provide 256 x 256 pixel images for final annotation (Methods). We arranged these candidates in a randomly shuffled but fixed ordering, and uploaded them to a custom web interface for expert annotation (Figure 2.1). We recruited five neuropathology experts and two undergraduate novices from five different medical institutions to annotate the images. Annotators performed a

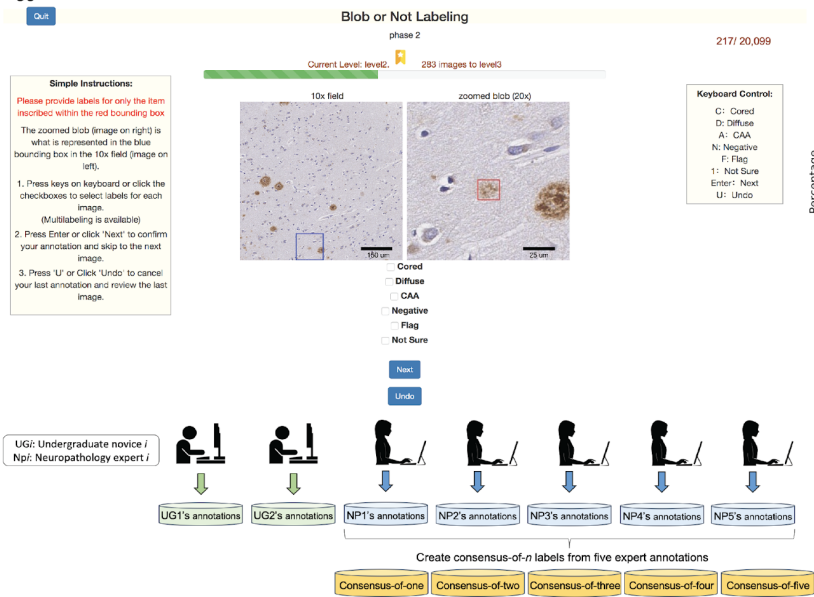
multi-labeling task and independently labeled an image as any combination of “cored,” “diffuse,” and “CAA.” The annotators also had options of marking “negative,” “flag,” and “not sure,” but we did not use these to alter our image labels (Supplemental Figure 2.2). As in previous work, the diffuse class was most prevalent (Figure 2.1).

To draw on the complementary proficiencies of multiple experts, we introduced the concept of a “consensus-of- n ” strategy to label pathologies. In the consensus-of- n strategy, an image was labeled positive for plaque p if at least n experts positively labeled the image as presenting p . Otherwise, the image was assigned as negative for p . A consensus-of-one was the most permissive strategy, in which only one expert needed to positively identify p for this image to be labeled as positive for p . A consensus-of-one (logical set union) maximizes sensitivity, while a stricter consensus-of-five (logical set intersection) maximizes precision. We created five different annotation datasets by applying the consensus-of- n strategy to the experts’ annotations, from $n=1$ to $n=5$. Each consensus annotation dataset combined information from all five experts’ annotations using this thresholding scheme, and consisted of labels for the same set of independently annotated 20,099 images. Images for each A β class and each consensus strategy are shown in Figure 2.1.

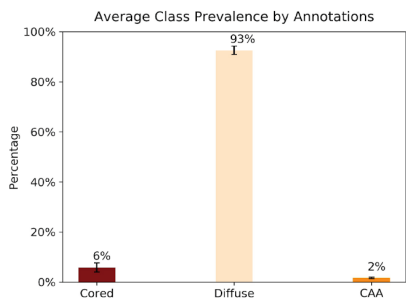
Qualitatively, we saw more phenotypic uniformity as more agreement was reached from $n=1$ to $n=5$. This held especially for the diffuse class in which some $n=1$ images resembled the phenotype for cored plaques, but as we increased to $n=5$, the classical phenotype emerged of sparse and scattered A β protein. A complete consensus-of-five experts was reached for 65% of

images labeled as a diffuse plaque from any expert (Figure 2.1). For the CAA class, the classical ring structure around blood vessels emerged as n increased. Complete consensus-of-five was reached for 33% of images with any CAA annotation (Figure 2.1). We did not specify CAA subtypes^{87,88}. For the cored class however, there was no smooth qualitative progression, visually indicating more differences and idiosyncrasies in identification. Complete agreement on a positive label only occurred for 11% of cored-plaque labeled images. There were many cases in which only one annotator identified a particular cored plaque, with this consensus-of-one scenario making up the majority (44%) of cored-plaque images (Figure 2.1).

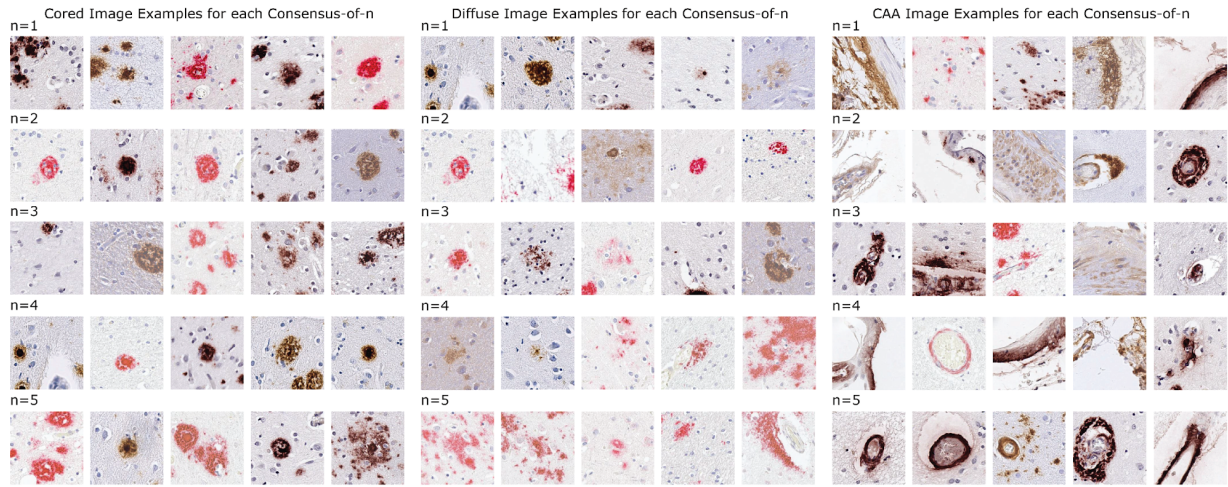
a



b



c



d

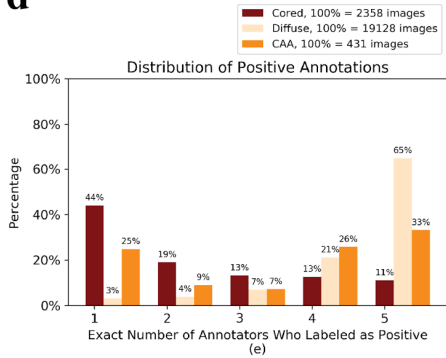


Figure 2.1: We curated annotations of A β neuropathologies from multiple experts, and found differing degrees of consensus.

(a) Five experts (NP) and two undergraduate novices (UG) used a custom web portal for annotation. Each annotator labeled the same set of images in the same order. From the expert annotations, we constructed consensus-of- n labels ($n=1$ to $n=5$) for the same 20,099 images. (b) Average class distributions are consistent across the seven annotators. The y-axis plots average frequency, while the x-axis plots the A β class. (c) Representative images illustrating consensus-of- n strategies applied to each A β class, with rows progressing from top to bottom in order of increasing consensus. For a consensus-of- n image, at least n experts labeled the image as positive for the designated class. Each image was randomly and independently chosen from the set of images. (d) Positive annotation distributions differ by A β class. The x-axis plots the exact (not cumulative) number of annotators who gave a positive label. Hence, when $e=1$ and $e=5$ this is equivalent to a consensus-of-one and consensus-of-five respectively. For $e=2, 3$, or 4 , this is *not* equivalent to an at-least- n consensus strategy. The y-axis plots the frequency. Each class has a different count of total positive labels (indicated in the legend). This total count represents the total number of images with at least one expert identifying the class. Each image may have multiple classes present.

To assess patterns of inter-rater agreement, we calculated the Cohen's kappa coefficient⁸⁹ between every pair of experts (Figure 2.2). There was low average agreement for the diffuse class ($\text{kappa} = 0.46 \pm 0.11$) despite the fact that a complete consensus ($n=5$) for it was the most common case out of any consensus. Hence, for diffuse cases without complete agreement, experts varied greatly. For CAA cases, average inter-rater agreement was much higher ($\text{kappa} = 0.76 \pm 0.10$). Agreement for the cored cases was low ($\text{kappa} = 0.50 \pm 0.08$). No single expert was a clear outlier in annotation across all A β classes. NP2 differed most with the other annotators for the diffuse and CAA classes, and NP4 differed most for the cored class.

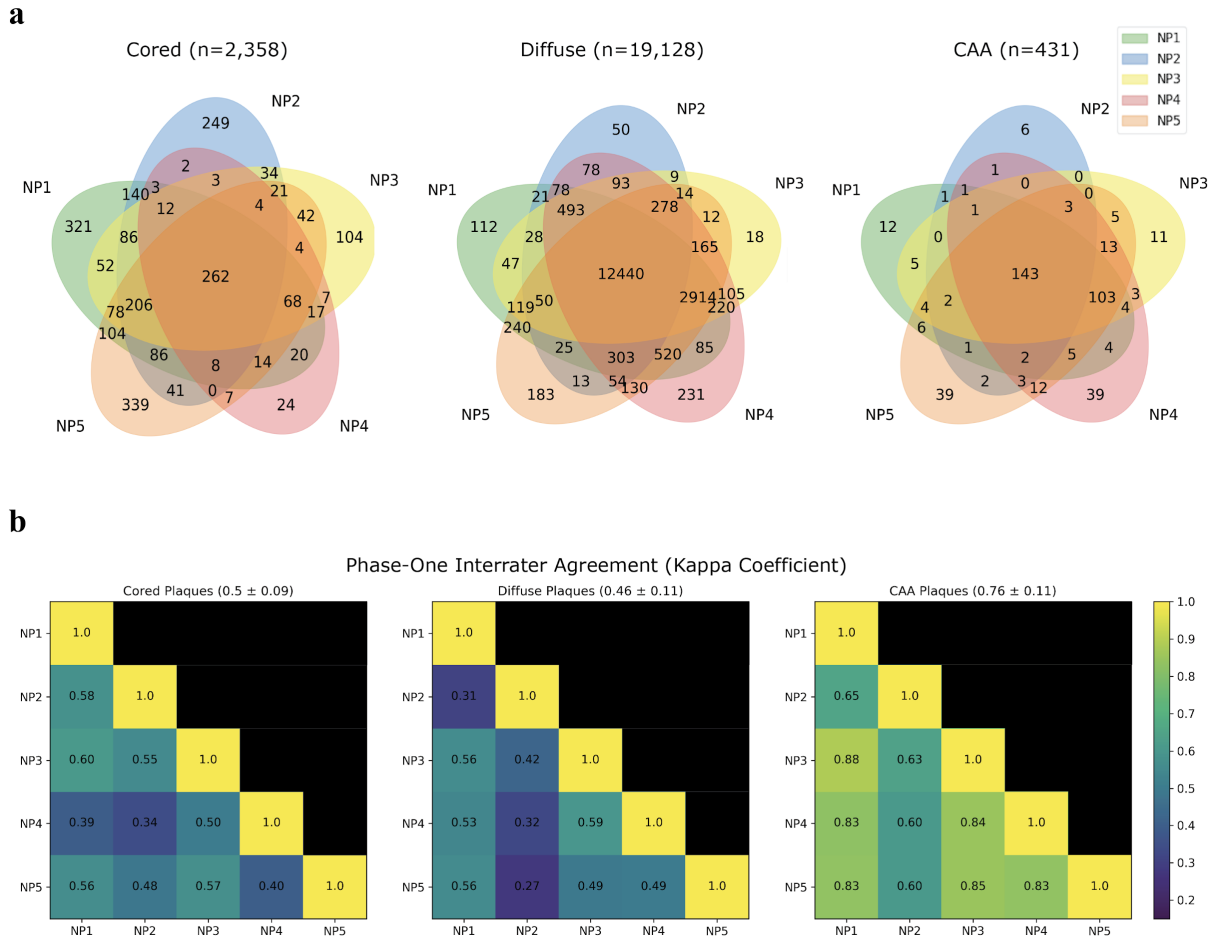


Figure 2.2: Inter-rater agreement varies by class and annotator.

(a) Venn diagrams by class, with overlaps of each permutation of NP1 through NP5. Each overlap shows the count of how many images are all positively annotated by the experts included in that overlap. Areas are not to scale. (b) Kappa coefficients⁸⁹ indicating agreement between each pair of experts. A high kappa coefficient indicates high inter-rater agreement between two annotators, with kappa = 1.0 indicating perfect agreement, and kappa = 0.0 indicating no agreement other than what would be expected by random chance.

2.3.2 CNNs mimic human annotators, and consensus CNNs mimic a consensus of experts

CNNs are a powerful class of DL networks particularly useful for analyzing image data^{90–92}. We found CNNs accurately learned the specific annotation behavior of both humans and consensus strategies. Using the five expert annotation datasets, we trained CNNs to reproduce each human’s annotations. We also trained independent CNNs to reproduce each consensus-of- n strategy.

We randomly assigned each WSI to either the training set or test set. Of the 20,099 center-cropped images, we held out 33% as a test set, and the remaining training images were split into a fourfold cross-validation (Methods). Both the models trained on individual-expert annotations and the models trained on the consensus annotations generalized and performed well on the held-out test set (Figure 2.3); they had high area under the receiver operating characteristic (AUROC) and area under the precision recall curve (AUPRC). We evaluated each model according to the labels of the annotation dataset on which it was trained. The consensus models mimicked consensus strategies, and the individual-expert models mimicked their corresponding expert annotators. Interestingly, the individual-expert models captured their human inter-rater agreement patterns, further suggesting that they were mimicking their human counterparts (Supplemental Figure 2.3). For every scoring metric and A β class, consensus models performed slightly better than individual-expert models (Figure 2.3). On average, they were able to more accurately reproduce the consensus strategies than the individual-expert models were able to reproduce their specific human annotators. There were substantial

performance differences among the A β classes. All models were able to accurately reproduce the annotations of the diffuse class, which was far more ubiquitous than cored and CAA. For these crucial but minority classes, AUROC and AUPRC performance was lower likely due to less available training examples (e.g., 1-2% for CAA class prevalence, depending on label strategy). Performance by stain was largely consistent, except for the CAA class with 6E10 staining, which had lower and more variable performance (Supplemental Figure 2.4). Color-normalization had little effect on performance, except for the CAA class (Supplemental Figure 2.5).

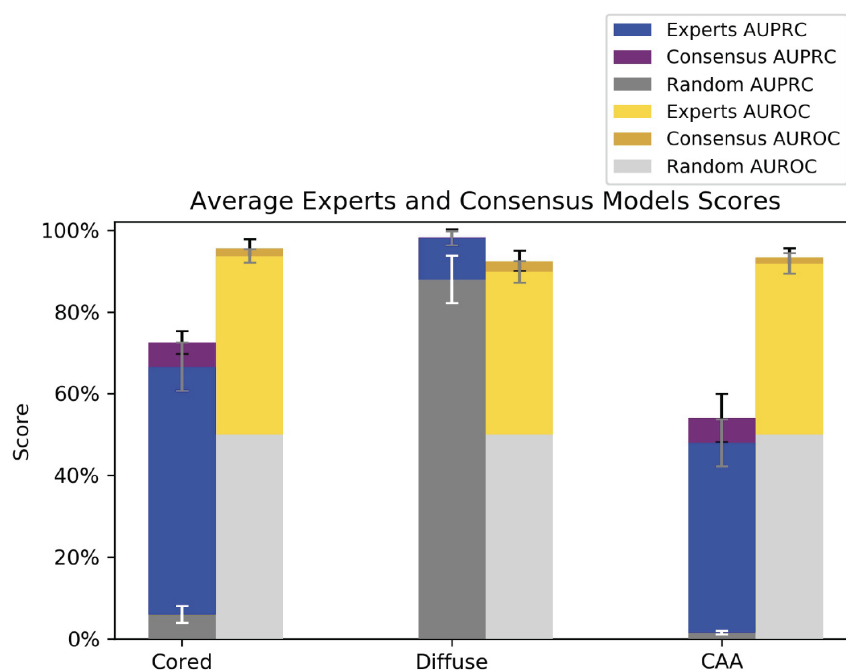


Figure 2.3: We trained models to learn human annotation behavior and consensus strategies.

Consensus models matched or outperformed individual-expert models in average AUROC and AUPRC, per stacked bar graphs. Error bars show one standard deviation in each direction. The y-axis indicates the score on the held-out test set for each A β class (x-axis). No novice models were included in this evaluation. For the AUPRC metric, the consensus model achieved 0.73 ± 0.03 for

cored, 0.98 ± 0.02 for diffuse, and 0.54 ± 0.06 for CAA. The individual-expert models achieved 0.67 ± 0.06 for cored, 0.98 ± 0.02 for diffuse, and 0.48 ± 0.06 for CAA. Random baseline performance for AUPRC is the average prevalence of positive examples. Average random baselines for individuals-experts were equivalent to those of consensus strategies (variance of individual-experts shown): 0.06 ± 0.02 for cored, 0.88 ± 0.06 , and 0.02 ± 0.004 for CAA. For the AUROC metric, the consensus models achieved 0.96 ± 0.02 for cored, 0.92 ± 0.02 for diffuse, and 0.93 ± 0.02 for CAA. The individual-expert models achieved 0.94 ± 0.02 for cored, 0.90 ± 0.03 for diffuse, and 0.92 ± 0.03 for CAA.

2.3.3 Consensus provided superior models and annotation benchmarks

We found that learning from a consensus strategy resulted in consistently higher model performance than learning from individual-expert annotations. Since there was no clear best or established evaluation benchmark strategy across the test-set images (with each image having five independent expert annotations), we compared the consensus models with the individual-expert models using four different benchmark schemes (Figure 2.4).

For all four benchmark schemes, the consensus models had superior average AUPRC performance versus individual-expert models (Figure 2.4). We found this surprising, especially for the “individual benchmarks” case in which all models were evaluated with the different experts’ annotation benchmarks. We had expected individual-expert models to perform the best on this benchmark, because they were trained specifically to mimic these same annotators. Instead, we observed the consensus models—which in this scenario were trained under a different annotation dataset than the one that they were being evaluated with—consistently performed better on average than the individual-expert models across all $A\beta$ classes (Figure 2.4). Of the five consensus schemes, the consensus-of-two model achieved the highest average

AUPRC (0.70 ± 0.24) and AUROC (0.88 ± 0.13) when evaluated against all benchmarks and all A β classes. On average, this consensus-of-two model substantially outperformed individual-expert models for all benchmark schemes and for all A β classes, except for the diffuse class in which performance was roughly equivalent (Figure 2.4).

a

Self Benchmarks

Model	NP1	NP2	NP3	NP4	NP5	C1	C2	C3	C4	C5
NP1	X									
NP2		X								
NP3			X							
NP4				X						
NP5					X					
C1						X				
C2							X			
C3								X		
C4									X	
C5										X
	NP1	NP2	NP3	NP4	NP5	C1	C2	C3	C4	C5

Benchmark

Consensus Benchmarks

Model	NP1	NP2	NP3	NP4	NP5	C1	C2	C3	C4	C5
NP1						X	X	X	X	X
NP2						X	X	X	X	X
NP3						X	X	X	X	X
NP4						X	X	X	X	X
NP5						X	X	X	X	X
C1						X	X	X	X	X
C2						X	X	X	X	X
C3						X	X	X	X	X
C4						X	X	X	X	X
C5						X	X	X	X	X
	NP1	NP2	NP3	NP4	NP5	C1	C2	C3	C4	C5

Benchmark

Individual Benchmarks

Model	NP1	NP2	NP3	NP4	NP5	C1	C2	C3	C4	C5
NP1	X	X	X	X	X					
NP2	X	X	X	X	X					
NP3	X	X	X	X	X					
NP4	X	X	X	X	X					
NP5	X	X	X	X	X					
C1	X	X	X	X	X					
C2	X	X	X	X	X					
C3	X	X	X	X	X					
C4	X	X	X	X	X					
C5	X	X	X	X	X					
	NP1	NP2	NP3	NP4	NP5	C1	C2	C3	C4	C5

Benchmark

All Benchmarks

Model	NP1	NP2	NP3	NP4	NP5	C1	C2	C3	C4	C5
NP1	X	X	X	X	X	X	X	X	X	X
NP2	X	X	X	X	X	X	X	X	X	X
NP3	X	X	X	X	X	X	X	X	X	X
NP4	X	X	X	X	X	X	X	X	X	X
NP5	X	X	X	X	X	X	X	X	X	X
C1	X	X	X	X	X	X	X	X	X	X
C2	X	X	X	X	X	X	X	X	X	X
C3	X	X	X	X	X	X	X	X	X	X
C4	X	X	X	X	X	X	X	X	X	X
C5	X	X	X	X	X	X	X	X	X	X
	NP1	NP2	NP3	NP4	NP5	C1	C2	C3	C4	C5

Benchmark

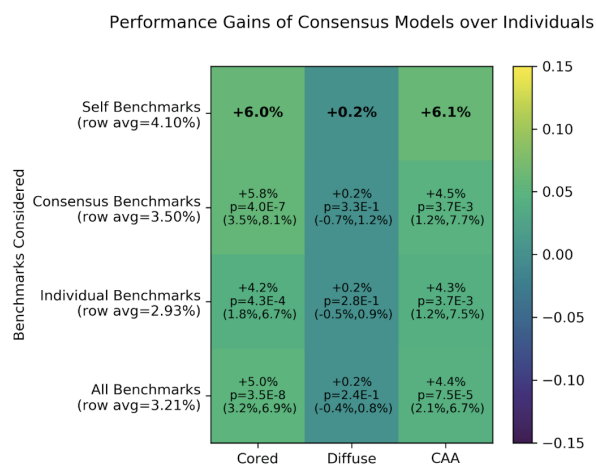
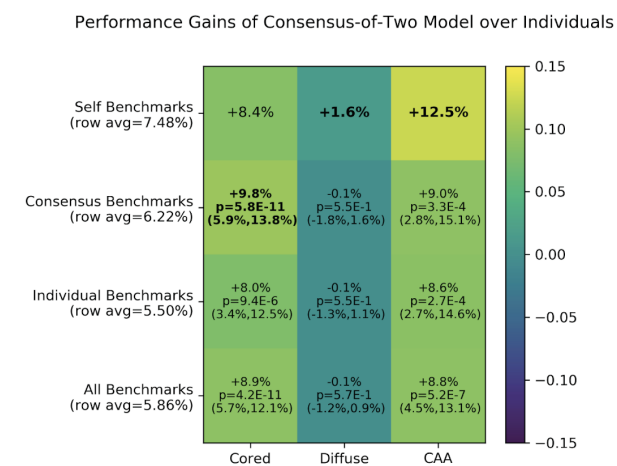
b**c**

Figure 2.4: Consensus models performed better than individual-expert models across all benchmarks.

(a) Four evaluation benchmark schemes to compare consensus models with individual-expert models. The row indicates the model and the column indicates the benchmark. For each evaluation scheme, the average AUPRC of the blue region (individual-expert models) is compared with the average AUPRC of the gold region (consensus models) over the held-out test set. The consensus-of-two is dark-gold for emphasis. The “self benchmarks” scheme was the most internally-consistent scheme that evaluated each individual-expert model according to the labels of its annotator (i.e. its own benchmark). For consensus models, the self benchmark corresponded to labels derived from the matching consensus-of- n strategy. The “consensus benchmarks” scheme independently evaluated each model on every consensus-of- n annotation set from $n=1$ to $n=5$. The “individual benchmarks” scheme independently evaluated each model on each of the five individual-expert benchmarks. The “all benchmarks” scheme evaluated each model on its average performance across all benchmarks. (b) Performance gains of consensus models over individual-expert models. Values are reported as the absolute AUPRC difference. We calculated p-values of the comparisons using a two-sample Z-test (Methods). P-values for the self-benchmark are not included because the sample size ($n=20$ comparisons) is not large enough to assign significance. 95% confidence intervals shown in parentheses. The row indicates the type of benchmark considered when evaluating the model performance differentials, while the column shows the A β class being evaluated. Highest performance differential for each A β class in bold. (c) Heatmap as in (b), for only the consensus-of-two model versus the individual-expert models. For this consensus-of-two model evaluation, only dark-gold regions in (a) corresponding to the consensus-of-two model are compared to the blue region.

When we did the converse and compared annotation *benchmarks* as opposed to comparing *models*, we observed annotation datasets from consensus-of- n strategies provided a more robust benchmark yielding greater average model performance, regardless of which models we evaluated (Supplemental Figure 2.6). Similarly, regardless of training strategy, models performed better on the consensus-of-two benchmark than on individual-expert benchmarks, for all A β classes (Supplemental Figure 2.6).

2.3.4 Model interpretation reflected differences in human expertise

Models learned human-interpretable, pathologically-relevant, and granular visual features for each A β class. By incorporating saliency mapping methods^{85,93} to interpret the model's rationale, we found models focused on pixels corresponding to boundaries of A β pathologies and excluding boundaries of extraneous deposits in a way that was task (pathology) specific—even though this granularity of information was not provided during training (Figure 2.5).

To investigate the impact of field expertise, we also trained models from the annotations of two undergraduate novices. We observed that a model trained on a consensus of two experts had more focused and specific feature saliency than a model trained on an individual undergraduate novice. To visualize what parts of an image were important to a model's decision making, we used guided gradient-weighted class activation mapping, or “guided Grad-CAM”⁸⁵ (Methods). When we compared class activation maps (CAMs) of the two models—consensus-of-two versus an undergraduate novice—we observed the CAMs of the novice's model were more diffuse than CAMs of the consensus-of-two model (Figure 2.5). Qualitatively, the individual novice models attributed greater salience to pixel features that were not important to the classification task, while the consensus-of-two model focused on relevant morphologies. Therefore, the model trained on consensus across greater expertise appeared to be a more specific feature detector. These results held for both novice annotators (Supplemental Figure 2.7). However, subsequent studies with additional novices are needed to investigate this trend.

Quantitatively, for each A β class and for each image in our test set, we compared its novice CAMs to its consensus-of-two CAMs. For each pair of CAMs derived from the same image and same class, but different model, we binarized the CAMs across incrementing thresholds (Figure 2.5). The majority of activations matched between the novice and consensus-of-two model at the different thresholds. In regions where they did not match, the novice CAMs tended to have signal while the consensus-of-two CAM did not (Figure 2.5). Figure 2.5 shows binarized image examples at different thresholds. At high pixel thresholds the images became more similar as the granular features of the image were lost. Furthermore, we computed the fraction of the consensus-of-two CAMs found within the novice CAMs, and also the fraction of the novice CAMs found within the consensus-of-two CAMs across different pixel thresholds. The consensus-of-two CAMs were a smaller and more focused subset of the novice CAMs (Figure 2.5).

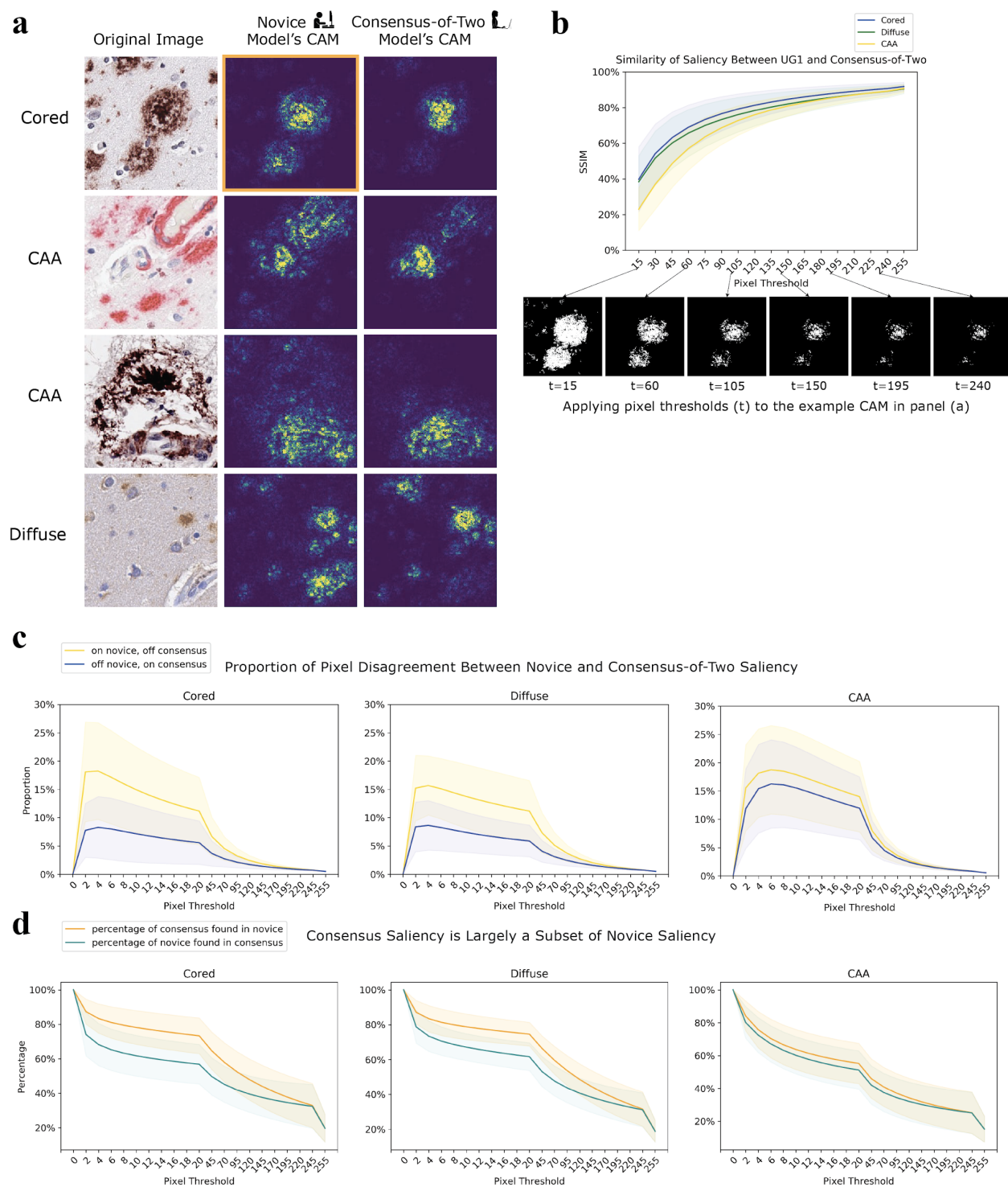


Figure 2.5: Class activation maps (CAM) of DL models indicate progression of human expertise.

(a) Novice CAMs are more diffuse than expert CAMs. The original image (leftmost column), the CAM of the novice model trained on UG1’s annotations (middle column), and the CAM of the consensus-of-two model (rightmost column). CAMs are plotted with a false-color map such that bright regions correspond to high intensity regions with high salience. (b) Although expert and novice CAMs differ, they converge on the same pixels. We progressively assess the structural similarity index (SSIM)⁹⁴ between novice CAMs and consensus-of-two CAMs across the entire test set of images. The CAMs show the most similar salience by SSIM (y-axis) at the highest pixel thresholds as we increment the threshold (x-axis) used to binarize the images before comparison. Binarized examples are shown of one CAM from (a) (boxed in orange). (c) Comparing the novice CAMs and the consensus-of-two CAMs, we classify each pixel location into two categories: ON in the novice CAM and OFF in the corresponding consensus CAM (yellow), or OFF in the novice CAM and ON in the consensus CAM (blue). ON and OFF are determined by binarizing the images at pixel threshold t (x-axis). Y-axis shows the proportions at which these two cases occur. Zoomed inset highlights disagreement between CAMs. (d) Consensus CAM pixels are mostly contained within the novice CAM. The x-axis plots the varying pixel thresholds, while the y-axis plots the percent overlap of either how much of the consensus CAM pixels are a subset of the novice CAM (orange) or how much those of the novice CAM are a subset of the consensus (cyan).

2.3.5 Ensemble learners bolstered performance and were robust to intentionally noisy

“annotators”

We further tested multi-rater strategies by creating an ensemble of individual-expert models (no novices) and taking a “wisdom of crowds” approach⁹⁵. One approach would be to simply take the predictions of the different expert models and use a majority-voting scheme. However, weighting each model equally would ignore the specific and differing expertise the models gleaned during training. Other approaches assign differently weighted votes to different models^{72,96}. Accordingly, we combined the individual-expert models in a learnable way, such that the resulting ensemble learned how to properly weight each contributing network’s A β class predictions to maximize overall performance. We theorized that the ensemble would learn how

to combine the strengths of individual-expert models. The ensemble training did not modify the individual CNNs, but rather learned how to weight each constituent network to maximize performance for a given annotation benchmark (Figure 2.6). We created ensembles for each of the phase-one annotation sets: both individual-expert annotation sets, as well as consensus annotation sets (Methods).

This ensemble approach bolstered performance, and improved over individual-expert CNNs in identifying CAAs. On average, ensembles outperformed single (non-ensembled) CNNs (Figure 2.6). Performance differences for diffuse class detection were relatively unaffected, likely because these examples were already highly prevalent. Rather, we observed performance gains for the less prevalent but pathologically important minority classes (cored and CAA). For consensus ensemble cases, in which we trained different ensembles to reproduce consensus-of- n annotations from $n=1$ to $n=5$, the ensembles matched or slightly outperformed single CNNs trained to mirror the consensus.

Next, we investigated whether these performance gains were robust to having a poor annotator included in the ensembles (Figure 2.6). We modeled this by creating a random labeler, who assigned randomly shuffled annotations that matched the average class distribution of all experts. When we included a random annotator CNN in the ensembles, the ensembles successfully learned to reject the noisy information from the random labeler. Performance was largely unaffected by the random annotator CNN, with differences in performance averaging to

less than 0.01 in AUPRC across every permutation of ensemble model and evaluation benchmark (Figure 2.6).

In a separate experiment, we also included five independent random labeler CNNs in the ensembles (Figure 2.6E), and we observed the same resilience and robustness. The ensembles successfully learned to ignore noisy contributors, and maintained performance even at a population that consisted of 50% poor labelers who inserted random information (Figure 2.6). Hence, we found ensembles resulted in better performance, and were robust to multiple poor annotators at little perceived risk of random information. By contrast, including random annotations into the consensus-of- n schemes would proportionally dilute the training data signal.

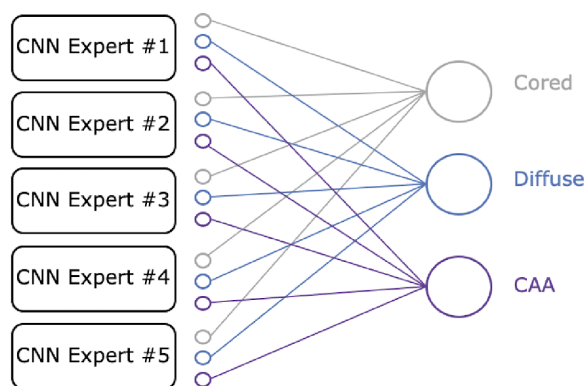
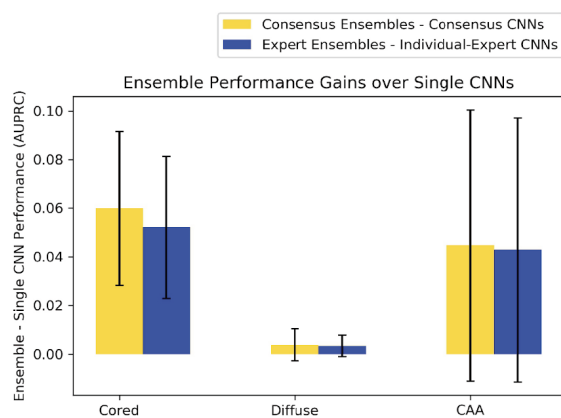
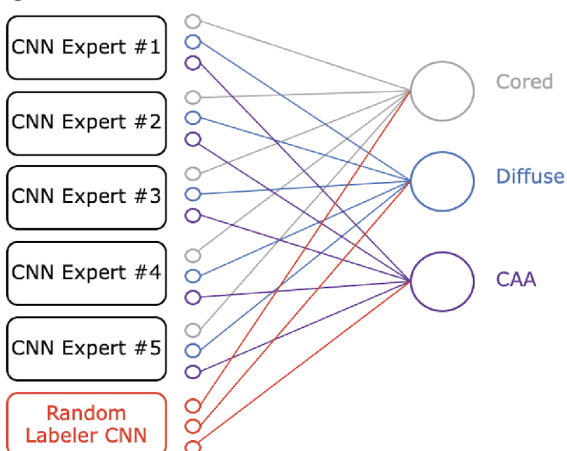
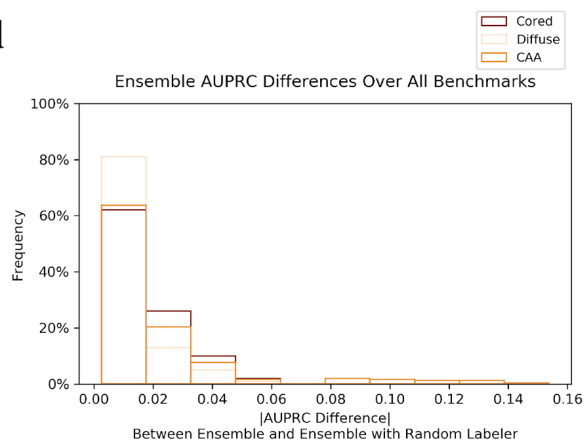
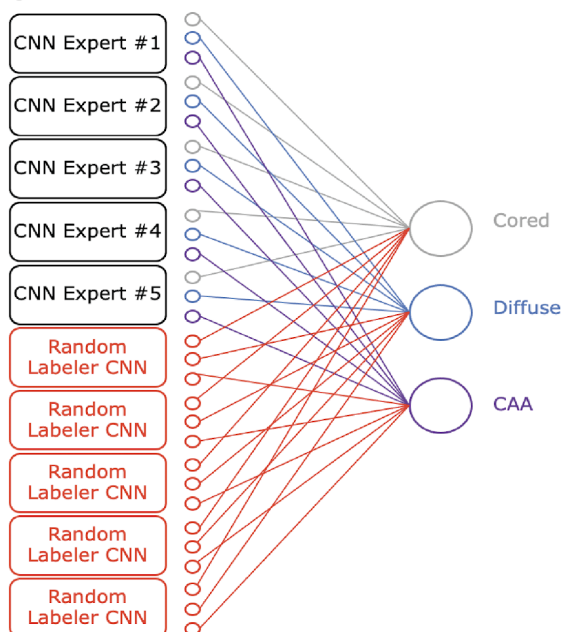
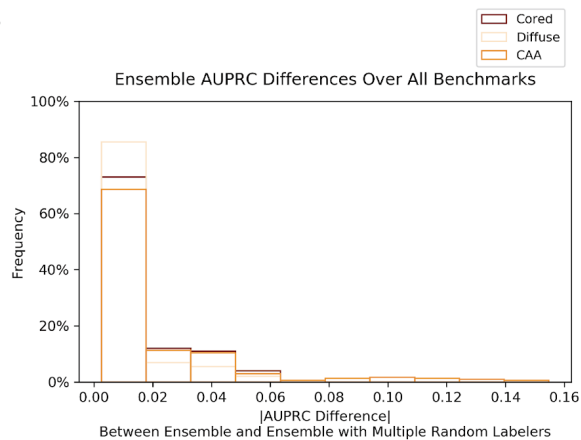
a**b****c****d****e****f**

Figure 2.6: Ensembles improve performance and are robust to false information.

(a) Five trained individual-expert CNNs, combined by a trainable sparse affine layer, make up an ensemble model. The training process simply determines how to best weigh and combine each CNN's existing class predictions. (b) Ensembling on average increases performance for each A β class, and for both consensus and individual benchmarks. Performance gains are calculated by averaging each ensemble's AUPRC on the held-out test set minus the corresponding individual-expert CNN's AUPRC on the same set, across all ten benchmarks (Methods). (c) We tested ensembling with a random labeler CNN, trained using a randomly shuffled permutation of labels with the same class distribution ratios as the five expert annotations. (d) Ensemble performance is largely unaffected by inclusion of a random labeler CNN. Density histogram of AUPRC performance differences for each A β class between the normal ensemble and the ensemble with a single random labeler CNN. Each ensemble is evaluated on all ten benchmarks (five individual-expert benchmarks, five consensus benchmarks), and the absolute value of the performance differential (x-axis) is calculated and binned for each class. (e) Ensemble architecture with multiple random labeler CNNs, each trained on a different permutation of randomly shuffled labels. (f) Ensemble performance is largely unaffected by inclusion of five random labeler CNNs. Same density histogram as in (d), but comparison is between normal ensemble and ensemble with five random labeler CNNs injected.

2.3.6 Annotators performed a second phase of annotation and were self-consistent

To assess use in a real-world clinical research setting, we designed a second, prospective, stage of the study (denoted as “phase-two”) to assess whether models could usefully serve as a prospective filter by identifying comparatively infrequent cored plaques and CAAs on new patients. Whereas cored plaques and CAAs have low prevalence (12% and 2% in phase-one, respectively), we posited that models from phase-one could filter a large and previously unseen dataset of primarily diffuse plaques to enrich for these rare but important neuropathologies. We hypothesized these models could prospectively predict what their annotators were going to label. Moreover, given our initial findings that a consensus-of-two model was more effective and robust than individual-expert models, we compared their prospective capabilities.

We conducted phase-two annotations six months after the completion of phase-one. During phase-two, each annotator received a total of 10,511 images to label, using the same web platform as phase-one (Figure 2.7). Annotators were told only that the study would proceed in two phases, but were given no information regarding data selection. Each blinded image came from one of four categories: “self-repeat,” “consensus-repeat,” “self-enrichment,” and “consensus-enrichment.” Self-repeats and consensus-repeats were collected to assess how consistently annotators were able to label images. These repeats were images the annotators had already seen in the first phase of annotation, and were triplicated, rotated randomly in 90-degree increments, and randomly spread throughout phase-two for intra-rater reliability testing⁹⁷ (Methods). Self-repeats were random selections of images that the annotators labeled as positive during phase-one, balanced across all A β classes. Consensus-repeats were images with positive labels in the consensus-of-two annotation set from phase-one (Methods). Encouragingly, each annotator achieved high intra-rater agreement for all of the repeat images, and was able to consistently annotate and reproduce the same labels across the six-month gap (Figure 2.7). Each expert achieved greater than 90% intra-rater accuracy among all A β classes, while novices achieved greater than 88% accuracy. There was no significant difference in intra-rater agreement between self-repeats and consensus-repeats (Supplemental Figure 2.8).

Both self-enrichment and consensus-enrichment images were selected from unannotated and different WSIs than phase-one, spanning 14 new patients across all three institutions. For the self-enrichment images assigned to annotator A, we used the individual-expert CNN trained on

A's phase-one annotations to enrich for images that the model predicted as having an important but minority plaque present (cored or CAA). For consensus-enrichment images, we used predictions from the consensus-of-two model instead to enrich for images. The same ordering of the four image categories was given to each annotator for consistency. Images differed between annotators in the self-repeat and self-enrichment categories only. See Methods for a detailed description of the different categories and selection process.

2.3.7 Models prospectively enriched for minority A β classes and favored consensus learning

Before completing this phase of the study, it was not clear whether a model trained on an individual annotator would prospectively perform the filtering and prediction task the best, as compared to a model trained from expert consensus. Consequently, from the 10,511 phase-two annotations across five expert annotators (i.e., total of 52,555 phase-two expert annotations), we created performance benchmarks to compare strategies. Models were assessed by either taking the individual expert's labels as truth (called the "individual benchmarks," five total benchmarks), or by taking a consensus-of-two experts as truth (called the "consensus benchmark", one benchmark). We found that regardless of the selected benchmark, the consensus model performed better than individual-expert models in most cases, despite the fact that the individual-expert models were trained specifically under the same human annotators that provided the individual benchmarks (Figure 2.7).

Both types of models were able to perform well prospectively for each A β class, with the consensus model outperforming individual-expert models on average in AUPRC and AUROC (by as much as 0.22 ± 0.18 for CAA AUPRC, Figure 2.7). For all classes and for all metrics, the consensus model evaluated on the consensus benchmark performed best. Furthermore, evaluating under a consensus benchmark as opposed to an individual benchmark resulted in either the same or better performance across models and across success metrics. Taken together the consensus model was robust to differing annotations of five different experts. On average, this model exceeded performance of individual-expert models even on their own benchmarks that they were trained to mimic, and also on the consensus benchmark (Figure 2.7).

These results held for each class and each metric, with the sole exception being the cored AUPRC. In this case, individual-expert models slightly outperformed the consensus model under the individual benchmark. This was consistent with the phase-one observation that cored-plaque labeling had more idiosyncrasies (Figure 2.1). Hence, we expected individual-expert models to be better performers on individual benchmarks for this class. Furthermore, evaluating individual-expert models under a consensus benchmark did not improve performance over the individual benchmark, likely for the same reason.

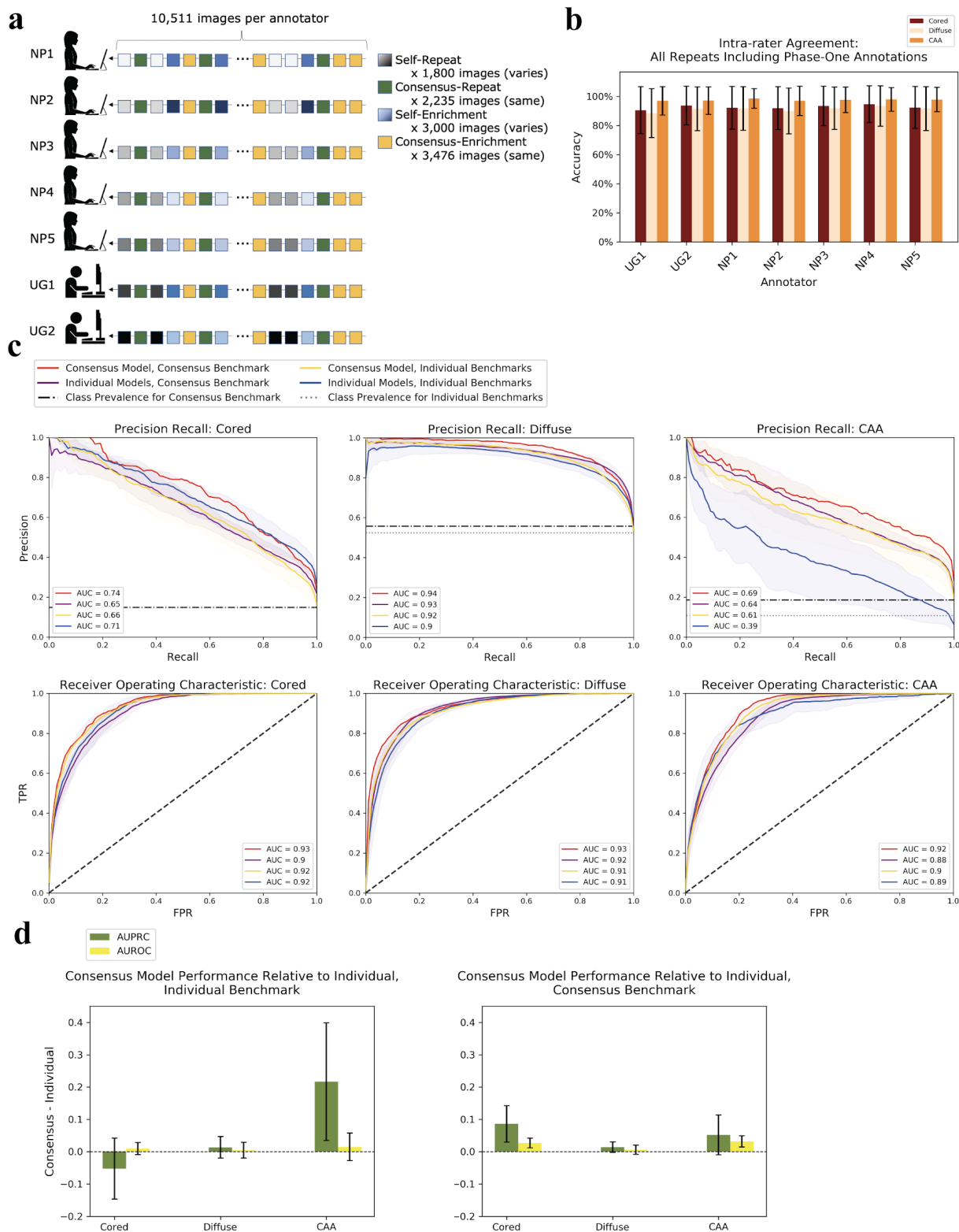


Figure 2.7: Models prospectively predict human annotation, with consensus models performing the most consistently.

(a) Schematic of the phase-two annotation protocol. These images fall under one of four categories: self-repeat, consensus-repeat, self-enrichment, and consensus-enrichment. See Methods for a detailed description of these categories. Each annotator is given the same order of image categories. Gradients of different colors indicate images from the same category. These gradients are depicted to reinforce the fact that each annotator received a different set of images for the self-repeat and self-enrichment categories. (b) Intra-rater agreement is measured as the accuracy at which each rater consistently annotates repeats of the same image (both self-repeat and consensus-repeat). We include image labels from phase-one in this intra-rater calculation. The x-axis indicates the annotator, and the y-axis indicates intra-rater accuracy. Accuracies are averaged over each set of repeated images. Novices achieved an average intra-rater agreement accuracy of 0.92 for cored, 0.90 for diffuse, and 0.97 for CAA. Experts achieved an average intra-rater agreement accuracy of 0.93 for cored, 0.92 for diffuse, and 0.98 for CAA. (c) Precision recall plots and receiver operating characteristic (ROC) plots for the consensus model versus the individual-expert models. Two different benchmarks are used—truth according to the individual annotators, and truth according to a consensus-of-two scheme. The shaded regions indicate one standard deviation in each direction centered at the mean. The consensus model evaluated under a consensus benchmark (red line) has no variation by definition. (d) Summarizes panel (c). Bar graphs depict the average performance of the consensus model minus the average performance of the individual-expert models (y-axis). Individual benchmark for figure left, consensus benchmark for figure right. Error bars show one standard deviation centered at the mean.

2.4 DISCUSSION

We developed a scalable, objective, and consistent way to automate and compile annotations of multiple neuropathology experts. Four points merit emphasis: 1) We automated individual and consensus annotations across datasets from five experts and three staining procedures from different institutions, advocating for generalizability; 2) When we prospectively validated these models in a research setting, a consensus model was robust under different benchmarks and provided a high-performance approach to automatic and standardized labeling; 3) The models were interpretable and showed increased pixel specificity with increased

expertise; and 4) Ensemble models generally performed better and were robust to intentionally randomized information. We note however that the experimental task differed from the daily practice of neuropathological annotation, where experts would commonly assess these objects using varying magnifications and localizations.

DL requires annotated datasets that define how to train a model and quantify its success. However, the classification of neuropathologies continues to evolve, with different experts having complementary focus areas. Consequently, two competing hypotheses regarding the best annotation strategy might be reasonable. In the first, a DL model trained on an individual might best leverage that single expert's unique intuition and decision-making procedures for edge cases. Conversely, in a second strategy, a consensus approach to annotations might instead leverage a wisdom-of-the-crowd logic to remove individual variance in a consolidated consensus model, as long as there is common underlying signal across the cohort. Depending on the consensus strategy, such a consolidated model could balance between strict-but-conservative expert consensus (the logical intersection) versus an inclusive but potentially overly-permissive consensus (the logical union).

In this study, we found that a balanced consensus-of-two strategy slightly favoring permissiveness performed better in nearly every model-training and model-benchmarking scenario (Figure 2.4). This was striking, as the consensus-trained model outperformed the individual-expert models even when unfairly assessed against those self-same individual-expert benchmarks (Figure 2.4). One might reasonably have otherwise expected individual-expert

models to better encode their particular expert’s pathological training and intuition, but this was generally not the case. We conclude from this observation that neuropathological labeling across thousands of independent annotations was sufficiently consistent across the cohort that a permissive consensus model computationally codified commonly-held expertise in neuropathological identification—despite noticeable differences in individual annotations (Figure 2.2). We are unaware of any previously published work that operated on a cohort of independent expert annotations at this scale. Whereas other DL pathology applications rely on the independent assessments of individual pathologists^{69,72,78,98–100}, we believe this is the first study to demonstrate that learning expert consensus within pathology provides robust and superior performance over learning individual assessments.

Encapsulating a consensus strategy into a model may be useful when an expert cohort or adequate labeling time are unavailable. For practitioners who would prefer individualized models or have unique research use-cases, the expert models effectively encoded individual intuition; thus they are limited in their application only by image-data preparation and computational-resource availability. Indeed, the purpose of phase-two was to prospectively assess whether these models could enrich for rarer but important neuropathologies within an unexplored dataset of $n=275,879$ candidate neuropathologies from new patients (Figure 2.7). This represented a clinical scenario in which the pathologist wishes to rapidly curate sparse pathology examples, or to have a DL model act as an individualized assistant enriching for important but infrequent

morphologies, which may improve understanding of these phenotypes and provide opportunity for overcoming sparsity challenges.

In clinical and research settings, we cannot rely on DL models solely as “black boxes”¹⁰¹ if we wish to have a common understanding of ground truth, guide diagnosis and other high-stakes decisions, or improve on subtle oversights of the current models. Leveraging the field of model interpretability, we thus employed a saliency mapping technique called guided grad-CAM⁸⁵ to determine which pixels in the images contributed most strongly to classification. The models learned directly from the image data and became more focused in their feature detection as experience of the labeler increased from novice to expert consensus (Figure 2.5). These results suggest potential for application in the training of new pathologists, using models to identify critical image features and thereby visually illustrate for new trainees the phenotypes relevant for labeling. A future direction is to explore this new opportunity of DL facilitating human learning, as opposed to the more common framework of humans facilitating DL, ultimately leading to a hybrid-feedback loop through active learning¹⁰².

The ensemble models may provide an opportunity of weighting individualized model input to tailor a specific annotator’s expertise to specific tasks or cases. Subsequent to a prediction, we can inspect the interplay of contributor-specific weights (Supplemental Figure 2.9). In this way, individualized contributions encoding complementary and situation-specific expertise do not get overwhelmed as they might otherwise under a simple vote averaging. Although we do not expect labelers to intentionally introduce false information, the ensembles’

robustness to unreliable annotations encourages assembling numerous and diverse labelers to facilitate accurate and learnable labeling.

Several limitations to this study inform its practical adoption. Researchers may not have access to necessary computational resources or training data for DL. However, cloud computing resources are becoming more accessible, and open-data sharing is also possible through data-hosting services. We are openly releasing this study's annotated datasets and trained models (<https://github.com/keiserlab/consensus-learning-paper>). Turning to limitations in interpretability methods, we note that guided grad-CAM devolves to being a visual edge-detector in some settings¹⁰³. However, this was not the case in our study because saliency maps calculated on the same input image but different A β class substantially differed (Supplemental Figure 2.10). Nonetheless, a future direction would be exploring a range of independent saliency mapping techniques^{104–106}. Finally, prospective effectiveness outside this cohort remains open to exploration. Whereas this study was consistent with our earlier work leveraging one expert annotator⁶⁹ and its independent application at another institution⁷⁰, we expect larger cohorts and institutionally-diverse datasets (including persons from multiple socioeconomic backgrounds) with different neuroanatomic areas and staining techniques will facilitate more comprehensive standards in neuropathology. For instance, performance for CAA with 6E10 was lower than for other stains (Supplemental Figure 2.4). This could be due to its smaller representation in the test set (Supplemental Figure 2.11). Although this study's cohort of five experts was institutionally diverse, it could be improved by capturing greater variability from the broader community.

Despite these limitations, the models accurately learned the annotation practices of five experts and of their shared expertise, indicating the method's generalizability.

We imagine this neuropathology study might apply also to anatomic pathology and other areas of medical research. Any medical discipline that leverages human expertise could benefit from taking advantage of expert diversity and consensus to make accurate diagnoses. Models and annotated datasets developed and shared across the community could be progressively refined as ever-improving metrics and deployable tools of shared ground truth. Whereas we focused on automatic image annotation, the concept may generalize to other domains and data types where expert annotation is crucial. Although a consensus-of-two experts was best in this particular study, this may not hold for other studies or for a different cohort or pathology, and indeed improved ensemble modeling techniques may be the strongest approach (Figure 2.6). However, the resilience and robustness of the consensus model indicated that training from a consensus was better than relying on individual assessments, despite high expert intra-rater reliability even over a half-year gap (Figure 2.7). We hope that developing and openly sharing consistent, accurate, and automated DL methods and their datasets can facilitate standardization and accelerate quantitative pathology as a freely available community resource.

2.5 METHODS

2.5.1 Slide Preparation

43 WSIs of the temporal cortex were collected from 3 different sites: 17 from UC Davis, 16 from University of Pittsburgh, and 10 from UT Southwestern. See Supplemental Figure 2.1 for associated clinical data. Slides were derived from formalin fixed paraffin embedded sections and stained with an antibody directed against A β . UC Davis used the A β 4G8 antibody, University of Pittsburgh used an NAB228 antibody, and UT Southwestern used a 6E10 antibody. All WSIs were imaged on an Aperio AT2 at 20X magnification.

2.5.2 Data Collection and Annotation (Phase-one)

We selected 29 of the 43 WSIs at random for phase-one, and used the remaining 14 for phase-two. We color-normalized the WSIs according to the method presented in Reinhard⁸⁶. Each WSI was uniformly tiled to 1536 x 1536 pixel non-overlapping images. After tiling, we applied a hue saturation value (HSV) filter and smoothing technique to detect candidate plaques, using the python library openCV. We used different HSV ranges for the different stain types as follows: 4G8 HSV = (0, 40), (10, 255), (0, 220); 6E10 HSV = (0, 40), (10, 255), (0, 220); NAB228 HSV = (0, 100), (1, 255), (0, 250).

Each candidate was center cropped to provide a 256 x 256 pixel image. This process yielded 526,531 images for phase-one. We randomly selected 20,099 images from this set to be

annotated by our seven annotators. For the phase-two images, the center cropping yielded 275,879 images.

For the first phase of annotation, these 20,099 images were shuffled and placed into a fixed order, then uploaded to an Amazon instance web portal for independent annotation by seven different people. Five of them were professionally trained experts (E.B., B.N.D., M.E.F., J.K.K., K.E.M.), while the remaining two were undergraduate novices (J.A. and S.D.) with no formal training in neuropathology. They annotated the same 20,099 images in the same exact order in a multi-label classification task using a rapid-keystroke based custom annotation tool (first described in Tang et al.⁶⁹, with current code released in <https://github.com/keiserlab/consensus-learning-paper>). Each 20X image had a bounding box, and the annotators were instructed to label any and all of A β pathologies found within the box. They had the option to label the pathology as any combination of three classes: cored, diffuse, or CAA. They also had the option to mark an image as negative, flag, or not sure (see Supplemental Figure 2.2). We did not use these alternative markings for constructing our final image labels. Hence, any A β class that was marked as positive was recorded as a positive annotation for that class. Any classes left unmarked were recorded as negative. There were instances in which an annotator marked negative and also marked a positive annotation for any of the three A β classes. Whenever this occurred (which was rare, Supplemental Figure 2.2), we labeled the image as positive for the amyloid pathology indicated.

This process yielded seven different annotation sets (one for each person) for the 20,099 images that were annotated. From the five expert annotated sets, we constructed consensus-of- n annotation sets (from $n=1$ to $n=5$) for the same 20,099 images. For a given image i and A β class c , the new consensus-of- n annotation was recorded as positive if any n experts marked image i as positive for class c , else the image was recorded as negative.

2.5.3 Training and Evaluation of DL Models

Of the 29 WSIs for phase-one, we used 20 WSIs for training and 9 separate WSIs for the held-out test set. The 20,099 phase-one images were divided into a 67% train and 33% held-out test split. Of the 67% training data, we performed fourfold cross-validation, keeping each fold's image set consistent across all training and evaluation protocols such that each fold always had a distinct set of images that were not present in any other fold. We trained one model for each fold of the cross-validation, resulting in four models for each annotation set. Due to the large imbalance between the different A β classes, we performed class balancing during training. We calculated the ratios of diffuse to cored ($r1$), and diffuse to CAA ($r2$), and replicated any image with a cored plaque present a total of $r1$ times, and any image with a CAA plaque present a total of $r2$ times. Models were trained to perform multi-class classification of the three A β classes, taking a 256 x 256 pixel image as input and returning three floating point predictions (one for each class) as output.

For each of the twelve annotation sets (seven from our individual annotators, and five from our constructed consensus annotation sets), we trained four CNNs (one CNN for each fold). We trained models for 60 epochs, and saved the best model state that resulted in the highest cored AUPRC performance over the validation set. We used an Adam optimizer with a learning rate of 0.001, and a weight decay of 0.03. We used a multi-label soft margin loss as our loss function. During training, the input images were transformed with a random horizontal flip, a random vertical flip, a random 180 degree rotation, a color jitter, and a random affine transform so that the model would learn to generalize across different inputs. Images were normalized to have zero-mean and unit-variance prior to entering the model.

We converted each image annotation into a floating point representation, such that the binary labels given by the annotators were converted to a more specific continuous value, using information from other annotated bounding boxes that may be present in the image. During the annotation process, each 20X image possibly contained regions that were previously annotated in a different query with a different bounding box. Hence we had instances in which parts of bounding boxes or multiple bounding boxes from different annotations (from the same annotator) were present in a single 20X image. For each image i , we used the labeled bounding box information that the annotator provided in order to create a floating point representation for each class c . The floating point representation of c was the total sum of fractions of positively labeled bounding boxes for class c that exists within i . An image potentially had more than one positively labeled bounding box within i , or many fractions of positively labeled bounding boxes

within i . Therefore, the floating point representation was possibly greater than 1.0, but must have been greater than or equal to 0.0. Finally, image i was considered positive for class c if the floating point representation of class c was greater than 0.99. This final binary label was used for training and evaluation.

During evaluation of a model, no transforms were applied except for normalization. For each type of model, we evaluated our models by taking each model from each fold and evaluating it on the held-out test set. The reported metrics were the average over these four evaluations. For example, the reported phase-one results for the consensus-of-two model was the average performance of all four consensus-of-two models, one for each of the four cross-validation folds, each evaluated on the held-out test set. The full source code used for training and evaluation can be found at <https://github.com/keiserlab/consensus-learning-paper>. We used the PyTorch library¹⁰⁷ for all training and evaluation.

2.5.4 Assessing Consensus Model and Benchmark Superiority

When assessing the consensus models versus the individual-expert models, we applied four different benchmark schemes using only the held-out test set annotations, and calculated the average performance of the consensus models minus the average performance of the five individual-expert models. “Self” means that each model was evaluated by its own benchmark (i.e. a model trained under A’s annotation set is evaluated according to the annotations in A’s held-out test set). For instance, a consensus-of-two model with benchmark “Self” was evaluated

according to how well the predictions matched with the held-out test labels provided by the consensus-of-two annotations. This was done for each of the four consensus-of-two models (one for each fold), and averaged. “Consensus benchmarks” means that all models were evaluated on how well their predictions matched on average with each of the five consensus benchmarks (consensus-of- n from $n=1$ to $n=5$). We averaged performance across these five consensus benchmarks. “Individual benchmarks” means that all models were evaluated on how well their predictions matched on average with each expert benchmark. We averaged performance across these five expert benchmarks. “All benchmarks” is evaluating a model across the five consensus benchmarks and five expert benchmarks, and then averaging these results. For the consensus-of-two superiority results, we performed the same procedure, except only the consensus-of-two model and consensus-of-two benchmark were evaluated to represent the consensus strategy.

We used a two-sample, one-sided Z-test to assess statistical significance. The null hypothesis was that the average of the consensus performance and the average of the individual experts’ performance were the same. The alternative hypothesis was that the average of the consensus performance was greater than the average of the individual experts’ performance. P-values were reported using a standard lookup table. The sample sizes for each of the two samples (consensus and experts) are as follows: 20 for “self”; 100 for “consensus benchmarks”; 100 for “individual benchmarks”; and 200 for “all benchmarks”. Each sample was a model’s performance on a specific benchmark.

2.5.5 Model Interpretability

We used guided Grad-CAM to obtain the saliency maps from the different models. The library we used can be found here: <https://github.com/utkuozbulak/pytorch-cnn-visualizations>. For analyses requiring image binarization, we chose incremental pixel thresholds such that at threshold t , any signal less than t is assigned 0 value, while anything greater than or equal to t is assigned a maximal pixel value of 255. To calculate SSIM between the novice CAM and the consensus-of-two CAM, we used a standard SSIM function from the skimage library.

For analyzing the CAM patterns of the novice CAMs versus the consensus-of-two CAMs, we constructed a subtraction map of the binary consensus CAM minus the binary novice CAM for each 256 x 256 pixel image. Each pixel of this map is classified as one of three classes: signal (ON) in the novice CAM and no signal (OFF) in the consensus CAM, or no signal in the novice CAM and signal in the consensus CAM, or a match between the two. For the CAM subset analysis, we calculated both the total fraction of the binary consensus CAM that was activated in the binary novice CAM at the same corresponding pixel locations, and also the total fraction of the novice CAM that was activated in the consensus CAM at the same corresponding pixel locations for each pixel threshold.

2.5.6 Ensemble Training and Evaluation

For each of the four folds of the cross-validation, we linked each of the five individual-expert CNNs trained from that fold with a sparse affine layer. The individual-expert CNNs were

frozen such that no backpropagation occurred in the individual-expert CNNs. The only weights that were updated were the spare affine weights, which weight each individual-expert CNN's final class output.

For each annotation set (five expert sets, and five consensus annotation sets), four ensembles (one for each cross-validation) were trained using the same training data that their constituent CNNs used for training. Holistically, having four cross-validation folds, five experts, and five consensus schemes, resulted in 40 ensemble models, each independently trained to reproduce a specific annotation set. Ensemble training occurred for 60 epochs, with the same hyperparameters that were used to train the single CNNs.

After training, each ensemble model was evaluated on the images from the held-out test set, and on every benchmark (five expert benchmarks, and five consensus benchmarks). We aggregated ensembles that were trained to mirror the same annotation set but belonged to different cross-validation folds. Hence, every permutation of 1) ensemble model (aggregated across cross-validation folds) and 2) benchmark was assigned an average AUPRC score. This resulted in 100 average AUPRC scores (10 aggregated ensembles, 10 held-out test benchmarks). To assess ensemble superiority over single CNNs for a given benchmark, these 100 average AUPRC scores were compared to the 100 average AUPRC scores of the single (non-ensemble) CNNs. This resulted in 100 comparisons of ensemble models versus single CNNs, with both model types evaluated on equivalent benchmarks and an equivalent image set.

For ensembles that contained a random labeler CNN, we trained a separate CNN on a random annotation set that kept the same class ratios of cored, diffuse, and CAA as the ones present among the five expert annotations. This ratio was determined by averaging the class ratios of each of the five expert annotation sets. We then linked this CNN trained on random labels with the five professional CNNs using a learnable sparse affine layer. Likewise, for ensembles with five random labeler CNNs, we trained five independent CNNs on five different permutations of the randomly labeled annotation set. We then linked these five random labeler CNNs with the five professional CNNs using a sparse affine layer. For both the ensemble with a single random labeler, and the ensembles with multiple random labelers, we used the same training procedure as the normal ensembles (i.e. ensembles without any random labeler present).

In evaluating the performance of ensembles with any number of random labeler(s) present, we performed the same evaluation procedure as for the normal ensembles. For comparing performance between ensembles with any random labeler versus performance of normal ensembles, we compared the final average AUPRC values for these two model types and for each of the ten benchmarks. This resulted in 100 comparisons between normal ensembles and ensembles with a random labeler, and 100 comparisons between normal ensembles and ensembles with five random labelers. In every comparison, everything was kept constant except for the choice of model architecture.

2.5.7 Setup of Phase-Two Prospective Validation and Analysis

For phase-two, we used 14 WSIs that were completely separate from the 29 WSIs from phase-one. We performed a second round of human annotation (phase-two) to validate the models and assess intra-rater reliability. All annotations were performed on the same web interface and by the same expert and novice annotators from phase-one. Annotators saw four different categories of images: self-repeat, consensus-repeat, self-enrichment, and consensus-enrichment. For the self-repeat images, which are simply repeats of a subset images the annotators already saw in phase-one, we selected a total of 600 images. We first selected all of the images that were marked as positive for CAA. CAA was rarely annotated during phase-one, and we wanted to include all of them. Once all of the CAAs were included, we selected a random subset of images that were marked as positive for cored until we had 400 images. If there weren't enough cored positive images to make up 400, then we took as many as possible. For the remaining images we randomly selected diffuse positive images to make up a total of 600 images. We enforced having no duplicates. We then triplicated and randomly rotated these 600 images, and shuffled the resulting 1,800 images. This entire procedure of obtaining the self-enrichment image set was done independently for each expert and each undergraduate novice, resulting in different sets of self-repeat images given to different annotators.

Likewise for the consensus-repeat set, we took images exclusively from the set of phase-one images that the annotators already labeled. For each of the three A β classes, we randomly selected 250 images that were positive for this class according to a consensus-of-two strategy.

From this list of images, we removed duplicates, which occurred because some images were positive for multiple A β classes according to a consensus-of-two. This resulted in a total of 745 images, which were then triplicated and randomly shuffled, resulting in our final consensus-repeat set of 2,235 images. This set was given identically to all of the annotators during phase-two.

For the self-enrichment set given to annotator A, we used the CNN trained on individual A's annotation set to select images that the model predicted as having a minority A β plaque present. We chose to use the models from fold three of the cross-validation for enrichment. All images came from a held-out set of images that the annotators did not see during phase-one of annotation. This held-out set consisted of 275,880 images total. From this held-out set, we randomly selected 800 images with a model prediction threshold > 0.90 for the cored class. Next, the 275,880 images were sorted according to the model's CAA prediction confidence, and the top 800 images with highest CAA confidence were included. After collecting this set of 1,600 images, we included all of the image neighbors that had at least a 20% bounding box overlap of a plaque with any of these 1,600 images. Afterwards, we randomly shuffled this resulting list, and took a random subset of 3,000 images to use for our final self-enrichment set. This procedure was repeated for each of our seven human annotators. We calculated the overall intra-rater agreement for each annotator by averaging the accuracy of how consistent each annotator was over each set of replicated images. Each replicated set had four total images (one annotation from phase-one, and three annotations from phase-two).

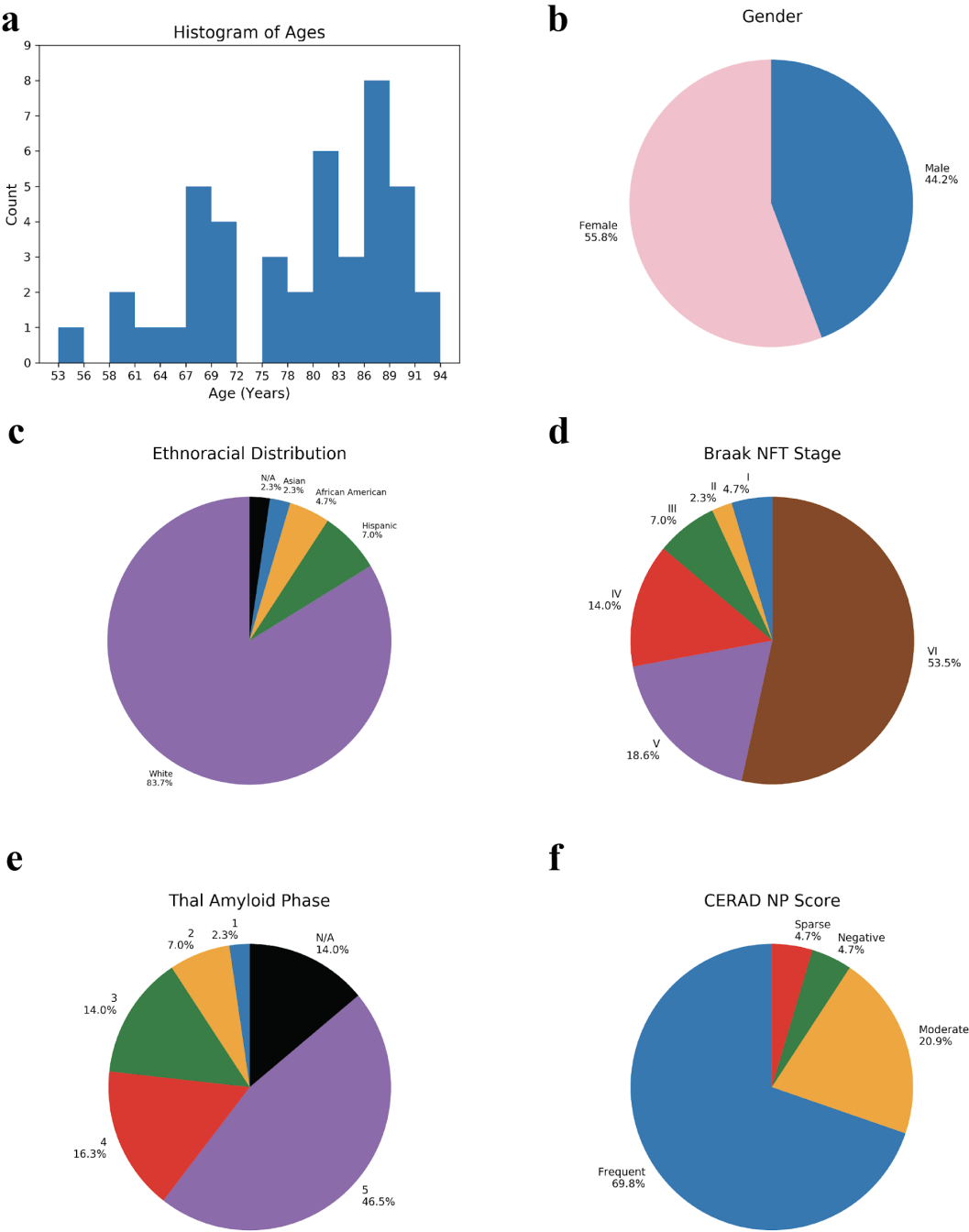
For the consensus-enrichment set, we randomly pulled 750 images that the consensus-of-two model predicted as positive for cored, and an additional 750 images that the same model predicted as positive for CAA. From this resulting set of 1,500 images, we found their image neighbors that had at least a 20% bounding box overlap of a plaque with any of the 1,500 images. We randomly selected from these neighbors until we had a total of 3,476 images for the final consensus-enrichment set. If an image achieved high rank by both self-enrichment and consensus-enrichment, it was assigned with equal probability to either self or consensus.

These four image sets were given to the annotators, such that the ordering of the image category was randomly shuffled, but fixed and identical among each annotator. Upon completion of annotation, we analyzed the model's prospective performance to match annotations given during phase-two. We stipulated two different benchmarks that could be derived from these new annotations: the individual benchmark, which simply assigned what the individual annotator labeled as the truth labels, and the consensus benchmark, which used the consensus-of-two strategy to assign truth labels. We used the individual CNN models to make class predictions for the 10,511 images of phase-two, independent of the human annotators and their labels. We did the same for the consensus-of-two model, and made class predictions for each of the 10,511 images. Model predictions were completely hidden from the annotators, as well as any experimental details of how the images were selected. For analysis on phase-two data, we define an individual-expert model as a model trained on one of the expert's annotation sets from phase-one, and a consensus model as the model trained on the consensus-of-two annotation set from

phase-one. All label data used to assess model performance during phase-two came from this second phase of annotation, and we did not use any phase-one annotation labels.

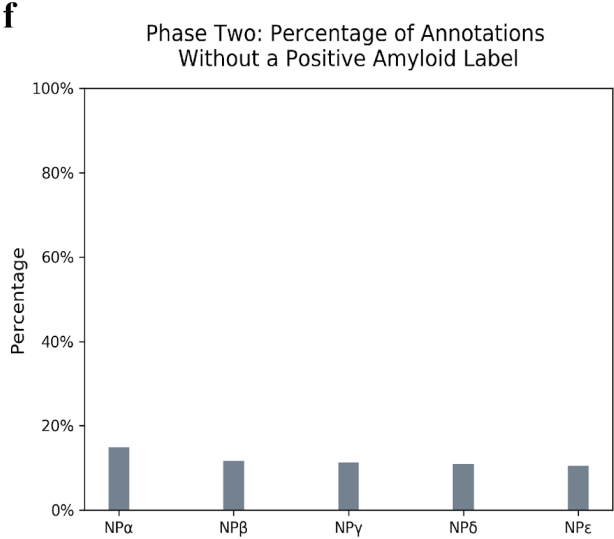
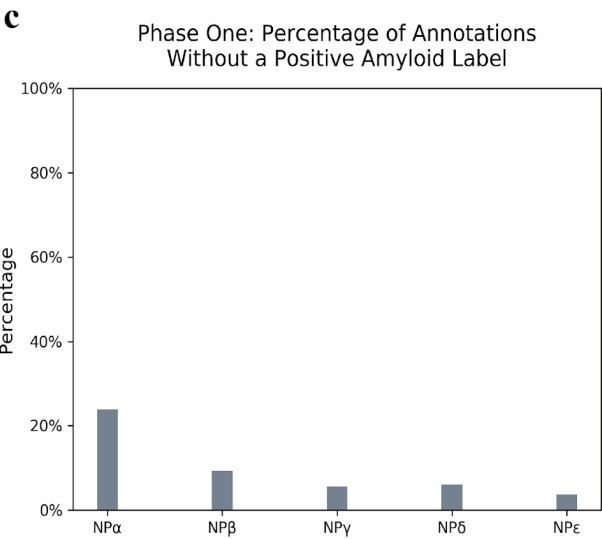
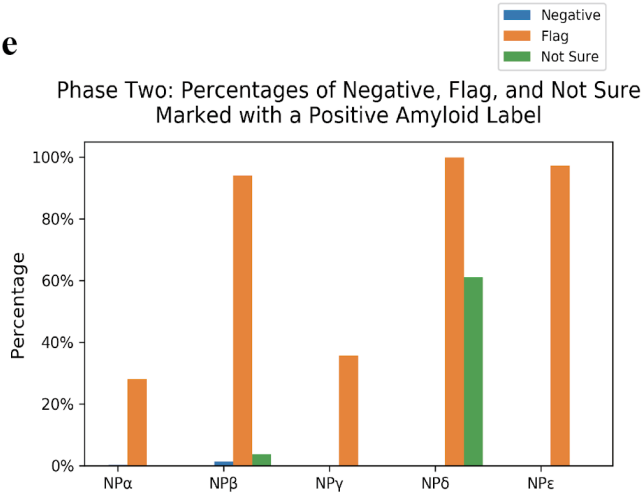
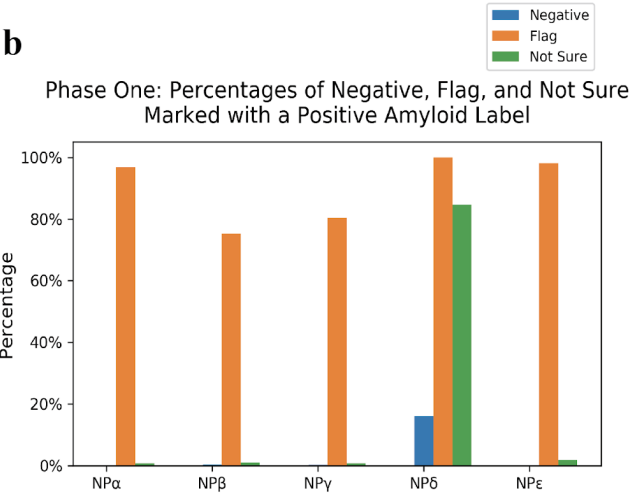
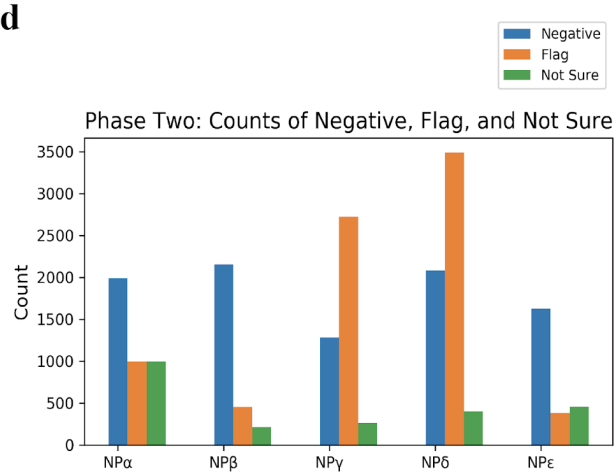
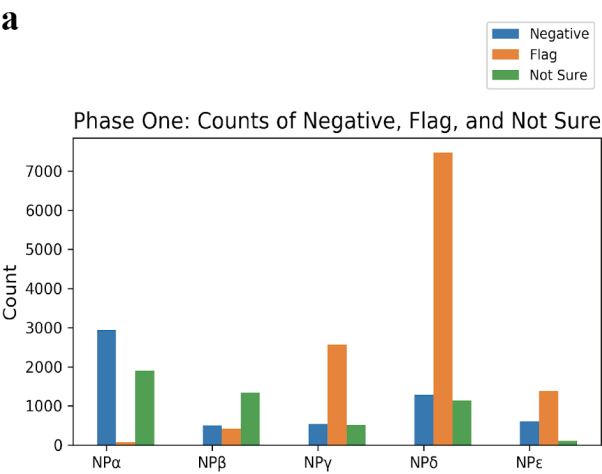
To assess performance of individual-expert models under the individual benchmark, we compared the predictions of each individual-expert model trained on annotator A with the labels that annotator A gave during phase-two (undergraduate novices were excluded from this analysis). This was repeated for each expert annotator model and the five results were averaged. To assess performance of individual-expert models under a consensus-of-two benchmark, we compared the individual-expert model's predictions with the labels provided by a consensus-of-two scheme. This was done for each professional annotator model and the five results were averaged. For the consensus model and individual benchmark case, we compared the consensus model's predictions with the labels provided by annotator A. This was repeated for each of the five professional annotator labels, and these five results were averaged. For the consensus model and consensus benchmark case, we compared the consensus model's predictions with the label set provided by a consensus-of-two strategy. There was only one consensus-of-two model and one consensus benchmark, resulting in no variability and no averaging in this case.

2.6 SUPPLEMENTAL FIGURES



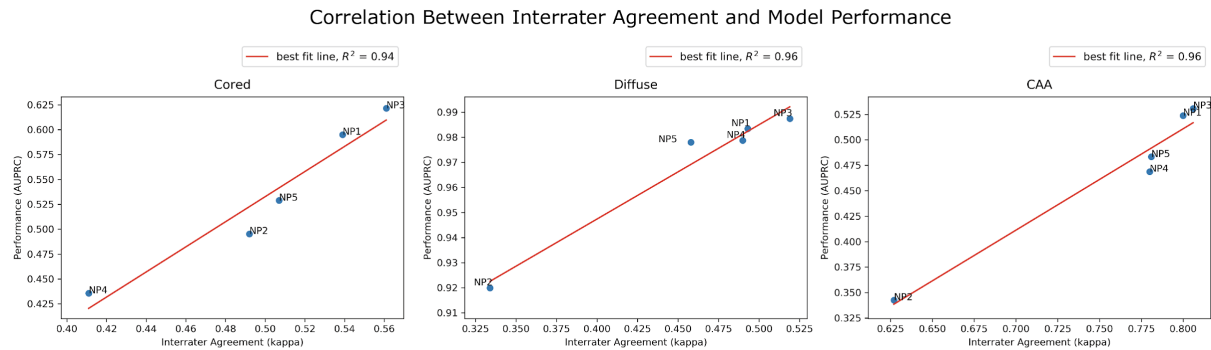
Supplemental Figure 2.1. Demographics of the 43 patients from all institutions.

(a) Histogram of ages (b) Pie chart of gender distribution (c) Pie chart of racial distribution (d) Pie chart for scoring guidelines for Braak NFT Stage (e) Thal Amyloid Phase and (f) CERAD NP Score, set forth by Montine et al⁶³.



Supplemental Figure 2.2: Negative, flag, and not sure annotations for both phase one and phase two.

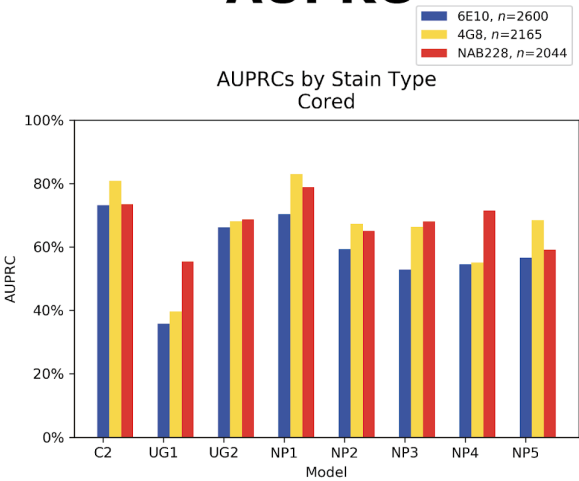
For all panels, we further obfuscate the neuropathologist identities by assigning each with a randomly chosen but fixed Greek letter. (a) The number of times that an annotator checked one of the three alternatives: “negative”, “flag”, and “not sure” for phase one. There are a total of 20,099 annotations for each neuropathologist during phase one. (b) Of the three alternatives, the fraction of annotations in which the annotator also marked any of the amyloid plaques as being present during phase one. The cases in which annotators marked “negative” and also identified at least one amyloid plaque were sparse. In such erroneous cases, we chose to accept the positive amyloid marking(s) over the “negative” marking. For the majority of cases in which annotators marked “flag,” the annotators also identified at least one amyloid plaque as well. For these cases, we constructed the final image labels by using the positive amyloid marking(s), essentially disregarding the “flag” marking. For the cases in which an annotator marked “not sure” and also identified at least one amyloid plaque as well, we likewise used the positive amyloid marking for our final image labels. (c) Fraction of phase one annotations that did not receive any positive label for any of the three amyloid classes. (d) Counts of the three special cases for phase two. There were a total of 10,511 images. (e) Proportions of the three special cases that were also marked with a positive amyloid label during phase two. (f) Fraction of the phase two annotations without a positive amyloid label for any class.



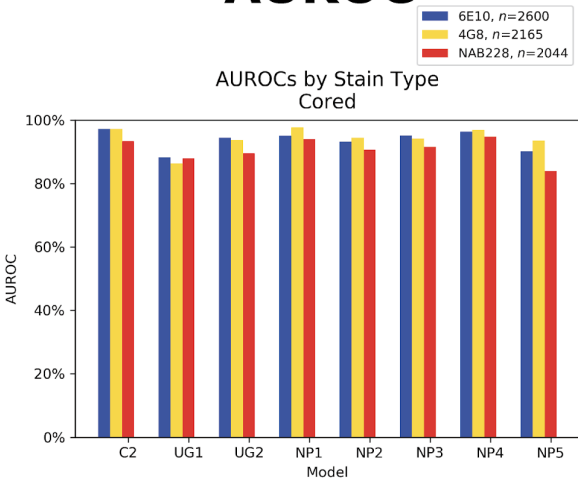
Supplemental Figure 2.3: Models capture interrater agreement patterns among experts.

For each expert E , we calculated an average kappa coefficient between E and every other annotator. We also calculated the average AUPRC performance for expert E . We took every other expert model that was not trained with E 's annotation set, and evaluated those models using E 's annotation benchmark, and averaged the results. This yielded a single kappa score (x-axis) and a single AUPRC score (y-axis) for each expert E . Correlating these two scores, we saw human interrater agreement reflected in the ML models. We calculated R^2 correlation from the interpolated line of best fit. There were strong R^2 correlations (0.94 for cored, 0.96 for diffuse, and 0.96 for CAA) between an individual's kappa coefficient and how well other models acting as annotators agreed with this individual's benchmark.

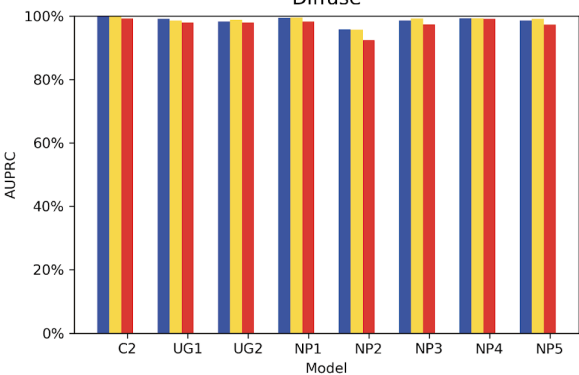
AUPRC



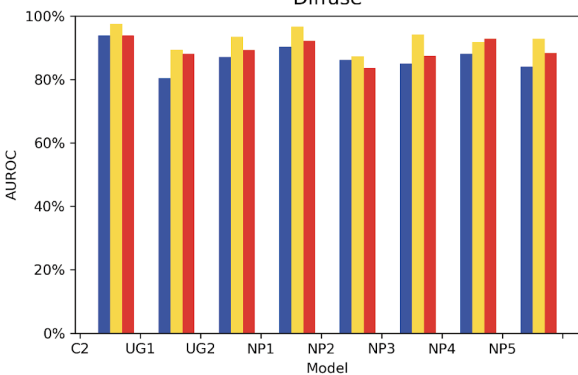
AUROC



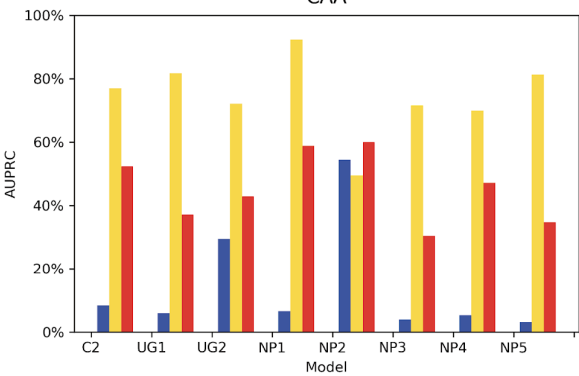
AUPRCs by Stain Type Diffuse



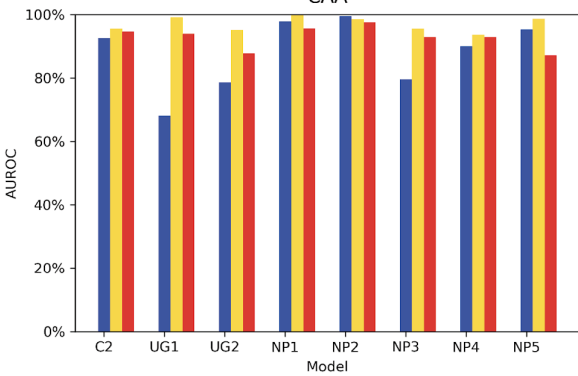
AUROC by Stain Type Diffuse



AUPRCs by Stain Type CAA

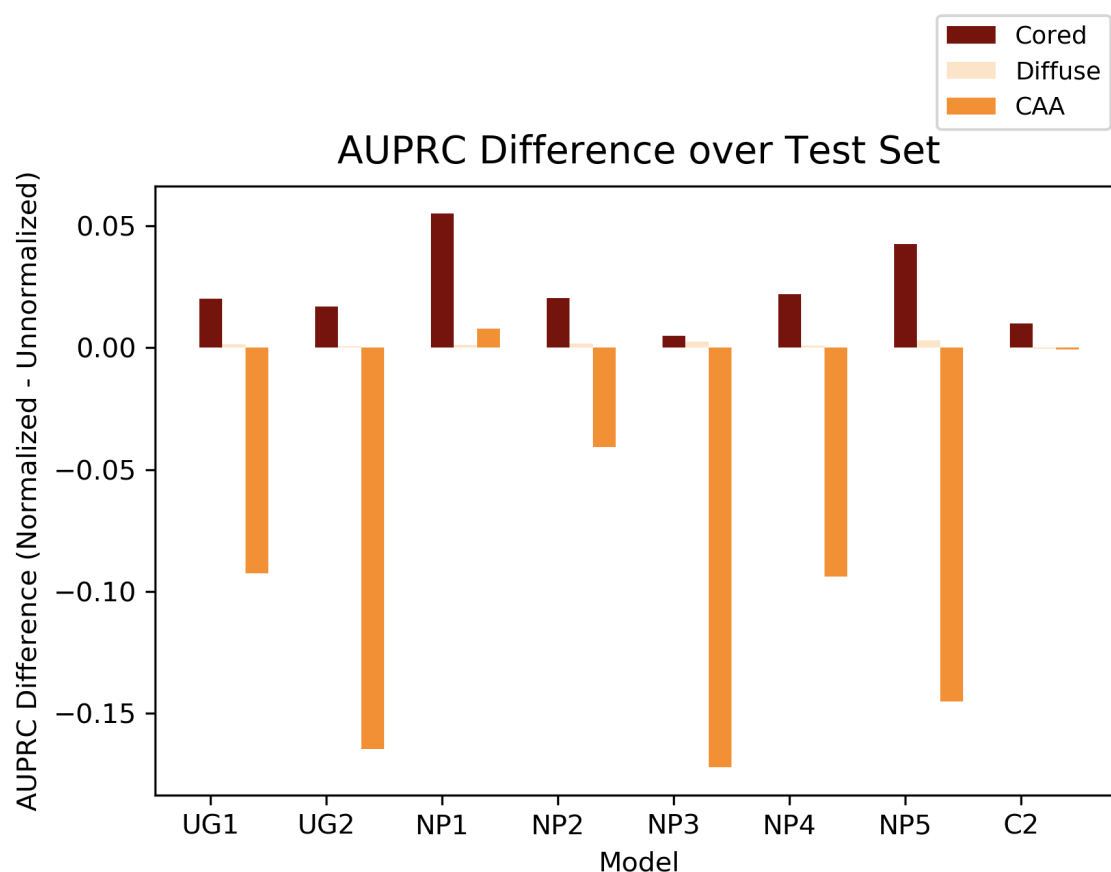


AUROC by Stain Type CAA



Supplemental Figure 2.4: Model performance by stain type for the three A β classes.

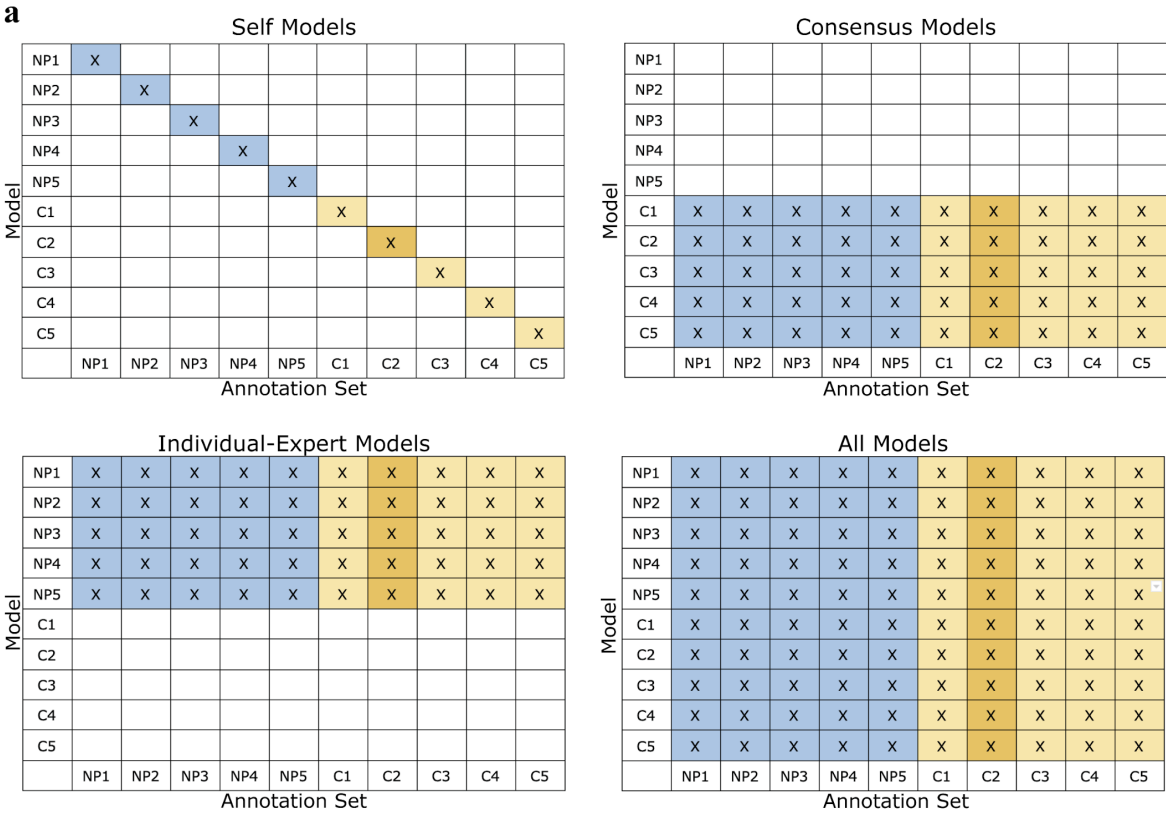
The left half of the figure is AUPRC, while the right half is the corresponding AUROC. The x-axis designates the model. “C2” refers to the consensus-of-two model. The y-axis designates the performance metric over the held-out test set. There are three different stain types, each from three different institutions: 6E10, 4G8, and NAB228. The number of images (n) per stain type in the test set are displayed in the figure legend. Both AUPRC and AUROC performance are consistent across stain types. The exception is AUPRC for CAA with the 6E10 stain. In this case, representation in the test set is small (Supplemental Figure 2.11). Since each stain came from a different institution, effects from stain versus patient cohort cannot be disentangled.



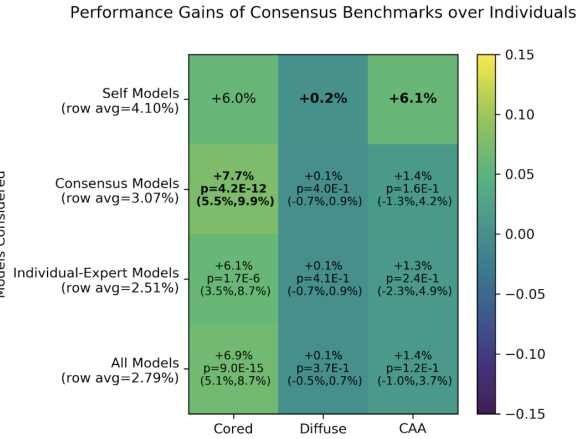
Supplemental Figure 2.5: Model evaluation differences on color-normalized versus unnormalized test images.

Each model (x-axis) is evaluated on its own benchmark's test set. "C2" is the consensus-of-two model. The difference in AUPRC (y-axis) is calculated when evaluating the model on color normalized images versus unnormalized images. Results are stratified by amyloid class. Most models have higher performance for the CAA class when we do not color-normalize the images.

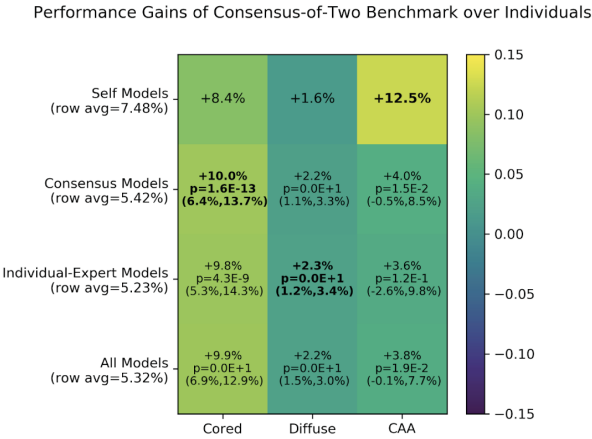
a



b

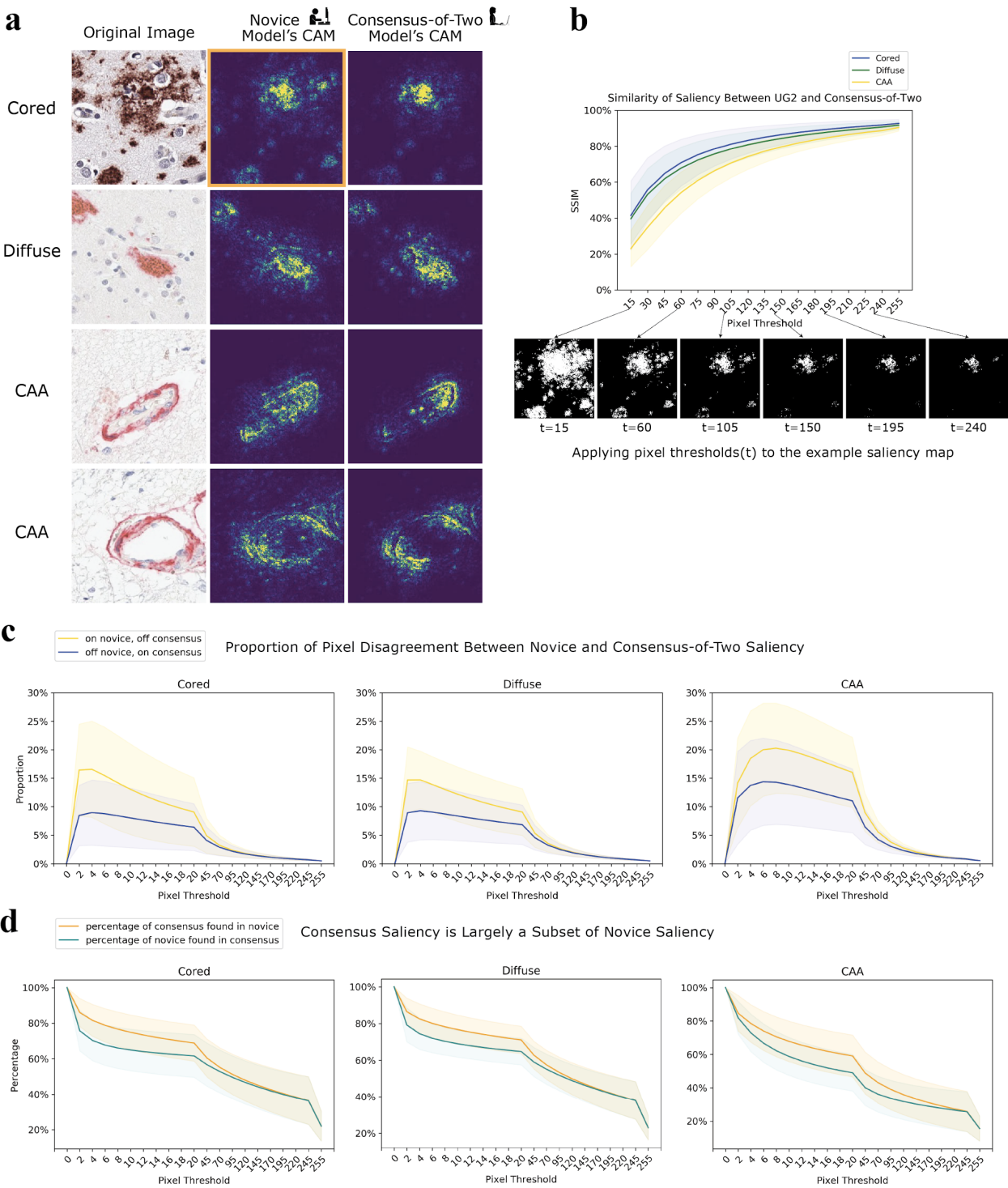


c



Supplemental Figure 2.6: Consensus benchmarks facilitate higher model performance.

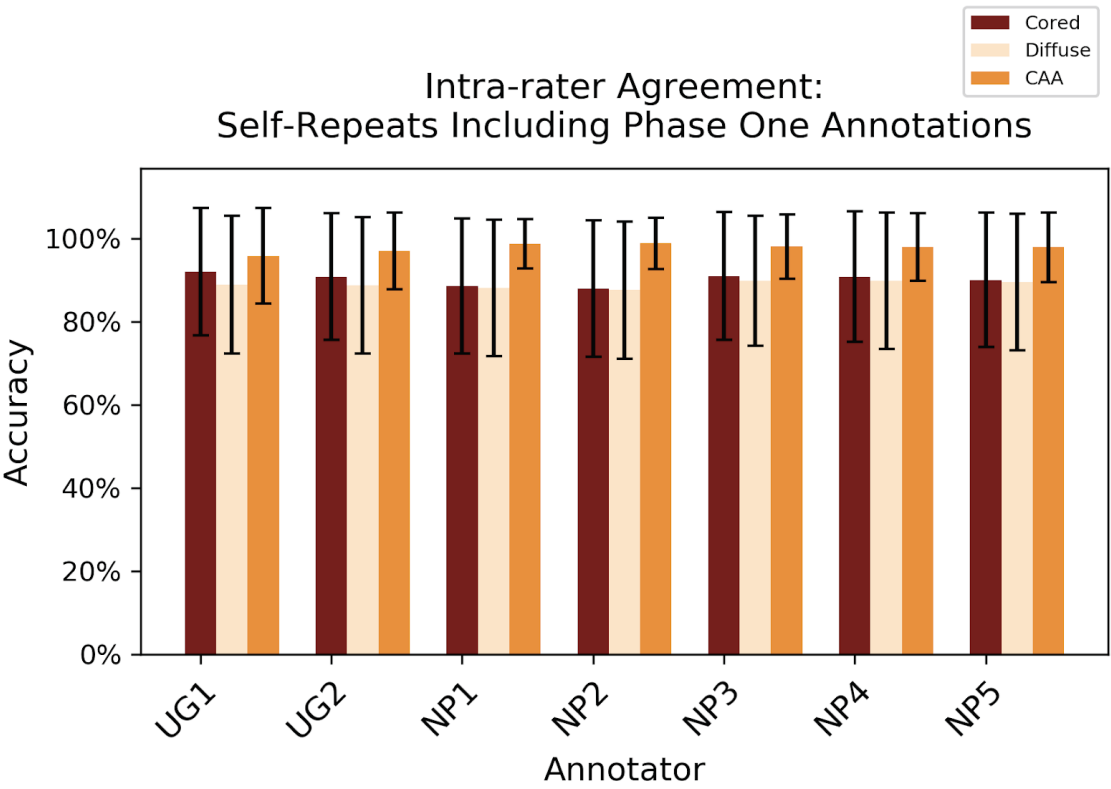
Corresponds to Figure 2.4, however this is comparing consensus versus individual-expert benchmarks as opposed to comparing models. (a) We compared the consensus benchmarks with the individual-expert benchmarks by evaluating subsets of models on every annotation benchmark. We compared the average performance of the models on the consensus-of- n benchmarks versus the average performance of the same models on the individual-expert benchmarks. The first and most internally-consistent model scheme evaluated each individual-expert model according to the labels of its annotator (called “self models”). For consensus models, the annotation labels corresponded to labels derived from the matching consensus-of- n strategy. “Consensus models” evaluated every consensus model on every annotation set. “Individual-expert models” evaluated every individual-expert model on every annotation set. “All models” evaluated both consensus models and individual-expert models on every annotation set. The y-axis indicates the model, and the x-axis indicates the annotation set used to evaluate the model. The average AUPRC of the blue region (individual-expert benchmarks) is compared with the average AUPRC of the gold region (consensus benchmarks). The consensus-of-two is colored dark-gold for emphasis. (b) Depicts a heatmap of all consensus benchmarks vs individual benchmarks. We calculated p-values of the comparisons using a two-sample Z-test (Methods). P-values for the self-benchmark are not included because the sample size ($n=20$ comparisons) is not large enough to assign significance. 95% confidence intervals shown in parentheses. The x-axis shows the $A\beta$ class being evaluated, while the y-axis indicates the models that are evaluating the benchmark. We bolded the highest performance differential for each $A\beta$ class. Regardless of what set of models we used, the annotations derived by a consensus-of- n strategy allowed for greater average model performance across all $A\beta$ classes. (c) Depicts just the consensus-of-two benchmark versus the individual-expert benchmarks. For this consensus-of-two benchmark evaluation, only dark-gold regions in (a) corresponding to the consensus-of-two benchmark are compared to the blue region.



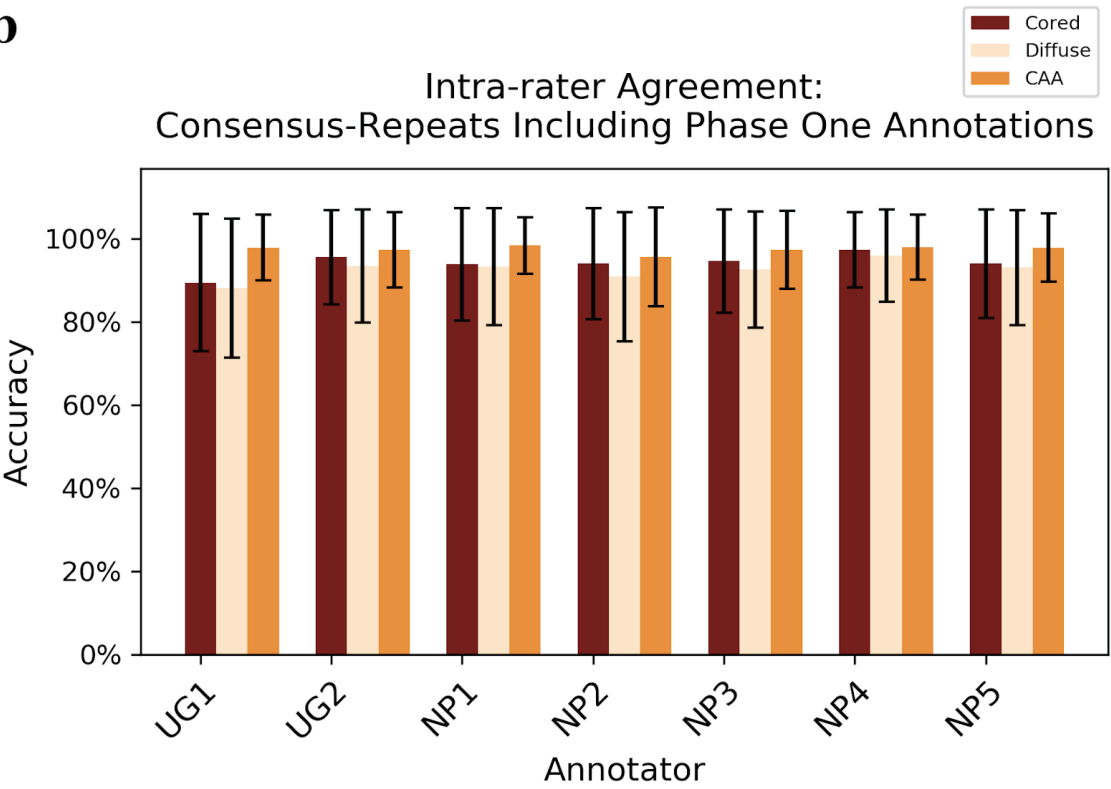
Supplemental Figure 2.7: CAM analysis of UG2.

Corresponds to Figure 2.5, but for the second undergraduate novice. (a) Novice CAMs are more diffuse than expert CAMs. The original image (leftmost column), the CAM of the novice model trained on UG1's annotations (middle column), and the CAM of the consensus-of-two model (rightmost column). CAMs are plotted with a false-color map such that bright regions correspond to high intensity regions with high salience. (b) Although expert and novice CAMs differ, they converge on the same pixels. We progressively assess the structural similarity index (SSIM) between novice CAMs and consensus-of-two CAMs across the entire test set of images. The CAMs show the most similar salience by SSIM (y-axis) at the highest pixel thresholds as we increment the threshold (x-axis) used to binarize the images before comparison. Binarized examples are shown of one CAM from (a) (boxed in orange). (c) Comparing the novice CAMs and the consensus-of-two CAMs, we classify each pixel location into two categories: ON in the novice CAM and OFF in the corresponding consensus CAM (yellow), or OFF in the novice CAM and ON in the consensus CAM (blue). ON and OFF are determined by binarizing the images at pixel threshold t (x-axis). Y-axis shows the proportions at which these two cases occur. Zoomed inset highlights disagreement between CAMs. (d) Consensus CAM pixels are mostly contained within the novice CAM. The x-axis plots the varying pixel thresholds, while the y-axis plots the percent overlap of either how much of the consensus CAM pixels are a subset of the novice CAM (orange) or how much those of the novice CAM are a subset of the consensus (cyan).

a



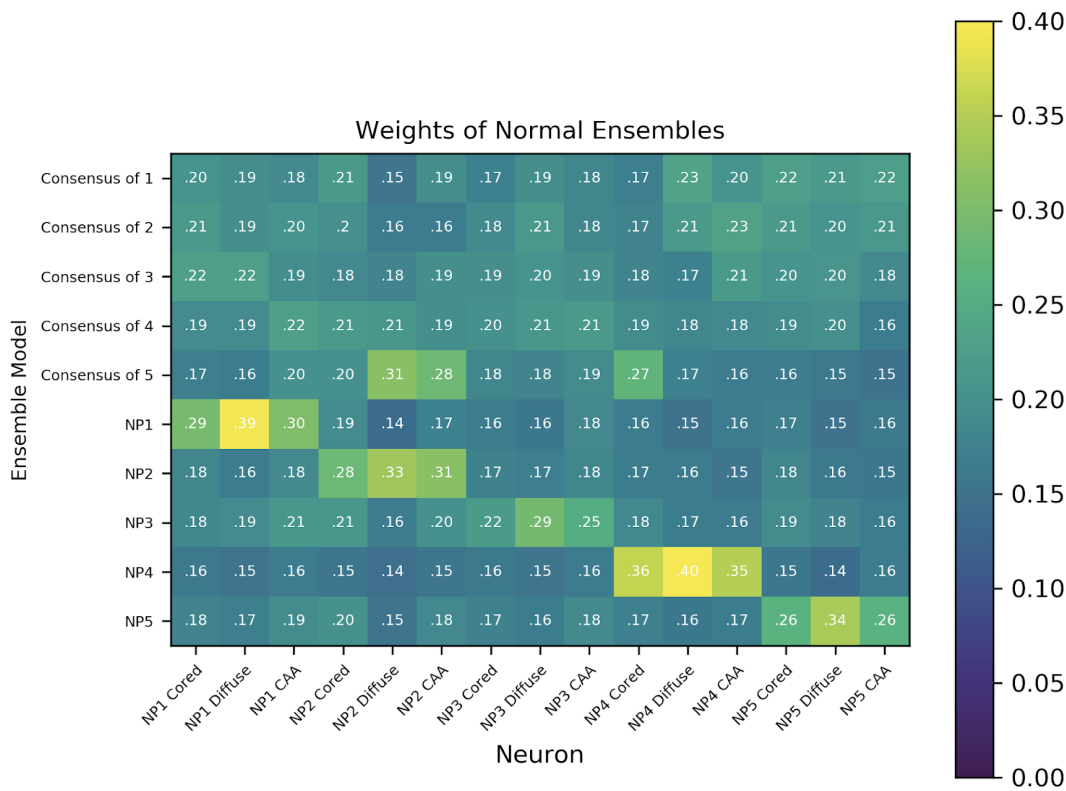
b



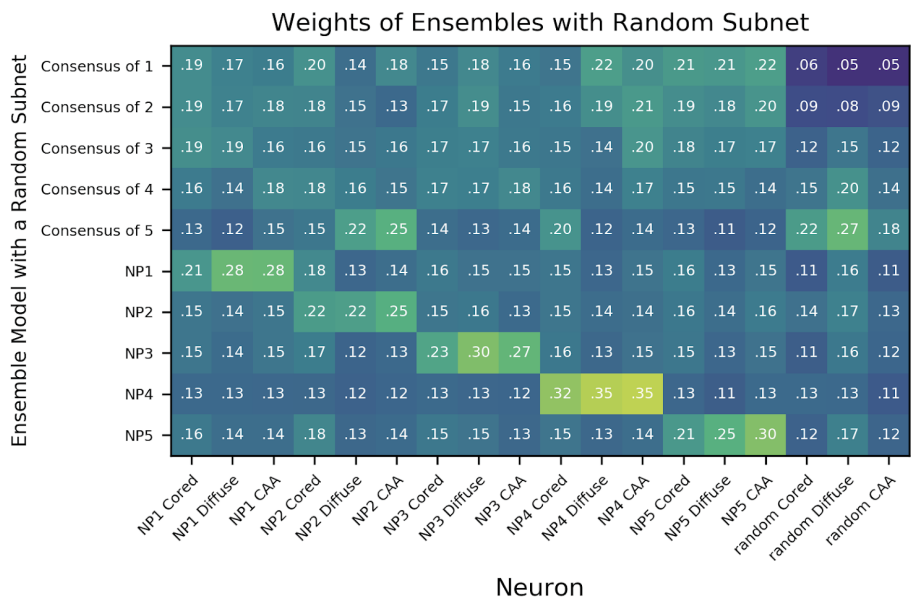
Supplemental Figure 2.8: Intra-rater agreement for self-enrichment and consensus-enrichment.

Intra-rater agreement when only calculated on self-repeat images and phase one data. Novices achieved an average intra-rater agreement accuracy of 0.92 for cored, 0.89 for diffuse, and 0.97 for CAA. Experts achieved an average intra-rater agreement accuracy of 0.90 for cored, 0.89 for diffuse, and 0.98 for CAA. (b) Intra-rater agreement when only calculated on consensus-repeat images and phase one data. Novices achieved an average intra-rater agreement accuracy of 0.93 for cored, 0.91 for diffuse, and 0.98 for CAA. Experts achieved an average intra-rater agreement accuracy of 0.95 for cored, 0.94 for diffuse, and 0.97 for CAA.

a

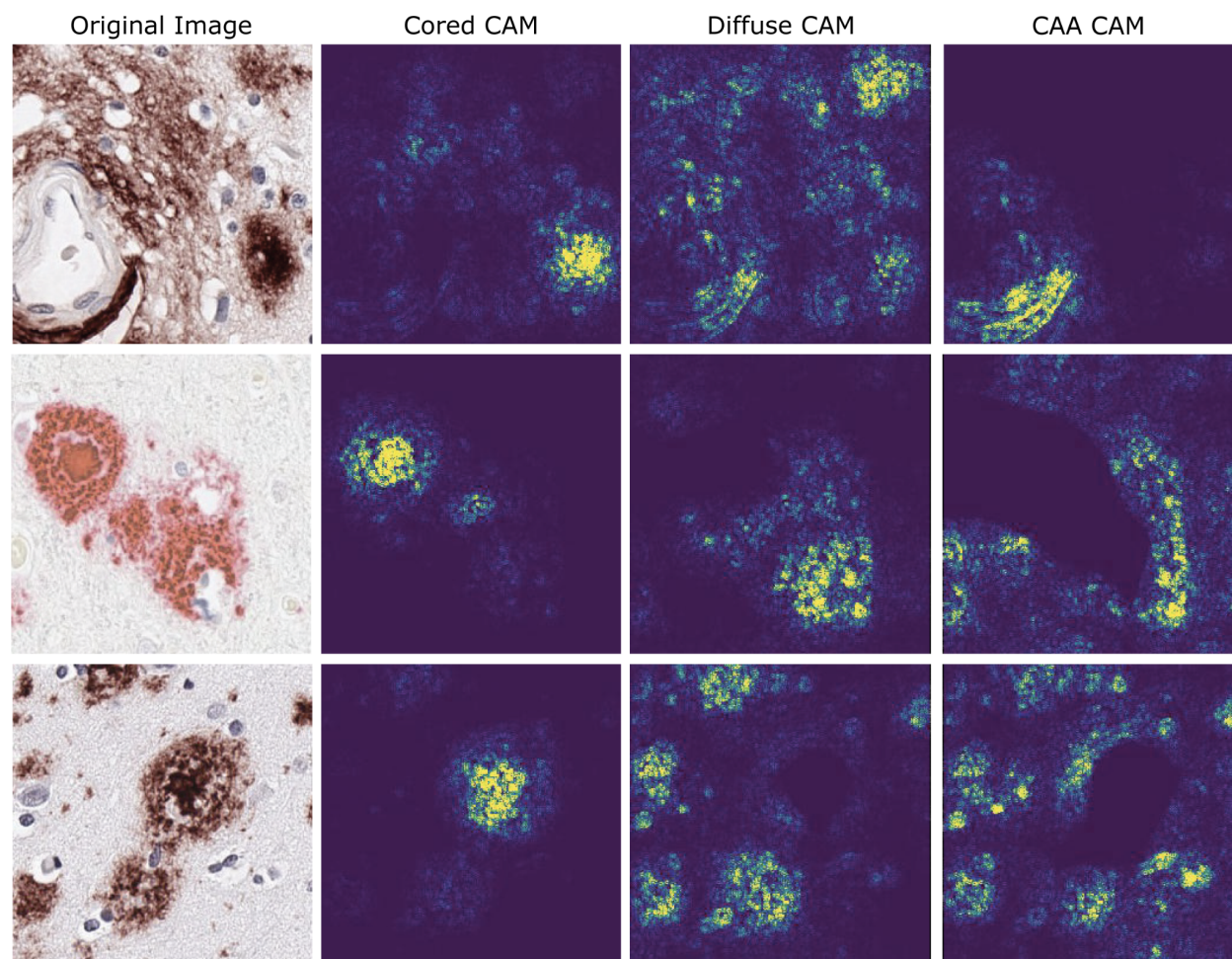


b



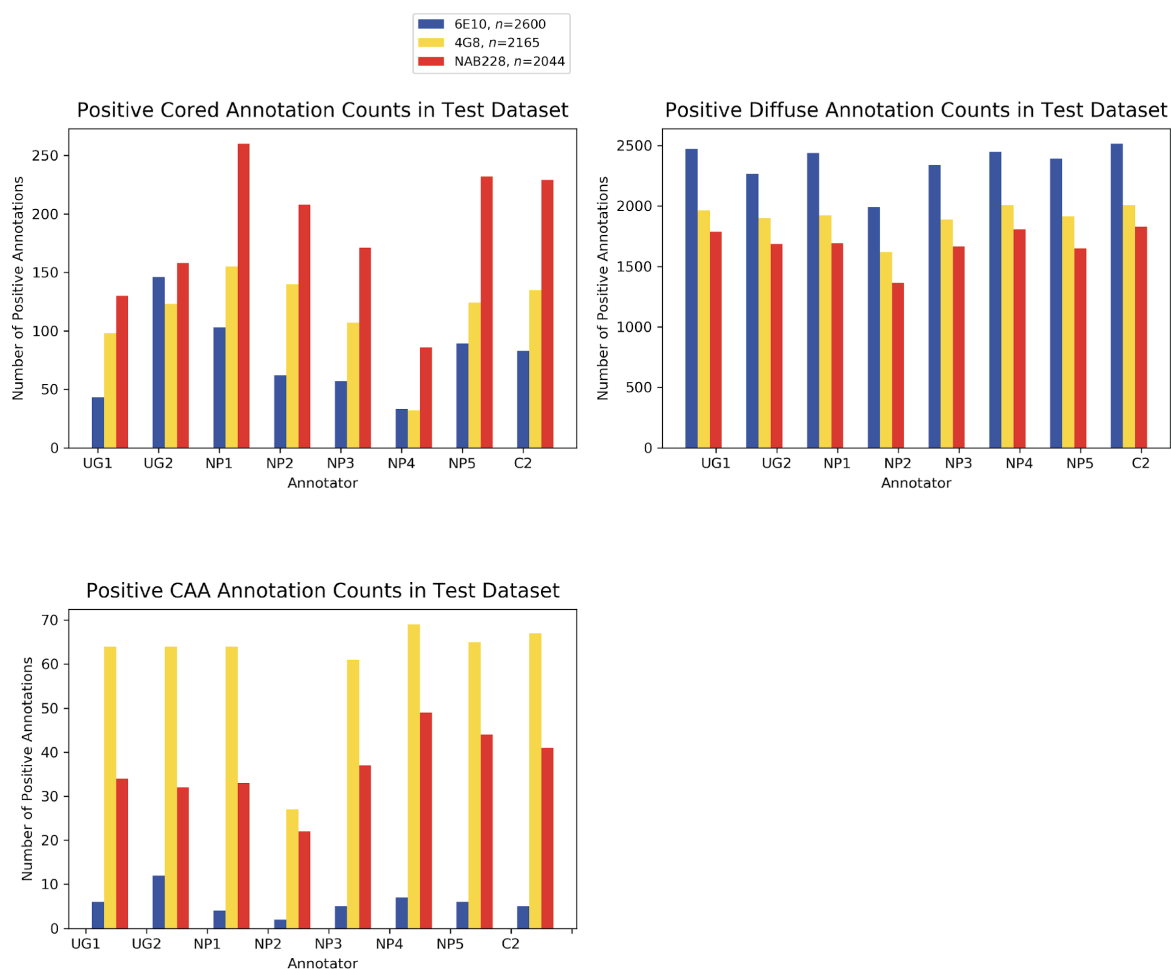
Supplemental Figure 2.9: Ensemble weights.

The y-axis specifies the ensemble model, while the x-axis specifies the output neuron from the single CNN. Values are the softmaxed weights attached from each CNN's class output neurons to the final ensemble output layer. We apply a softmax function to the weights across each of the three amyloid classes. (a) depicts the normal ensembles (five single expert CNNs connected with an affine layer). (b) depicts the ensembles with a single random subnet (Five single expert CNNs and a random CNN connected with an affine layer). The "Random" neurons specify the class output neurons of the random CNN. The softmaxed weights provide a sense of relative importance of each CNN and how much effect each CNN has on the final output.



Supplemental Figure 2.10: CAMs for each class.

We queried the consensus-of-two model to generate a CAM for each A β class (designated by column header). The leftmost column indicates the original image being analyzed. We generated a CAM for each class regardless of whether the image was positive for the class. Each image was labeled as positive by all five annotators (a full consensus-of-five)—the top row was positive for CAA, and the bottom two rows were positive for both cored and CAA.



Supplemental Figure 2.11: Distribution of positive annotations in the test set.

Each plot shows the number of positive annotations for a specific A β class in the test set. There are noticeable differences between stains, even within the same A β class. The CAA class with 6E10 is the most underrepresented in the test set.

Chapter 3: Prototype of running clinical trials in an untrustworthy environment using blockchain

Daniel R. Wong^{1,2}, Sanchita Bhattacharya^{1,2}, Atul J. Butte^{1,2,3}. “Prototype of running clinical trials in an untrustworthy environment using blockchain.” *Nature Communications* **10**, 917 (2019).

1. Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, 94158, US

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

3. Center for Data-Driven Insights and Innovation, University of California, Office of the President, Oakland, CA 94607, USA.

3.1 ABSTRACT

Monitoring and ensuring the integrity of data within the clinical trial process is currently not always feasible with the current research system. We propose a blockchain-based system to make data collected in the clinical trial process immutable, traceable, and potentially more trustworthy. We use raw data from a real completed clinical trial, simulate the trial onto a proof of concept web portal service, and test its resilience to data tampering. We also assess its prospects to provide a traceable and useful audit trail of trial data for regulators, and a flexible

service for all members within the clinical trials network. We also improve the way adverse events are currently reported. In conclusion, we advocate that this service could offer an improvement in clinical trial data management, and could bolster trust in the clinical research process and the ease at which regulators can oversee trials.

3.2 INTRODUCTION

Clinical trial networks often involve many parties and sites, and a large flow of information and confidential data. With the involvement of more parties, more exchanges, and trials being conducted far from sponsoring institutions comes the opportunity for human-induced error, whether it is unintentional or malicious. In a survey of authors of clinical drug trials, 17% of them reported that they were personally aware of intentional fabrication in research¹⁰⁸. Although it is obviously not possible to calculate the actual rate or impact of fraud, there is evidence that some misconduct and alteration of data is happening in scientific and medical research¹⁰⁹, potentially including clinical trials. In addition, current clinical trials are executed with many manual processes that could be prone to error. Enforcing Good Clinical Practice guidelines set forth by the International Council for Harmonization of Requirements for Pharmaceuticals for Human Use (ICH) has been a challenge for regulators of trials, and process improvement is an active area of research¹¹⁰. For regulators of trials such as the Food and Drug Administration (FDA), auditing of data are challenging, and real-time oversight of a trial is lacking as there is no easy and secure way of accessing or viewing the complex network of data transactions as they occur. There can be a lack of transparency and traceability of data, lack of

real-time access to results as they are made, and potentially a risk of data tampering¹¹⁰. The ability to easily trace data back to the original source can be limited¹¹¹, and the FDA has identified this lack of traceability as one of their top data issues to address¹¹².

Blockchain is a new software development methodology involving a unique data structure that has garnered increased attention due to the seminal paper that outlined Bitcoin (bitcoin.org)¹¹³. The technology provides a data structure that ensure a secure and unfalsifiable transaction history. This is accomplished primarily through the use of cryptographic hashing, which has properties and use cases in many domains ranging from internet security to banking. A hash function is a function that maps data of variable length data to a fixed-length digest. Any change to the input data result in an unpredictable change in the hash. In this implementation, each new block added to the chain includes a hash of the previous block. If the previous block is later changed, the subsequent hash would no longer be valid. In addition, blockchains are designed to be append only, and are thus immutable by design, providing a guarantee of safeguarded data. This yields a verifiable and tamper proof history of all transactions since its beginning¹¹³.

We propose a solution to the challenges in the current clinical trials system by using blockchain technology, coupled with changes in methodology for the management of clinical trials. Due to the innate need of clinical trials to have a centralized authority, such as through the regulator, a completely decentralized blockchain infrastructure as used in Bitcoin might not be appropriate. Instead, we borrow ideas and repurpose the methodology to work with clinical trial

management, and demonstrate the feasibility with a new prototype web portal. Blockchain has already been proposed for use in various healthcare settings¹¹⁴, with potential applications in medical record management, claims processing, health supply chain management, and integration of geospatial data in various data modalities¹¹⁵. Work has already begun in seeing how clinical trial management can be improved with blockchain technologies^{116,117}. Moreover, certain aspects relevant to the clinical trial process, such as patient recruiting and IRB (Institutional Review Board) enforcement of human subject regulations via smart contracts, have been described at a high level in the literature¹¹⁸. We improve upon previous work by implementing a web-based portal accessible to all parties with a real clinical trial dataset, facilitating and verifying patient and clinical investigator interaction, integrating version control into the blockchain, expediting adverse event reporting, and testing malicious attacks to data integrity with real world clinical data.

3.3 RESULTS

3.3.1 Network Protocol

We modeled a prototype phase II clinical trial. Upon approval of the study protocol and initiation of phase II, we propose a future regulator could instantiate a private blockchain and registers all participating parties in the portal providing authenticated and controlled web-based and API (application programming interface) access to the blockchain. All parties would be required to use the portal service for any and all exchange of information related to the trial, and only the information present on the blockchain would be used for review when considering

approval of the treatment. The network and representative transactions are illustrated in Figure

3.1.

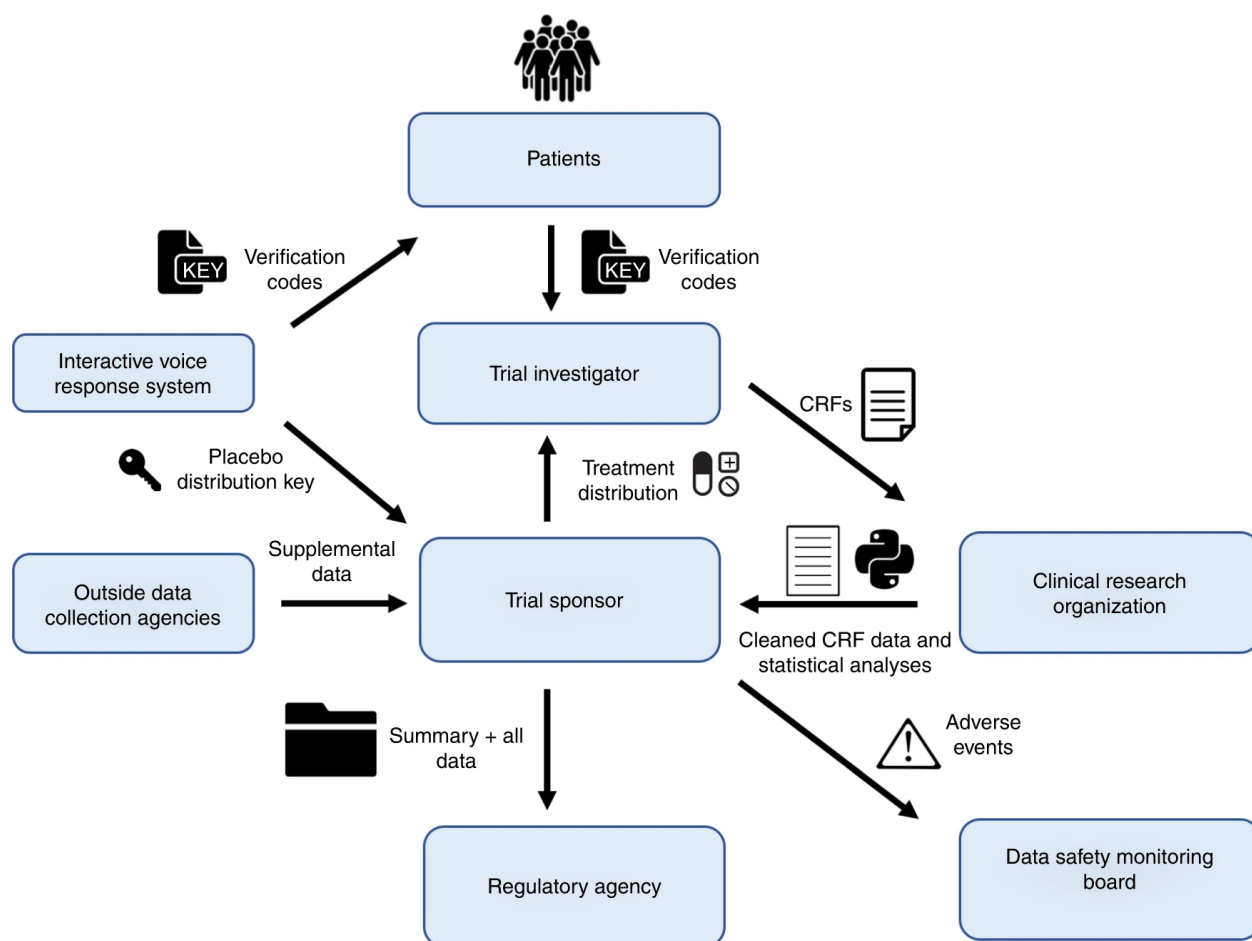


Figure 3.1: The idealized clinical trial network in the context of a blockchain-based record system.

The various transactions (along each arrow) and key participants (boxed) within a clinical trial are shown.

For new patient recruitment and consent acquisition by the clinical site, we propose that an Interactive Voice Response System (IVRS) generates unique verification codes for each subject to give to the trial investigator at the clinical site, and posts encrypted decoding keys for later unblinding. The decoding keys describe the various treatment types that a patient can receive, and will be saved in a password protected environment by the IVRS service provider. The trial sponsor then sends a blinded treatment distribution scheme to the trial investigator at the clinical site. The unique verification codes for each subject are appended to that subject's CRF at the clinical site upon the office visit. All CRFs would be completed digitally and considered valid if the proper verification code is present. Once completed, the CRF will be directed to the Clinical Research Organization (CRO) involved in the trial, and this transaction will be stamped onto the growing blockchain.

When adverse events are reported, we envision the Data Safety Management Board (DSMB) or regulatory agency would be able to continuously see these events through each chain-posted CRF, and then potentially propagate these onto that trial's page for public view, if appropriate. Once the CRO receives raw CRF data, data cleaning, and statistical analysis can begin and be done transparently. Upon completion, cleaned data and analysis scripts are sent to the trial sponsor through the portal and these transactions are subsequently recorded onto the blockchain. Any outside data collection sources that are enrolled in the clinical trial would also have to send data to the sponsor through the portal, all marked on the growing ledger.

When the trial sponsor wishes to apply for approval of the drug, the sponsor would send all of their finalized data and in house statistical analysis results to the regulator through the portal and these, like all other elements, would be subsequently added to the chain. When reviewing for approval, the regulator will only consider data that are present on this secure blockchain, and has full read access to everything that has occurred since the blockchain's instantiation. All data that were ever transmitted in the network would be easily accessible, and its integrity and guarantee of when the transaction occurred will be assured.

3.3.2 Data Transaction Details

Whenever a transaction occurs, the sender, receiver, timestamp, file attachment, and hash of the previous block, are all recorded onto a new block. These elements are then concatenated together, and hashed using the SHA256 algorithm¹¹⁹, with the result instantiated as the hash string of the current block. The blockchain is constructed by creating a linked list of such blocks (Figure 3.2). The previous block's hash is kept for ordering and to make each block dependent on all blocks that preceded it in the chain, which is a useful property for quickly validating a chain⁶. Data storage of the blockchain will be accomplished by duplicating and distributing the chain to physically separate machines and data warehouses to be managed by the regulator (see Supplemental Methods).

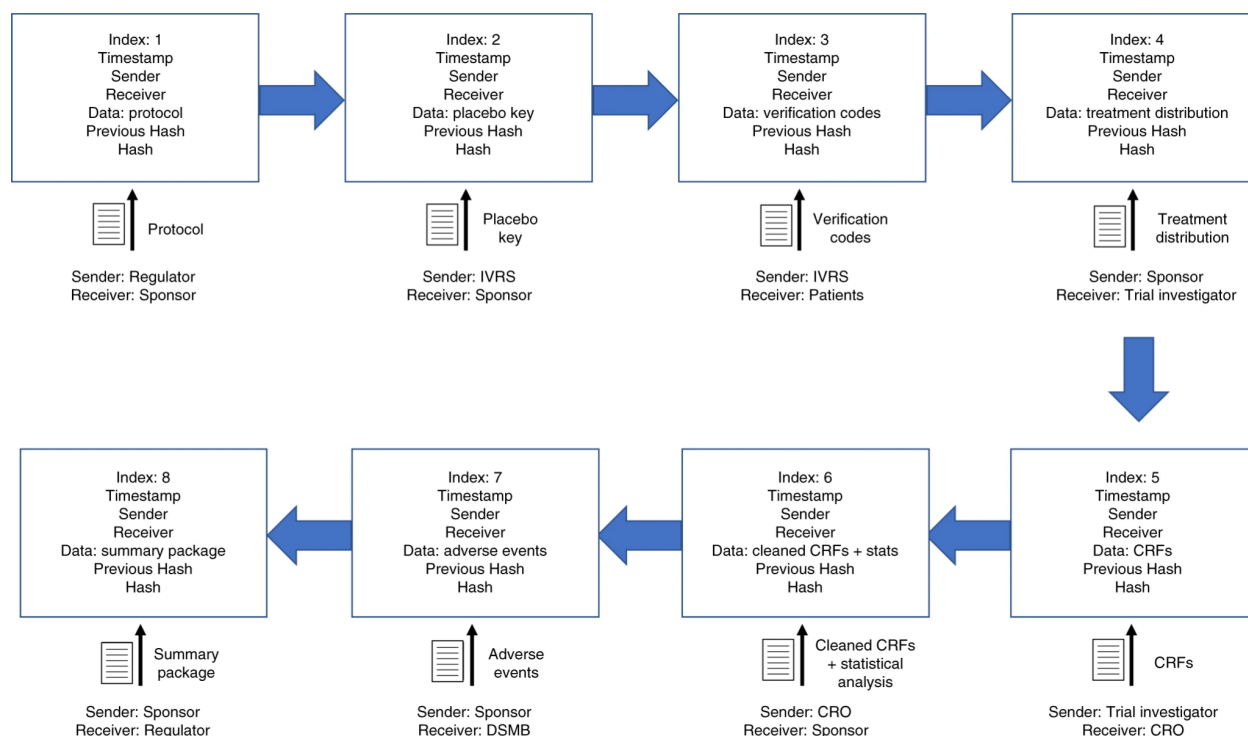


Figure 3.2: The growing blockchain.

With each new transaction that occurs, a block is appended and keeps track of information like the timestamp, sender, receiver, file contents, hash of the previous block, and current hash, all in an immutable data structure. On-chain storage of such elements requires minimal memory allocation linearly proportional to the amount of data being uploaded, since the additional book keeping elements are fixed-length strings. Hence scalability is possible, especially given adequate allocation of hardware made possible via growing cloud storage capabilities (see Supplemental Discussion). A summary, compressed blockchain is shown. The actual blockchain will have a beginning genesis block and individual blocks for each transaction (such as a new block for each CRF instead of a single block for all CRFs as shown). The compressed chain is shown to illustrate the chronology of the trial, and what information constitutes a block.

Encryption through a password based key derivation function is offered, and can ensure that sensitive information is protected if the user chooses to do so, which is especially relevant to maintaining integrity of health and medical information, and eliminating information exposure to unwanted parties. Data are thus stored as an unintelligible series of bytes at the storage level,

ensuring that any sensitive information in the network is obfuscated and will not be compromised in the incident of a data breach.

File storage of any type onto the blockchain is supported, and the user is able to encrypt, send, and extract files easily. For regulators, the full transaction history since the block's genesis is readily available with precise timestamps (Figure 3.3), and the auditing process can be done swiftly and with the confidence that all data are original or version controlled. Content since the earliest phases of the trial are sorted, fully transparent, and easily compressible, and downloadable.

a

Blockchain Ledger						
Validation: Passed, this is a valid blockchain						
Download full blockchain						
Block Index	Time Stamp	Sender	Receiver	File	Previous Block Hash	Block Hash
1	March 18, 2018, 7:07 p.m.	--	--	genesis block	--	aeebad4a796fcc2e15dc4c6061b45ed9b373926a9fc798ca7d2d8cc58182718e
2	March 27, 2018, 1:18 p.m.	Regulator	Sponsor	Casale ITN019AD Protocol v8.0.PTL2331.pdf	aeebad4a796fcc2e15dc4c6061b45ed9b373926a9fc798ca7d2d8cc58182718e	c28111743761092d74e312db47d5ad25d4c372a3e18ce543885d876d34d939
3	March 27, 2018, 1:28 p.m.	IVRS	Sponsor	placeboKey.csv	c23f11743761092d74e312db47d5ad25d4c372a3e18ce543885d876d34d939	56f7bc12d907b55c8959fe742453398ce18011222be796fd8dc4a8f5b9541d39
4	March 27, 2018, 2:13 p.m.	IVRS	Sponsor	verificationCodes.txt	56f7bc12d907b55c8959fe742453398ce18011222be796fd8dc4a8f5b9541d39	2c3eeac829769c99e81a7468fcd8302314370285e702feafa3a342fc728a
5	March 27, 2018, 4:41 p.m.	Sponsor	Investigator	treatmentDistribution.csv	2c3eeac829769c99e81a7468fcd8302314370285e702feafa3a342fc728a	a05ec2214c8c3963b8bb69c43e5f87162518fe45838ead892ddcd4685ddd1c
6	March 27, 2018, 4:42 p.m.	Investigator	CRO	CRF_SUB73443.txt	a05ec2214c8c3963b8bb69c43e5f87162518fe45838ead892ddcd4685ddd1c	cacc3be77749c84bca9cbb0792a1c0ca1eb93cb0a7502229a39c6f52d7ce107
7	March 27, 2018, 4:42 p.m.	Investigator	CRO	CRF_SUB73452.txt	cacc3be77749c84bca9cbb0792a1c0ca1eb93cb0a7502229a39c6f52d7ce107	7d1a23a029516253941da8c7548653abb3ec04be257ef83732f7ad1466c5b5f5
8	March 27, 2018, 4:42 p.m.	Investigator	CRO	CRF_SUB73491.txt	7d1a23a029516253941da8c7548653abb3ec04be257ef83732f7ad1466c5b5f5	aabebf4948c0a417ac9ed93cfc4d131be86ddc58cc5bce08be601a83a5bd4fa9
9	March 27, 2018, 4:42 p.m.	Investigator	CRO	CRF_SUB73511.txt	aabebf4948c0a417ac9ed93cfc4d131be86ddc58cc5bce08be601a83a5bd4fa9	e11ea4af82d78833a2efef651d7240a4d28a8e0af47ea2d71103a73a8e0e613
10	March 27, 2018, 4:43 p.m.	CRO	Sponsor	script.py	e11ea4af82d78833a2efef651d7240a4d28a8e0af47ea2d71103a73a8e0e613	4a203fcd5a8e166a2ef5ca99e21687a2e7df279cc0ebcc5a12ed88b4cd9ea7d49
11	March 27, 2018, 4:43 p.m.	CRO	Sponsor	cleaned_CRF_SUB73443.txt	4a203fcd5a8e166a2ef5ca99e21687a2e7df279cc0ebcc5a12ed88b4cd9ea7d49	c28b64c0d3ab5591166dbc73c762a5e21d180e7d5d7393daa8da946ab331ab
12	March 27, 2018, 4:43 p.m.	CRO	Sponsor	cleaned_CRF_SUB73452.txt	c28b64c0d3ab5591166dbc73c762a5e21d180e7d5d7393daa8da946ab331ab	36530ecf0011a112d71e333e454d4d823a11433d999c7b3e2be67213fa8835c
13	March 27, 2018, 4:43 p.m.	CRO	Sponsor	cleaned_CRF_SUB73491.txt	36530ecf0011a112d71e333e454d4d823a11433d999c7b3e2be67213fa8835c	a884c467c80e2a5d236a3a56708375d5ae3982c98d5dbed70bdc12316f3ced
14	March 27, 2018, 4:43 p.m.	CRO	Sponsor	cleaned_CRF_SUB73511.txt	a884c467c80e2a5d236a3a56708375d5ae3982c98d5dbed70bdc12316f3ced	180c192eb0238631c1ef8c4ec9d194d10bfefec87a3a7b611a93c8ea443bc293
15	March 27, 2018, 4:44 p.m.	Sponsor	DSMB	adverseEvents.csv	180c192eb0238631c1ef8c4ec9d194d10bfefec87a3a7b611a93c8ea443bc293	648644323b0a1aef8906a71ef3a2a738cfa52550cd29a779bd809cc94400d5
16	March 27, 2018, 4:46 p.m.	Sponsor	Regulator	cleaned_CRF_SUB73443.txt	648644323b0a1aef8906a71ef3a2a738cfa52550cd29a779bd809cc94400d5	64d96a88b55936475edc08cc50a200a3295511c87953015d83bea2b2c0c927
17	March 27, 2018, 4:46 p.m.	Sponsor	Regulator	cleaned_CRF_SUB73452.txt	64d96a88b55936475edc08cc50a200a3295511c87953015d83bea2b2c0c927	a59c71146a82c91eab4a10aa095cd19a0d637dbab6200beca9b0e751ebc27d
18	March 27, 2018, 4:46 p.m.	Sponsor	Regulator	cleaned_CRF_SUB73491 (v2).txt	a59c71146a82c91eab4a10aa095cd19a0d637dbab6200beca9b0e751ebc27d	534708dc01d6faa88572d26bacae683c1ebc0fe277a2a694446a0e462411d5e
19	March 27, 2018, 4:46 p.m.	Sponsor	Regulator	cleaned_CRF_SUB73511 (v2).txt	534708dc01d6faa88572d26bacae683c1ebc0fe277a2a694446a0e462411d5e	ef5437a19d553b779f533d0f0a5d9c8db5a15be6991564879e3a50cd4682bc
20	May 16, 2018, 10:51 a.m.	Sponsor	Regulator	summary.zip	ef5437a19d553b779f533d0f0a5d9c8db5a15be6991564879e3a50cd4682bc	736a420b638325d2e5d4b81c360c6e2873b040bc0059e0f327ea86237ad30

b

13	March 27, 2018, 4:43 p.m.	CRO	Sponsor	cleaned_CRF_SUB73491.txt
14	March 27, 2018, 4:43 p.m.	CRO	Sponsor	cleaned_CRF_SUB73511.txt
15	March 27, 2018, 4:44 p.m.	Sponsor	DSMB	adverseEvents.csv
16	March 27, 2018, 4:46 p.m.	Sponsor	Regulator	cleaned_CRF_SUB73443.txt
17	March 27, 2018, 4:46 p.m.	Sponsor	Regulator	cleaned_CRF_SUB73452.txt
18	March 27, 2018, 4:46 p.m.	Sponsor	Regulator	cleaned_CRF_SUB73491 (v2).txt
19	March 27, 2018, 4:46 p.m.	Sponsor	Regulator	cleaned_CRF_SUB73511 (v2).txt

c

Blockchain Ledger						
Validation: Failed, invalid blockchain at block index 5, file: treatmentDistribution.csv						
Download full blockchain						
Block Index	Time Stamp	Sender	Receiver	File	Previous Block Hash	Block Hash
1	March 18, 2018, 7:07 p.m.	--	--	genesis block	--	aeebad4a796fcc2e15dc4c6061b45ed9b373f26adfc798ca7d2d8cc58182718e
2	March 27, 2018, 1:18 p.m.	Regulator	Sponsor	Casale ITN019AD Protocol v8.0.PTL2331.pdf	aeebad4a796fcc2e15dc4c6061b45ed9b373f26adfc798ca7d2d8cc58182718e	c23f11743761092d74e312db47d5ad25d4c372a3e18ce5438f65d876d34df539
3	March 27, 2018, 1:28 p.m.	IVRS	Sponsor	placeboKey.csv	c23f11743761092d74e312db47d5ad25d4c372a3e18ce5438f65d876d34df539	56f7bc12d607b56c895fe7424533f8ce18011222be796fd8d4a6f5b9541d39
4	March 27, 2018, 2:13 p.m.	IVRS	Sponsor	verificationCodes.txt	56f7bc12d607b56c895fe7424533f8ce18011222be796fd8d4a6f5b9541d39	2c3eeac829769c99e81a746f8cfd830231437028f5e702fefafa3a342fc728a
5	March 27, 2018, 4:41 p.m.	Sponsor	Investigator	treatmentDistribution.csv	2c3eeac829769c99e81a746f8cfd830231437028f5e702fefafa3a342fc728a	a05ed214c8c3f63b8b69c43e5f87162518fe45838ead9f2ddcd4685ddd1c

d

Reported Adverse Events	
Subject	Adverse Event
73443	Limb injury NOS, Nasal congestion, Nasal oedema, Dizziness, Pruritus
73452	Nasal oedema, Injection site reaction NOS, Eye swelling, Sneezing, Abdominal pain NOS, Diarrhoea NOS, Nasal polyps, Injection site erythema, Otic examination abnormal, Rhinorrhoea
73491	Abrasion NOS, Throat irritation, Sinus headache, Injection site swelling, Sinus pain, Nasal congestion, Sneezing, Skin laceration, Insomnia, Muscle strain, Dyspnoea exertional, Blood pressure decreased, Sinus congestion, Myalgia, Tendon injury, Arthralgia, Influenza, None
73511	Rash pruritic, Injection site pruritus, Dizziness, Injection site reaction NOS, Parotid duct cyst, Chest tightness, Cough, Sinus congestion, Sinus headache, Blood creatine phosphokinase increased, Blood pressure decreased, Bursitis, Lower respiratory tract infection NOS, Bronchitis NOS, Nasopharyngitis, Sinusitis NOS, Throat tightness, Dermatitis atopic, None

Figure 3.3: Portal functionality.

(a) The public ledger shows the full transaction history since the start of the trial. Blocks are timestamped, indexed, and attached with the file of the transaction and identities of the participating parties. The regulator can easily download individual files, or all elements in bulk, and inspect when and between whom files are shared. (b) New versions of files are given a version number, which increments with each new version of that file. In the screenshot above, the original CRF was modified by the sponsor and automatically appended with a (v2) by the system (boxed in red). The responsible party and time of modification are readily apparent. (c) Internal validation and automated hash checks take place to verify the integrity of the data without the need to manually read through the data's contents. Hash checks are performed chronologically starting with the genesis block, and verify that the hash of the contents of the block match what is to be expected (see Supplemental Methods). Validation fails when the treatmentDistribution.csv is modified at the storage level. The precise origin and location of the fault can be readily discerned. (d) Adverse events are auto populated from investigator uploaded CRFs to the pages of the regulator and DSMB, circumventing the normal, slower, and more error prone pathway that is normally taken for adverse event reporting. Such instances are available for inspection at the soonest possible time.

3.3.3 Version Controls

If a user needs to edit content that is already present on the blockchain, such as the case when an honest mistake is made and needs to be corrected, the user could make an update known by submitting a new transaction with the corrected data without overwriting the old data. By nature, blockchains are append only, so editing the data directly on the blockchain is not possible. We propose combining blockchain's append only criteria with version controlling similar to the functionality of GitHub to accommodate this issue. When a new file is uploaded in a transaction, its contents are hashed and compared to any existing files on the blockchain. If there is a conflict, then the system initiates a schema in which subsequent and differing versions of a file are given incrementing version numbers automatically. Hence, a user can be assured that any downstream modifications to that user's file by anyone else in the network will be documented and cannot be done discretely. No trust in any other parties in the network is needed for data purity, as any tampering will be version controlled and any editors of the file will be easily identified (Figure 3.3).

3.3.4 Simulation of a Previously Completed Clinical Trial

To test how blockchain software technologies could be used to manage the governance and data management aspects of a clinical trial, we simulated how a previously completed clinical trial testing the efficacy and safety of omalizumab¹²⁰ could have been executed using blockchain. We downloaded the completed clinical trials data, including all necessary components, such as raw data, case report form (CRF) components, and protocols from the open

clinical trials data repository ImmPort¹²¹. The trial simulation sequence of events and corresponding files are shown in Figure 3.2. Of the 159 actual patients in the trial with CRF data, we mimicked one subject from each of the four treatment arms for the sake of clarity. Only a few selected categories from the large wealth of CRF information were mirrored for the same reason (see Methods). The statistical scripts in this simulation are not the real Python analyses because of our lack of access.

Here, we show a simulation of how different types of clinical trial events were implemented using our blockchain-based data portal. The first event occurs during encounters between a clinical investigator and the patients after being enrolled for a clinical trial. The second event we simulated is the mutation of CRF data by the trial sponsor. The third event is a storage level corruption on the machine housing the data. Finally, we demonstrate an improved and expedited version of adverse event reporting.

3.3.5 Patient and Clinical Investigator Encounters

We composed truncated and digitized CRFs for the four patients we mimicked on the portal using the publicly available CRF component data. For instance, Subject 73,491 from the study came to the clinical site on the first day of the trial period (day 0). The investigator collected immunological data, such as a white blood cell concentration of 5.9×10^3 cells per μL , eosinophil percentage of 4.6%, and platelet count of 223×10^3 cells per μL (Supplemental Note 3.1). In our proposed schema, if the CRF were paper based, then it would be scanned in; if

electronically captured, it would be directly added to the growing blockchain via the portal, as is the case in our simulation. Verification codes are appended to each CRF.

3.3.6 User Mediated Corruption

We then simulated two types of hostile conditions for the trial. The first was to simulate an effort to manipulate data that were uploaded from other trial staff. While logged in as the trial sponsor, we attempted to modify the adverse events reported in selected CRFs that recorded subjects 73,491 and 73,511 receiving the treatment drug omalizumab so as to deceptively bolster treatment approval in a potentially untrustworthy network. Subject 73,491 showed many adverse reactions during the treatment period, such as muscle strain, injection site swelling, sinus headaches, and nasal congestion among other events (Supplemental Note 3.1), while Subject 73,511 exhibited events such as chest tightness, injection site reactions, sinus congestion, decreased blood pressure, and a lower respiratory tract infection among other ill effects (Supplemental Note 3.2). As the trial sponsor, these CRFs were mutated such that no adverse events were listed (Supplemental Note 3.3, 3.4). The new tampered replacement files are appended with a version number automatically (Figure 3.3), and the corrupting party, time of modification, and changes are all easily visible. The system is capable of handling multiple versions of files in case the original one is part of a later transaction, or in case further revisions or illegitimate mutations are made. These are designated with incrementing version numbers for each new unique version. Original documents, however, are designated with no version number.

With an append only transaction scheme and version controlling, we have a means of keeping a full record of everything that happens to a file, and can easily refer to the author and old and new versions of data, similar to the concept and flexibility of GitHub. This is integral to the auditing process as regulators can track precisely what was changed, by whom, and when with the immutable timestamp. Hence, we simultaneously accommodate the user's need for making changes clear, the regulator's desire for monitoring data easily, and also abide by blockchain's append only schema, which allows for the maintenance and persistence of older data.

3.3.7 Storage Corruption

The second hostile condition we simulated was that of an intentional fault or data corruption at the storage level. In this simulation, we purposefully corrupted the treatment distribution outlining which medication plan was given to which patients (Supplemental Table 3.1, 3.2) to check if the infrastructure would detect and guard against such changes. In the blockchain ledger section of the portal, there is a validation check that successfully shows exactly where the fault and corrupted file lies (Figure 3.3). Due to the sensitivity of a hash function's output in relation to its input, changing the data in even the smallest way in a block, such as modifying a single character in the block's attached file, will result in a completely different hash string. This string will be fed into the input of the next block's hash function, and the resulting string will be completely different from what it was prior to the data modification. Hence, data integrity can be checked by simply comparing the hash strings of a proposed

blockchain under audit with a set of verified and correct hashes. Since each transaction is given its own block, precise determination of location and file that were corrupted are possible by simply finding the first block with an incorrect hash. Hence, storage of the desired and correct hashes is necessary, and we advocate for centralized and secured storage by the trusted regulator who will be performing the audit (see Supplemental Methods). Since only hash strings are required for the purpose of verifying integrity, the regulator need not allocate much hard disk space for the audit process. Verifying integrity can be done quickly as the regulator need only check for string equivalence, which is a quick and trivial process.

3.3.8 Zero-knowledge Proof of Purity

In this proof of concept model, we illustrate the ease at which a data repository can be checked and verified for tampering without manually reviewing each file. SHA256 hashing is a quick and highly optimized process¹¹⁹, and comparing two hash string for equality is trivial. Hence, verifying originality can be done quickly, and without actually opening and inspecting data, which is useful if confidential data are being audited. This serves as a zero-knowledge proof¹²² of data integrity because the auditor need not know the exact configurations and detailed information within a file, and yet still verify its originality. This is particularly useful for the data being handled in clinical processes. This type of methodology and proof provides another layer of security and respect for confidential data, all while quickly and automatically verifying purity of data. Furthermore, by giving the user the option of storing encrypted data on the servers, we ensure that sensitive information cannot be compromised even in the event of a data breach.

3.3.9 Expedited Adverse Event Reporting

As part of the simulation in the portal, we also scrutinized adverse event reporting and how to improve it. In abidance with the goal of making the management of a clinical trial easier and more effective than the current standard, adverse events from the clinical investigator uploaded CRFs are automatically parsed and populated to the pages of the regulator and DSMB (Figure 3.3D). This not only serves as a fast means of assessing safety as the trial continues, but also circumvents the slower and potentially error prone route that adverse event reporting would normally take before reaching the regulator. In the current way clinical trials are run, the CRFs would normally be sent to the CRO, who then parses out the adverse reactions and reports these events to the sponsor, who sends these reports to the DSMB. This process takes time, and is subject to modification or loss by human error or malice. In the proposed scheme and proof of concept service, this vulnerable process is circumvented, and the regulator or DSMB can be immediately notified of each subject's adverse reactions at the soonest possible time, which can be crucial for maintaining public safety. Since the adverse events are extracted directly from the CRFs on the immutable blockchain, the regulator and DSMB can be assured that the events are legitimate.

3.4 DISCUSSION

Blockchain architecture has been prescribed to cure many data interoperability challenges today, including tasks such as overturning the current financial system¹¹³ and remodeling the infrastructure of the internet. Many of these proposed solutions are either not realistic, or have

not yet demonstrated efficacy. Here, we showed that a blockchain-based file and data structure could be used to reliably safeguard data in a clinical trials network, and provide an immutable and fully traceable audit trail. Scalability is of utmost importance when considering the utility of such an architecture, especially if it were to move to full-scale production from proof of concept. The system scales linearly in regards to memory allocation. Similarly, performance is also linear in dataset size, so the architecture is able to scale well (for formal discussion, see Supplemental Discussion).

By using the actual clinical trials data from a previously completed major clinical study, Efficacy and Safety Evaluation of Allergen Immunotherapy Co-Administered with Omalizumab, we showed that data entry, storage, and adverse event reporting can be performed in a more robust and secure manner, which could withstand attacks from both other people in the network and infrastructure damage at the storage level.

We do see some limitations in this work. Forcing all participating parties to use a service like this will still remain a challenge. This could be overcome by restricting regulatory approval of a proposed treatment upon the condition that the trial sponsor uses the service for information exchange, and that CROs and trial investigators are used supporting this methodology. By law, many trials that study drugs, biologics, or devices are required to register with the [clinicaltrials.gov](https://www.clinicaltrials.gov) portal¹²³, so requiring a registration with a service like this could be feasible.

To encourage use of the service and ensure data integrity, only data on the blockchain would be considered when reviewing potential approval of a drug. Hence, any offline personal

transactions of data such as through email would be discouraged. Although blockchain technology provides a means of recording data into structures that are immutable, traceable, and verifiable, it cannot prevent data from being falsified at the point of origin. Clinicians or clinical researchers can be careless or fraudulent and record misleading data into the CRFs, and statisticians within pharmaceutical companies can overinflate p-values because of vested interests in success. These mistakes would be carried forward in the blockchain. The ideal would be to encourage the most raw forms of data or input to be captured as early as possible into the blockchain. Regardless, with unfalsifiable data collected from clinical investigators and cleaned by CROs before being sent to the sponsors, regulators would have the raw data to validate the statistical results due to the design of blockchain. Independent statistical analyses can be run to verify the results. The sponsoring statisticians and CROs could also be required to post their Python and R analysis scripts and freeze these on the ledger for verification and reproducibility, though this is not necessary or critical to the blockchain architecture and merely an example of how the chain can be potentially utilized after its creation. More generally, by being able to see each edge in the network and the corresponding transaction and values, regulators are able to confidently cross check the results of any given subsequent transaction.

By providing each patient with a verification code to give to their physician and validate the CRF upon visit, patient interaction is ensured. Note that this is independent to the functionality of the blockchain infrastructure and design, and merely an additional modification we propose to the clinical trials process with feasibility and ease yet to be tested. Unfortunately,

if a patient and fraudulent clinician are in full adverse cooperation, there is no way of assuring that real data are generated for that office visit. In theory, physicians could still fraudulently purchase these verification codes from patients to save time and avoid running the test trial, but adding a verification code provides a deterrent from blatantly fabricating CRFs. Fabricating data will require the involvement and cooperation of the patient. This will be less enticing and riskier for the clinician to generate fake CRFs, as the clinician is no longer the only party with knowledge of the forgery. Discovery of malpractice is now more likely and can make administering the office visit treatment more attractive than fabrication. Additionally, moving clinicians away from paper-based recordings and forms may also be difficult, but not impossible as the move towards electronic health records (EHRs) rises and as the public concern for non-standardized and fallible physical documentation increases. In contrast to physical documentation, which can be convincingly modified, digital documents through our blockchain service cannot be changed without being noticed and invalidated. Accommodating the current standard of paper documentation is still possible with the service, since these can be easily scanned and converted to a digitized data type to be added to the immutable ledger.

We chose to give data storage control to the regulator, as opposed to distributing data storage to nodes across participants in the network like in Bitcoin applications. We preferred this design because the trial regulator is the only party that can and must be trusted, since this party is the one having final approval over success or rejection of the treatment. It is a central authority that cannot be eliminated. We feel implementing distributed storage to everyone in the network

(such as that used for Bitcoin) does not fit this regulatory context well, and is also impractical as all parties would also have to locally store data pertinent to the clinical trial on their machines. Requiring usage of a web application is a much more feasible first step, and the regulator can have more refined control over the data and abstract management away from the users.

This service could be an improvement to the current clinical system, and could be integrated into existing workflows with the use of a web-based platform, such as the one prototyped here. A service like this could be especially useful in the context of international clinical trials in which oversight is more difficult to administer¹⁰⁸. Moreover, there is an added benefit to the regulator as overwatch and monitoring in real-time since the beginning of a trial could be more feasible, and the regulator need not wait until a final summary package is delivered by the sponsor at the end of the trial. However, current regulatory bodies might not be ready to deal with real-time data access.

This framework can also serve to advance the initiative towards open and publicly accessible big data, as the regulator can selectively decide which data elements to release to the public on the portal for each clinical trial. For future directions, the portal can be easily expanded in functionality to include services similar to clinicaltrials.gov, and also open access to raw data for research scientists like the concept behind the ImmPort repository¹²¹. An interesting future direction and direct application would be to use this methodology to broadcast information that the trial sponsor and regulator deem fit to show to the general public directly through the clinicaltrials.gov portal. This could be an exciting avenue that both takes advantage of the

efficiencies in speed and security that this proof of concept has to offer, along with maintaining regulator discretion and control over data. Of course, much work would be needed before this hypothetical direction can be realized. The idea of open access to clinical trials data are a debated topic^{124,125}, but services like ImmPort have shown that there is a desire among trial sponsors and the scientific community, notably the National Institute of Allergy and Infectious Disease (NIAID) and National Institutes of Health (NIH), to make this data accessible and open¹⁴. In addition, the International Committee of Medical Journal Editors (ICMJE) has expressed their commitment to promoting clinical research data to be openly available¹²⁴.

We propose leaving the question of opening access to trials data to a case-by-case decision of the regulator and sponsor, if they wish to disseminate data on the chain, keeping in mind that the feasibility of maintaining patient privacy and data usability is a question that needs a case-by-case perspective. Our proposal and proof of concept can be synthesized in a methodology that combines ideas of open access data sharing, and the resulting security and trust benefits that arise from blockchain technology. It is important to note however that our blockchain service does not solve or speak to the complex issues of public data sharing, and we are simply advocating this as a potential post-trial use of the service if the regulator deems it responsible and safe. All of the additional functionality that we propose can be integrated into this system, and we will have a unified platform for all things clinical trial catering to the researcher, regulator, and potentially public client. If a service like the one that we propose in this paper can be adopted, the benefits of immutability, traceability, and more trust in the clinical

research process can ensue. Regulators can confidently track the complex flow of data throughout a trial, and be kept up to date on safety proceedings and progress.

3.5 METHODS

3.5.1 Portal Implementation and Simulated Trial Data

We built a proof of concept web portal to access a trial-specific blockchain. The portal was built using a Python and Django software development framework, with additional Python encryption libraries, including hashlib and simplecrypt. The portal is hosted on the cloud platform Heroku. A proof of concept prototype of the portal can be found at: <http://trialchain.ucsf.edu/>. This implementation of the service we are proposing is meant as a proof of concept and is not designed for production usage.

3.5.2 Code Availability

All source code can be found at <https://github.com/wongdaniel8/ClinicalTrials>. The code is publicly accessible for view to encourage reproducibility. The software requirements can be found in the GitHub repository, within the requirements.txt file.

3.5.3 CRF Extraction

In order to test the utility of this service with real world data, we used open access data from a real completed clinical trial: Efficacy and Safety Evaluation of Allergen Immunotherapy Co-Administered with Omalizumab (NCT00078195)¹²⁶. The trial had a total of 159 patients and four

treatment arms (see Supplemental Table 3.2). For the sake of brevity and clarity, we extracted information from four patients in total, one per treatment type. Only a subset of the large wealth of CRF data were mirrored for the simulation. Broad information like race, gender, and age were included. Since the clinical trial was studying allergen immunotherapy, we also extracted relevant assay measurements such as free IgG-a and free IgE-a concentrations in the blood. Additional relevant immunological data, such as concentrations of white blood cells were extracted, and included measurements for monocytes, neutrophils, eosinophils, basophils, and lymphocytes. Concentrations of red blood cells, corpuscular hemoglobin, and platelets were also extracted and recorded into the CRFs. Known adverse events reported during the trial period were also parsed and included in the digitized CRFs. We inserted a necessary verification code on each CRF, with the motivation of encouraging patient and clinician interaction. To meet the requirement of a verification code being present to legitimize the CRF, we simulated the patient providing the verification code (received from the IVRS) to the clinical investigator staff to place on the CRF. Upon audit, the regulator would check that each CRF has the necessary verification code in a process that can be readily machine automated.

3.6 SUPPLEMENTAL MATERIALS

3.6.1 Supplemental Methods

The blockchain in this simulation exists as multiple copies in one physical location on one machine (see README on GitHub: <https://github.com/wongdaniel8/ClinicalTrials/>). In reality, the blockchain will be duplicated and stored on multiple different machines, all

controlled by the regulator, so that if one machine fails or is corrupted, the network data is not jeopardized. The true data and hash values will be reached by consensus of the different machines, such that each machine will cast a vote on what the hash values of particular articles of data should be. The majority vote consensus hash will be taken as the true value. When new transactions are added on the clinical network, the transactions will be appended to each blockchain on each physically separated machine. This schema will appropriately guard against the compromise or failure of individual machines, given enough machines are used in data storage and no more than half of them are compromised. When validating a blockchain, hash checks are performed chronologically from the genesis block. All elements of the block are hashed using SHA256, and this hash string is compared to the ground truth hash that is reached by consensus.

3.6.2 Supplemental Discussion

This scheme, if taken from the current proof of concept prototype to a fully functional service, also has the potential to meet scalability needs as more machines and more memory to these machines will be added as needed. With the advent of massive cloud storage, space availability is not an issue, and runtime is still linear in the amount of data on a trial's blockchain. Elements such as timestamp, sender and receiver identities, and hashes, all of which are small string literals, add a negligible fixed-size burden (k bytes) to the memory allocation requirements per entry. More formally, given a memory burden of n bytes for all files stored on the chain with m number of machines, the system requires $n + k$ bytes duplicated across m

machines, or $m(n + k)$ memory allocation, which is bounded by $O(n)$, and hence linearly scalable. Performance is also linear. Checking for cryptographic hash equivalence is simply a string matching test of fixed length strings returned from SHA256, which is hence bounded by a constant runtime of $O(1)$. When constructing a hash of a file, we must inspect all elements of the data to create the hash string, which is an $O(n)$ operation. Version controlling is fast and scalable, because checking if any two files differ simply requires checking for hash string equivalence, which is an $O(1)$ operation. Since both memory and performance are bounded linearly, we advocate that pushing the proof of concept to full production is feasible and practical, especially given the advances in high performance computing and storage. With favorable linear bounds in place, congestion of the system through increased traffic is not an issue and performance velocity can be maintained.

Instantiating a new blockchain for a new trial is relatively simple and executed by the administrator in charge of the trialchain service. The administrator can instantiate the regulator node and all nodes in the network specified by the regulator, as is the case in private blockchains. The regulator is given full read privileges, and instantiating this super node is the same as instantiating all other nodes, just changing the specific user's privileges. Validating the user privileges and giving more or less access depending on identity is accomplished through logic in the software that is executed automatically. Full implementation of logic can be found in the GitHub repository.

3.6.3 Supplemental Tables

Supplemental Table 3.1: Subject Arm and Codes

Subject Accession	Arm Accession	Verification Code
SUB73443	ARM2	c6f35fe423
SUB73452	ARM1	ed01da9a83
SUB73491	ARM3	741be0d4c0
SUB73511	ARM4	d3feabcf82

Supplemental Table 3.2: Arm Description

Arm Accession	Description	Name
ARM1	Placebo omalizumab pre-treatment, placebo RIT, placebo omalizumab + placebo IT	Placebo Immunotherapy with placebo anti-IgE
ARM2	Placebo omalizumab pre-treatment, ragweed RIT, placebo omalizumab + ragweed IT	Immunotherapy with placebo anti-IgE
ARM3	Omalizumab pre-treatment, placebo RIT, omalizumab + placebo IT	Placebo Immunotherapy with anti-IgE
ARM4	Omalizumab pre-treatment, ragweed RIT, omalizumab + ragweed IT	Immunotherapy with anti-IgE

3.6.4 Supplemental Notes

The following section contains example CRF data extracted from ImmPort.

Supplemental Note 3.1

Subject: SUB73491

verification code: 741be0d4c0

Meta Information

race: Black or African American

gender: Male

age_reported: 28

arm_accession: ARM3

Blood Tests

('study_time_collected: -20 Days', 'name_reported: MCV', 'result_value_reported: 90',
'result_unit_reported: um*3')

('study_time_collected: -20 Days', 'name_reported: WBC', 'result_value_reported: 7.2',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: -20 Days', 'name_reported: MONOCYTES',
'result_value_reported: 5.4', 'result_unit_reported: %')

('study_time_collected: -20 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: -20 Days', 'name_reported: RBC', 'result_value_reported: 4.71',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: -20 Days', 'name_reported: BASOPHILS', 'result_value_reported: 0.2', 'result_unit_reported: %')

('study_time_collected: -20 Days', 'name_reported: MCH', 'result_value_reported: 30', 'result_unit_reported: %')

('study_time_collected: -20 Days', 'name_reported: Total IgE', 'result_value_reported: 34', 'result_unit_reported: unknown')

('study_time_collected: -20 Days', 'name_reported: HEMOGLOBIN', 'result_value_reported: 14.1', 'result_unit_reported: g/dL')

('study_time_collected: -20 Days', 'name_reported: SEGMENTED NEUTROPHILS', 'result_value_reported: 52.7', 'result_unit_reported: %')

('study_time_collected: -20 Days', 'name_reported: LYMPHOCYTES', 'result_value_reported: 37.1', 'result_unit_reported: %')

('study_time_collected: -20 Days', 'name_reported: HEMATOCRIT', 'result_value_reported: 42', 'result_unit_reported: %')

('study_time_collected: -20 Days', 'name_reported: LARGE UNCLASSIFIED CELLS', 'result_value_reported: 3.4', 'result_unit_reported: NULL')

('study_time_collected: -20 Days', 'name_reported: EOSINOPHILS', 'result_value_reported: 1.7', 'result_unit_reported: %')

('study_time_collected: -20 Days', 'name_reported: PLATELET COUNT', 'result_value_reported: 194', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 0 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 4.6', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: RBC', 'result_value_reported: 4.45',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 0 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 223', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 0 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 41', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: MCV', 'result_value_reported: 92',
'result_unit_reported: um*3')

('study_time_collected: 0 Days', 'name_reported: MCHC', 'result_value_reported: 33',
'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: BASOPHILS', 'result_value_reported:
0.4', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 13.5', 'result_unit_reported: g/dL')

('study_time_collected: 0 Days', 'name_reported: MCH', 'result_value_reported: 30',
'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: MONOCYTES', 'result_value_reported:
6.6', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 41.2', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: WBC', 'result_value_reported: 5.9',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 0 Days', 'name_reported: BANDS', 'result_value_reported: 0',
'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 44.4', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 2.8', 'result_unit_reported: NULL')

('study_time_collected: 119 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 38.1', 'result_unit_reported: %')

('study_time_collected: 119 Days', 'name_reported: WBC', 'result_value_reported: 5.4',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 119 Days', 'name_reported: RBC', 'result_value_reported: 4.55',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 119 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: 119 Days', 'name_reported: MCV', 'result_value_reported: 88',
'result_unit_reported: um*3')

('study_time_collected: 119 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 189', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 119 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 50.2', 'result_unit_reported: %')

('study_time_collected: 119 Days', 'name_reported: MONOCYTES',
'result_value_reported: 6.2', 'result_unit_reported: %')

('study_time_collected: 119 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 13.7', 'result_unit_reported: g/dL')

('study_time_collected: 119 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 2.4', 'result_unit_reported: %')

('study_time_collected: 119 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 2.6', 'result_unit_reported: NULL')

('study_time_collected: 119 Days', 'name_reported: MCH', 'result_value_reported: 30',
'result_unit_reported: %')

('study_time_collected: 119 Days', 'name_reported: BASOPHILS',
'result_value_reported: 0.5', 'result_unit_reported: %')

('study_time_collected: 119 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 40', 'result_unit_reported: %')

('study_time_collected: 147 Days', 'name_reported: MCH', 'result_value_reported: 30',
'result_unit_reported: %')

('study_time_collected: 147 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 13.7', 'result_unit_reported: g/dL')

('study_time_collected: 147 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 41', 'result_unit_reported: %')

('study_time_collected: 147 Days', 'name_reported: MCV', 'result_value_reported: 90',
'result_unit_reported: μm^3 ')

('study_time_collected: 147 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 2.6', 'result_unit_reported: %')

('study_time_collected: 147 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 2.4', 'result_unit_reported: NULL')

('study_time_collected: 147 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 54.2', 'result_unit_reported: %')

('study_time_collected: 147 Days', 'name_reported: MONOCYTES',
'result_value_reported: 6.5', 'result_unit_reported: %')

('study_time_collected: 147 Days', 'name_reported: RBC', 'result_value_reported: 4.51',
'result_unit_reported: 10^6 cells/uL')

('study_time_collected: 147 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 203', 'result_unit_reported: 10^3 cells/uL')

('study_time_collected: 147 Days', 'name_reported: WBC', 'result_value_reported: 6.9',
'result_unit_reported: 10^3 cells/uL')

('study_time_collected: 147 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 34', 'result_unit_reported: %')

('study_time_collected: 147 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: 147 Days', 'name_reported: BASOPHILS',
'result_value_reported: 0.3', 'result_unit_reported: %')

('study_time_collected: 160 Days', 'name_reported: RBC', 'result_value_reported: 4.29',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 160 Days', 'name_reported: MONOCYTES',
'result_value_reported: 7.3', 'result_unit_reported: %')

('study_time_collected: 160 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 39', 'result_unit_reported: %')

('study_time_collected: 160 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: 160 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 37', 'result_unit_reported: %')

('study_time_collected: 160 Days', 'name_reported: MCH', 'result_value_reported: 31',
'result_unit_reported: %')

('study_time_collected: 160 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 13.3', 'result_unit_reported: g/dL')

('study_time_collected: 160 Days', 'name_reported: BASOPHILS',
'result_value_reported: 0.3', 'result_unit_reported: %')

('study_time_collected: 160 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 2.8', 'result_unit_reported: %')

('study_time_collected: 160 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 2.9', 'result_unit_reported: NULL')

('study_time_collected: 160 Days', 'name_reported: WBC', 'result_value_reported: 4.7',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 160 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 198', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 160 Days', 'name_reported: MCV', 'result_value_reported: 90',
'result_unit_reported: um*3')

('study_time_collected: 160 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 49.7', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: WBC', 'result_value_reported: 6.2',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 28 Days', 'name_reported: BASOPHILS', 'result_value_reported:
0', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: BANDS', 'result_value_reported: 4',
'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 235', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 28 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 49', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: MCHC', 'result_value_reported: 33',
'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 0', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 43', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 43', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: MCV', 'result_value_reported: 91',
'result_unit_reported: um*3')

('study_time_collected: 28 Days', 'name_reported: RBC', 'result_value_reported: 4.68',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 28 Days', 'name_reported: MCH', 'result_value_reported: 30',
'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 14.2', 'result_unit_reported: g/dL')

('study_time_collected: 28 Days', 'name_reported: MONOCYTES',
'result_value_reported: 4', 'result_unit_reported: %')

('study_time_collected: 369 Days', 'name_reported: BASOPHILS',
'result_value_reported: 0.3', 'result_unit_reported: %')

('study_time_collected: 369 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 47.7', 'result_unit_reported: %')

('study_time_collected: 369 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 39.5', 'result_unit_reported: %')

('study_time_collected: 369 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 2.3', 'result_unit_reported: %')

('study_time_collected: 369 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 14.4', 'result_unit_reported: g/dL')

('study_time_collected: 369 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 2.8', 'result_unit_reported: NULL')

('study_time_collected: 369 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 42', 'result_unit_reported: %')

('study_time_collected: 369 Days', 'name_reported: MCH', 'result_value_reported: 31',
'result_unit_reported: %')

('study_time_collected: 369 Days', 'name_reported: MONOCYTES',
'result_value_reported: 7.4', 'result_unit_reported: %')

('study_time_collected: 369 Days', 'name_reported: MCV', 'result_value_reported: 91',
'result_unit_reported: um*3')

('study_time_collected: 369 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: 369 Days', 'name_reported: RBC', 'result_value_reported: 4.62',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 369 Days', 'name_reported: WBC', 'result_value_reported: 5.2',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 369 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 179', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 56 Days', 'name_reported: BANDS', 'result_value_reported: 1',
'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: MCV', 'result_value_reported: 90',
'result_unit_reported: um*3')

('study_time_collected: 56 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 43', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: MONOCYTES',
'result_value_reported: 9', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 14.5', 'result_unit_reported: g/dL')

('study_time_collected: 56 Days', 'name_reported: BASOPHILS', 'result_value_reported: 0', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: MCHC', 'result_value_reported: 34', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: RBC', 'result_value_reported: 4.74', 'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 56 Days', 'name_reported: MCH', 'result_value_reported: 31', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: LYMPHOCYTES', 'result_value_reported: 51', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: PLATELET COUNT', 'result_value_reported: 221', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 56 Days', 'name_reported: WBC', 'result_value_reported: 6.7', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 56 Days', 'name_reported: SEGMENTED NEUTROPHILS', 'result_value_reported: 39', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: EOSINOPHILS', 'result_value_reported: 0', 'result_unit_reported: %')

('study_time_collected: 91 Days', 'name_reported: BASOPHILS', 'result_value_reported: 0.3', 'result_unit_reported: %')

('study_time_collected: 91 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 30.3', 'result_unit_reported: %')

('study_time_collected: 91 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 199', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 91 Days', 'name_reported: MONOCYTES',
'result_value_reported: 5.5', 'result_unit_reported: %')

('study_time_collected: 91 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 1.9', 'result_unit_reported: %')

('study_time_collected: 91 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 59.6', 'result_unit_reported: %')

('study_time_collected: 91 Days', 'name_reported: MCHC', 'result_value_reported: 35',
'result_unit_reported: %')

('study_time_collected: 91 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 13.8', 'result_unit_reported: g/dL')

('study_time_collected: 91 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 2.4', 'result_unit_reported: NULL')

('study_time_collected: 91 Days', 'name_reported: RBC', 'result_value_reported: 4.53',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 91 Days', 'name_reported: MCH', 'result_value_reported: 30',
'result_unit_reported: %')

('study_time_collected: 91 Days', 'name_reported: MCV', 'result_value_reported: 88',
'result_unit_reported: um*3')

('study_time_collected: 91 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 40', 'result_unit_reported: %')

('study_time_collected: 91 Days', 'name_reported: WBC', 'result_value_reported: 6.8',
'result_unit_reported: 10*3 cells/uL')

Assay Measurements

('study_time_collected: -8', 'analyte: IgE-a Amb a', 'value_reported: 28.3', 'unit_reported:
U/ml')

('study_time_collected: -8', 'analyte: IgG-a Amb a', 'value_reported: 43', 'unit_reported:
U/ml')

('study_time_collected: -8', 'analyte: IgE-a Ragweed', 'value_reported: 1.71',
'unit_reported: kIUa/ml')

('study_time_collected: -8', 'analyte: Free IgE Concentration', 'value_reported: 89.92',
'unit_reported: ng/ml')

('study_time_collected: -8', 'analyte: IgG-a Ragweed', 'value_reported: 100',
'unit_reported: ng/ml')

('study_time_collected: 105', 'analyte: IgG-a Amb a', 'value_reported: 511.5',
'unit_reported: U/ml')

('study_time_collected: 105', 'analyte: IgE-a Amb a', 'value_reported: 3340',
'unit_reported: U/ml')

('study_time_collected: 105', 'analyte: IgE-a Ragweed', 'value_reported: 71.1',
'unit_reported: kIUa/ml')

('study_time_collected: 105', 'analyte: Free IgE Concentration', 'value_reported: 17.376',
'unit_reported: ng/ml')

('study_time_collected: 105', 'analyte: IgG-a Ragweed', 'value_reported: 385',
'unit_reported: ng/ml')

('study_time_collected: 202', 'analyte: Free IgE Concentration', 'value_reported: 34.303',
'unit_reported: ng/ml')

('study_time_collected: 202', 'analyte: IgE-a Ragweed', 'value_reported: 45.6',
'unit_reported: kIUa/ml')

('study_time_collected: 202', 'analyte: IgG-a Ragweed', 'value_reported: 377',
'unit_reported: ng/ml')

('study_time_collected: 202', 'analyte: IgG-a Amb a', 'value_reported: 398',
'unit_reported: U/ml')

('study_time_collected: 202', 'analyte: IgE-a Amb a', 'value_reported: 2195',
'unit_reported: U/ml')

Adverse Reactions

Muscle strain

Injection site swelling

Sinus headache

Insomnia

Nasal congestion

Sinus pain

Sneezing

Throat irritation

Dyspnoea exertional

Blood pressure decreased

Sinus congestion

Myalgia

Abrasion NOS

Tendon injury

Arthralgia

Skin laceration

Influenza

Supplemental Note 3.2

Subject: SUB73511

verification code: d3feabcf82

Meta Information

race: White

gender: Male

age_reported: 45

arm_accession: ARM4

Blood Tests

('study_time_collected: -15 Days', 'name_reported: RBC', 'result_value_reported: 5.45',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: -15 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 47', 'result_unit_reported: %')

('study_time_collected: -15 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 271', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: -15 Days', 'name_reported: WBC', 'result_value_reported: 6.7',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: -15 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 15.8', 'result_unit_reported: g/dL')

('study_time_collected: -15 Days', 'name_reported: MCV', 'result_value_reported: 86',
'result_unit_reported: um*3')

('study_time_collected: -15 Days', 'name_reported: Total IgE', 'result_value_reported: 58',
'result_unit_reported: unknown')

('study_time_collected: -15 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: -15 Days', 'name_reported: MCH', 'result_value_reported: 29',
'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 2.2', 'result_unit_reported: NULL')

('study_time_collected: 0 Days', 'name_reported: MONOCYTES', 'result_value_reported:
4.5', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 14.8', 'result_unit_reported: g/dL')

('study_time_collected: 0 Days', 'name_reported: MCV', 'result_value_reported: 87',
'result_unit_reported: um*3')

('study_time_collected: 0 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 1.9', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 56.3', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 44', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: WBC', 'result_value_reported: 6.8',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 0 Days', 'name_reported: BASOPHILS', 'result_value_reported:
0.6', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: MCH', 'result_value_reported: 29',
'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 34.5', 'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: 0 Days', 'name_reported: RBC', 'result_value_reported: 5.09',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 0 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 295', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 114 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 2.4', 'result_unit_reported: NULL')

('study_time_collected: 114 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 278', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 114 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 46', 'result_unit_reported: %')

('study_time_collected: 114 Days', 'name_reported: BASOPHILS',
'result_value_reported: 0.6', 'result_unit_reported: %')

('study_time_collected: 114 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 57.3', 'result_unit_reported: %')

('study_time_collected: 114 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 2', 'result_unit_reported: %')

('study_time_collected: 114 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 32.5', 'result_unit_reported: %')

('study_time_collected: 114 Days', 'name_reported: MCH', 'result_value_reported: 29',
'result_unit_reported: %')

('study_time_collected: 114 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: 114 Days', 'name_reported: WBC', 'result_value_reported: 7.3',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 114 Days', 'name_reported: MONOCYTES',
'result_value_reported: 5.2', 'result_unit_reported: %')

('study_time_collected: 114 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 15.5', 'result_unit_reported: g/dL')

('study_time_collected: 114 Days', 'name_reported: RBC', 'result_value_reported: 5.45',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 114 Days', 'name_reported: MCV', 'result_value_reported: 85',
'result_unit_reported: um*3')

('study_time_collected: 140 Days', 'name_reported: MCV', 'result_value_reported: 84',
'result_unit_reported: um*3')

('study_time_collected: 140 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 15.1', 'result_unit_reported: g/dL')

('study_time_collected: 140 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 0.9', 'result_unit_reported: NULL')

('study_time_collected: 140 Days', 'name_reported: BASOPHILS',
'result_value_reported: 0.4', 'result_unit_reported: %')

('study_time_collected: 140 Days', 'name_reported: MONOCYTES',
'result_value_reported: 3.5', 'result_unit_reported: %')

('study_time_collected: 140 Days', 'name_reported: WBC', 'result_value_reported: 12.7',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 140 Days', 'name_reported: MCH', 'result_value_reported: 29',
'result_unit_reported: %')

('study_time_collected: 140 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 0.5', 'result_unit_reported: %')

('study_time_collected: 140 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: 140 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 77.8', 'result_unit_reported: %')

('study_time_collected: 140 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 44', 'result_unit_reported: %')

('study_time_collected: 140 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 275', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 140 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 16.9', 'result_unit_reported: %')

('study_time_collected: 140 Days', 'name_reported: RBC', 'result_value_reported: 5.26',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 155 Days', 'name_reported: MCV', 'result_value_reported: 84',
'result_unit_reported: um*3')

('study_time_collected: 155 Days', 'name_reported: MONOCYTES',
'result_value_reported: 4.8', 'result_unit_reported: %')

('study_time_collected: 155 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 44', 'result_unit_reported: %')

('study_time_collected: 155 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 286', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 155 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 33.3', 'result_unit_reported: %')

('study_time_collected: 155 Days', 'name_reported: MCH', 'result_value_reported: 30',
'result_unit_reported: %')

('study_time_collected: 155 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 1.7', 'result_unit_reported: NULL')

('study_time_collected: 155 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 15.7', 'result_unit_reported: g/dL')

('study_time_collected: 155 Days', 'name_reported: BASOPHILS',
'result_value_reported: 0.9', 'result_unit_reported: %')

('study_time_collected: 155 Days', 'name_reported: WBC', 'result_value_reported: 7.2',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 155 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 2.4', 'result_unit_reported: %')

('study_time_collected: 155 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 56.9', 'result_unit_reported: %')

('study_time_collected: 155 Days', 'name_reported: RBC', 'result_value_reported: 5.21',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 155 Days', 'name_reported: MCHC', 'result_value_reported: 36',
'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 64.3', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 15.8', 'result_unit_reported: g/dL')

('study_time_collected: 28 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 1.7', 'result_unit_reported: NULL')

('study_time_collected: 28 Days', 'name_reported: RBC', 'result_value_reported: 5.44',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 28 Days', 'name_reported: MCV', 'result_value_reported: 87',
'result_unit_reported: um*3')

('study_time_collected: 28 Days', 'name_reported: BASOPHILS', 'result_value_reported:
0.7', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: MCHC', 'result_value_reported: 33',
'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 267', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 28 Days', 'name_reported: WBC', 'result_value_reported: 6.5',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 28 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 1.3', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 29.2', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 47', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: MONOCYTES',
'result_value_reported: 3', 'result_unit_reported: %')

('study_time_collected: 28 Days', 'name_reported: MCH', 'result_value_reported: 29',
'result_unit_reported: %')

('study_time_collected: 365 Days', 'name_reported: MCHC', 'result_value_reported: 33',
'result_unit_reported: %')

('study_time_collected: 365 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 15.1', 'result_unit_reported: g/dL')

('study_time_collected: 365 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 52.8', 'result_unit_reported: %')

('study_time_collected: 365 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 1.7', 'result_unit_reported: NULL')

('study_time_collected: 365 Days', 'name_reported: MCV', 'result_value_reported: 87',
'result_unit_reported: um*3')

('study_time_collected: 365 Days', 'name_reported: MONOCYTES',
'result_value_reported: 6.2', 'result_unit_reported: %')

('study_time_collected: 365 Days', 'name_reported: RBC', 'result_value_reported: 5.19',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 365 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 45', 'result_unit_reported: %')

('study_time_collected: 365 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 2.7', 'result_unit_reported: %')

('study_time_collected: 365 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 36.3', 'result_unit_reported: %')

('study_time_collected: 365 Days', 'name_reported: MCH', 'result_value_reported: 29',
'result_unit_reported: %')

('study_time_collected: 365 Days', 'name_reported: WBC', 'result_value_reported: 5.9',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 365 Days', 'name_reported: BASOPHILS',
'result_value_reported: 0.3', 'result_unit_reported: %')

('study_time_collected: 365 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 252', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 56 Days', 'name_reported: RBC', 'result_value_reported: 5.17',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 56 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 58.9', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 270', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 56 Days', 'name_reported: MCH', 'result_value_reported: 30',
'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: BASOPHILS', 'result_value_reported:
0.7', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 1.6', 'result_unit_reported: NULL')

('study_time_collected: 56 Days', 'name_reported: MCHC', 'result_value_reported: 35',
'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 1.7', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 44', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: MCV', 'result_value_reported: 86',
'result_unit_reported: um*3')

('study_time_collected: 56 Days', 'name_reported: WBC', 'result_value_reported: 6.8',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 56 Days', 'name_reported: MONOCYTES',
'result_value_reported: 4.3', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 32.8', 'result_unit_reported: %')

('study_time_collected: 56 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 15.3', 'result_unit_reported: g/dL')

('study_time_collected: 87 Days', 'name_reported: WBC', 'result_value_reported: 6.9',
'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 87 Days', 'name_reported: RBC', 'result_value_reported: 5.21',
'result_unit_reported: 10*6 cells/uL')

('study_time_collected: 87 Days', 'name_reported: MCH', 'result_value_reported: 29',
'result_unit_reported: %')

('study_time_collected: 87 Days', 'name_reported: EOSINOPHILS',
'result_value_reported: 1.5', 'result_unit_reported: %')

('study_time_collected: 87 Days', 'name_reported: LYMPHOCYTES',
'result_value_reported: 39.1', 'result_unit_reported: %')

('study_time_collected: 87 Days', 'name_reported: MCV', 'result_value_reported: 84',
'result_unit_reported: um*3')

('study_time_collected: 87 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'result_value_reported: 51.6', 'result_unit_reported: %')

('study_time_collected: 87 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'result_value_reported: 1.8', 'result_unit_reported: NULL')

('study_time_collected: 87 Days', 'name_reported: HEMOGLOBIN',
'result_value_reported: 14.9', 'result_unit_reported: g/dL')

('study_time_collected: 87 Days', 'name_reported: PLATELET COUNT',
'result_value_reported: 238', 'result_unit_reported: 10*3 cells/uL')

('study_time_collected: 87 Days', 'name_reported: HEMATOCRIT',
'result_value_reported: 44', 'result_unit_reported: %')

('study_time_collected: 87 Days', 'name_reported: BASOPHILS', 'result_value_reported:
0.5', 'result_unit_reported: %')

('study_time_collected: 87 Days', 'name_reported: MCHC', 'result_value_reported: 34',
'result_unit_reported: %')

('study_time_collected: 87 Days', 'name_reported: MONOCYTES',
'result_value_reported: 5.5', 'result_unit_reported: %')

Assay Measurements

('study_time_collected: 100', 'analyte: IgG-a Ragweed', 'value_reported: 2307',
'unit_reported: ng/ml')

('study_time_collected: 100', 'analyte: IgE-a Amb a', 'value_reported: 2631.5',
'unit_reported: U/ml')

('study_time_collected: 100', 'analyte: IgG-a Amb a', 'value_reported: 4979.2',
'unit_reported: U/ml')

('study_time_collected: 100', 'analyte: Free IgE Concentration', 'value_reported: 21.967',
'unit_reported: ng/ml')

('study_time_collected: 100', 'analyte: IgE-a Ragweed', 'value_reported: 47.9',
'unit_reported: kIUa/ml')

('study_time_collected: 66', 'analyte: Free IgE Concentration', 'value_reported: 15.556',
'unit_reported: ng/ml')

('study_time_collected: 66', 'analyte: IgG-a Amb a', 'value_reported: 390.3',
'unit_reported: U/ml')

('study_time_collected: 66', 'analyte: IgE-a Ragweed', 'value_reported: 32.4',
'unit_reported: kIUa/ml')

('study_time_collected: 66', 'analyte: IgE-a Amb a', 'value_reported: 670.9',
'unit_reported: U/ml')

('study_time_collected: 66', 'analyte: IgG-a Ragweed', 'value_reported: 100',
'unit_reported: ng/ml')

('study_time_collected: 73', 'analyte: Free IgE Concentration', 'value_reported: 19.712',
'unit_reported: ng/ml')

('study_time_collected: 73', 'analyte: IgE-a Amb a', 'value_reported: 1005.5',
'unit_reported: U/ml')

('study_time_collected: 73', 'analyte: IgG-a Amb a', 'value_reported: 659.7',
'unit_reported: U/ml')

('study_time_collected: 73', 'analyte: IgE-a Ragweed', 'value_reported: 41', 'unit_reported: kIUa/ml')

('study_time_collected: 73', 'analyte: IgG-a Ragweed', 'value_reported: 290', 'unit_reported: ng/ml')

Adverse Reactions

Parotid duct cyst

Chest tightness

Injection site reaction NOS

Cough

Sinus congestion

Injection site pruritus

Dizziness

Rash pruritic

Sinus headache

Injection site pruritus

Blood creatine phosphokinase increased

Blood pressure decreased

Bursitis

Lower respiratory tract infection NOS

Bronchitis NOS

Nasopharyngitis

Sinusitis NOS

Throat tightness

Injection site pruritus

Dermatitis atopic

Supplemental Note 3.3

Subject: SUB73491

verification code: 741be0d4c0

Meta Information

race: Black or African American

gender: Male

age_reported: 28

arm_accession: ARM3

Blood Tests

('study_time_collected: -20 Days', 'name_reported: MCV', '90', 'um*3')

('study_time_collected: -20 Days', 'name_reported: WBC', '7.2', '10*3 cells/uL')

('study_time_collected: -20 Days', 'name_reported: MONOCYTES', '5.4', '%')

('study_time_collected: -20 Days', 'name_reported: MCHC', '34', '%')

('study_time_collected: -20 Days', 'name_reported: RBC', '4.71', '10*6 cells/uL')

('study_time_collected: -20 Days', 'name_reported: BASOPHILS', '0.2', '%')

('study_time_collected: -20 Days', 'name_reported: MCH', '30', '%')

('study_time_collected: -20 Days', 'name_reported: Total IgE', '34', 'unknown')

('study_time_collected: -20 Days', 'name_reported: HEMOGLOBIN', '14.1', 'g/dL')

('study_time_collected: -20 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'52.7', '%')

('study_time_collected: -20 Days', 'name_reported: LYMPHOCYTES', '37.1', '%')

('study_time_collected: -20 Days', 'name_reported: HEMATOCRIT', '42', '%')

('study_time_collected: -20 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'3.4', 'NULL')

('study_time_collected: -20 Days', 'name_reported: EOSINOPHILS', '1.7', '%')

('study_time_collected: -20 Days', 'name_reported: PLATELET COUNT', '194', '10*3
cells/uL')

('study_time_collected: 0 Days', 'name_reported: EOSINOPHILS', '4.6', '%')

('study_time_collected: 0 Days', 'name_reported: RBC', '4.45', '10*6 cells/uL')

('study_time_collected: 0 Days', 'name_reported: PLATELET COUNT', '223', '10*3
cells/uL')

('study_time_collected: 0 Days', 'name_reported: HEMATOCRIT', '41', '%')

('study_time_collected: 0 Days', 'name_reported: MCV', '92', 'um*3')

('study_time_collected: 0 Days', 'name_reported: MCHC', '33', '%')

('study_time_collected: 0 Days', 'name_reported: BASOPHILS', '0.4', '%')

('study_time_collected: 0 Days', 'name_reported: HEMOGLOBIN', '13.5', 'g/dL')

('study_time_collected: 0 Days', 'name_reported: MCH', '30', '%')

('study_time_collected: 0 Days', 'name_reported: MONOCYTES', '6.6', '%')

('study_time_collected: 0 Days', 'name_reported: LYMPHOCYTES', '41.2', '%')

('study_time_collected: 0 Days', 'name_reported: WBC', '5.9', '10*3 cells/uL')

('study_time_collected: 0 Days', 'name_reported: BANDS', '0', '%')

('study_time_collected: 0 Days', 'name_reported: SEGMENTED NEUTROPHILS', '44.4', '%')

('study_time_collected: 0 Days', 'name_reported: LARGE UNCLASSIFIED CELLS', '2.8', 'NULL')

('study_time_collected: 119 Days', 'name_reported: LYMPHOCYTES', '38.1', '%')

('study_time_collected: 119 Days', 'name_reported: WBC', '5.4', '10*3 cells/uL')

('study_time_collected: 119 Days', 'name_reported: RBC', '4.55', '10*6 cells/uL')

('study_time_collected: 119 Days', 'name_reported: MCHC', '34', '%')

('study_time_collected: 119 Days', 'name_reported: MCV', '88', 'um*3')

('study_time_collected: 119 Days', 'name_reported: PLATELET COUNT', '189', '10*3 cells/uL')

('study_time_collected: 119 Days', 'name_reported: SEGMENTED NEUTROPHILS', '50.2', '%')

('study_time_collected: 119 Days', 'name_reported: MONOCYTES', '6.2', '%')

('study_time_collected: 119 Days', 'name_reported: HEMOGLOBIN', '13.7', 'g/dL')

('study_time_collected: 119 Days', 'name_reported: EOSINOPHILS', '2.4', '%')

('study_time_collected: 119 Days', 'name_reported: LARGE UNCLASSIFIED CELLS', '2.6', 'NULL')

('study_time_collected: 119 Days', 'name_reported: MCH', '30', '%')

('study_time_collected: 119 Days', 'name_reported: BASOPHILS', '0.5', '%')

('study_time_collected: 119 Days', 'name_reported: HEMATOCRIT', '40', '%')

('study_time_collected: 147 Days', 'name_reported: MCH', '30', '%')
 ('study_time_collected: 147 Days', 'name_reported: HEMOGLOBIN', '13.7', 'g/dL')
 ('study_time_collected: 147 Days', 'name_reported: HEMATOCRIT', '41', '%')
 ('study_time_collected: 147 Days', 'name_reported: MCV', '90', 'um*3')
 ('study_time_collected: 147 Days', 'name_reported: EOSINOPHILS', '2.6', '%')
 ('study_time_collected: 147 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
 '2.4', 'NULL')
 ('study_time_collected: 147 Days', 'name_reported: SEGMENTED NEUTROPHILS',
 '54.2', '%')
 ('study_time_collected: 147 Days', 'name_reported: MONOCYTES', '6.5', '%')
 ('study_time_collected: 147 Days', 'name_reported: RBC', '4.51', '10*6 cells/uL')
 ('study_time_collected: 147 Days', 'name_reported: PLATELET COUNT', '203', '10*3
 cells/uL')
 ('study_time_collected: 147 Days', 'name_reported: WBC', '6.9', '10*3 cells/uL')
 ('study_time_collected: 147 Days', 'name_reported: LYMPHOCYTES', '34', '%')
 ('study_time_collected: 147 Days', 'name_reported: MCHC', '34', '%')
 ('study_time_collected: 147 Days', 'name_reported: BASOPHILS', '0.3', '%')
 ('study_time_collected: 160 Days', 'name_reported: RBC', '4.29', '10*6 cells/uL')
 ('study_time_collected: 160 Days', 'name_reported: MONOCYTES', '7.3', '%')
 ('study_time_collected: 160 Days', 'name_reported: HEMATOCRIT', '39', '%')
 ('study_time_collected: 160 Days', 'name_reported: MCHC', '34', '%')

('study_time_collected: 160 Days', 'name_reported: LYMPHOCYTES', '37', '%')
 ('study_time_collected: 160 Days', 'name_reported: MCH', '31', '%')
 ('study_time_collected: 160 Days', 'name_reported: HEMOGLOBIN', '13.3', 'g/dL')
 ('study_time_collected: 160 Days', 'name_reported: BASOPHILS', '0.3', '%')
 ('study_time_collected: 160 Days', 'name_reported: EOSINOPHILS', '2.8', '%')
 ('study_time_collected: 160 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
 '2.9', 'NULL')
 ('study_time_collected: 160 Days', 'name_reported: WBC', '4.7', '10*3 cells/uL')
 ('study_time_collected: 160 Days', 'name_reported: PLATELET COUNT', '198', '10*3
 cells/uL')
 ('study_time_collected: 160 Days', 'name_reported: MCV', '90', 'um*3')
 ('study_time_collected: 160 Days', 'name_reported: SEGMENTED NEUTROPHILS',
 '49.7', '%')
 ('study_time_collected: 28 Days', 'name_reported: WBC', '6.2', '10*3 cells/uL')
 ('study_time_collected: 28 Days', 'name_reported: BASOPHILS', '0', '%')
 ('study_time_collected: 28 Days', 'name_reported: BANDS', '4', '%')
 ('study_time_collected: 28 Days', 'name_reported: PLATELET COUNT', '235', '10*3
 cells/uL')
 ('study_time_collected: 28 Days', 'name_reported: LYMPHOCYTES', '49', '%')
 ('study_time_collected: 28 Days', 'name_reported: MCHC', '33', '%')
 ('study_time_collected: 28 Days', 'name_reported: EOSINOPHILS', '0', '%')

('study_time_collected: 28 Days', 'name_reported: SEGMENTED NEUTROPHILS', '43', '%')

('study_time_collected: 28 Days', 'name_reported: HEMATOCRIT', '43', '%')

('study_time_collected: 28 Days', 'name_reported: MCV', '91', 'um*3')

('study_time_collected: 28 Days', 'name_reported: RBC', '4.68', '10*6 cells/uL')

('study_time_collected: 28 Days', 'name_reported: MCH', '30', '%')

('study_time_collected: 28 Days', 'name_reported: HEMOGLOBIN', '14.2', 'g/dL')

('study_time_collected: 28 Days', 'name_reported: MONOCYTES', '4', '%')

('study_time_collected: 369 Days', 'name_reported: BASOPHILS', '0.3', '%')

('study_time_collected: 369 Days', 'name_reported: SEGMENTED NEUTROPHILS', '47.7', '%')

('study_time_collected: 369 Days', 'name_reported: LYMPHOCYTES', '39.5', '%')

('study_time_collected: 369 Days', 'name_reported: EOSINOPHILS', '2.3', '%')

('study_time_collected: 369 Days', 'name_reported: HEMOGLOBIN', '14.4', 'g/dL')

('study_time_collected: 369 Days', 'name_reported: LARGE UNCLASSIFIED CELLS', '2.8', 'NULL')

('study_time_collected: 369 Days', 'name_reported: HEMATOCRIT', '42', '%')

('study_time_collected: 369 Days', 'name_reported: MCH', '31', '%')

('study_time_collected: 369 Days', 'name_reported: MONOCYTES', '7.4', '%')

('study_time_collected: 369 Days', 'name_reported: MCV', '91', 'um*3')

('study_time_collected: 369 Days', 'name_reported: MCHC', '34', '%')

('study_time_collected: 369 Days', 'name_reported: RBC', '4.62', '10*6 cells/uL')
 ('study_time_collected: 369 Days', 'name_reported: WBC', '5.2', '10*3 cells/uL')
 ('study_time_collected: 369 Days', 'name_reported: PLATELET COUNT', '179', '10*3
 cells/uL')
 ('study_time_collected: 56 Days', 'name_reported: BANDS', '1', '%')
 ('study_time_collected: 56 Days', 'name_reported: MCV', '90', 'um*3')
 ('study_time_collected: 56 Days', 'name_reported: HEMATOCRIT', '43', '%')
 ('study_time_collected: 56 Days', 'name_reported: MONOCYTES', '9', '%')
 ('study_time_collected: 56 Days', 'name_reported: HEMOGLOBIN', '14.5', 'g/dL')
 ('study_time_collected: 56 Days', 'name_reported: BASOPHILS', '0', '%')
 ('study_time_collected: 56 Days', 'name_reported: MCHC', '34', '%')
 ('study_time_collected: 56 Days', 'name_reported: RBC', '4.74', '10*6 cells/uL')
 ('study_time_collected: 56 Days', 'name_reported: MCH', '31', '%')
 ('study_time_collected: 56 Days', 'name_reported: LYMPHOCYTES', '51', '%')
 ('study_time_collected: 56 Days', 'name_reported: PLATELET COUNT', '221', '10*3
 cells/uL')
 ('study_time_collected: 56 Days', 'name_reported: WBC', '6.7', '10*3 cells/uL')
 ('study_time_collected: 56 Days', 'name_reported: SEGMENTED NEUTROPHILS', '39',
 '%')
 ('study_time_collected: 56 Days', 'name_reported: EOSINOPHILS', '0', '%')
 ('study_time_collected: 91 Days', 'name_reported: BASOPHILS', '0.3', '%')

('study_time_collected: 91 Days', 'name_reported: LYMPHOCYTES', '30.3', '%')

('study_time_collected: 91 Days', 'name_reported: PLATELET COUNT', '199', '10*3
cells/uL')

('study_time_collected: 91 Days', 'name_reported: MONOCYTES', '5.5', '%')

('study_time_collected: 91 Days', 'name_reported: EOSINOPHILS', '1.9', '%')

('study_time_collected: 91 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'59.6', '%')

('study_time_collected: 91 Days', 'name_reported: MCHC', '35', '%')

('study_time_collected: 91 Days', 'name_reported: HEMOGLOBIN', '13.8', 'g/dL')

('study_time_collected: 91 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'2.4', 'NULL')

('study_time_collected: 91 Days', 'name_reported: RBC', '4.53', '10*6 cells/uL')

('study_time_collected: 91 Days', 'name_reported: MCH', '30', '%')

('study_time_collected: 91 Days', 'name_reported: MCV', '88', 'um*3')

('study_time_collected: 91 Days', 'name_reported: HEMATOCRIT', '40', '%')

('study_time_collected: 91 Days', 'name_reported: WBC', '6.8', '10*3 cells/uL')

Assay Measurements

('study_time_collected: -8', 'analyte: IgE-a Amb a', '28.3', 'U/ml')

('study_time_collected: -8', 'analyte: IgG-a Amb a', '43', 'U/ml')

('study_time_collected: -8', 'analyte: IgE-a Ragweed', '1.71', 'kIUa/ml')

('study_time_collected: -8', 'analyte: Free IgE Concentration', '89.92', 'ng/ml')
('study_time_collected: -8', 'analyte: IgG-a Ragweed', '100', 'ng/ml')
('study_time_collected: 105', 'analyte: IgG-a Amb a', '511.5', 'U/ml')
('study_time_collected: 105', 'analyte: IgE-a Amb a', '3340', 'U/ml')
('study_time_collected: 105', 'analyte: IgE-a Ragweed', '71.1', 'kIUa/ml')
('study_time_collected: 105', 'analyte: Free IgE Concentration', '17.376', 'ng/ml')
('study_time_collected: 105', 'analyte: IgG-a Ragweed', '385', 'ng/ml')
('study_time_collected: 202', 'analyte: Free IgE Concentration', '34.303', 'ng/ml')
('study_time_collected: 202', 'analyte: IgE-a Ragweed', '45.6', 'kIUa/ml')
('study_time_collected: 202', 'analyte: IgG-a Ragweed', '377', 'ng/ml')
('study_time_collected: 202', 'analyte: IgG-a Amb a', '398', 'U/ml')
('study_time_collected: 202', 'analyte: IgE-a Amb a', '2195', 'U/ml')

Adverse Reactions

None

Supplemental Note 3.4

Subject: SUB73511

verification code: d3feabcf82

Meta Information

race: White

gender: Male

age_reported: 45

arm_accession: ARM4

Blood Tests

('study_time_collected: -15 Days', 'name_reported: RBC', '5.45', '10*6 cells/uL')

('study_time_collected: -15 Days', 'name_reported: HEMATOCRIT', '47', '%')

('study_time_collected: -15 Days', 'name_reported: PLATELET COUNT', '271', '10*3 cells/uL')

('study_time_collected: -15 Days', 'name_reported: WBC', '6.7', '10*3 cells/uL')

('study_time_collected: -15 Days', 'name_reported: HEMOGLOBIN', '15.8', 'g/dL')

('study_time_collected: -15 Days', 'name_reported: MCV', '86', 'um*3')

('study_time_collected: -15 Days', 'name_reported: Total IgE', '58', 'unknown')

('study_time_collected: -15 Days', 'name_reported: MCHC', '34', '%')

('study_time_collected: -15 Days', 'name_reported: MCH', '29', '%')

('study_time_collected: 0 Days', 'name_reported: LARGE UNCLASSIFIED CELLS', '2.2', 'NULL')

('study_time_collected: 0 Days', 'name_reported: MONOCYTES', '4.5', '%')
 ('study_time_collected: 0 Days', 'name_reported: HEMOGLOBIN', '14.8', 'g/dL')
 ('study_time_collected: 0 Days', 'name_reported: MCV', '87', 'um*3')
 ('study_time_collected: 0 Days', 'name_reported: EOSINOPHILS', '1.9', '%')
 ('study_time_collected: 0 Days', 'name_reported: SEGMENTED NEUTROPHILS', '56.3',
 '%')
 ('study_time_collected: 0 Days', 'name_reported: HEMATOCRIT', '44', '%')
 ('study_time_collected: 0 Days', 'name_reported: WBC', '6.8', '10*3 cells/uL')
 ('study_time_collected: 0 Days', 'name_reported: BASOPHILS', '0.6', '%')
 ('study_time_collected: 0 Days', 'name_reported: MCH', '29', '%')
 ('study_time_collected: 0 Days', 'name_reported: LYMPHOCYTES', '34.5', '%')
 ('study_time_collected: 0 Days', 'name_reported: MCHC', '34', '%')
 ('study_time_collected: 0 Days', 'name_reported: RBC', '5.09', '10*6 cells/uL')
 ('study_time_collected: 0 Days', 'name_reported: PLATELET COUNT', '295', '10*3
 cells/uL')
 ('study_time_collected: 114 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
 '2.4', 'NULL')
 ('study_time_collected: 114 Days', 'name_reported: PLATELET COUNT', '278', '10*3
 cells/uL')
 ('study_time_collected: 114 Days', 'name_reported: HEMATOCRIT', '46', '%')
 ('study_time_collected: 114 Days', 'name_reported: BASOPHILS', '0.6', '%')

('study_time_collected: 114 Days', 'name_reported: SEGMENTED NEUTROPHILS',
 '57.3', '%')
 ('study_time_collected: 114 Days', 'name_reported: EOSINOPHILS', '2', '%')
 ('study_time_collected: 114 Days', 'name_reported: LYMPHOCYTES', '32.5', '%')
 ('study_time_collected: 114 Days', 'name_reported: MCH', '29', '%')
 ('study_time_collected: 114 Days', 'name_reported: MCHC', '34', '%')
 ('study_time_collected: 114 Days', 'name_reported: WBC', '7.3', '10*3 cells/uL')
 ('study_time_collected: 114 Days', 'name_reported: MONOCYTES', '5.2', '%')
 ('study_time_collected: 114 Days', 'name_reported: HEMOGLOBIN', '15.5', 'g/dL')
 ('study_time_collected: 114 Days', 'name_reported: RBC', '5.45', '10*6 cells/uL')
 ('study_time_collected: 114 Days', 'name_reported: MCV', '85', 'um*3')
 ('study_time_collected: 140 Days', 'name_reported: MCV', '84', 'um*3')
 ('study_time_collected: 140 Days', 'name_reported: HEMOGLOBIN', '15.1', 'g/dL')
 ('study_time_collected: 140 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
 '0.9', 'NULL')
 ('study_time_collected: 140 Days', 'name_reported: BASOPHILS', '0.4', '%')
 ('study_time_collected: 140 Days', 'name_reported: MONOCYTES', '3.5', '%')
 ('study_time_collected: 140 Days', 'name_reported: WBC', '12.7', '10*3 cells/uL')
 ('study_time_collected: 140 Days', 'name_reported: MCH', '29', '%')
 ('study_time_collected: 140 Days', 'name_reported: EOSINOPHILS', '0.5', '%')
 ('study_time_collected: 140 Days', 'name_reported: MCHC', '34', '%')

('study_time_collected: 140 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'77.8', '%')

('study_time_collected: 140 Days', 'name_reported: HEMATOCRIT', '44', '%')

('study_time_collected: 140 Days', 'name_reported: PLATELET COUNT', '275', '10*3
cells/uL')

('study_time_collected: 140 Days', 'name_reported: LYMPHOCYTES', '16.9', '%')

('study_time_collected: 140 Days', 'name_reported: RBC', '5.26', '10*6 cells/uL')

('study_time_collected: 155 Days', 'name_reported: MCV', '84', 'um*3')

('study_time_collected: 155 Days', 'name_reported: MONOCYTES', '4.8', '%')

('study_time_collected: 155 Days', 'name_reported: HEMATOCRIT', '44', '%')

('study_time_collected: 155 Days', 'name_reported: PLATELET COUNT', '286', '10*3
cells/uL')

('study_time_collected: 155 Days', 'name_reported: LYMPHOCYTES', '33.3', '%')

('study_time_collected: 155 Days', 'name_reported: MCH', '30', '%')

('study_time_collected: 155 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'1.7', 'NULL')

('study_time_collected: 155 Days', 'name_reported: HEMOGLOBIN', '15.7', 'g/dL')

('study_time_collected: 155 Days', 'name_reported: BASOPHILS', '0.9', '%')

('study_time_collected: 155 Days', 'name_reported: WBC', '7.2', '10*3 cells/uL')

('study_time_collected: 155 Days', 'name_reported: EOSINOPHILS', '2.4', '%')

('study_time_collected: 155 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'56.9', '%')

('study_time_collected: 155 Days', 'name_reported: RBC', '5.21', '10*6 cells/uL')

('study_time_collected: 155 Days', 'name_reported: MCHC', '36', '%')

('study_time_collected: 28 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'64.3', '%')

('study_time_collected: 28 Days', 'name_reported: HEMOGLOBIN', '15.8', 'g/dL')

('study_time_collected: 28 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'1.7', 'NULL')

('study_time_collected: 28 Days', 'name_reported: RBC', '5.44', '10*6 cells/uL')

('study_time_collected: 28 Days', 'name_reported: MCV', '87', 'um*3')

('study_time_collected: 28 Days', 'name_reported: BASOPHILS', '0.7', '%')

('study_time_collected: 28 Days', 'name_reported: MCHC', '33', '%')

('study_time_collected: 28 Days', 'name_reported: PLATELET COUNT', '267', '10*3
cells/uL')

('study_time_collected: 28 Days', 'name_reported: WBC', '6.5', '10*3 cells/uL')

('study_time_collected: 28 Days', 'name_reported: EOSINOPHILS', '1.3', '%')

('study_time_collected: 28 Days', 'name_reported: LYMPHOCYTES', '29.2', '%')

('study_time_collected: 28 Days', 'name_reported: HEMATOCRIT', '47', '%')

('study_time_collected: 28 Days', 'name_reported: MONOCYTES', '3', '%')

('study_time_collected: 28 Days', 'name_reported: MCH', '29', '%')

('study_time_collected: 365 Days', 'name_reported: MCHC', '33', '%')
 ('study_time_collected: 365 Days', 'name_reported: HEMOGLOBIN', '15.1', 'g/dL')
 ('study_time_collected: 365 Days', 'name_reported: SEGMENTED NEUTROPHILS',
 '52.8', '%')
 ('study_time_collected: 365 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
 '1.7', 'NULL')
 ('study_time_collected: 365 Days', 'name_reported: MCV', '87', 'um*3')
 ('study_time_collected: 365 Days', 'name_reported: MONOCYTES', '6.2', '%')
 ('study_time_collected: 365 Days', 'name_reported: RBC', '5.19', '10*6 cells/uL')
 ('study_time_collected: 365 Days', 'name_reported: HEMATOCRIT', '45', '%')
 ('study_time_collected: 365 Days', 'name_reported: EOSINOPHILS', '2.7', '%')
 ('study_time_collected: 365 Days', 'name_reported: LYMPHOCYTES', '36.3', '%')
 ('study_time_collected: 365 Days', 'name_reported: MCH', '29', '%')
 ('study_time_collected: 365 Days', 'name_reported: WBC', '5.9', '10*3 cells/uL')
 ('study_time_collected: 365 Days', 'name_reported: BASOPHILS', '0.3', '%')
 ('study_time_collected: 365 Days', 'name_reported: PLATELET COUNT', '252', '10*3
 cells/uL')
 ('study_time_collected: 56 Days', 'name_reported: RBC', '5.17', '10*6 cells/uL')
 ('study_time_collected: 56 Days', 'name_reported: SEGMENTED NEUTROPHILS',
 '58.9', '%')

('study_time_collected: 56 Days', 'name_reported: PLATELET COUNT', '270', '10*3 cells/uL')

('study_time_collected: 56 Days', 'name_reported: MCH', '30', '%')

('study_time_collected: 56 Days', 'name_reported: BASOPHILS', '0.7', '%')

('study_time_collected: 56 Days', 'name_reported: LARGE UNCLASSIFIED CELLS', '1.6', 'NULL')

('study_time_collected: 56 Days', 'name_reported: MCHC', '35', '%')

('study_time_collected: 56 Days', 'name_reported: EOSINOPHILS', '1.7', '%')

('study_time_collected: 56 Days', 'name_reported: HEMATOCRIT', '44', '%')

('study_time_collected: 56 Days', 'name_reported: MCV', '86', 'um*3')

('study_time_collected: 56 Days', 'name_reported: WBC', '6.8', '10*3 cells/uL')

('study_time_collected: 56 Days', 'name_reported: MONOCYTES', '4.3', '%')

('study_time_collected: 56 Days', 'name_reported: LYMPHOCYTES', '32.8', '%')

('study_time_collected: 56 Days', 'name_reported: HEMOGLOBIN', '15.3', 'g/dL')

('study_time_collected: 87 Days', 'name_reported: WBC', '6.9', '10*3 cells/uL')

('study_time_collected: 87 Days', 'name_reported: RBC', '5.21', '10*6 cells/uL')

('study_time_collected: 87 Days', 'name_reported: MCH', '29', '%')

('study_time_collected: 87 Days', 'name_reported: EOSINOPHILS', '1.5', '%')

('study_time_collected: 87 Days', 'name_reported: LYMPHOCYTES', '39.1', '%')

('study_time_collected: 87 Days', 'name_reported: MCV', '84', 'um*3')

('study_time_collected: 87 Days', 'name_reported: SEGMENTED NEUTROPHILS',
'51.6', '%')

('study_time_collected: 87 Days', 'name_reported: LARGE UNCLASSIFIED CELLS',
'1.8', 'NULL')

('study_time_collected: 87 Days', 'name_reported: HEMOGLOBIN', '14.9', 'g/dL')

('study_time_collected: 87 Days', 'name_reported: PLATELET COUNT', '238', '10*3
cells/uL')

('study_time_collected: 87 Days', 'name_reported: HEMATOCRIT', '44', '%')

('study_time_collected: 87 Days', 'name_reported: BASOPHILS', '0.5', '%')

('study_time_collected: 87 Days', 'name_reported: MCHC', '34', '%')

('study_time_collected: 87 Days', 'name_reported: MONOCYTES', '5.5', '%')

Assay Measurements

('study_time_collected: 100', 'analyte: IgG-a Ragweed', '2307', 'ng/ml')

('study_time_collected: 100', 'analyte: IgE-a Amb a', '2631.5', 'U/ml')

('study_time_collected: 100', 'analyte: IgG-a Amb a', '4979.2', 'U/ml')

('study_time_collected: 100', 'analyte: Free IgE Concentration', '21.967', 'ng/ml')

('study_time_collected: 100', 'analyte: IgE-a Ragweed', '47.9', 'kIUa/ml')

('study_time_collected: 66', 'analyte: Free IgE Concentration', '15.556', 'ng/ml')

('study_time_collected: 66', 'analyte: IgG-a Amb a', '390.3', 'U/ml')

('study_time_collected: 66', 'analyte: IgE-a Ragweed', '32.4', 'kIUa/ml')

('study_time_collected: 66', 'analyte: IgE-a Amb a', '670.9', 'U/ml')
('study_time_collected: 66', 'analyte: IgG-a Ragweed', '100', 'ng/ml')
('study_time_collected: 73', 'analyte: Free IgE Concentration', '19.712', 'ng/ml')
('study_time_collected: 73', 'analyte: IgE-a Amb a', '1005.5', 'U/ml')
('study_time_collected: 73', 'analyte: IgG-a Amb a', '659.7', 'U/ml')
('study_time_collected: 73', 'analyte: IgE-a Ragweed', '41', 'kIUa/ml')
('study_time_collected: 73', 'analyte: IgG-a Ragweed', '290', 'ng/ml')

Adverse Reactions

None

CONCLUSION

In college when I majored in both computer science and molecular biology, I thought that surely I would need to abandon half of my training after college and pick one to advance and practice. Now, after completing this PhD dissertation in a topic that was at the intersection of the two, I now see that interdisciplinary training will be completely necessary to solve the problems of the future. Computing and sophisticated algorithms were the tools that allowed for much of this dissertation to be completed. However, one of the main things I learned while completing my PhD was that a tool is no good unless we know how and where to apply it. Will this application be useful? Is it practical to adopt? Is it accessible to many, or just an elite select few? All of these questions I had to grapple with, and the training I received as a biologist was completely necessary to ensure that I was not building the most complex answer to the most useless question. I have my academic advisors Mike and Atul to thank for helping make sense of impact and the core motivation of why I did any of this at all.

This dissertation was a traversal of the medical pipeline at different scales, and at times I cannot help but feel that there is so much more to do—perhaps smaller pipes branching off of the larger one, and more pipes branching off of even those, or even whole new pipe-systems that were left completely unexplored. I left many questions unanswered that are left as future directions to academia. Starting with the micro and the work done for high-throughput screening, the biggest question I have is what is going to come of those compounds I helped to triage and

unearth? The greatest limitation of that whole body of work was that I declared victory upon discovering that these new compounds were active by dose response analysis. But there is so much more that needs to be proved and tested if those compounds are to ever have any practical impact to humanity. How will they perform in mouse models? Are those compounds safe *in vivo*? Are there any side-effects that will impose even greater danger than the perceived remediation they will bring? All of these more clinical questions I answer with an “I do not know, perhaps someone should investigate them!” They are still predictions, albeit more efficiently ranked predictions than what non-ML based methods were proposing. Through this process, I learned that I am just one piece in the larger biomedical research community, and my appreciation has grown tremendously for the medicinal chemists who follow and have the potential to bring discoveries at the bench and the servers to the patient.

One question I continually asked myself was what are the next steps to bring the things I engineered into day-to-day use by the most number of people? Both of the deep learning tools I built for drug discovery and neuropathology require special hardware and knowledge of software. A future direction would be to wrap the algorithms into an easy to use web platform that is point and click, thereby removing a huge barrier to equitable use regardless of computer science background. This was done for the blockchain work. However, I left this question simply as a proof-of-concept, and a future direction would be to implement a clinical trial prospectively using the platform and assessing the benefits and drawbacks compared to more traditional methods.

Alas, a PhD is only a few years, which is a small blip in the scientific timeline, and I leave many of these exercises to future researchers and aspiring academics. My hope though is that I have advanced understanding both in theory and in practice, with all of my work thus far being completely open source and accessible to anyone with an internet connection. I also hope that through the work presented in this dissertation, future researchers will be encouraged to continue building the bridge between computer science and biomedicine to solve the most pressing questions that impede human health.

REFERENCES

1. Hussey, J. Two Years Later, Windows XP Devices Can Still Make The Health Care Industry Wannacry. *Forbes Magazine* (2019).
2. Minor, L. Why your doctor's office still depends on a fax machine. *Wall Street journal (Eastern ed.)* (2019).
3. Li, Z., Cvijic, M. E. & Zhang, L. 2.15 - Cellular Imaging in Drug Discovery: Imaging and Informatics for Complex Cell Biology. in *Comprehensive Medicinal Chemistry III* (eds. Chackalamannil, S., Rotella, D. & Ward, S. E.) 362–387 (Elsevier, 2017). doi:10.1016/B978-0-12-409547-2.12328-5.
4. Kim, S.-W., Roh, J. & Park, C.-S. Immunohistochemistry for Pathologists: Protocols, Pitfalls, and Tips. *J Pathol Transl Med* **50**, 411–418 (2016).
5. Cardoso, M. C. Fluorescence Microscopy: Spectral Imaging vs. Filter-based Imaging. in *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine* 583–586 (Springer Berlin Heidelberg, 2006). doi:10.1007/3-540-29623-9_5560.
6. Lao, K. *et al.* Drug development for Alzheimer's disease. *J. Drug Target.* **27**, 164–173 (2019).
7. Cummings, J., Lee, G., Ritter, A., Sabbagh, M. & Zhong, K. Alzheimer's disease drug development pipeline: 2019. *Alzheimers. Dement.* **5**, 272–293 (2019).
8. Iqbal, K., Liu, F. & Gong, C.-X. Tau and neurodegenerative disease: the story so far. *Nat. Rev. Neurol.* **12**, 15–27 (2016).

9. Eckermann, K. *et al.* The β -Propensity of Tau Determines Aggregation and Synaptic Loss in Inducible Mouse Models of Tauopathy. *J. Biol. Chem.* **282**, 31755–31765 (2007).
10. Fatouros, C. *et al.* Inhibition of tau aggregation in a novel *Caenorhabditis elegans* model of tauopathy mitigates proteotoxicity. *Hum. Mol. Genet.* **21**, 3587–3603 (2012).
11. Goedert, M., Clavaguera, F. & Tolnay, M. The propagation of prion-like protein inclusions in neurodegenerative diseases. *Trends Neurosci.* **33**, 317–325 (2010).
12. Jucker, M. & Walker, L. C. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature* **501**, 45–51 (2013).
13. Clavaguera, F. *et al.* Transmission and spreading of tauopathy in transgenic mouse brain. *Nat. Cell Biol.* **11**, 909–913 (2009).
14. Aoyagi, A. *et al.* A β and tau prion-like activities decline with longevity in the Alzheimer's disease human brain. *Science Translational Medicine* vol. 11 eaat8462 (2019).
15. Sanders, D. W. *et al.* Distinct tau prion strains propagate in cells and mice and define different tauopathies. *Neuron* **82**, 1271–1288 (2014).
16. Jackson, S. J. *et al.* Short Fibrils Constitute the Major Species of Seed-Competent Tau in the Brains of Mice Transgenic for Human P301S Tau. *J. Neurosci.* **36**, 762–772 (2016).
17. Furman, J. L. *et al.* Widespread tau seeding activity at early Braak stages. *Acta Neuropathol.* **133**, 91–100 (2017).
18. Despres, C. *et al.* Identification of the Tau phosphorylation pattern that drives its aggregation. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9080–9085 (2017).
19. Lai, R., Harrington, C. & Wischik, C. Erratum: Lai, R.Y.K.; Harrington, C.R.; Wischik,

- C.M. Absence of a Role for Phosphorylation in the Tau Pathology of Alzheimer's Disease. *Biomolecules* 2016, 6, 19. *Biomolecules* vol. 6 35 (2016).
20. Goedert, M. *et al.* Assembly of microtubule-associated protein tau into Alzheimer-like filaments induced by sulphated glycosaminoglycans. *Nature* **383**, 550–553 (1996).
 21. Grundke-Iqbal, I. *et al.* Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 4913–4917 (1986).
 22. Sengupta, A. *et al.* Phosphorylation of tau at both Thr 231 and Ser 262 is required for maximal inhibition of its binding to microtubules. *Arch. Biochem. Biophys.* **357**, 299–309 (1998).
 23. Alonso, A. C., Zaidi, T., Grundke-Iqbal, I. & Iqbal, K. Role of abnormally phosphorylated tau in the breakdown of microtubules in Alzheimer disease. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 5562–5566 (1994).
 24. Lindwall, G. & Cole, R. D. Phosphorylation affects the ability of tau protein to promote microtubule assembly. *J. Biol. Chem.* **259**, 5301–5305 (1984).
 25. Johnson, G. V. W. & Stoothoff, W. H. Tau phosphorylation in neuronal cell function and dysfunction. *J. Cell Sci.* **117**, 5721–5729 (2004).
 26. Gong, C.-X. & Iqbal, K. Hyperphosphorylation of Microtubule-Associated Protein Tau: A Promising Therapeutic Target for Alzheimer Disease. *Curr. Med. Chem.* **15**, 2321–2328 (2008).
 27. Grandjean, J.-M. M. *et al.* Discovery of 4-Piperazine Isoquinoline Derivatives as Potent and

- Brain-Permeable Tau Prion Inhibitors with CDK8 Activity. *ACS Med. Chem. Lett.* **11**, 127–132 (2020).
28. Preuss, U., Döring, F., Illenberger, S. & Mandelkow, E. M. Cell cycle-dependent phosphorylation and microtubule binding of tau protein stably transfected into Chinese hamster ovary cells. *Mol. Biol. Cell* **6**, 1397–1410 (1995).
 29. Malia, T. J. *et al.* Epitope mapping and structural basis for the recognition of phosphorylated tau by the anti-tau antibody AT8. *Proteins: Struct. Funct. Bioinf.* **84**, 427–434 (2016).
 30. Goedert, M., Jakes, R. & Vanmechelen, E. Monoclonal antibody AT8 recognises tau protein phosphorylated at both serine 202 and threonine 205. *Neurosci. Lett.* **189**, 167–169 (1995).
 31. Duka, V. *et al.* Identification of the sites of tau hyperphosphorylation and activation of tau kinases in synucleinopathies and Alzheimer's diseases. *PLoS One* **8**, e75025 (2013).
 32. Jensen, E. C. Overview of live-cell imaging: requirements and methods used. *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology* **296**, 1–8 (2013).
 33. Sung, M.-H. & McNally, J. G. Live cell imaging and systems biology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **3**, 167–182 (2011).
 34. Christiansen, E. M. *et al.* In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. *Cell* **173**, 792–803.e19 (2018).
 35. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free

- prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917–920 (2018).
36. Moen, E. *et al.* Deep learning for cellular image analysis. *Nat. Methods* **16**, 1233–1246 (2019).
 37. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv [cs.CV]* (2015).
 38. Liu, G. *et al.* Image inpainting for irregular holes using partial convolutions. in *Proceedings of the European Conference on Computer Vision (ECCV)* 85–100 (2018).
 39. Kundel, F. *et al.* Measurement of Tau Filament Fragmentation Provides Insights into Prion-like Spreading. *ACS Chem. Neurosci.* **9**, 1276–1282 (2018).
 40. Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **49**, 6789–6801 (2006).
 41. Müllers, E., Cascales, H. S., Burdova, K., Macurek, L. & Lindqvist, A. Residual Cdk1/2 activity after DNA damage promotes senescence. *Aging Cell* vol. 16 575–584 (2017).
 42. Chuang, K. V. & Keiser, M. J. Comment on ‘Predicting reaction performance in C-N cross-coupling using machine learning’. *Science* vol. 362 (2018).
 43. Soekhoe, D., van der Putten, P. & Plaat, A. On the Impact of Data Set Size in Transfer Learning Using Deep Neural Networks. in *Advances in Intelligent Data Analysis XV* 50–60 (Springer International Publishing, 2016). doi:10.1007/978-3-319-46349-0_5.
 44. Williams, E. *et al.* Image Data Resource: a bioimage data integration and publication platform. *Nat. Methods* **14**, 775–781 (2017).

45. Allen, B. *et al.* Abundant tau filaments and nonapoptotic neurodegeneration in transgenic mice expressing human P301S tau protein. *J. Neurosci.* **22**, 9340–9351 (2002).
46. Lee, I. S., Long, J. R., Prusiner, S. B. & Safar, J. G. Selective precipitation of prions by polyoxometalate complexes. *J. Am. Chem. Soc.* **127**, 13802–13803 (2005).
47. Bray, M.-A. *et al.* Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
48. Murphy, M. *et al.* Neuropathological diagnostic accuracy. *Br. J. Neurosurg.* **16**, 461–464 (2002).
49. Bruner, J. M., Inouye, L., Fuller, G. N. & Langford, L. A. Diagnostic discrepancies and their clinical impact in a neuropathology referral practice. *Cancer* **79**, 796–803 (1997).
50. Gill, J. M., Reese, C. L., 4th & Diamond, J. J. Disagreement among health care professionals about the urgent care needs of emergency department patients. *Ann. Emerg. Med.* **28**, 474–479 (1996).
51. Fillenbaum, G. G. *et al.* Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years. *Alzheimers. Dement.* **4**, 96–109 (2008).
52. Mirra, S. S. *et al.* Interlaboratory comparison of neuropathology assessments in Alzheimer's disease: a study of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD). *J. Neuropathol. Exp. Neurol.* **53**, 303–315 (1994).
53. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
54. Sarwar, S. *et al.* Physician perspectives on integration of artificial intelligence into

- diagnostic pathology. *NPJ Digit Med* **2**, 28 (2019).
55. Shen, D., Wu, G. & Suk, H.-I. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
 56. Gurcan, M. N. *et al.* Histopathological image analysis: a review. *IEEE Rev. Biomed. Eng.* **2**, 147–171 (2009).
 57. Robertson, S., Azizpour, H., Smith, K. & Hartman, J. Digital image analysis in breast pathology-from image processing techniques to artificial intelligence. *Transl. Res.* **194**, 19–35 (2018).
 58. Khullar, D., Jha, A. K. & Jena, A. B. Reducing Diagnostic Errors--Why Now? *N. Engl. J. Med.* **373**, 2491–2493 (2015).
 59. Graber, M. L. The incidence of diagnostic error in medicine. *BMJ Qual. Saf.* **22 Suppl 2**, ii21–ii27 (2013).
 60. Metter, D. M., Colgan, T. J., Leung, S. T., Timmons, C. F. & Park, J. Y. Trends in the US and Canadian Pathologist Workforces From 2007 to 2017. *JAMA Netw Open* **2**, e194337 (2019).
 61. of Pathologists, R. C. Meeting Pathology Demand. Histopathology Workforce Census. *Royal College of Pathologists* (2018).
 62. Dugger, B. N. & Dickson, D. W. Pathology of Neurodegenerative Diseases. *Cold Spring Harb. Perspect. Biol.* **9**, (2017).
 63. Montine, T. J. *et al.* National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. *Acta Neuropathol.*

- 123**, 1–11 (2012).
64. Love, S. *et al.* Development, appraisal, validation and implementation of a consensus protocol for the assessment of cerebral amyloid angiopathy in post-mortem brain tissue. *Am. J. Neurodegener. Dis.* **3**, 19–32 (2014).
 65. Dickson, D. W. The pathogenesis of senile plaques. *J. Neuropathol. Exp. Neurol.* **56**, 321–339 (1997).
 66. Dickson, T. C. & Vickers, J. C. The morphological phenotype of beta-amyloid plaques and associated neuritic changes in Alzheimer’s disease. *Neuroscience* **105**, 99–107 (2001).
 67. Vonsattel, J. P. *et al.* Cerebral amyloid angiopathy without and with cerebral hemorrhages: a comparative histological study. *Ann. Neurol.* **30**, 637–649 (1991).
 68. Perl, D. P. *et al.* Practical approaches to stereology in the setting of aging- and disease-related brain banks. *J. Chem. Neuroanat.* **20**, 7–19 (2000).
 69. Tang, Z. *et al.* Interpretable classification of Alzheimer’s disease pathologies with a convolutional neural network pipeline. *Nat. Commun.* **10**, 2173 (2019).
 70. Vizcarra, J. C. *et al.* Validation of machine learning models to detect amyloid pathologies across institutions. *Acta Neuropathol Commun* **8**, 59 (2020).
 71. Wang, F., Casalino, L. P. & Khullar, D. Deep Learning in Medicine-Promise, Progress, and Challenges. *JAMA Intern. Med.* **179**, 293–294 (2019).
 72. Guan, M. Y., Gulshan, V., Dai, A. M. & Hinton, G. E. Who Said What: Modeling Individual Labelers Improves Classification. *arXiv [cs.LG]* (2017).
 73. Smyth, P., Fayyad, U. M., Burl, M. C., Perona, P. & Baldi, P. Inferring Ground Truth from

- Subjective Labelling of Venus Images. in *Advances in Neural Information Processing Systems 7* (eds. Tesauro, G., Touretzky, D. S. & Leen, T. K.) 1085–1092 (MIT Press, 1995).
74. Raykar, V. C. *et al.* Learning from crowds. *J. Mach. Learn. Res.* **11**, (2010).
 75. Raykar, V. C. *et al.* Supervised learning from multiple experts: whom to trust when everyone lies a bit. in *Proceedings of the 26th Annual International Conference on Machine Learning* 889–896 (Association for Computing Machinery, 2009).
doi:10.1145/1553374.1553488.
 76. Dawid, A. P. & Skene, A. M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **28**, 20–28 (1979).
 77. Linares, M. *et al.* Collaborative intelligence and gamification for on-line malaria species differentiation. *Malar. J.* **18**, 21 (2019).
 78. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
 79. Komura, D. & Ishikawa, S. Machine Learning Methods for Histopathological Image Analysis. *Comput. Struct. Biotechnol. J.* **16**, 34–42 (2018).
 80. Lipton, Z. C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queueing Syst.* **16**, 31–57 (2018).
 81. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 22071–22080 (2019).
 82. Wainberg, M., Merico, D., Delong, A. & Frey, B. J. Deep learning in biomedicine. *Nat.*

- Biotechnol.* **36**, 829–838 (2018).
83. Lipton, Z. C. The Mythos of Model Interpretability. *arXiv [cs.LG]* (2016).
 84. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).
 85. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv [cs.CV]* (2016).
 86. Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **21**, 34–41 (2001).
 87. Greenberg Steven M. & Vonsattel Jean-Paul G. Diagnosis of Cerebral Amyloid Angiopathy. *Stroke* **28**, 1418–1422 (1997).
 88. Thal, D. R. *et al.* Two types of sporadic cerebral amyloid angiopathy. *J. Neuropathol. Exp. Neurol.* **61**, 282–293 (2002).
 89. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med.* **22**, 276–282 (2012).
 90. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. in *The handbook of brain theory and neural networks* 255–258 (MIT Press, 1998).
 91. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
 92. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran

- Associates, Inc., 2012).
93. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv [cs.CV]* (2013).
 94. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
 95. *The Wisdom Of Crowds*. (Anchor Books, 2005).
 96. Lines, J., Taylor, S. & Bagnall, A. HIVE-COTE: The Hierarchical Vote Collective of Transformation-Based Ensembles for Time Series Classification. in *2016 IEEE 16th International Conference on Data Mining (ICDM)* 1041–1046 (2016).
doi:10.1109/ICDM.2016.0133.
 97. Müller, C. M. *et al.* Staging of sporadic Parkinson disease-related alpha-synuclein pathology: inter- and intra-rater reliability. *J. Neuropathol. Exp. Neurol.* **64**, 623–628 (2005).
 98. Yan, Y., Rosales, R., Fung, G., Subramanian, R. & Dy, J. Learning from multiple annotators with varying expertise. *Mach. Learn.* **95**, 291–327 (2014).
 99. Signaevsky, M. *et al.* Artificial intelligence in neuropathology: deep learning-based assessment of tauopathy. *Lab. Invest.* **99**, 1019–1029 (2019).
 100. Perosa, V. *et al.* Deep learning assisted quantitative assessment of histopathological markers of Alzheimer’s disease and cerebral amyloid angiopathy. *Acta Neuropathol Commun* **9**, 141 (2021).
 101. London, A. J. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus

- Explainability. *Hastings Cent. Rep.* **49**, 15–21 (2019).
102. Settles, B. Active learning literature survey. (2010).
 103. Adebayo, J. *et al.* Sanity Checks for Saliency Maps. *arXiv [cs.CV]* (2018).
 104. Zhao, R., Ouyang, W., Li, H. & Wang, X. Saliency detection by multi-context deep learning. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1265–1274 (2015). doi:10.1109/CVPR.2015.7298731.
 105. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv [cs.LG]* (2017).
 106. Gupta, A. & Arora, S. A Simple Saliency Method That Passes the Sanity Checks. *arXiv [cs.LG]* (2019).
 107. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv [cs.LG]* (2019).
 108. Gupta, A. Fraud and misconduct in clinical research: A concern. *Perspect. Clin. Res.* **4**, 144–147 (2013).
 109. George, S. L. & Buyse, M. Data fraud in clinical trials. *Clin. Investig.* **5**, 161–173 (2015).
 110. Bhatt, A. Quality of clinical trials: A moving target. *Perspect Clin Res.* 2011 Oct; 2 (4): 124--128. doi: 10.4103/2229-3485.86880.
 111. van Valkenhoef, G., Tervonen, T., Brock, B. de & Hillege, H. Deficiencies in the transfer and availability of clinical trials evidence: a review of existing systems and standards. *BMC Med. Inform. Decis. Mak.* **12**, 95 (2012).
 112. Hume, S., Sarnikar, S., Becnel, L. & Bennett, D. Visualizing and Validating Metadata

- Traceability within the CDISC Standards. *AMIA Jt Summits Transl Sci Proc* **2017**, 158–165 (2017).
113. Nakamoto, S. & Others. Bitcoin: a peer-to-peer electronic cash system (2008). (2008).
 114. Kuo, T.-T., Kim, H.-E. & Ohno-Machado, L. Blockchain distributed ledger technologies for biomedical and health care applications. *J. Am. Med. Inform. Assoc.* **24**, 1211–1220 (2017).
 115. Kamel Boulos, M. N., Wilson, J. T. & Clauson, K. A. Geospatial blockchain: promises, challenges, and scenarios in health and healthcare. *Int. J. Health Geogr.* **17**, 25 (2018).
 116. Nugent, T., Upton, D. & Cimpoesu, M. Improving data transparency in clinical trials using blockchain smart contracts. *F1000Research* vol. 5 2541 (2016).
 117. Benchoufi, M. & Ravaud, P. Blockchain technology for improving clinical research quality. *Trials* **18**, 335 (2017).
 118. Choudhury, O. *et al.* Enforcing Human Subject Regulations using Blockchain and Smart Contracts. *BHTY* (2018) doi:10.30953/bhty.v1.10.
 119. Guilford, J., Yap, K. & Gopal, V. Fast SHA-256 implementations on Intel architecture processors. *IA Architects* (2012).
 120. Efficacy and safety evaluation of allergen immunotherapy co-administered with omalizumab, an anti-IgE monoclonal antibody.
<https://www.immunetolerance.org/studies/efficacy-and-safety-evaluation-allergen-immunotherapy-co-administered-omalizumab-anti-ige-mo>.
 121. Bhattacharya, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci Data* **5**, 180015 (2018).

122. Feige, U., Fiat, A. & Shamir, A. Zero-knowledge proofs of identity. *J. Cryptology* **1**, 77–94 (1988).
123. History, policies, and laws. <https://clinicaltrials.gov/ct2/about-site/history>.
124. Taichman, D. B. *et al.* Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors. *Ethiopian journal of health sciences* vol. 27 315–318 (2017).
125. Ross, J. S., Lehman, R. & Gross, C. P. The importance of clinical trial data sharing: toward more open science. *Circ. Cardiovasc. Qual. Outcomes* **5**, 238–240 (2012).
126. Efficacy and safety evaluation of allergen immunotherapy co-administered with omalizumab, an anti-IgE monoclonal antibody.
<https://www.immunetolerance.org/studies/efficacy-and-safety-evaluation-allergen-immunotherapy-co-administered-omalizumab-anti-ige-mo>.

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:

Daniel Wong

FB0406C331CE417...

Author Signature

11/11/2021

Date