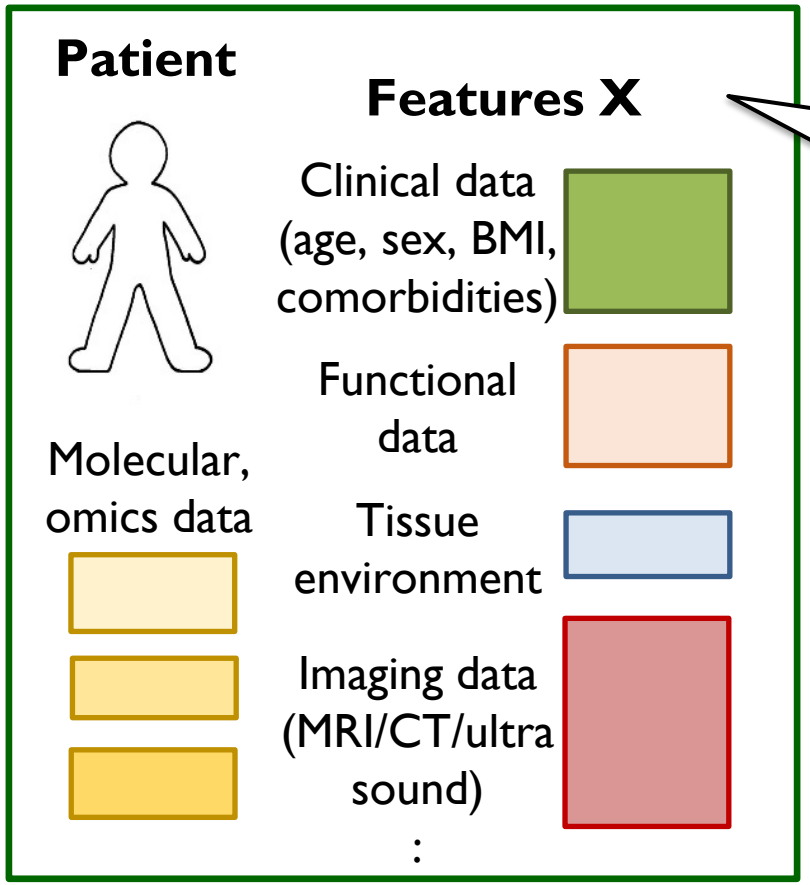
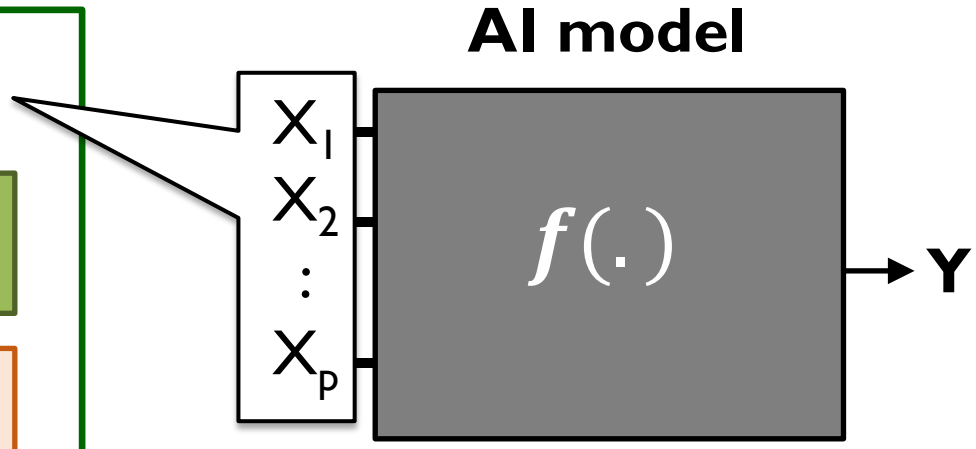
Explainable, Transparent, and Trustworthy AI for Biomedicine

Su-In Lee

Boeing Endowed Professor
Paul G. Allen School of Computer Science & Engineering
University of Washington, Seattle

A common AI problem in biomedicine:
Given features **X**, predict **Y.**





Outcomes Y

Regenerative therapy response

"favorable"

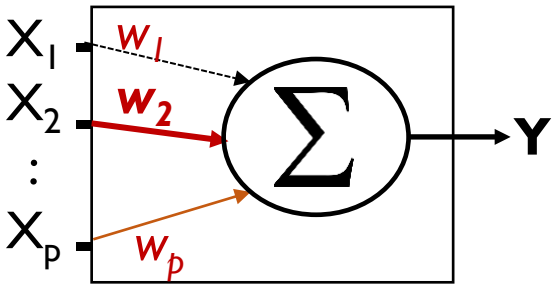
- **A more actionable question:** Which components of X should we intervene on to improve Y?
- **The missing piece:** Why was a certain prediction made?
- **The challenge:** AI models are often black boxes.
- **Our solution: Explaiable AI (XAI)**

Our solution was 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



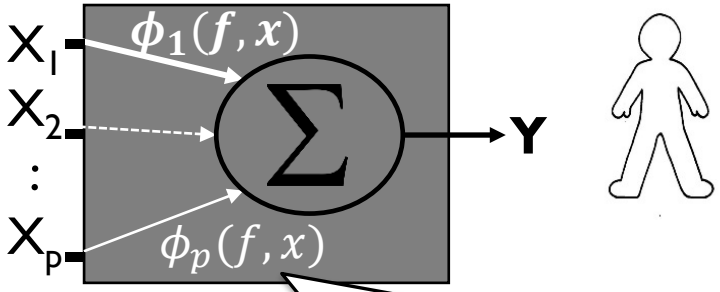
Complex model f (.)

Black Box

Our approach, SHAP

(SHapley Additive exPlanations)

For a particular prediction

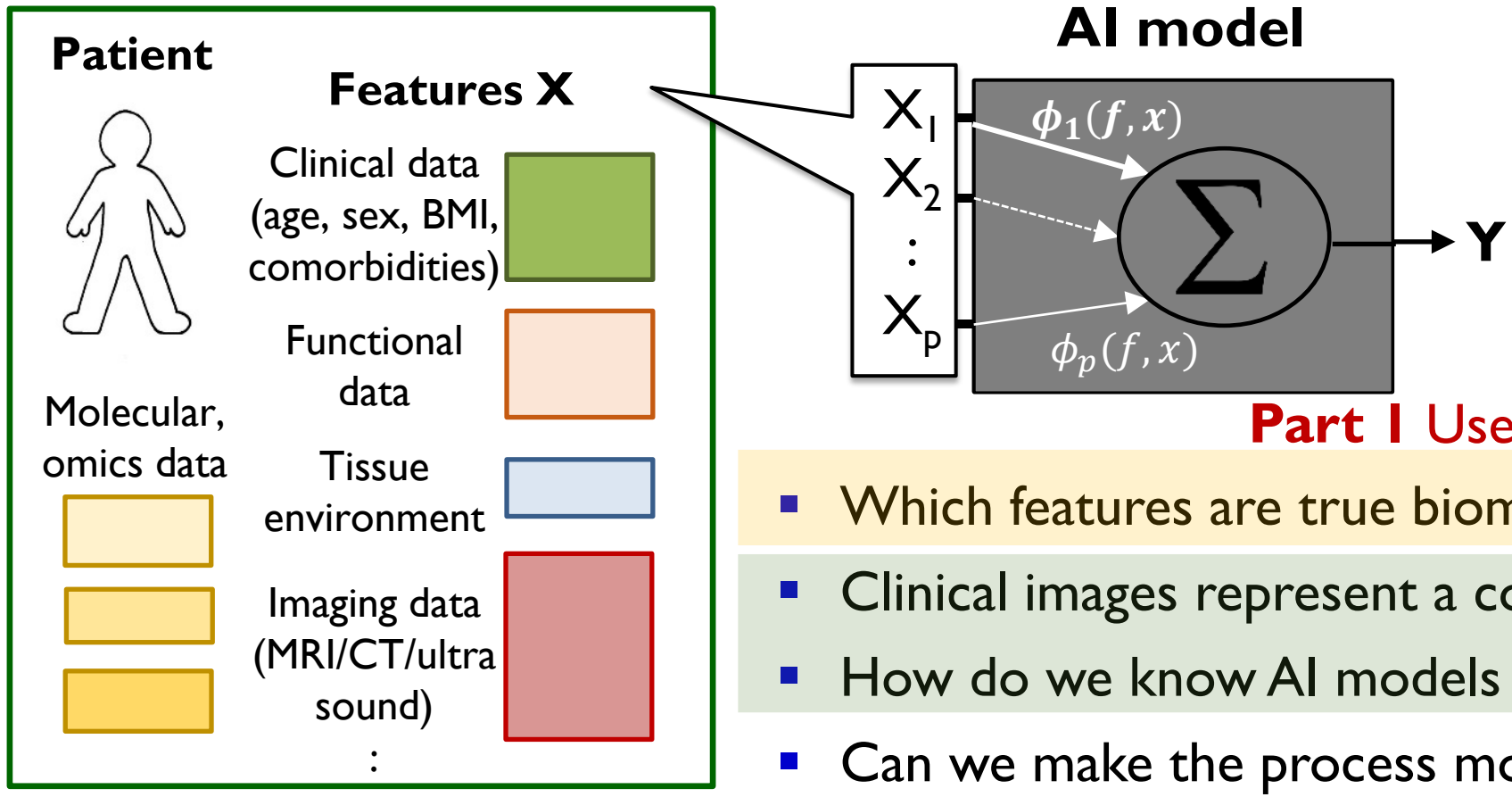


SHAP can estimate feature importance for a particular prediction for any model.

Scott, CSE PhD'19

Lundberg & Lee. ***Neural Information Processing Systems (NeurIPS)*** Oral (2017) – Cited 50,000+ times

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



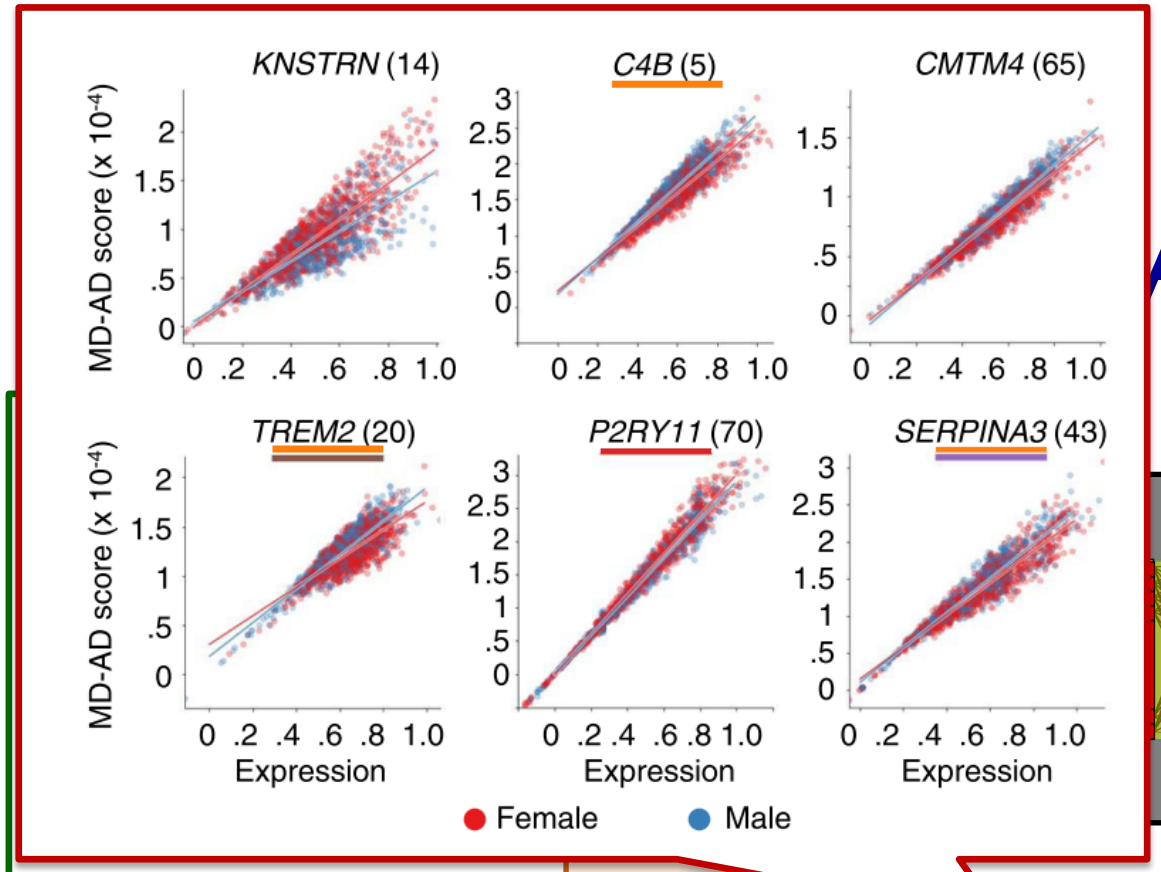
Part 1 Use case: Alzheimer's disease

- Which features are true biomarkers or disease drivers?
- Clinical images represent a complex modality. **Part 2**
- How do we know AI models are performing correctly?
- Can we make the process more cost-aware?

Lundberg et al. ***Nature Machine Intelligence***, *2020* – Featured on the Cover

mechanistic explanation

AD) phenotypes



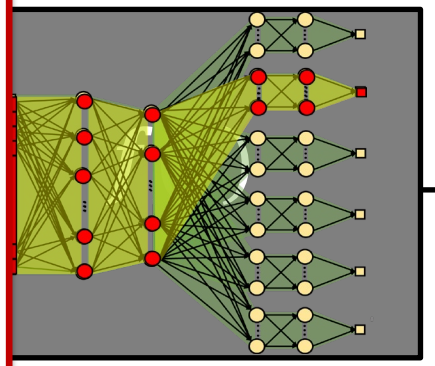
MSBB (*n*=879)

ROSMAP (*n*=542)

Demographic Data

Clinical data

AI model

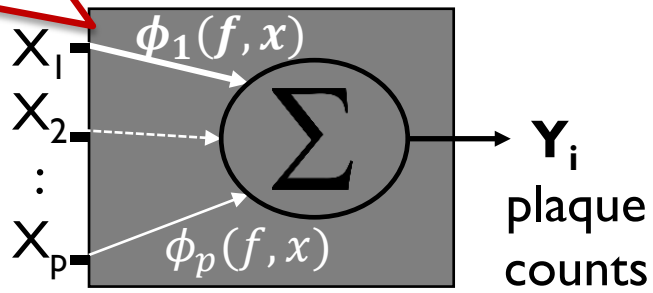


Y

Outcomes Y

Neuropathological phenotypes:

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



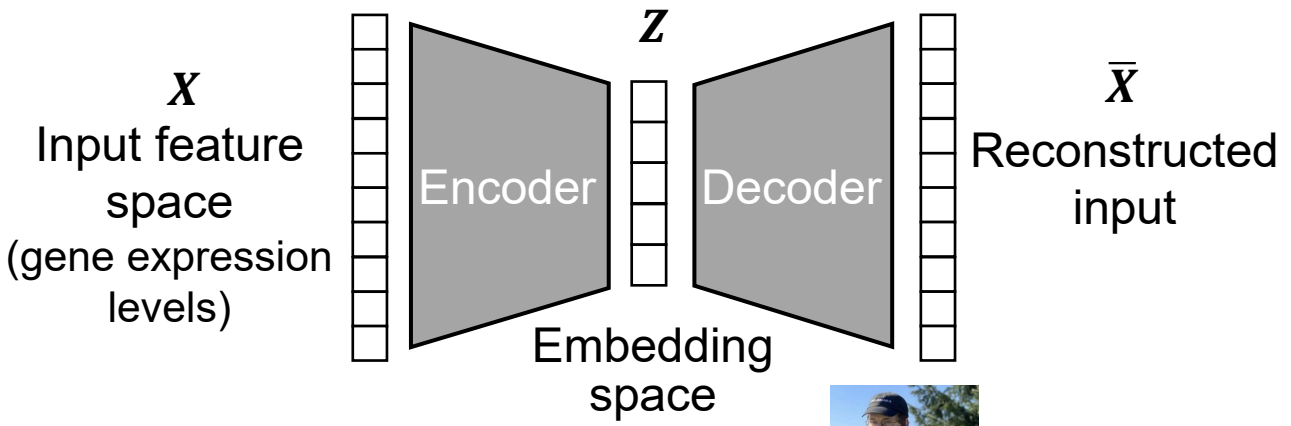
- Estimate each gene's contribution to AD neuropathologies
  - Previously unknown sex-differential immune response (***microglia activity***) in AD

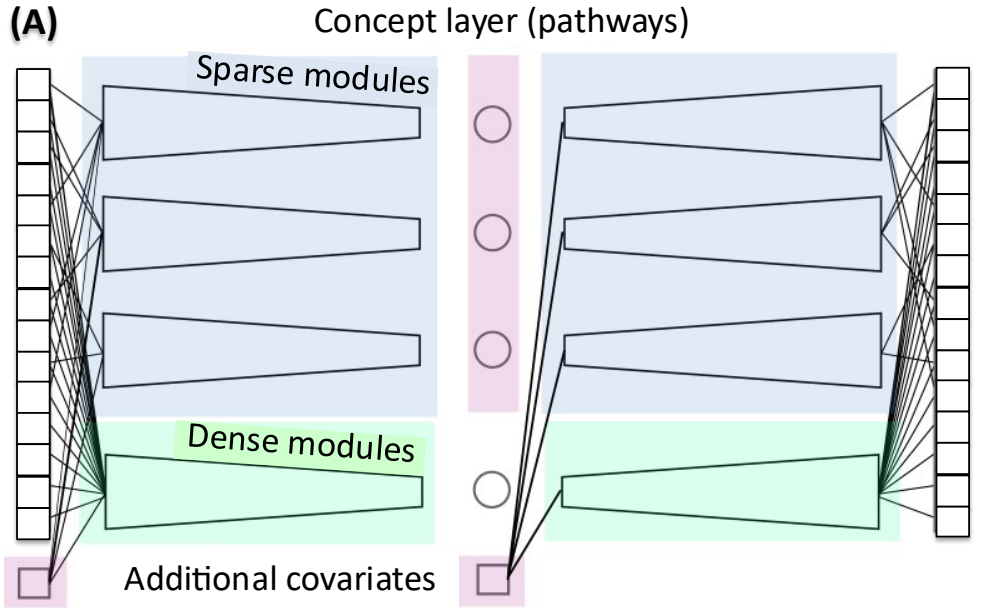
Beebe-Wang et al. ***Nature Communications***, 2021

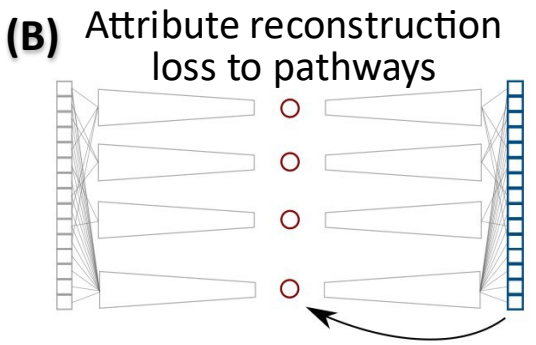
CSE PhD'22

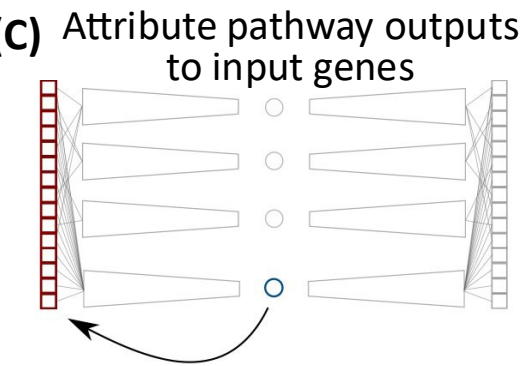
Biologically interpretable AI modeling further advances data-driven discovery

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

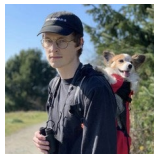








Janizek et al. ***Genome Biology***, 2023



Joe, UW MSTP/CSE PhD'22, Residency at Stanford Radiology

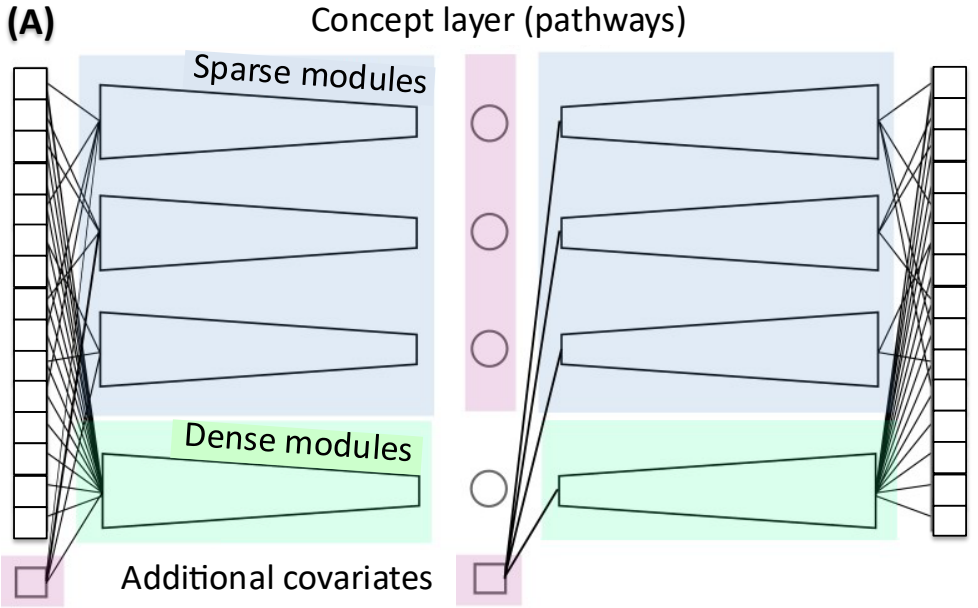
Collaboration with UW Laboratory Medicine & Pathology (Matt Kaeberlein)

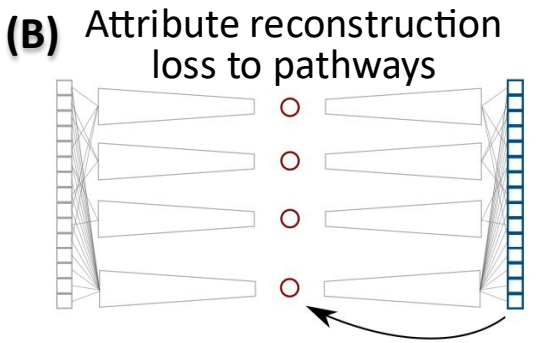
Biologically interpretable AI modeling identifies experimentally validated AD therapeutic targets

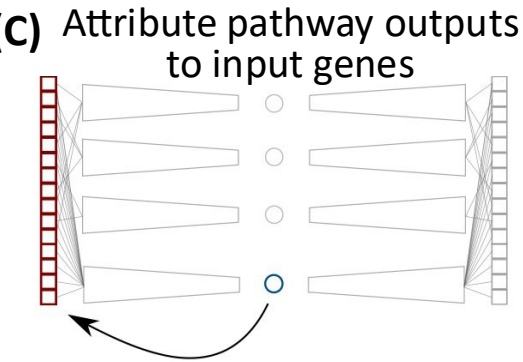
- We applied our approach to extended bulk RNAseq datasets from AD study cohorts
- We identified mitochondrial complex I as a potential mediator for tolerance to Aβ toxicity
  - *In vivo* validation in a transgenic *C. elegans* model expressing Aβ done by Matt Kaeberlein's lab

A promising pharmacological avenue!

Capsaicin



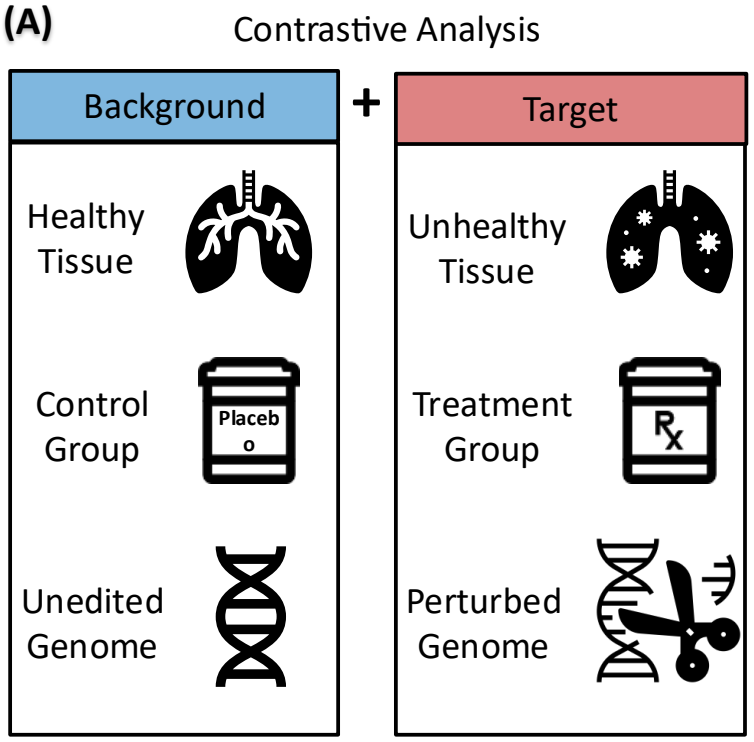


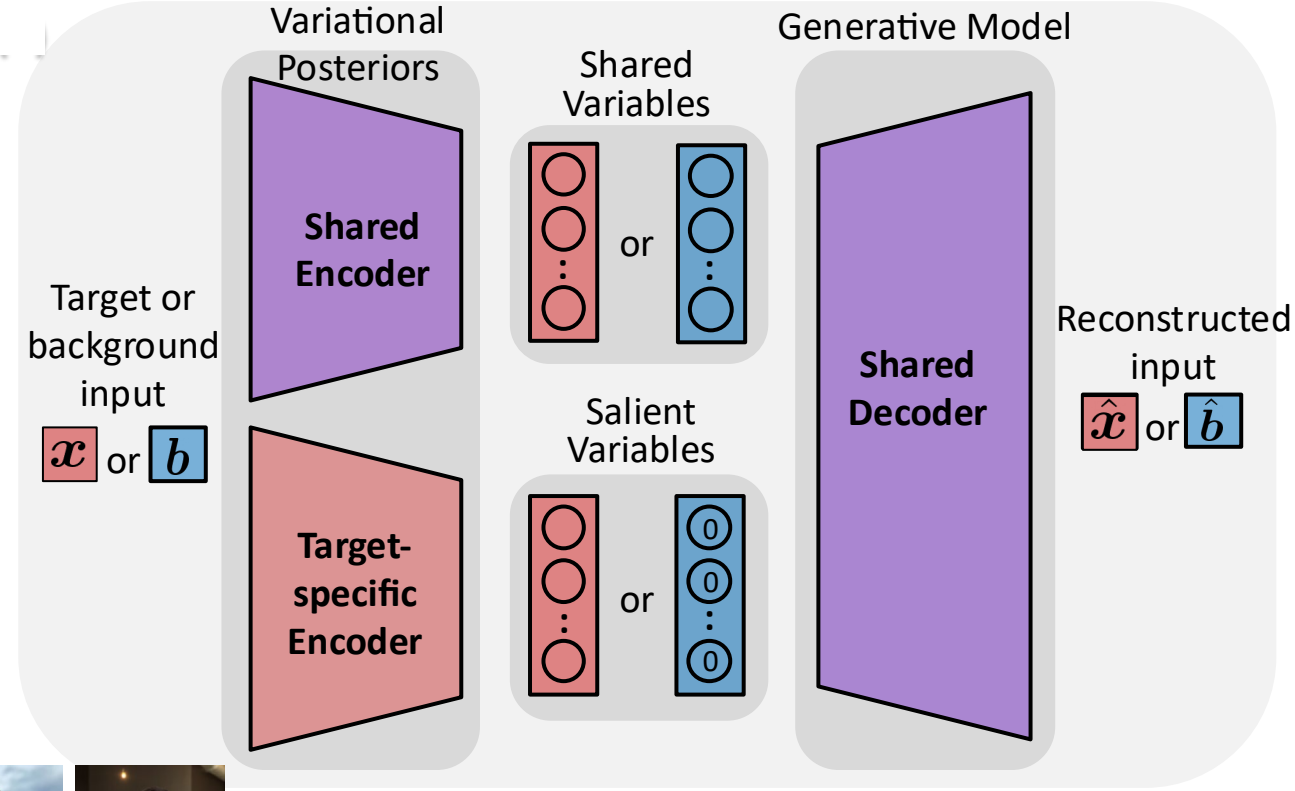


Janizek et al. ***Genome Biology***, 2023

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

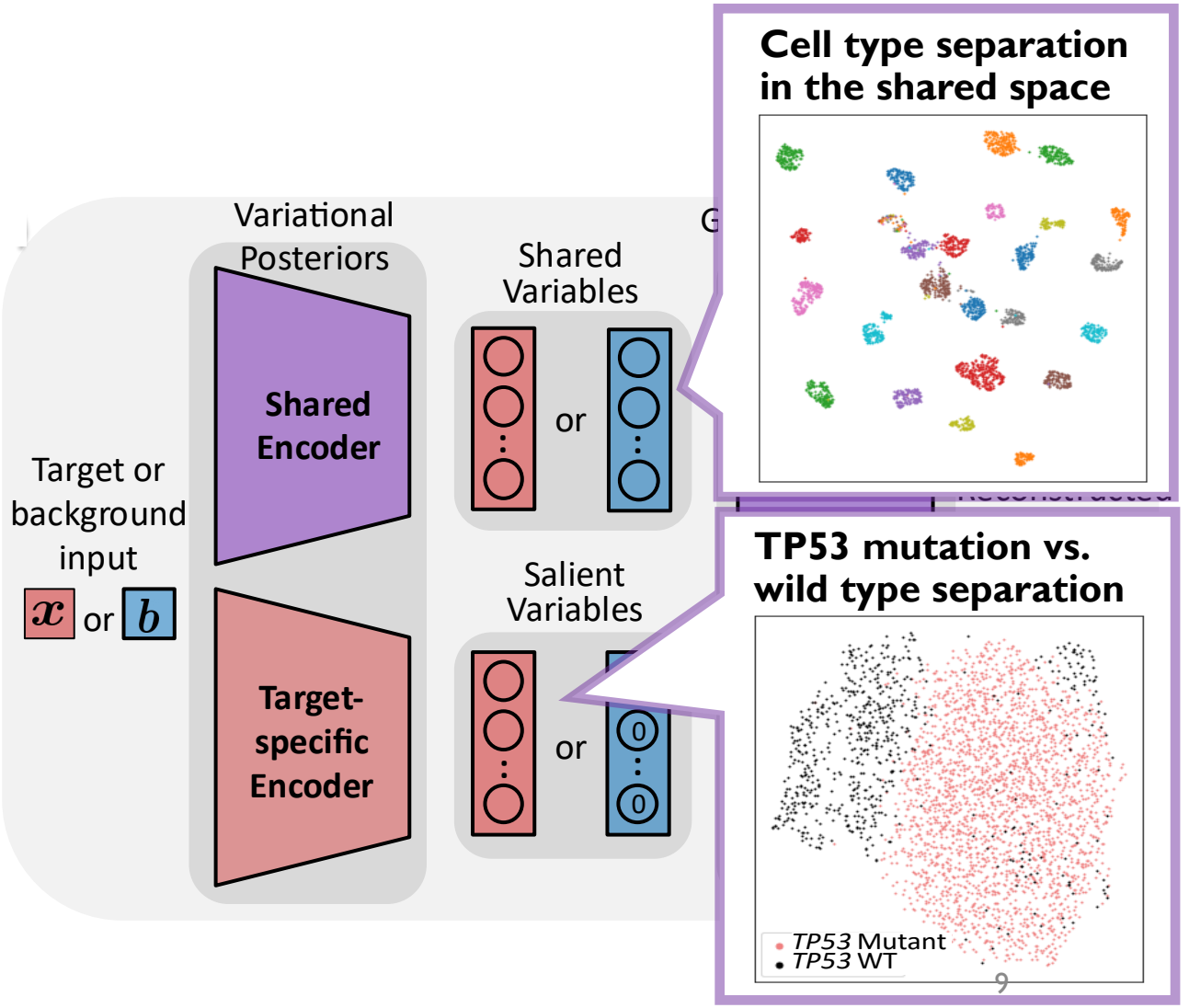
Ethan & Chris, CSE PhD

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

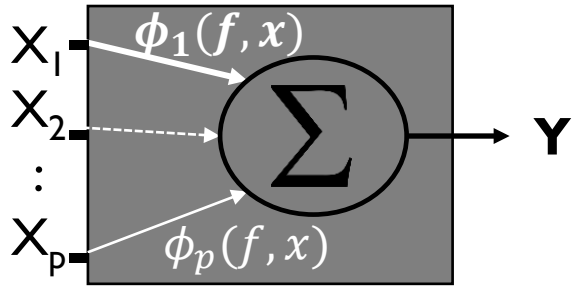
Important implications for personalized medicine!

- 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 1 – Identifying disease-driving genes and processes
  - Unveiling neurodegenerative disease insights with explainable AI
    [*Nature Comm'21 ; Genome Biology'23; Nature Methods'23*]

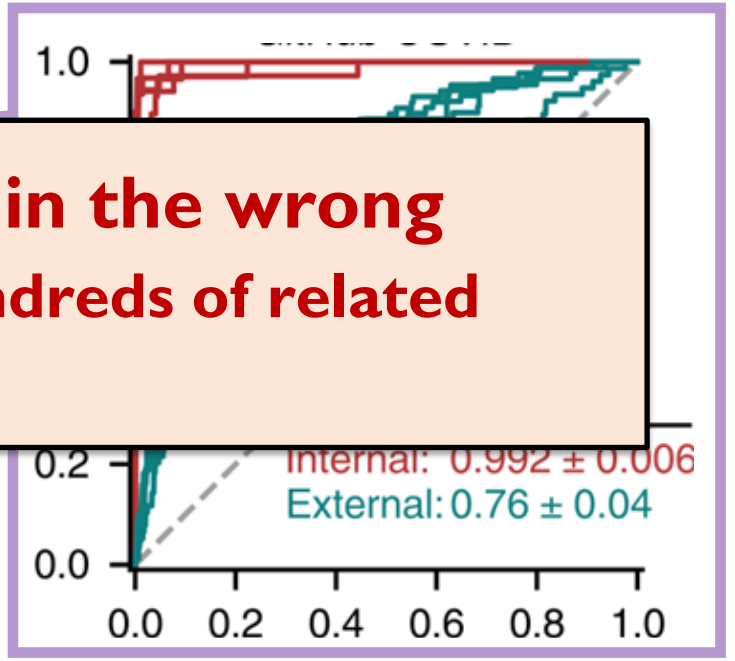
- Part 2 – Auditing AI models
  - Feature attributions [*Nature MI*'21 featured in *Nature*'22]
  - Counterfactual explanations (via generative AI) [*Nature BME*'25 cover; *Lancet*'24]
  - Concept-based explanations (via foundation models) [*Nature Medicine*'24]

Our explainable AI techniques are generalizable to a wide range of biomedical problems, including regenerative medicine.

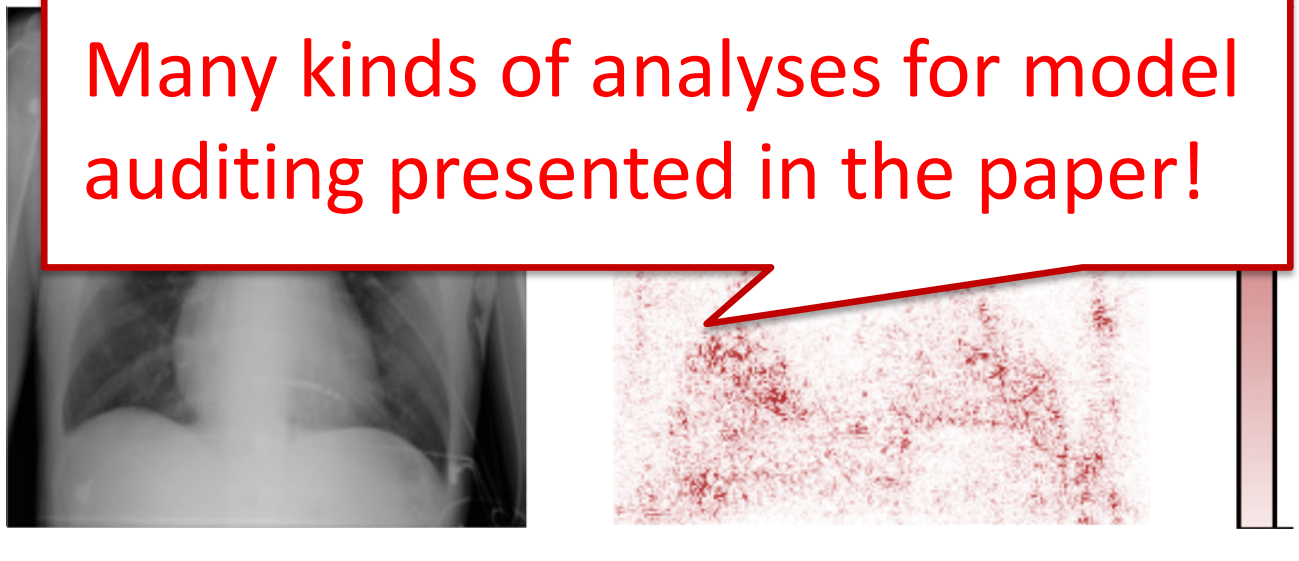
Auditing AI for COVID-19 detection using XAI

- 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

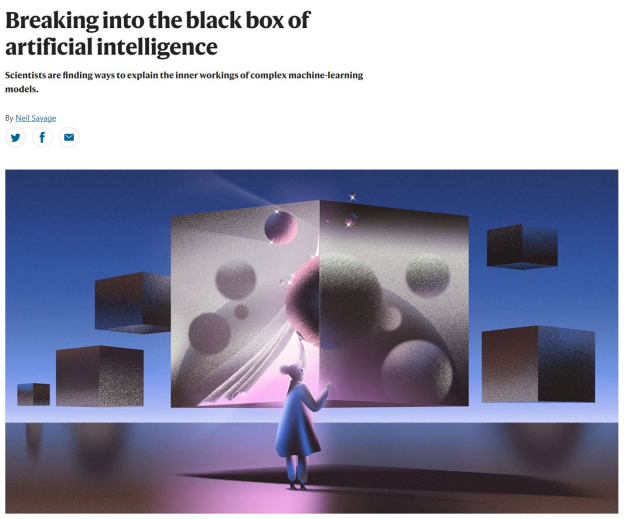
✗ medical devices should not predict negative status

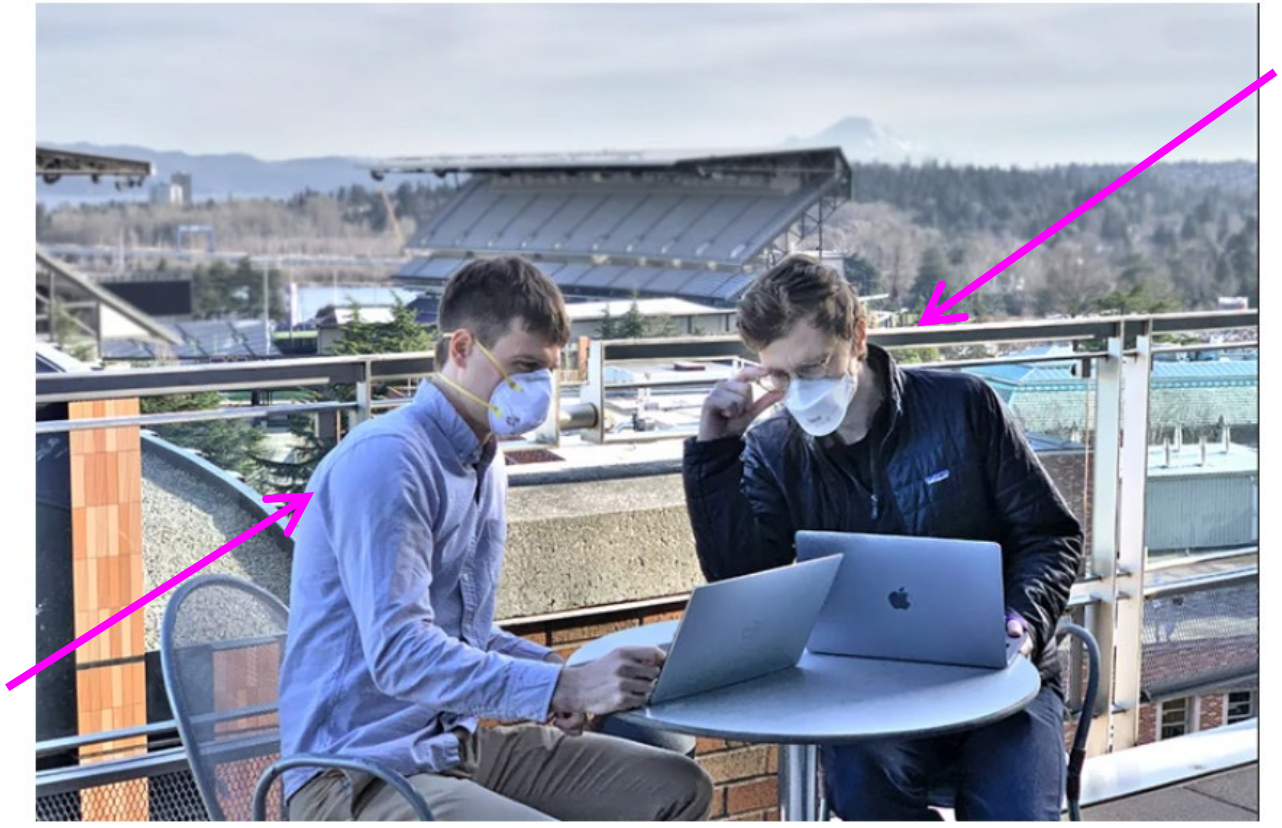
MSTP / CSE PhD

DeGrave,* Janizek* et al. ***Nature Machine Intelligence***, 2021; featured in ***Nature***, 2022

Our AI auditing work featured in *Nature*

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



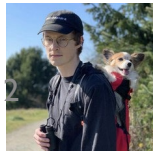


UW MSTP / CSE PhD **Joe Janizek** (Just matched to Stanford)

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

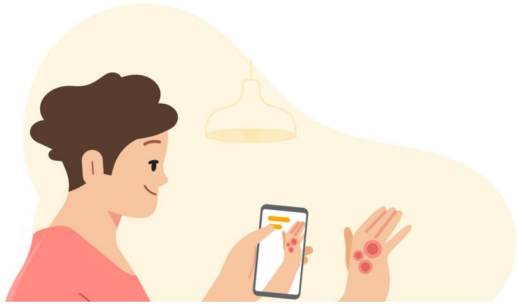
Alex DeGrave and Joseph Janizek are students on the Medical Scientist Training Program at the University of Washington, in Seattle. Credit: Alex DeGrave

DeGrave,* Janizek* et al. ***Nature Machine Intelligence***, 2021



Collaboration with Stanford DBDS / Dermatology (Roxana Daneshjou)

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

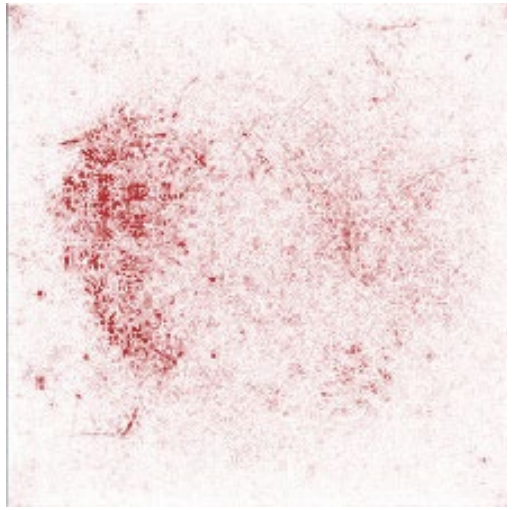


- Auditing AI models to predict skin cancer
  - Five models – 2 academic models, 2 commercial devices, and 1 competition winner
- Technical challenges – saliency maps often do not work

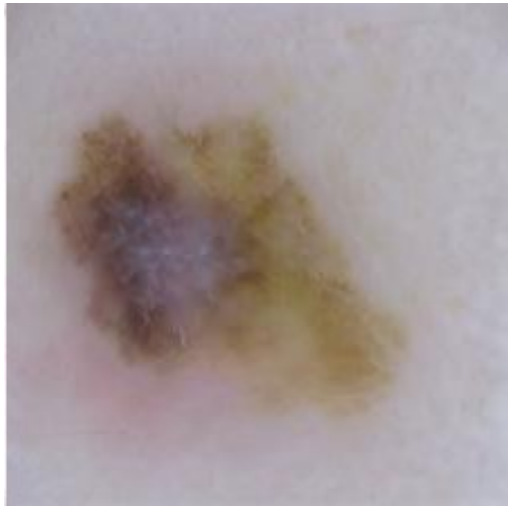
Original image

Predicted: benign

Saliency map



Modified image



Predicted: malignant

Our solution:

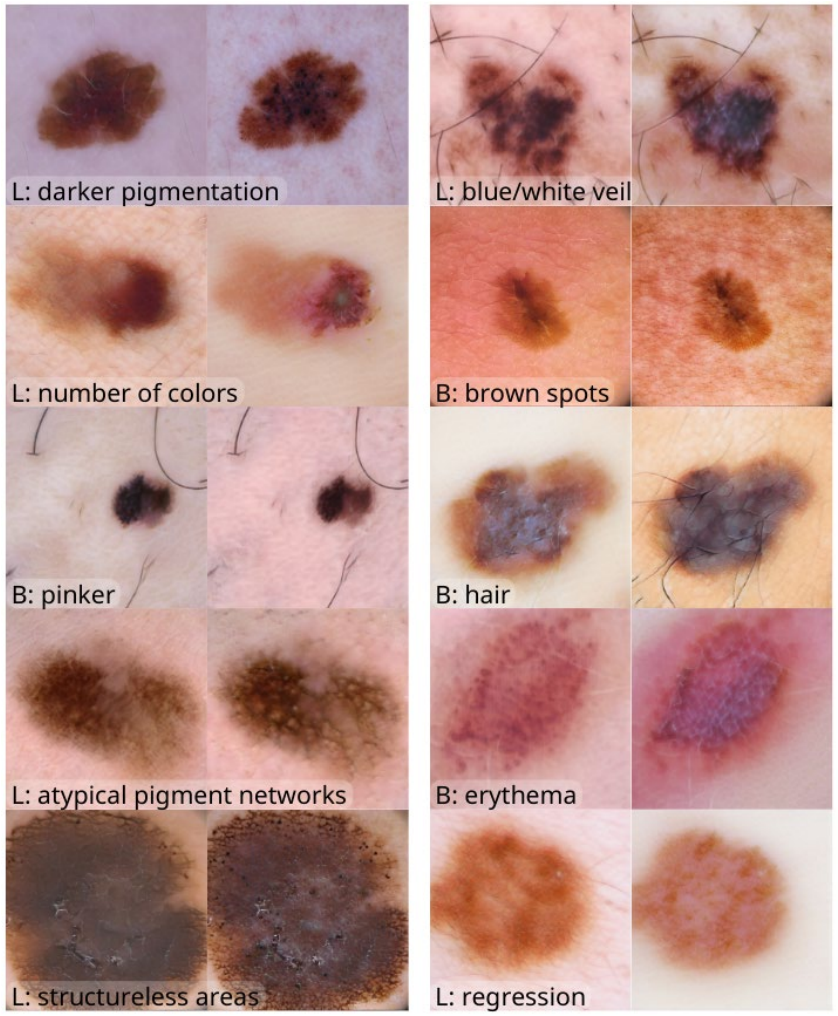
- Generate counterfactual images ***from the AI model***
- Systematic characterization by experts: Drs. Roxana Daneshjou, and Zhuo Ran Cai (Stanford)

DeGrave et al. (*Nature Biomedical Engineering*)

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

Collaboration with Stanford DBDS / Dermatology (Roxana Daneshjou)

How do dermatology AI systems make decisions on dermoscopic in



Attribute predominant in:
- Malignant counterfactuals
- Benign counterfactuals
- Neither/Varies by image

Fraction of counterfactual pairs:
- 100%
- 50%
- 25%
- 10%

Attribute category:
L: Lesion
B: Background

Alex, MSTP / CSE PhD

Degrave, Ran Cai, Janizek, Daneshjou,* and Lee* ***Nature Biomedical Engineering***, 2025

The *Lancet* perspective

"The clinical potential of counterfactual AI"

by Su-In Lee and Eric Topol

Nature Reviews Bioengineering, 2025

Medical AI transparency in the entire life cycle

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.

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

Further reading

DeGrave AJ, Cai ZR, Janizek Daneshjou R, Lee SI. Audit inference processes of me image classifiers by levera generative AI and the exp of physicians. *Nat Biomed* 2023; published online De https://doi.org/10.1038/s41551-023-01160-9

nature reviews bioengineering

https://doi.org/10.1038/s44222-025-00363-w

Review article

Check for updates

Transparency of medical artificial intelligence systems

Chanwoo Kim[1,2], Soham U. Gadgil[1,2] & Su-In Lee[1]

Abstract

Medical artificial intelligence (AI) systems hold promise for transforming healthcare by supporting clinical decision-making in diagnostics and treatment. The effective deployment of medical AI requires trust among key stakeholders – including patients, providers, developers and regulators – which can be built by ensuring transparency in medical AI, including in its design, operation and outcomes. However, many AI

Sections

Introduction

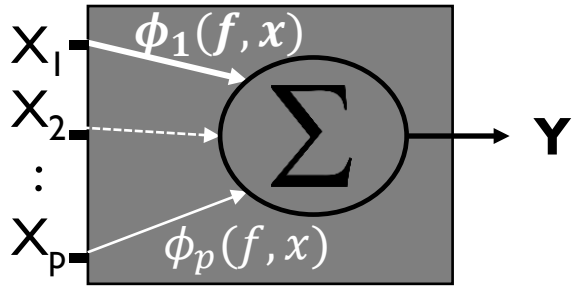
Data transparency

Model transparency

Deployment transparency

Outlook

Outline – Two parts



- Part 1 – Identifying disease-driving genes and processes
  - Unveiling neurodegenerative disease insights with explainable AI
    [*Nature Comm'21; Genome Biology'23; Nature Methods'23*]

- Part 2 – Auditing AI models
  - Feature attributions [*Nature MI*'21 featured in *Nature*'22]
  - Counterfactual explanations (via generative AI) [*Nature BME*'25 cover; *Lancet*'24]
  - Concept-based explanations (via foundation models) [*Nature Medicine*'24]

Our explainable AI techniques are generalizable to a wide range of biomedical problems, including regenerative medicine.

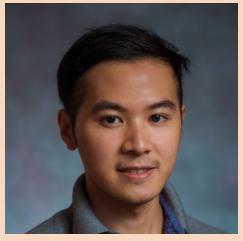
AI for bioMedical Sciences (AIMS) Lab

Nicasia Beebe-Wang (CSE PhD)

Ian Covert (former CSE PhD)

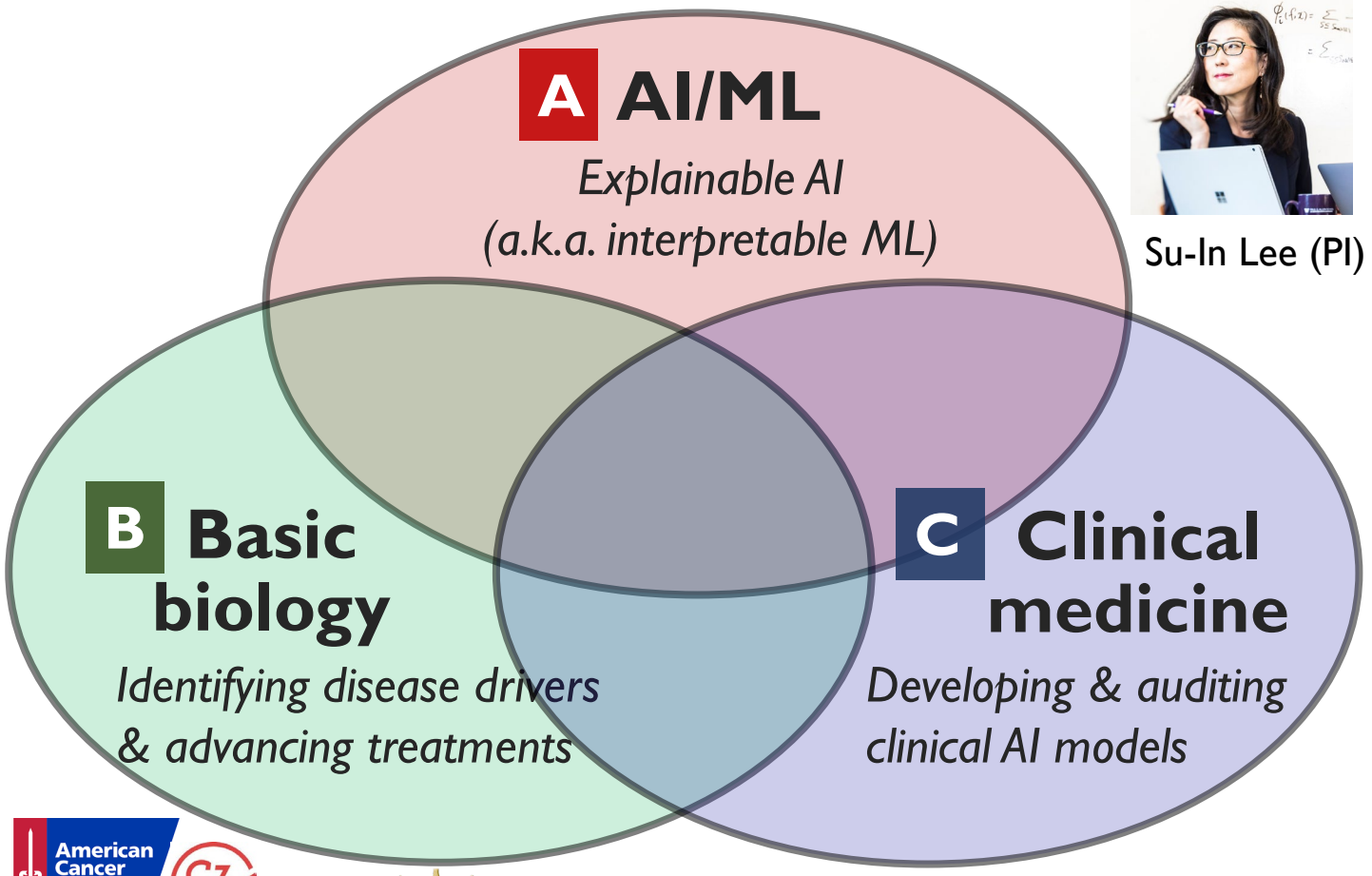
Wei Qiu (CSE PhD)

Chris Lin (CSE PhD)



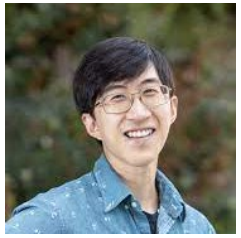
Mingyu Lu, MD (CSE PhD)

Patrick Yu (CSE PhD)



Su-In Lee (PI)

UW MSTP



Hugh Chen (former 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)

American Cancer Society

NIH

NSF

Genentech
A Member of the Roche Group

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),