Explainable AI for health: where we are and how to move forward

Su-In Lee

Paul G. Allen Professor

Paul G. Allen School of Computer Science & Engineering

University of Washington, Seattle

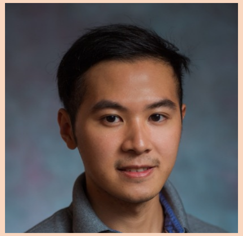
AI for bioMedical Sciences (AIMS) Lab


Nicasia Beebe-Wang (CSE PhD)

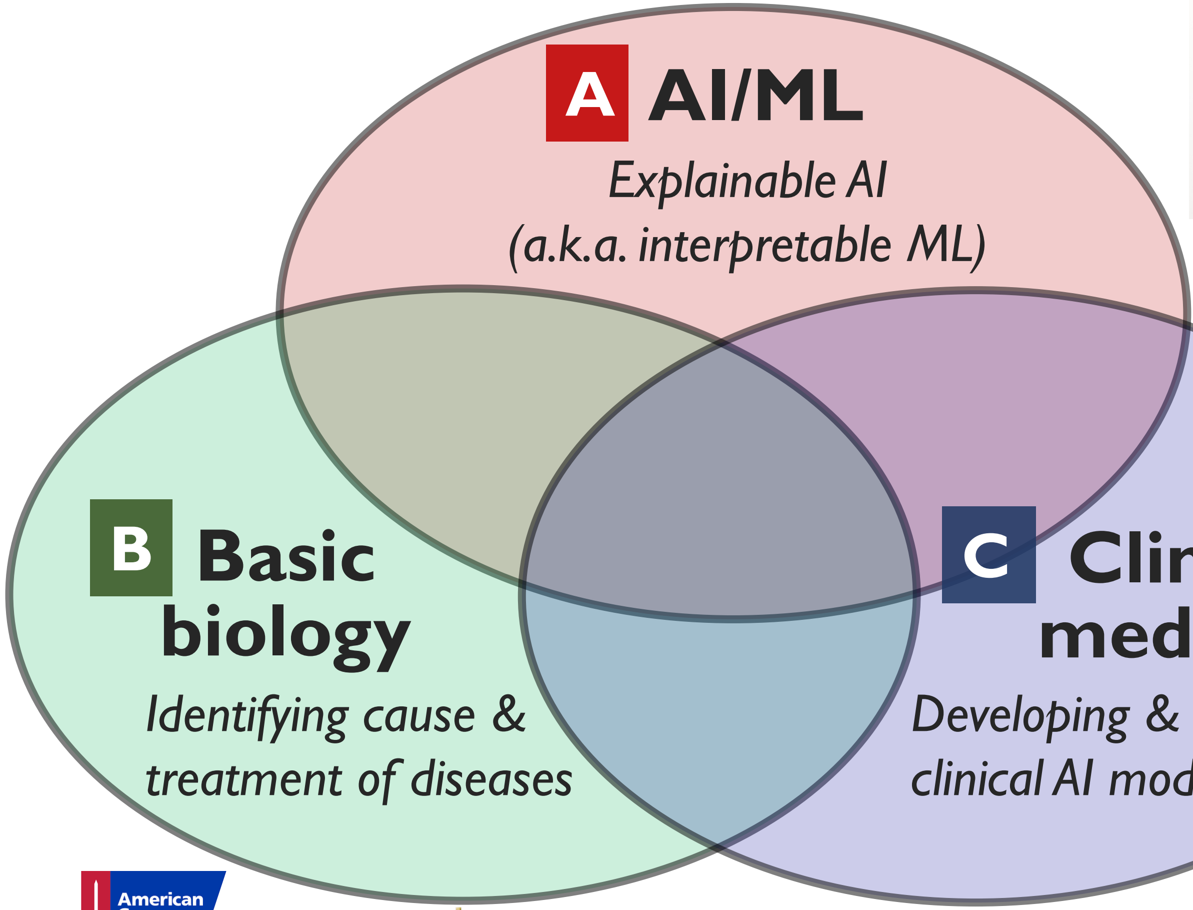

Ian Covert (CSE PhD)


Wei Qiu (CSE PhD)


Chris Lin (CSE PhD)

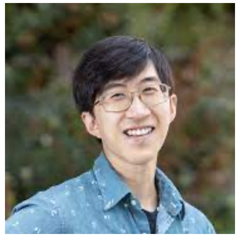

Mingyu Lu, MD (CSE PhD)


Patrick Yu (CSE PhD)




Su-In Lee (PI)

UW MSTP


Hugh Chen (CSE PhD)


Joe Janizek (MSTP, CSE PhD; matched to Stanford)


Ethan Weinberger (CSE PhD)


Alex DeGrave (MSTP, CSE PhD)


Chanwoo Kim (CSE PhD)

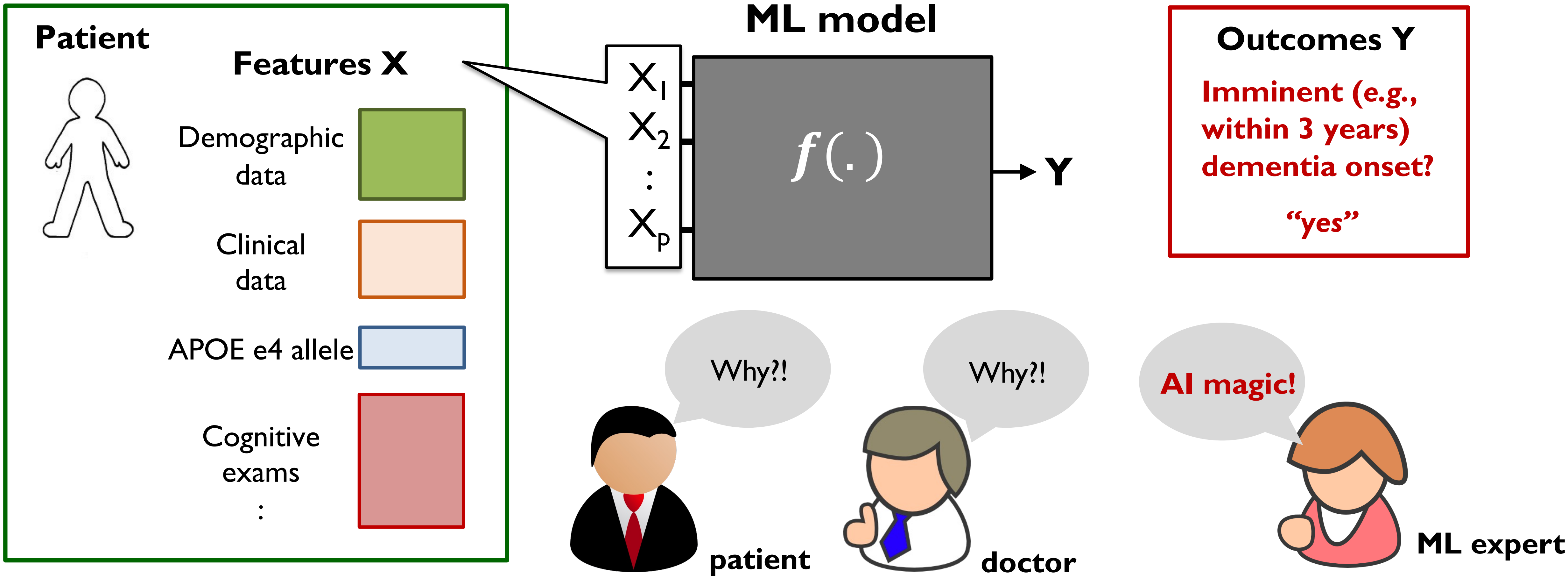

Soham Gadgil (CSE PhD)

Previous members: Ben Logsdon (postdoc), Safiye Celik (CSE PhD'18), Scott Lundberg (CSE PhD'19), Parmita Mehta (CSE PhD'20), Gabe Erion (MSTP, CSE PhD'21; now Harvard Medical School for residency),

Outline – Two parts

- Part 1 – The significance of explainable AI in biomedical sciences
  - Demystifying the biological age
  - Unveiling neurodegenerative disease insights with explainable AI
- Part 2 – Advancing beyond explaining models
  - Cancer therapy design for precision oncology
  - Model auditing
  - Cost-aware clinical AI

Explainable AI (XAI): Accurately predicting an outcome is vital, but the critical question revolves around ***why***.



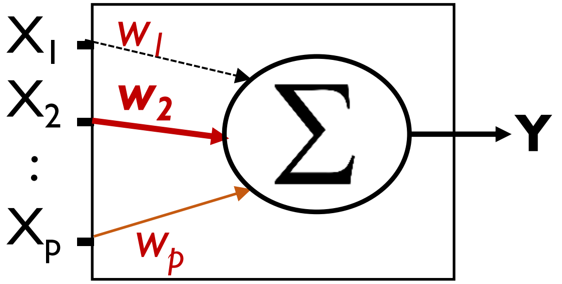
Lundberg et al. ***Nature Machine Intelligence***, *2020* – Featured on the Cover   Beebe-Wang et al. ***IEEE JBHI***, 2021

Our solution is to fundamentally advance AI research to make a prediction ***with explanations***

- Accuracy vs. interpretability
  - Simple models often lead to lower performance.
  - Complex models are often considered to be a black box.

Linear model

X: Features **Y:** Outcome



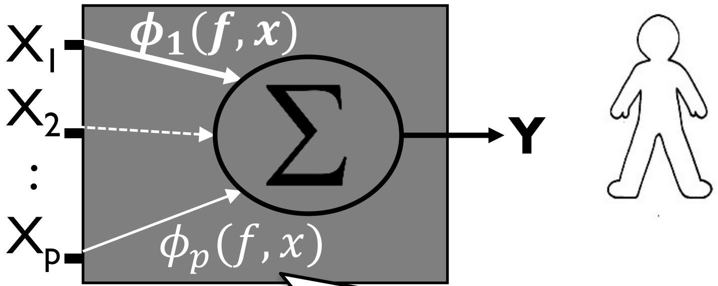
Complex model f (.)

Black Box

Our approach, SHAP

(SHapley Additive exPlanations)

For a particular prediction



SHAP can estimate feature importance for a particular prediction for any model.

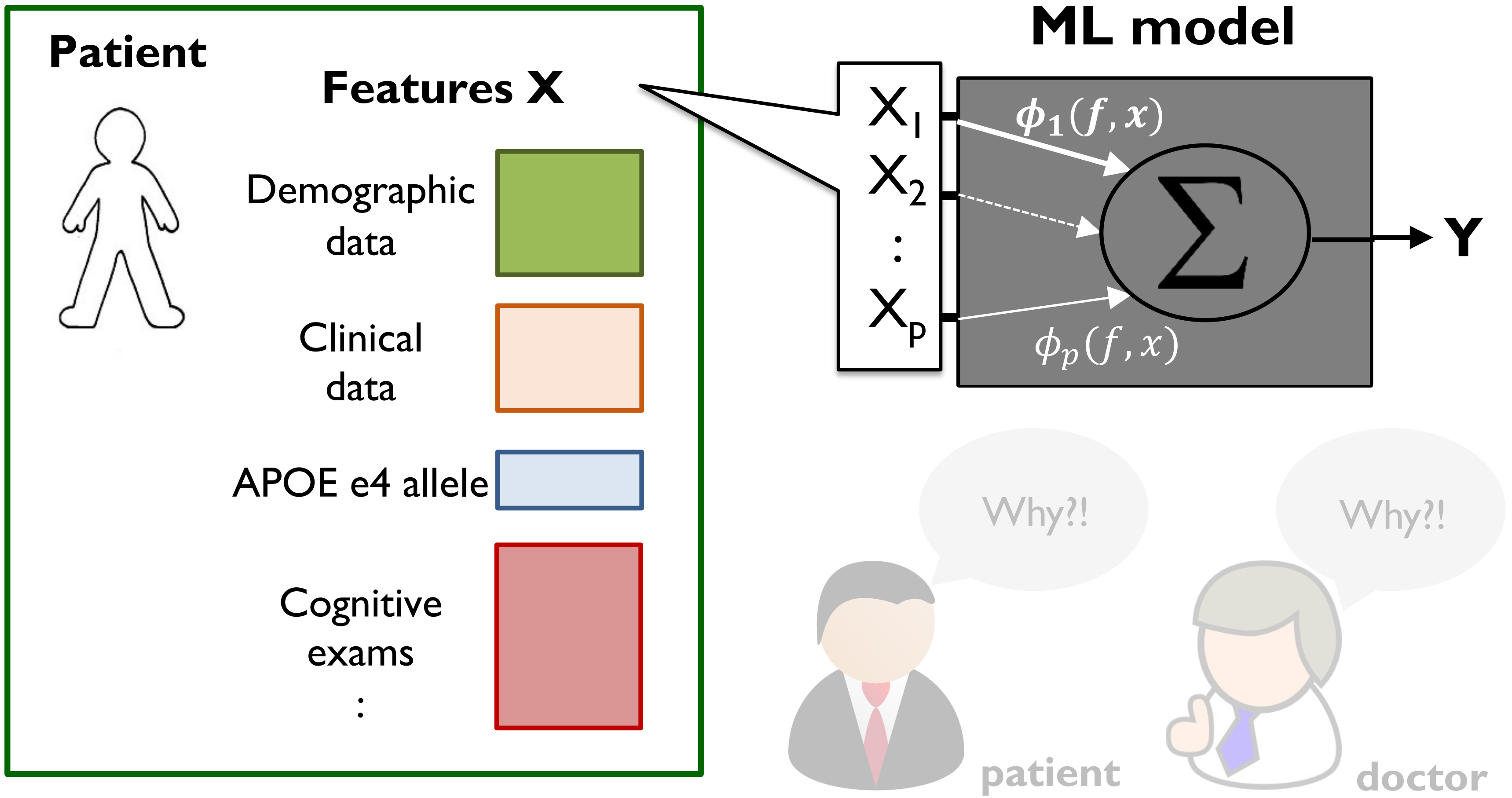
Scott, CSE PhD'19

Lundberg & Lee. ***Neural Information Processing Systems (NeurIPS)*** Oral (Dec 2017) – Cited 20,000+ times

C

Collaboration with UW Nephrology

Explainable AI (XAI): Accurately predicting an outcome is vital, but the critical question revolves around ***why***.



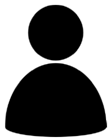
Lundberg et al. ***Nature Machine Intelligence***, *2020* – Featured on the Cover Beebe-Wang et al. ***IEEE JBHI***, 2021

C

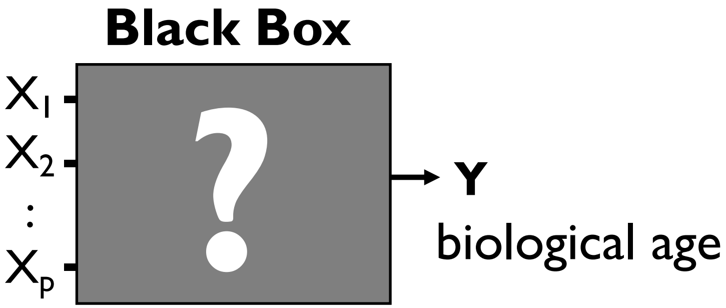
XAI for interpretable biological age

- ENABL (ExplaiNAble BioLogical)
  Age clock
  - Estimates an individual's biological age
  - Trained using the UK biobank data from 0.5M people based on 825 features:

From first principles of movement



demographics
lab tests
exam results
lifestyle
:



Wei, CSE PhD

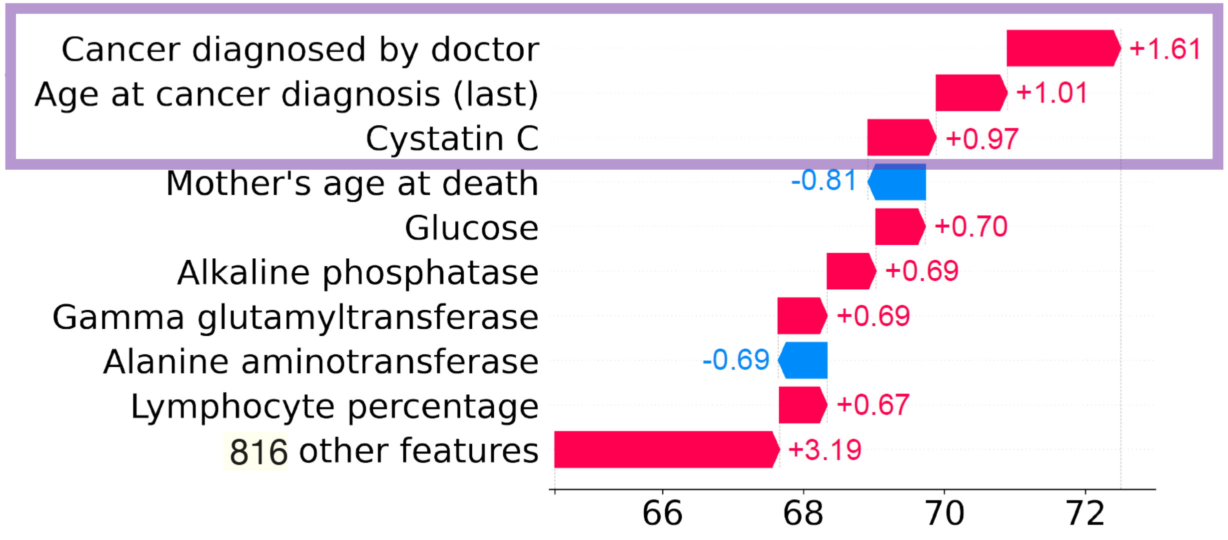
Qiu et al. ***Nature Comm. Medicine***, 2022

Qiu et al. ***Lancet Healthy Longevity***, 2023 – Featured on the Cover

C

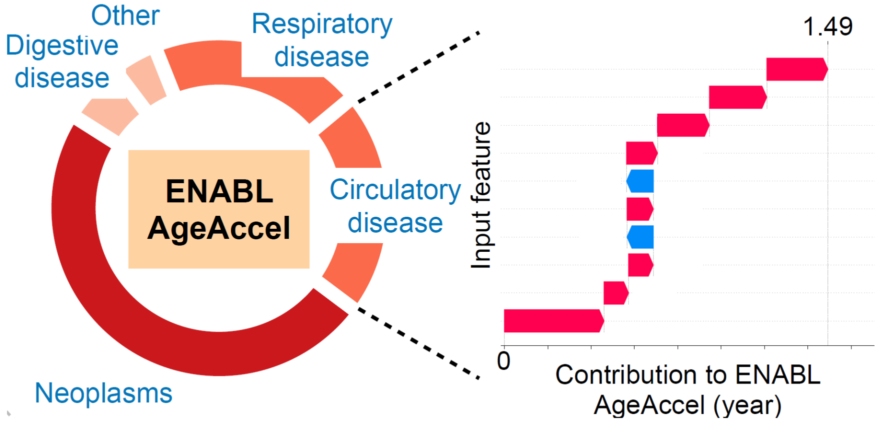
UW Nathan Shock Center for Basic Biology of Aging (AI Core Director role)

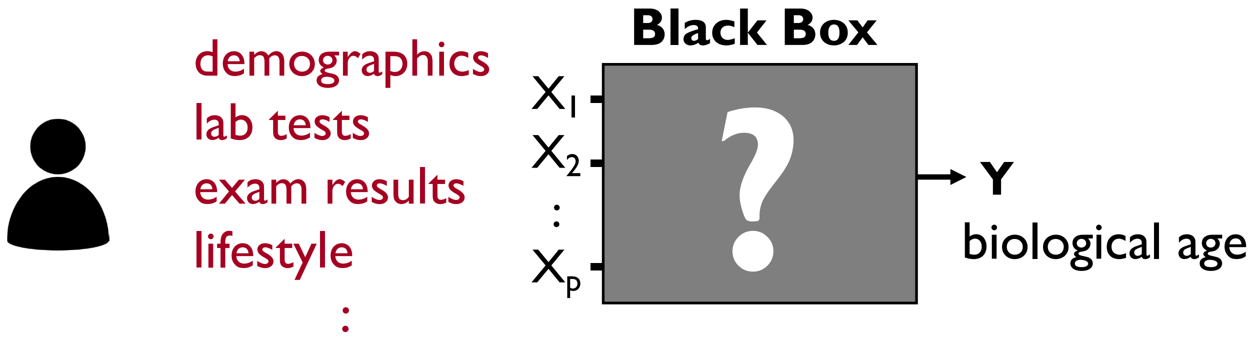
Explainable AI for interpretable biological age



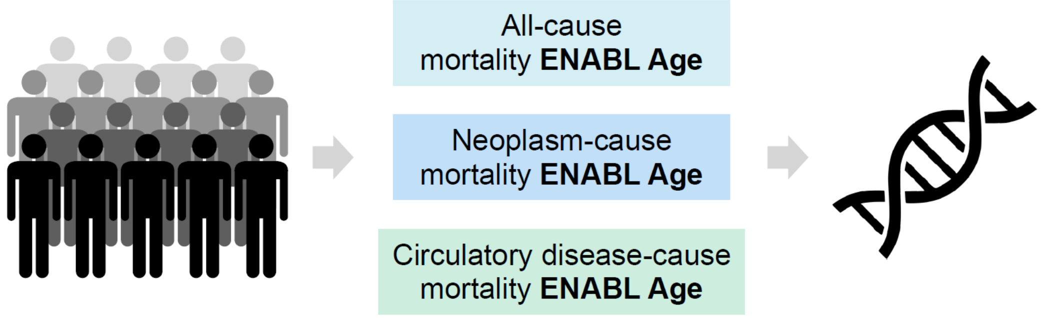
Chronological Age = 65 **ENABL Age = 72.51**

Impact of mortality causes on all-cause mortality





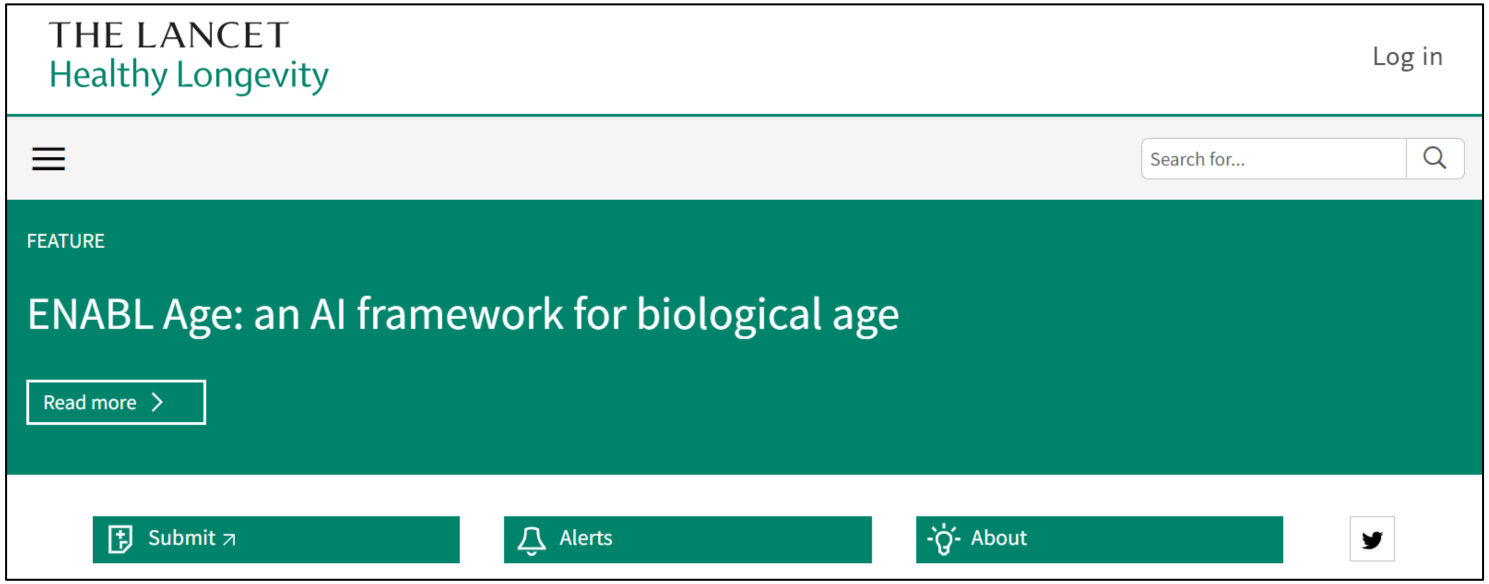
Genome-wide association study

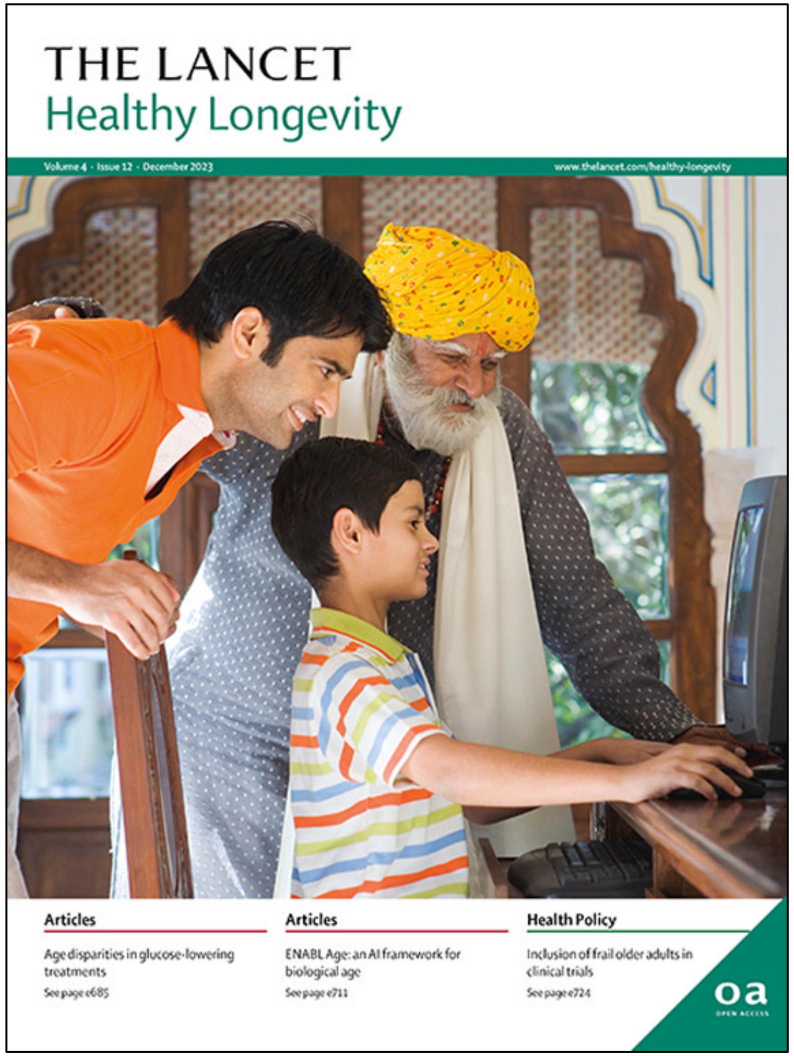


Qiu et al. ***Nature Comm. Medicine***, 2022 Qiu et al. ***Lancet Healthy Longevity***, 2023 – Featured on the Cover

ENABL age paper is now featured on the cover of Lancet Healthy Longevity.

- Please check it out!





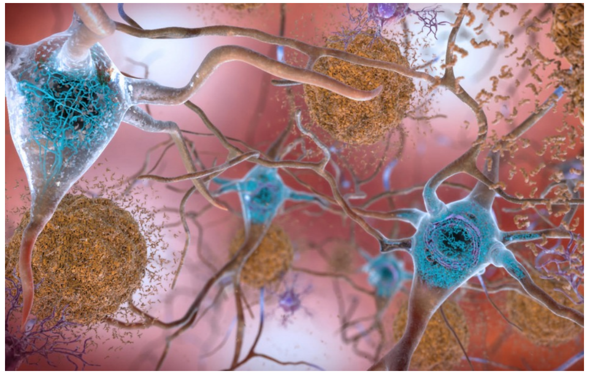
Wei, CSE PhD

B

Alzheimer's disease (AD)

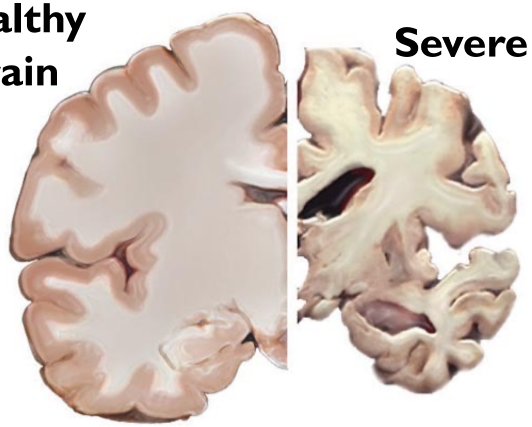
- 6th most common cause of death in the US
- No long-term effective therapy exists to delay or prevent onset of progression
- AD lacks effective treatments due to limited understanding of *early cellular pathways* leading to end-stage pathologies like amyloid-β (Aβ) and tau.

Amyloid-β (Aβ)



Tau

Healthy brain



Severe AD

B

Collaboration with Sara Mostafavi (UW)

The key question is the ***mechanistic explanation*** of complex neuropathological phenotypes

Postmortem brain

ACT
(n=337)

MSBB
(n=879)

ROSMAP
(n=542)

Features X

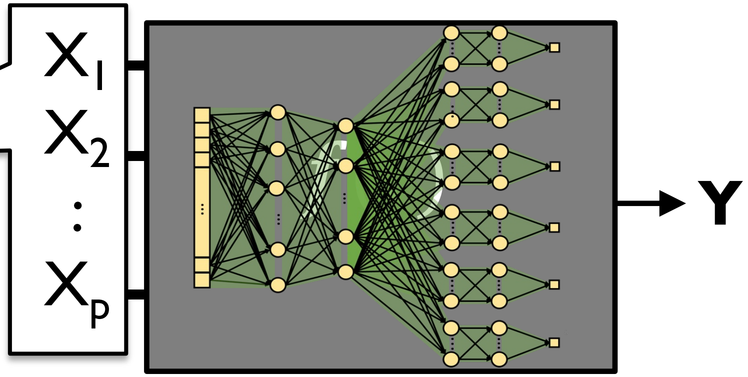
RNAseq

ATACseq

Demographic Data

Clinical data

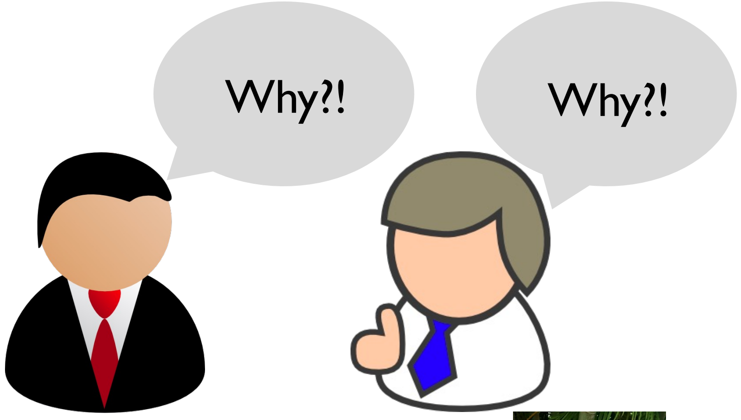
ML model



Outcomes Y

Neuropathological phenotypes:

Aß, Tau, CERAD score, Plaque counts, Braak stage, tangle counts



Neurobiologists

- Advantages
  - No need to harmonize phenotypes across cohorts
  - Ability to capture complex, non-linear gene-phenotype relationships

Nicasia, CSE PhD'22

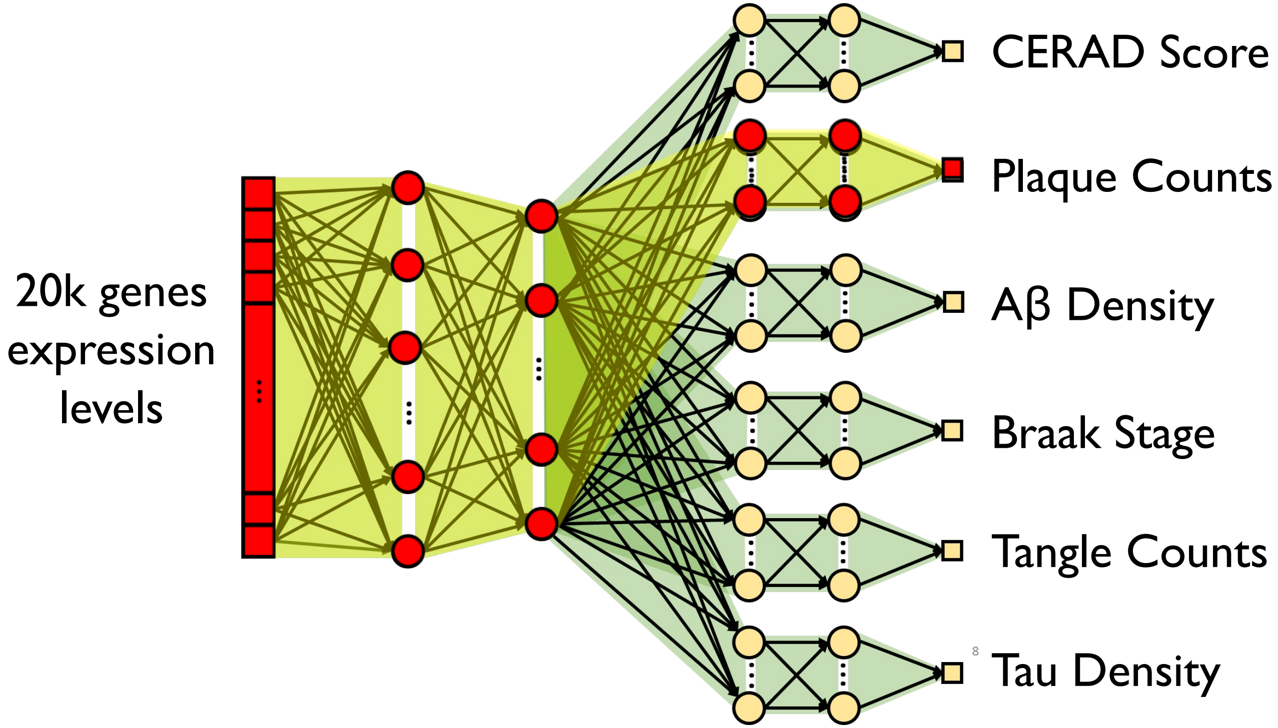
Beebe-Wang et al. ***Nature Communications***, 2021

B

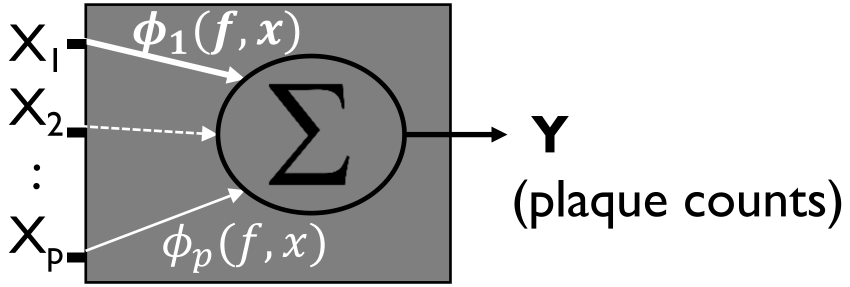
Collaboration with Sara Mostafavi (UW)

Explainable AI (XAI) enhances neurodegenerative disease research in multiple ways

- Our robust model trained across multiple cohorts was successfully validated, even in mouse brain and human blood datasets



- Using XAI, we can estimate each gene's contribution to AD neuropathologies
  - Previously unknown sex-specific associations btw. immune response genes and AD neuropathologies

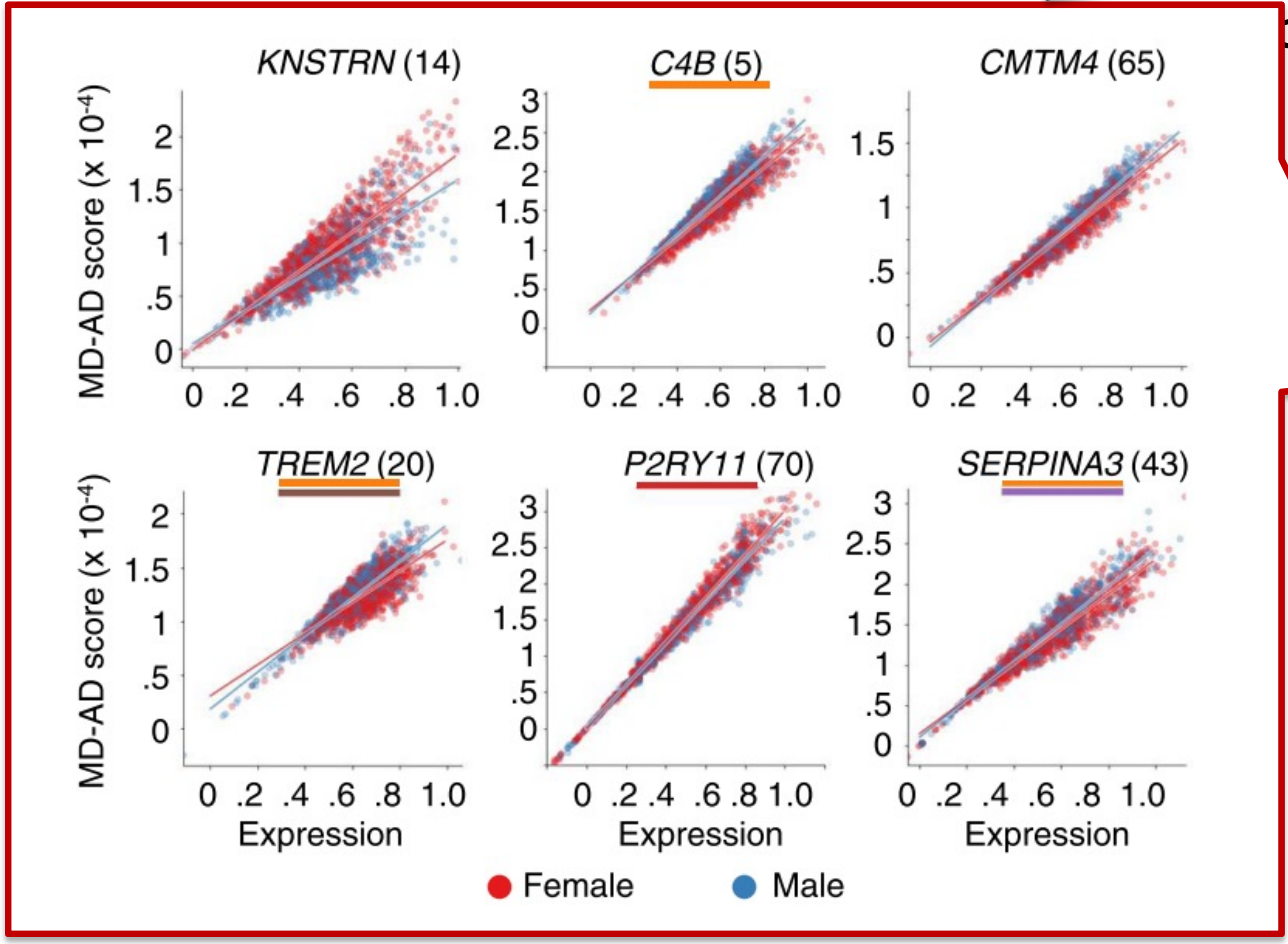


Beebe-Wang et al. ***Nature Communications***, 2021 Janizek et al. ***Nature Biomedical Engineering***, 2023

B

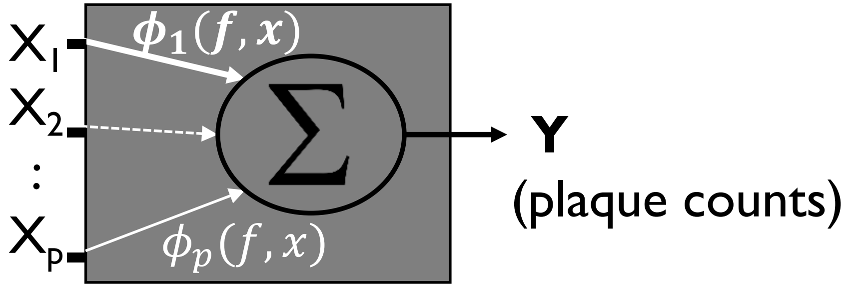
Explainable AI (X
disease research

XAI may capture patterns related to ***sex-differential microglia activity***.



...orts was successfully validated, even in

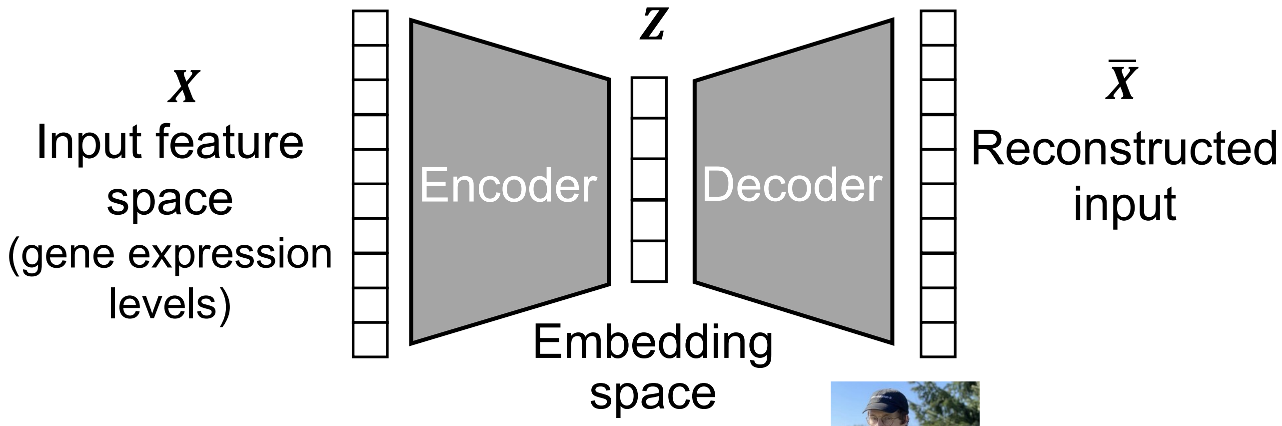
- Using XAI, we can estimate each gene's contribution to AD neuropathologies
  - Previously unknown sex-specific associations btw. immune response genes and AD neuropathologies

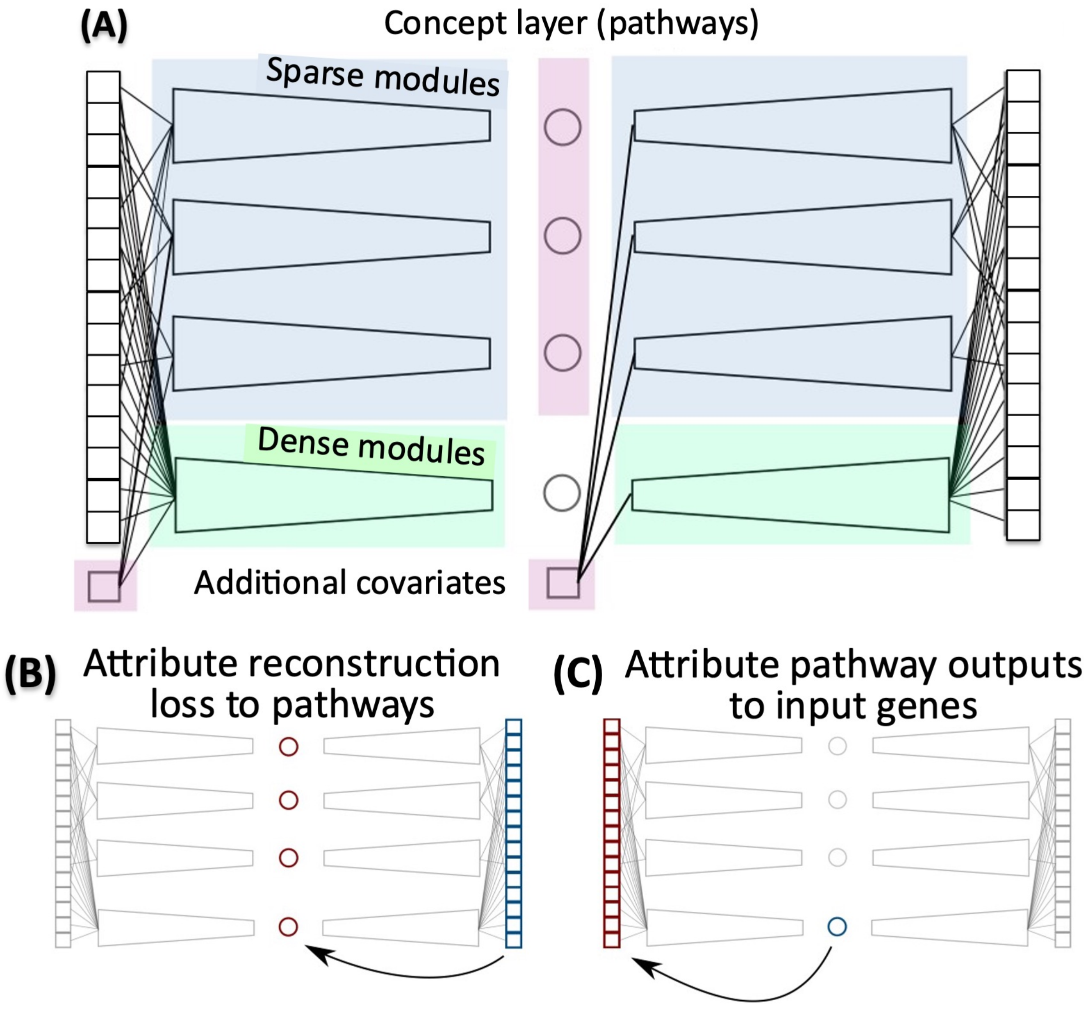


Beebe-Wang et al. ***Nature Communications***, 2021 Janizek et al. ***Nature Biomedical Engineering***, 2023

Biologically interpretable AI modeling further advances data-driven discovery

- Individual genes are not as interpretable as functional units (e.g., pathway)
- Unsupervised modeling enables the incorporation of unlabeled data
  - XAI can pinpoint crucial genes that explain the expression variation within the dataset





Janizek et al. ***Genome Biology***, 2023

Joe, UW MSTP/CSE PhD'22 (Matched in Radiology at Stanford)

B

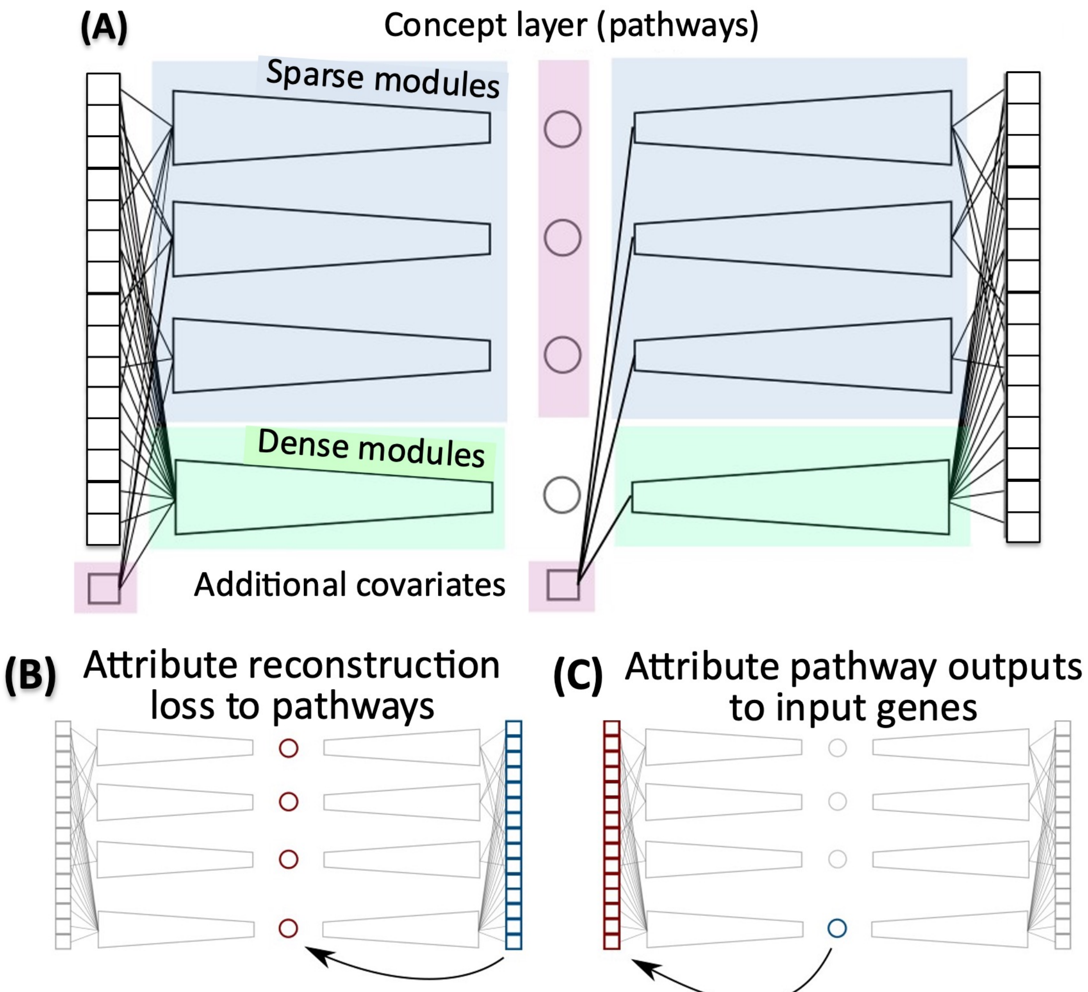
Collaboration with UW Laboratory Medicine & Pathology (Matt Kaeberlein)

Biologically interpretable modeling identifies experimentally validated AD therapeutic targets

- We applied our approach to extended bulk RNAseq datasets from AD study cohorts
- We identified mitochondrial complex I as a potential mediator for tolerance to Aβ toxicity
  - *In vivo* validation in a transgenic *C. elegans* model expressing Aβ done by Matt Kaeberlein's lab

A promising pharmacological avenue!

Capsaicin

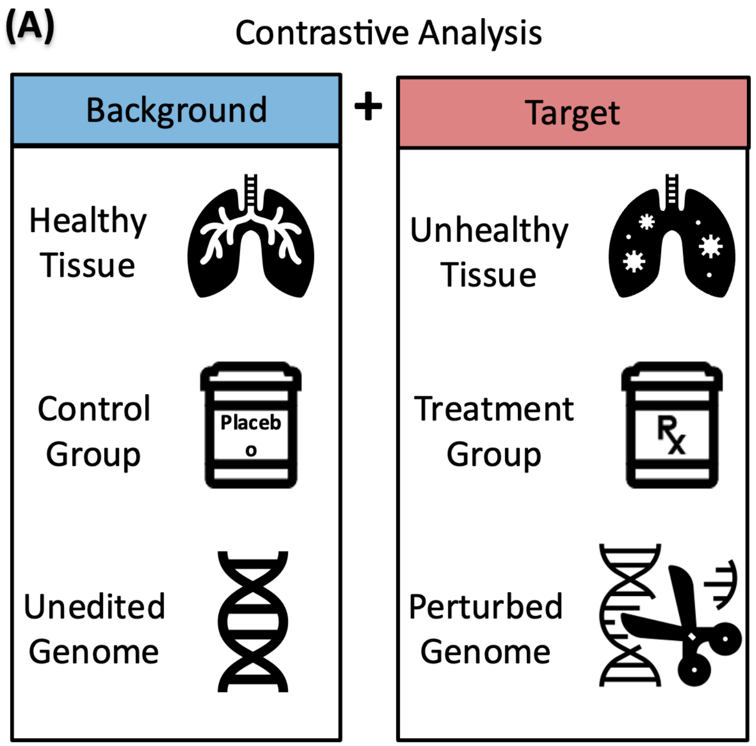


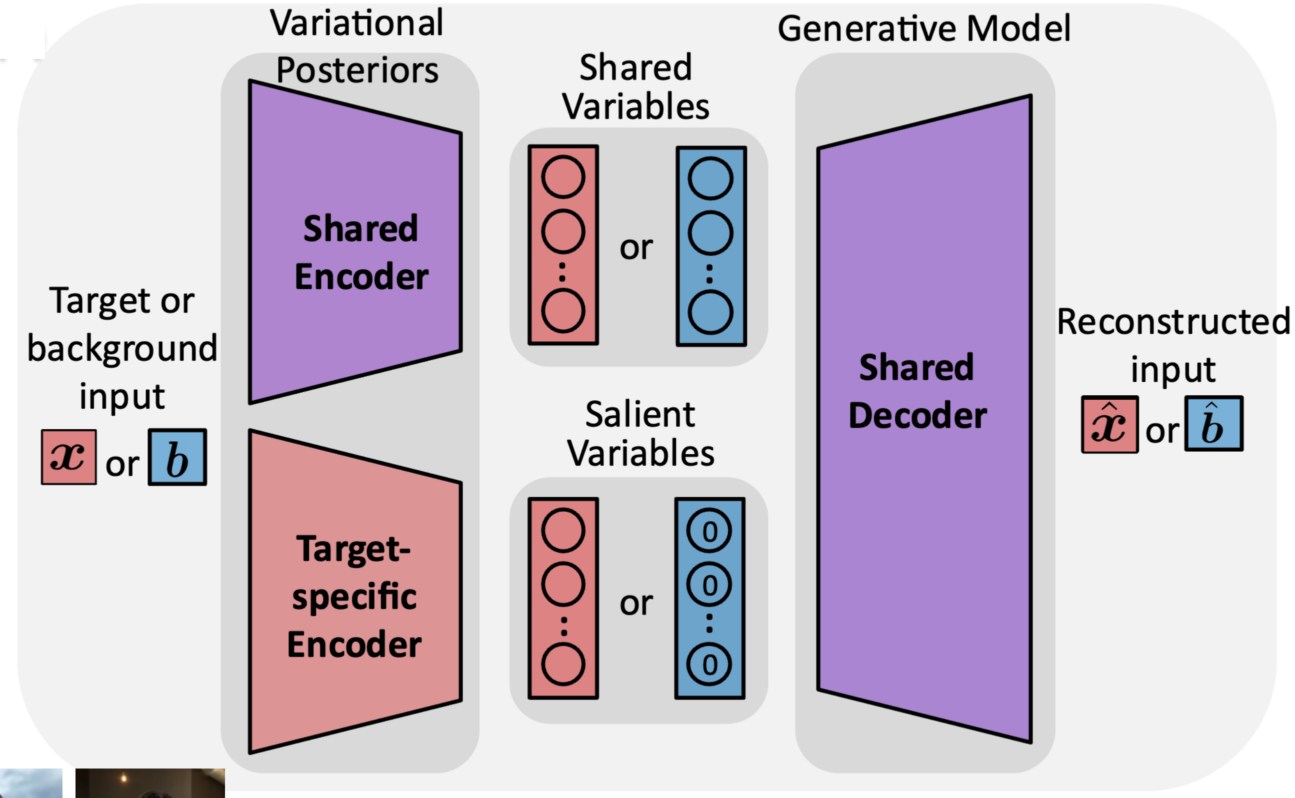
Janizek et al. ***Genome Biology***, 2023

B

Contrastive modeling enhances interpretability

- Single-cell datasets are often collected to investigate differences in cellular state between background cells and those under specific treatments





Weinberger,* Lin,* and Lee. ***Nature Methods***, 2023

Ethan & Chris, CSE PhD

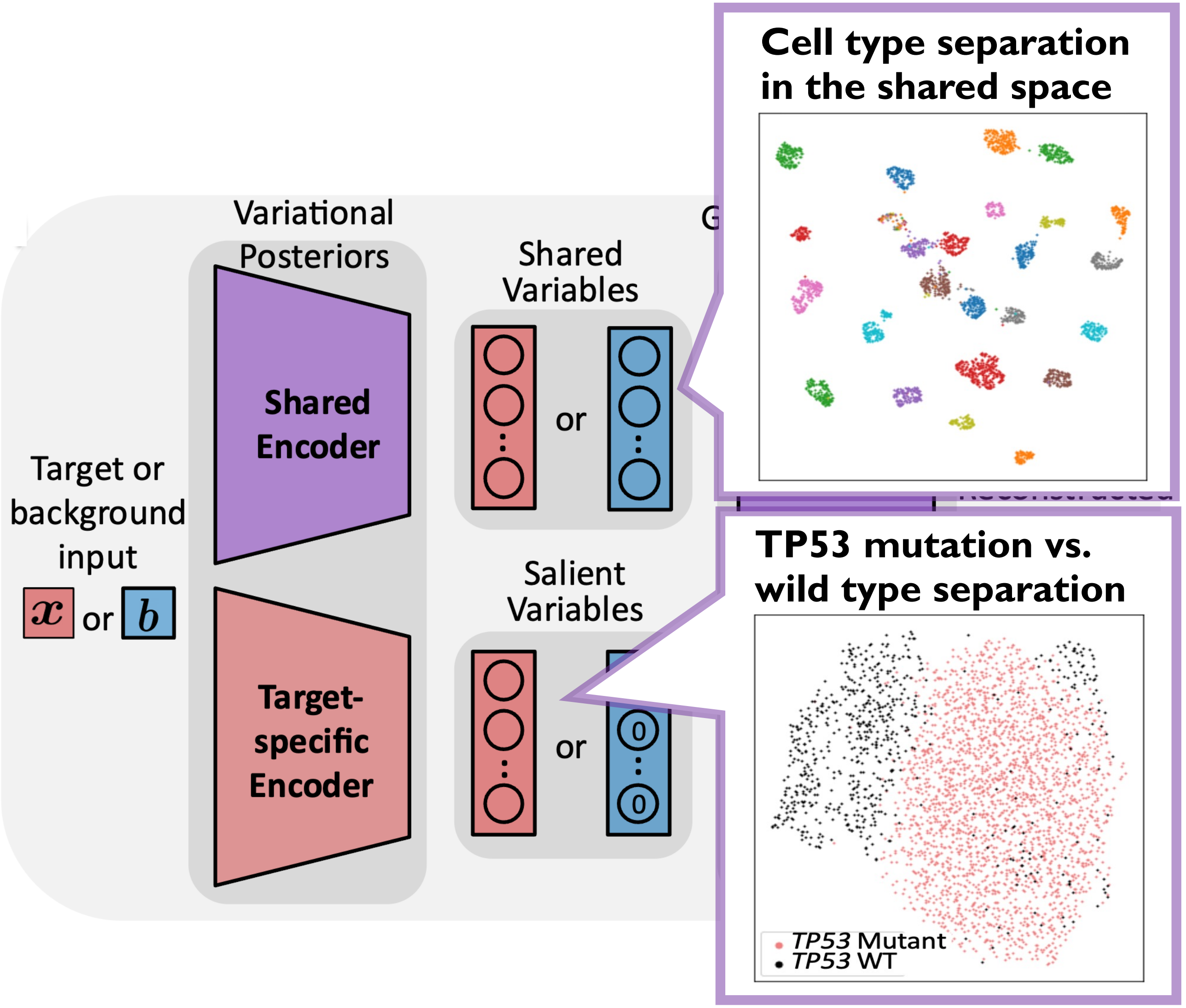
B

Contrastive modeling enhances interpretability

- Cancer cells treated with idasanutlin *vs.* untreated as background
  - Cells behave differently in salient space depending on their TP53 mutation status

Important implications for personalized medicine!

- How about AD *vs.* control brain tissue?
  - What drives neurodegeneration (in collaboration with Jessica Young)
  - What drives biological aging process? (Jessica Young & Suman Jayadev)



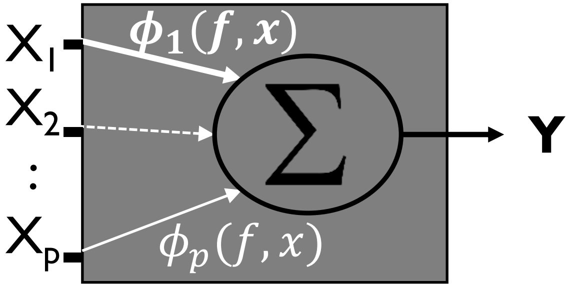
Weinberger,* Lin,* and Lee. ***Nature Methods***, 2023

Outline – Two parts

- Part I – What explainable AI can do in biological research
  - Demystifying the biological age
  - Unveiling neurodegenerative disease insights with explainable AI

- Part II – Beyond explaining models
  - Cancer therapy design for precision oncology
  - Model auditing
  - Cost-aware clinical AI

Beyond interpreting models…



– Cancer therapy design for precision oncology [*Nature BME'23*]

– AI auditing [*Nature MI'21, Nature BME'23, Nature Medicine'24*]

radiology, dermatology

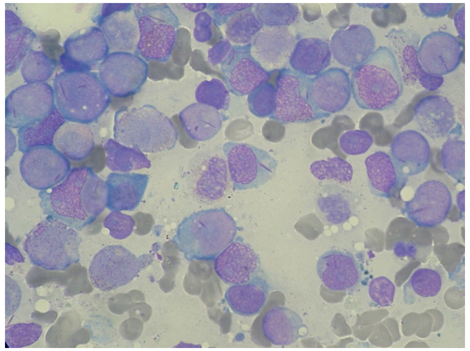
– Cost-aware clinical AI [*Nature BME'22*]

emergency medicine, critical medicine

B

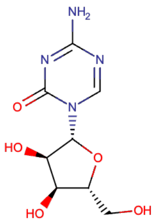
Collaboration with Harvard Medical School (Kamila Naxerova)

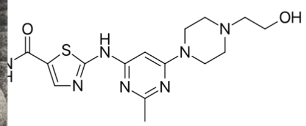
Explainable AI to design cancer therapy

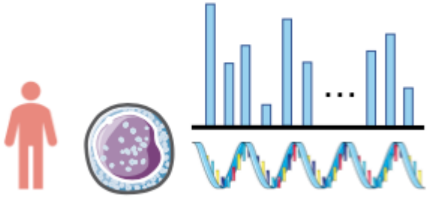


- Cancers are increasingly treated by combination therapy
  - Choosing drugs tha
  - ways
  - Greater efficacy
  - Fewer side-effects
- Choosing optimal c
  - Explanations to th
  - portant

Hundreds of individual drugs







AML Gene Expression

irs of Drugs

MSTP/CSE PhD'22 (got matched to Stanford Radiology)

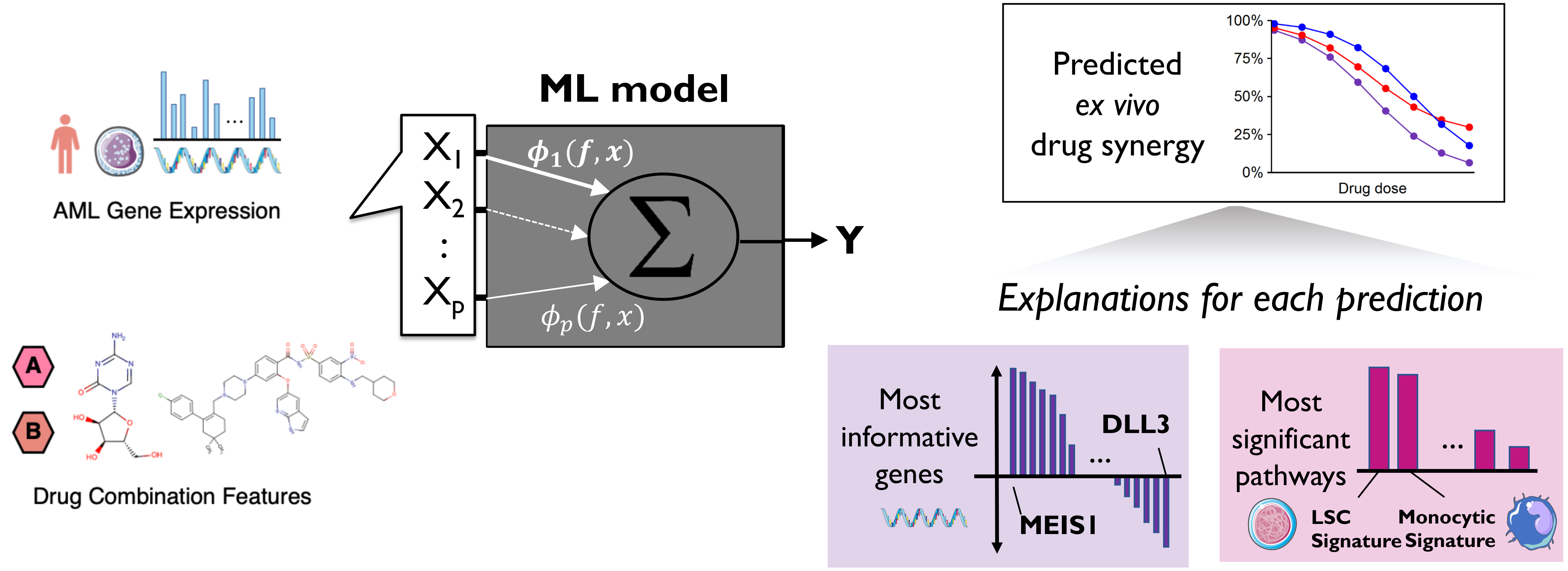
Janizek et al. ***Nature Biomedical Engineering***, 2023 Lee*, Celik*, et al. ***Nature Comm.***, 2018

B

Collaboration with Harvard Medical School (Kamila Naxerova)

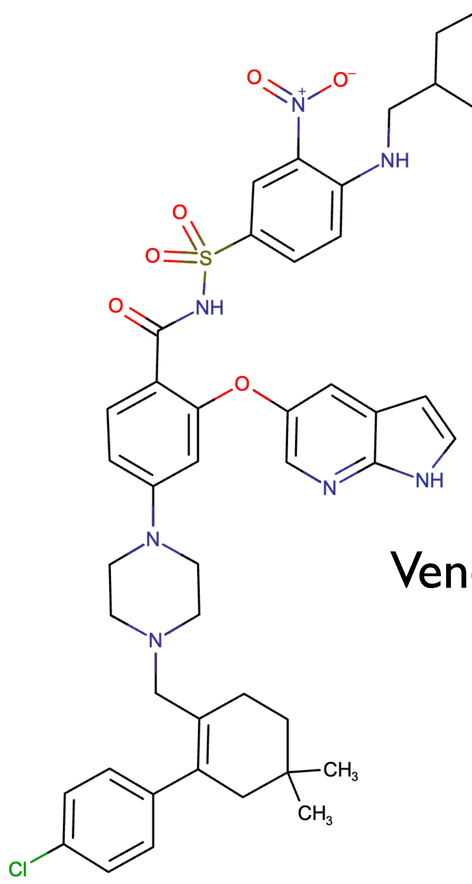
Explainable AI to design cancer therapy

- EXPRESS: Explainable prediction of drug synergy

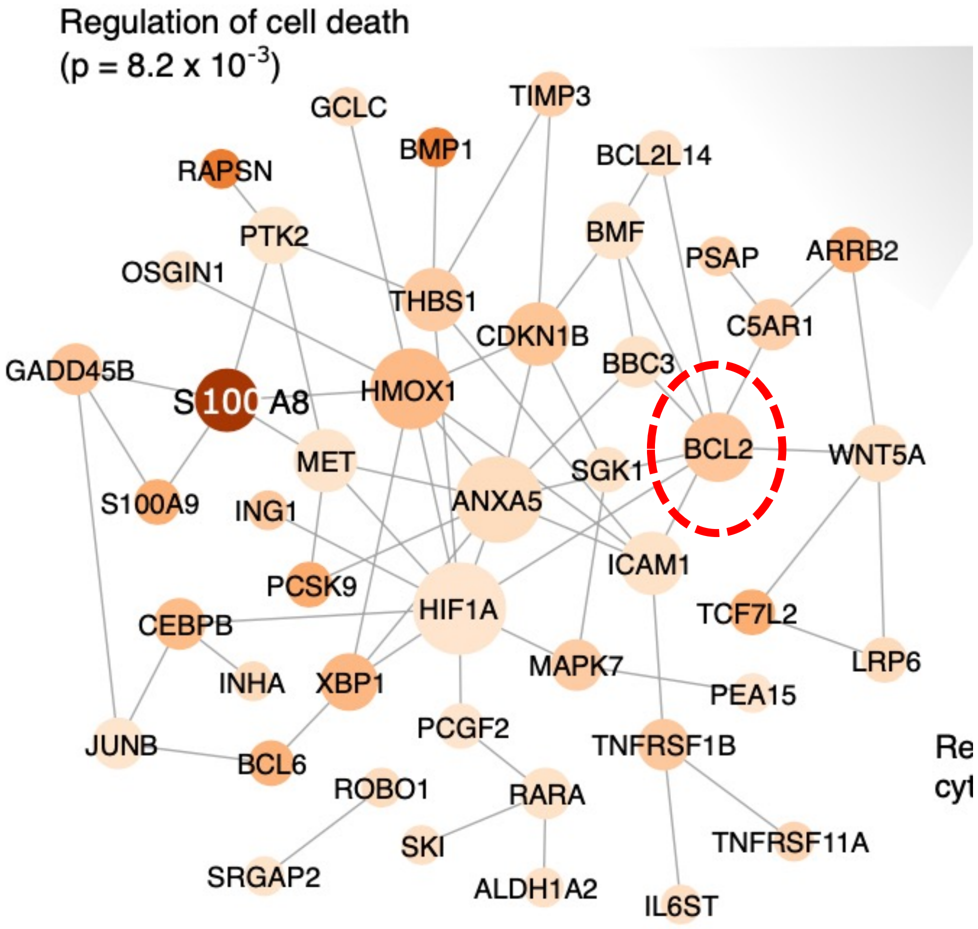


Janizek et al. ***Nature Biomedical Engineering***, 2023 Lee*, Celik*, et al. ***Nature Comm.***, 2018

Interpretability allows us to validate our model's decisions



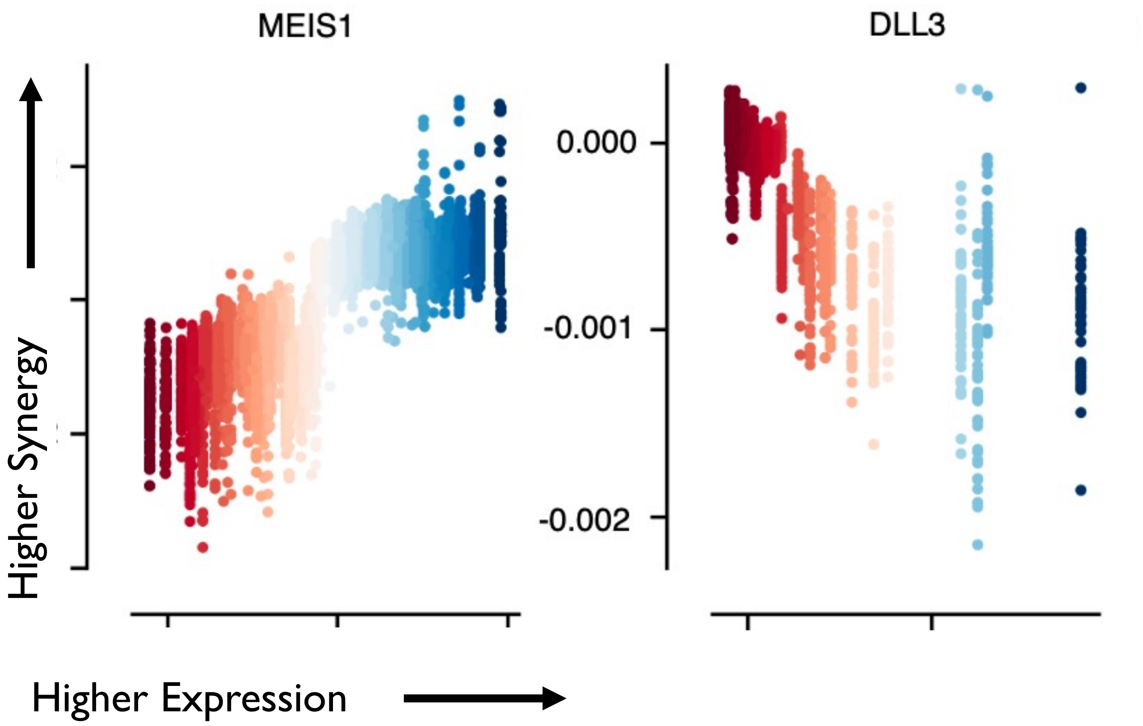
Venetoclax, BCL-2 inhibitor



Janizek et al. ***Nature Biomedical Engineering***, 2023   Lee*, Celik*, et al. ***Nature Comm.***, 2018

Collaboration with Harvard Medical School (Kamila Naxerova)

Interpretability uncovers transcription programs underlying drug synergy



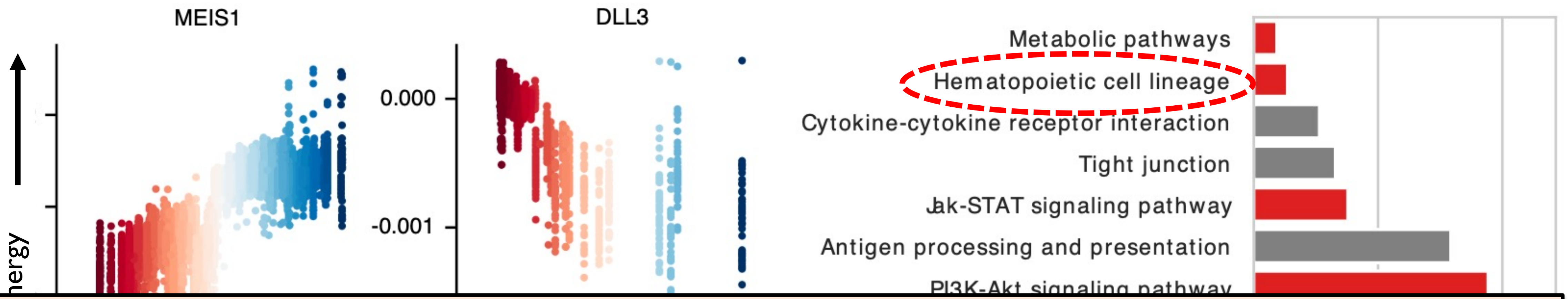
Linked to prognosis

Related to hematopoietic differentiation

Janizek et al. ***Nature Biomedical Engineering***, 2023 Lee*, Celik*, et al. ***Nature Comm.***, 2018

Collaboration with Harvard Medical School (Kamila Naxerova)

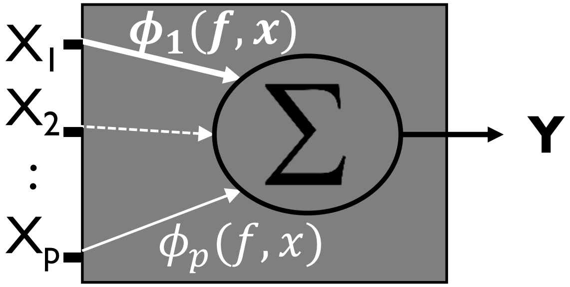
Interpretability uncovers transcription programs underlying drug synergy



"Stemness" can be considered as an "axis" to design combination therapies – Two drugs that target different differentiation stages of cancer are likely effective.

Janizek et al. ***Nature Biomedical Engineering***, 2023   Lee*, Celik*, et al. ***Nature Comm.***, 2018

Beyond interpreting models…

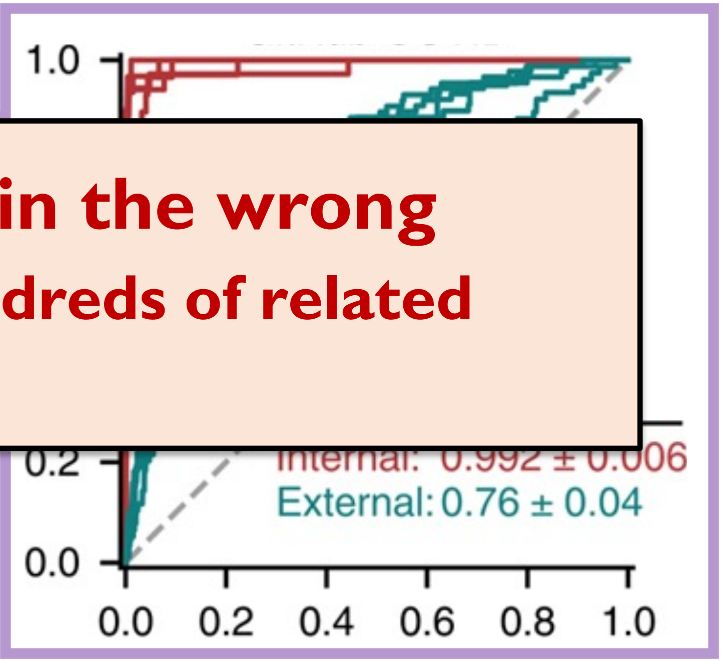


- Cancer therapy design for precision oncology *[Nature BME'23]*
- AI auditing *[Nature MI'21, Nature BME'23, Nature Medicine'24]*
  radiology, dermatology
- Cost-aware clinical AI *[Nature BME'22]*
  emergency medicine, critical medicine

C

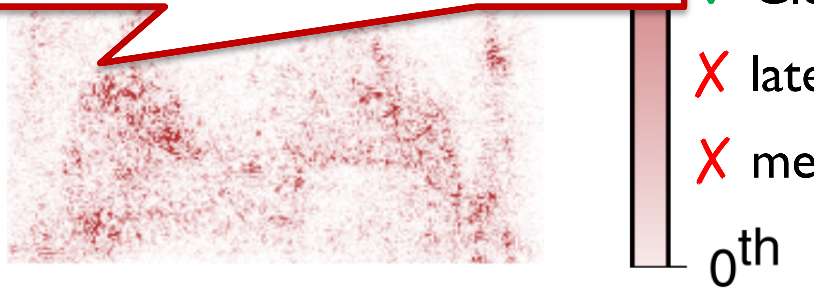
Auditing AI for COVID-19 detection using XAI

- Many published AI models that detect COVID-19



XAI helped us to stop the field from moving in the wrong direction – There were 6 published papers and hundreds of related models out there that learned the shortcuts.

Many kinds of analyses for model auditing presented in the paper!



✓ Clear lung bases predict negative COVID-19 status

✗ laterality markers should not predict negative status

✗ medical devices should not predict negative status

MSTP / CSE PhD

DeGrave,* Janizek* et al. ***Nature Machine Intelligence***, 2021 Cited 440+, Featured in *Nature*, 2022

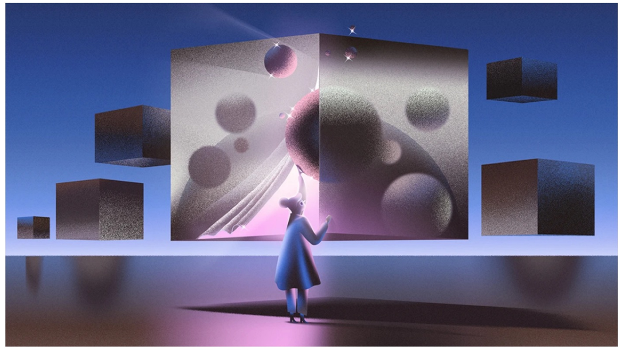
Our AI auditing work featured in *Nature*

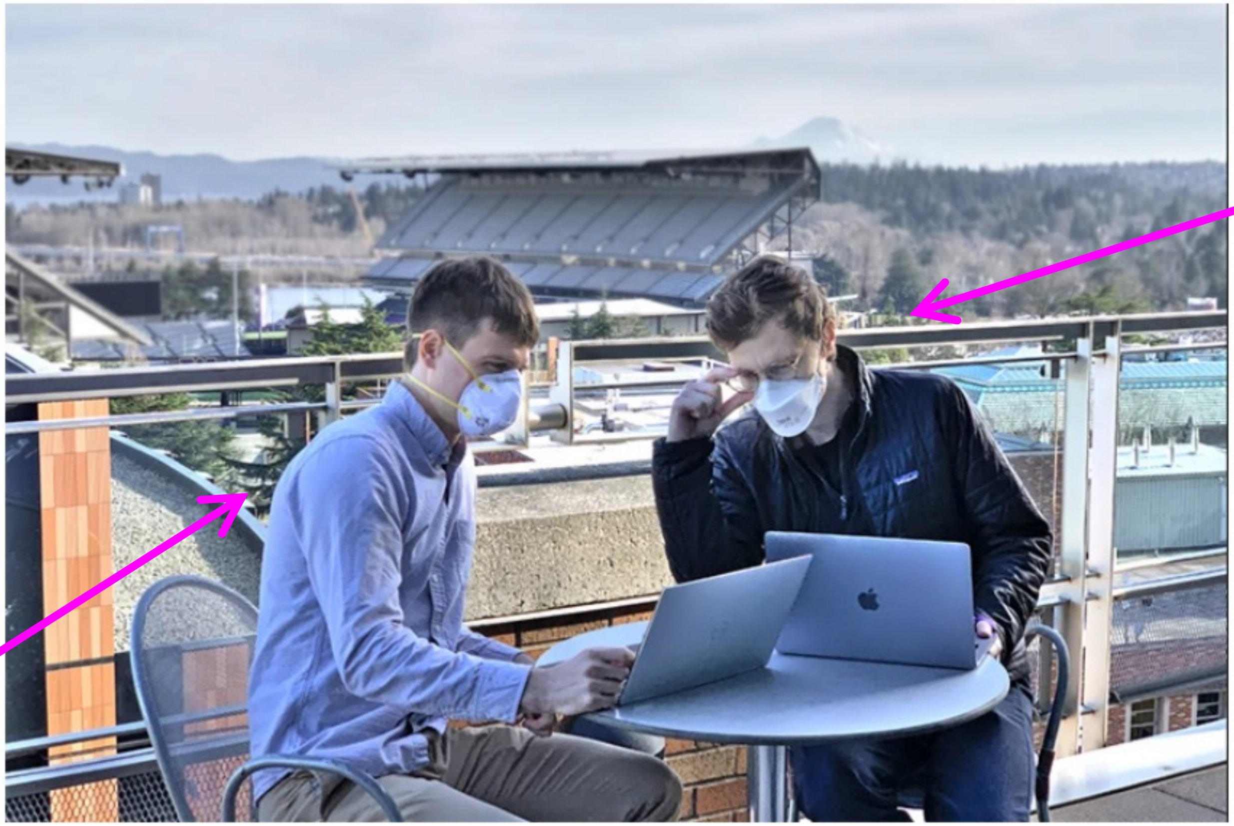
- "Breaking into the black box of artificial intelligence" *Nature* Outlook

Breaking into the black box of artificial intelligence

Scientists are finding ways to explain the inner workings of complex machine-learning models.

By Neil Savage





Alex DeGrave and Joseph Janizek are students on the Medical Scientist Training Program at the University of Washington, in Seattle. Credit: Alex DeGrave

UW MSTP / CSE PhD **Joe Janizek** (residency at Stanford)

UW MSTP/CSE PhD student **Alex Degrave**

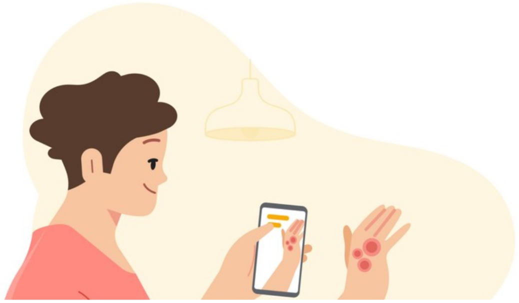
DeGrave,* Janizek* et al. ***Nature Machine Intelligence***, 2021 Cited 440+, Featured in *Nature*, 2022

C

Collaboration with Stanford Dermatology (Roxana Daneshjou)

Further digging into the flaws in the reasoning processes of clinical AI – dermatology

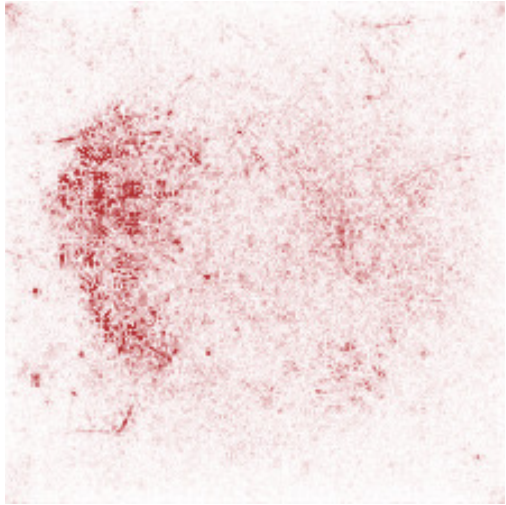
- Auditing AI models to predict skin cancer
  - Five models – 2 academic models, 2 commercial devices, and 1 competition winner
- Technical challenges – saliency maps often do not work



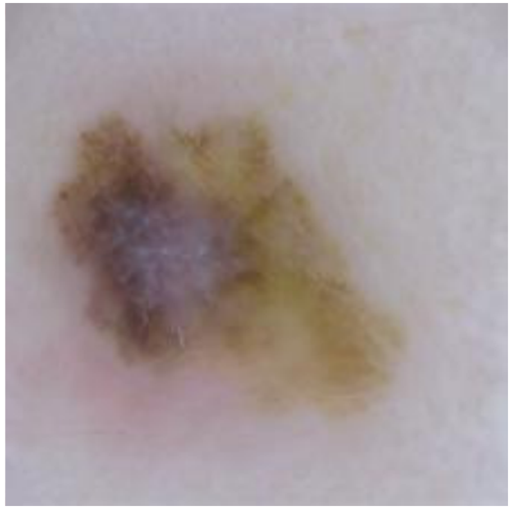
Original image

Predicted: benign

Saliency map



Modified image



Predicted: malignant

Our solution #1

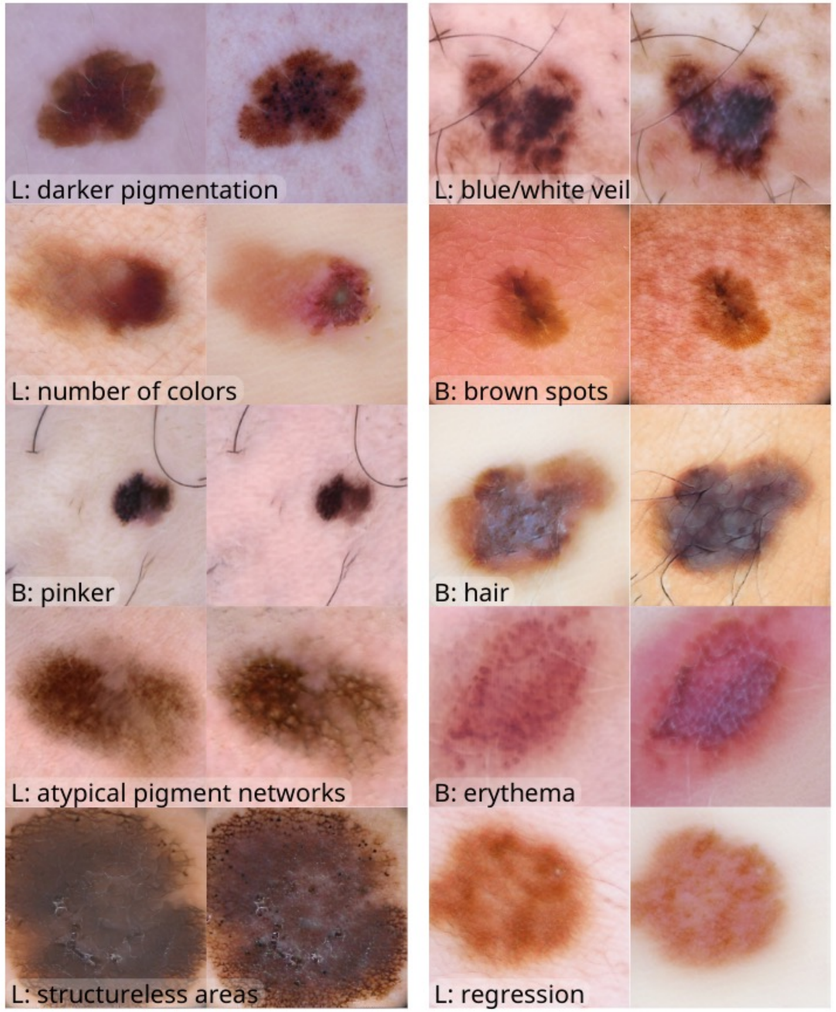
- Generate counterfactual images ***from the AI model***
- Systematic characterization by experts: Drs. Roxana Daneshjou, and Zhuo Ran Cai (Stanford)

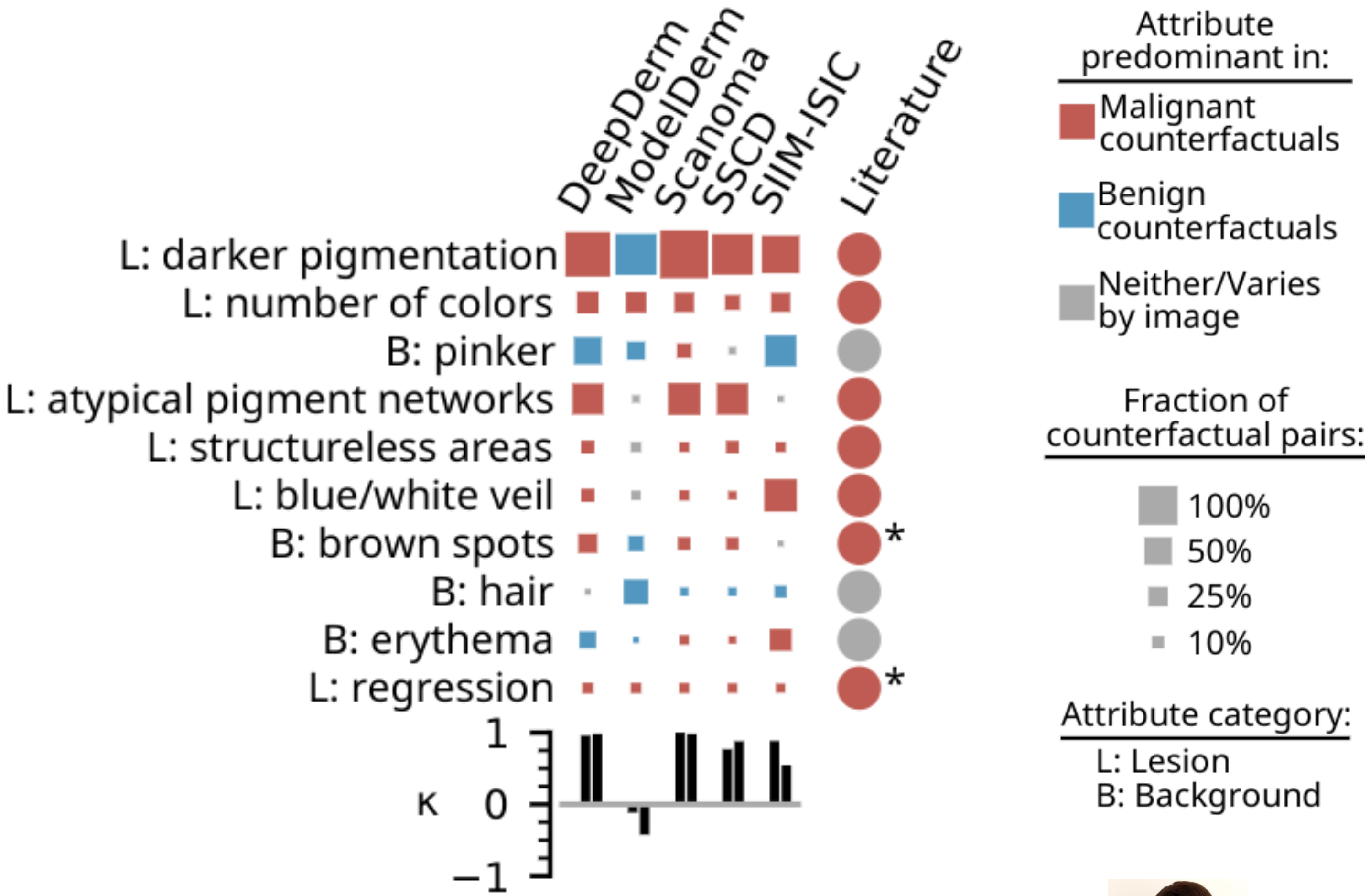
DeGrave et al. (*Nature Biomedical Engineering*)

Kim et al. (*Nature Medicine*, 2024)

Collaboration with Stanford Dermatology (Roxana Daneshjou)

How do dermatology AI systems make decisions on dermoscopic images?





Degrave, Ran Cai, Janizek, Daneshjou,* and Lee* ***Nature Biomedical Engineering***, 2023

MSTP / CSE PhD

The *Lancet* perspective (Feb 2024)

- Broader promises of counterfactual AI

"The clinical potential of counterfactual AI"

by Su-In Lee* and Eric Topol

Digital medicine

The clinical potential of counterfactual AI models

Clinicians frequently use conditional reasoning for treatment decisions by envisioning potential outcomes for patients. This is counterfactual thinking, exploring "what if" scenarios. Developments in generative artificial intelligence (AI) enable us to simulate this patient-level reasoning at the data level, opening new opportunities for science and health care. We term this approach counterfactual AI.

This approach is exemplified by use of counterfactual images in dermatology. Using AI, original skin images were modified to resemble melanoma guided by the decision-making process of a particular AI-based dermatological classifier. Dermatologists were then tasked with identifying clinically relevant features in the counterfactual images of melanoma and normal conditions. This process elucidated the reasoning processes of five AI-based dermatological classifiers. This data-centric counterfactual AI aligns the reasoning processes of AI classifiers with human clinicians' intuition, establishing a new approach to auditing clinical AI classifiers. Model auditing provides insights into the performance of deployed clinical AI classifiers for patients, clinicians, regulators, and data scientists.

Such uses of counterfactual AI prompt a crucial question: how might patient data change under specific conditions such as genetic mutations, treatments, time, or ageing? This exploration leads to intriguing scenarios, including forecasting the progression of clinical images or other data types over time for a particular treatment, potentially providing prognostic insights, or simulating the impact of genetic mutations to enhance our comprehension of disease mechanisms and treatment outcomes. This could present a frontier for future research. For example, personalised T-cell receptor sequence design for immunotherapy offers possibilities for new treatment strategies. Moreover, counterfactual AI has the potential to fill data gaps for rare diseases or under-represented groups, aiding the development of more inclusive and comprehensive health-care solutions. Furthermore, counterfactual AI could spur innovation in scientific hypothesis generation for drug discovery and development, potentially leading to breakthroughs in urgent areas such as Alzheimer's disease. Research suggests it could generate data on specific pathological conditions and conduct in-silico synthetic lethality testing for novel combination therapies.

An unexpected synergy is emerging as data-centric counterfactual AI contributes to the interpretation and auditing of clinical AI models. There are challenges in understanding the decision-making processes of many AI models. Saliency maps or, more broadly, feature attribution methods, are commonly used for model interpretation, indicating the areas of an image (or other data types) that the AI model focuses on (figure). Yet they provide only a partial view of the inner workings of complex AI models, impeding efforts to identify flaws in clinical AI reasoning processes. Counterfactual AI expands the scope of explainable AI by providing counterfactual images that elicit specific outcome predictions from complex AI classifiers (figure), enabling humans to grasp more comprehensive insights into the reasoning processes of these classifiers. Collaborating with clinicians, counterfactual AI could unearth previously unnoticed image attributes. Research indicates that by partnering with AI methods capable of automatically annotating images with an array of semantically meaningful concepts, counterfactual AI can systematically probe AI classifiers about how these concepts affect their decision-making processes.

Counterfactual AI in medicine faces ethical concerns and challenges related to fairness, data quality, and generalisability. Obtaining high-quality, diverse datasets is difficult. Generalising to new data is also problematic, particularly across diverse patient populations and health-care settings. Moreover, ethical and regulatory issues, including patient privacy concerns about the use of training data, must be addressed to ensure responsible AI deployment in health care.

What should we do to fully leverage the potential of counterfactual AI to advance scientific and therapeutic discovery? Generative AI operates through complex models that necessitate explanation. The reciprocal relation between generative AI and explainable AI is essential: generative AI informs the development of explainable AI; explainable AI aids in understanding generative AI models. By focusing on these principles, we can ensure that "what if" AI models are transparent and interpretable, facilitating their effective use in biomedical endeavours.

*Su-In Lee, Eric J Topol

Paul G Allen School of Computer Science & Engineering, University of Washington, Seattle, WA 98195, USA (S-IL); Scripps Research Translational Institute, La Jolla, CA, USA (EJT)

suinlee@cs.washington.edu

Further reading

DeGrave AJ, Cai ZR, Janizek JD, Daneshjou R, Lee SI. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *Nat Biomed Eng* 2023; published online Dec 28. https://doi.org/10.1038/s41551-023-01160-9

DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell* 2021; **3:** 610–19

Kim C, Gadgil SU, DeGrave AJ, et al. Transparent medical image AI via an image-text foundational model grounded in medical literature. *Nat Med* 2024 (in press)

S-IL is supported by an NIH grant (R01 AG061132) and is the recipient of a 2024 International Society for Computational Biology (ISCB) Innovator Award. EJT is supported by the NIH/National Center for Advancing Translational Sciences grant UL1TR001114.

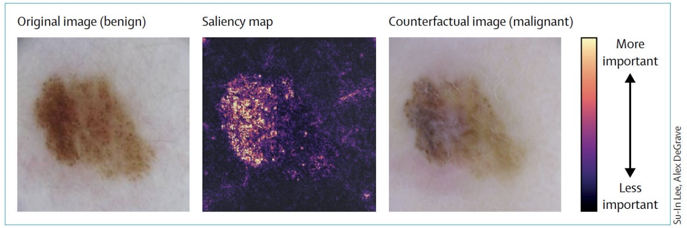
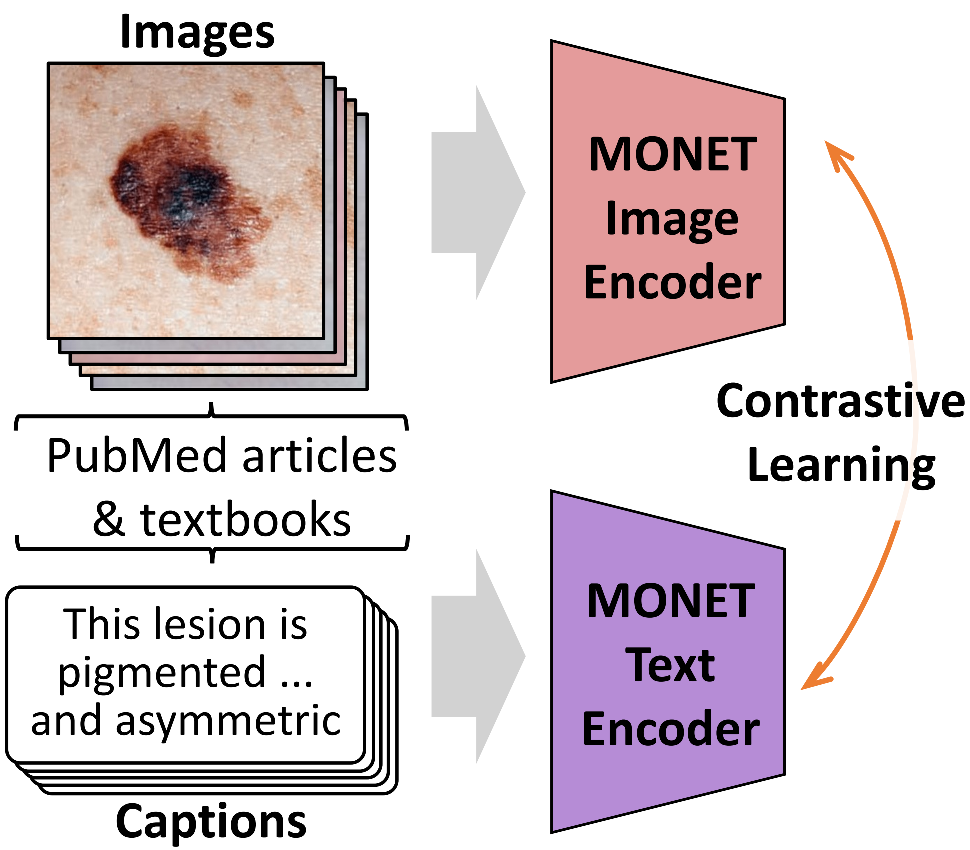


***Figure:* Auditing dermatology AI model with counterfactual AI**
A saliency map indicates little about an AI system for detecting melanoma, whereas counterfactual AI reveals that the system relied on the colour and pattern of pigmentation to determine that this lesion is benign.

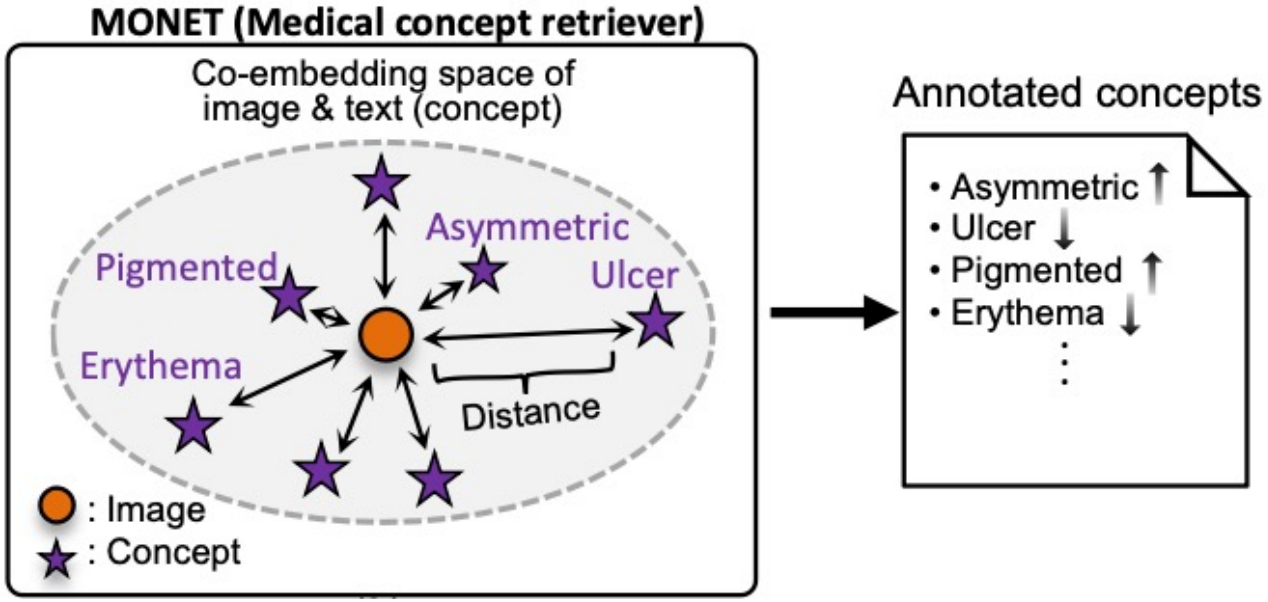
Collaboration with Stanford Dermatology (Roxana Daneshjou)

Fostering transparent AI via an ***image-text foundation model*** grounded in medical literature

- Finetune the CLIP (contrastive language-image pretraining) model



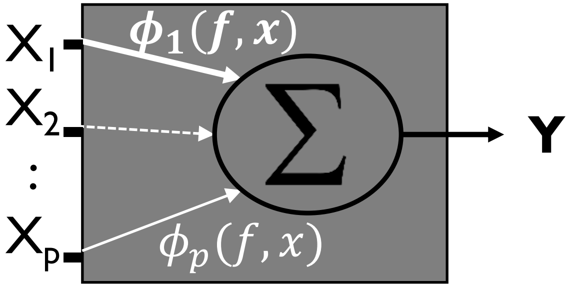
- Automatic concept annotation:
  - For each image,



Chanwoo, CSE PhD

Kim et al. ***Nature Medicine***, 2024

Beyond interpreting models…



- Cancer therapy design for precision oncology *[Nature BME'23]*
- AI auditing *[Nature MI'21, Nature BME'23, Nature Medicine'24]*
  radiology, dermatology
- Cost-aware clinical AI *[Nature BME'22]*
  emergency medicine, critical medicine

C

Collaboration with UW Emergency Medicine (Nathan White)

Explainable AI enables "cost-aware" AI (CoAI)

One year ago …

Gabe, MSTP/CSE PhD'21

(now Harvard for residency in EM)

Erion et al. ***Nature Biomedical Engineering***, 2022 - Featured in *Nature Comp. Science*, 2022

C

Collaboration with UW Emergency Medicine (Nathan White)

Explainable AI enables "cost-aware" AI (CoAI)

- Gathering features is often costly. (e.g., time, money, etc)
  - Acute traumatic coagulopathy (ATC), a dangerous bleeding disorder in trauma patients (failure to clot)
  - ATC is time sensitive – often requires massive transfusion and earlier transfusion leads to better outcomes

- In collaboration with Nathan White, we used our trauma registry dataset
  - 14,000 emergency room visits and 46 features from the trauma registry of Harborview Medical Center, an urban level-I trauma centre
- CoAI combines XAI-based feature importance with feature cost (time)
  - Time cost survey from clinicians, medical directors, EMTs, etc

Gabe, MSTP/CSE PhD'21
(now Harvard for residency in EM)

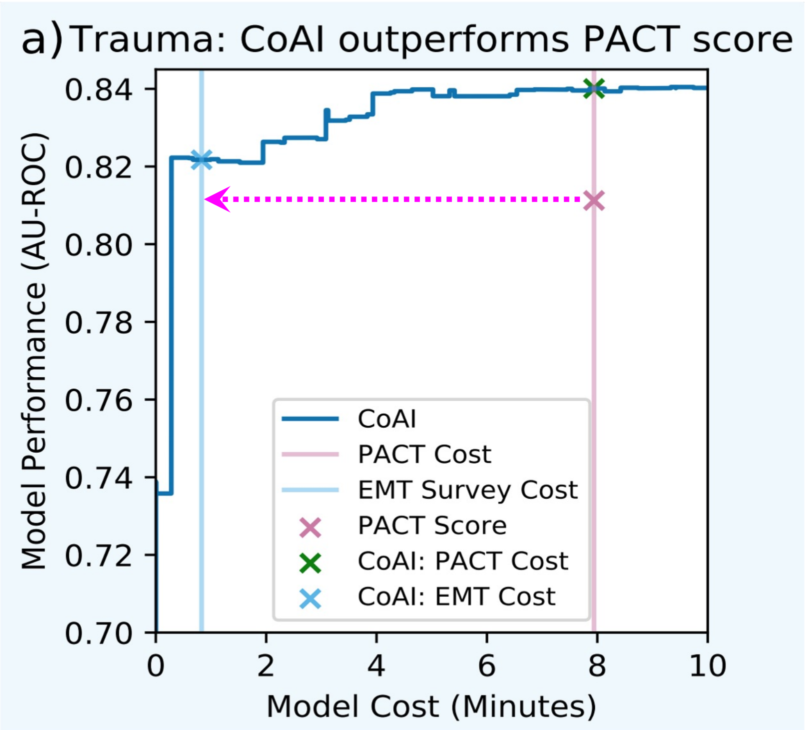
Erion et al. ***Nature Biomedical Engineering***, 2022 - Featured in *Nature Comp. Science*, 2022

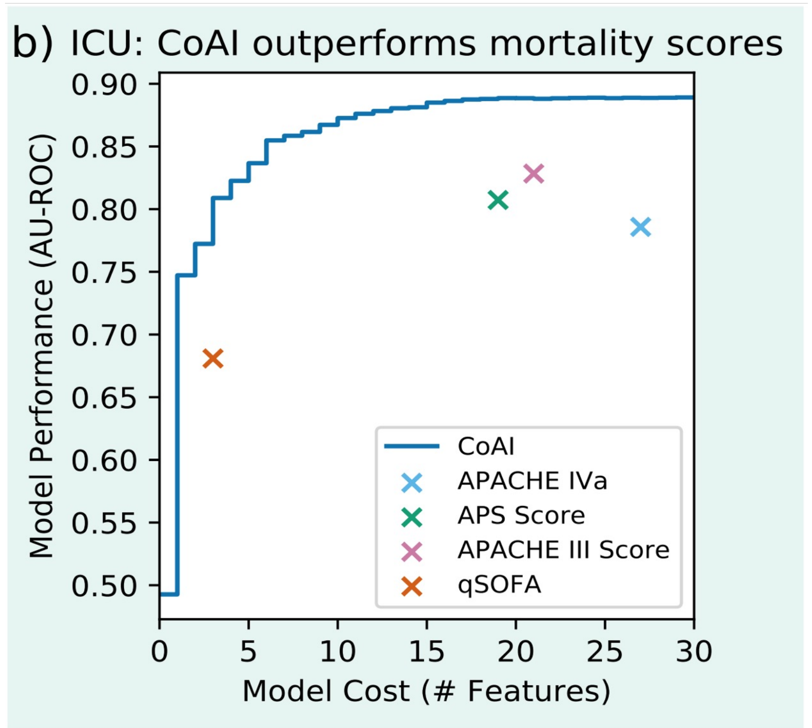
C

Collaboration with UW Emergency Medicine (Nathan White)

Explainable AI enables "cost-aware" AI (CoAI)

- CoAI improves *both* cost & accuracy
  - As accurate as the existing PACT score with <1 mins (vs. 8 mins) of feature gathering time





- CoAI is a general framework
  - Improves many existing clinical risk scores when applied to ICU mortality prediction

Erion et al. ***Nature Biomedical Engineering***, 2022 - Featured in *Nature Computational Science*, 2022

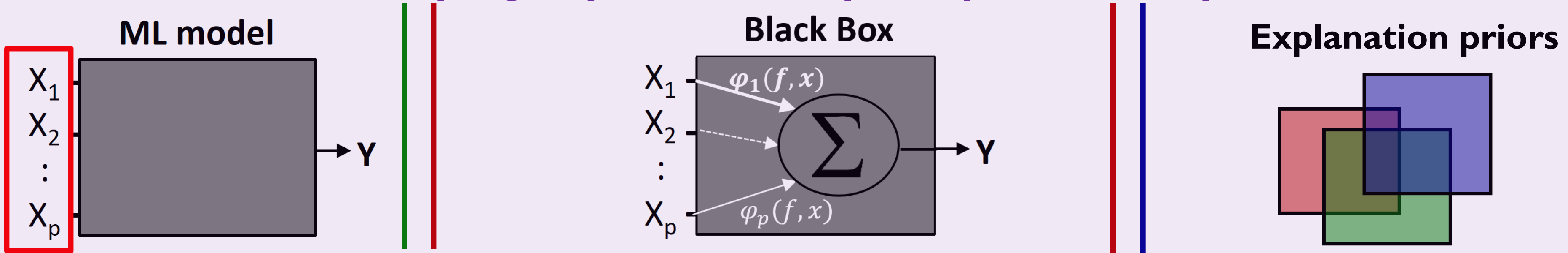
Explainable AI for biomedical sciences & beyond

Medicine & healthcare
anesthesia care, emergency medicine, critical care, nephrology, dermatology & biological age

Cancer biology & precision medicine

Alzheimer's disease therapeutic target discovery

Developing explainable AI principles techniques

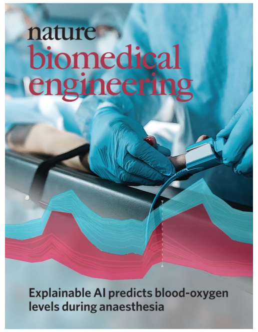


Learn interpretable features | **Make interpretable predictions** | **Learn explainable models**

Clinical medicine ← **Basic biology**

ICLR'24; ***NeurIPS***'23; ***NeurIPS***'23; ***Nature MI***'23; ***ICLR***'23; ***ICLR***'23; ***ICML***'23; ***AISTATS***, 2022; ***ICLR***, 2022; ***Nature MI***, 2021; ***JMLR***, 2021; ***Nature Comm***,. 2022; ***JMLR***, 2021; ***NeurIPS***, 2020; ***Nature MI*** (cover), 2020; ***NeurIPS***, 2020; ***AISTATS***, 2020; ***NeurIPS*** (oral), Dec 2017

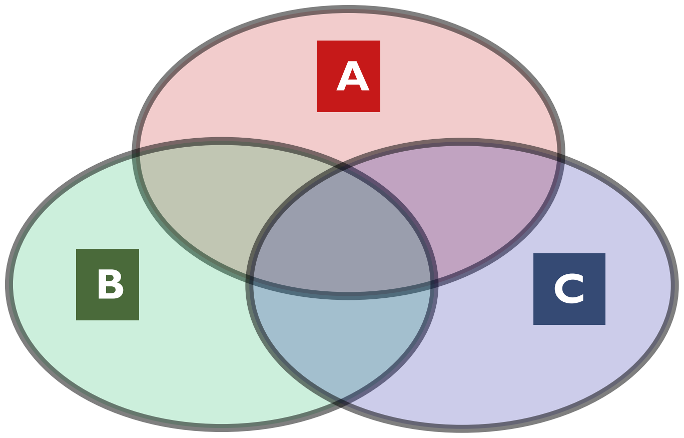
Nature Medicine, 2024; ***Lancet***, 2024; ***Nature Methods***, 2023; ***Genome Biology***, 2023; ***Nature BME 2023***; ***Lancet Healthy Longevity***, 2023 (cover); ***Nature BME** 2023*; ***Nature Comm. Medicine***, 2022; ***Nature BME***, 2022; ***Nature Comm.***, 2021; ***Nature MI***, 2021; ***Nature Comm***, 2018; ***Nature BME*** (cover), 2018



More about our research can be found at: https://aims.cs.washington.edu/publications

- A tip for navigating our publication site

Which field does the paper aim to advance?



AIMS Lab Home · Su-In Lee · **Publications** · Research · People · Education

A AI/ML
B AI in Biology
C AI in Clinical Medicine

J Journal publications
C Conference publications

* indicates equal contribution

Under Review

C J Dissection of medical AI reasoning processes via physician and generative-AI collaboration
Alex J. DeGrave, Zhuo Ran Cai, Joseph D. Janizek, Roxana Daneshjou*, and Su-In Lee*
In Press, *Nature Biomedical Engineering*
medRxiv

C J Fostering transparent medical image AI via an image-text foundation model grounded in medical literature
Chanwoo Kim, Soham U. Gadgil, Alex J. DeGrave, Zhuo Ran Cai, Roxana Daneshjou*, and Su-In Lee*
In Revision, *Nature Medicine*
medRxiv

A C Estimating Conditional Mutual Information for Dynamic Feature Selection
Soham U. Gadgil*, Ian Covert*, Su-In Lee
Under Review, ICLR'24
arXiv

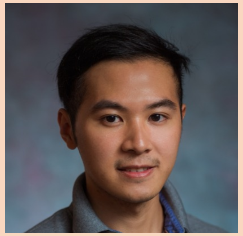
AI for bioMedical Sciences (AIMS) Lab

Nicasia Beebe-Wang (CSE PhD)

Ian Covert (CSE PhD)

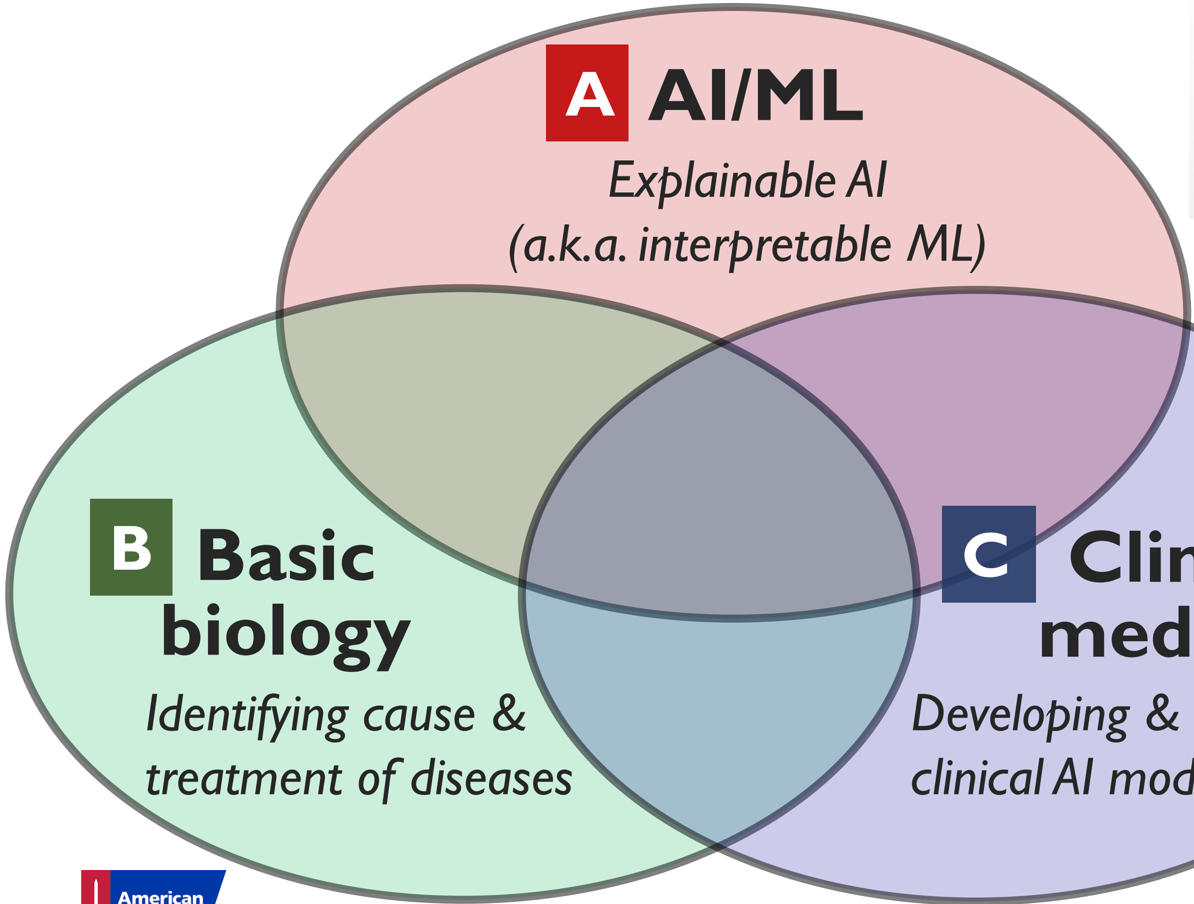
Wei Qiu (CSE PhD)

Chris Lin (CSE PhD)



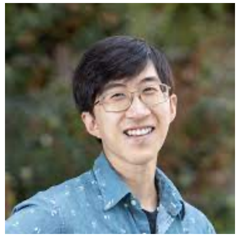
Mingyu Lu, MD (CSE PhD)

Patrick Yu (CSE PhD)



Su-In Lee (PI)

UW MSTP



Hugh Chen (CSE PhD)

Joe Janizek (MSTP, CSE PhD; matched to Stanford)

Ethan Weinberger (CSE PhD)

Alex DeGrave (MSTP, CSE PhD)

Chanwoo Kim (CSE PhD)

Soham Gadgil (CSE PhD)

Previous members: Ben Logsdon (postdoc), Safiye Celik (CSE PhD'18), Scott Lundberg (CSE PhD'19), Parmita Mehta (CSE PhD'20), Gabe Erion (MSTP, CSE PhD'21; now Harvard Medical School for residency),