

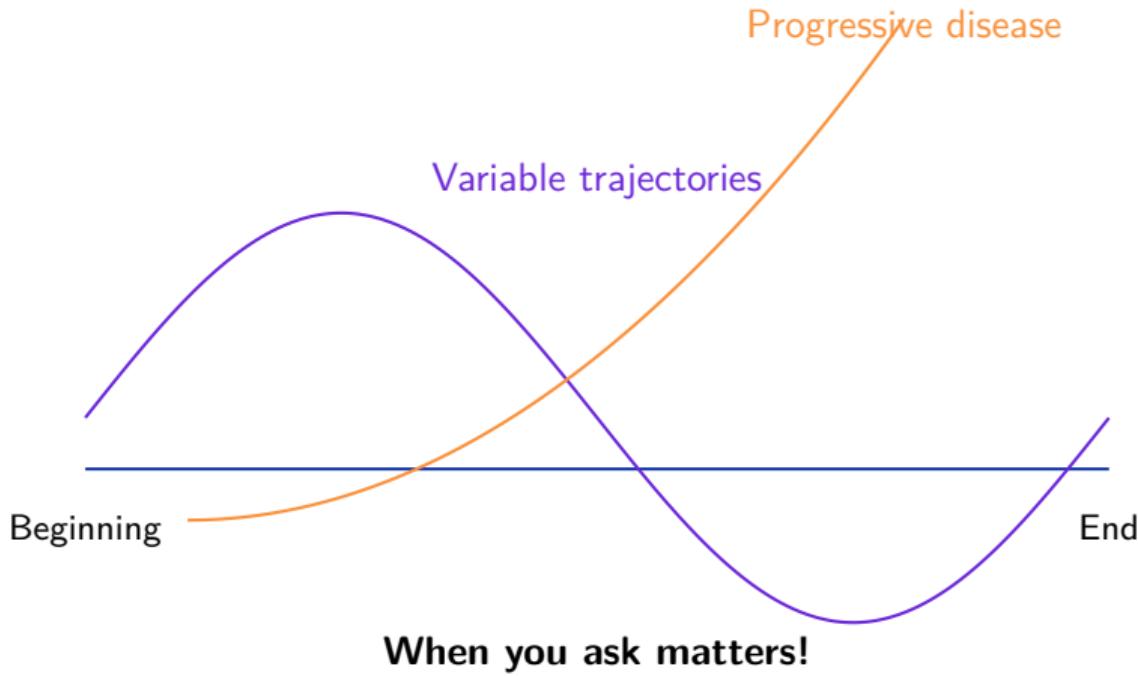
Sarah M. Urbut

Massachusetts General Hospital & Broad Institute

January 28, 2026



*"Every person is a story... and disease trajectories follow narrative arcs"*





## Traditional Approach

- Study one disease at a time
- Simple risk factors (age, sex)
- Static models
- Misses biological complexity

## Reality

- Diseases **co-occur** in patterns
- Shared biological pathways
- Risk evolves over lifetime
- Need to **borrow strength** across diseases

## Example

Metabolic syndrome: Hypertension, Type 2 Diabetes, and Coronary Artery Disease don't appear randomly—they cluster together in time and share genetic basis.

## Cox Proportional Hazards

$$h(t|\mathbf{X}) = h_0(t) \exp(\beta^T \mathbf{X})$$

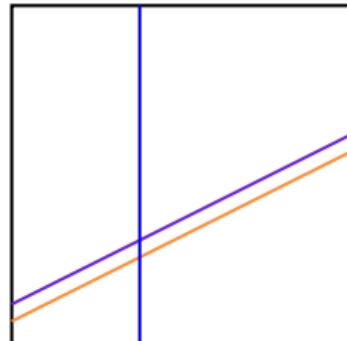
### Assumes:

- Relative risk is **constant over time**
- Hazards remain proportional

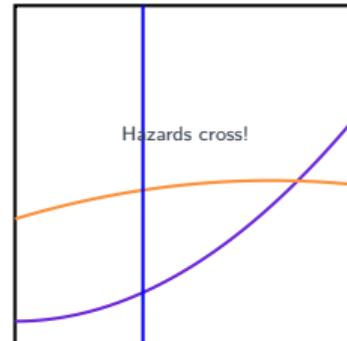
### Reality: Non-Proportional

- Same individual, different ages → different risk
- Risk factors change roles over lifetime
- Early disease ≠ all disease early

Proportional



Non-Proportional



## What We Have

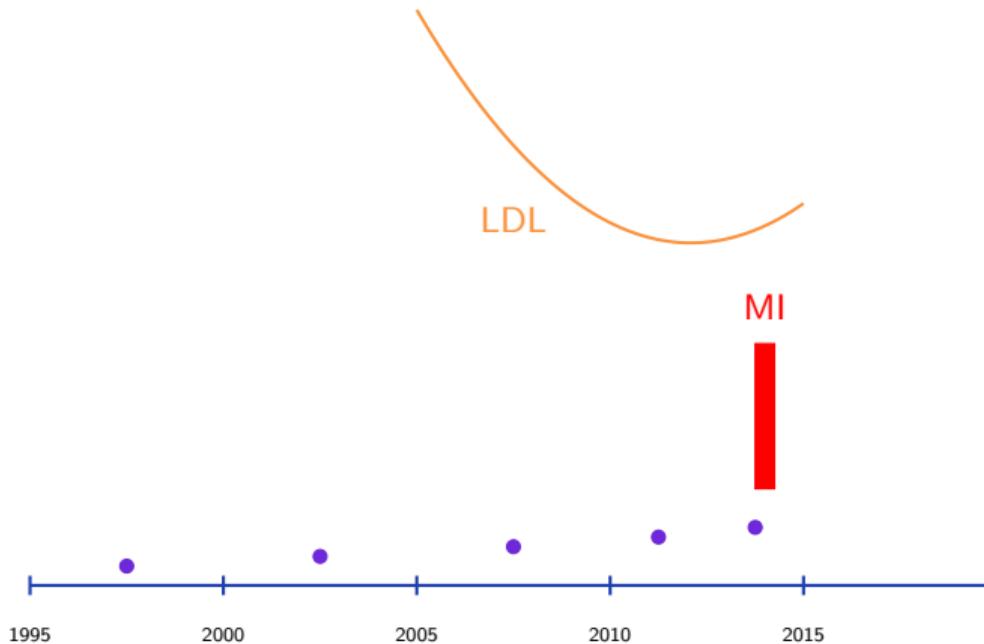
- **Longitudinal EHR:** Diagnoses over time for millions of patients
- **Genetics:** PRS for multiple diseases
- **Complex patterns:** Diseases co-occur, evolve, interact

## What We Need

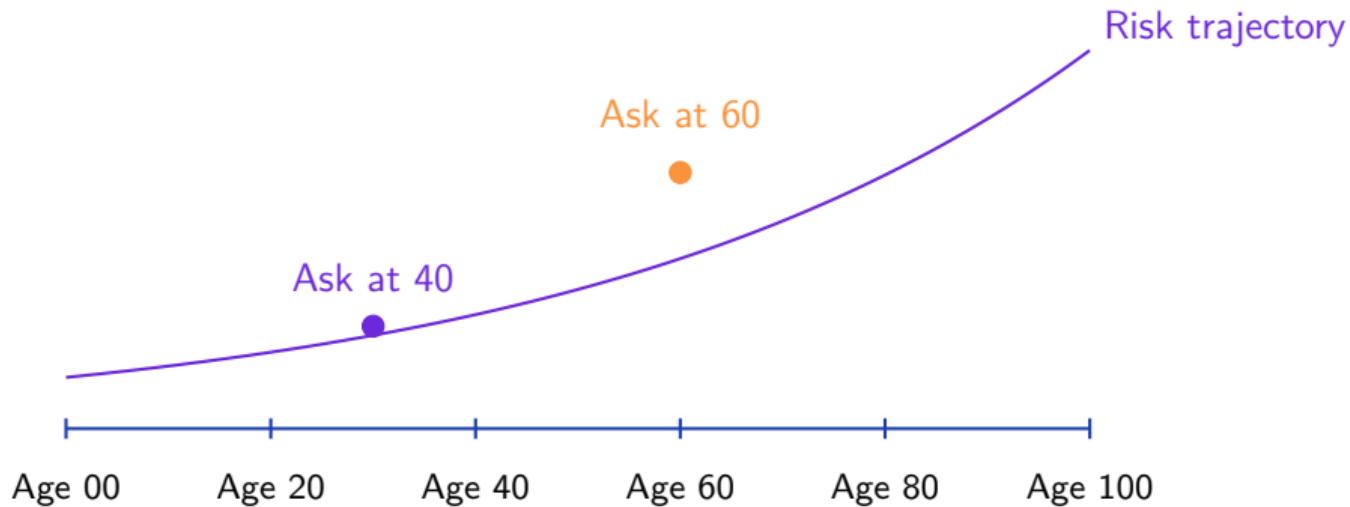
- ① Discover latent patterns of disease co-occurrence
- ② Model time-varying trajectories for individuals
- ③ Integrate genetics to inform individual risk
- ④ Predict future disease while learning biology

## Solution: Bayesian hierarchical model

## Patient Journey: MI at age 57



Can we predict before the event?



**Same individual, different answers**

*The question changes as we move through life*



## Discovery and Prediction together

$$P(\Pi \mid \text{Diagnoses}) \propto P(\text{Diagnoses} \mid \Pi) \cdot p(\Pi)$$

- Continuously updated posteriors
- Individual data likelihood (EHR)
- Population signatures (prior)

## The Bayesian Philosophy

$$\mathbb{P}(\text{Model} \mid \text{Data}) \propto \mathbb{P}(\text{Data} \mid \text{Model}) \cdot \mathbb{P}(\text{Model})$$

Prior	Likelihood	Posterior
<b>Population knowledge</b> Disease signatures from all patients	<b>Individual data</b> This patient's diagnoses over time	<b>Updated beliefs</b> Personalized risk prediction
Patterns	Diagnoses	Updated

Continuous updating as we walk through the lifetime

## The Key Insight

Diseases don't occur independently—they cluster in **signatures** representing shared biological pathways

### Example Signatures

#### Metabolic Signature:

- Type 2 Diabetes
- Hypertension
- Coronary Artery Disease
- Obesity

#### Inflammatory Signature:

- Rheumatoid Arthritis
- Inflammatory Bowel Disease
- Psoriasis

#### Why This Matters

- **Borrow strength:** Learn from similar patients
- **Predict multiple diseases:** Learn signature, predict all
- **Biological interpretation:** Signatures have genetic basis
- **Efficiency:** Don't model each disease separately

## The Core Idea: Mixture Model

For individual  $i$ , disease  $d$ , at time  $t$ :

$$\pi_{i,d,t} = \kappa \cdot \sum_{k=1}^K \theta_{i,k,t} \cdot \text{sigmoid}(\phi_{k,d,t})$$

### Breaking It Down

- $\theta_{i,k,t}$ : How much does individual  $i$  load on signature  $k$  at time  $t$ ? (**Individual-specific**)
- $\phi_{k,d,t}$ : How strongly is disease  $d$  associated with signature  $k$  at time  $t$ ? (**Population-level**)
- Multiple signatures can contribute

### Key Innovation

This is a **mixture of probabilities**, not probability of mixture

#### Why this matters:

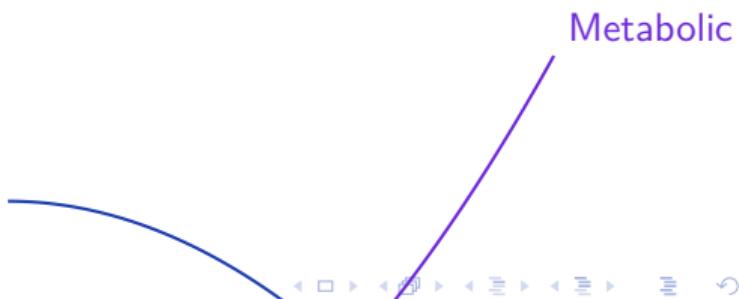
- Enables direct risk prediction
- Unlike topic models (allocation-based)

## Individual Signature Loadings Change Over Time

Each individual's association with signatures evolves:

$$\theta_{i,k,t} = \frac{\exp(\lambda_{i,k,t})}{\sum_{k'=1}^K \exp(\lambda_{i,k',t})}$$

- Softmax ensures  $\sum_k \theta_{i,k,t} = 1$  (proper mixture)
- $\lambda_{i,k,t}$  evolves smoothly over time



## Why Time-Varying Matters

- Young person: Low metabolic signature loading

We Model  $\lambda_{i,k,t}$  as a Gaussian Process

$$\lambda_{i,k} \sim \mathcal{GP}(\text{mean function}, K_\lambda)$$

Why GP?

- Ensures **smooth trajectories** over time (no wild jumps)
- Flexible—can capture any smooth function
- Handles irregular observation times
- Natural for longitudinal data

But what's the mean function?

This is where **genetics comes in!**

## Genetics Modify the Mean Trajectory

$$\lambda_{i,k} \sim \mathcal{GP}(r_k + \mathbf{g}_i^\top \boldsymbol{\Gamma}_k, K_\lambda)$$

### Components

- $r_k$ : Population baseline for signature  $k$
- $\mathbf{g}_i$ : Individual's PRS vector
- $\boldsymbol{\Gamma}_k$ : **How genetics affect signature  $k$  (learned!)**
- $K_\lambda$ : Temporal smoothness

### Why This Matters

#### High CAD PRS?

- $\boldsymbol{\Gamma}_{\text{metabolic}}$  tells us how this shifts your metabolic signature trajectory
- Higher genetic loading  $\rightarrow$  earlier/stronger signature activation
- This provides **biological interpretation**

High PRS

## Signature-Disease Associations Are Shared

$$\phi_{k,d} \sim \mathcal{GP}(\mu_d + \psi_{k,d}, K_\phi)$$

- $\mu_d$ : Disease baseline (population prevalence)
- $\psi_{k,d}$ : How strongly signature  $k$  associates with disease  $d$
- **Learned from all patients** (not individual-specific)
- Time-varying: associations can change over life course

### Example

Metabolic signature strongly associated with:

- T2D:  $\psi_{\text{metabolic,T2D}} = +2.3$
- CAD:  $\psi_{\text{metabolic,CAD}} = +1.8$
- Cancer:  $\psi_{\text{metabolic,Cancer}} = +0.1$

### Why This Works

- **Borrow strength**: Learn patterns from millions
- **Interpretability**: Signatures have clear meaning
- **Efficiency**: Shared parameters across all individuals

## The Likelihood: Discrete-Time Survival

For individual  $i$ , disease  $d$ :

$$\ell_{i,d} = \sum_{t < E_{i,d}} \log(1 - \pi_{i,d,t}) + Y_{i,d,t} \log(\pi_{i,d,t}) + (1 - Y_{i,d,t}) \log(1 - \pi_{i,d,t})$$

### At Risk

Before event time  
Person hasn't gotten disease yet  
 $(1 - \pi)$  terms

### Event

Disease occurs at time  $t$   
 $\log(\pi)$  contribution

### Censored

Person leaves study  
Last observation at  $E_{i,d}$   
 $(1 - \pi)$  at enrollment

## Why This Matters

**Properly handles censoring** (critical for EHR data!)  
and enables **direct prediction** (unlike topic models)

## Individual Disease Probability

$$\pi_{i,d,t} = \kappa \cdot \sum_{k=1}^K \theta_{i,k,t} \cdot \text{sigmoid}(\phi_{k,d,t})$$

Individual Component (Time-Varying)

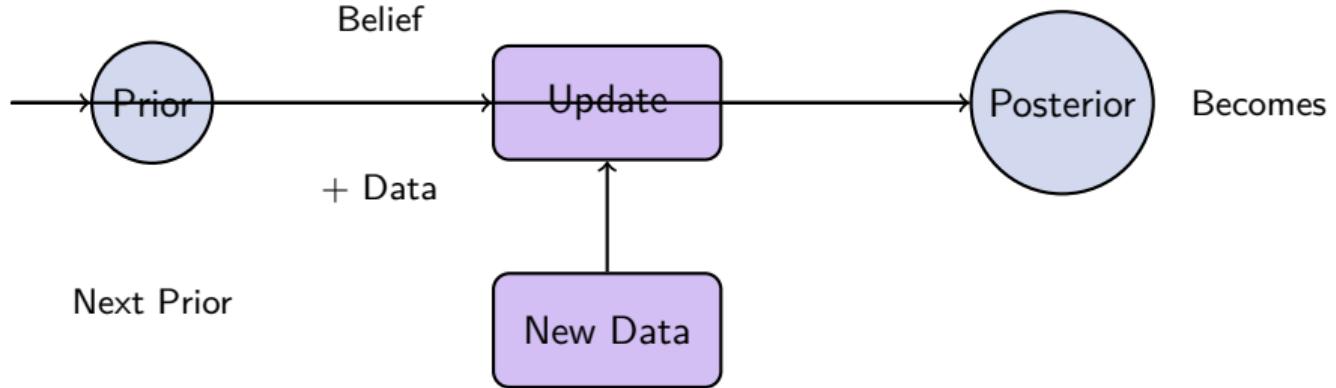
$$\begin{aligned}\theta_{i,k,t} &= \text{softmax}(\lambda_{i,k,t}) \\ \lambda_{i,k} &\sim \mathcal{GP}(r_k + \mathbf{g}_i^\top \boldsymbol{\Gamma}_k, K_\lambda)\end{aligned}$$

Population Component (Shared)

$$\phi_{k,d} \sim \mathcal{GP}(\mu_d + \psi_{k,d}, K_\phi)$$

Genetics + Population Patterns + Time = Personalized Risk





## Key Insight

At each time point  $t$ , we observe new diagnoses and update our beliefs about future risk:

$$\mathbb{P}(\pi_{i,d,t} \mid \text{Data up to } t) \propto \mathbb{P}(\text{Data up to } t \mid \pi_{i,d,t}) \cdot \mathbb{P}(\pi_{i,d,t})$$

This is what happens naturally—we continuously learn from data

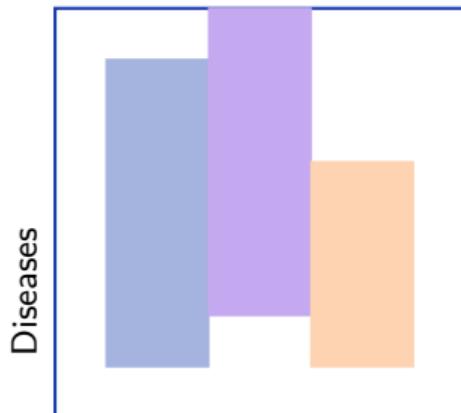


## Learned Signatures

- Latent patterns of disease co-occurrence
- Each signature has characteristic timing
- Some signatures activate early, others later in life
- Genetics inform individual predisposition

## Example Patterns

- **Metabolic** → early adulthood onset
- **Inflammatory** → mid-life activation
- **Cancer** → age-dependent patterns



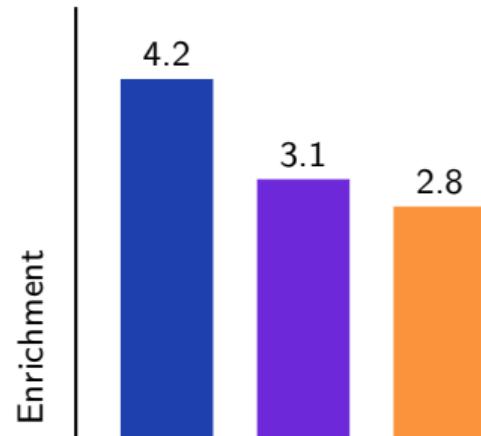
Signatures

Signature-Disease Matrix

## Key Findings

- Cardiovascular PRS → strongly associated with metabolic signature
- Psychiatric PRS → associated with neuro/psychiatric signatures
- Genetic correlations reveal shared architecture

PRS-Signature Associations



## ASCVD Predictions

- **10-year:** ALADYNOLLI 0.737 vs PCE 0.683  
*+7.9% improvement*
- **30-year:** ALADYNOLLI 0.708 vs PREVENT 0.650  
*+9.0% improvement*

## vs Cox Baseline

- Parkinson's: *+35.2%*
- CKD: *+33.2%*
- Stroke: *+31.6%*
- ASCVD: *+16.3%*

## vs Delphi-2M

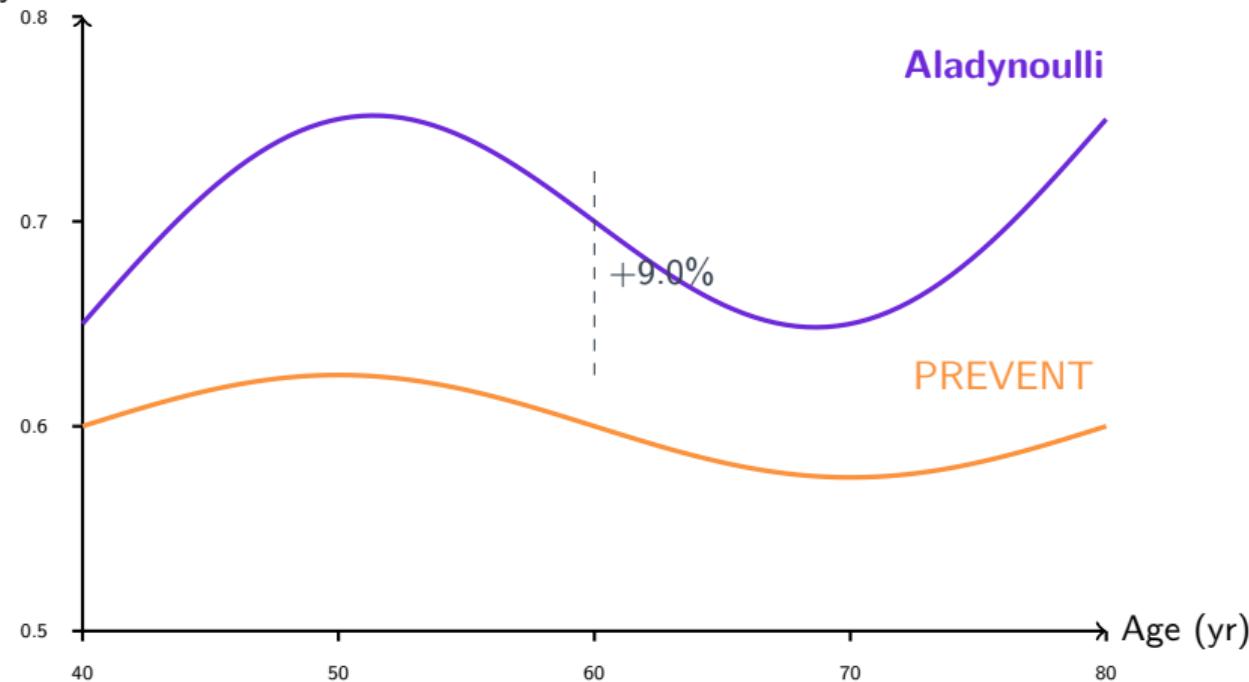
- Outperforms on 15/28 diseases
- Largest gains: Parkinson's (+35%), CKD (+33%)

## Dynamic Risk Assessment

Model updates predictions as patients age and develop conditions

## 30-year ASCVD Risk Prediction

Dynamic AUC

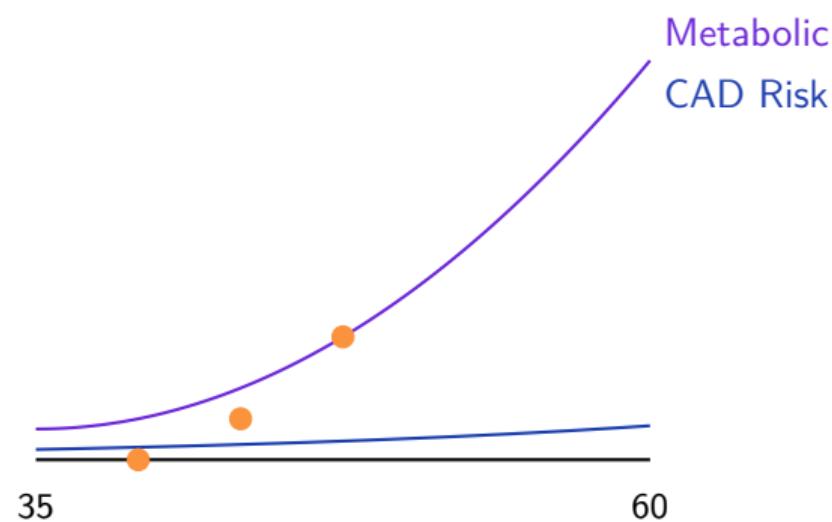


## Patient Timeline

- Age 35: Enrollment
- Age 40: Hypertension
- Age 45: Type 2 Diabetes
- Age 50: Coronary Artery Disease
- Age 55: Ongoing monitoring

## Signature Evolution

- Metabolic signature loading increases over time
- Genetic predisposition (PRS) influences trajectory
- Risk predictions continuously updated



Posterior beliefs evolve as new information arrives



### ✓ Unified Framework

Simultaneous genomic discovery and clinical prediction

### ↗ Dynamic

Properly models lifetime risk evolution with Bayesian updating

### ⊗ Interpretability

Signatures provide biological meaning through genetic architecture

### ◎ Scalable

Works across diseases and biobanks

**Bayesian framework enables both discovery and prediction**

## High Performance

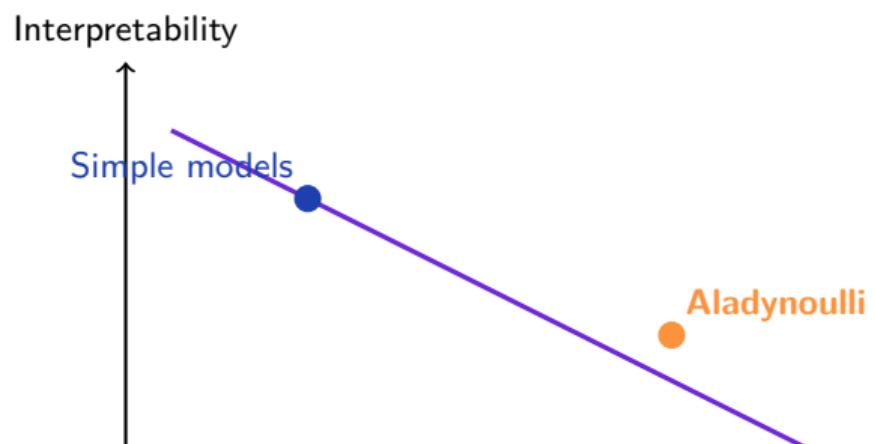
- Highly precise predictions
- Complex hierarchical structure
- Multiple interacting components

## High Interpretability

- Simple, explainable models
- Easy clinical communication
- Transparent decision-making

**Challenge:** Difficult to communicate

**Challenge:** Lower predictive power



## Integration

- Imaging (CAC, CT-coronary)
- AI-enhanced features (ECG, TTE)
- Multi-omics data
- Genomics + environment

## Applications

- Intervention modeling (digital twins)
- Personalized screening
- Drug repurposing
- Therapeutic targeting

## Vision

A model that combines genomics, opportunistic imaging, and AI-processed signals for comprehensive risk assessment



# Thank You!

- surbut@mgh.harvard.edu
  - aladynoulli.hms.harvard.edu

*Collaborators: P. Natarajan, G. Parmigiani, A. Gusev, and team*