Combining Clinical Phenotypes and DNA Foundation Models to Predict Human Disease

Lucila Ohno-Machado

Waldemar von Zedtwitz Professor of Medicine and Chair,

Department of Biomedical Informatics and Data Science

Deputy Dean for Biomedical Informatics

Yale SCHOOL OF MEDICINE

MIRACUM-DIFUTURE Symposium, Mannheim
6/30/26

Many thanks to the symposium organizers and attendees!

Acknowledging colleagues from the Yale Department of Biomedical Informatics & Data Science (Hua Xu, Daniella Meeker, Rui Zhu, …), Evo2 developers (Brian Hie et al), data sources, collaborators in various institutions

Topics

1. Generative AI
2. Clinical + Bio-informatics = Medical AI
3. Reflections

AI in Medicine is not new

80s – 90s
Expert Systems
(MYCIN, Internist)

→

90s – 00s
Neural Networks

→

00s – 10s
Other Machine
Learning Methods

↓

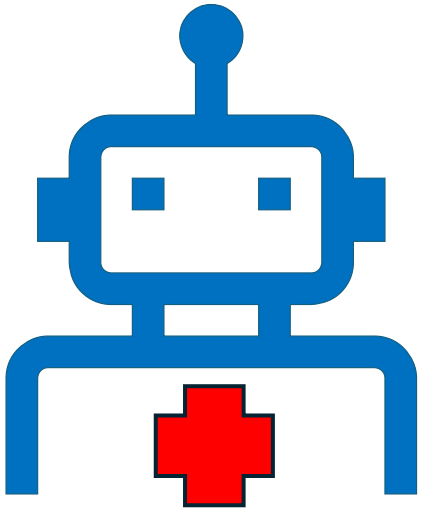
10s – 20s
Neural Networks
(deep learning)

→

20s – 30s
Generative AI
(LLMs)

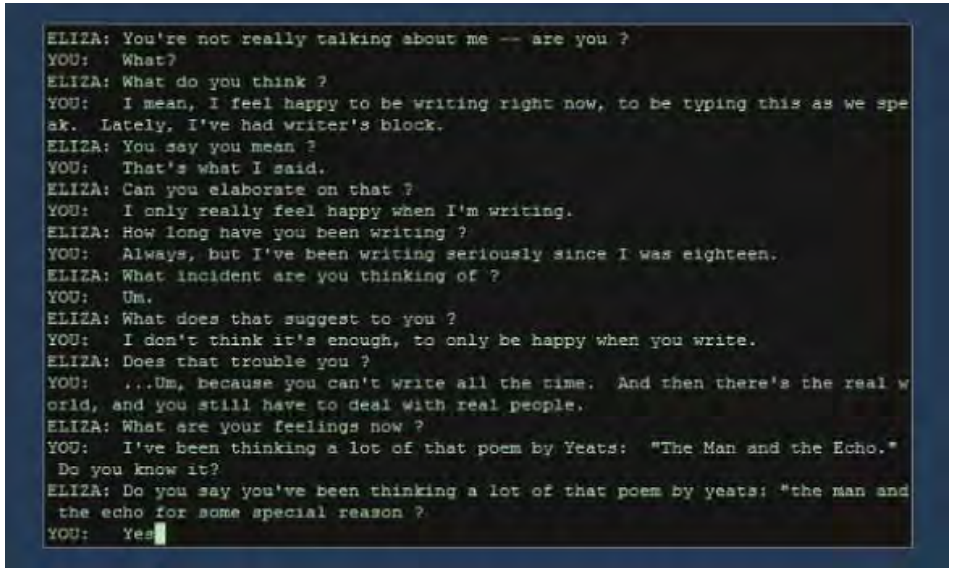
→

30s – 40s
…



Eliza, first chatbot (1966)

Weizenbaum, Joseph "ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine". Communications of the ACM; Volume 9 , Issue 1 (January 1966): p 36-45.



https://www.masswerk.at/elizabot/

ELIZA: Would it matter to you ?
YOU: Yes, a lot!
ELIZA: You are sure.
YOU: Yes, I am sure.
ELIZA: I see.
YOU: Are you complicated?
ELIZA: Perhaps I am complicated in your fantasies.
YOU: Indeed
ELIZA: That is interesting. Please continue.
YOU: On what?
ELIZA: What comes to mind when you ask that ?
YOU: That you are a bot.
ELIZA: What makes you think I am a bot ?
YOU: You sound like one.
ELIZA: You're not really talking about me -- are you ?
YOU: Yes.
ELIZA: I understand.
YOU: Where were you born?
ELIZA: What suggests that I was born ?

Talk Reset

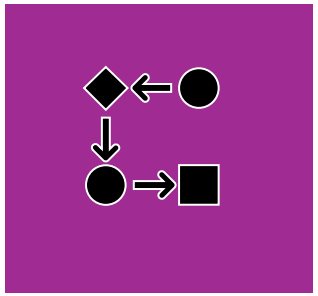
https://thoughtcatalog.com/oliver-miller/2012/08/a-conversation-with-eliza/

Expert Systems

- CASNET (Glaucoma)
- MYCIN (Infectious Diseases)
- INTERNIST-1 (Internal Medicine)

- Rule-based
- Knowledge-based (not data-driven)

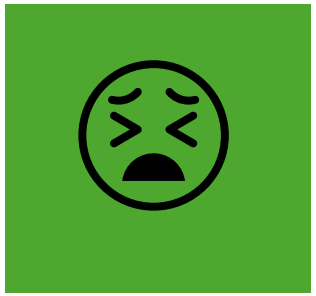
What do healthcare organizations have been using AI for?



Optimize services

Increase patient and provider satisfaction

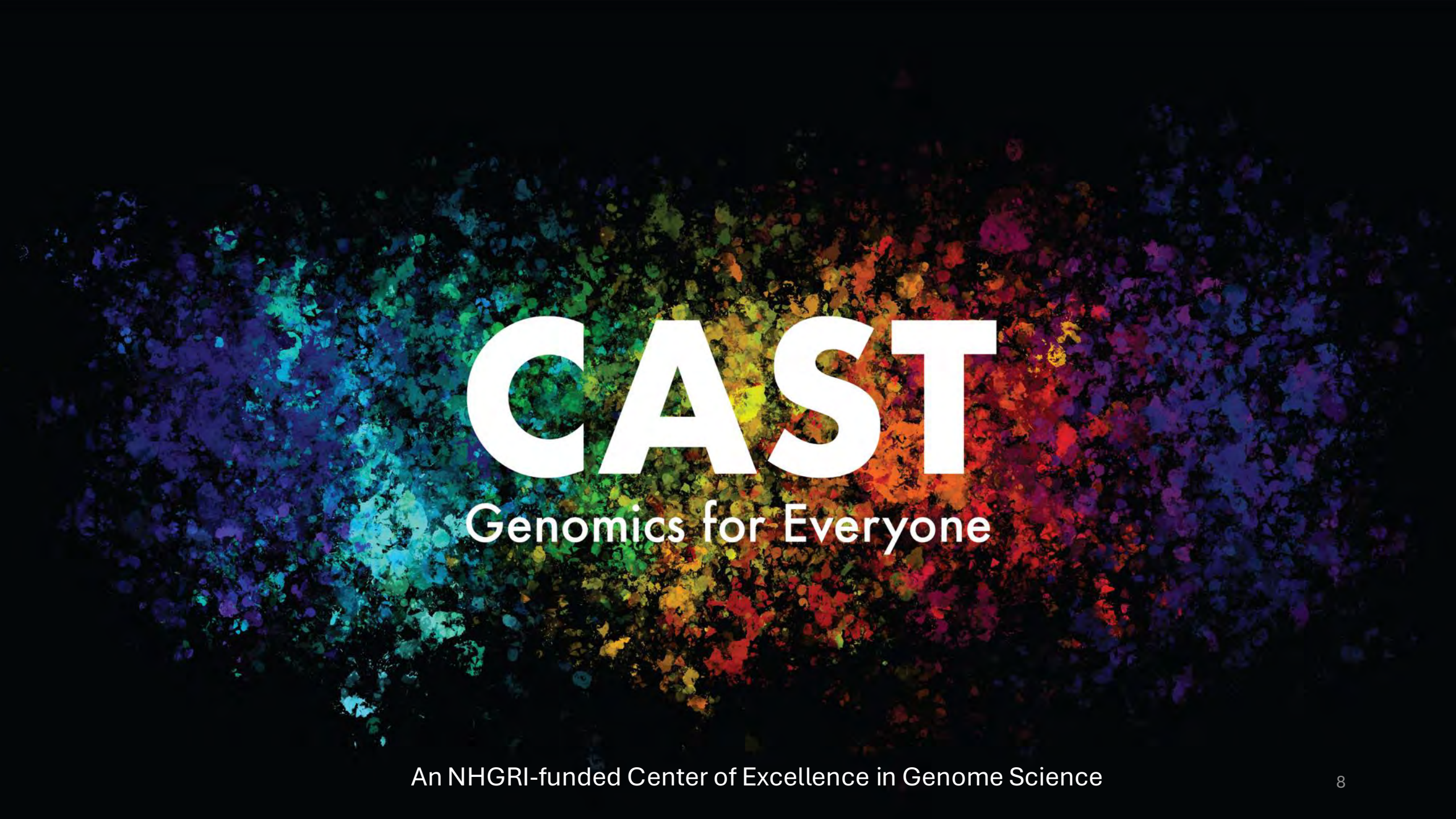
Increase efficiency



Decrease errors

Reduce diagnostic, therapy selection errors

Promote safety



An NHGRI-funded Center of Excellence in Genome Science

Disease risk is influenced by many factors, the more parameters, the more data are needed

Trait
e.g. LDL, cancer diagnosis

Genetic determinants

Variant effects
SNPs, HLA, TRs,…

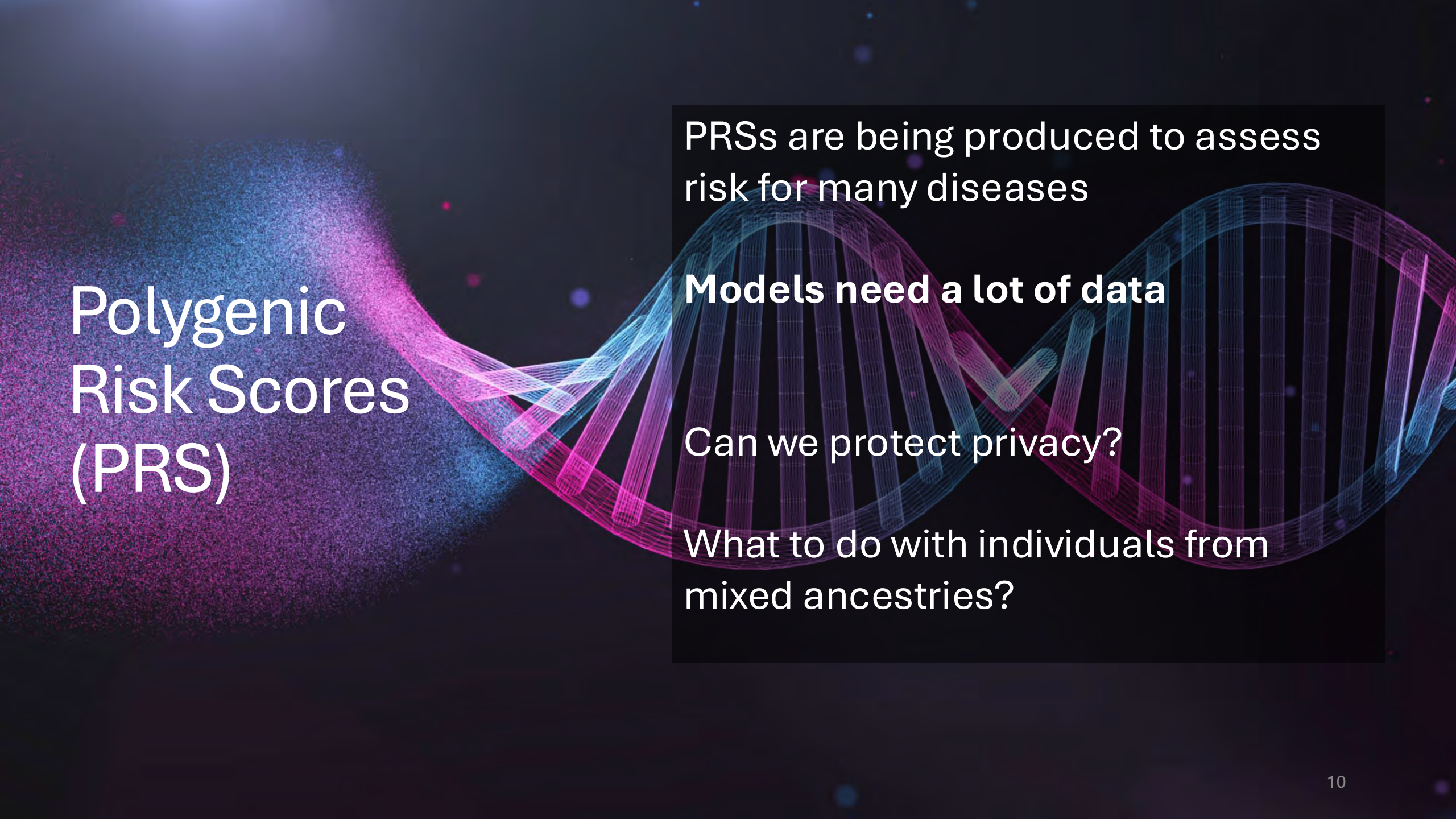
Ancestry
SNPs, PCs,…

Social determinants
income, education, diet,…

$$E[y_i] \sim X_i\beta + Q_i\gamma + C_i\delta$$

→ **Algorithms and tools to analyze data that can stay in their enclaves**

Polygenic Risk Scores (PRS)



PRSs are being produced to assess risk for many diseases

Models need a lot of data

Can we protect privacy?

What to do with individuals from mixed ancestries?

Opportunities: Organizing Academic Consortia

Precision Medicine
Federated Learning


covid19questions.org
Covid-19 Clinical Data Consult
Home
Questions & Answers
Answers in Progress
Ask Questions
About Us
COVID-19 Data Discovery from Clinical Records
Press here for an important note

Covid-19 Clinical Data Consult

In 2020

- 928,255 patients tested
- 59,074 diagnosed with COVID-19
- 19,022 hospitalized
- 2,591 deceased

- Data from >45M patients for comparisons

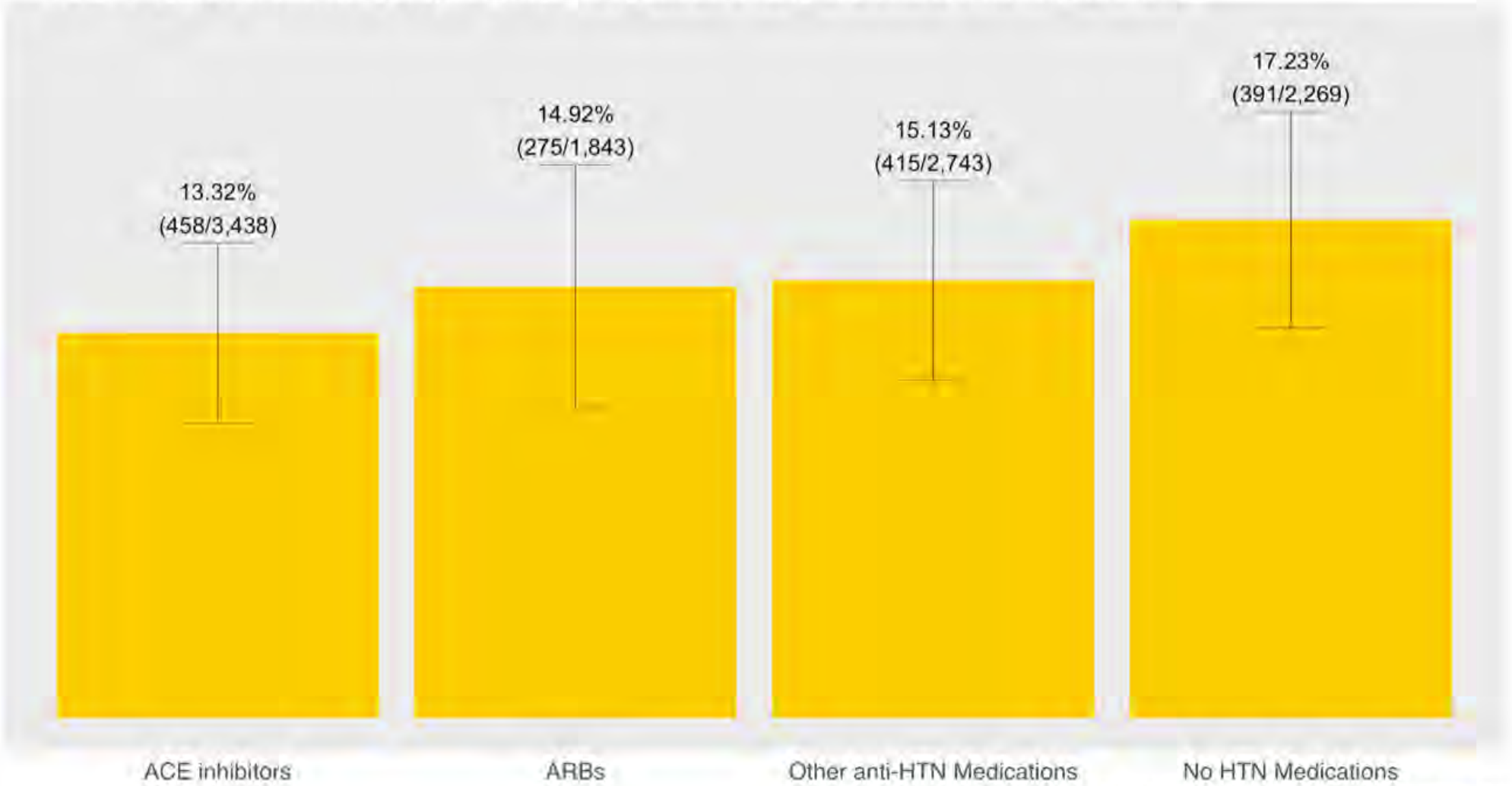
Moore Foundation award

In-hospital mortality of hypertensive patients per medication use group*

* Patients with medication use within one year prior to hospital encounter

Numbers in parentheses = in-hospital deaths / total patients per group

ACE = Angiotensin-converting enzyme, ARBs = Angiotensin II receptor blockers, HTN = Hypertension

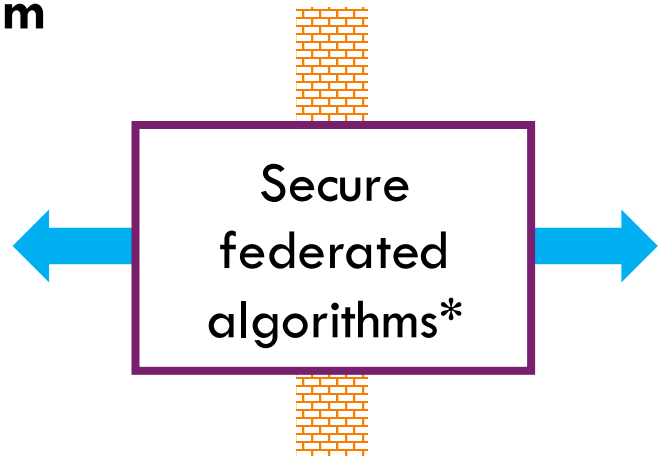


Source: 10,293 adult patients from 10 institutions
Data retrieved October 15, 2020 - November 03, 2020

Distributed Analysis for Precision Medicine

All of Us (AOU) Research Program

Million Veteran Program (MVP)



>600k participants
Health and genomic data (>400k WGS)

>950k participants
Health and genomic data

Access limited to AoU analysis platform

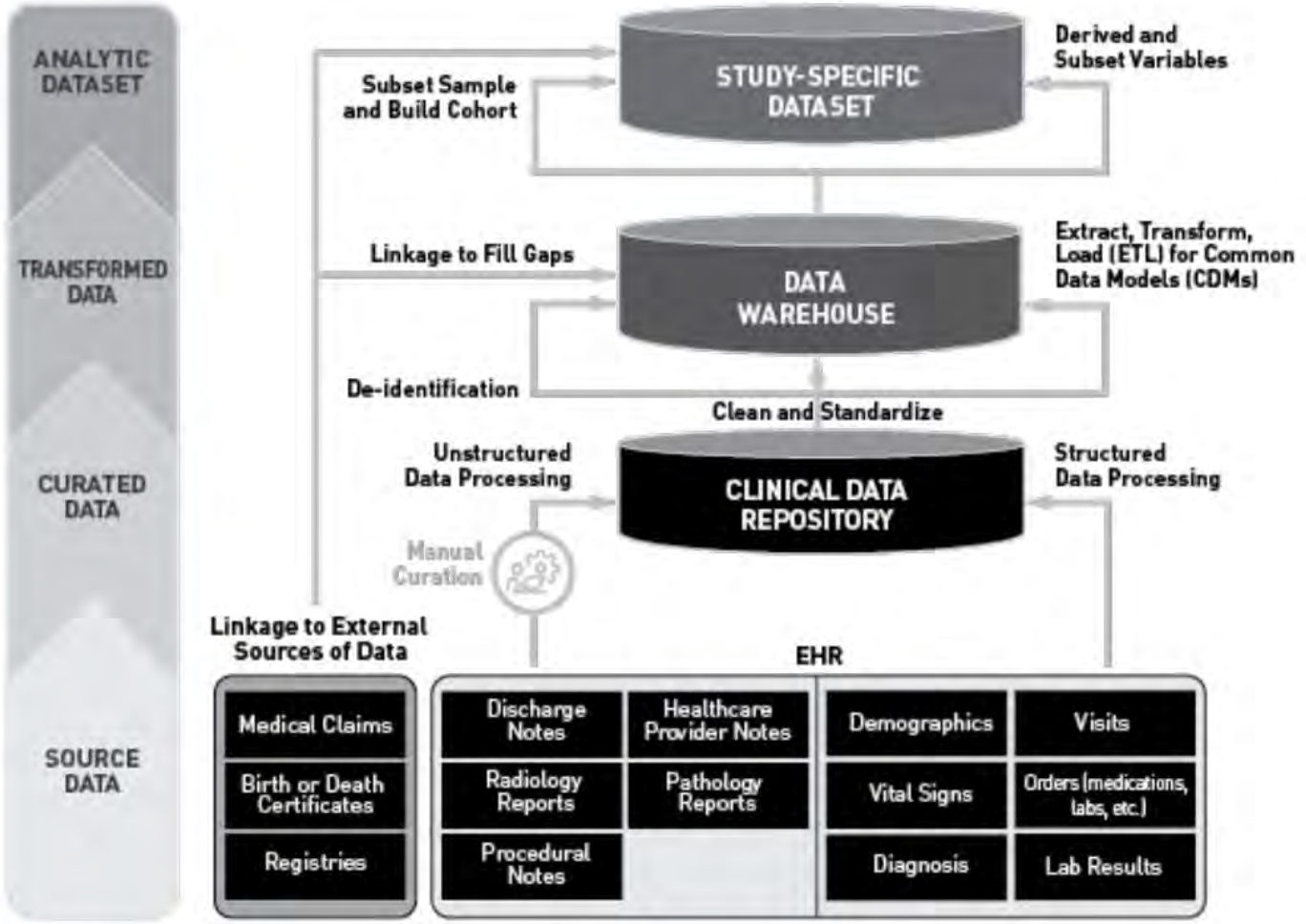
Access limited to VA analysis platform

training LLMs in these enclaves is not currently possible

Electronic Health Records (EHRs) for Clinical Studies

- EHRs (and linked data) becomes an enabling resource for Real World Evidence generation

FDA RWD Guidance 2021 *



Slide from Dr. Xu

* Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products – Draft Guidance by FDA, September 2021

Deep learning's imagination problem

Deep learning creates bizarre albeit fascinating images (Source: deepdreamgenerator.com)

Generative Adversarial Networks

Objective

Make fake data indistinguishable from real data

1. Generator

   Produces fake data

2. Discriminator Network

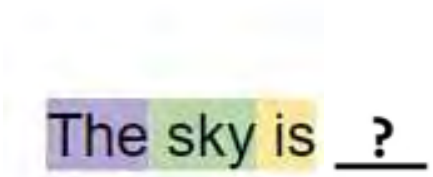
   Classifies fake vs. real

Synthetic images produced by StyleGAN, a GAN created by Nvidia researchers.
Image Credit: Nvidia

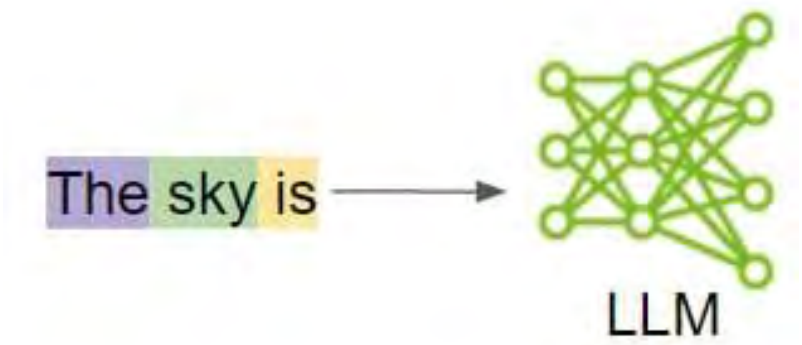
Examples of Generative AI

- Enhancing a Large Language Model (LLMs) for Medical Contents
- Using an existing DNA LLM to study disease

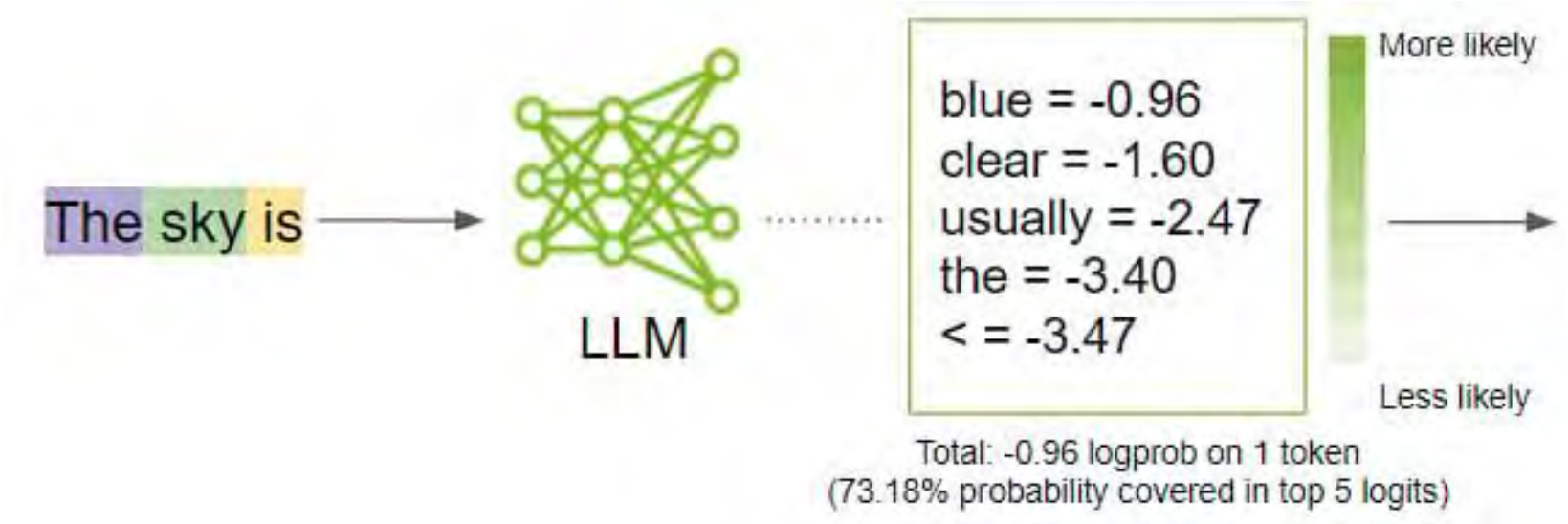
Large language model (LLM)



Large language model (LLM)



Large language model (LLM)



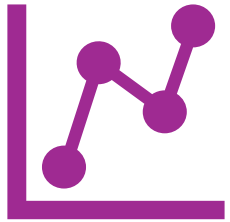
blue

How are healthcare organizations using generative AI?

Building access to Gen AI with contracts that do not allow training on prompts

Summaries (of visits, history)

Text generation (provider inbox management)



Clinical Research: Training existing foundational models with medical data

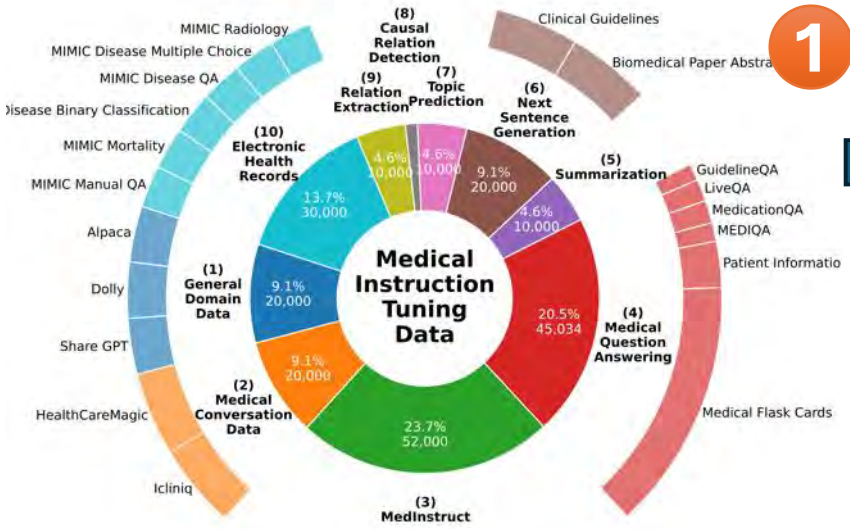
Published medical cases as examples

Biomedical literature

Real observations (wait!)

Me LLaMA – A Medical Large Language Model (LLM)

129 Billion Token Medical Corpus



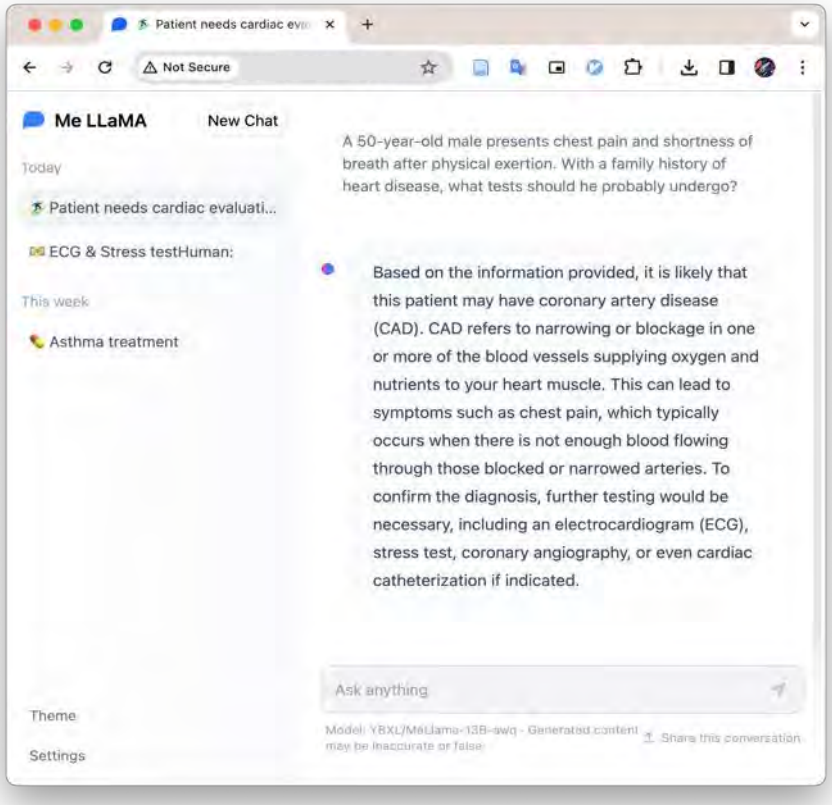
1 Continued Pre-Training

Medical Foundation Models

2 Instruction Tuning

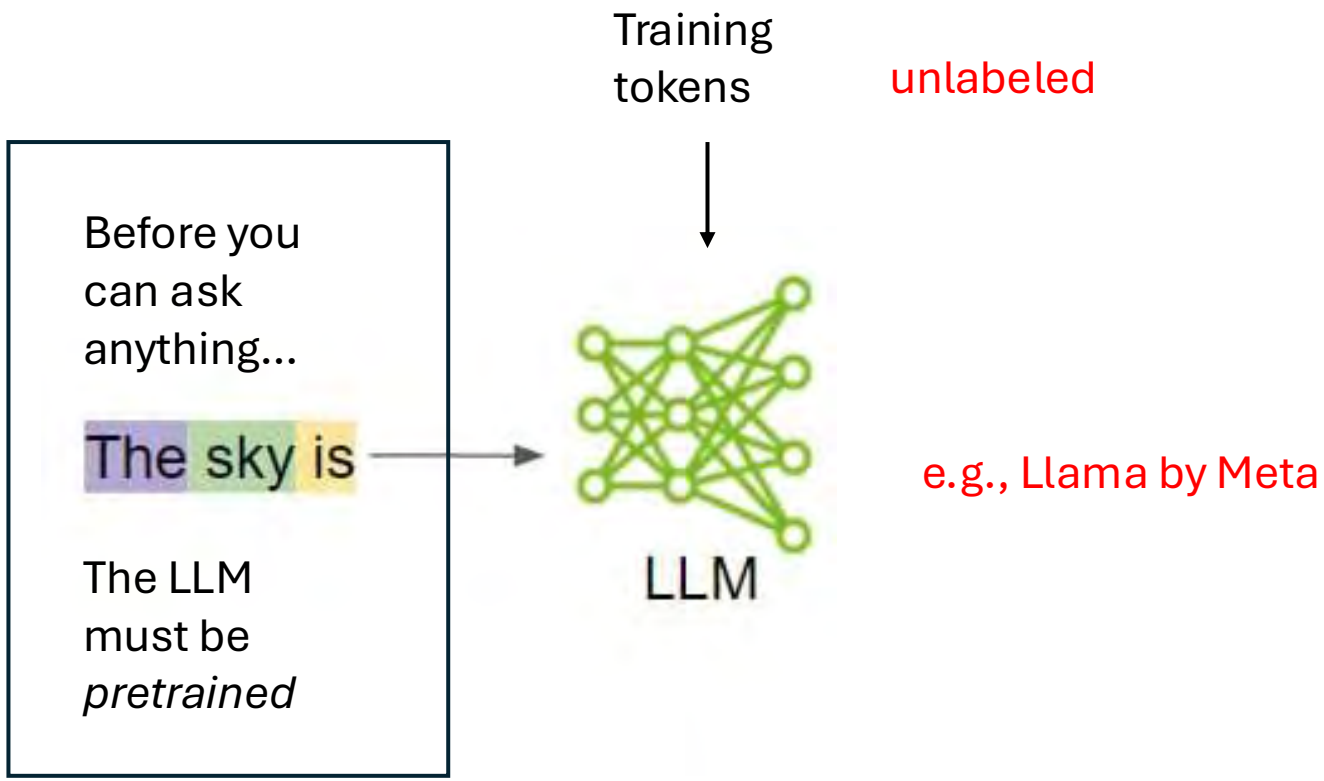
Me LLaMA Model

3 Interface



Slide from Dr. Hua Xu

Pretraining an LLM



Team

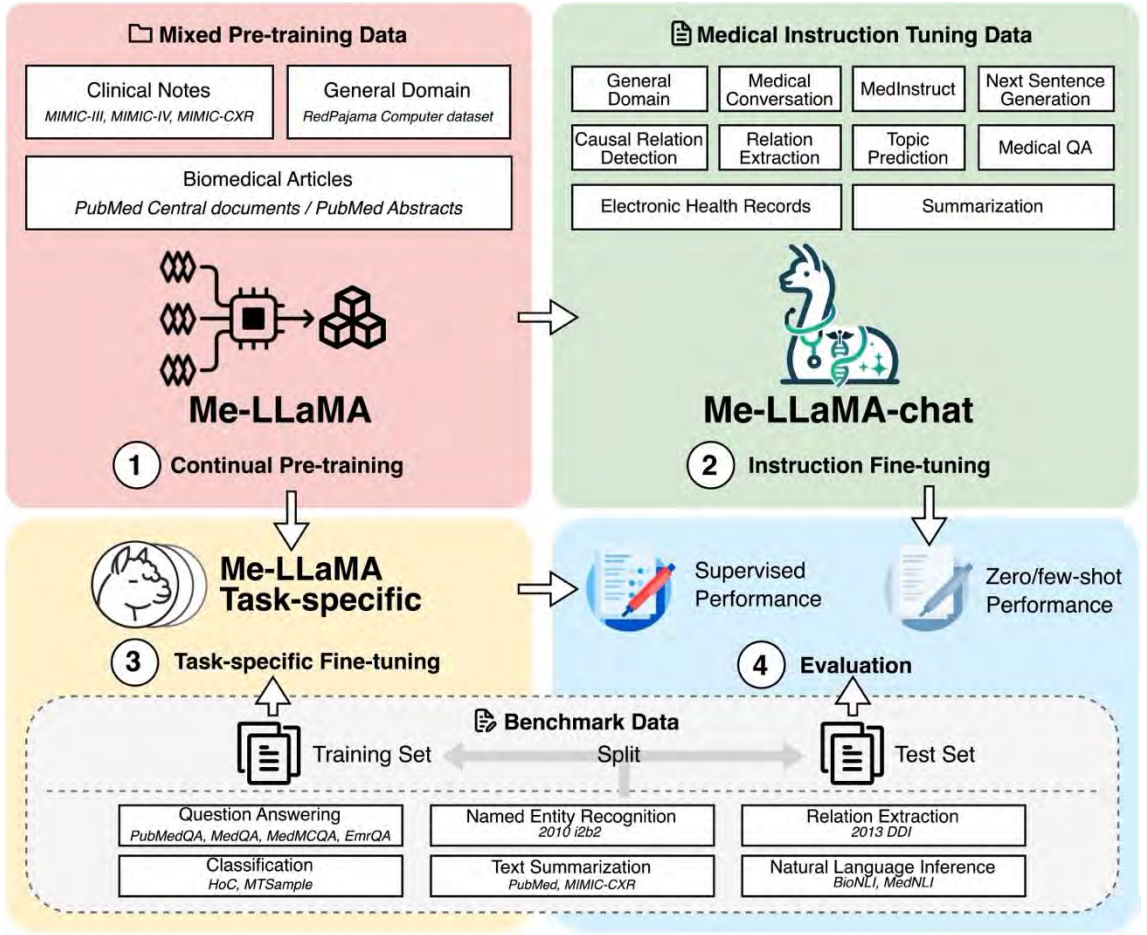
Yale University School of Medicine

University of Florida

University of Texas Health

Me LLaMA: Foundation Biomedical LLMs

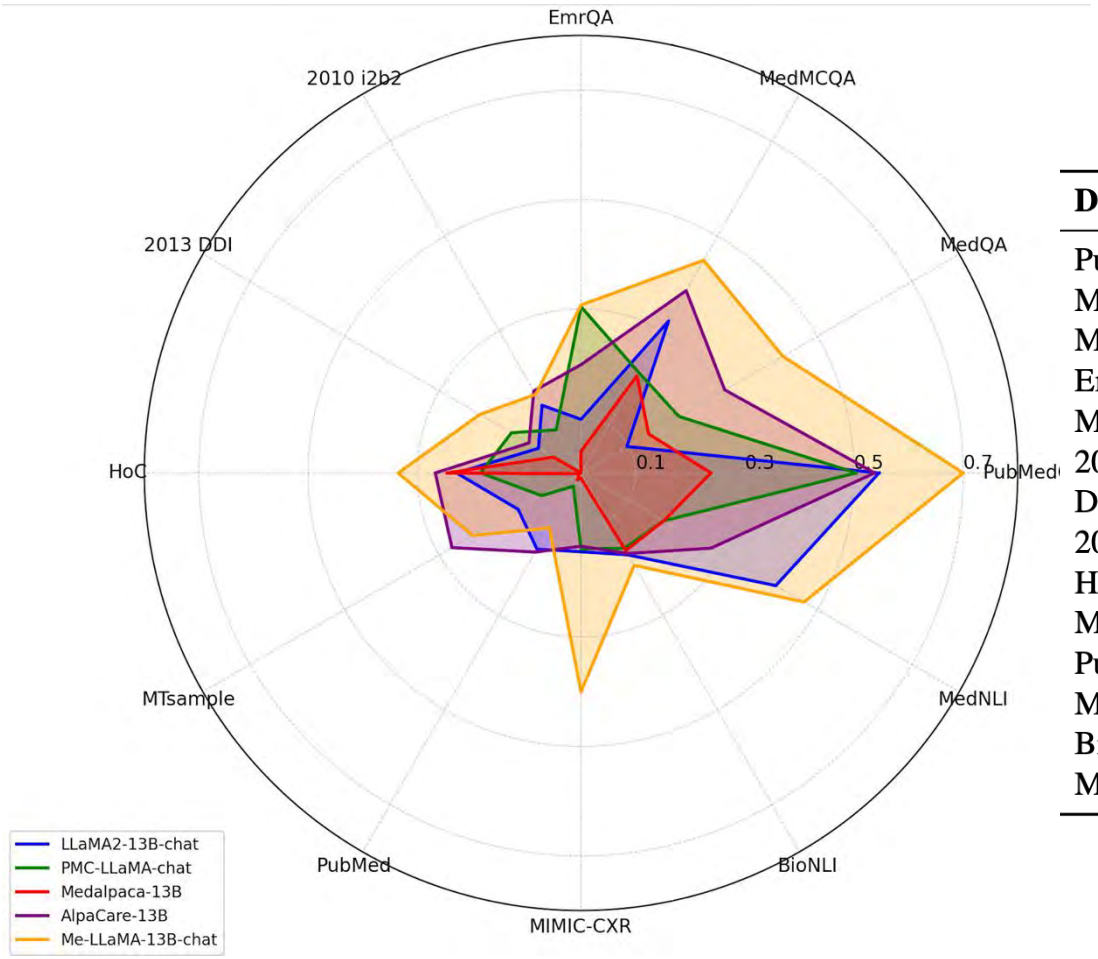
- Continual pre-training of LLaMA2: Trained on 129B tokens of biomedical data, with 100,000+ GPU hours
- Instruction fine-tuning: Trained on 200K+ medical QA pairs, with 1,000+ GPU hours
- Task-specific fine-tuning: Trained and evaluated on 6 tasks, 12 datasets
- Available at 13B and 70B models



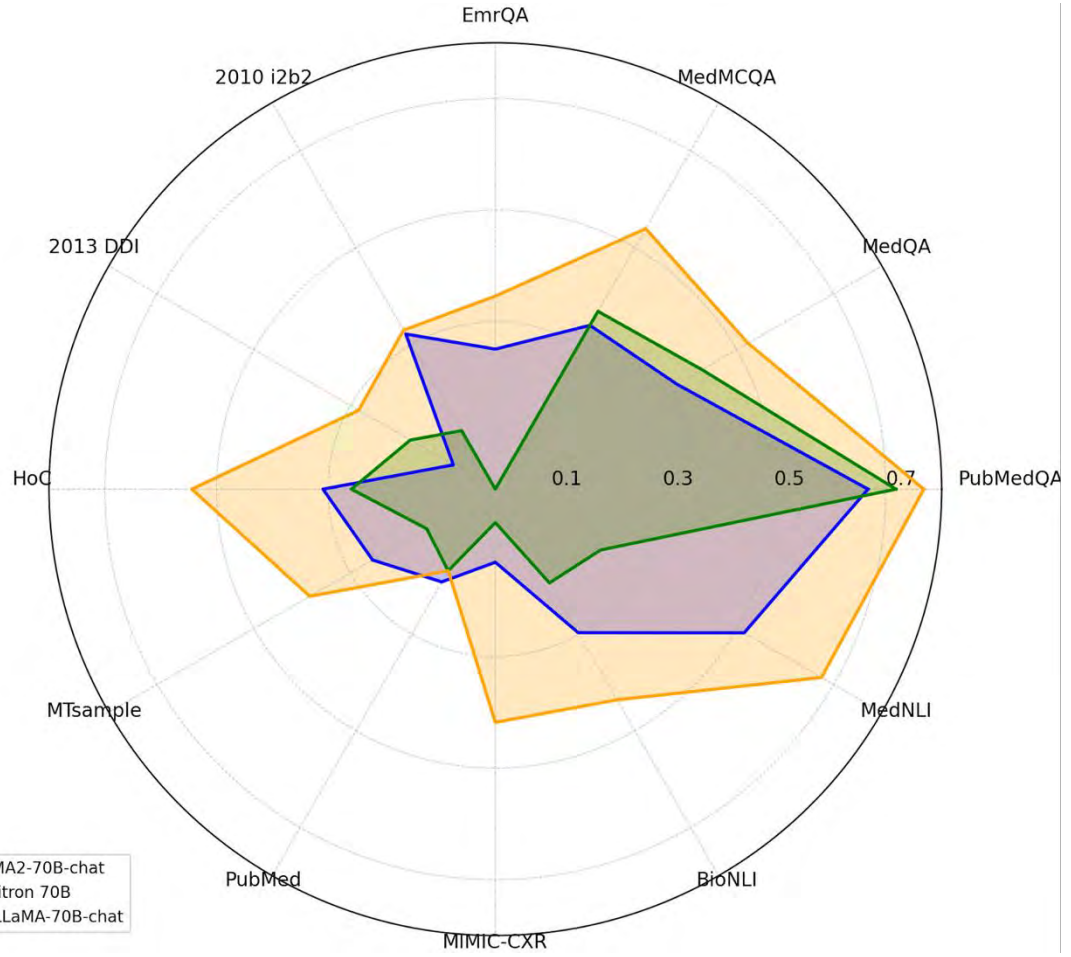
Xie, Q., *et al.* Medical foundation large language models for comprehensive text analysis and beyond. *npj Digit. Med.* **8**, 141 (2025)

Slide from Dr. Xu

Me LLAMA: Outperform Existing Open Medical LLMs



Data	Task
PubMedQA	QA
MedQA	QA
MedMCQA	QA
EmrQA	QA
MMLU	QA
2012 i2b2	NER
DDI2013	RE
2018 n2c2	RE
HoC	CF
MTSample	CF
PubMedSum	TS
MIMIC-CXR	TS
BioNLI	NLI
MedNLI	NLI



Best on 9 out of 12 datasets on 13B

Best on 11 out of 12 datasets on 70B

Slide from Dr. Xu

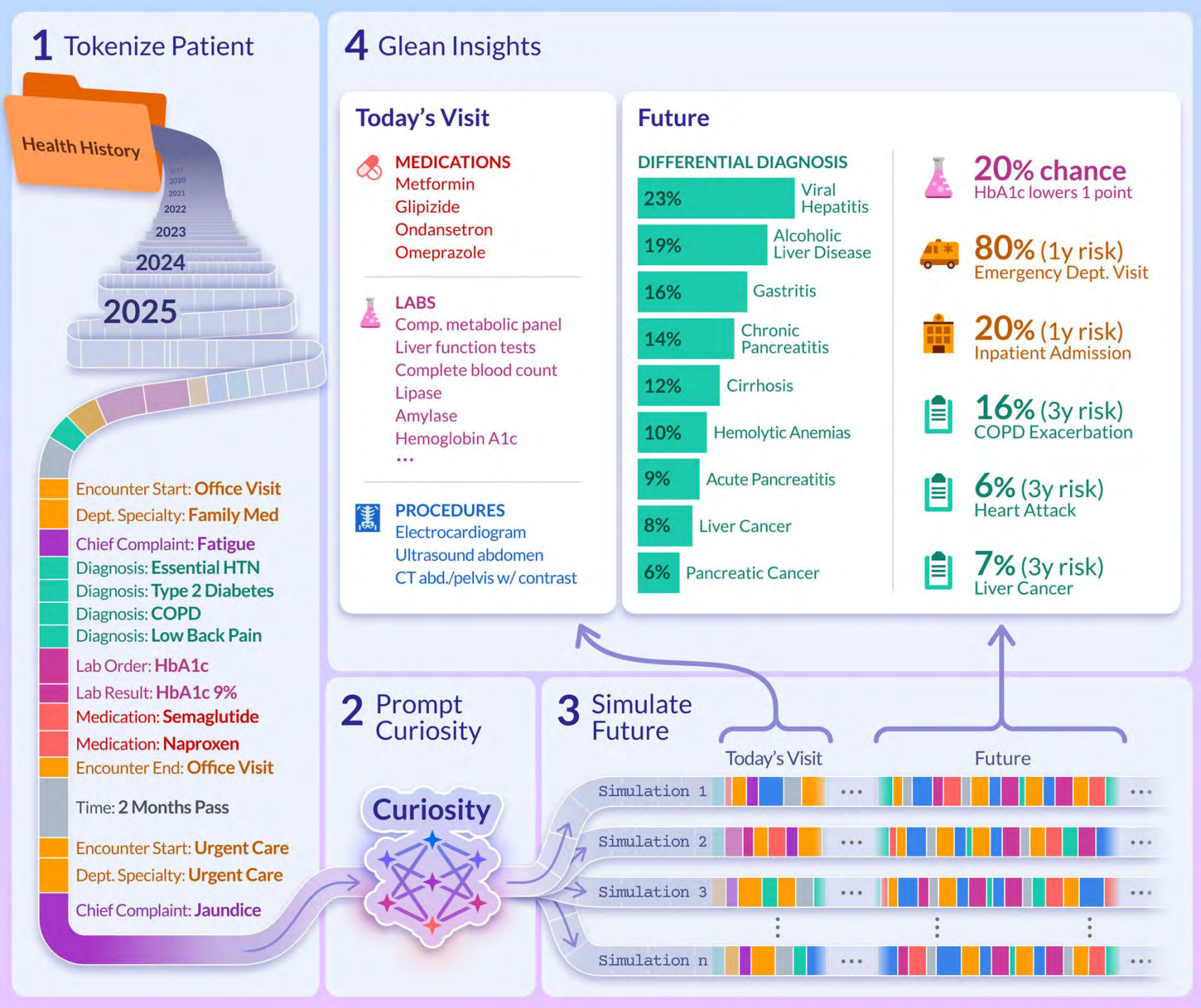
Curiosity: Cosmos-Based Foundation Model

Generative Medical Event Models Improve with Scale

Shane Waxler*[1], Paul Blazek*[1], Davis White*[1], Daniel Sneider*[1],
Kevin Chung[1], Mani Nagarathnam[1], Patrick Williams[1], Hank Voeller[1], Karen Wong[1],
Matthew Swanhorst[1], Sheng Zhang[2], Naoto Usuyama[2], Cliff Wong[2], Tristan Naumann[2],
Hoifung Poon[2], Andrew Loza[3], Daniella Meeker[3, 4], Seth Hain[1], and Rahul Shah†[1]

[1]Epic Systems
[2]Microsoft Research
[3]Yale School of Medicine
[4]Cosmos Governing Council

Slide from Dr. Meeker



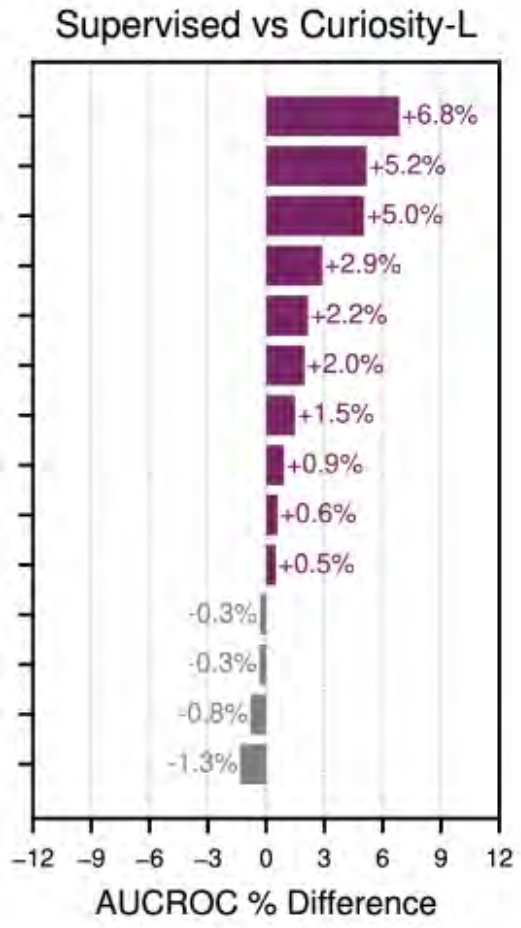
Curiosity

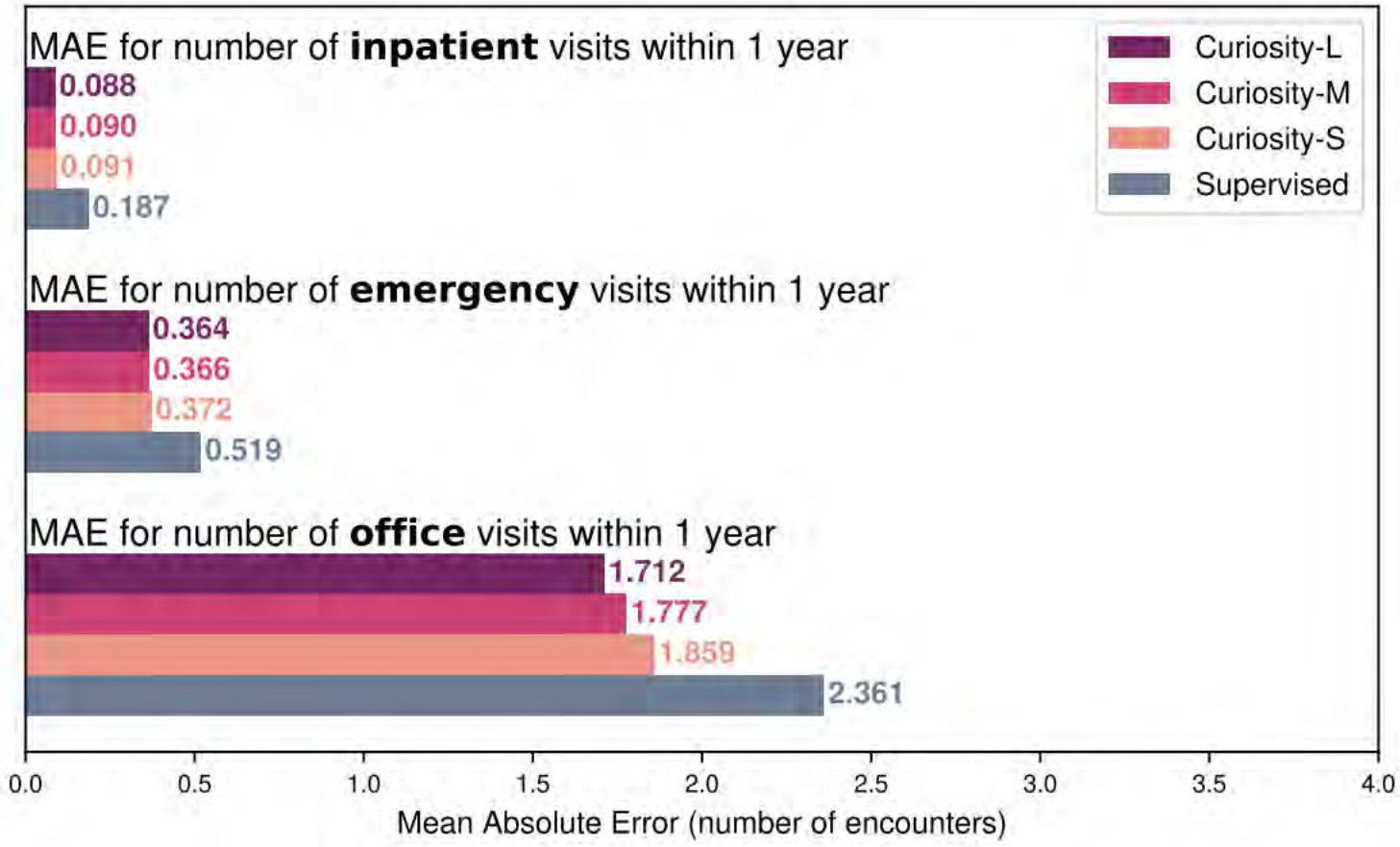
A Foundation model for EMR data

LLM-style "next word" prediction, but for coded sequences of medical events instead of lexical data

Slide from Dr. Meeker

One model, predicting any clinical event with comparable accuracy to “hand-crafted” ML models on 78 prediction tasks





Slide from Dr. Meeker

Challenges

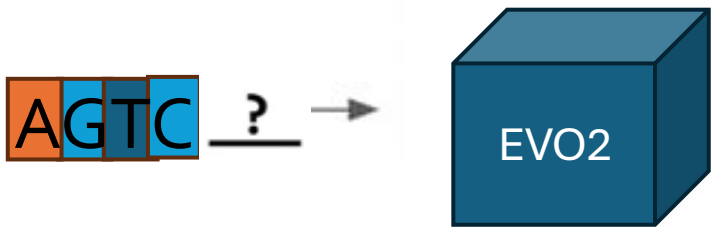
- Privacy (LLMs memorize cases)
- Black box nature of AI
- Computational infrastructure is demanding

Opportunities

- Zero shot performance
- Multimodal training
- Hypothesis generation

DNA LLM

DNA LLM

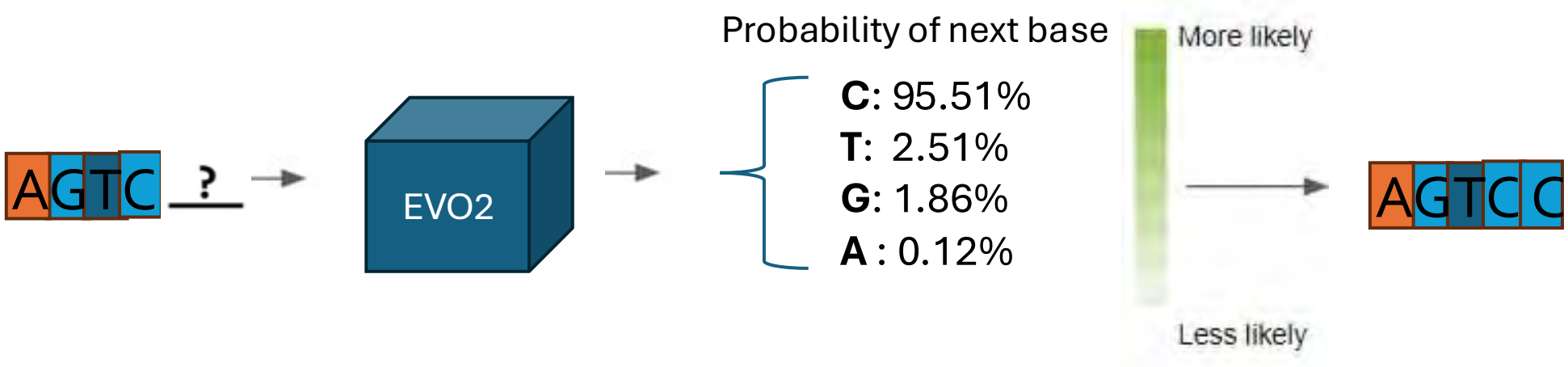


Genome modeling and design across all domains of life with Evo 2

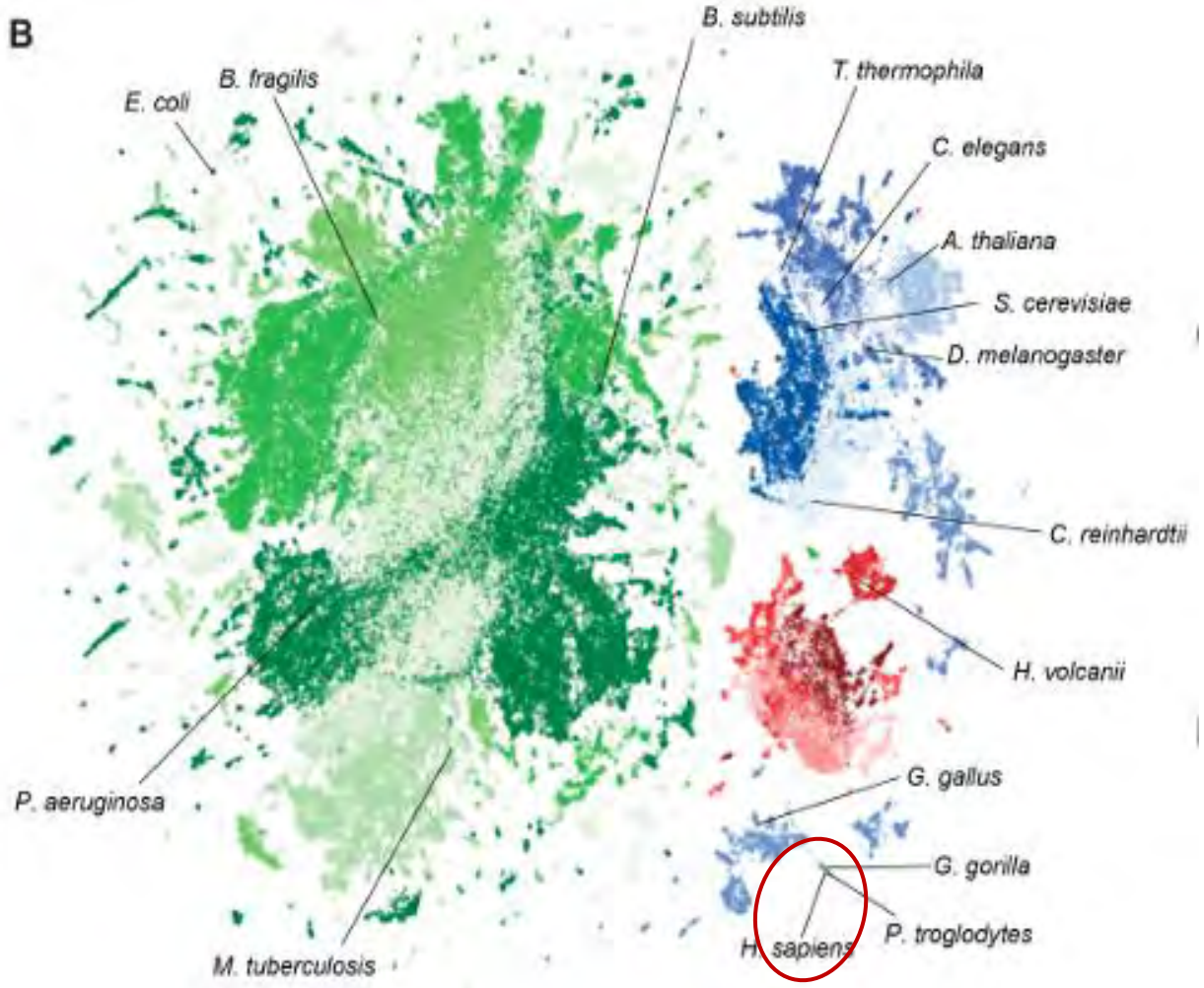
Garyk Brixi, …Brian L. Hie

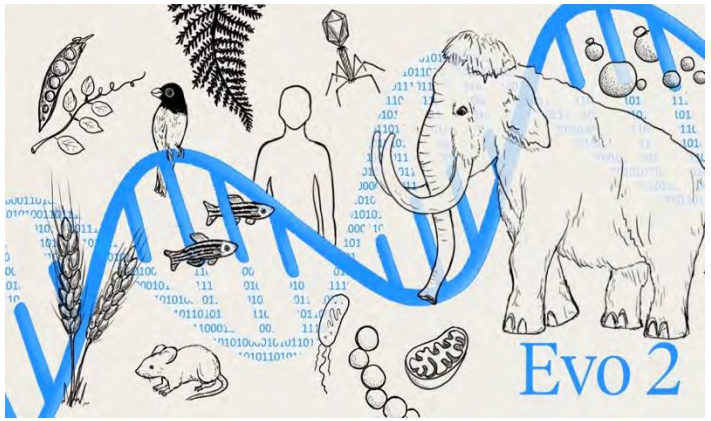
bioRxiv 2025.02.18.638918; doi: https://doi.org/10.1101/2025.02.18.638918

DNA LLM


AGTC ?
EVO2
Probability of next base
C: 95.51%
T: 2.51%
G: 1.86%
A : 0.12%
More likely
Less likely
AGTCC

An LLM model for eukaryotic and prokaryotic DNA – EVO 2

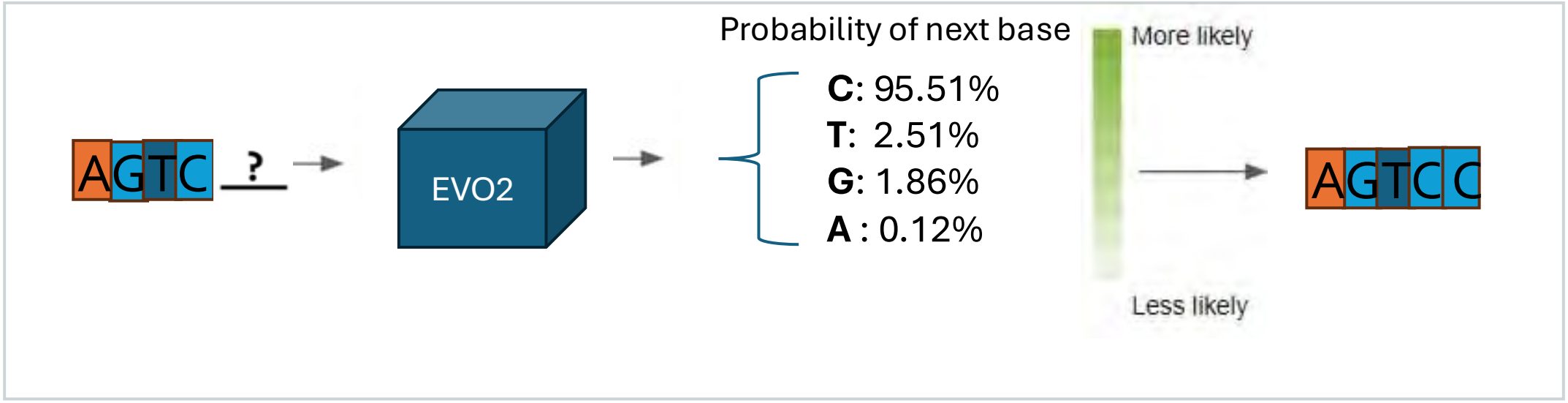




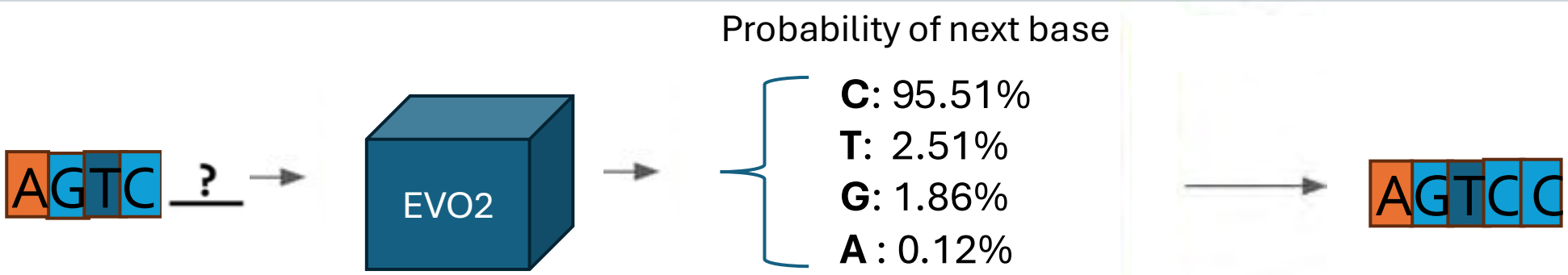
Brian Hie, 21 Feb 2025, bioRxiv

- Trained on **9.3 trillion** tokens (~3000 the human genome reference) from bacteria, archaea, eukaryotes, and phages.
- **40 and 7-billion-parameter** genomic foundation model for 1Mbp DNA input

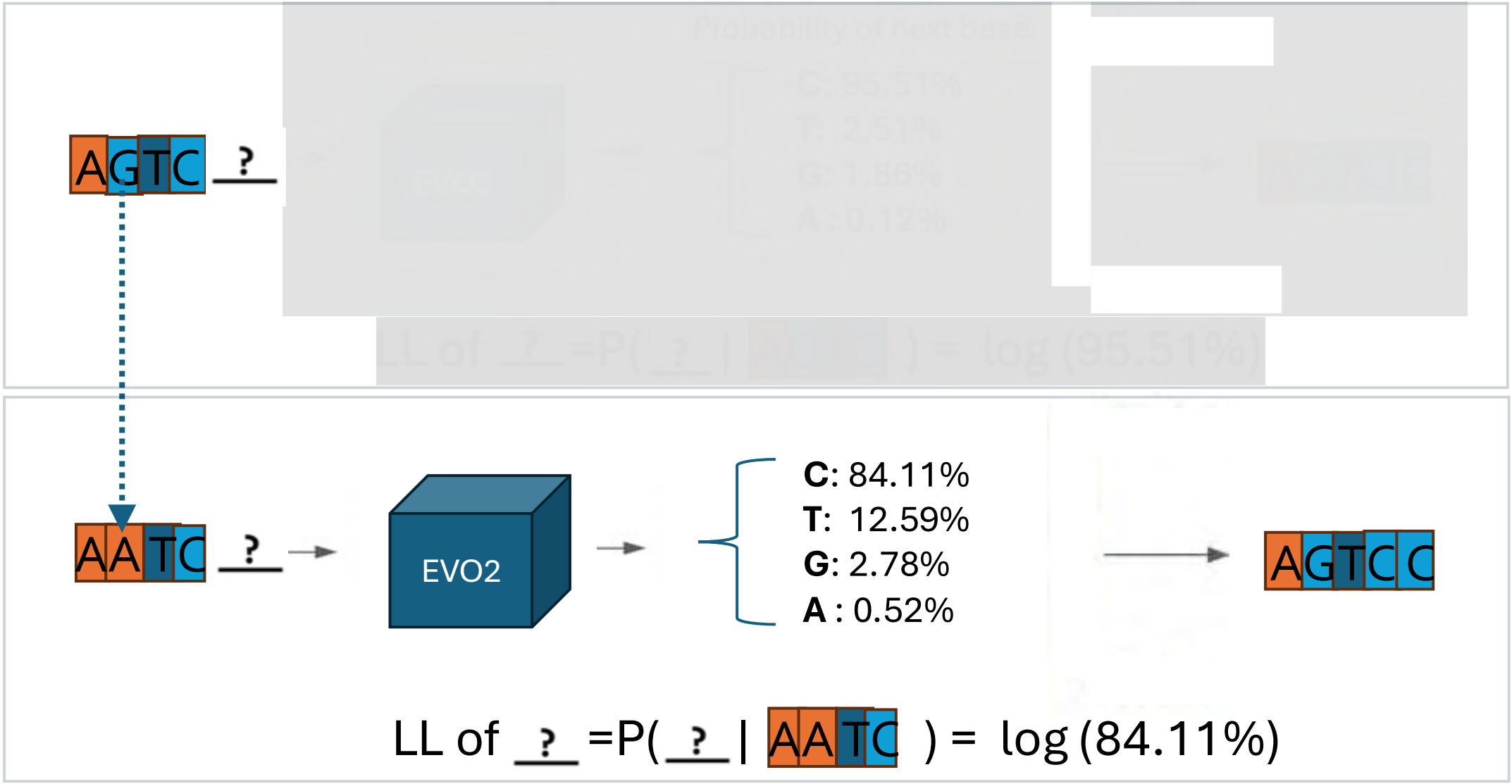
Single variant log-likelihood (LL) & delta log-likelihood



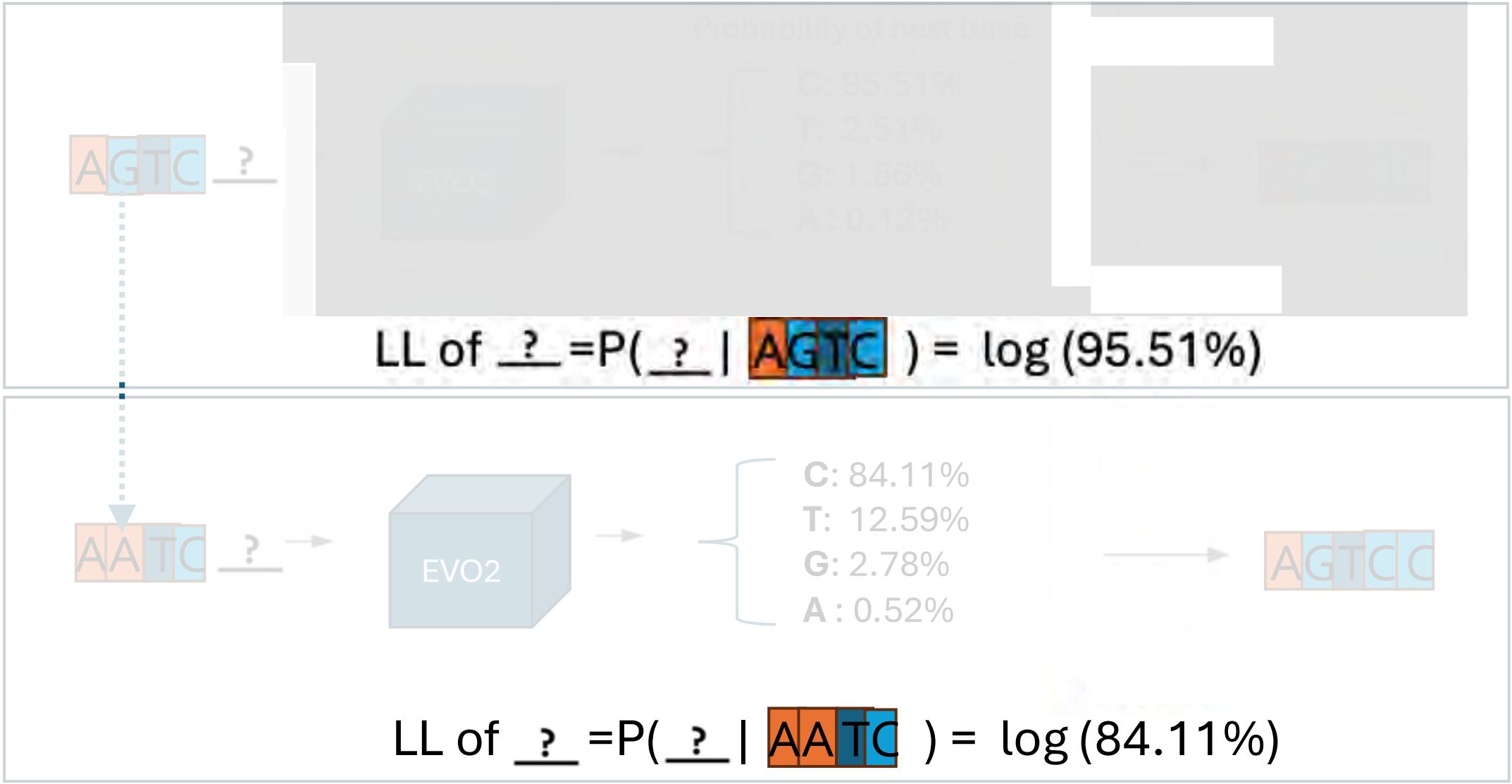
Single variant log-likelihood (LL) & delta log-likelihood



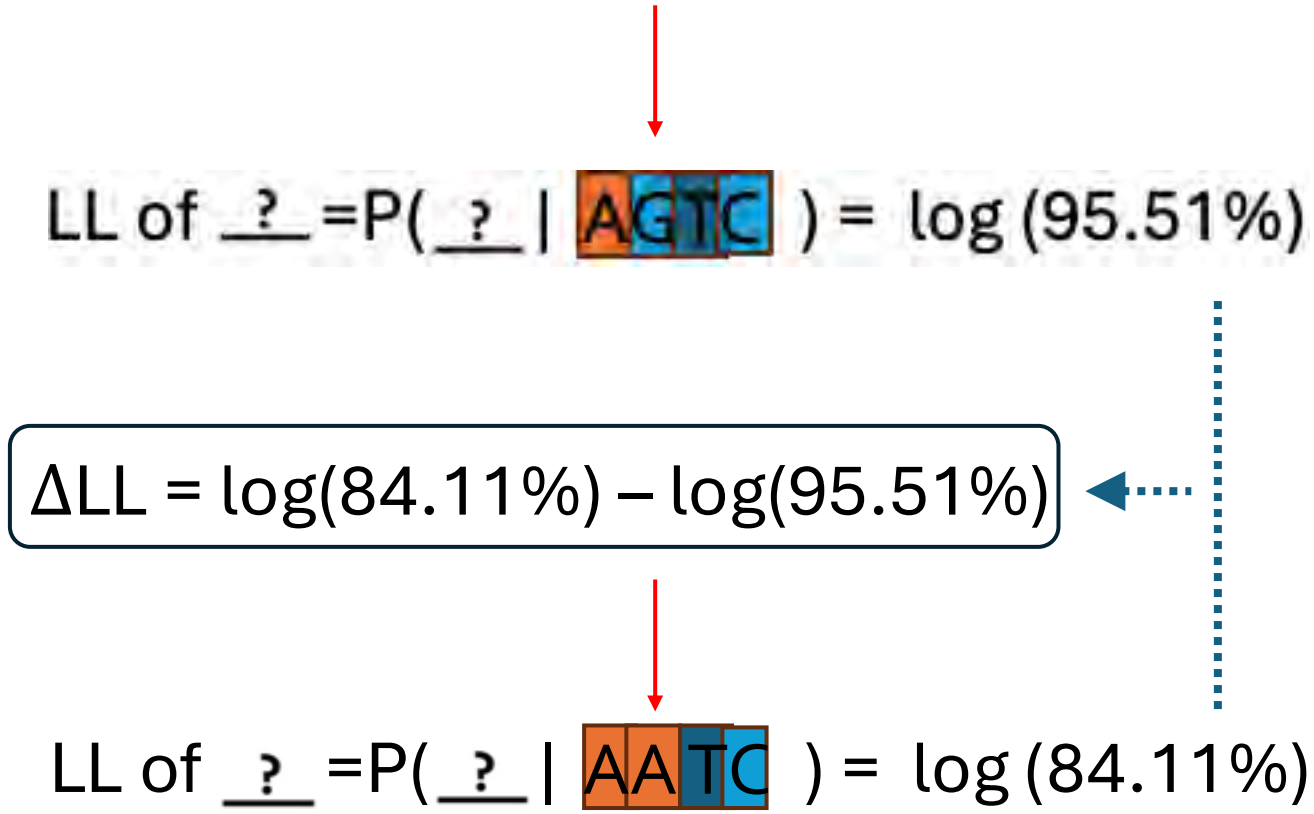
Single variant log-likelihood (LL) & delta log-likelihood



Single variant log-likelihood (LL) & delta log-likelihood

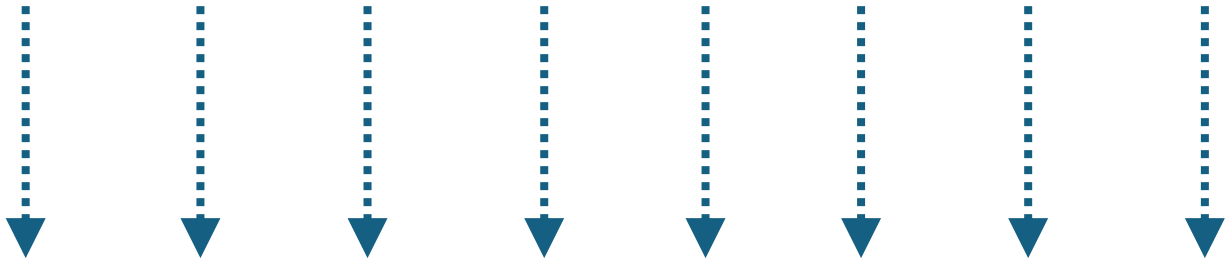


Single variant log-likelihood (LL) & delta log-likelihood



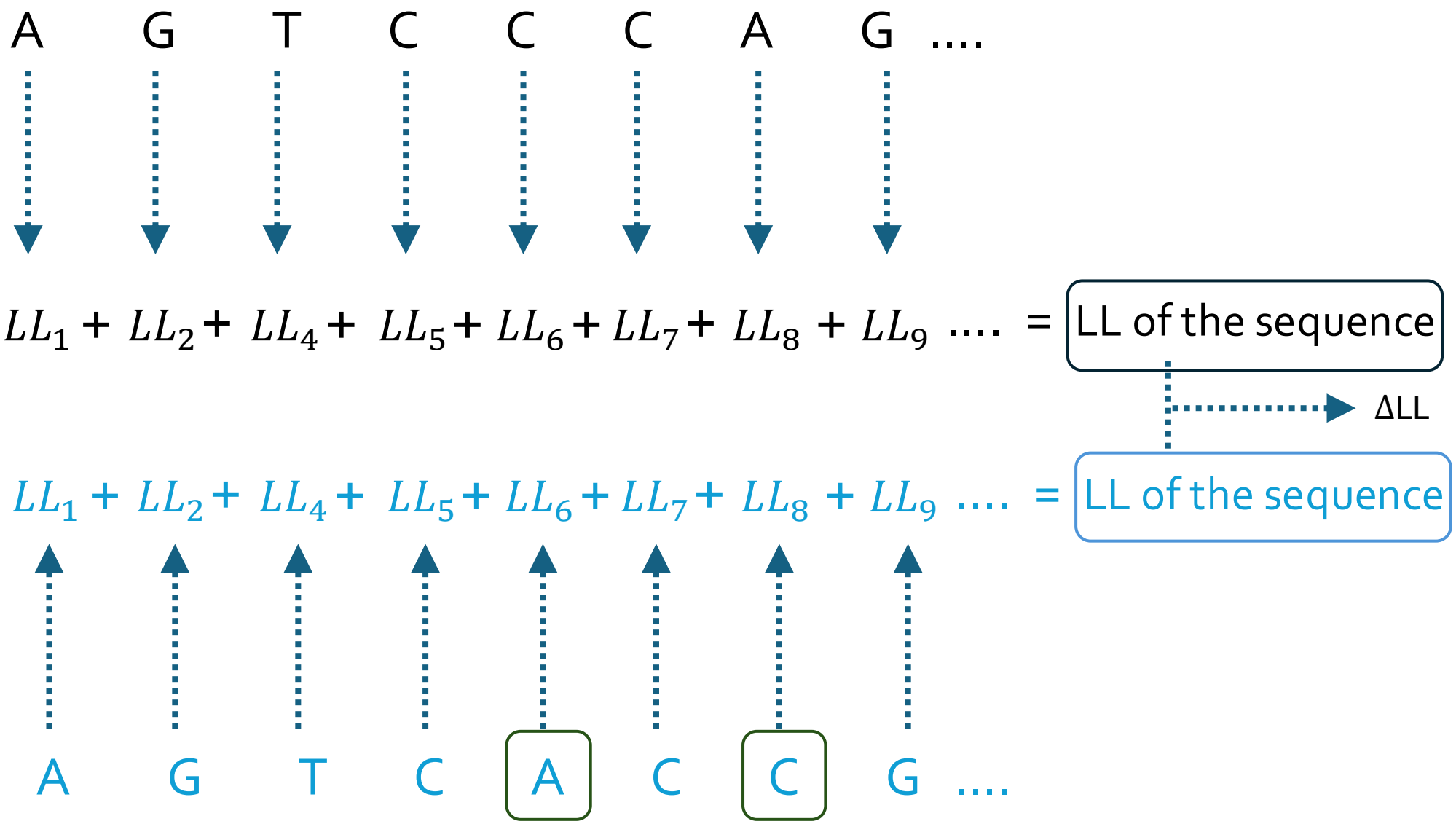
Sequence log-likelihood

A G T C C C A G ....

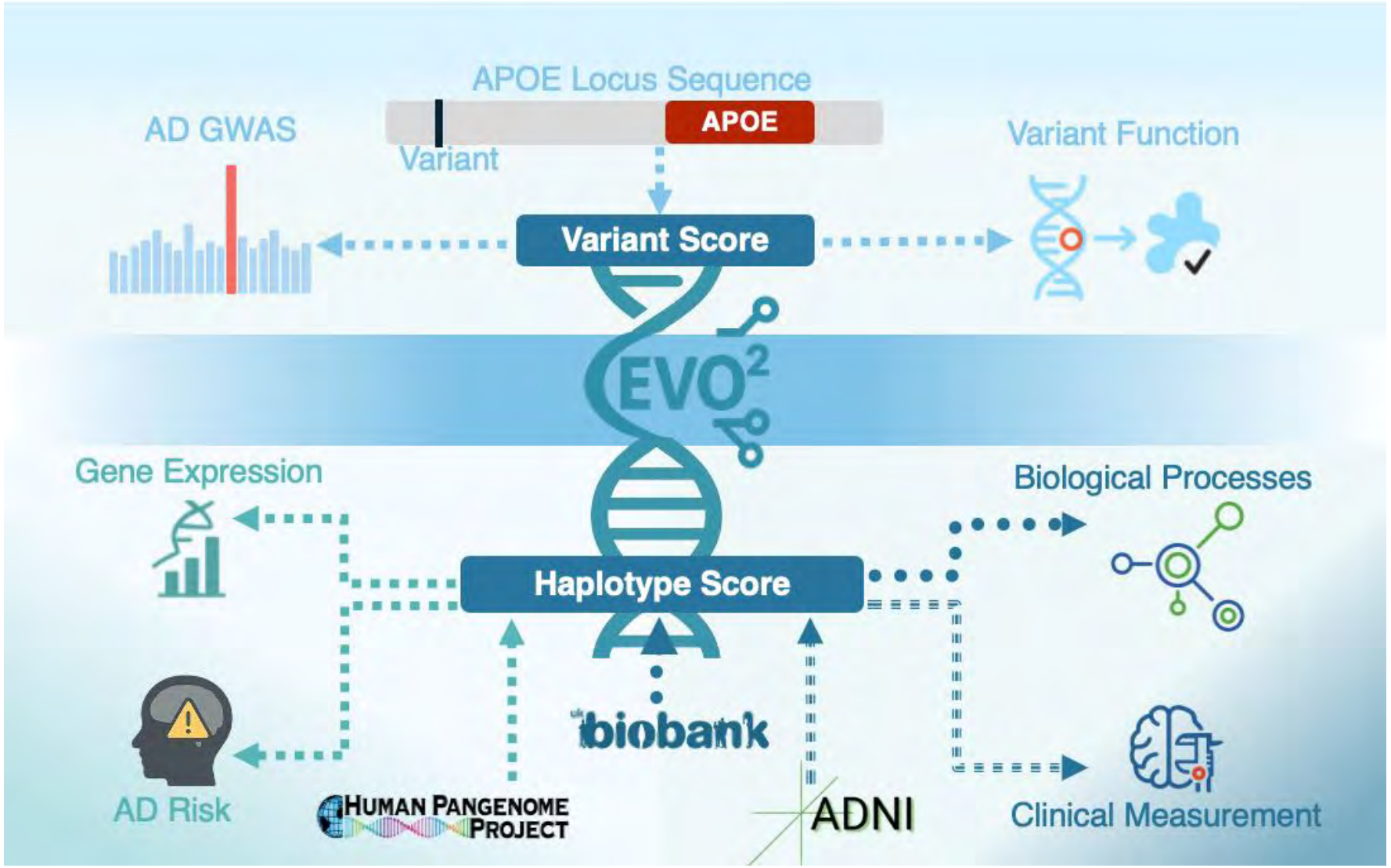


$LL_1 + LL_2 + LL_4 + LL_5 + LL_6 + LL_7 + LL_8 + LL_9 \ldots.$ = LL of the sequence

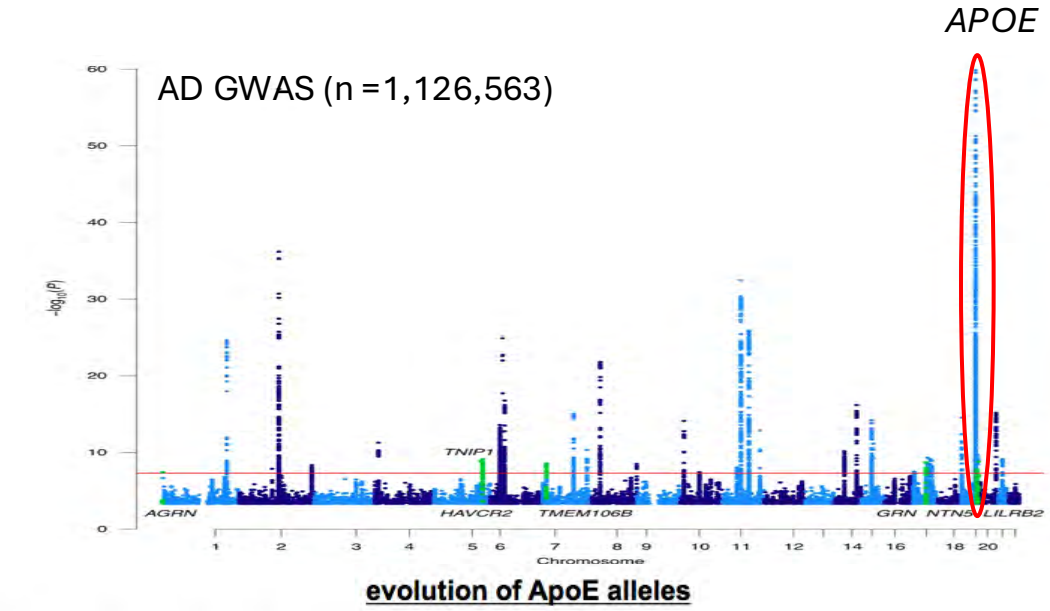
Sequence log-likelihood & delta log-likelihood

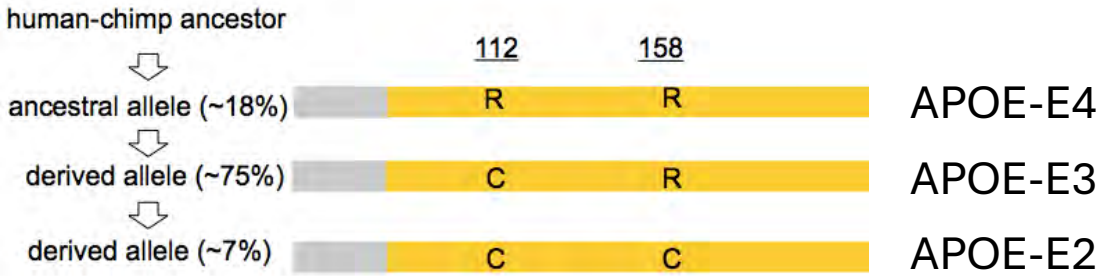


Using EVO 2 without new training



Alzheimer's disease and APOE

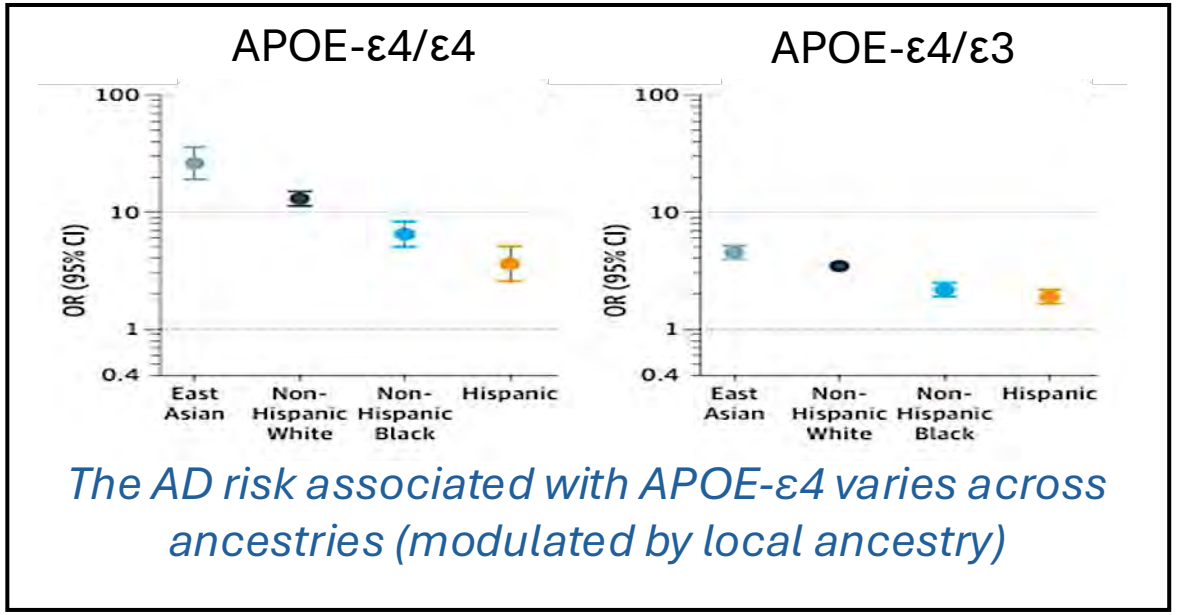




APOE-E4

APOE-E3

APOE-E2

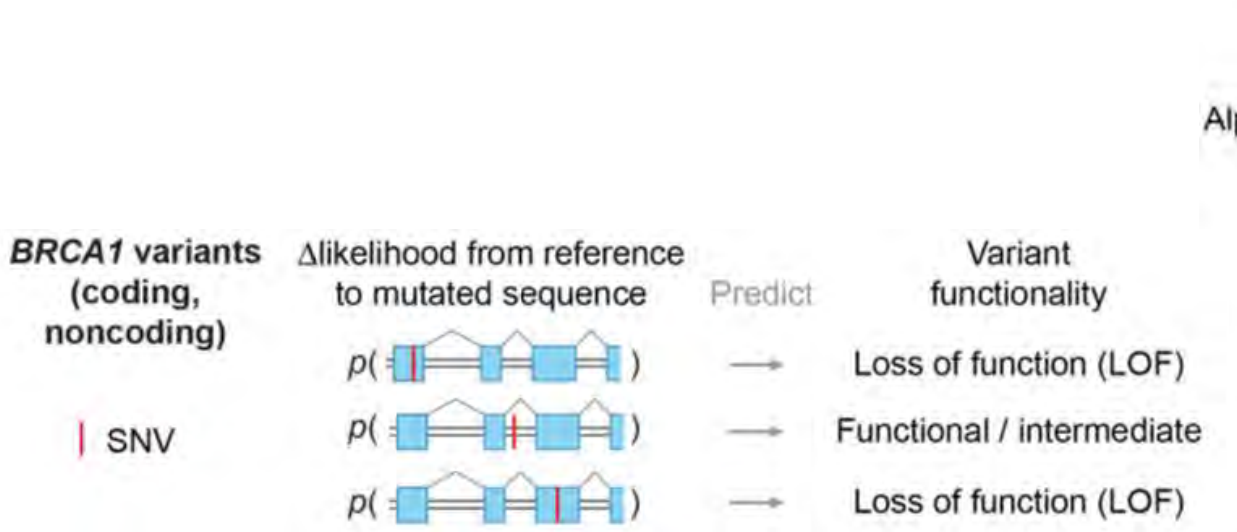


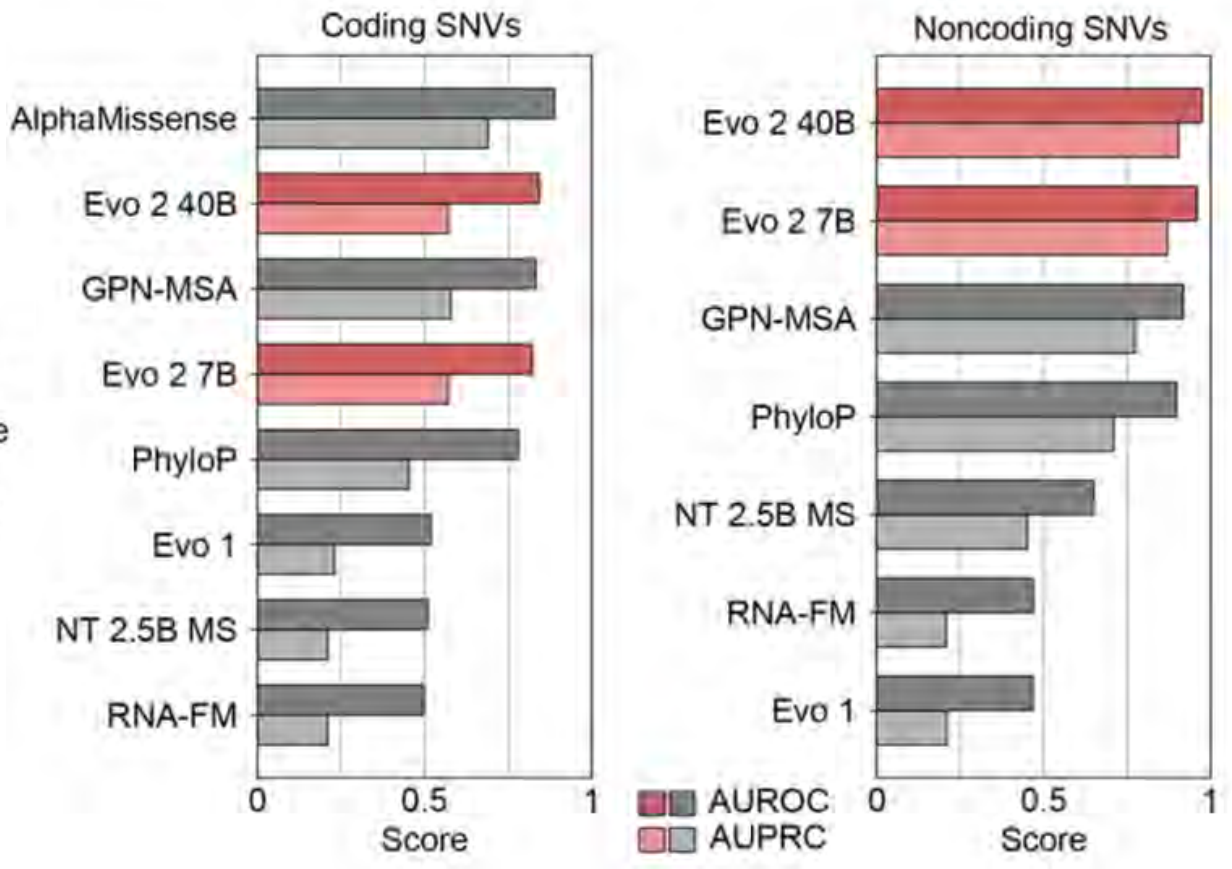
The AD risk associated with APOE-ε4 varies across ancestries (modulated by local ancestry)

*Individuals of **African ancestry** exhibit **lower APOE-E4–associated risk** for AD compared to those of European ancestry (modulated by local ancestry)*

doi:10.1001/jamaneurol.2023.3599

EVO2 and BRCA1





Genome modeling and design across all domains of life with Evo 2. Garyk Brixi, …Brian L. Hie
bioRxiv 2025.02.18.638918; doi: https://doi.org/10.1101/2025.02.18.638918

EVO2 Human Genomics Results

GRCh38 Reference Genome Only

Research Questions

Can EVO2 be applied to population/disease genomics research?

Team

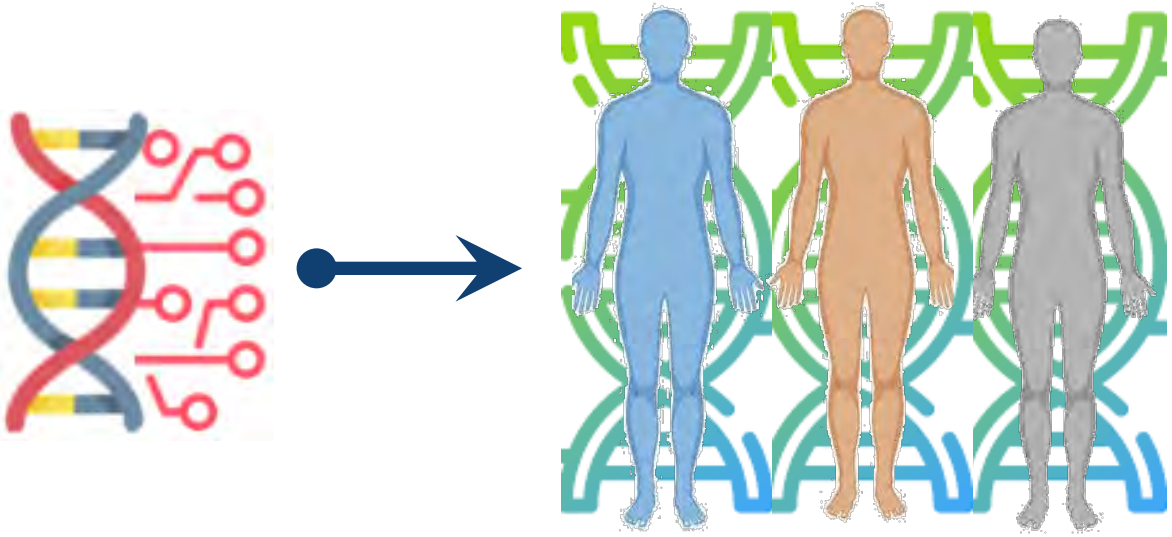
Yale University School of Medicine

University of Toronto SickKids Hospital

University College London

Indiana University Bloomington

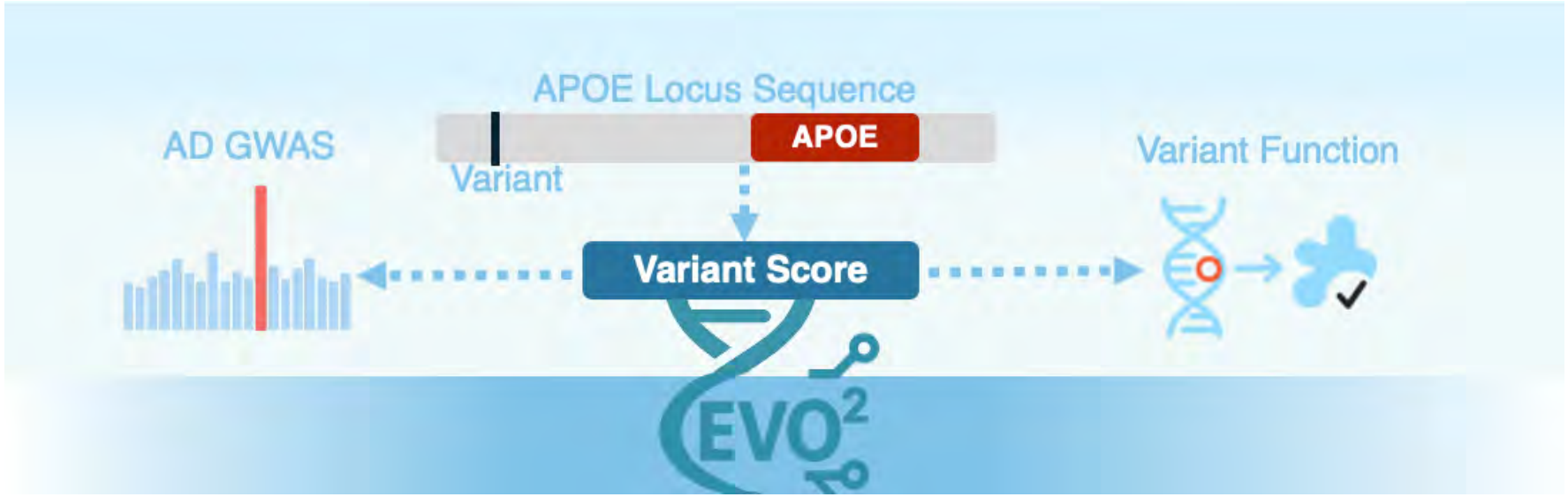
Evaluate EVO 2 on a population



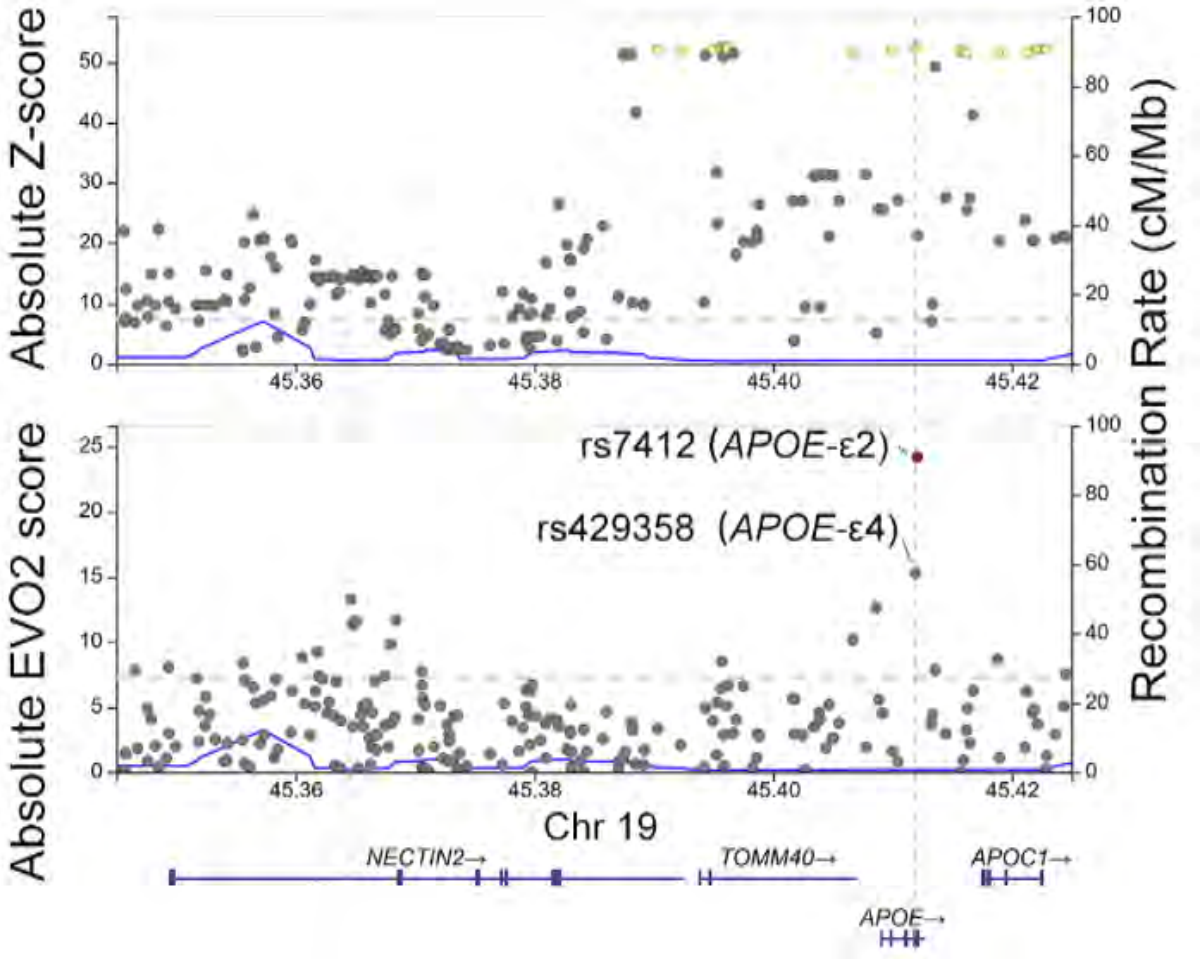
	Pos 1	Pos 2	...	Disease related measurement
Ind1	A	G	...	0
Ind2	T	A	...	1
Ind3	A	C	...	1
...	...	...	...	...
...	...	...	...	...
IndN	A	0	...	0

Zhu R et al. Advancing Human Population Genomics with DNA Foundation Models. medRxiv

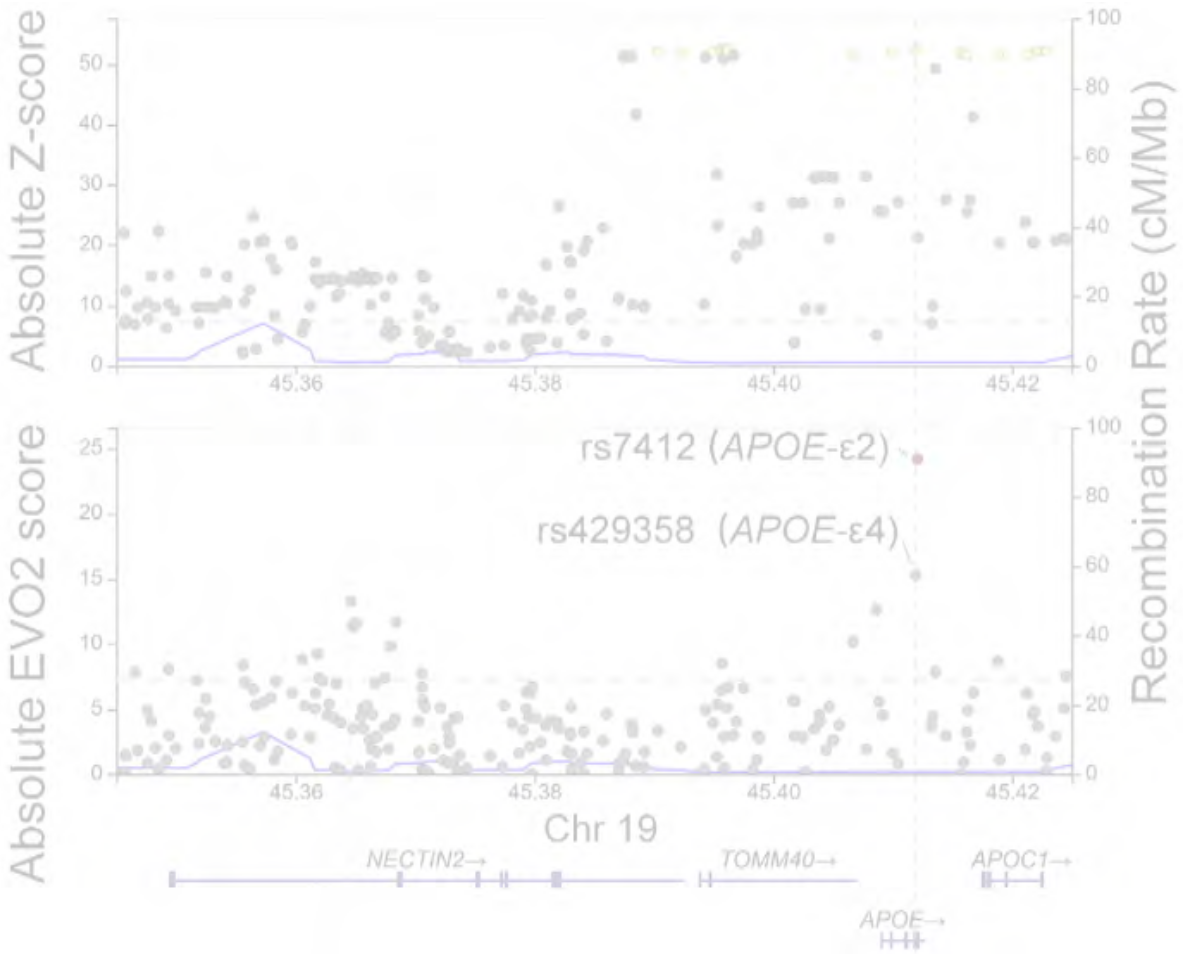
Zero shot performance

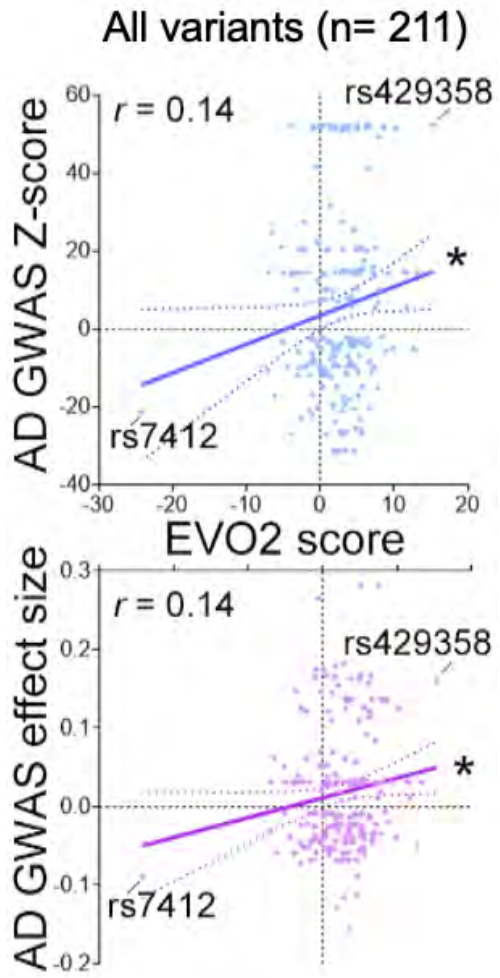


Alzheimer’s Disease (AD) GWAS and EVO2 score

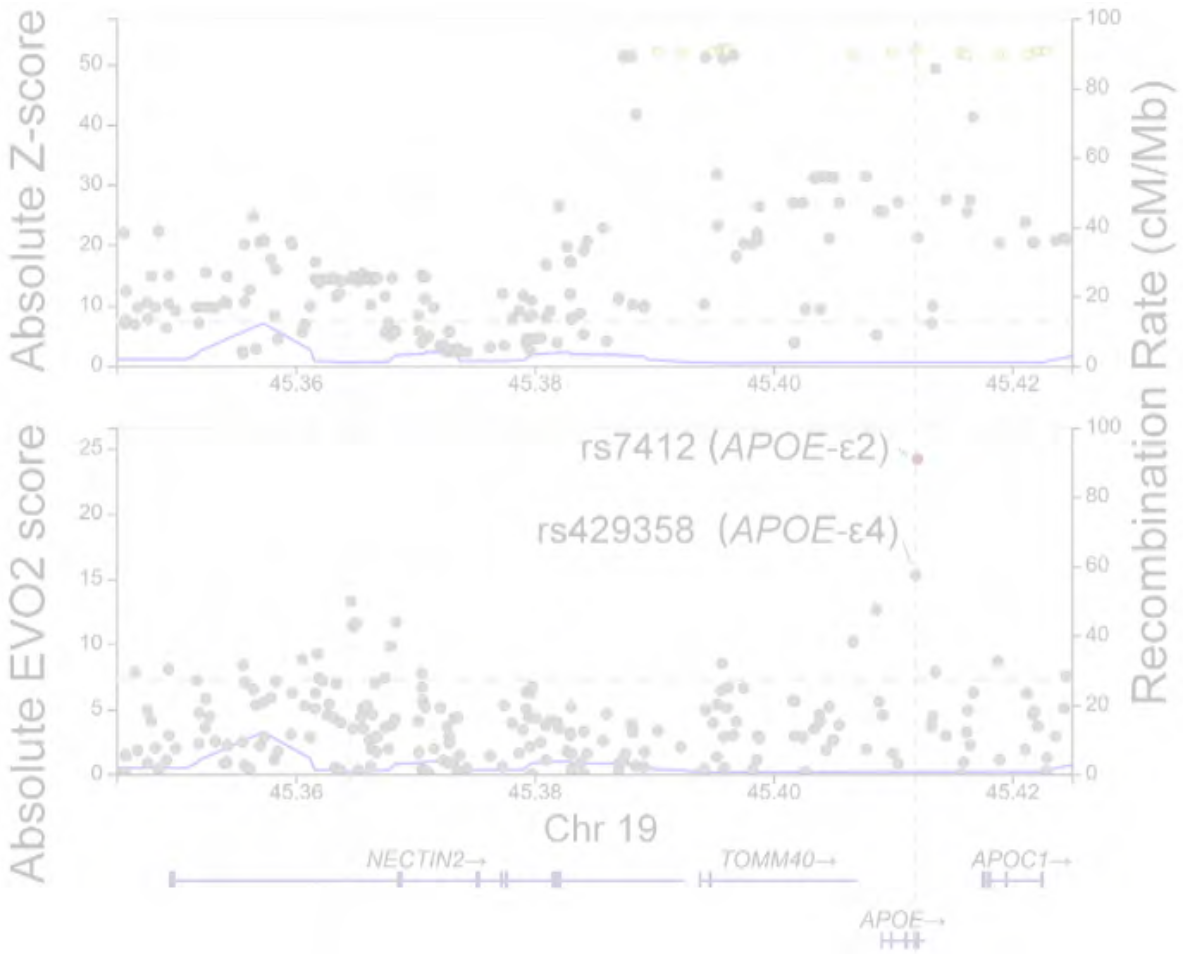


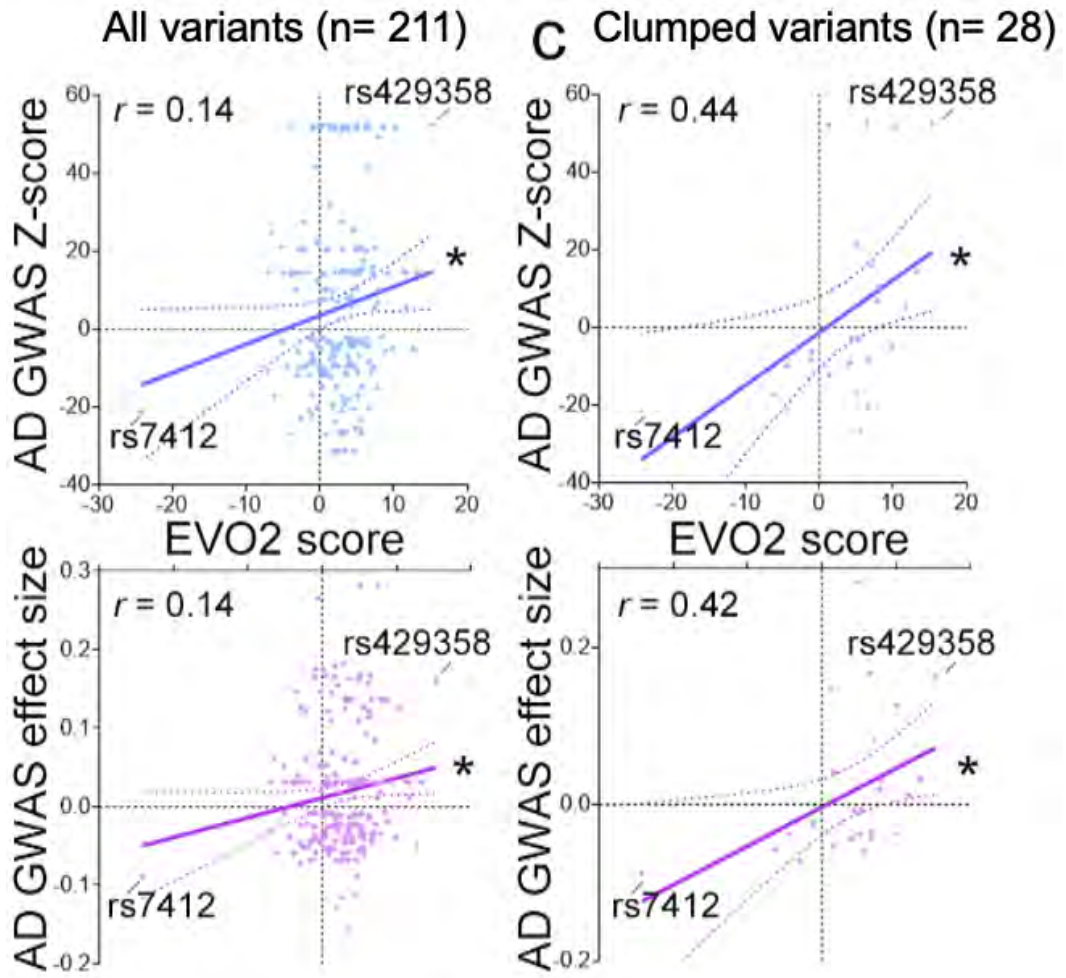
EVO2 score vs AD GWAS



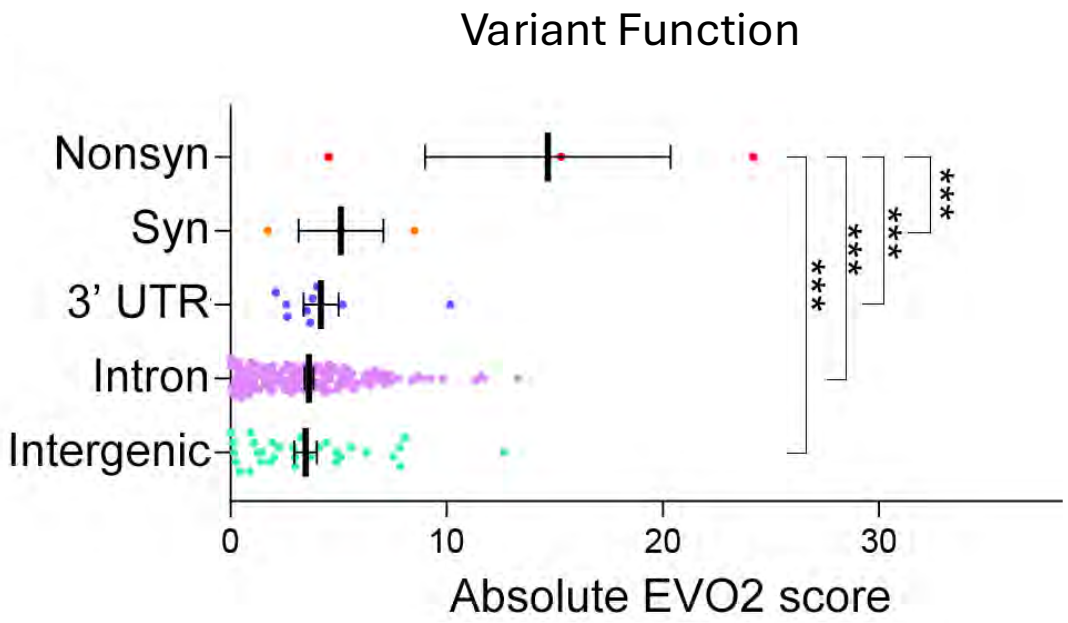


EVO2 score vs AD GWAS



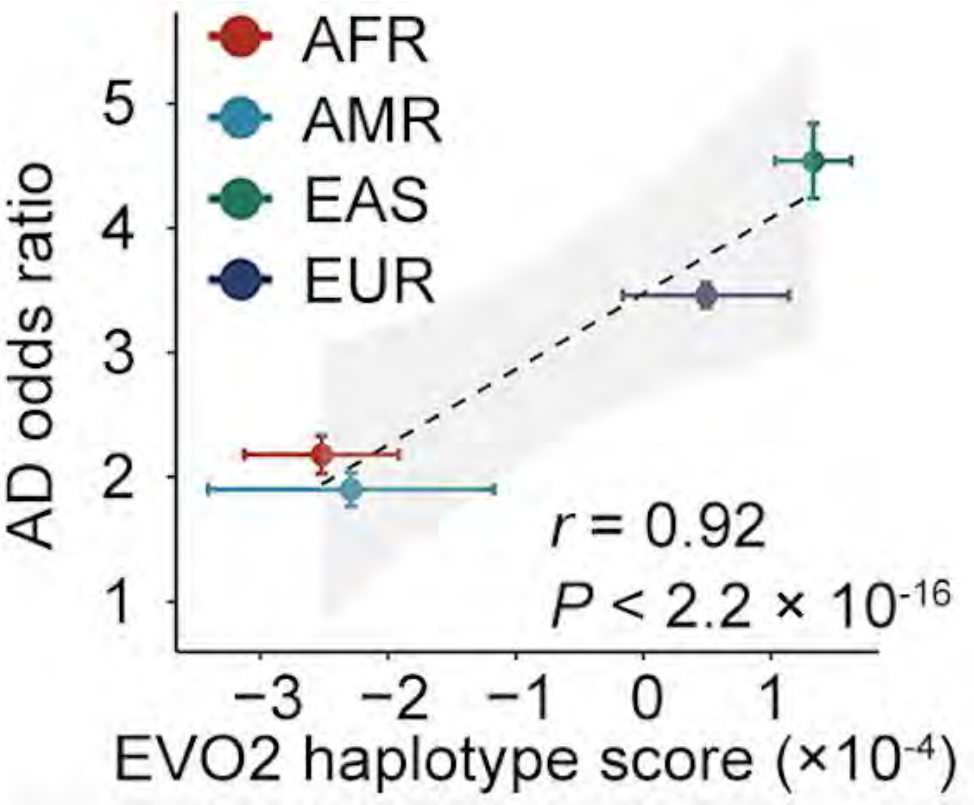


EVO2 Scores Prioritize Variants Based on Functional Impact



Nonsyn: nonsynonymous variant
Syn: synonymous variant
3' UTR: 3' untranslated region
Intron: intronic region
Intergenic: Intergenic region

EVO2 Haplotype Scores in the General Population and Across Diverse Ancestry Backgrounds

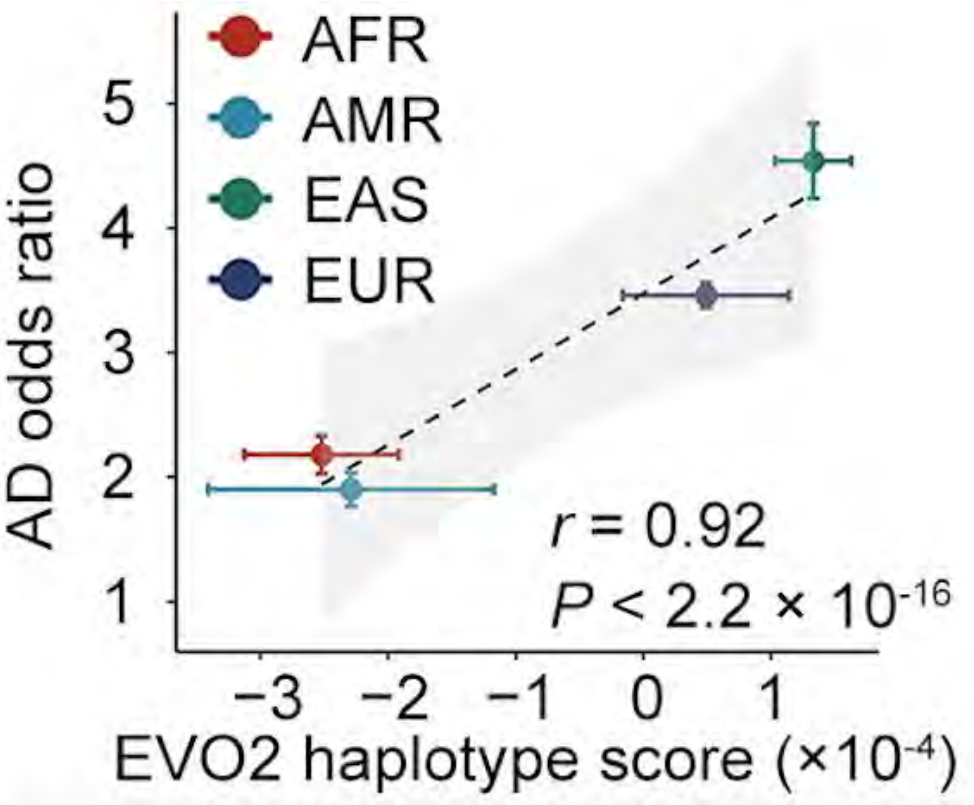


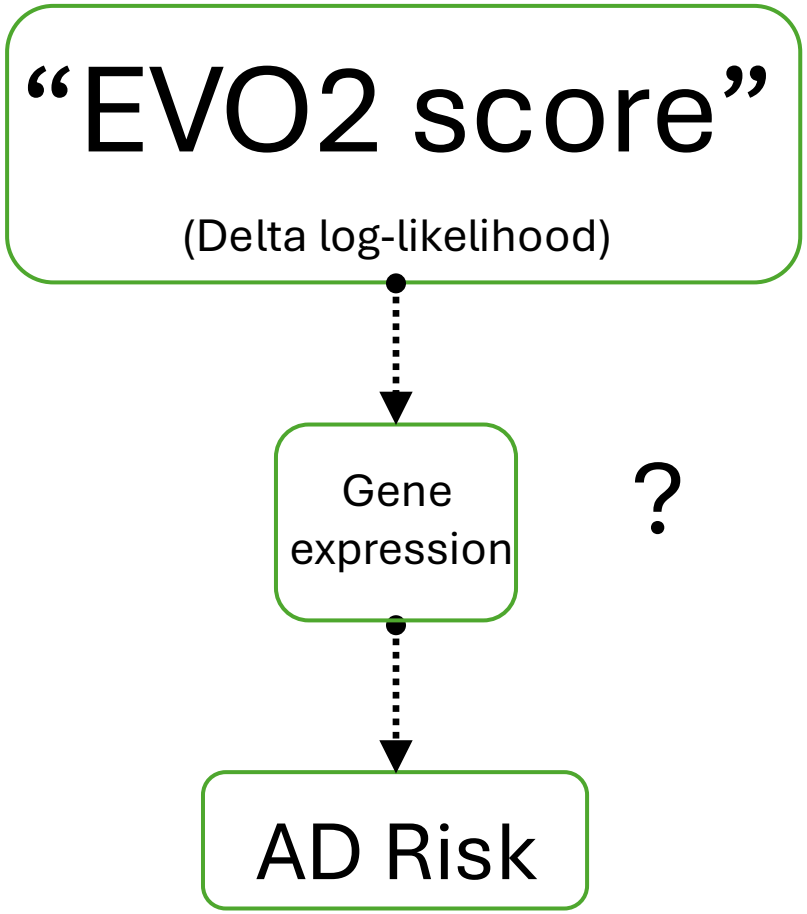
"EVO2 score"

(Delta log-likelihood)

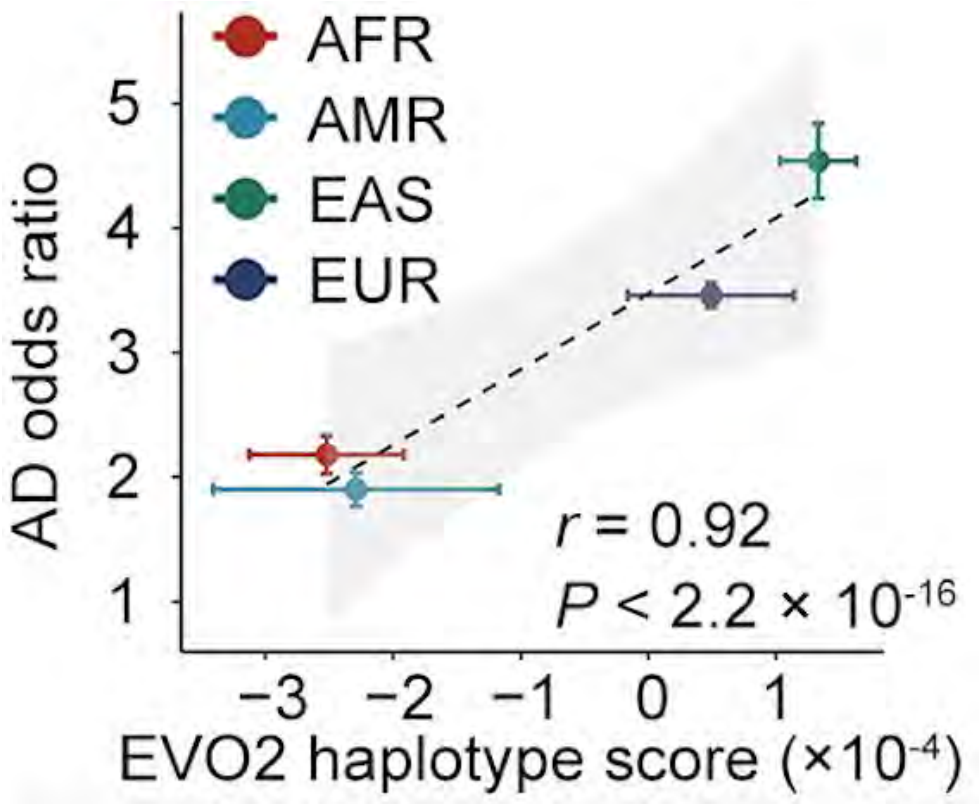
AD Risk

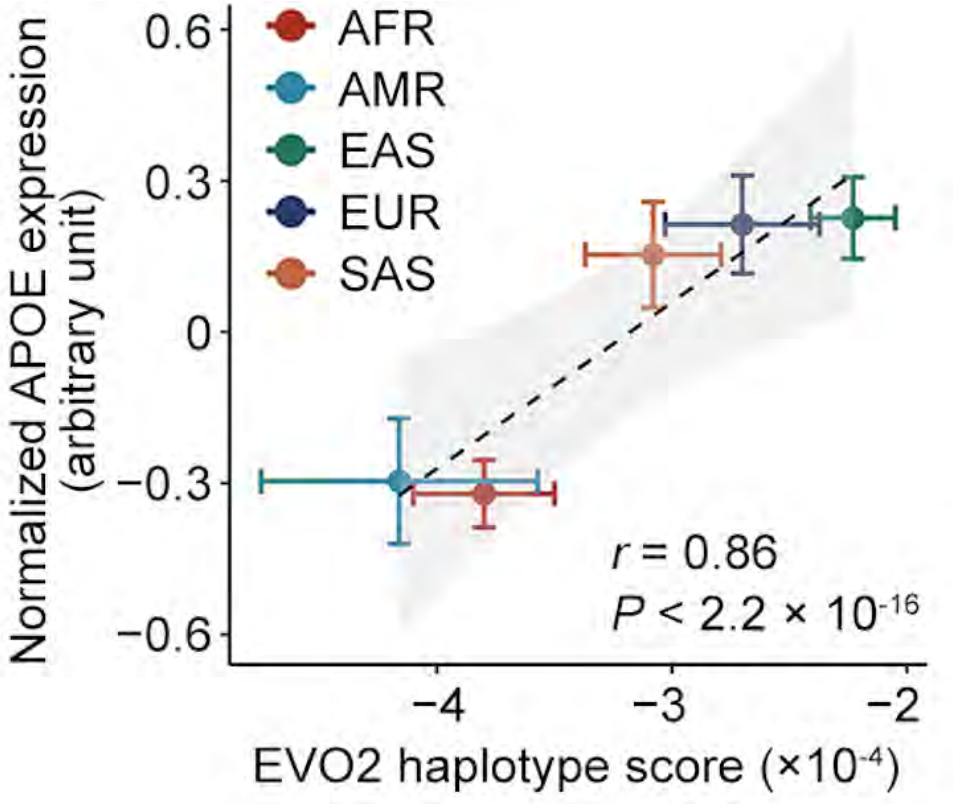
EVO2 Haplotype Scores in the General Population and Across Diverse Ancestry Backgrounds



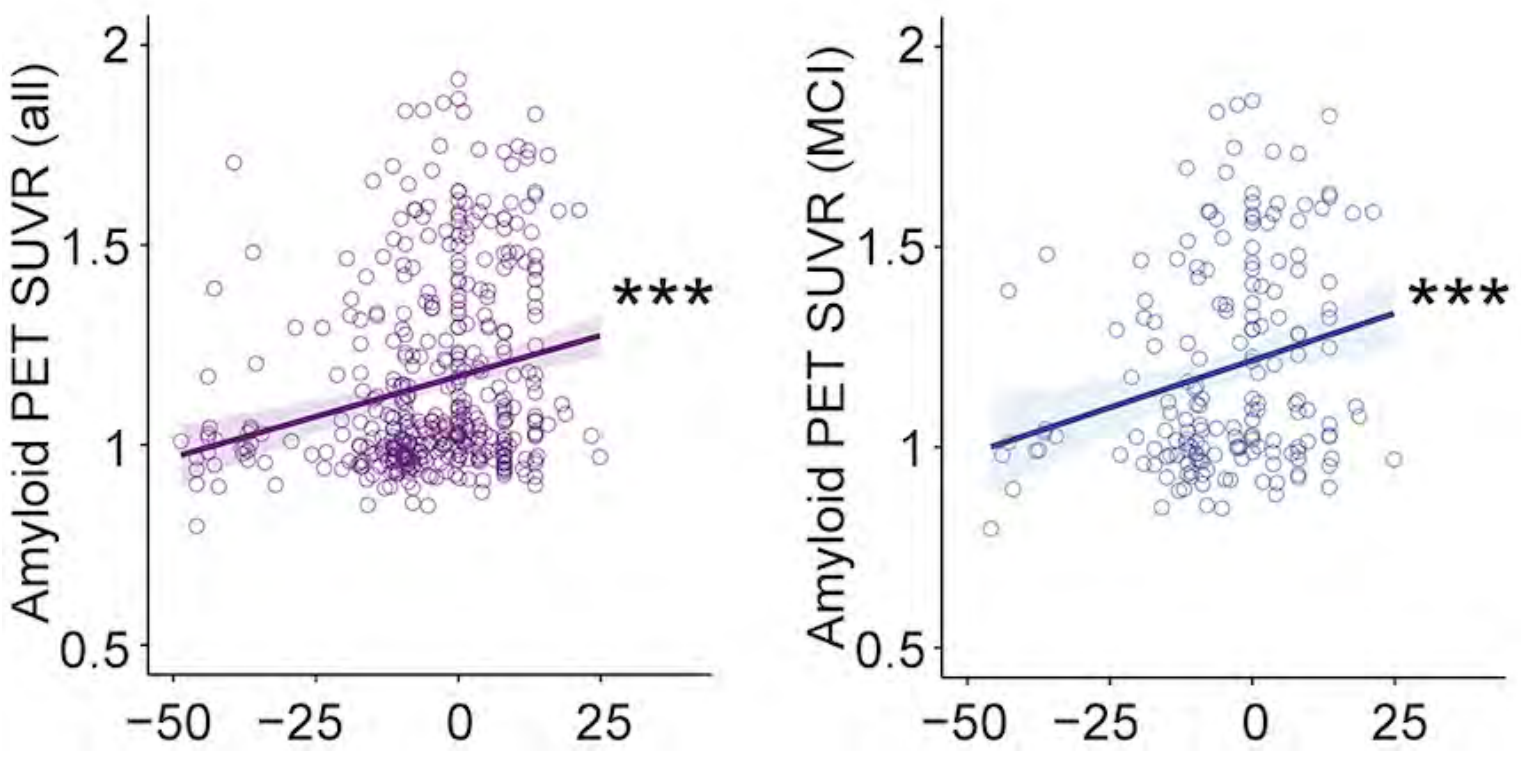


EVO2 Haplotype Scores in the General Population and Across Diverse Ancestry Backgrounds





EVO2 Haplotype Scores Correlate with Key AD Endophenotypes ADNI



Generative AI has great potential in medicine, but some important challenges remain

Privacy risks of Gen AI models have not been fully characterized

Models and data are no longer easy to separate

Natural intelligence still needed to make the best use of artificial intelligence

there is no AI free lunch

Slide by Dr. Gemini

generate a picture of an AI free lunch at a hospital with patient and doctor

Thank you!