2026 Translational Bioinformatics Year-in-Review

@proftatonetti @tatonetti.bsky.social

Nicholas Tatonetti, PhD, FACMI
Professor of Computational Biomedicine
Cedars-Sinai Medical Center

2026 AMIA Amplify - Denver

Disclosures

- NIH, FDA, DoD, Pfizer, AstraZeneca, Janssen, Amgen, PhARMA Foundation, CARI Health, WorldQuant Foundry
- Editor-in-Chief of BioData Mining
- I am influenced by my professional and personal network and experiences
- Biggest conflict: I am a geek for translational bioinformatics, methods that solve problems, computational medicine 😍

Goals

- Review trends in the translational bioinformatics literature
- Create a “snapshot” of what the field is doing now (Spring 2026)
- Recognize innovative work and identify opportunities for the future

Process

- Follow the literature throughout the year (i.e. my lab's #papers channel)
- Work with the talented and generous **AMIA Year-in-Review Committee**
- Triage all papers from a set of relevant journals since Jan 2025
  - Evaluate papers on a set of TBI criteria, score on:
    - Informatics Novelty, Application Importance, **Wow** Factor (total 0-9)
- I then take these scores and select papers to highlight in 1-5 slides

Caveats

- Translational bioinformatics =

  Informatics methods that link
  biological entities (genes, proteins, cells, small molecules)
  to **clinical entities** (drugs, diseases, symptoms, etc.)
  — or vice versa.

- Covers about 14 months (Jan 2025 - Mar 2026)

- Focused on human biology

- **What's NOT included:**

  - Amazing biology with straightforward informatics

  - Amazing informatics but no link between the clinical and the molecular

  - Perspectives, reviews (for the most part)

This is all thanks to…

The 2026 AMIA Year-in-Review Team!

Thank you so much for the incredible work towards this year's TBI YIR 2026!

Dr. Nicholas Tatonetti
Cedars-Sinai Medical Center, LA

Biniam Garomsa
Emory University, GA

Akinchan Bharadwaj
Icahn School of Medicine at Mount Sinai, NY

Farhan Quadir
University of Chicago, IL

Harshal Singh Chauhan
University of Maryland, Baltimore County, MD

Kexin Sun
University of Pittsburgh, PA

Ya Lin Chen
University of Washington, WA

Pratishtha Guckhool
Drexel University, PA

Nick Reid
University of Washington Seattle, WA

Ishita Vasudev
Icahn School of Medicine at Mount Sinai, NY

Syllas Otutey
Michigan Technological University, MI

Daelyn Richards
University of Colorado, CO

Hadi Ul Bashar
Marshall University, WV

Nora J. Gilliam
University of Rochester, NY

Akeemat Ayinla
UTHealth Houston, TX

Pushkala Jayaraman
Icahn School of Medicine at Mount Sinai, NY

Quan Minh Nguyen
University of Pennsylvania, PA

Toshika Talele
Arizona State University, AZ

Roopa Santhoshi Gatta
Prime Healthcare, CA

Renee Motew
University of Florida, FL

Rishik Kondadadi
University of Michigan, MI

Mu Yang
University of Washington, WA

Xiaoyu Sun
Yale University, CT

Final List

- 1,988 papers triaged!; 611 reviewed and scored by the committee
  - Over double the number of articles screened!
  - Over double the number of articles reviewed! Wow!
- 20 presented here + 7 shout outs + 2 pieces of brain candy
  - Apologies for those I missed, misunderstood, or misjudged, biases/mistakes are all mine
- 4 TBI topics:
  - **Reading Biology in Context**: *spatial, single-cell tumor ecology*
  - **Learning from Molecular Structures:** *using high-D data to classify, diagnose, and prognosticate*
  - **AI-Guided Therapeutic Discovery:** *nominate drugs, targets, ligands, and more*
  - **New Engines for Molecular Reasoning:** *agents, general purpose models, PLMs*
- All authors are mentioned if ≤3, all first authors otherwise
- Slides will be posted to www.tatonettlab.org and linked to my Bluesky and other social media accounts

Here we go…

Maps - *Yeah Yeah Yeahs*

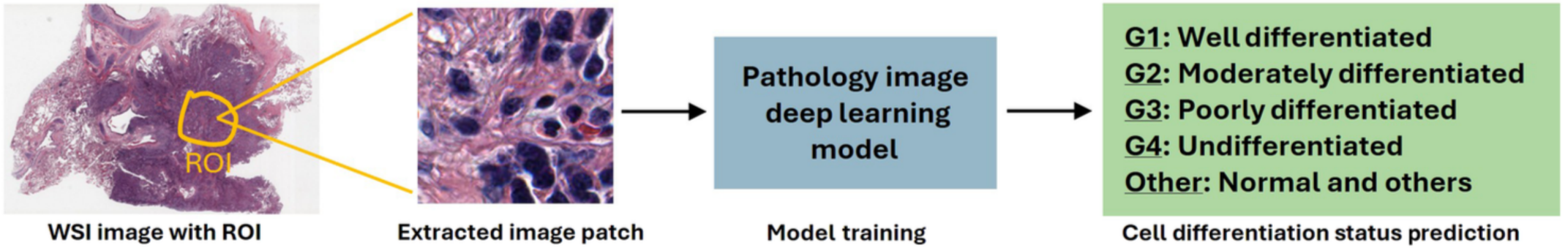
Reading Biology in Context

spatial, single-cell, tumor ecology, and disease progression as trajectories or populations

Image-based inference of tumor cell trajectories enables large-scale cancer progression analysis (Liu, Cai, Rong et al, *Science Advances*)

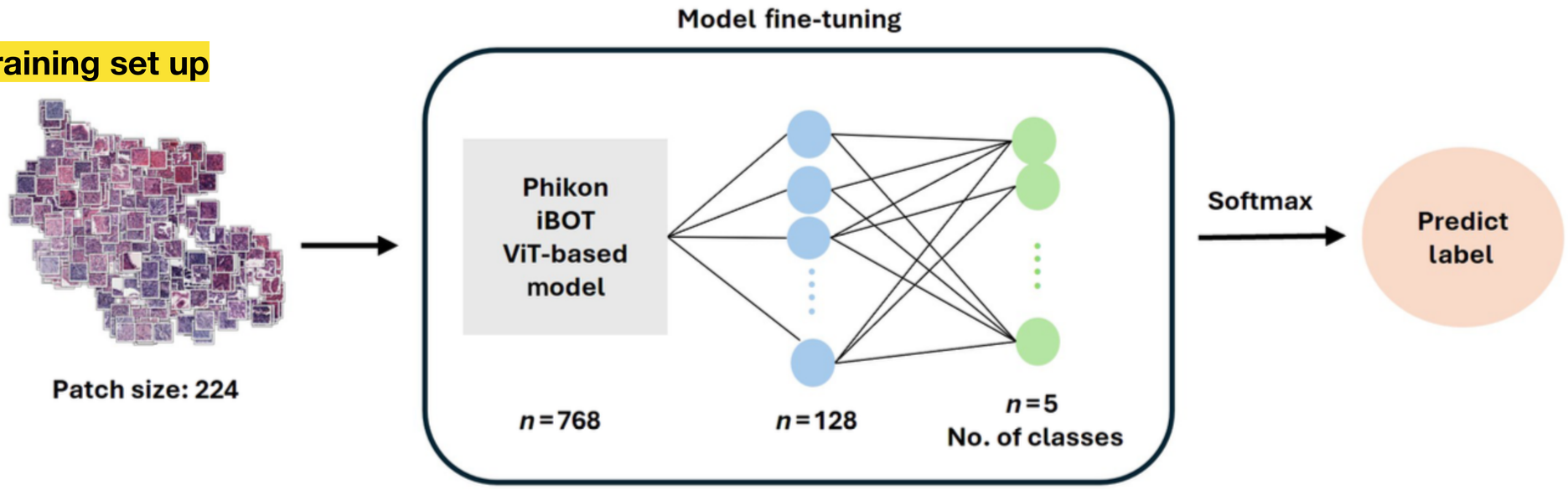
- **Goal**: Infer tumor progression dynamics from routine H&E pathology instead of relying on expensive single-cell or spatial molecular assays
- **Method**: Fine-tune a histopathology foundation model to classify lung adenocarcinoma differentiation status, extract image features, infer image-based pseudotime, and combine pseudotime + spatial organization + heterogeneity into a **tumor progression score**
- **Result**: Image-derived progression metrics stratified survival across three lung adenocarcinoma cohorts
- **Conclusion**: Routine pathology slides may contain enough morphology and spatial information to approximate tumor progression trajectories at scale — what's NOT in an H&E slide?

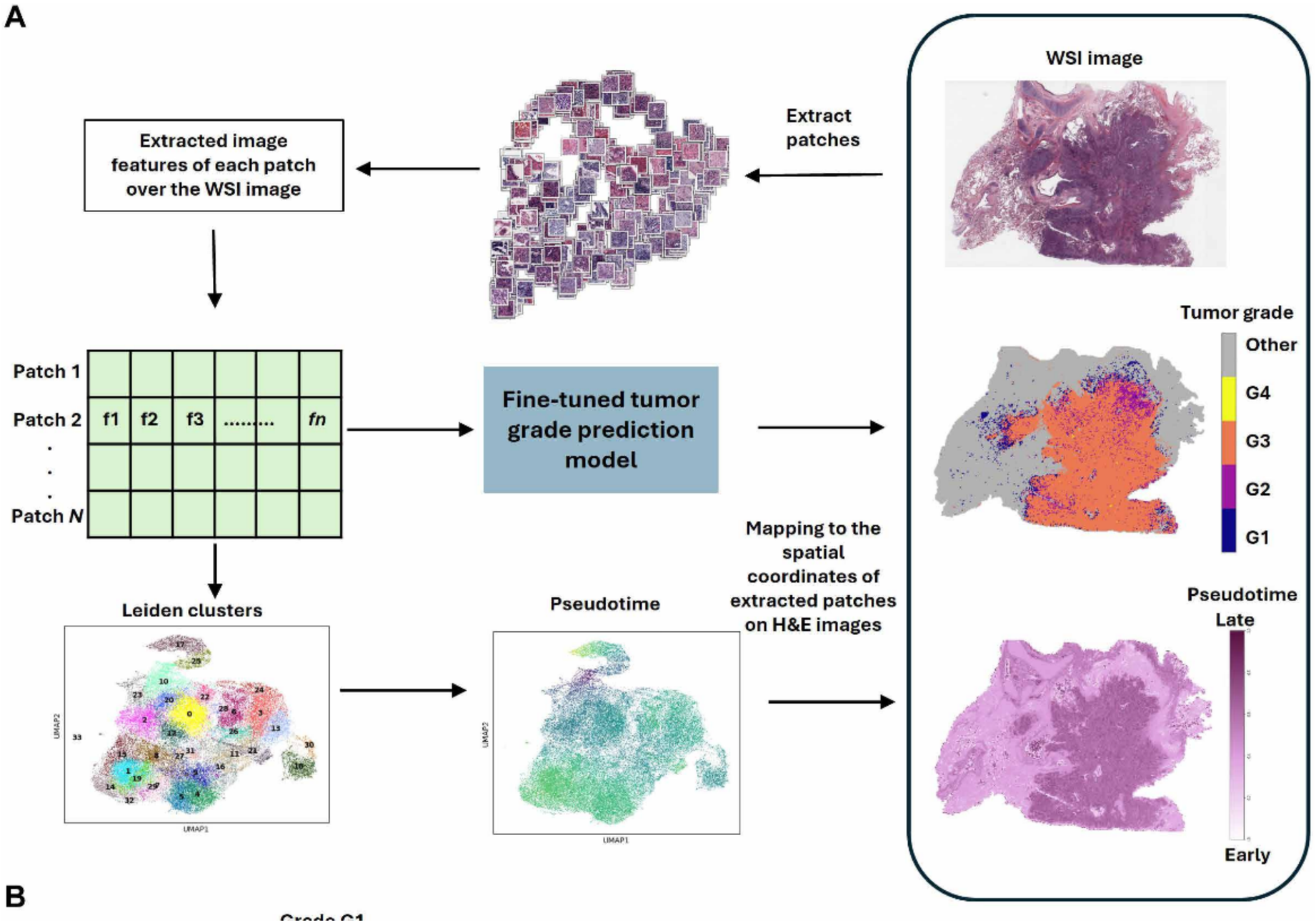
A **Format of available data**



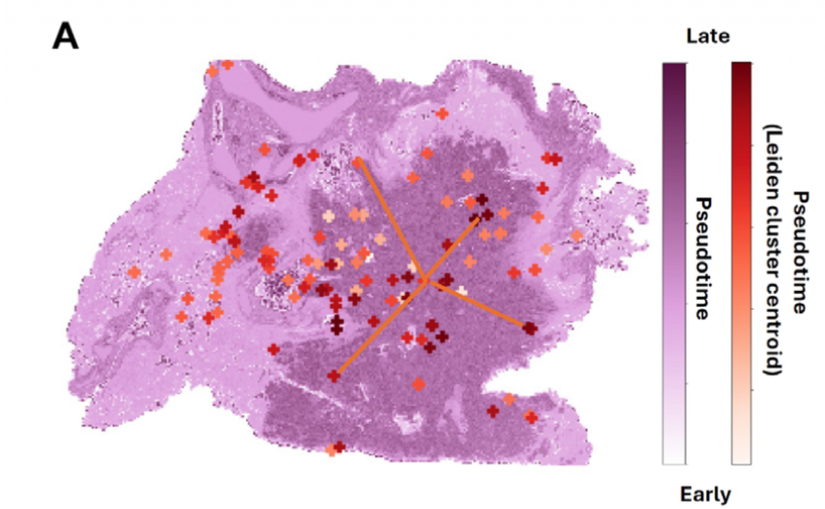
B

Model training set up

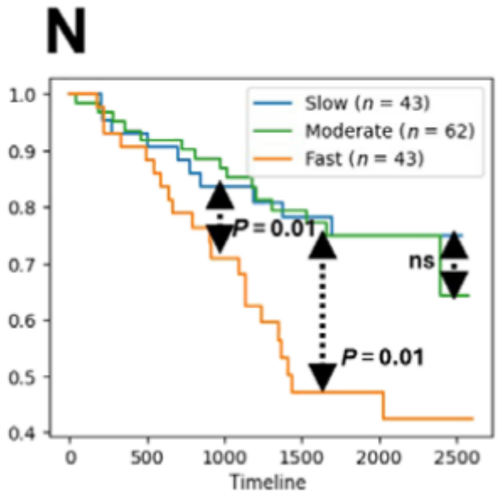




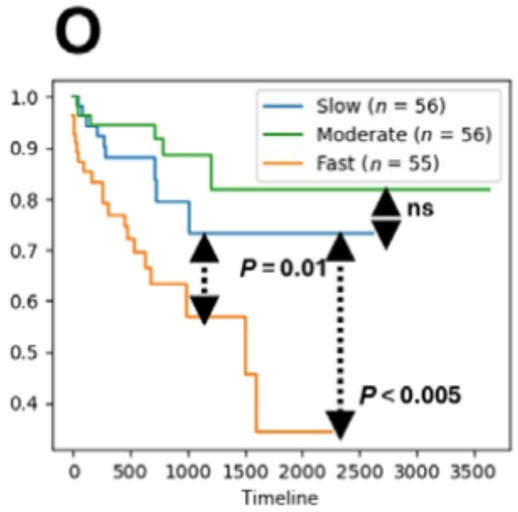
Applying the model allows you to overlay critical metadata like psuedotime



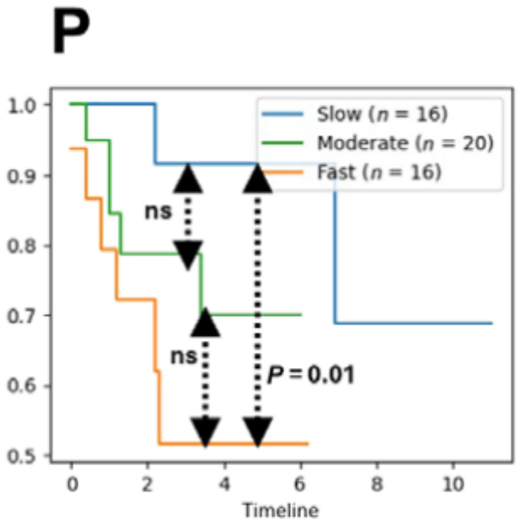
Predicted labels separate populations on a survival curve



NLST dataset



LUAD dataset



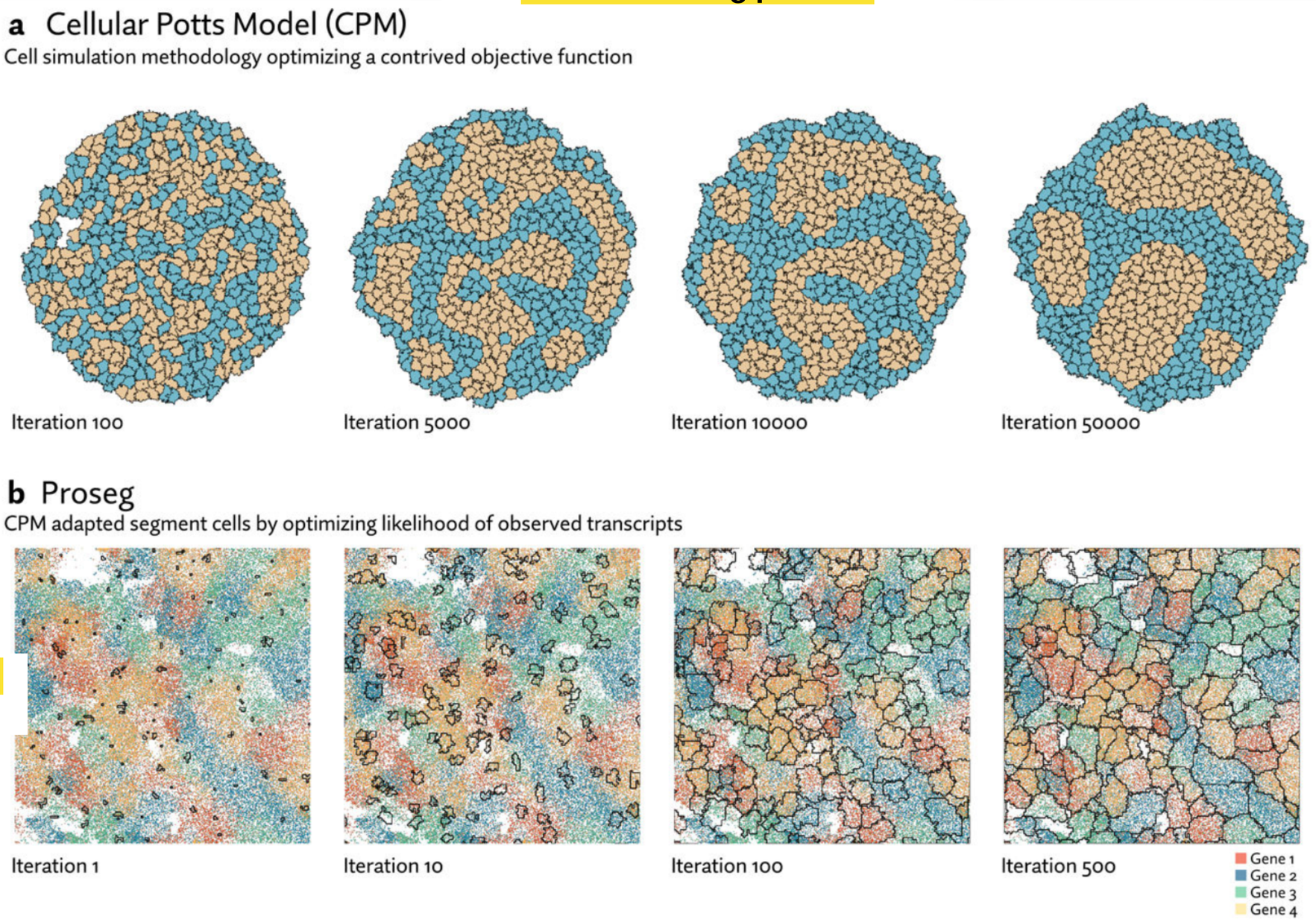
SPORE dataset

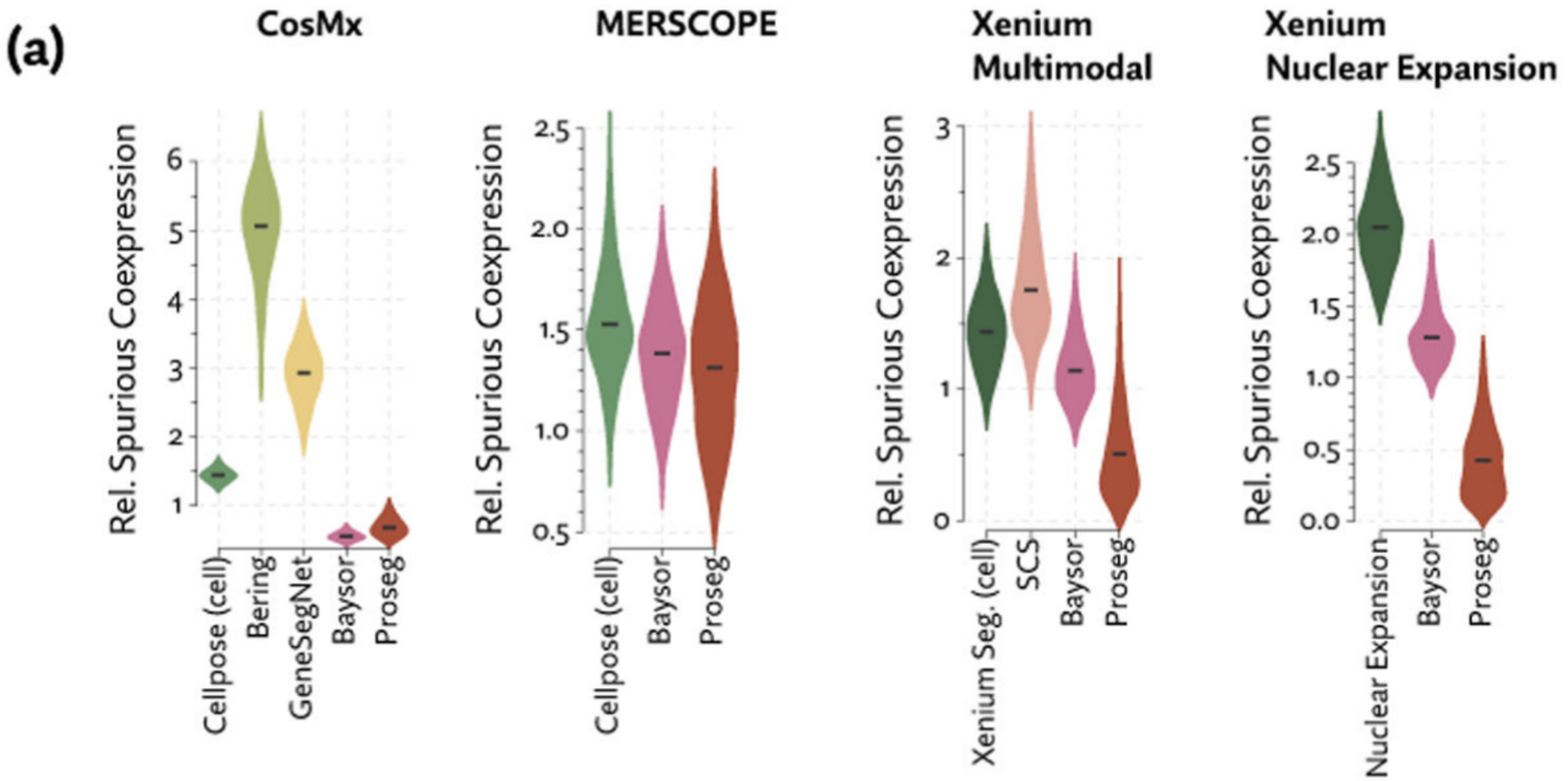
Cell Simulation as Cell Segmentation (Jones, Elz, Hadadianpour et al, *Nature Methods*)

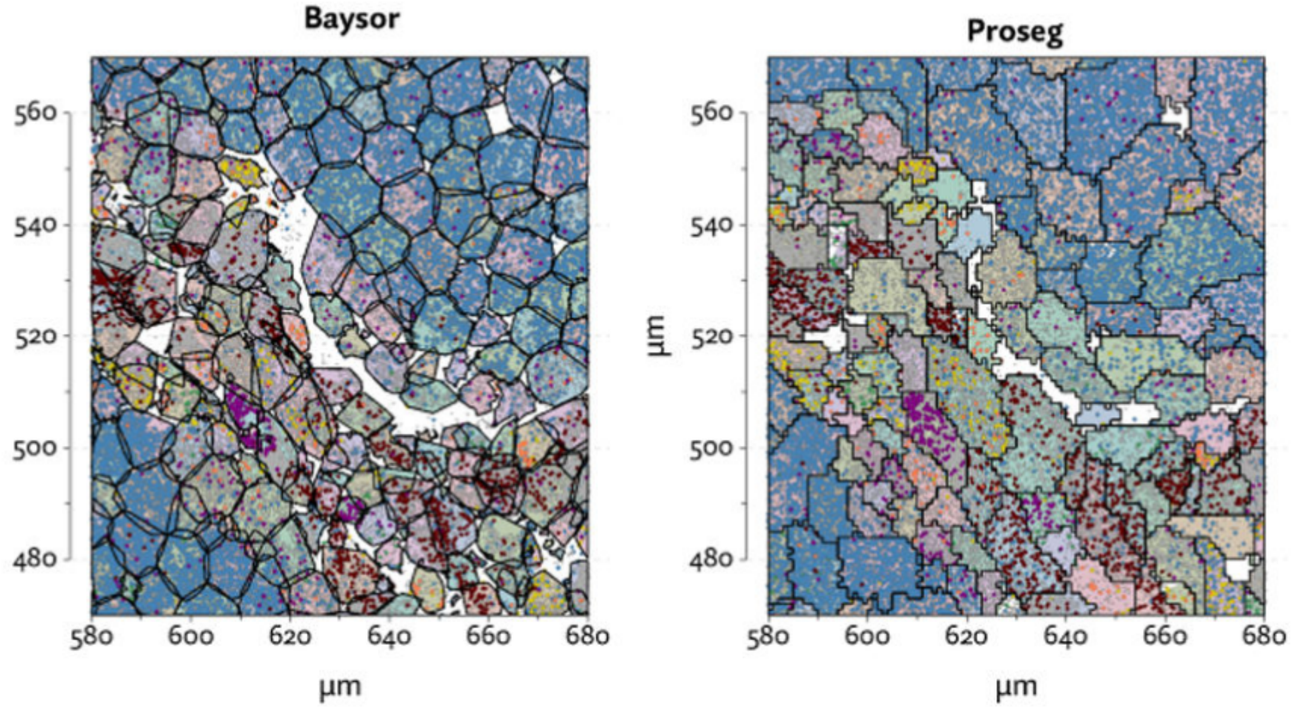
- **Goal**: segmentations errors in single-cell transcriptions can lead to massive misinterpretations of the data; fix this
- **Method**: Introduce Proseg, an unsupervised probabilistic segmentation method inspired by Cellular Potts models; initialize from nuclei, then simulate plausible cell boundaries that best explain observed transcript locations and expression patterns
- **Result**: Across MERSCOPE, CosMx, and Xenium datasets, Proseg reduced spurious co-expression, assigned more transcripts than image-based methods, avoided Baysor-like over-segmentation, and improved detection of tumor-infiltrating neutrophils and T cells
- **Conclusion**: Segmentation is not preprocessing trivia- it can determine the biology you think you discovered

Cellular Potts is a classic cell/ tissue modeling problem

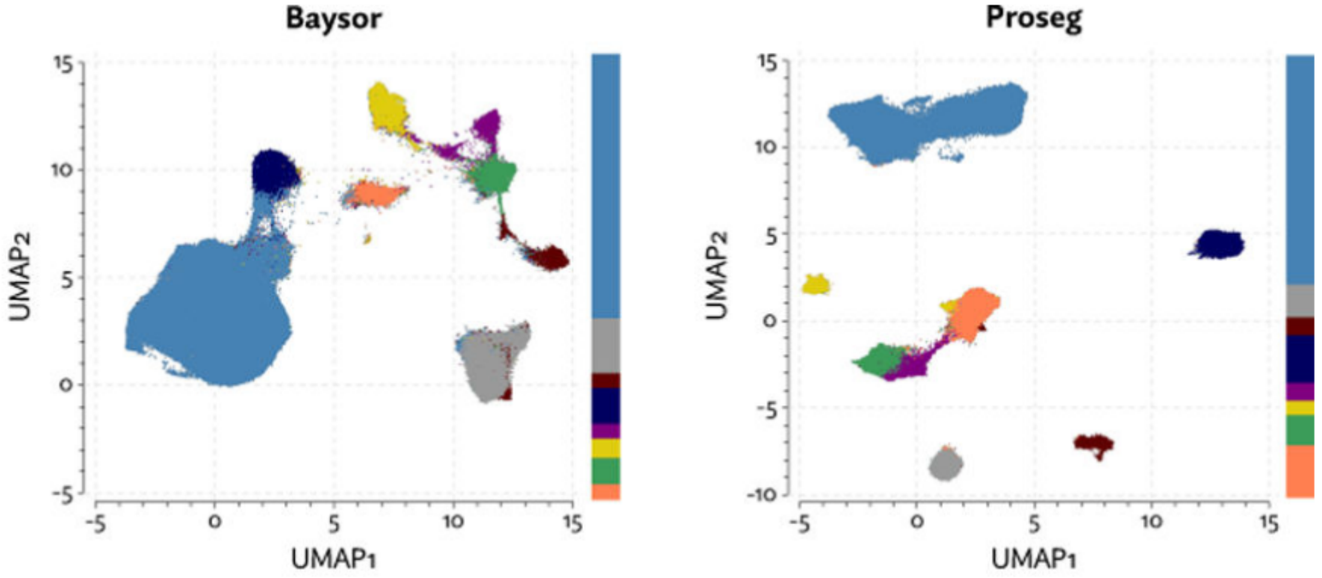
Proseg adapts Cellular Potts to make better boundaries




(a)
CosMx
MERSCOPE
Xenium Multimodal
Xenium Nuclear Expansion
Rel. Spurious Coexpression
Cellpose (cell)
Bering
GeneSegNet
Baysor
Proseg
SCS
Xenium Seg. (cell)
Nuclear Expansion

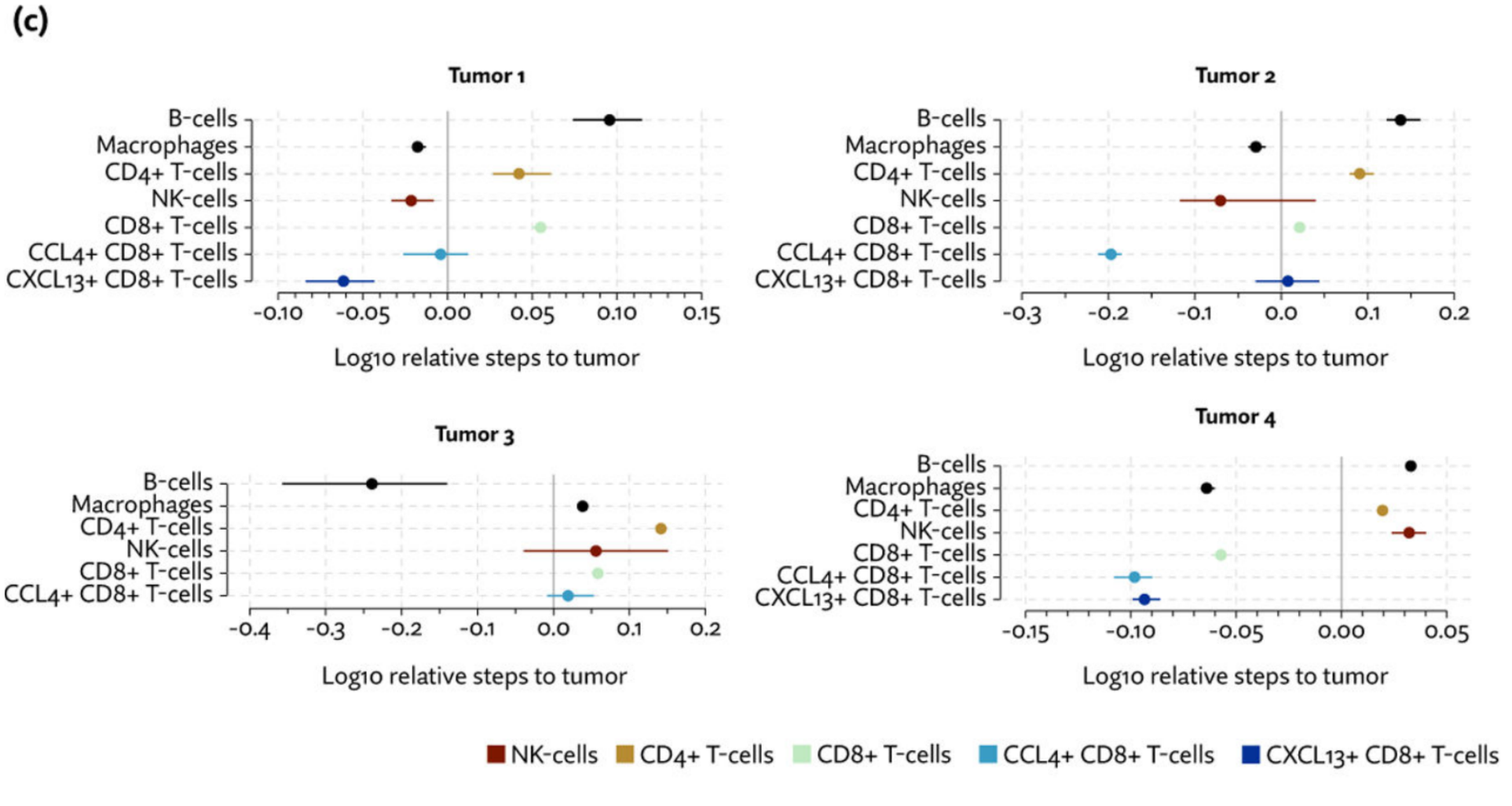


Produces pretty different cell boundaries



But they get much better separation

Use it to discovery some new stuff about T-cell infiltration



Single-cell polygenic risk scores dissect cellular and molecular heterogeneity of complex human diseases (Zhang, Shu, Zhou, Rubin-Sigler et al., *Nature Biotechnology*)

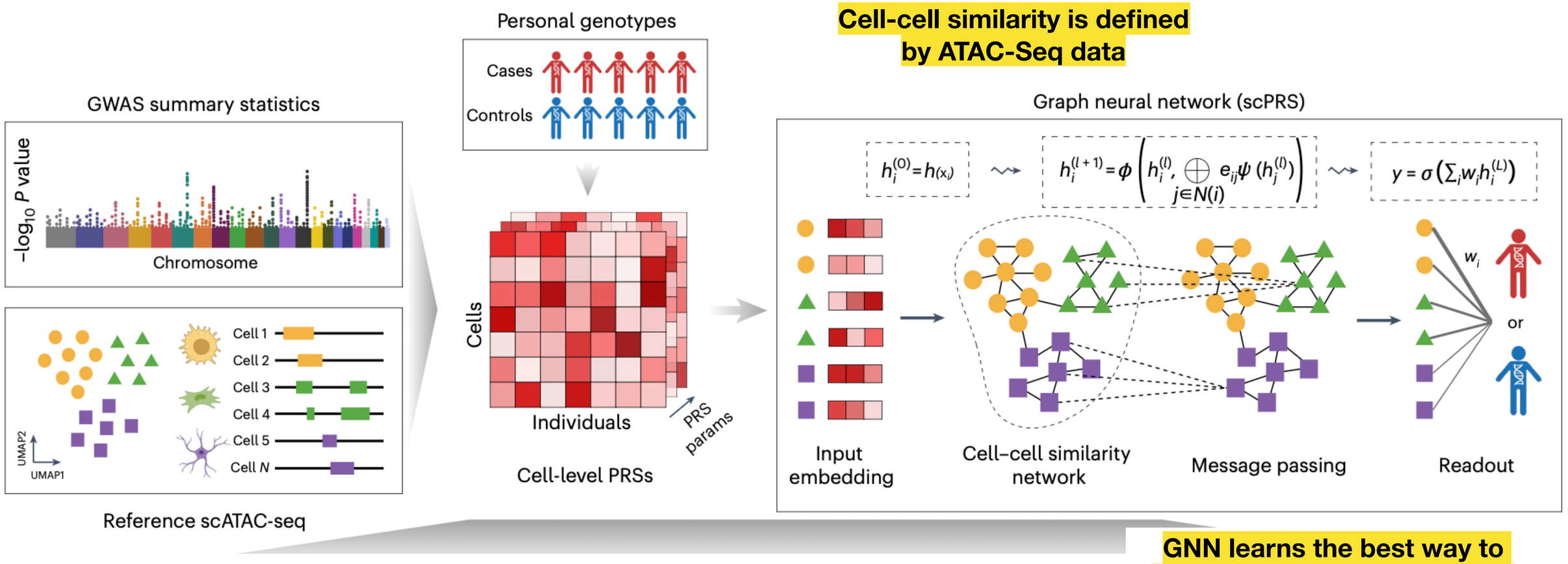
- **Goal**: Make polygenic risk scores more biologically interpretable by resolving genetic risk into disease-relevant cells, variants, and regulatory programs

- **Method**: Compute cell-level PRSs using disease GWAS + reference scATAC/snATAC-seq, then use a GNN over a cell–cell similarity graph to denoise, aggregate, and predict disease risk

- **Result**: scPRS improved risk prediction across HCM, AD, severe COVID-19, and T2D, while prioritizing known and novel disease-relevant cell populations

- **Conclusion**: PRS sweeps a lot under the rug, this could be a way to sift through that dirt and find some disease mechanisms

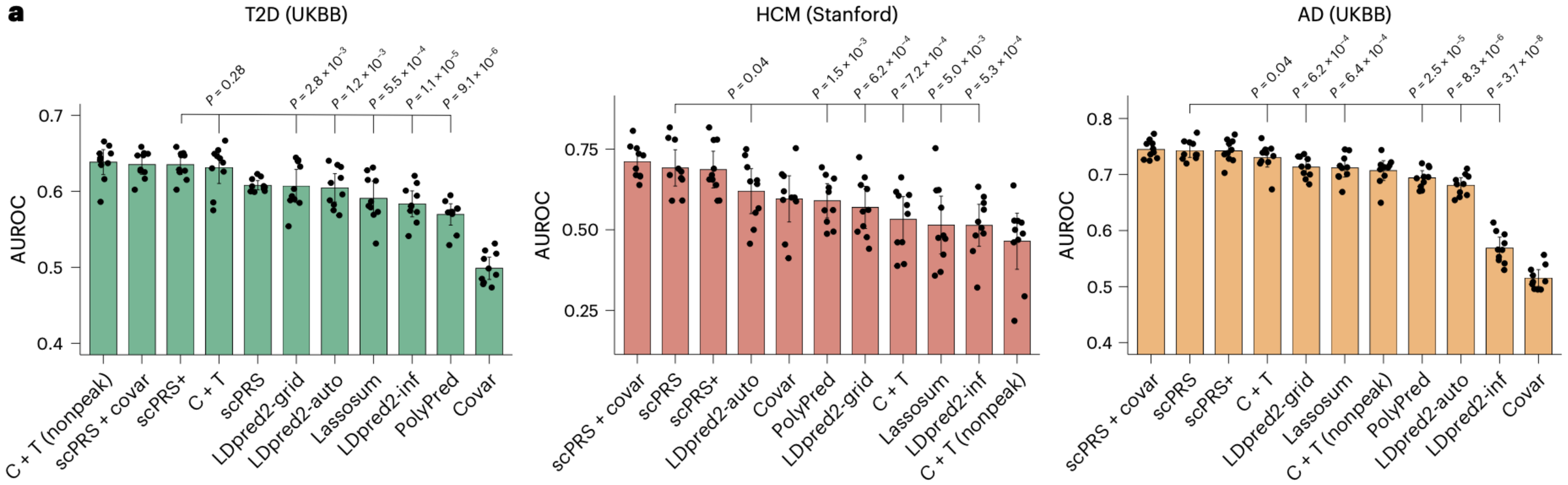
Use the disease GWAS and single-cell ATAC-seq to train a GNN



Cell-cell similarity is defined by ATAC-Seq data

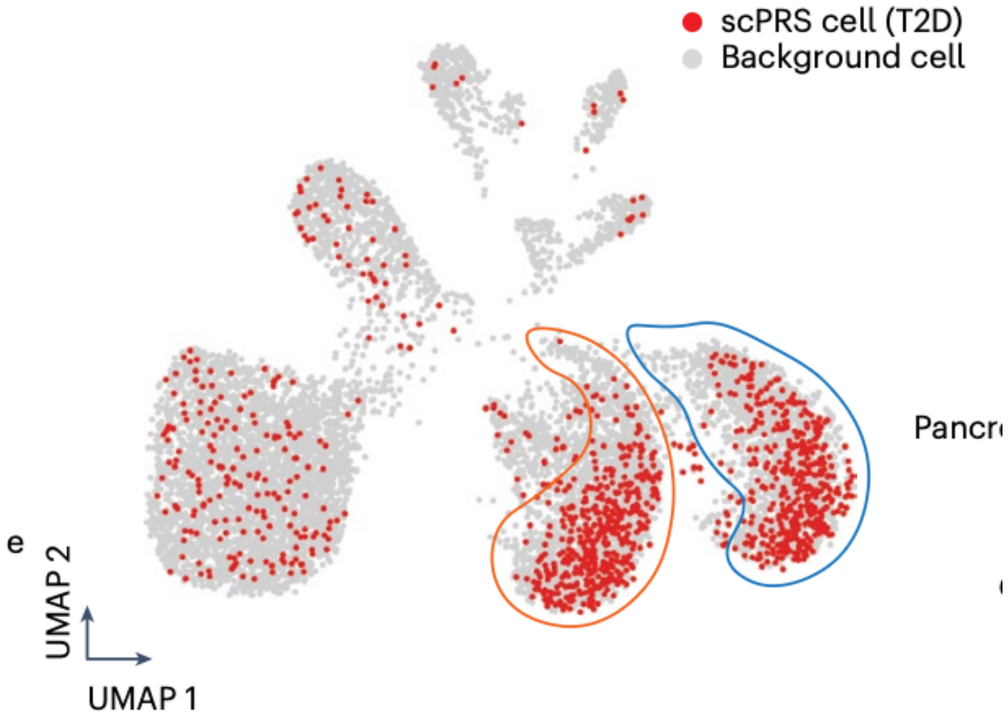
GNN learns the best way to smooth and denoise the data

scPRS performed well against baselines and other methods

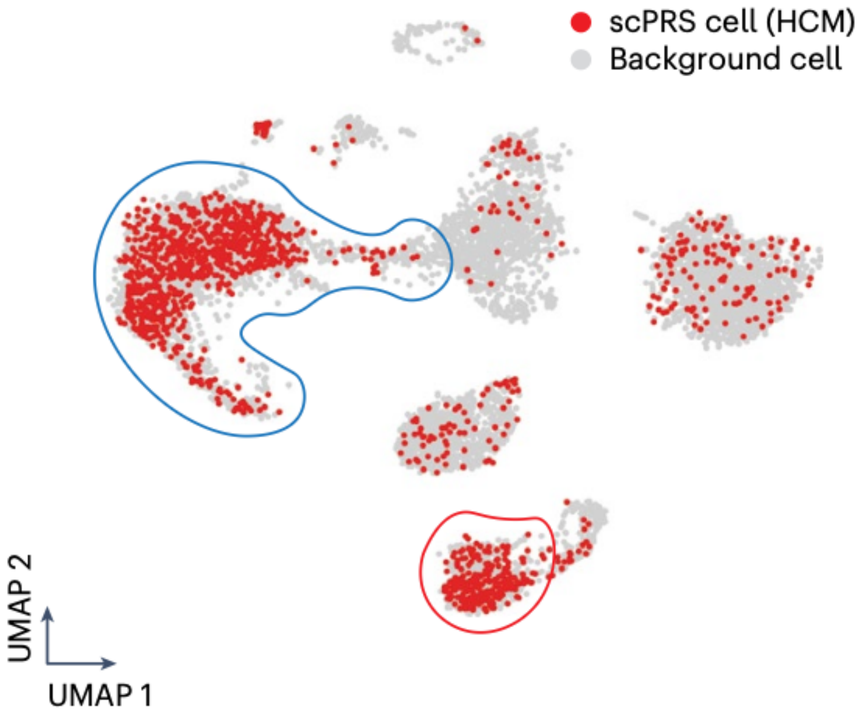


scPRS cells are disease-driver cells

— a cell with consistently high learned model weight



T2DM

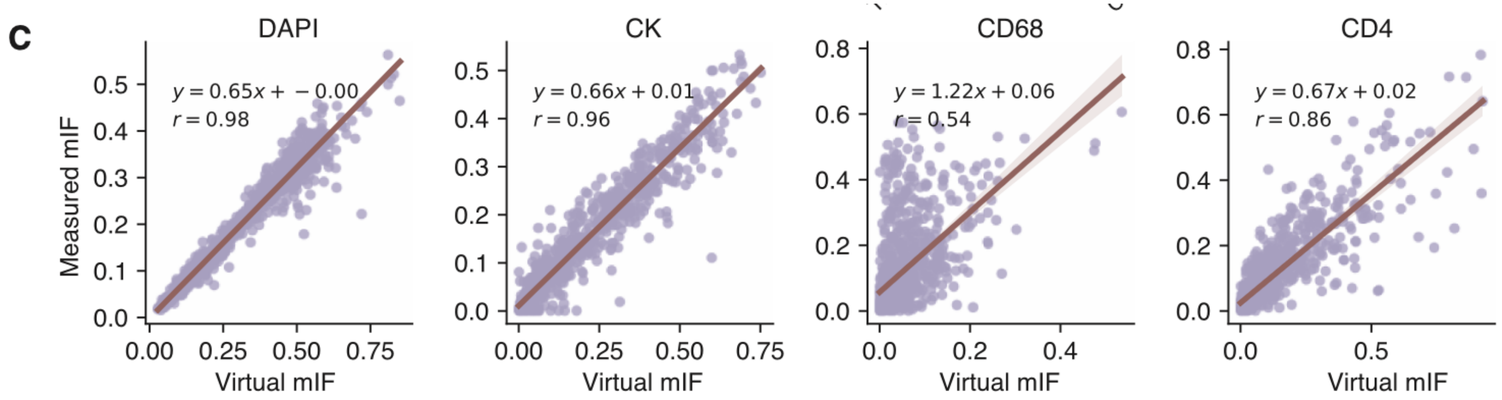


Hypertrophic Cardiomyopathy

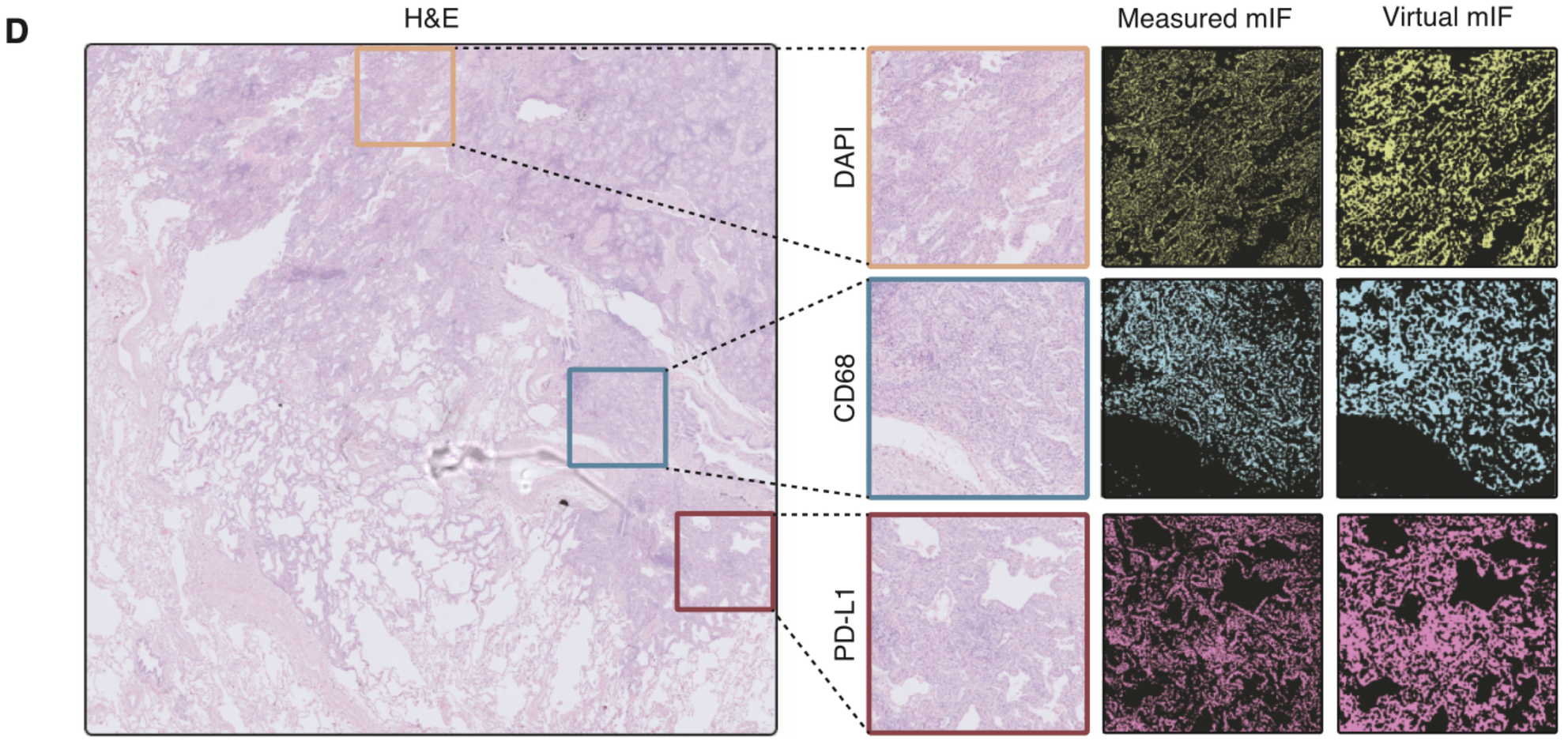
Multimodal AI generates virtual population for tumor microenvironment modeling (Valanarasu, Xu, Usuyama, Kim, et al., *Cell*)

- **Goal**: Multiplex immunofluorescence can reveal tumor immune microenvironments, but it is too expensive and low-throughput for population-scale studies - scale it using AI!
- **Method**: Train GigaTIME, a cross-modal AI model, on paired H&E and mIF images from 40 million cells to translate routine pathology slides into virtual mIF across 21 protein channels
- **Result**: Generated 299,376 virtual mIF slides for 14,256 patients across 24 cancers, identifying 1,234 protein–biomarker associations and survival/staging signatures
- **Conclusion**:
  - H&E slides contain a lot more information than just what we're using!
  - This is exciting, but how far can this go? At some level real data needs to be captured 🤔

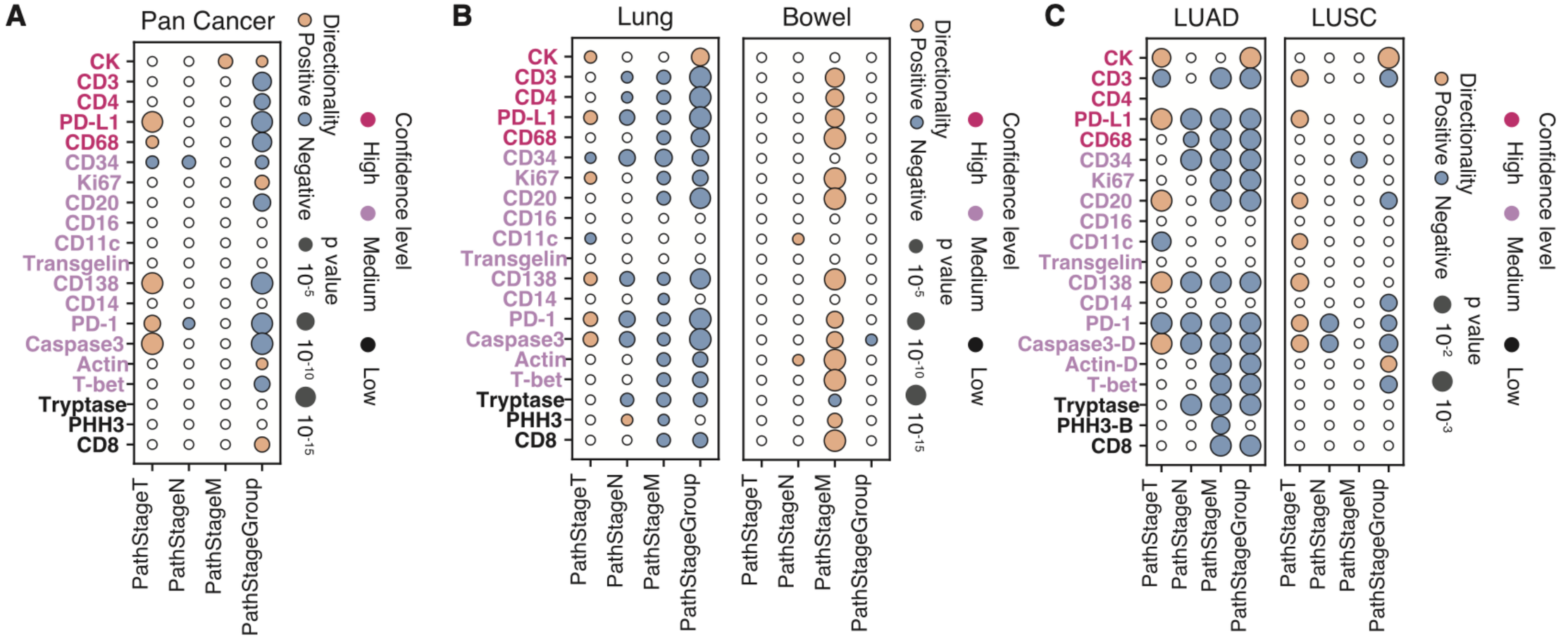
Not all perfect…



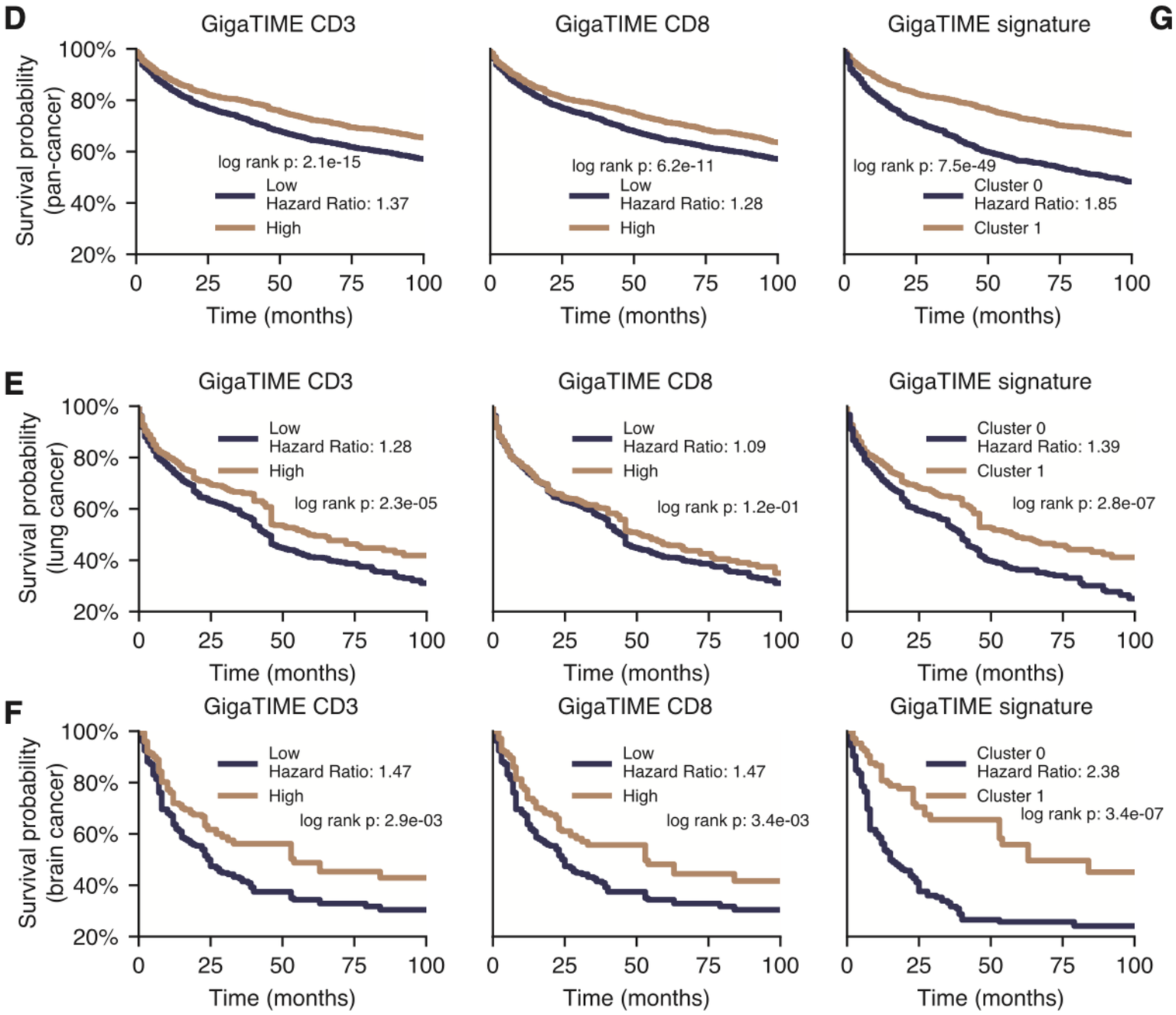
mIF=multiplex immunofluorescence



Virtual markers are predictive of stage

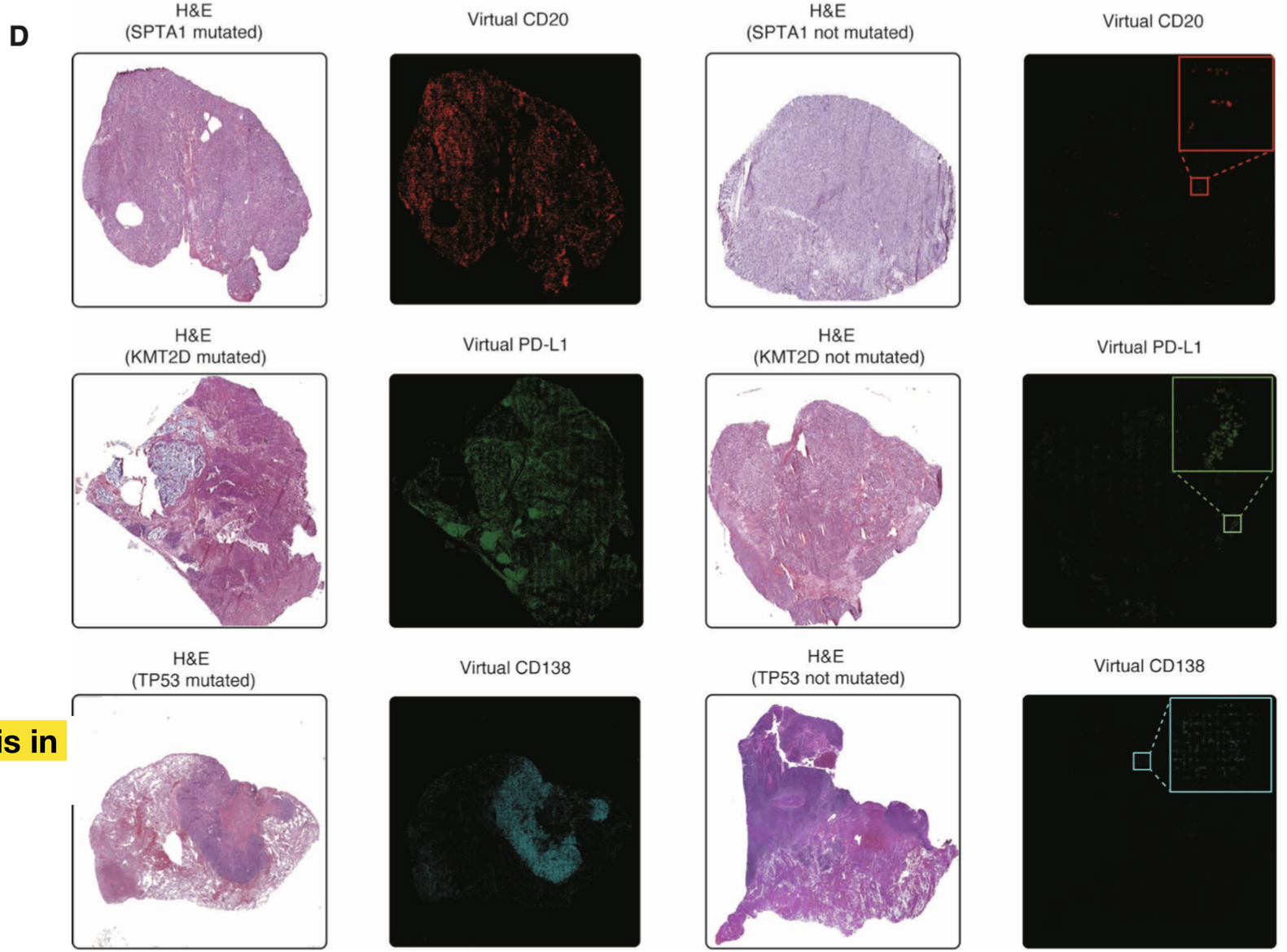


And outcomes



Should be expressed here

Shouldn't be



All this information is in the H&E!!!!!

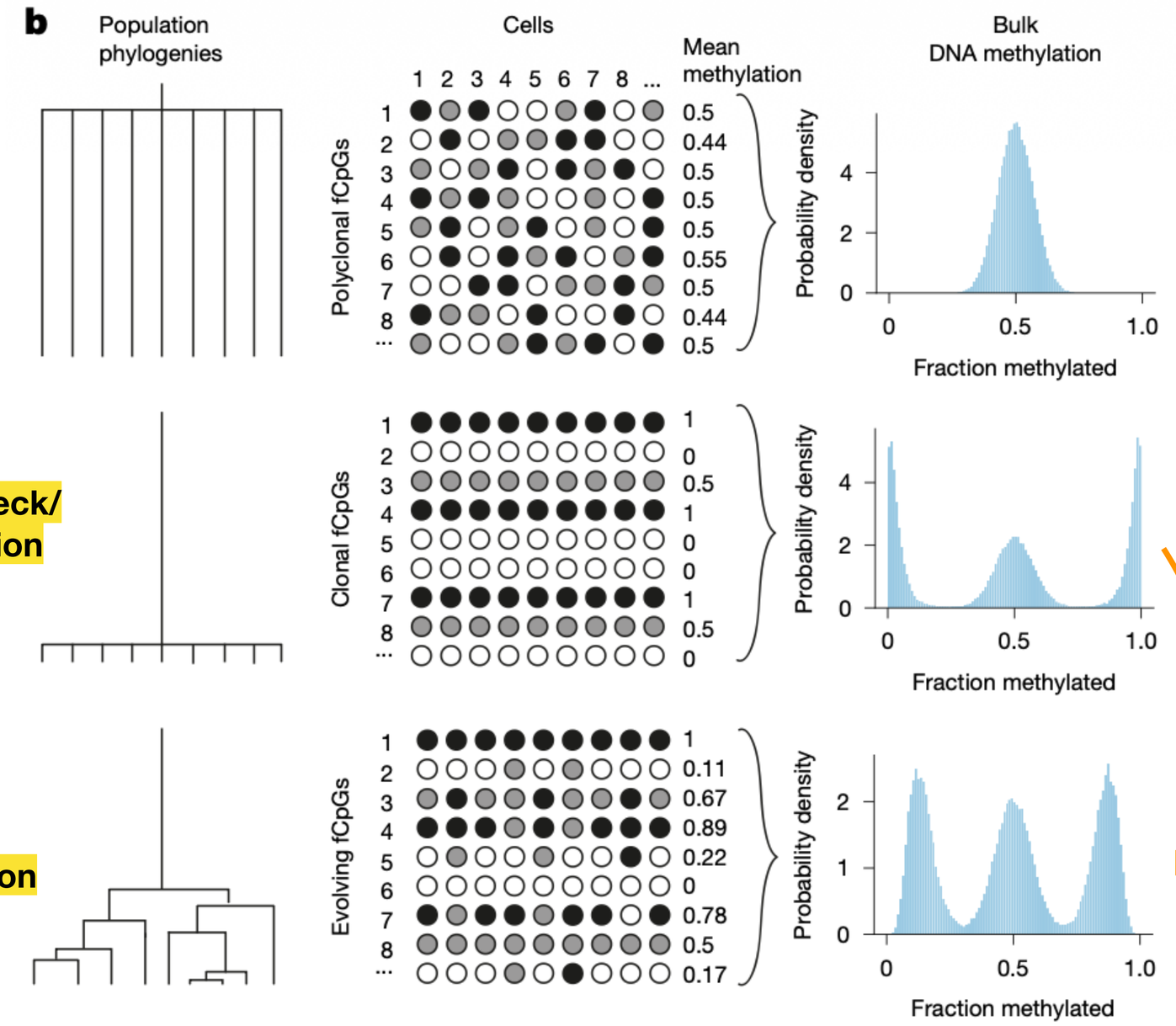
Fluctuating DNA methylation tracks cancer evolution at clinical scale (Gabbutt, Duran-Ferrer, Grant, et al., *Nature*)

- **Goal**: Infer the evolutionary history of human cancers from a single clinical tumor sample
- **Method**: Develop **EVOFLUx**
  - Bayesian model that uses naturally fluctuating CpG methylation sites as lineage “barcodes” to estimate tumour growth rate, cancer age, epimutation rates, subclonality, and independent tumour origins from bulk methylation data
- **Result**:
  - Applied to 1,976 lymphoid cancer samples
  - recovered disease-specific evolutionary histories, identified rare subclonal selection and independent primaries, and reconstructed longitudinal phylogenies
  - showed that faster inferred CLL growth predicts shorter time to first treatment
- **Conclusion**: Bulk methylation arrays may contain a hidden record of cancer evolution

“Healthy” state

Recent bottleneck/ clonal expansion

Tumor evolution



Difference between these tells you age, growth rate, sub clonal evo

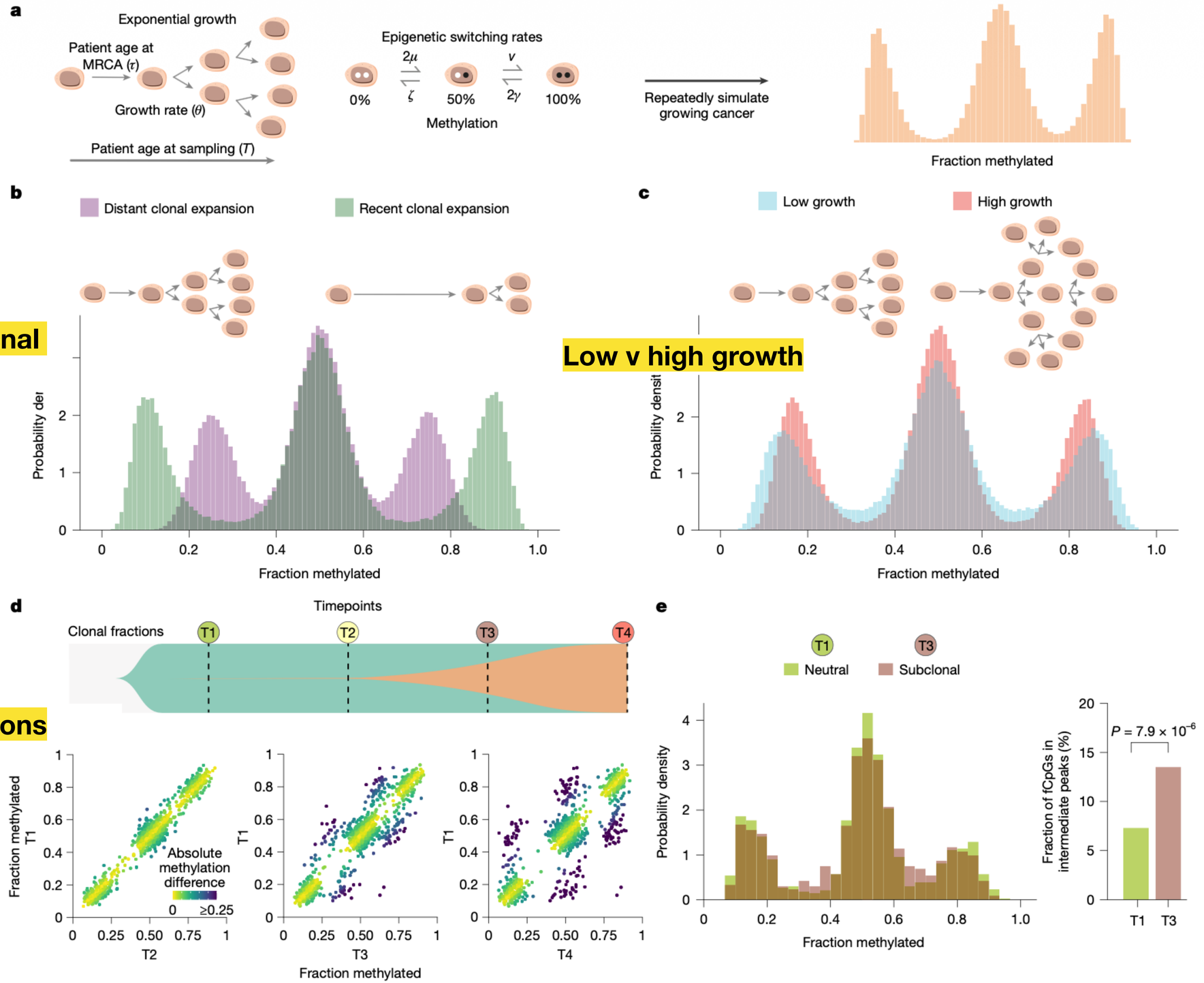
#IS25 #YIR25 𝕏@proftatone

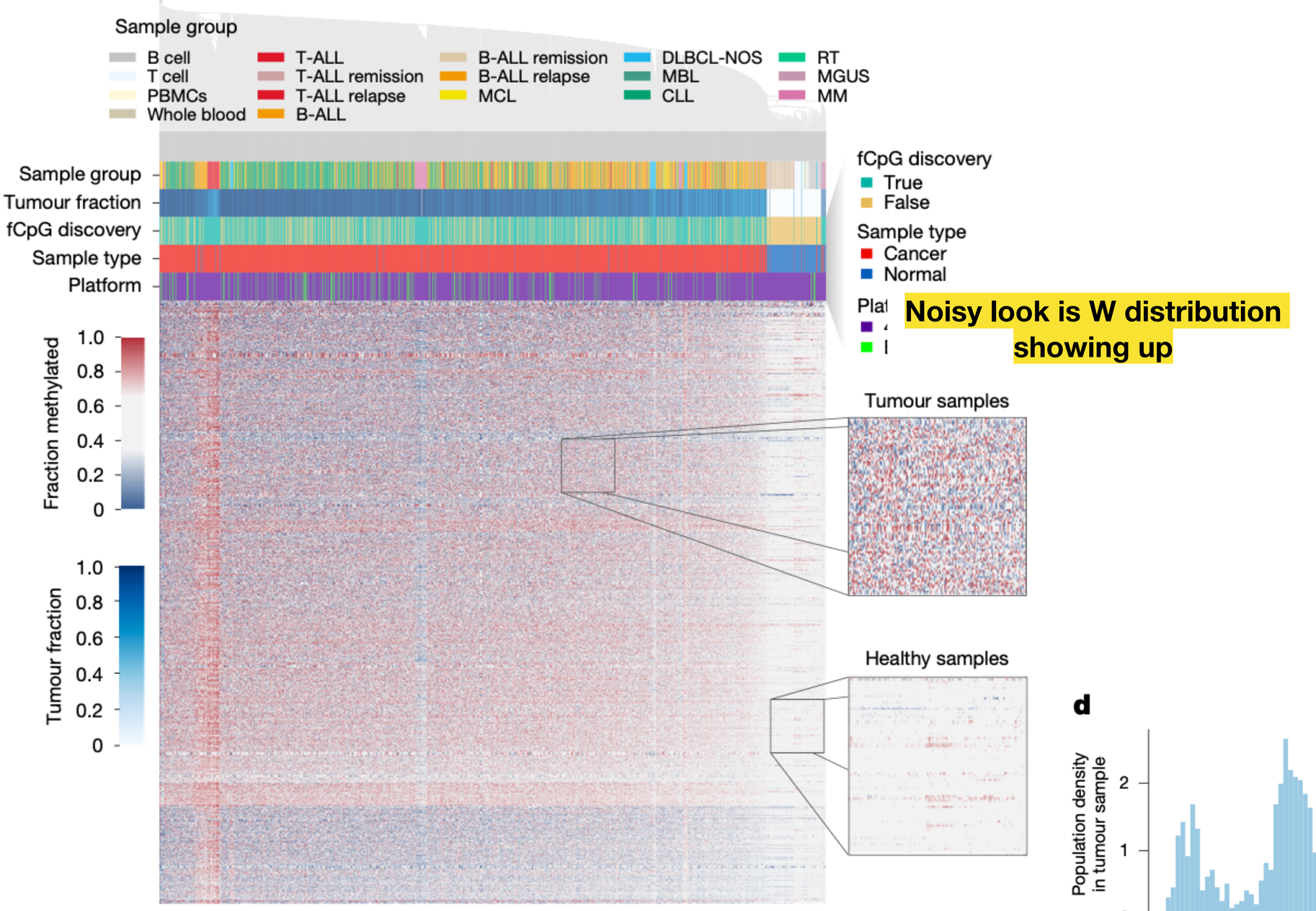
Simulations to generate lots of methylation "barcodes"

Distant v recent clonal expansion

Low v high growth

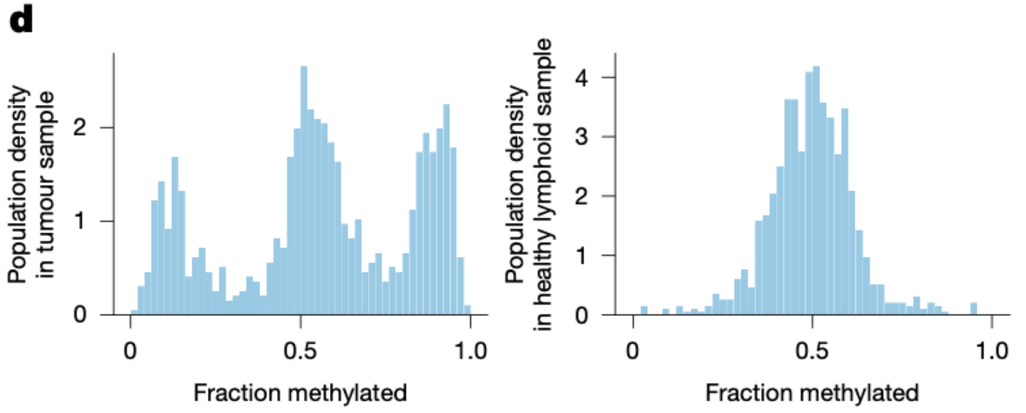
Clonal fractions

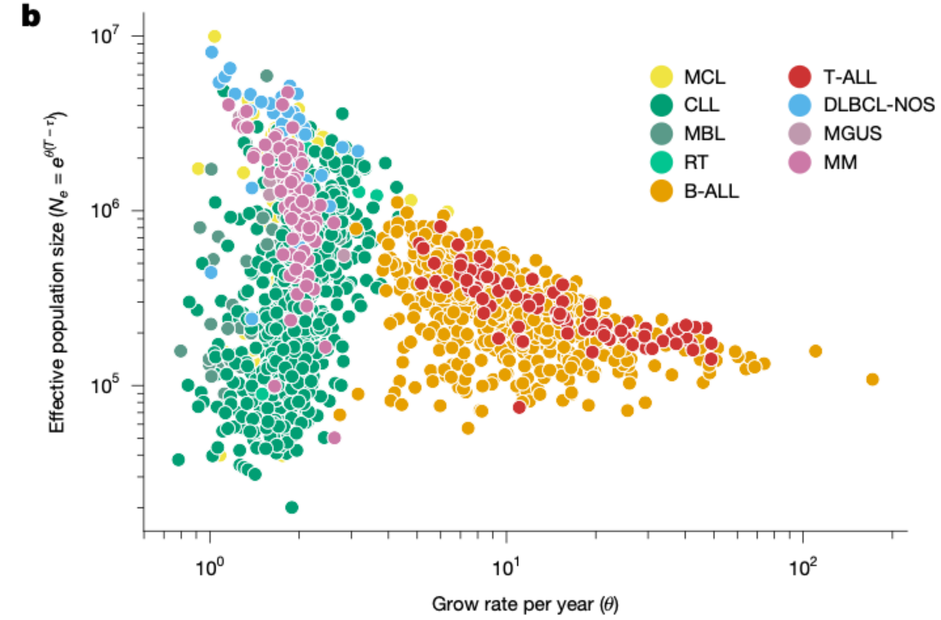


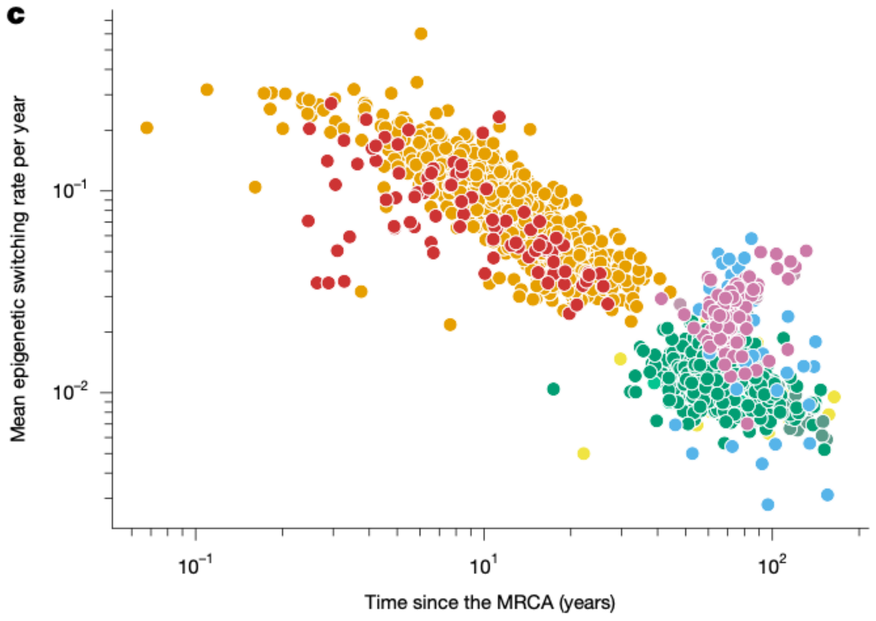


Noisy look is W distribution showing up

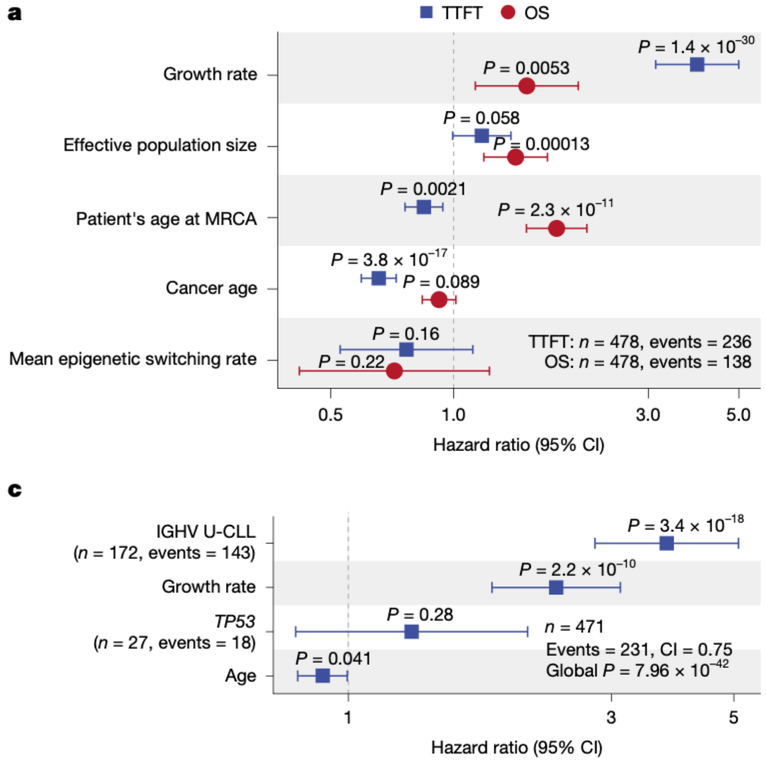
Real data now:

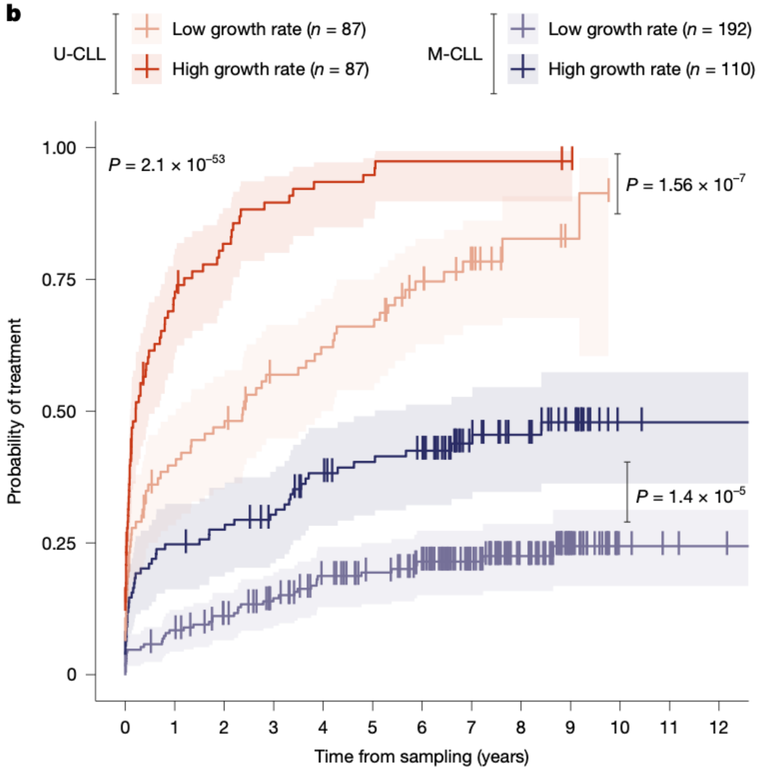






Estimated parameters are predictive of clinical outcomes





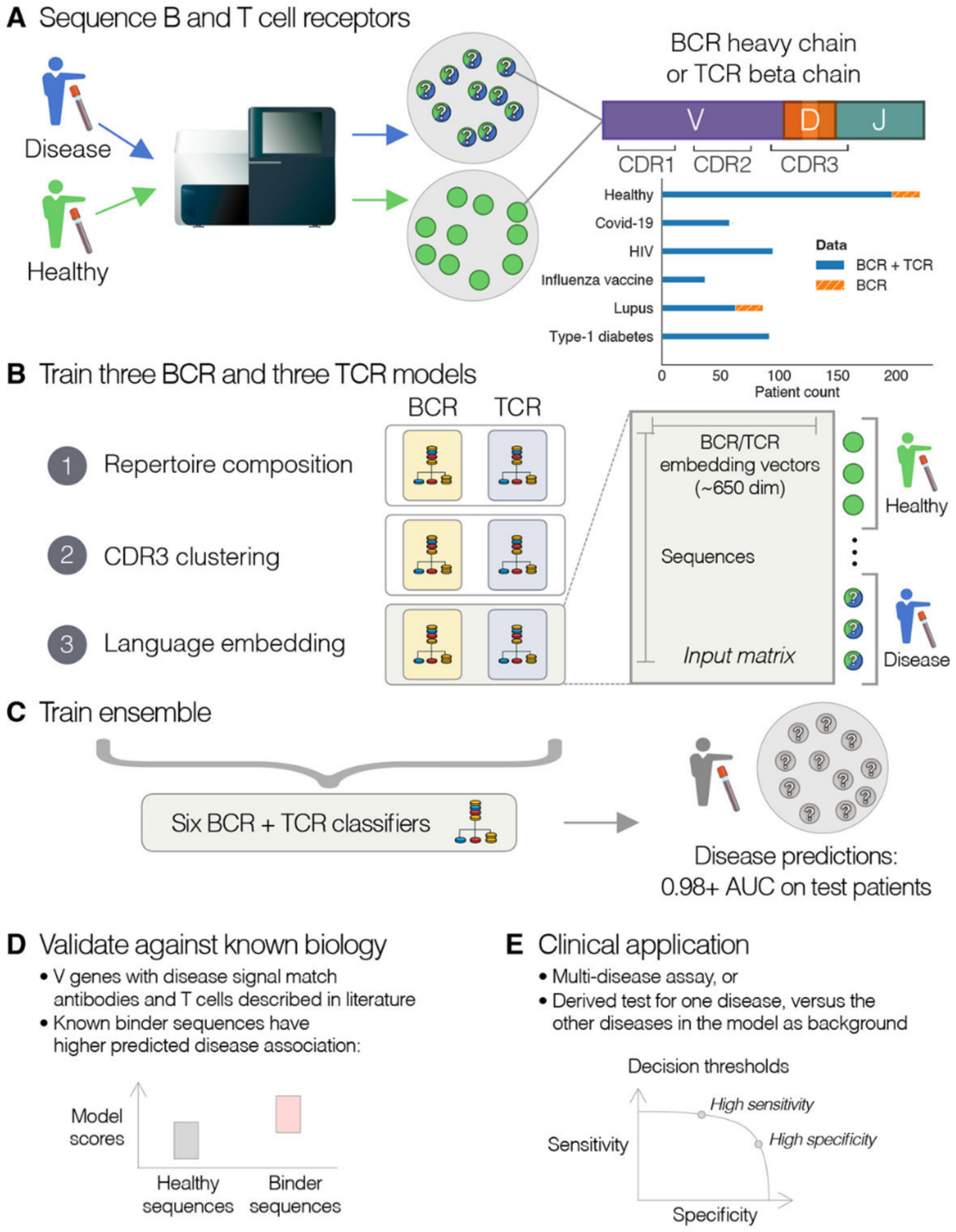
Somebody that I used to know - *Gotye*

Learning from Molecular Structures

using high-dimensional molecular or clinical-molecular data to classify, diagnose, or prognosticate

Disease diagnostics using machine learning of B cell and T cell receptor sequences (Zaslavsky, Craig, Michuda et al, *Science*)

- **Goal**: Use the adaptive immune system's "memory" of antigen exposures as a diagnostic signal for infection, autoimmunity, and vaccine response
- **Method**: Mal-ID integrates B-cell and T-cell repertoires using gene/isotype features, CDR3 sequence clusters, ESM-2 protein language model embeddings, and an ensemble classifier
- **Result**: Classified six immune states from 593 individuals with AUROC = 0.986; BCR + TCR together outperformed either alone.
- **Conclusion**: Immune receptor sequencing may become a general-purpose diagnostic readout



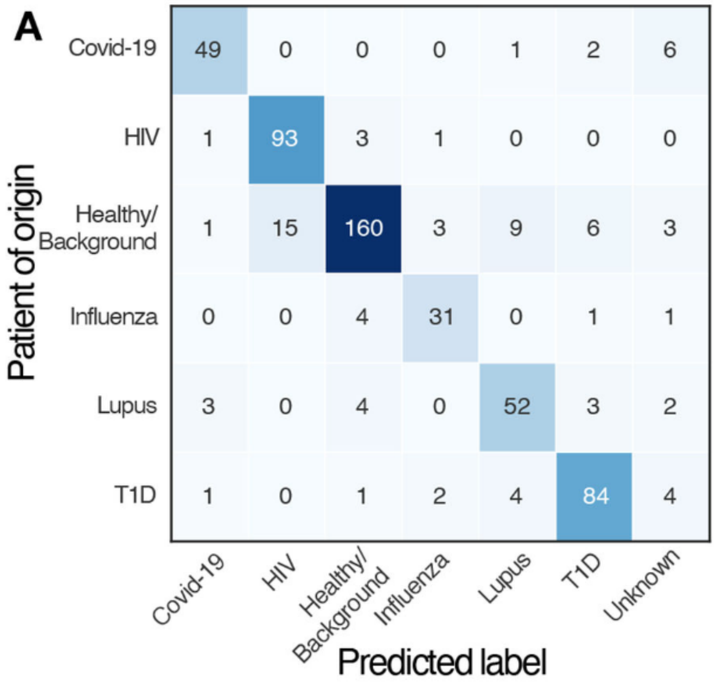
Sequence the relevant BCR and TCR chains

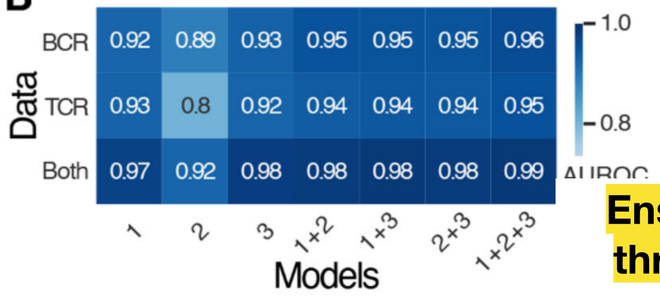
Build three models (Model 3 includes ESM-2)

Train an ensemble classifier
Validate against held-out set

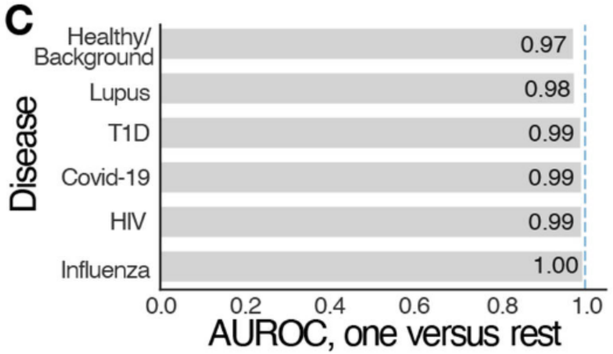
Validate against known biology and in clinical applications

Great confusion matrix

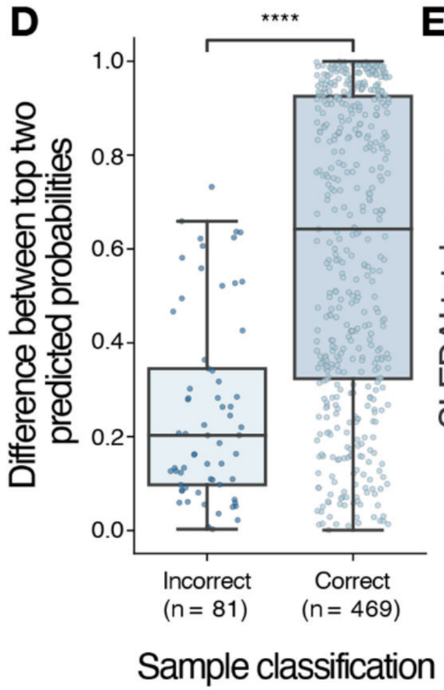


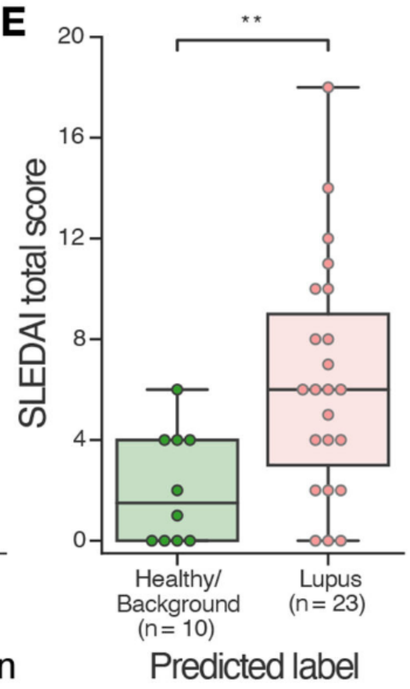


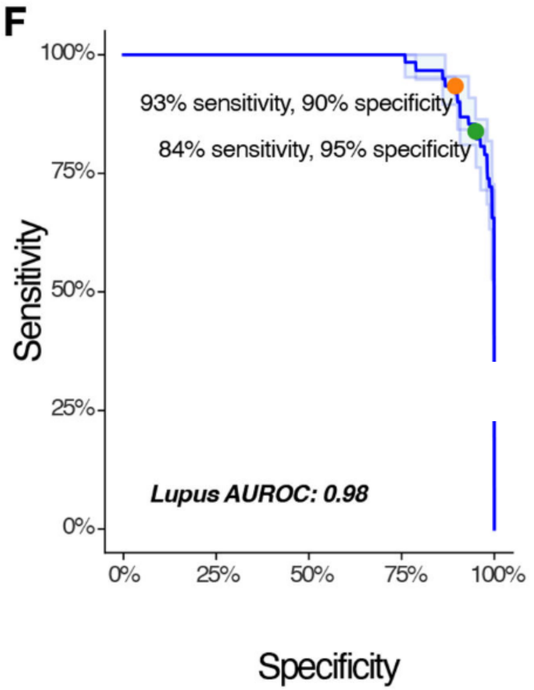
Ensemle model include all three models works best!



Correct classifications have bigger probability differences (more confidence)





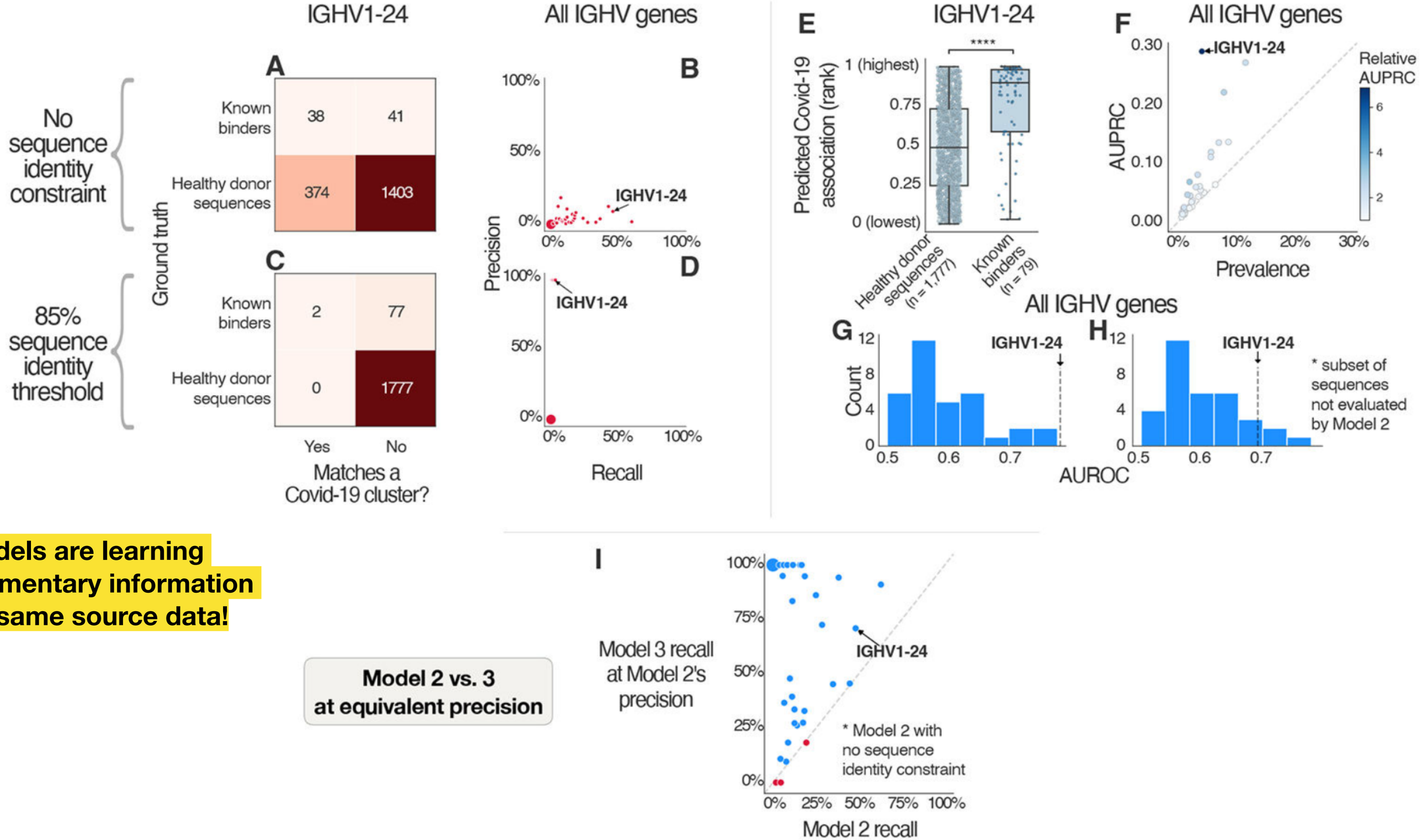


Lupus, classified.

#IS

Model 2. CDR3 Clustering

Model 3. Language embedding



Models are learning complementary information from same source data!

Model 2 vs. 3 at equivalent precision

LITERATURE HIGHLIGHT

Improving Polygenic Risk Prediction Performance by Integrating Electronic Health Records through Phenotype Embedding

Authors: Xu et al. | Journal: American Journal of Human Genetics (AJHG)

Presented by: Ishita Vasudev, MD

BronxCare Health System | Icahn School of Medicine at Mount Sinai

Informatics Novelty: 5/5 Application Importance: 5/5 Presentability: 5/5

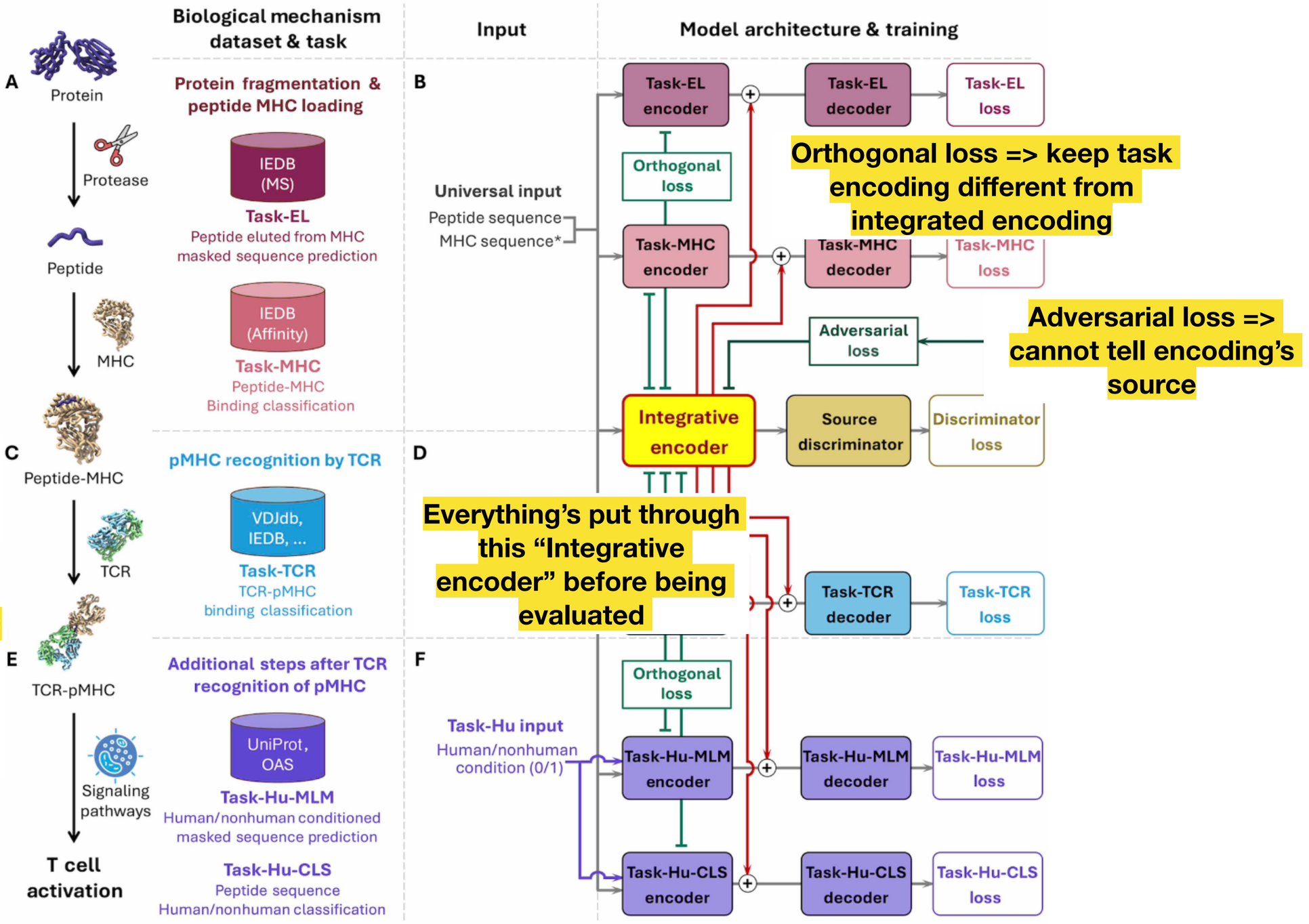
T-SCAPE: T cell immunogenicity scoring via cross-domain aided predictive engine (Kim, Jung, Lee, et al., *Science Advances*)

- **Goal**: Predict whether peptide-MHC pairs will trigger T cell activation even though the data we have are very very sparse
- **Method**: Pretrain on imperfect but biologically adjacent tasks across the immunogenicity cascade, then fine-tune on scarce T cell activation data
- **Result**: Outperformed or matched leading immunogenicity predictors on leakage-controlled neoantigen, infectious disease, and antidrug antibody benchmarks; especially strong for infectious disease epitopes
- **Conclusion**: Weak proxy labels can become useful when they are organized by mechanism rather than treated as generic extra data

Organized T-cell activation into a biological cascade

Identified data for each of these parts where data is more plentiful

Note: The labels for MHC binding are not substitutes for T-cell activation; they are weakly related only!

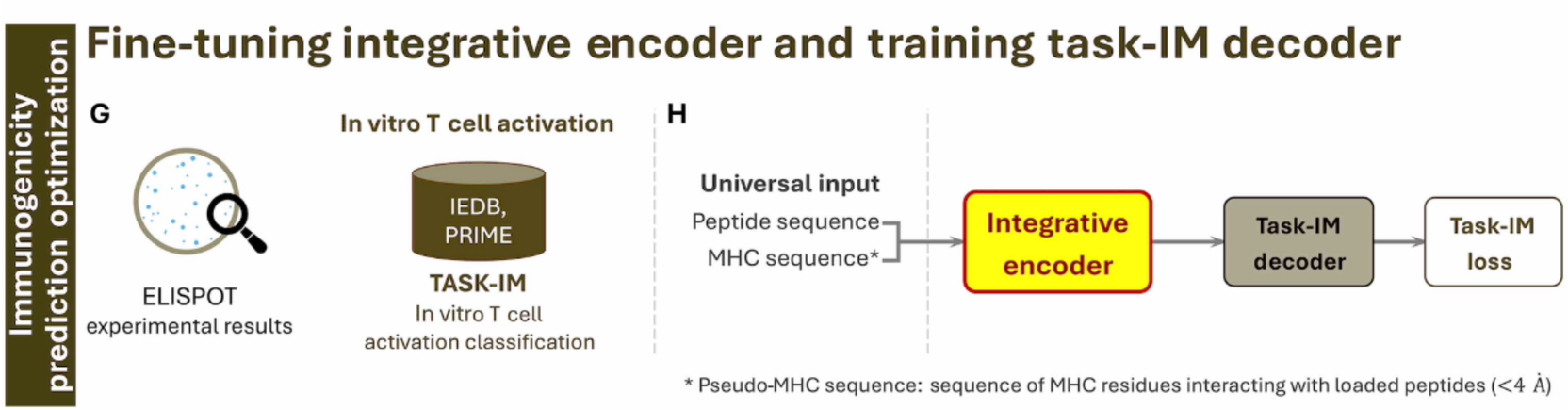


Orthogonal loss => keep task encoding different from integrated encoding

Adversarial loss => cannot tell encoding's source

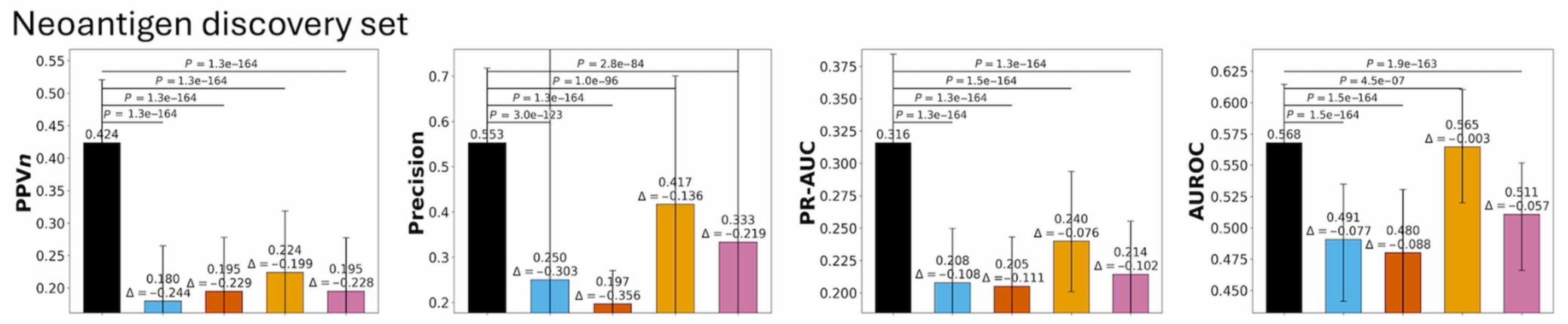
Everything's put through this "Integrative encoder" before being evaluated

Equipped with this integrative encoder, now the precious T-cell activation data can be used for fine-tuning



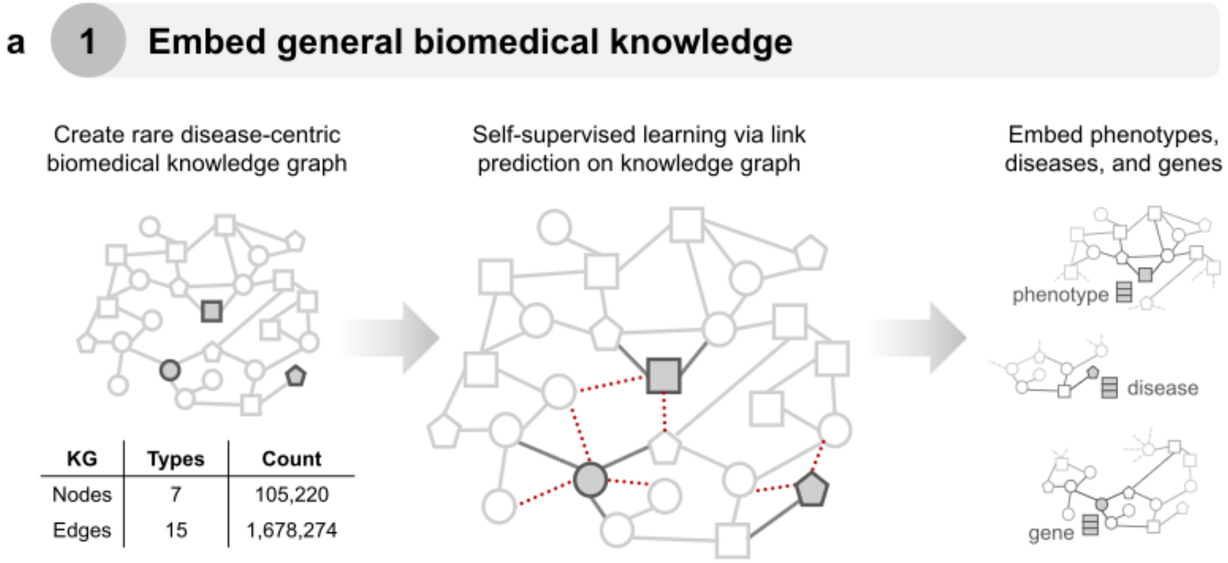
Not shown: they go in depth on how the signals can be weak and possibly misleading and how they protect against all kinds of data leakage.

Ablation analysis shows that you really need all steps of the cascade for this to work



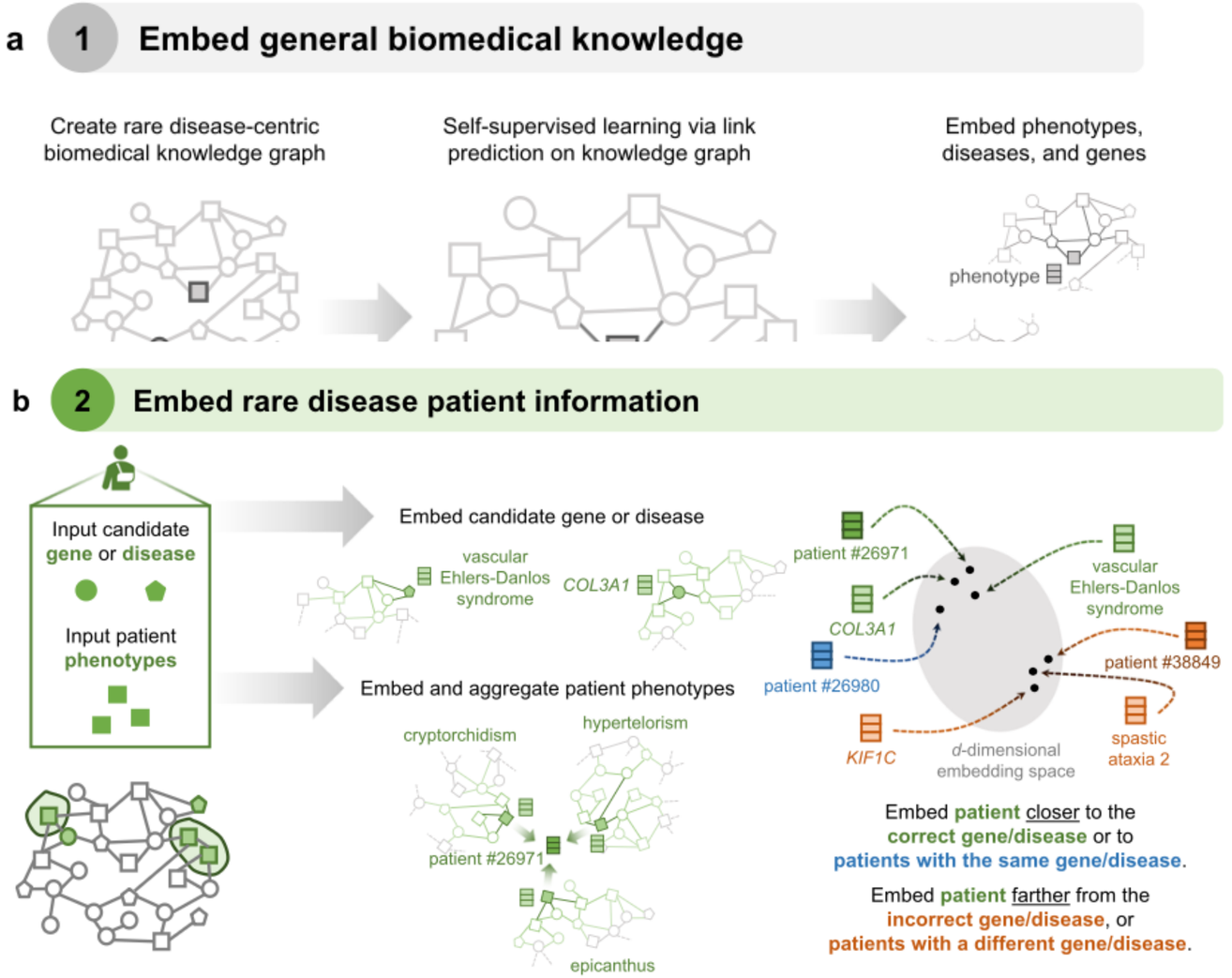
Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases (Alsentzer, Li, Kobren, Noori, UDN, Kohane & Zitnik, *npj Digital Medicine*)

- **Goal**: Help diagnose rare genetic diseases when there are too few patients per disease to train conventional deep learning models
- **Method**: SHEPHERD embeds patient HPO phenotypes, candidate genes, diseases, and similar patients in a rare-disease knowledge graph using few-shot geometric deep learning
- **Result**: In an external UDN cohort, SHEPHERD ranked the causal gene first in 40% of expert-curated cases and in the top five for 85%; it also retrieved “patients-like-me” and characterized novel disease presentations
- **Conclusion**: Rare diseases are everywhere, but examples are not. SHEPHERD uses the graph to fill in the missing supervision

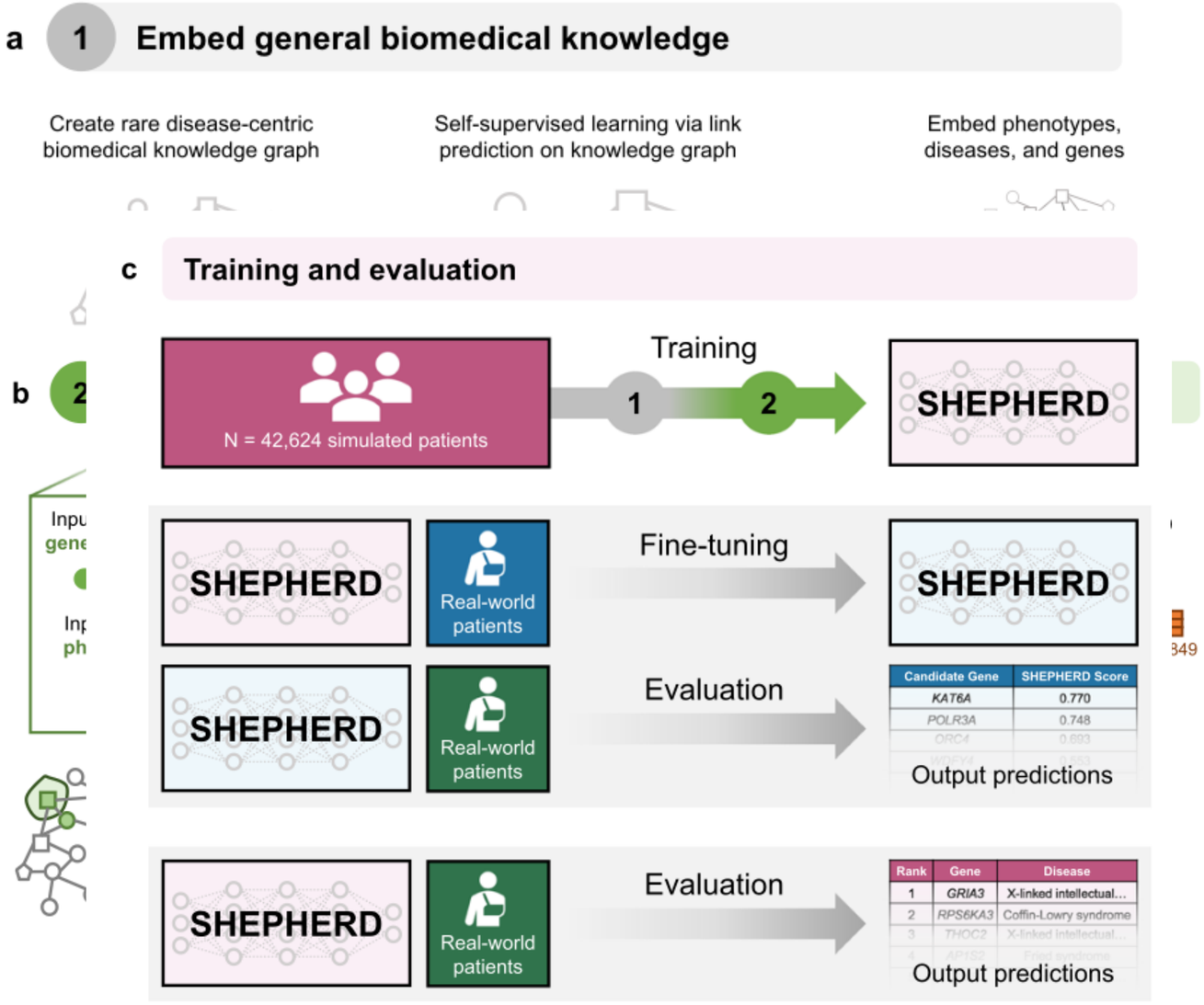


KG	Types	Count
Nodes	7	105,220
Edges	15	1,678,274

Start with a biomedical knowledge graph -> do self-supervised learning on link prediction to get embeddings



Embed rare disease patient information and set up constrastive loss function



Generate synthetic rare patient data to train the model

42k simulated patients

Start with 2k canonical Mendelian disease, then:

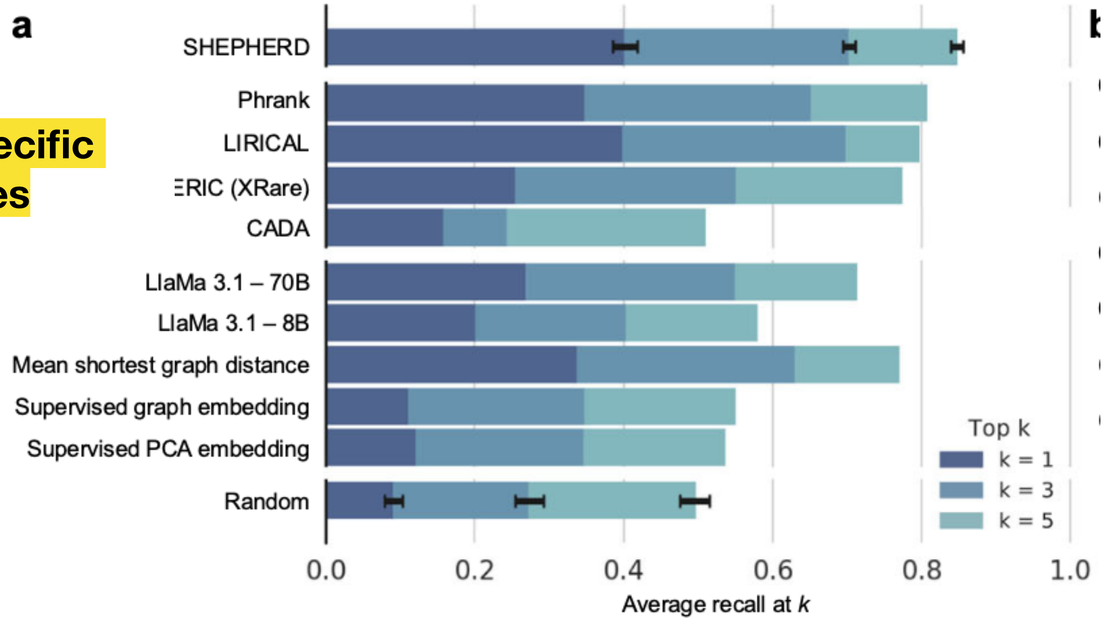
- **phenotype dropout**
- **Phenotype corruption (reduce specificity)**
- **Phenotype noise (add unrelated terms)**
- **Gene distractors (plausible but incorrect genes)**

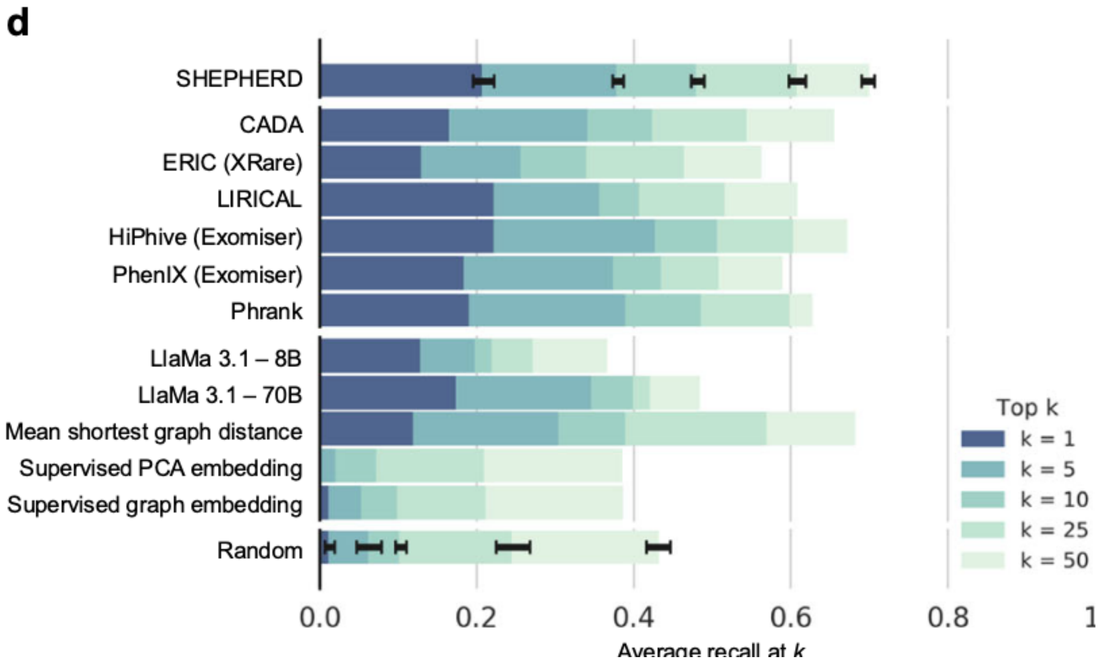
Compared to:

Domain specific baselines

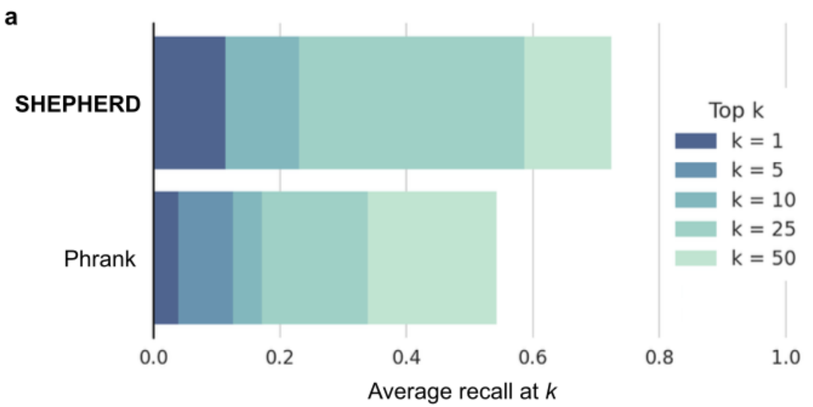
Language Models

Trad ML/graph methods



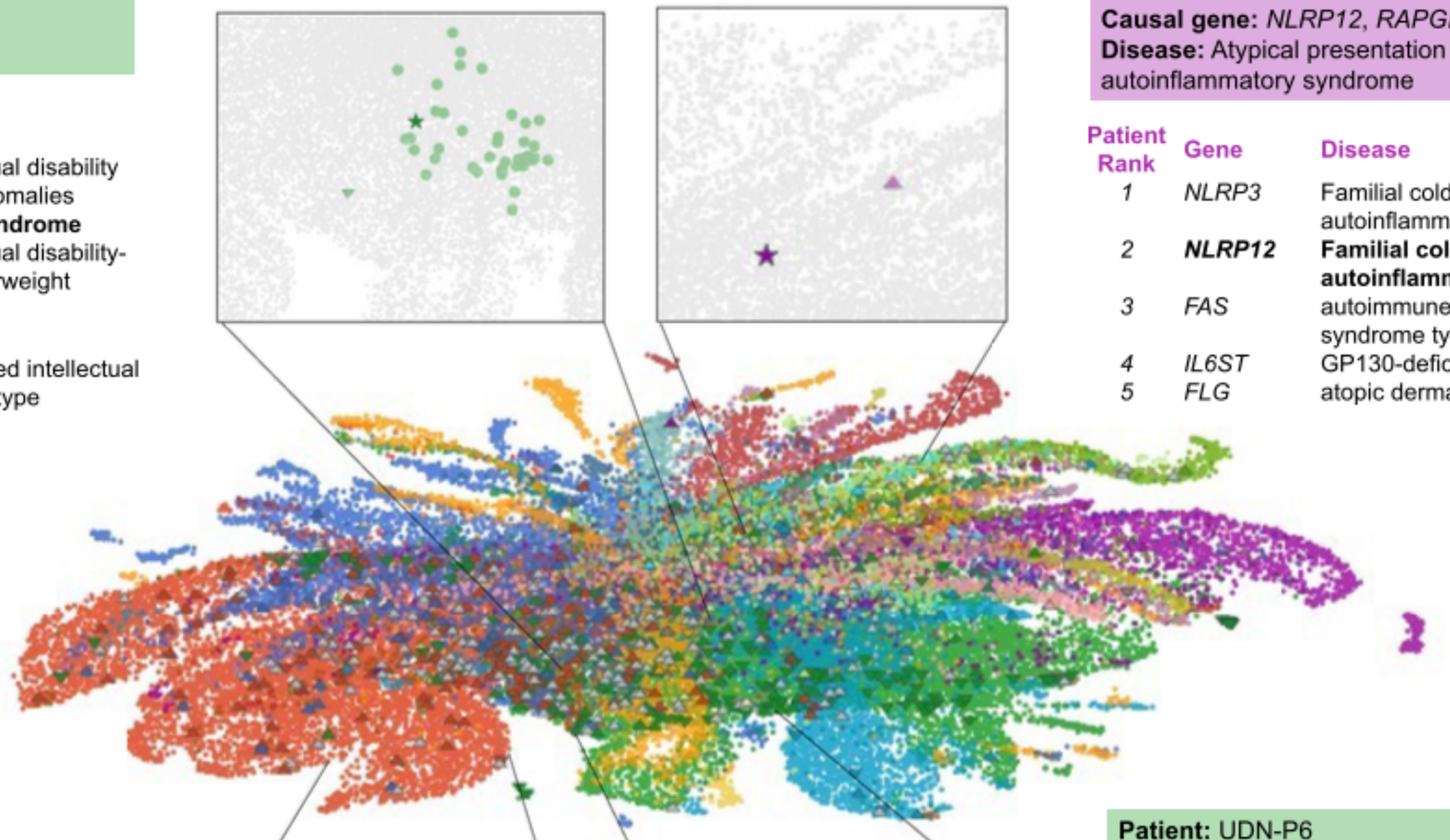


Outperforms baselines in patients-like-me analysis



Patient: UDN-P3 ***Patient Card***
Causal gene: *RPS6KA3*
Disease: Coffin-Lowry syndrome

Patient Rank	Gene	Disease
1	*GRIA3*	X-linked intellectual disability due to GRIA3 anomalies
2	***RPS6KA3***	**Coffin-Lowry syndrome**
3	*THOC2*	X-linked intellectual disability-short stature-overweight syndrome
4	*AP1S2*	Fried syndrome
5	*SMS*	Syndromic X-linked intellectual disability Snyder type



Patient: UDN-P5 ***Patient Card***
Causal gene: *NLRP12*, *RAPGEFL1*
Disease: Atypical presentation of familial cold autoinflammatory syndrome

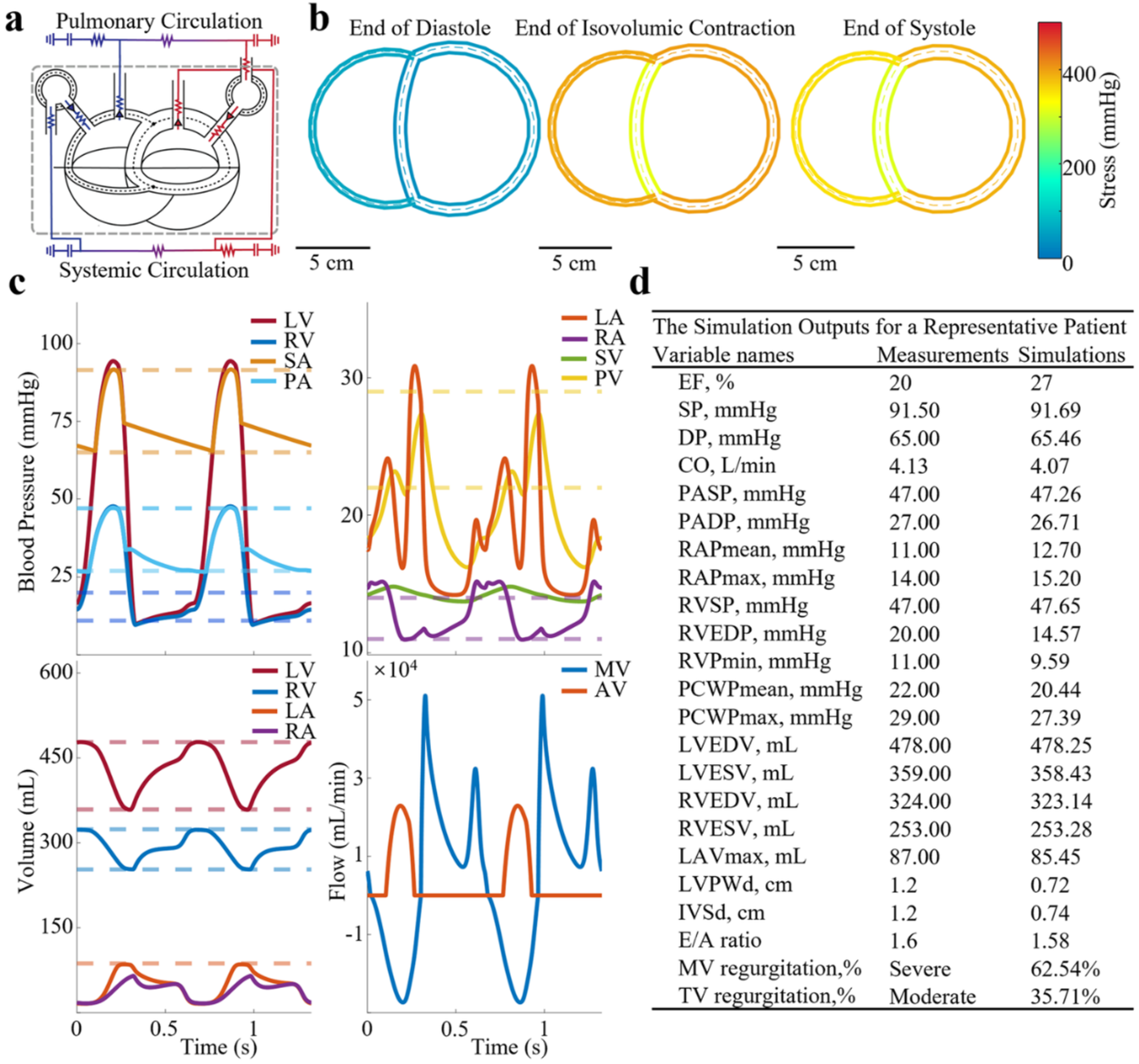
Patient Rank	Gene	Disease
1	*NLRP3*	Familial cold-induced autoinflammatory syndrome 1
2	***NLRP12***	**Familial cold-induced autoinflammatory syndrome 2**
3	*FAS*	autoimmune lymphoproliferative syndrome type 1
4	*IL6ST*	GP130-deficient hyper-IgE syndrome
5	*FLG*	atopic dermatitis 2

Patient: UDN-P6 ***Patient Card***

Identification of digital twins to guide interpretable AI for diagnosis and prognosis in heart failure (Gu et al, *npj Digital Medicine*)

- **Goal**: Heart failure is heterogeneous; build patient-specific digital twins that make prognosis more interpretable than clinical variables alone
- **Method**: Fit mechanistic cardiovascular models to EHR + TTE + CMR + right-heart catheterization data from 343 HF patients; use digital twin-derived features for clustering and random survival forest risk prediction
- **Result**: Digital twins identified mechanistically distinct HF phenogroups and improved prognostic modeling, especially when combined with clinical features
- **Conclusion**: Digital twins are maturing beyond a simulation toy — to something that can transform messy clinical measurements into usable physiological features

Used data that can be derived from the EHR to fit an established heart model



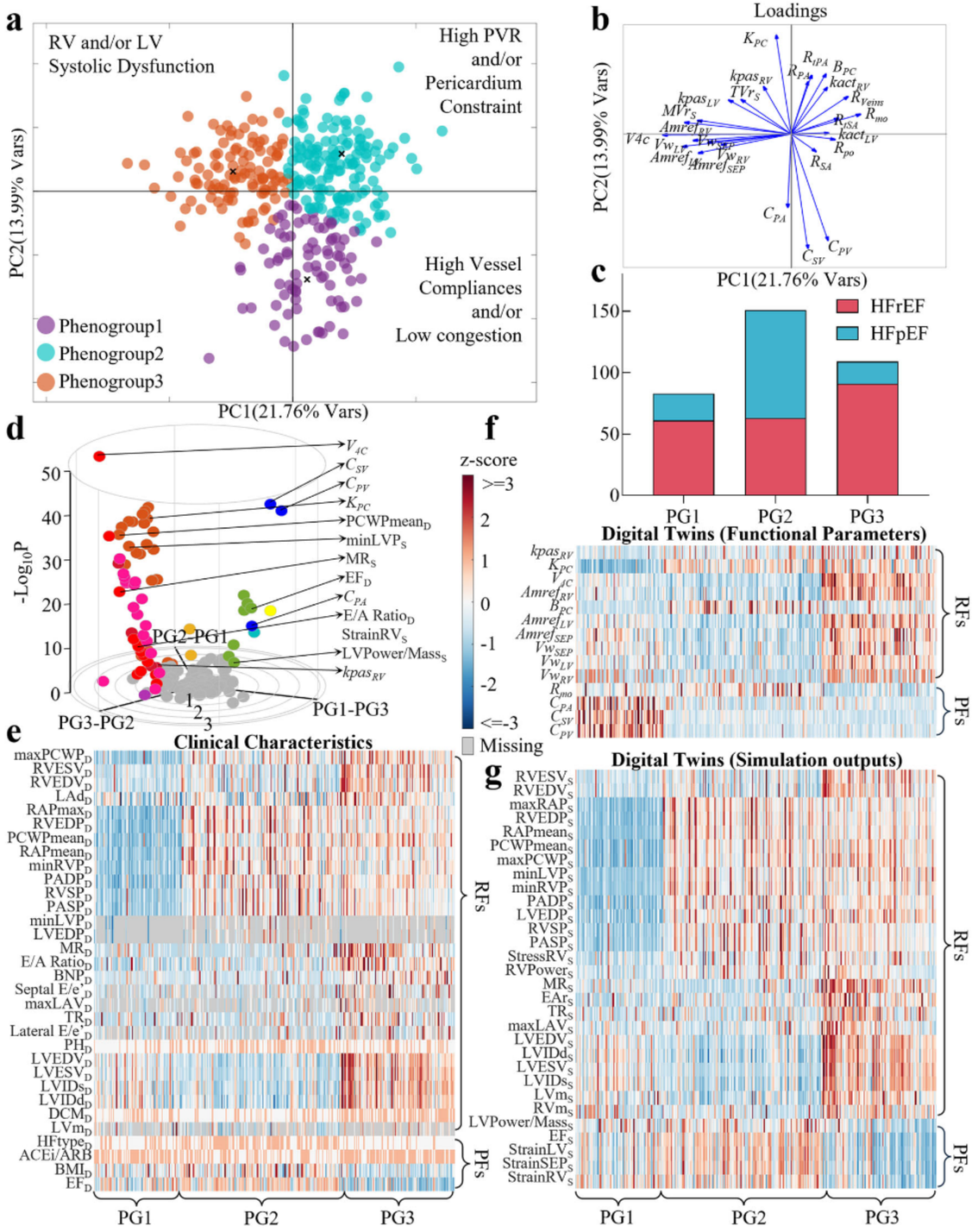
d

The Simulation Outputs for a Representative Patient

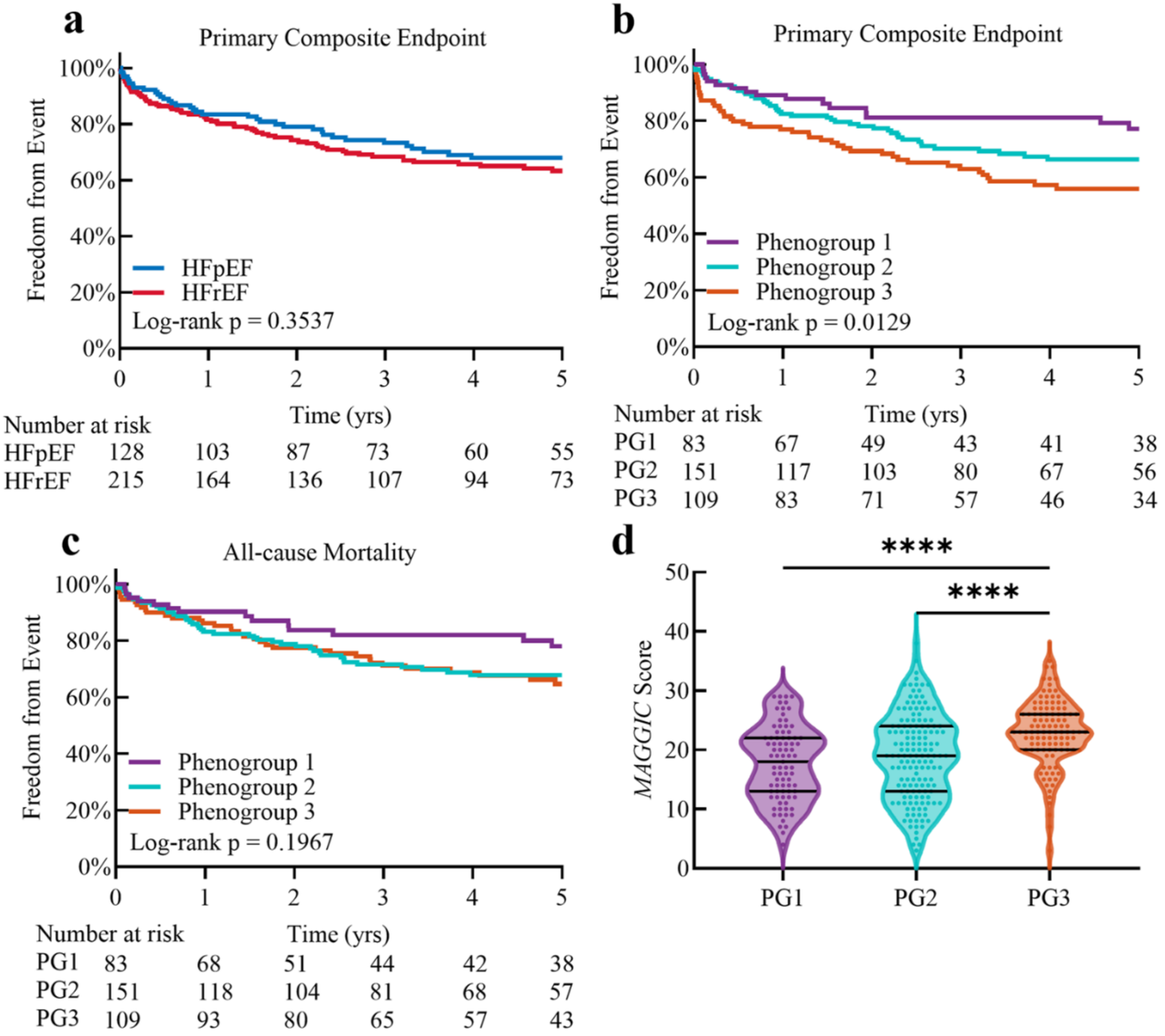
Variable names	Measurements	Simulations
EF, %	20	27
SP, mmHg	91.50	91.69
DP, mmHg	65.00	65.46
CO, L/min	4.13	4.07
PASP, mmHg	47.00	47.26
PADP, mmHg	27.00	26.71
RAPmean, mmHg	11.00	12.70
RAPmax, mmHg	14.00	15.20
RVSP, mmHg	47.00	47.65
RVEDP, mmHg	20.00	14.57
RVPmin, mmHg	11.00	9.59
PCWPmean, mmHg	22.00	20.44
PCWPmax, mmHg	29.00	27.39
LVEDV, mL	478.00	478.25
LVESV, mL	359.00	358.43
RVEDV, mL	324.00	323.14
RVESV, mL	253.00	253.28
LAVmax, mL	87.00	85.45
LVPWd, cm	1.2	0.72
IVSd, cm	1.2	0.74
E/A ratio	1.6	1.58
MV regurgitation,%	Severe	62.54%
TV regurgitation,%	Moderate	35.71%

#IS25 #YIR25 𝕏@prof

Digital twins found distinct phenotypes

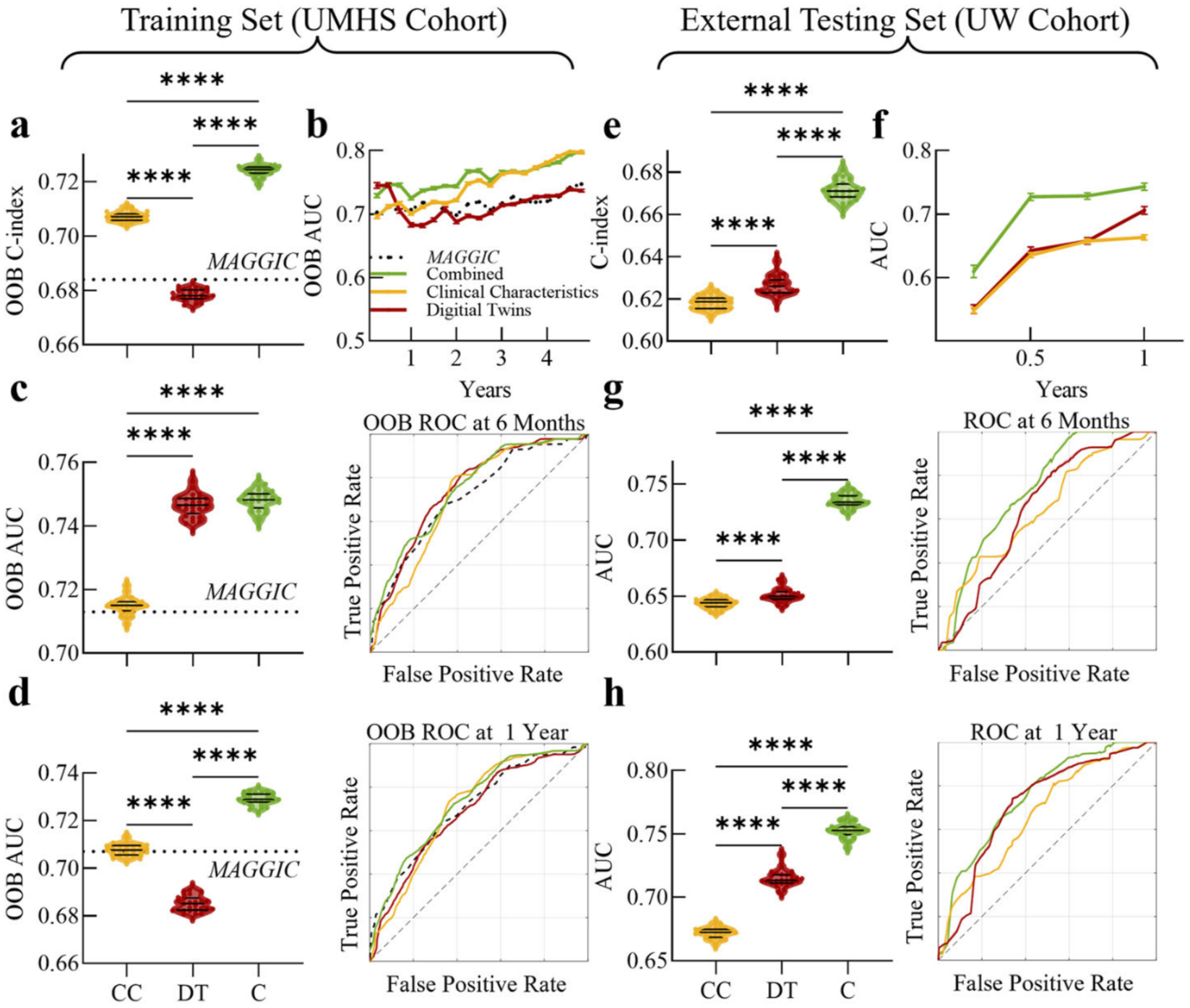


And simulation outputs matched clinical characteristics

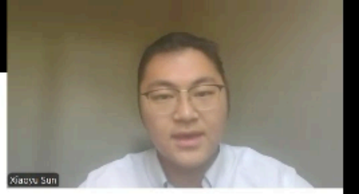


Digital twin phenotypes better differentiate in survival analysis

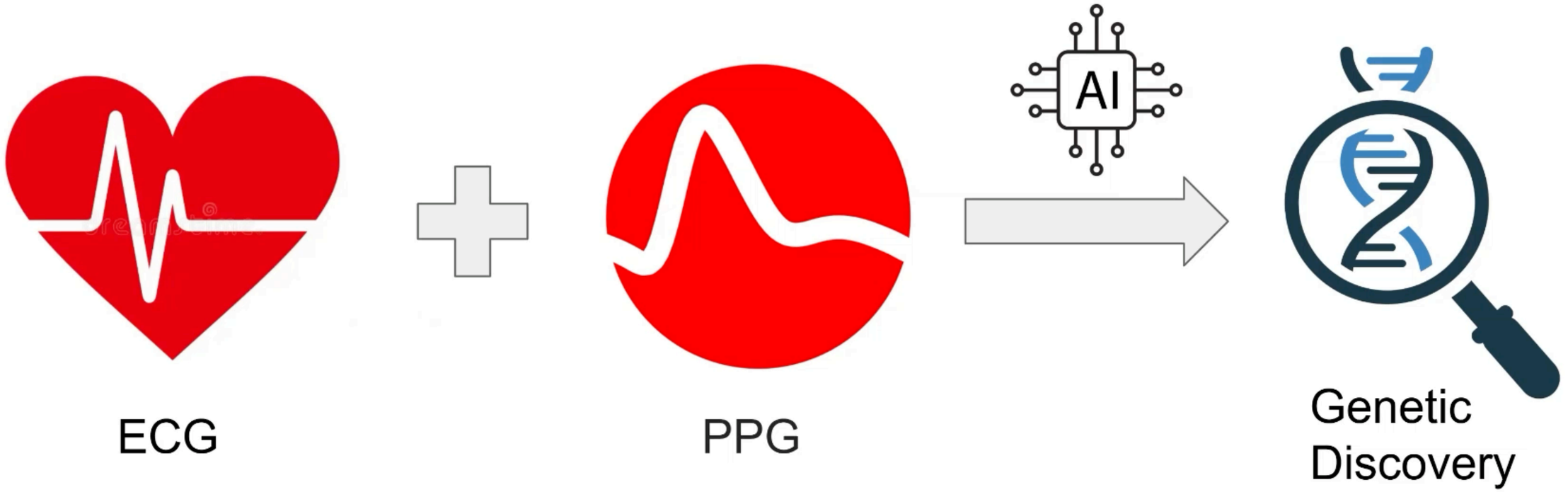
#I



Digital twin phenotypes hold up in external data



Waveforms reveal cardiovascular genetics



Zhou, Yuchen et al. “Applying multimodal AI to physiological waveforms improves genetic prediction of cardiovascular traits.” American journal of human genetics vol. 112,7 (2025): 1562-1579. doi:10.1016/j.ajhg.2025.05.015

Toxic - *Britney Spears*

AI-Guided Perturbation and Therapeutic Discovery

computational methods that nominate drugs, targets, ligands, perturbations, or interventions

CLL to Richter syndrome: Integrating network strategies with experiments elucidating disease drivers and personalized therapies (Maier et al., *Science Advances*)

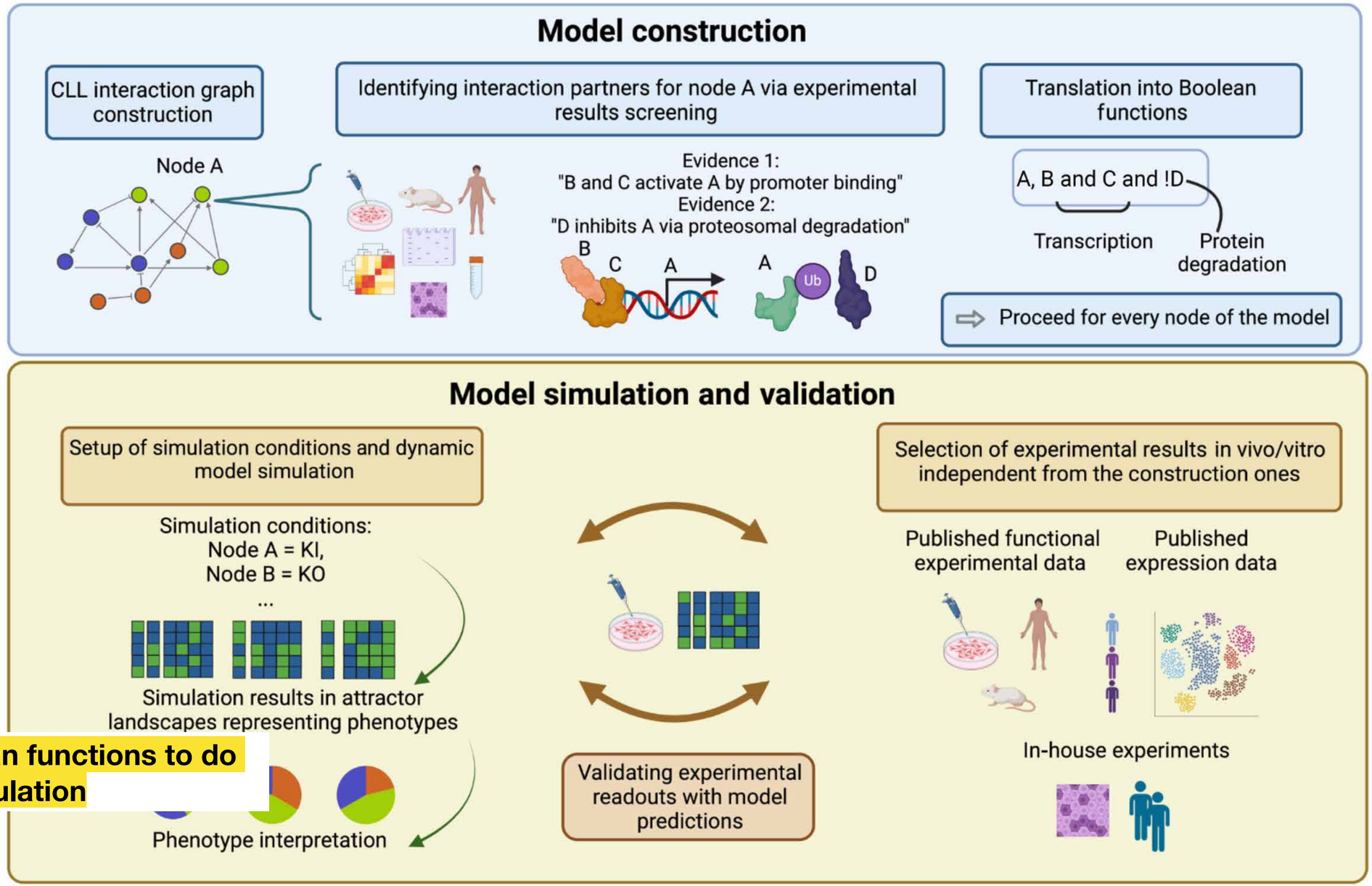
- **Goal**: Understand how chronic lymphocytic leukemia transforms into aggressive Richter syndrome, where experimental models are limited and prognosis is poor

- **Method**: Built a 49-node Boolean logic model from 228 publications, simulated CLL and RS perturbation states, and validated against mouse models, scRNA-seq, and a CLL/RS FFPE patient cohort

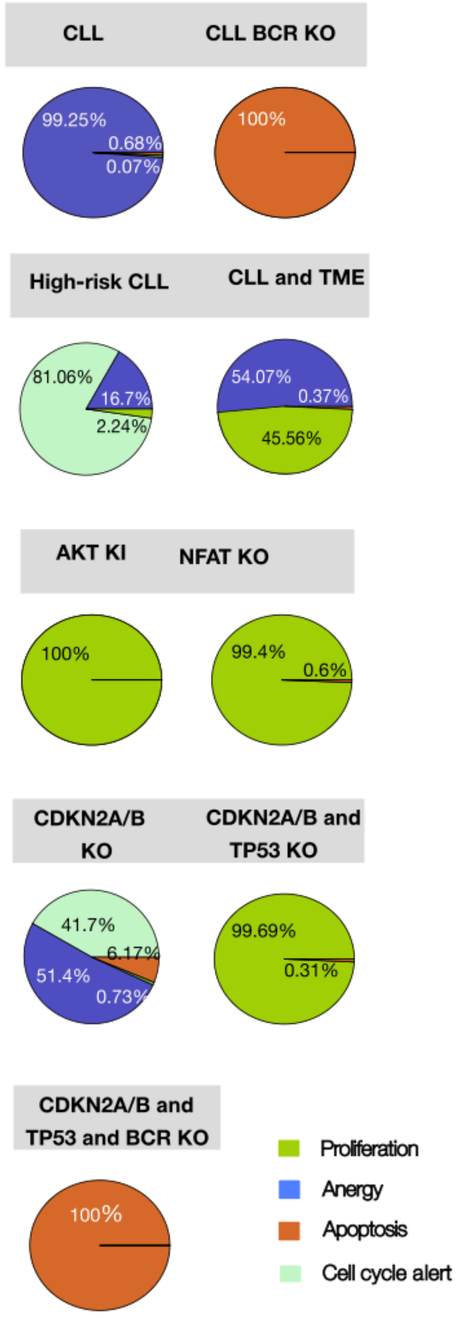
- **Result**: The model recovered known RS-like attractors and nominated BMI1 activation + TP53 loss as a transformation-driving combination; RS samples with TP53 lesions showed higher BMI1 expression

- **Conclusion**: Traditional systems bio modeling is great; This makes me dream up a future where agents can go out, read the literature, and build these models

Transform statements form 228 papers to boolean functions

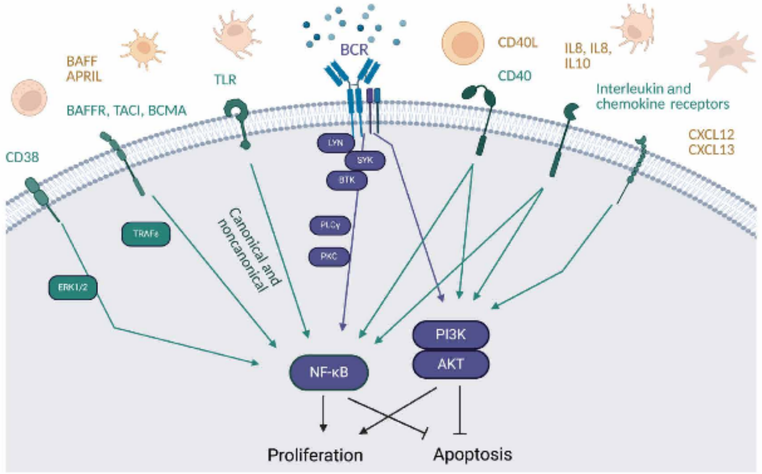


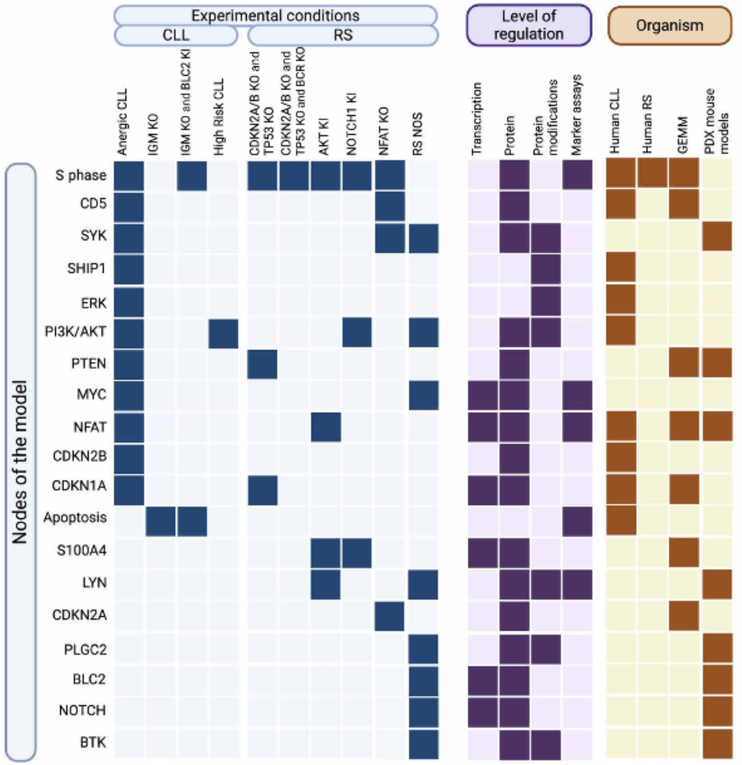
Use the boolean functions to do simulation



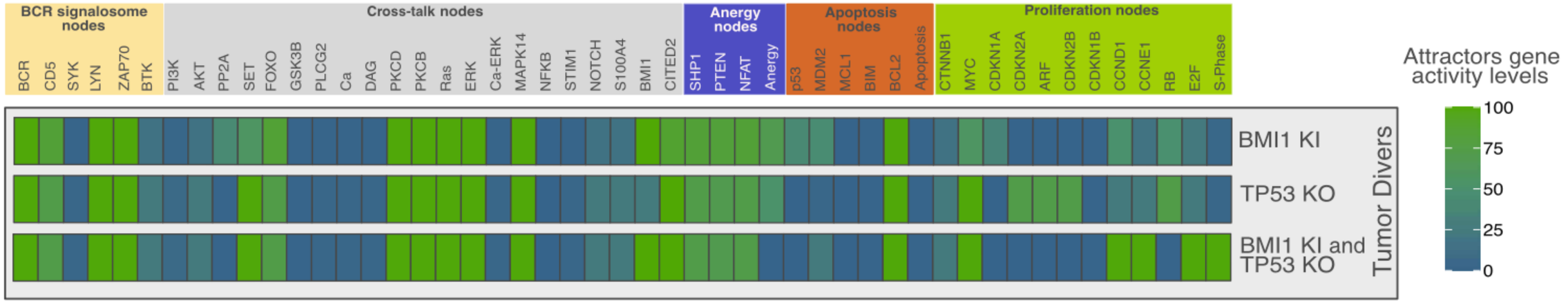
Produces reasonable simulation results

Use model to make prediction that BMI1 knock in and TP53 KO will lead to prliferation

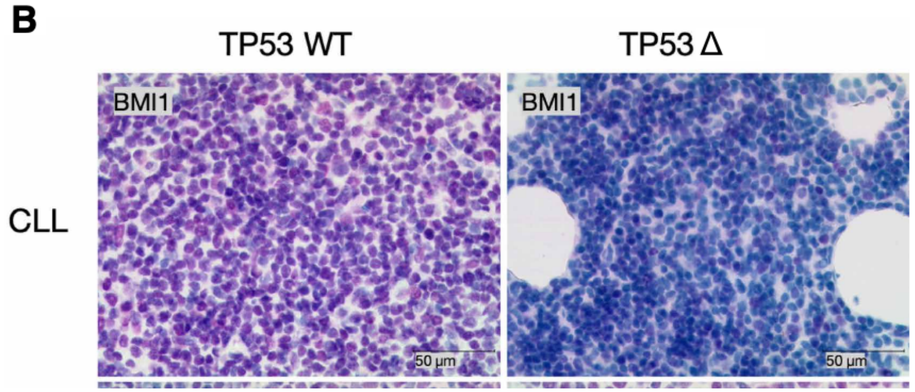




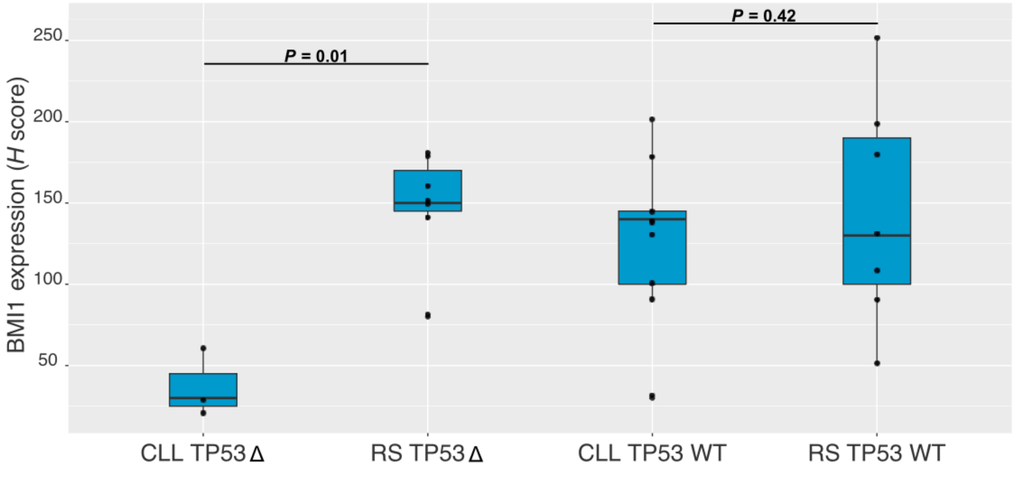
This is their simulation output



Stains show relationship between BMI1 knock in and TP53 KO



And in patient samples with CLL and RS



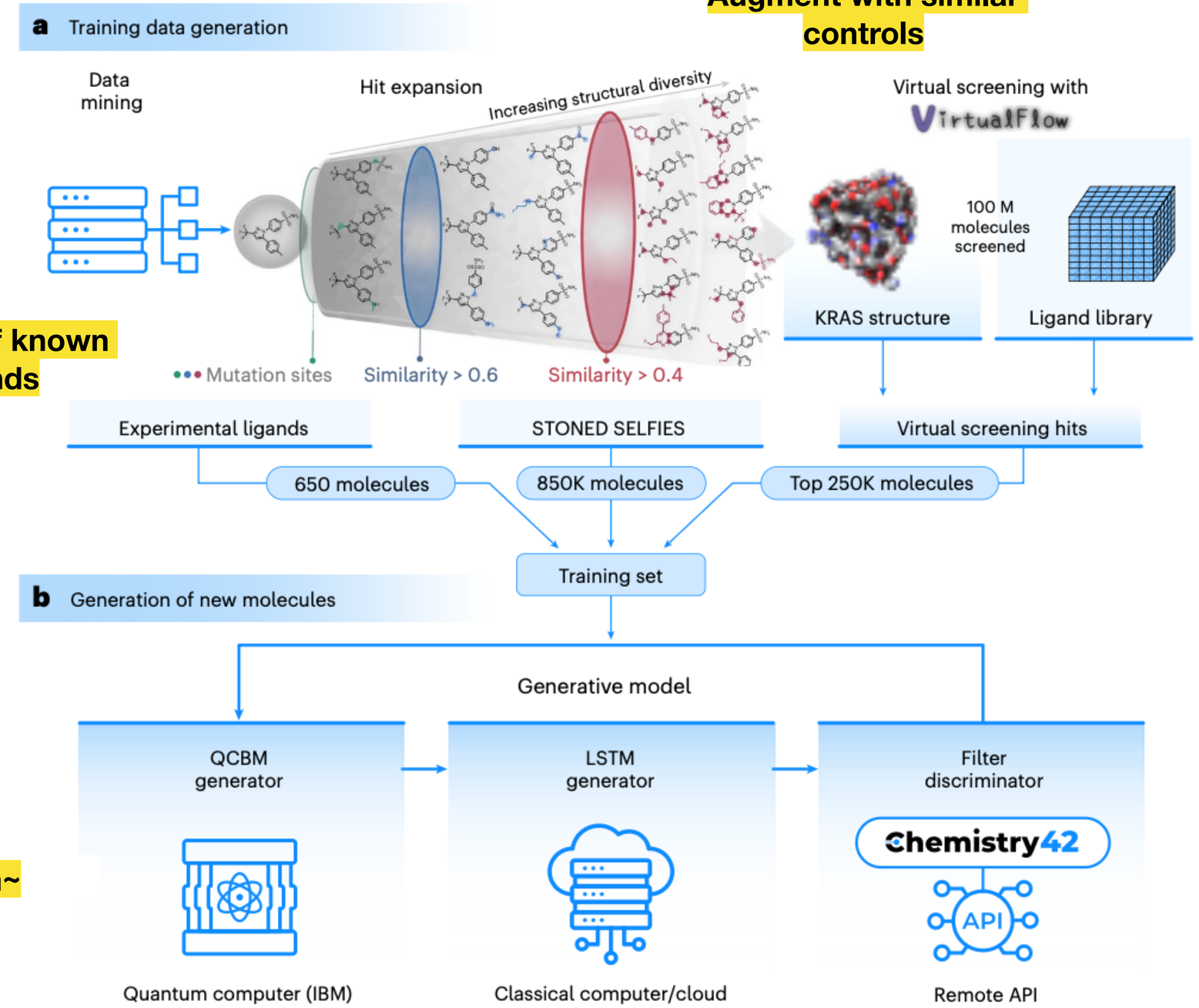
Quantum-computing-enhanced algorithm unveils potential KRAS inhibitors (Vakili, Gorgulla, Snider, et al., *Nature Biotechnology*)

- **Goal**: Test whether a hybrid quantum–classical generative model can design small molecules against KRAS, a historically difficult oncology target
- **Method**: Train an LSTM + quantum circuit Born machine on known KRAS inhibitors, SELFIES-expanded analogs, and virtual-screening hits; filter/rank generated molecules with molecular docking
- **Result**: Generated 1M candidate molecules per model class, synthesized 15, and found two compounds with micromolar KRAS/RAS activity in SPR and cell-based assays
- **Conclusion**: A real experimental foothold for quantum-enhanced drug discovery — exciting, no doubt, but still more about feasibility than superiority

Augment with similar controls

Take a bunch of known KRAS ligands

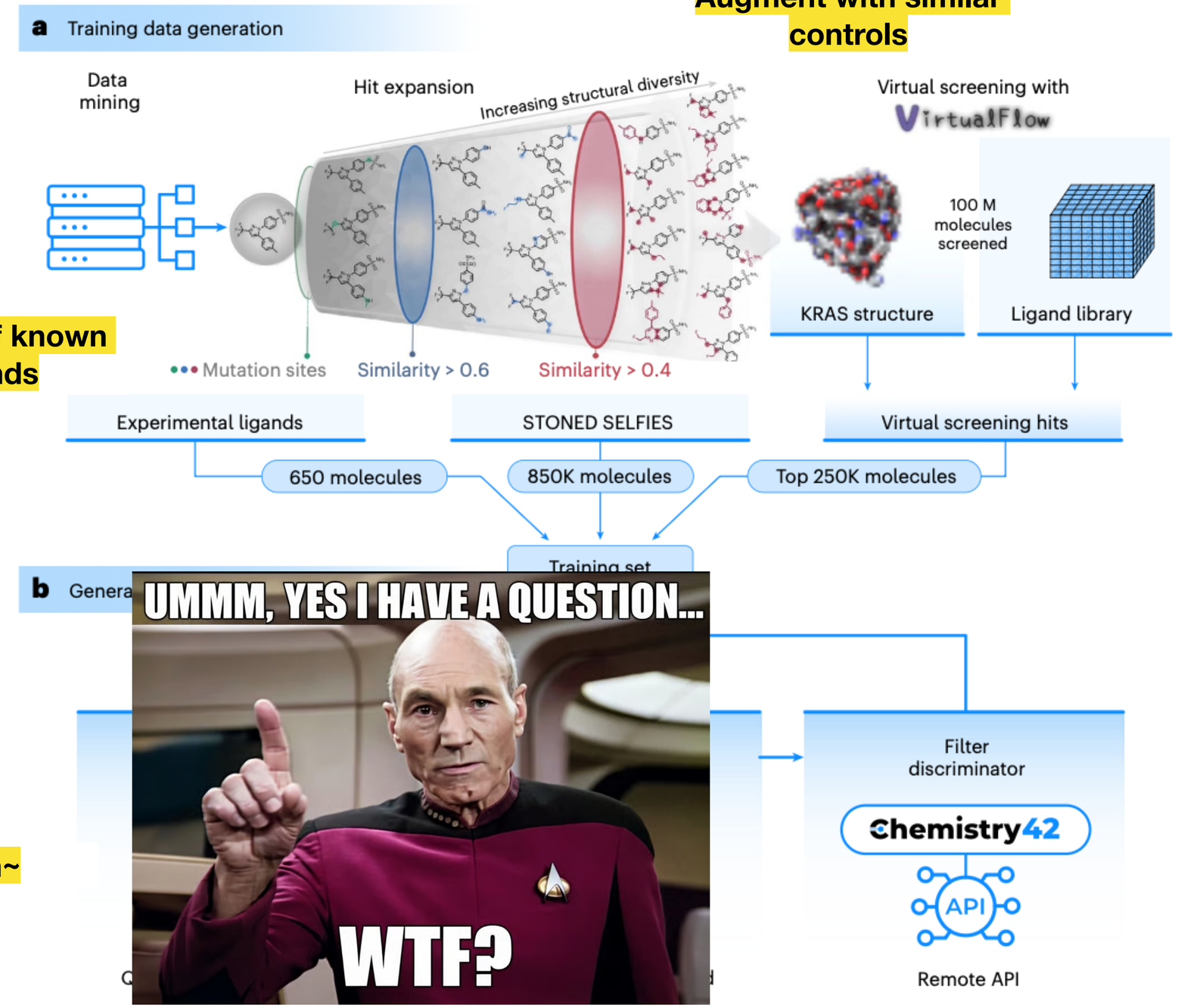
Do ~Quantum~



Augment with similar controls

Take a bunch of known KRAS ligands

Do ~Quantum~



okay eli5 what exactly the QCBM does

The QCBM is basically a quantum **random-number generator** that has been **trained to prefer "good" random numbers**.

why is this better for generating a distribution then learning an optimal distribution to generate on a classical compute architecture?

The authors' argument is:

- **Molecular generation is a distribution-learning problem**
- **QCBMs may represent some complicated distributions compactly making sampling more efficient**
- **That prior then biases the classical molecule generator**
- **Empirically, in their benchmark, the quantum prior helped —** improved success rate by **21.5%**

and why does a quantum computer navigate this high d space more efficiently?

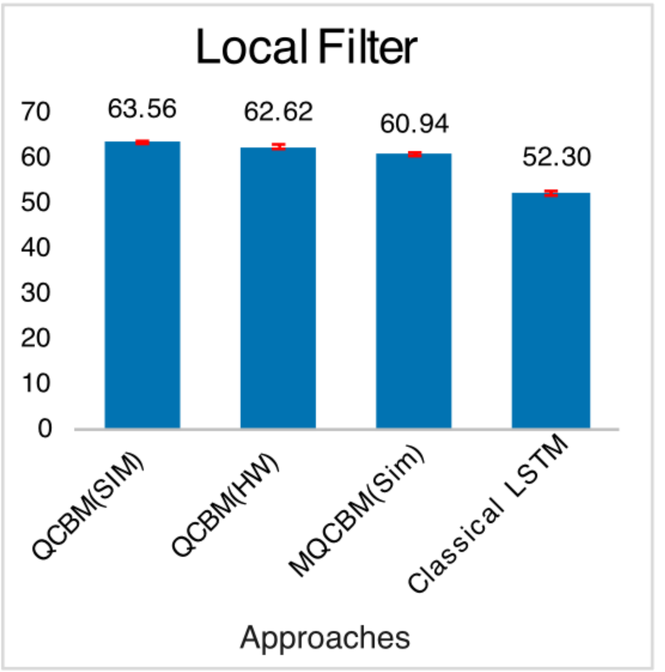
Classically, to model a complicated joint distribution over many binary variables, you often need to learn lots of pairwise and higher-order dependencies. If variable 1 matters only in combination with variables 7, 12, and 15, and that pattern changes depending on variable 3, the model has to represent a lot of conditional structure. Entangled quantum states naturally represent correlated variables. So the sales pitch is: **some correlated high-dimensional probability distributions may be represented with fewer parameters or sampled more naturally by a quantum circuit than by a simple classical model**.

-ChatGPT May 18th, 2026

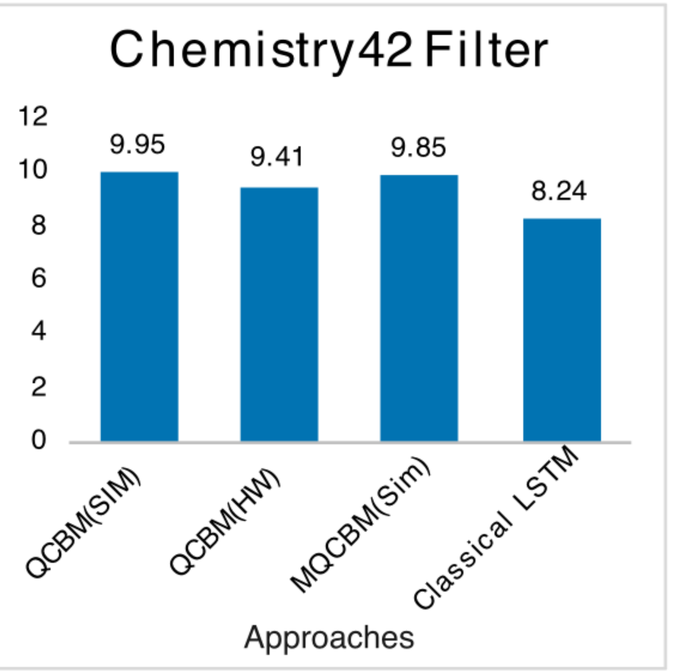
#IS25

Clasical vs. Quantum Modelling

(A)

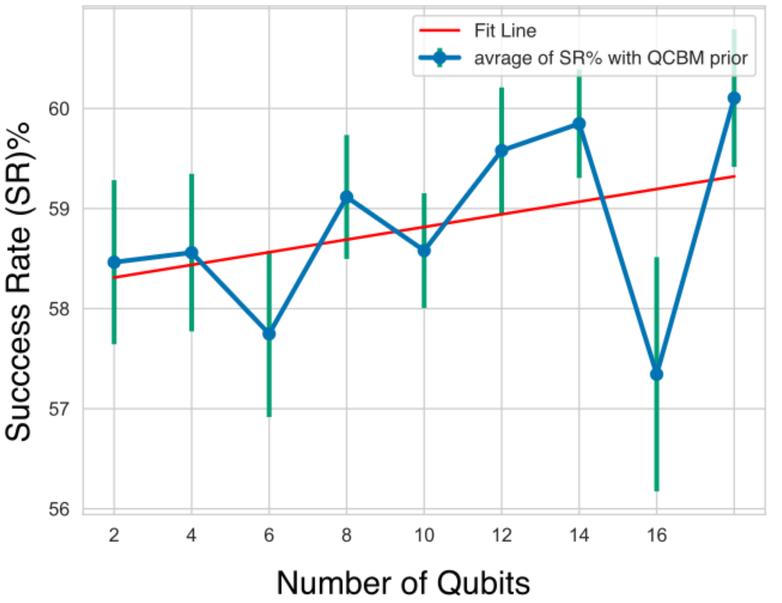


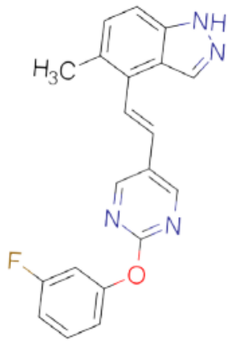
(B)



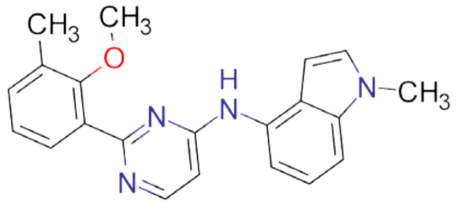
Quantum was more efficient!

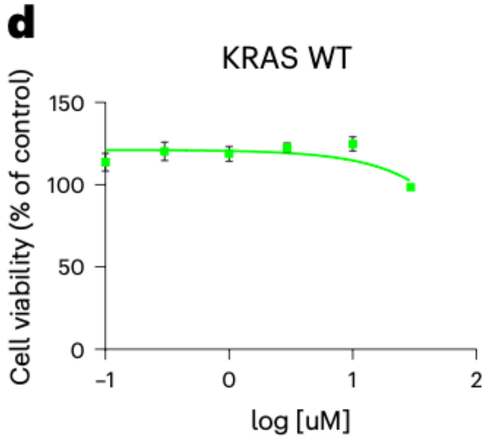
(C)



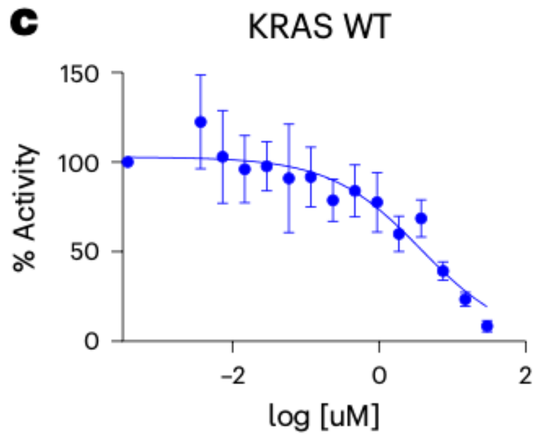


Synthesized the predicted molecules

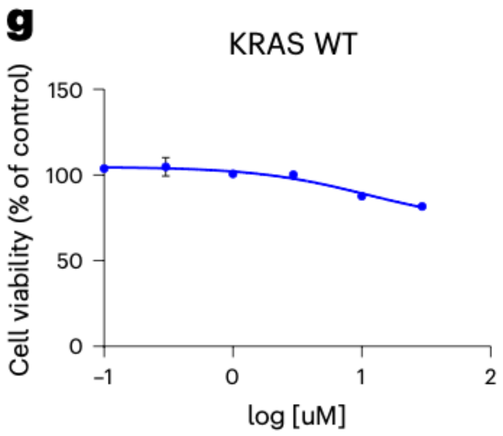


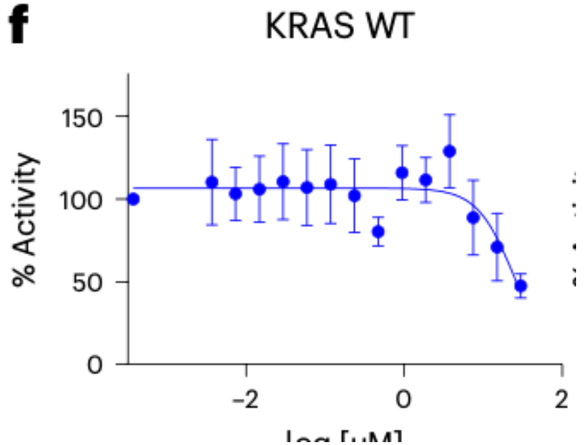


They don't totally kill the cells



They do inhibit KRAS

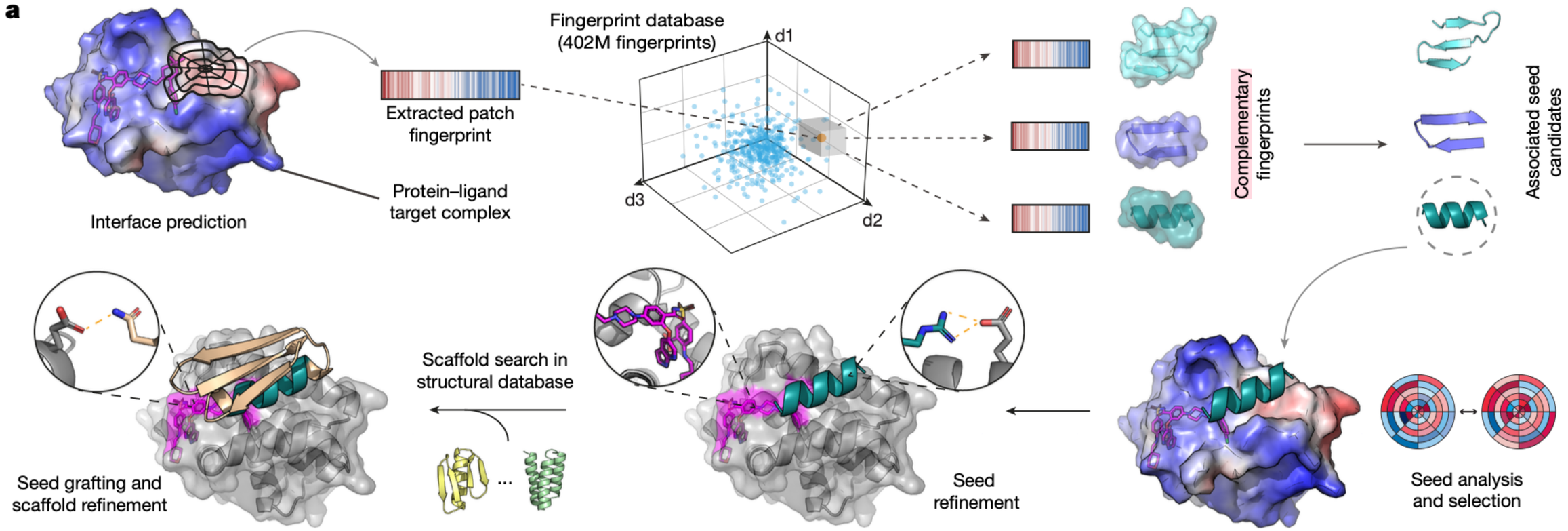




Targeting protein–ligand neosurfaces with a generalizable deep learning tool (Marchand, Buckley, Schneuing, et al., *Nature*)

- **Goal**: Design proteins that bind only when a target protein is complexed with a small molecule - chemically induced protein interactions as programmable ON-switches
- **Method**: Extend MaSIF surface fingerprints to protein–ligand "neosurfaces," then search 402M complementary surface patches, refine/graft binding seeds, and redesign interfaces with Rosetta
- **Result**: Designed binders against Bcl2–venetoclax, DB3–progesterone, and PDF1–actinonin with ligand-dependent binding, high specificity, mutational support, and structural/biophysical validation
- **Conclusion**: Opens up a new class of drug-controllable biologics, with the potential for extreme molecular specificity and safer therapeutic control

MaSIF learns embeddings such that binding partners are close in space

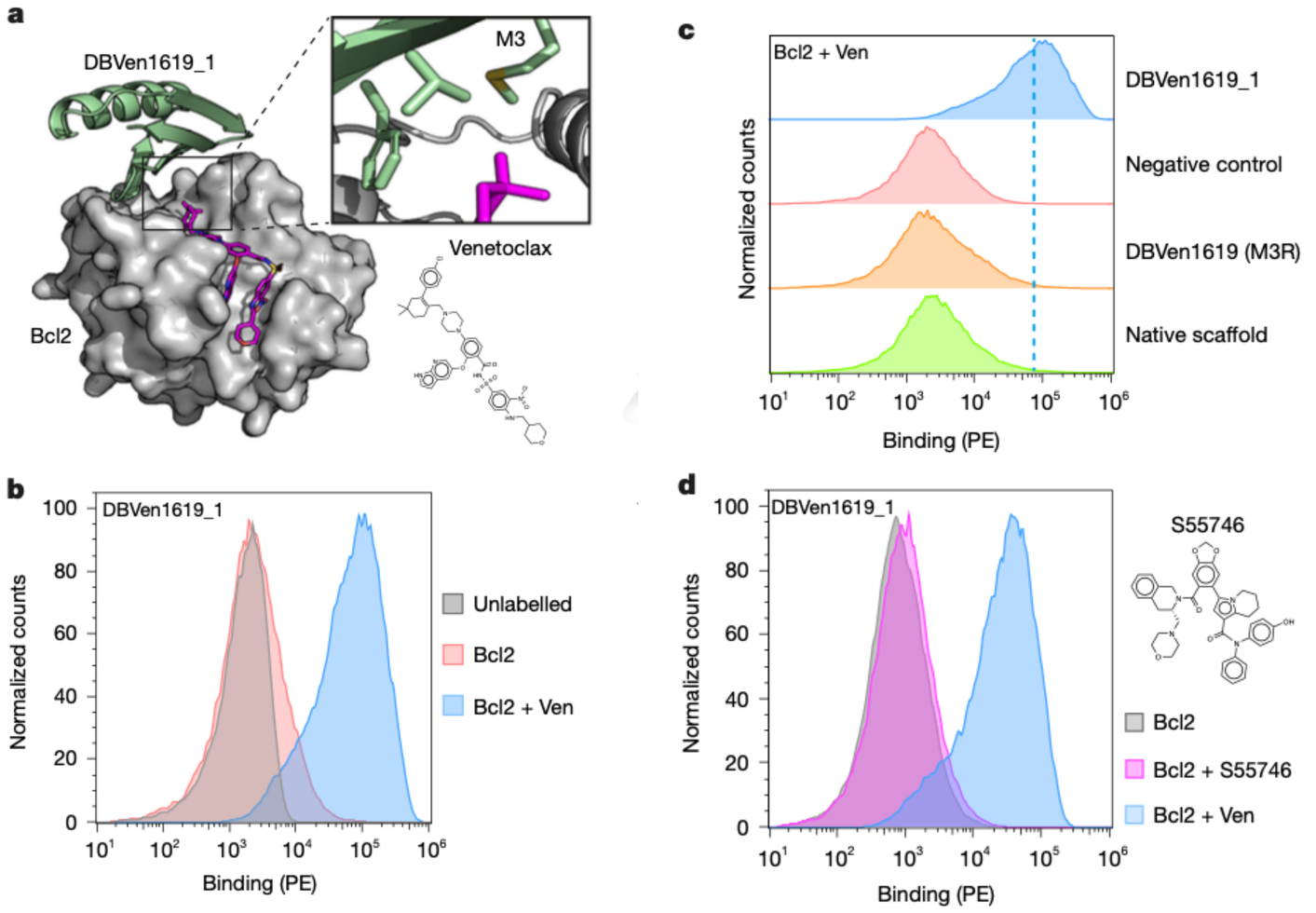


Scaffold is identified and seed is grafted on

Seeds are peptides with the right fingerprint

Experimental validation of small-molecule dependent binders

Binds only when small molecule is present



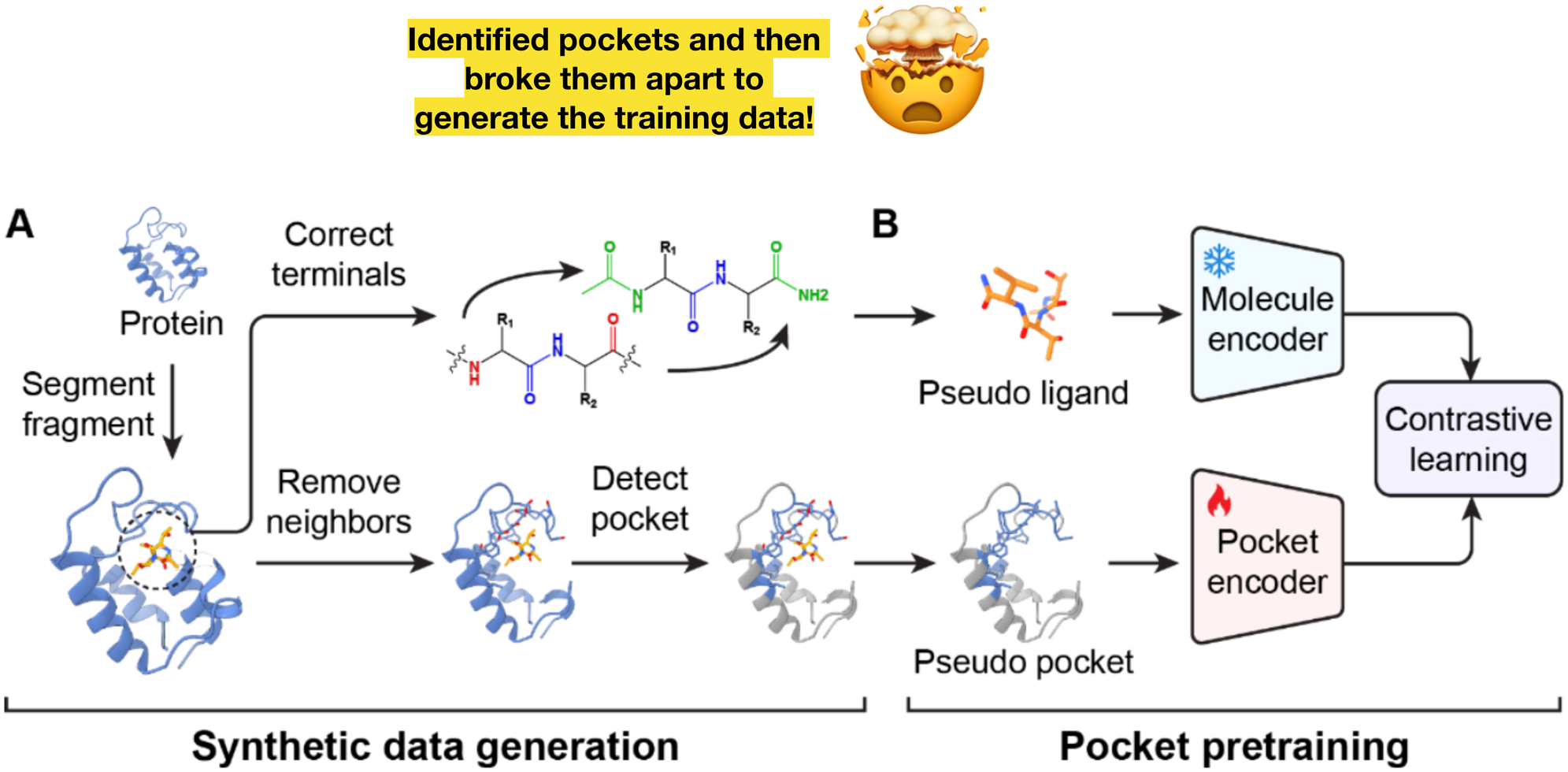
...and not when another small molecule is there

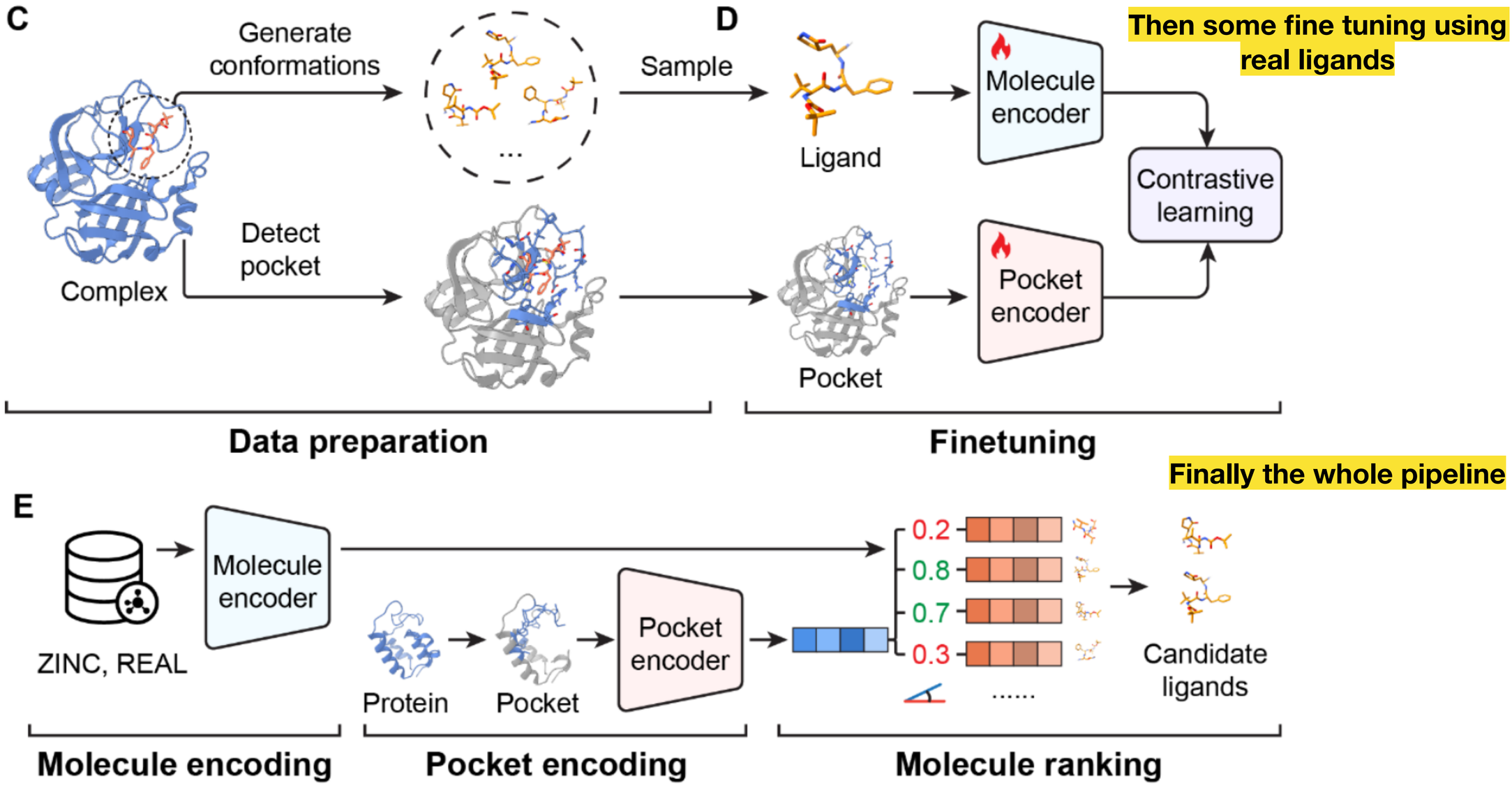
Shift right—> binding!

Deep contrastive learning enables genome-wide virtual screening (Jia, Gao, Tan et al, *Science*)

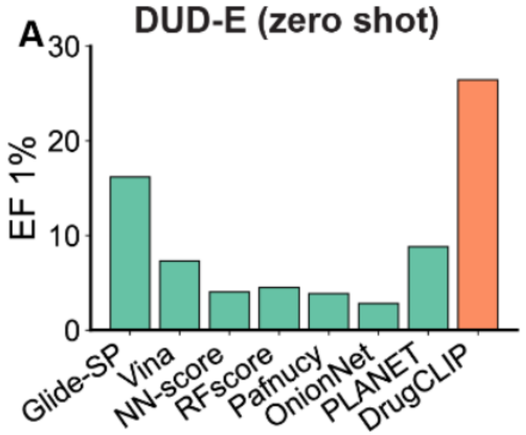
- **Goal**: Make genome-wide virtual screening feasible by moving beyond target-by-target molecular docking
- **Method**: DrugCLIP learns aligned embeddings of protein pockets and small molecules using contrastive learning, turning virtual screening into dense retrieval
- **Result**: Outperformed docking and other ML methods, screened ~10,000 human proteins against 500M molecules, and experimentally validated hits for 5HT2AR, NET, and additional targets
- **Conclusion**: The post-AlphaFold drug discovery bottleneck may be shifting from structure prediction to scalable search across protein and chemical space

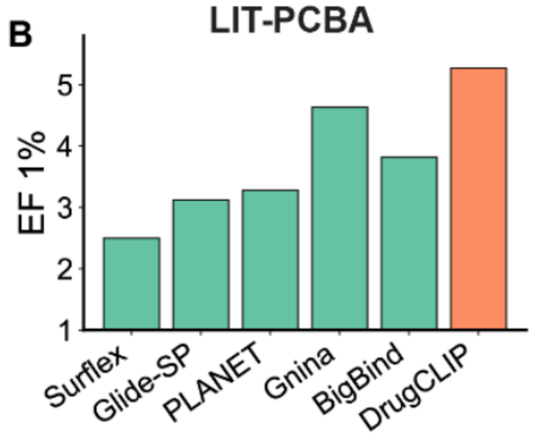
Identified pockets and then broke them apart to generate the training data!

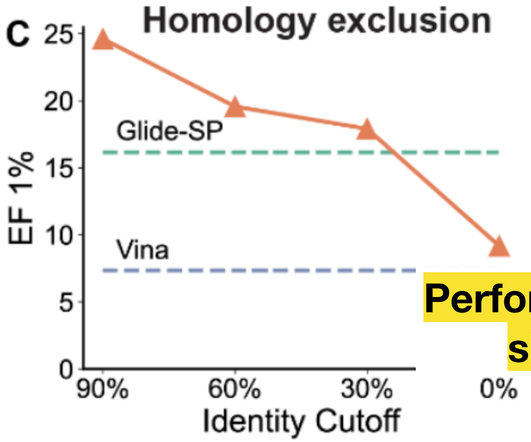



C
Generate conformations
Complex
Detect pocket
...
Data preparation
D
Sample
Ligand
Molecule encoder
Contrastive learning
Pocket
Pocket encoder
Finetuning
Then some fine tuning using real ligands
Finally the whole pipeline
E
ZINC, REAL
Molecule encoder
Protein
Pocket
Pocket encoder
0.2
0.8
0.7
0.3
.....
Candidate ligands
Molecule encoding
Pocket encoding
Molecule ranking

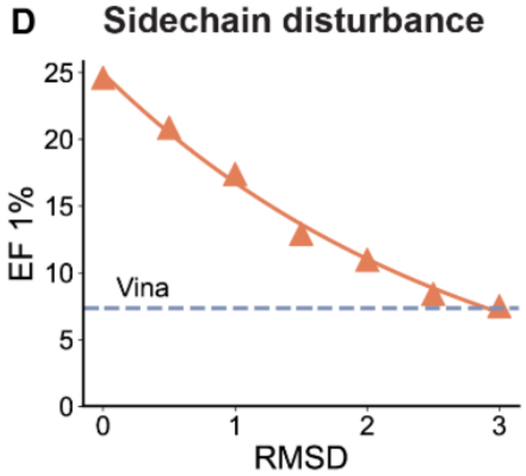
Best in class against two evaluation standards

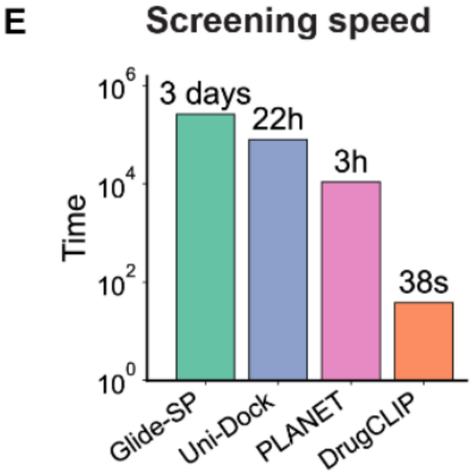


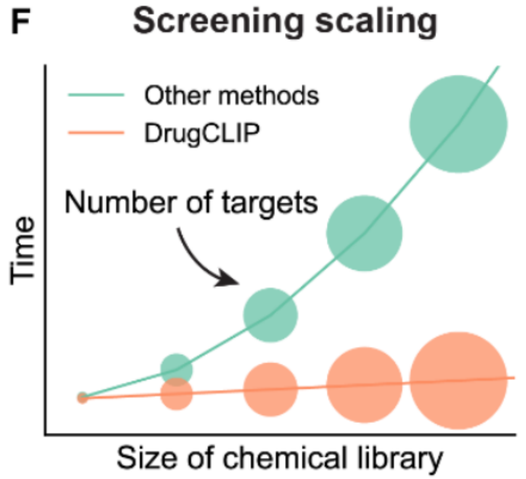




Performance drops when no similar targets exist





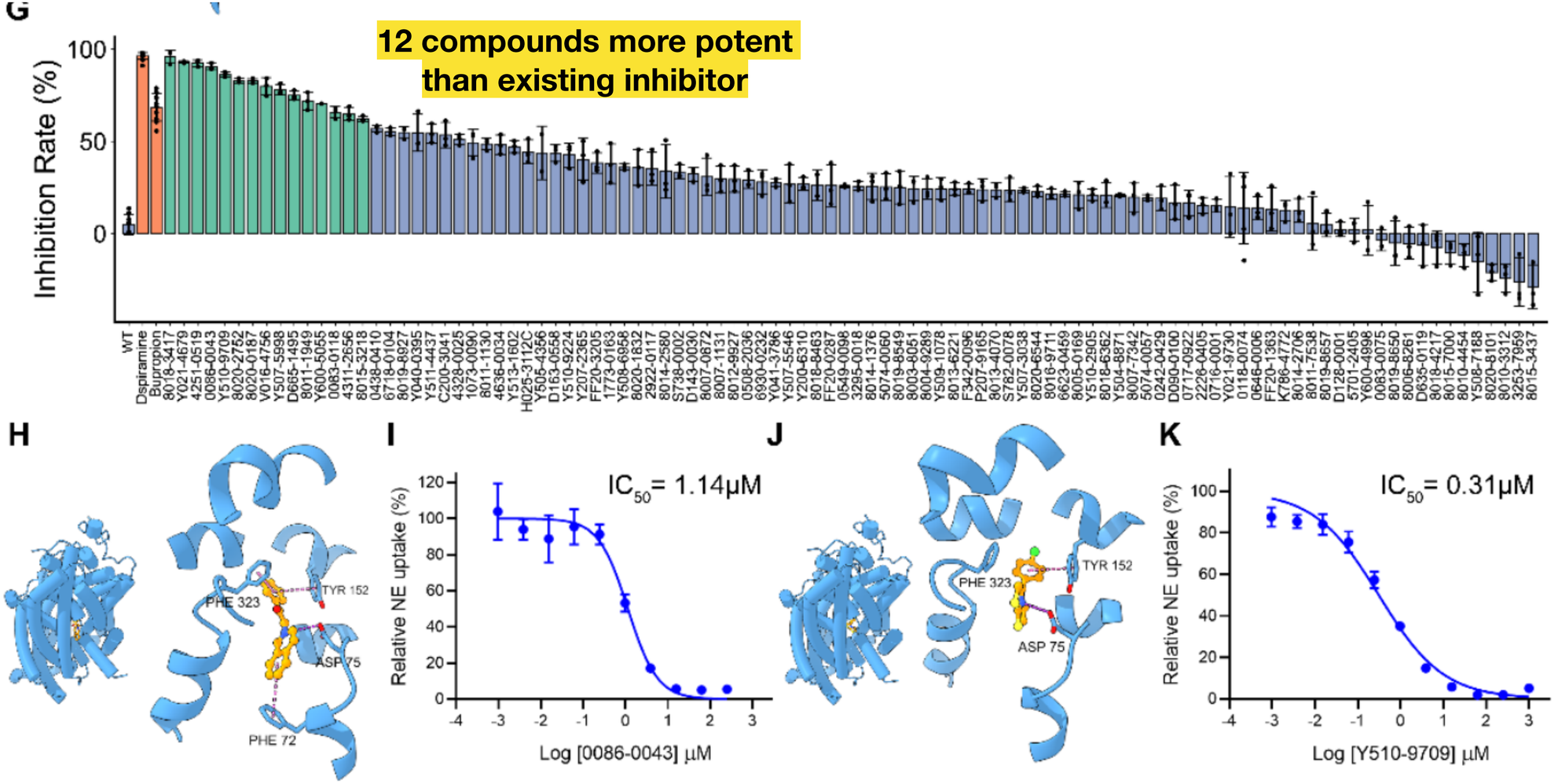


Performance drops when side chains are changed

So fast!

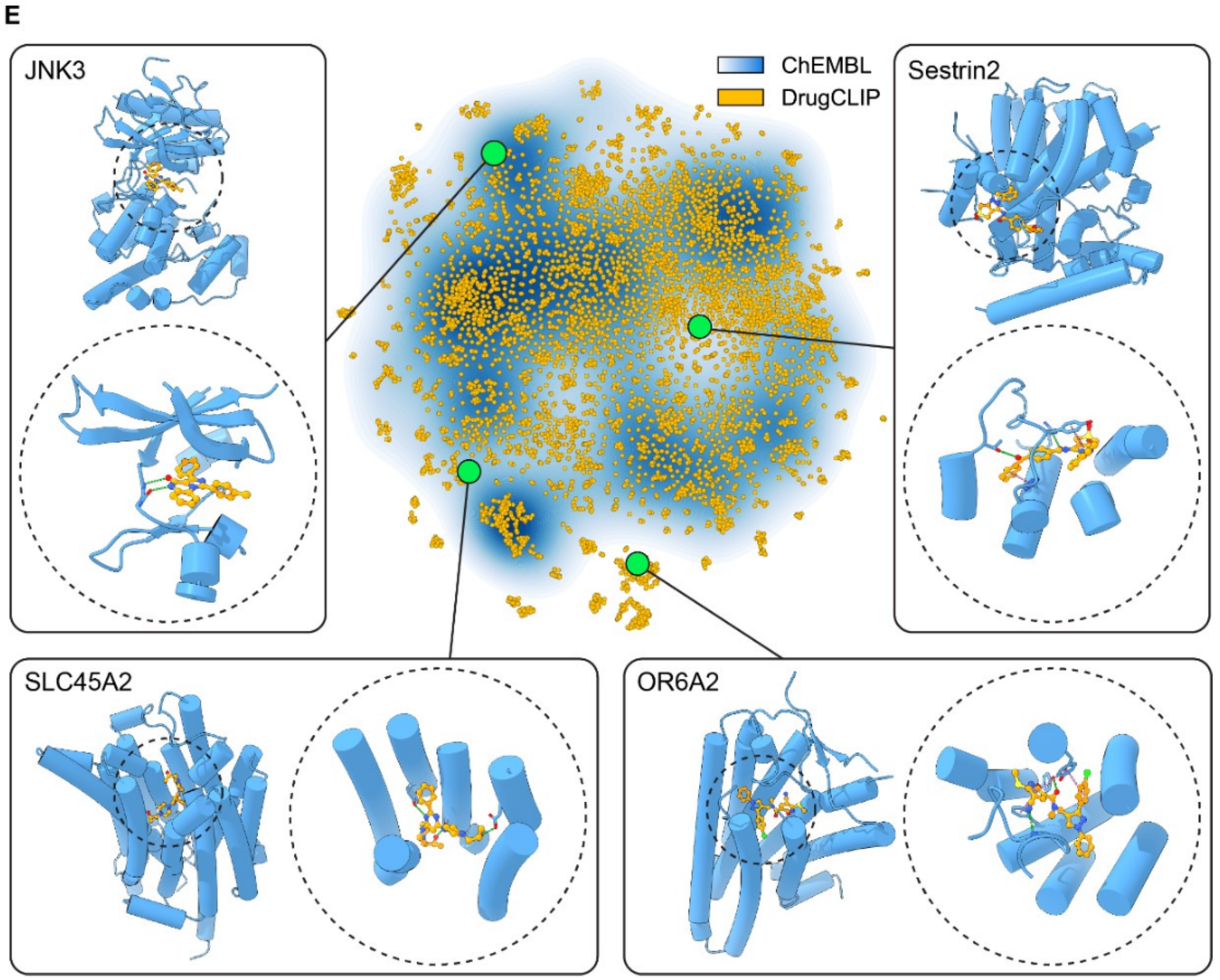
Predicting novel inhibitors of NET

12 compounds more potent than existing inhibitor



Two hits experimentally validated with IC50 and Cryo-EM structures

Virtual screen all the things!



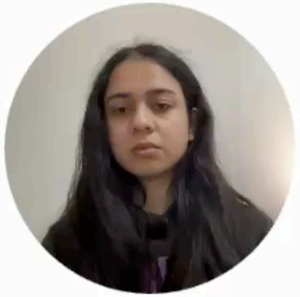
RESEARCH ARTICLE | DRUG DISCOVERY

Active learning framework leveraging transcriptomics identifies modulators of disease phenotypes

BENJAMIN DEMEO, CHARLOTTE NESBITT, SAMUEL A. MILLER, DANIEL B. BURKHARDT, INNA LIPCHINA, DORIS FU, PETER HOLDERRIETH, DAVID KIM, SERGEY KOLCHENKO, ARTUR SZALATA, ISHAN GUPTA, CHRISTINE KERR, THOMAS PFEFER, RAZIEL ROJAS-RODRIGUEZ, SUNIL KUPPASSANI, LAURENS KRUIDENIER, PARUL B. DOSHI, MAHDI ZAMANIGHOMI, JAMES J. COLLINS, ALEX K. SHALEK, FABIAN J. THEIS, AND MAURICIO CORTES

fewer Authors Info & Affiliations

SCIENCE • 23 Oct 2025 • Vol 390, Issue 6776 • DOI: 10.1126/science.adi8577



Harder, Better, Faster, Stronger - *Daft Punk*

New Engines for Molecular Reasoning

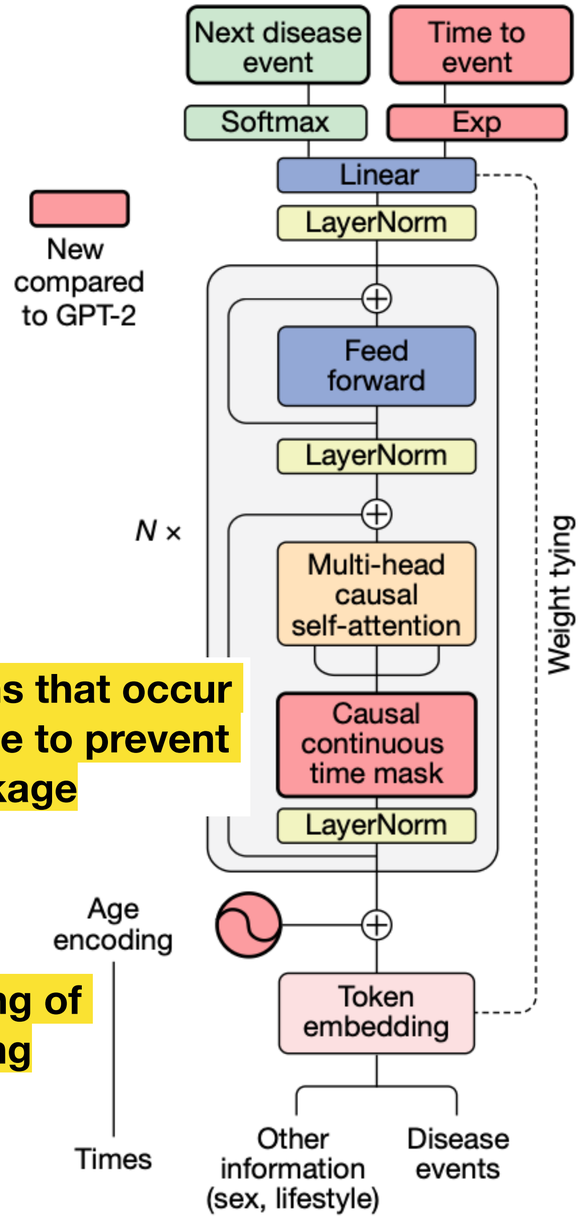
general-purpose models, agents, protein language models, and reusable biological AI infrastructure

 https://doi.org/10.1038/s41586-025-09529-3

Learning the natural history of human disease with generative transformers (Shmatko, Jung, Gaurav, et al., *Nature*)

- **Goal**: Model the full natural history of human disease, not one disease at a time, using longitudinal health records
- **Method**: Modify GPT-2 for continuous-time health trajectories with age encodings, disease/lifestyle tokens, and a time-to-next-event output head
- **Result**: Delphi-2M predicted rates for >1,000 diseases, generalized from UK Biobank to 1.9M Danish individuals, and sampled plausible future disease trajectories
- **Conclusion**: Health records can be treated like a generative language of disease progression — looking forward to how this can be linked more directly to molecular entities

Modified the architecture of GPT-2



Block tokens that occur at same time to prevent leakage

Replace positional encoding of tokens with age encoding

d

Input:
Age: Token
0.0: Male

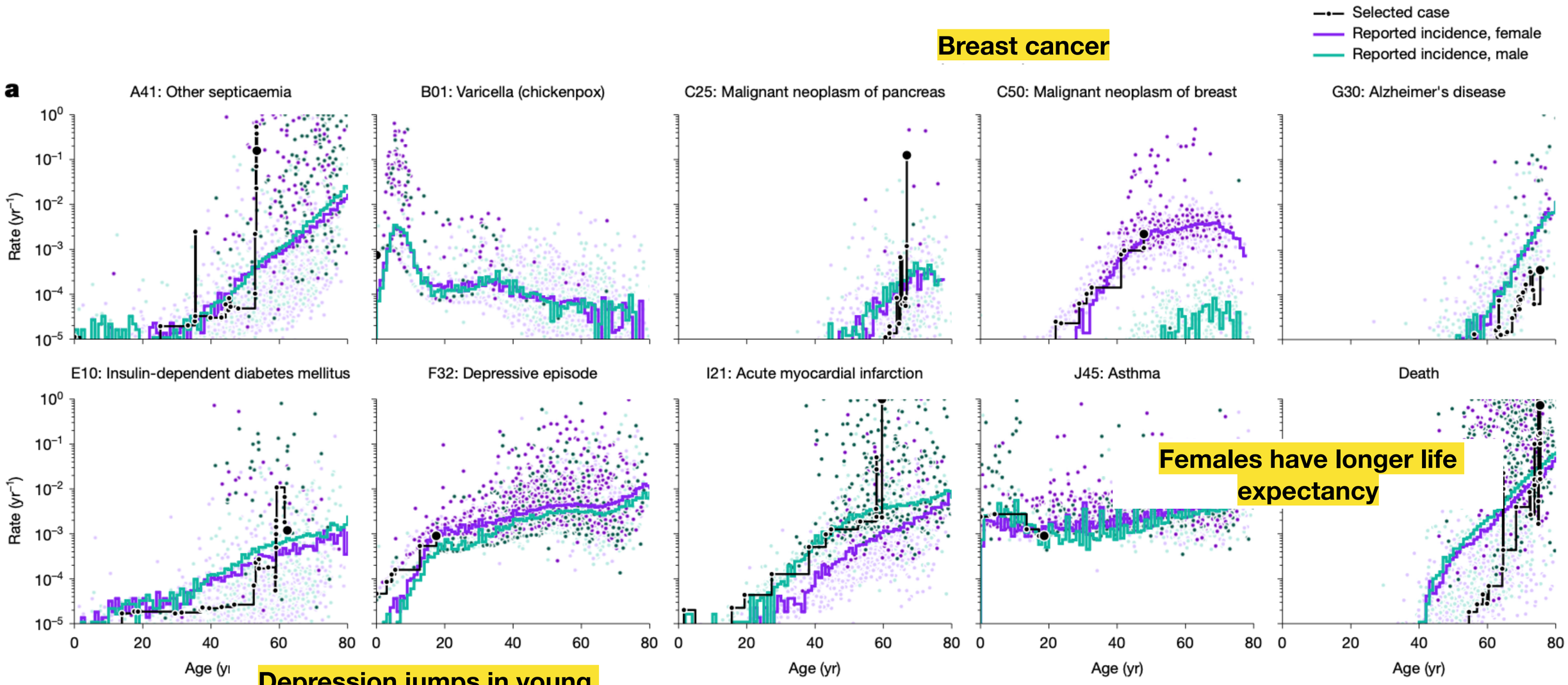
Replace next token with next token + time to token (event)

15.0: No event
20.0: No event
20.0: G43 migr
21.0: E73 lactose intolerance
22.0: B27 infectious mononucleosis
25.0: No event
28.0: J11 influenza, virus not identified
30.0: No event
35.0: No event
40.0: No event
41.0: Smoking low
41.0: BMI mid
41.0: Alcohol low
42.0: No event

**Transforms model output to disease rates over time

Output:
43.2: No event
43.5: M54 dorsalgia
44.6: I86 varicose veins of other sites
50.4: K52 other non-infective gastroenteritis and colitis
52.2: H83 other diseases of inner ear
53.9: J22 unspecified acute lower respiratory infection
54.5: L30 other dermatitis
55.3: No event
57.5: L50 urticaria
59.4: K62 other diseases of anus and rectum
...
69.8: J90 pleural effusion, not elsewhere classified
70.0: K21 gastro-oesophageal reflux disease
70.1: K76 other diseases of liver
70.3: I10 essential primary hypertension
70.4: M85 other disorders of bone density and structure
70.7: M81 osteoporosis without pathological fracture
71.2: J98 other respiratory disorders
72.1: J80 adult respiratory distress syndrome
72.2: No event
72.7: Death

Plotting model output (which is also disease rates)

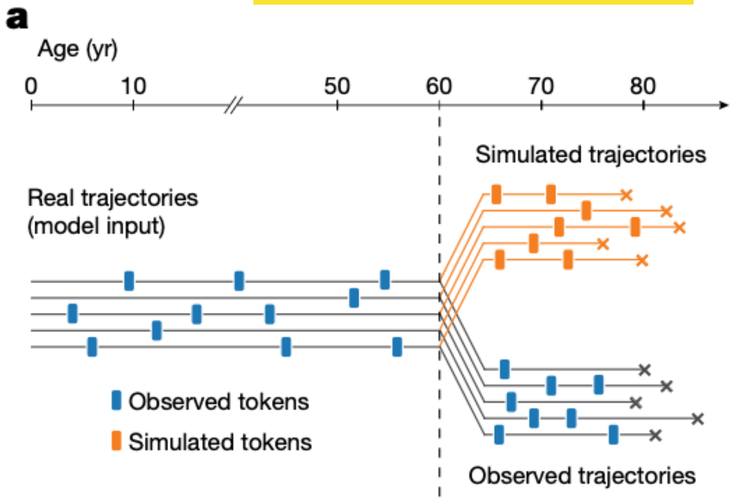


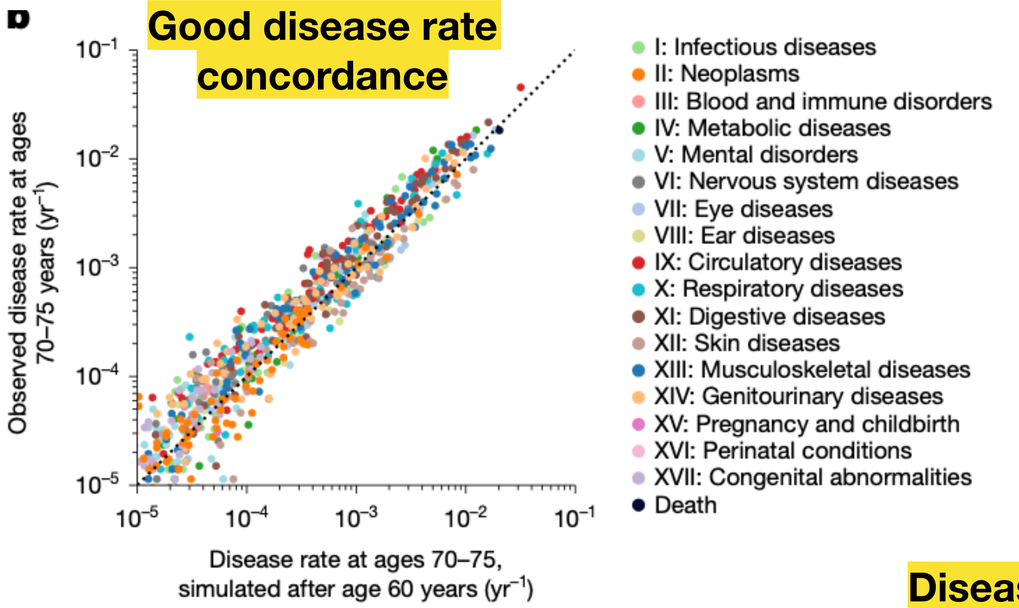
Breast cancer

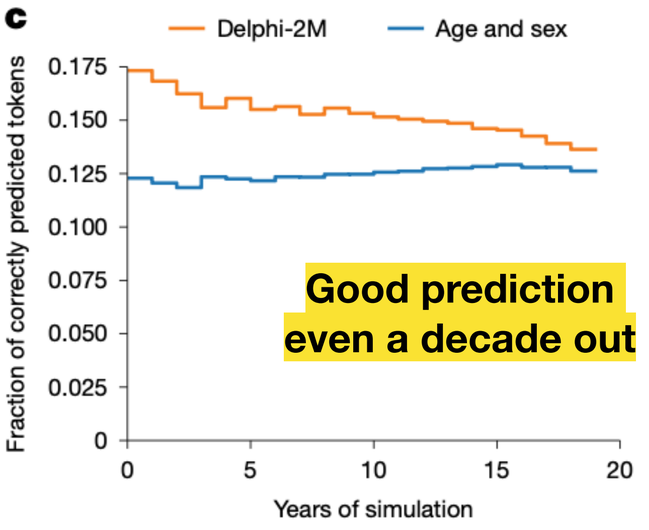
Females have longer life expectancy

Depression jumps in young adults and continues

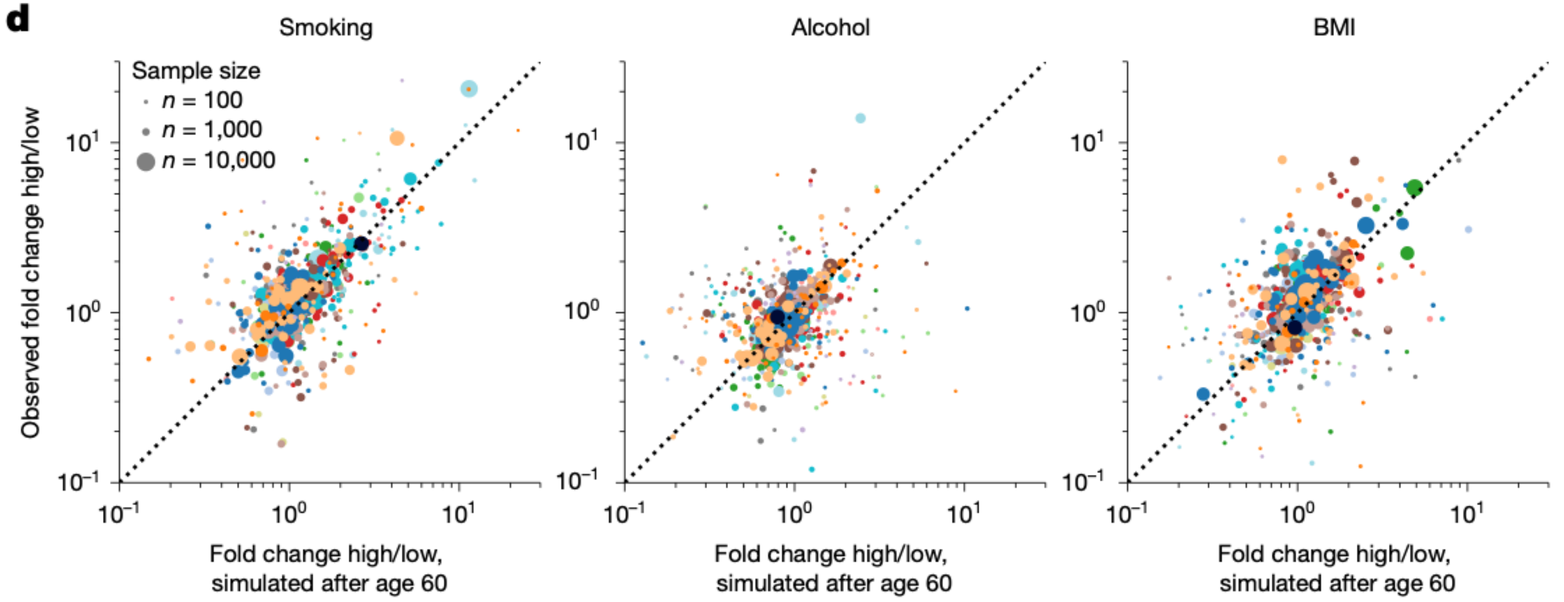
Simulate trajectories post 60 -> validate on real data

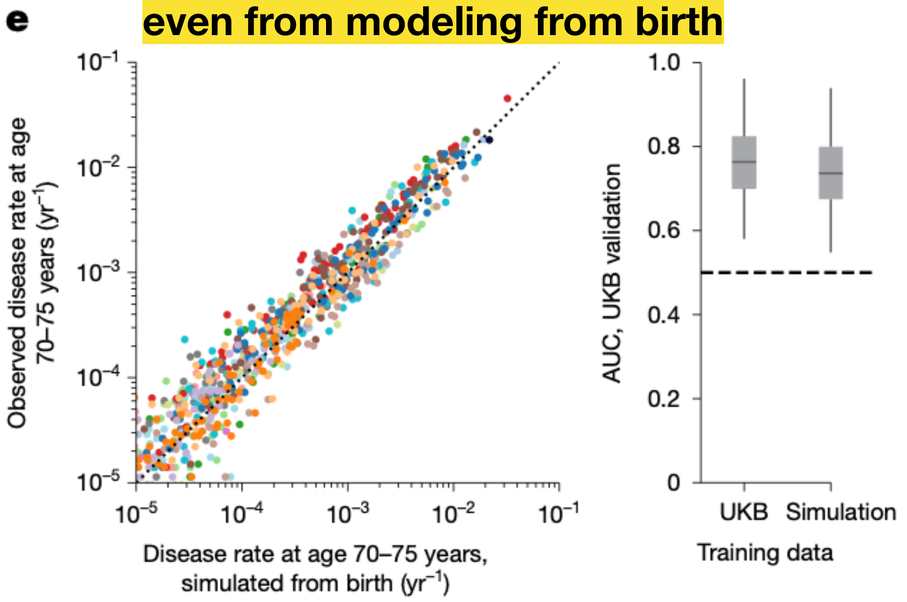




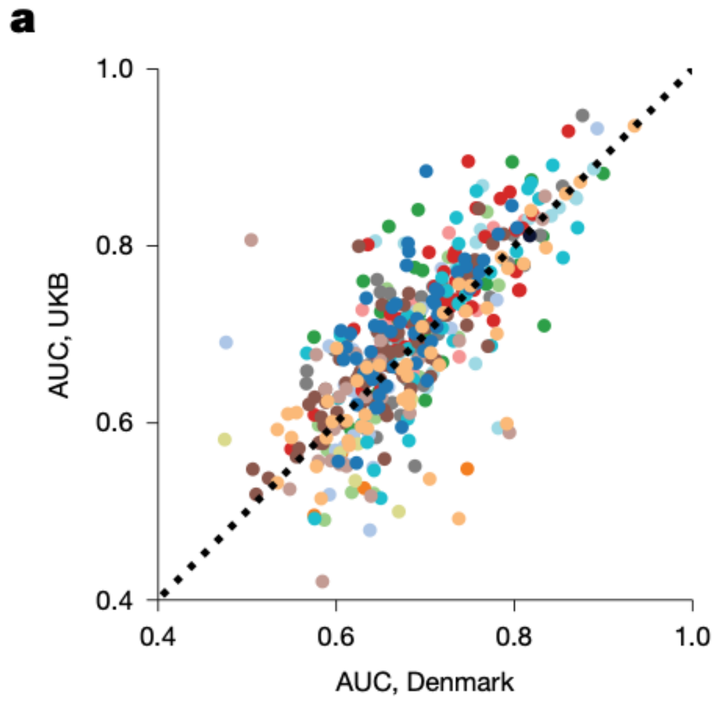


Disease rate concordance holds up even from modeling from birth

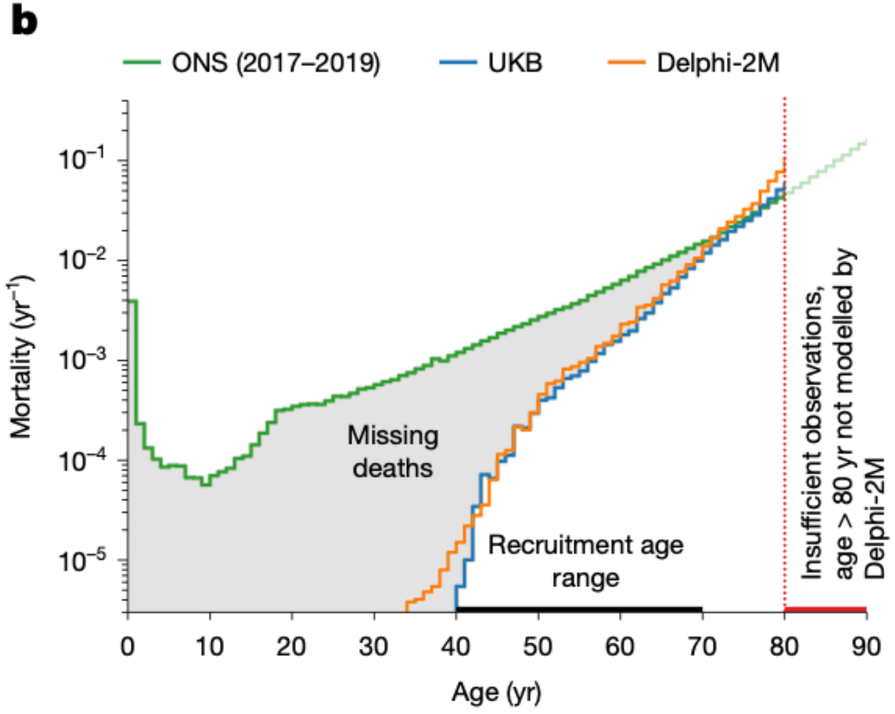




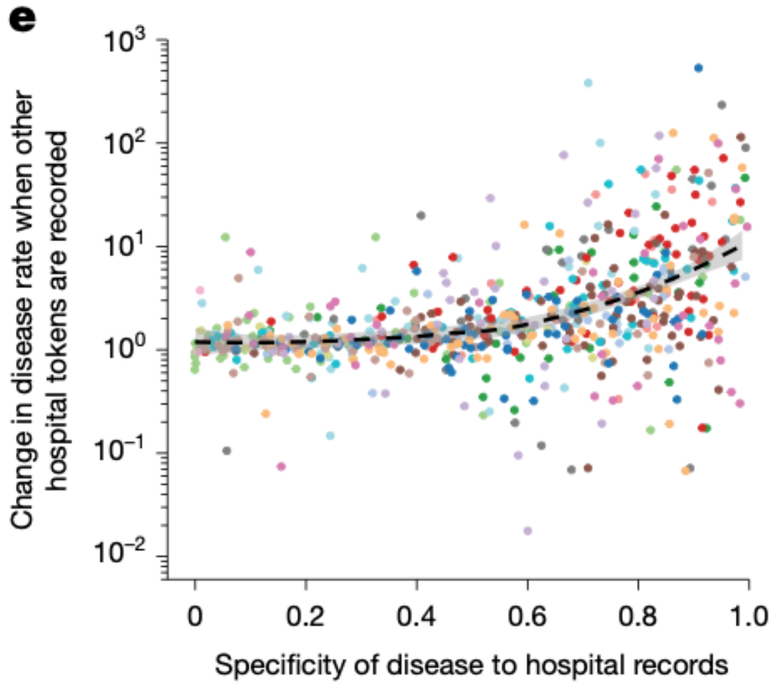
Validates against independent data



But can only learn from the data it's given — here it underestimates deaths for people under the age of 40



Disease rates increase after hospital token — learning health system grammar, not necessarily human biology

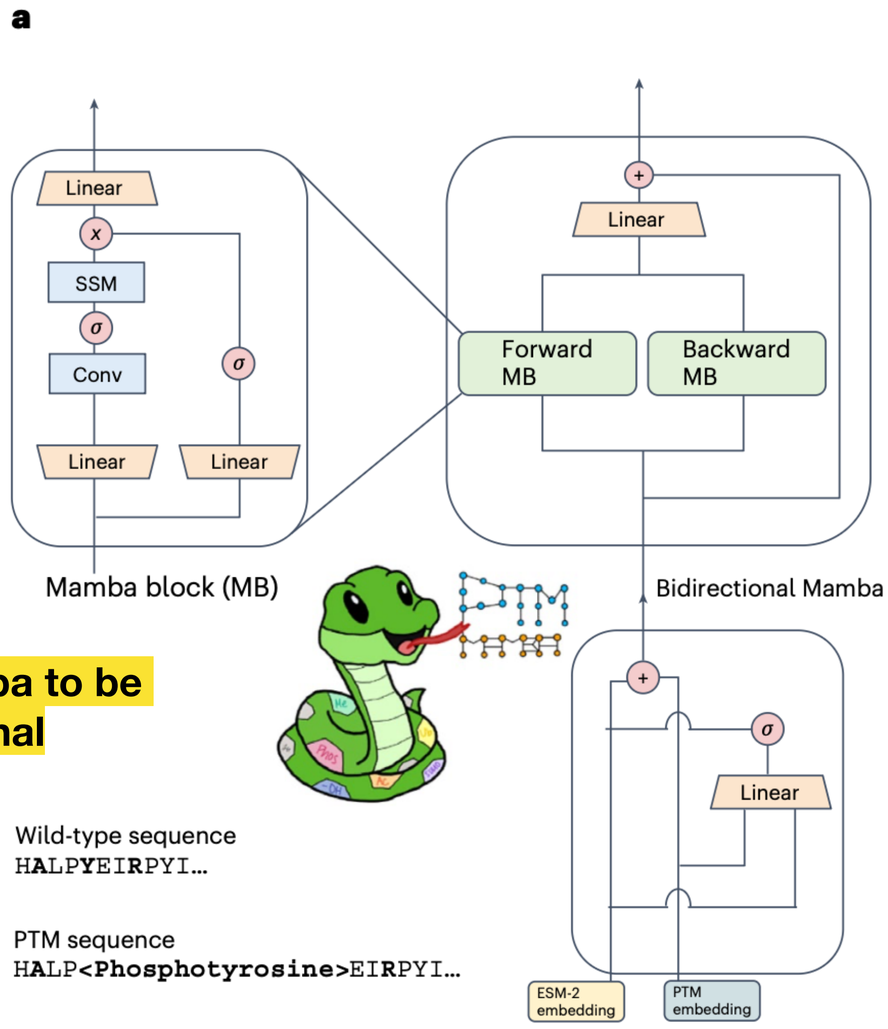


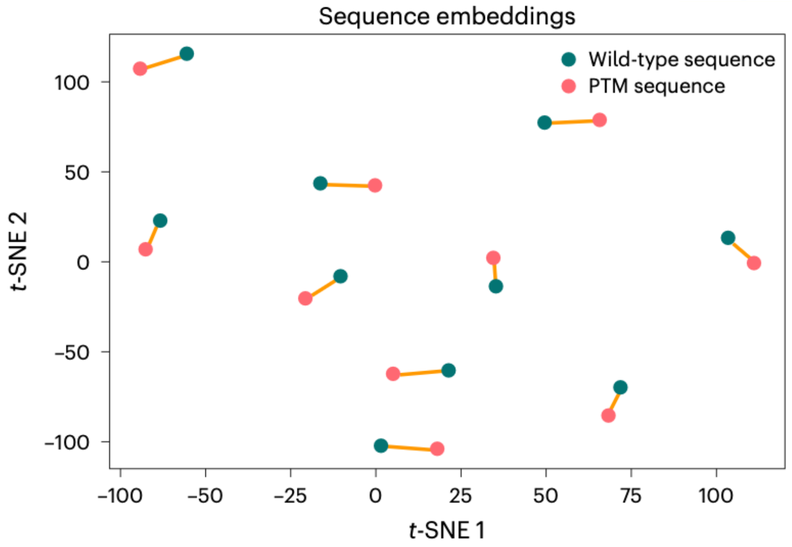
PTM-Mamba: a PTM-aware protein language model with bidirectional gated Mamba blocks (Peng, Wang, Chen et al, *Nature Methods*)

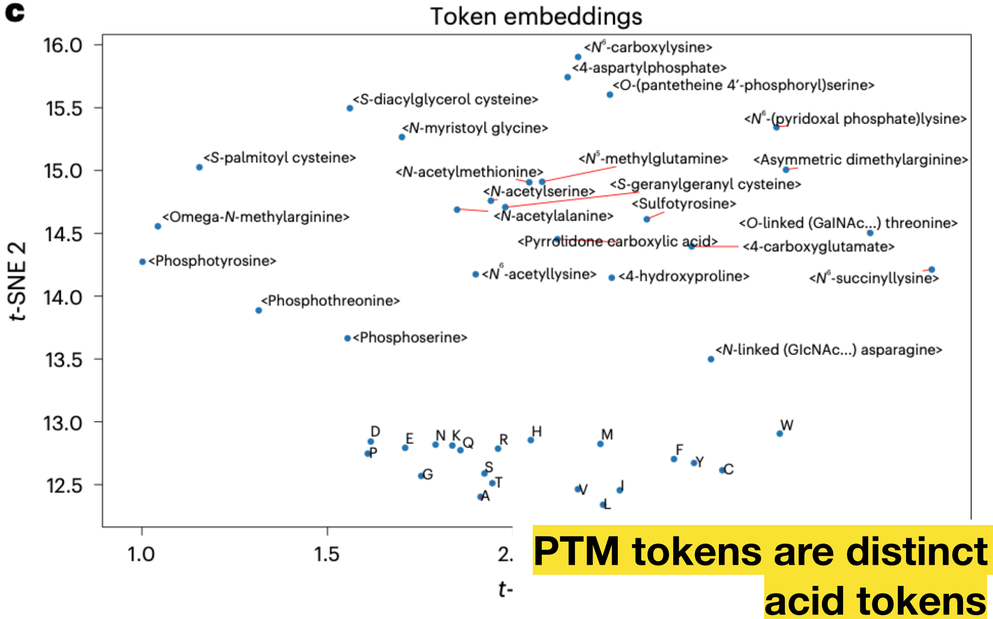
- **Goal**: Protein language models have learned a lot about proteins, but mostly ignore post-translational modifications - address this gap
- **Method**: Add explicit PTM tokens to protein sequences and combine ESM-2 embeddings with a bidirectional Mamba model using a gated fusion mechanism
- **Result**: PTM-Mamba improves PTM-aware tasks including disease association, druggability, PTM effects on PPIs, and zero-shot PTM prediction
- **Conclusion**: Finding ways to jam more data into our protein foundation models can only be good, right?

PTM sequences are close in embedding space to non PTM sequences

Modified Mamba to be bidirectional

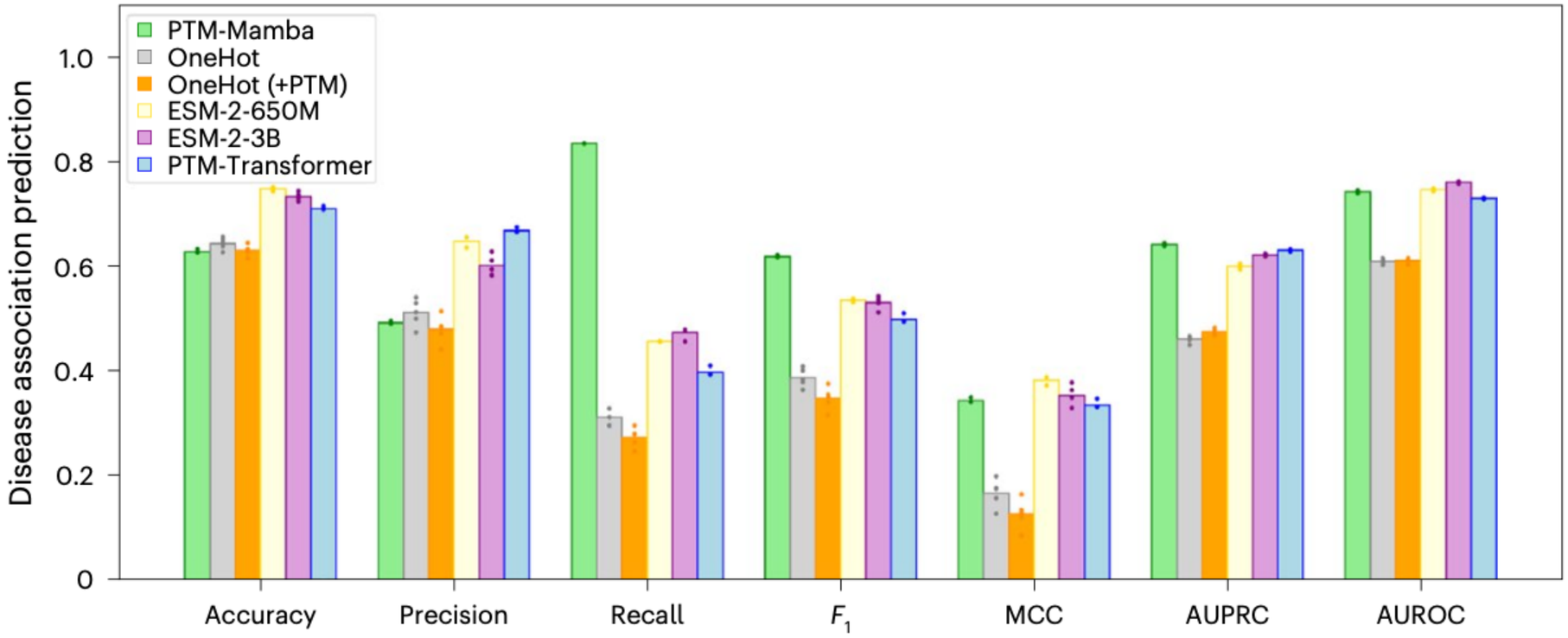




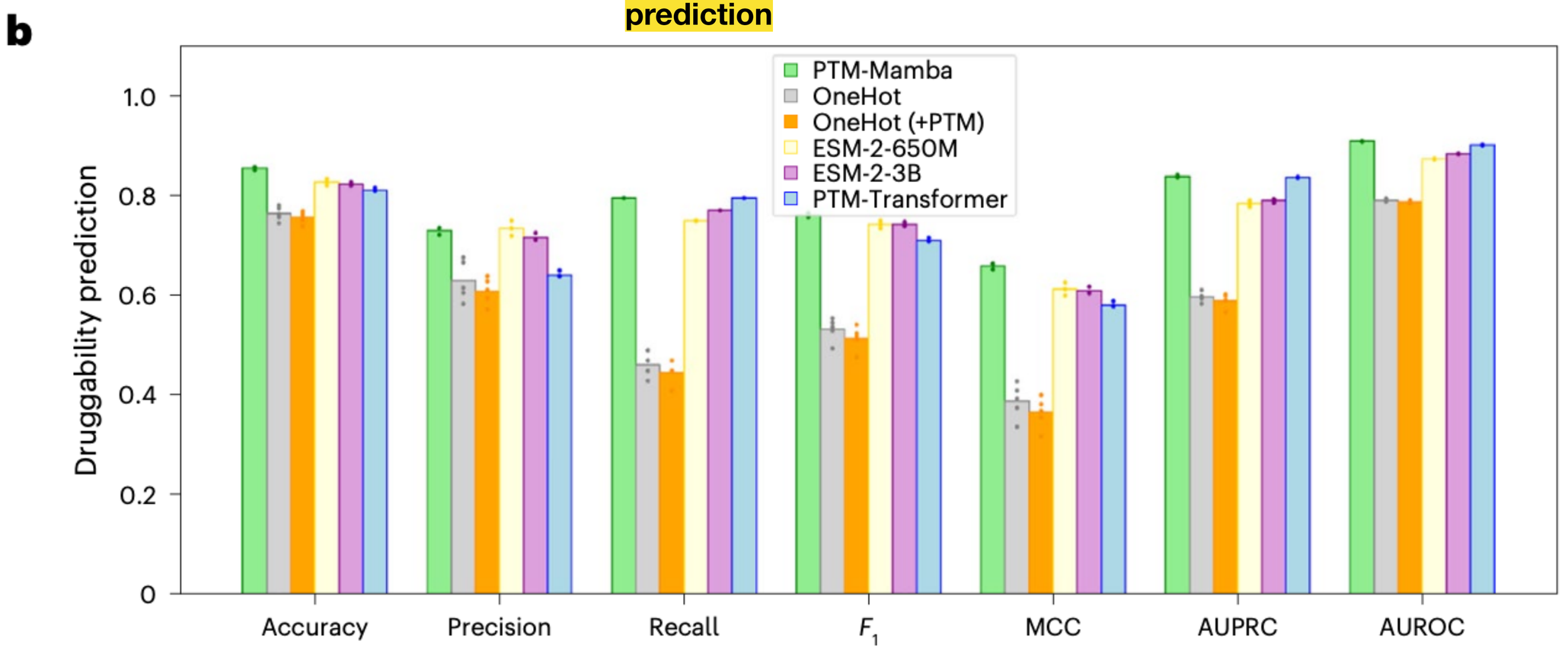


PTM tokens are distinct from amino acid tokens

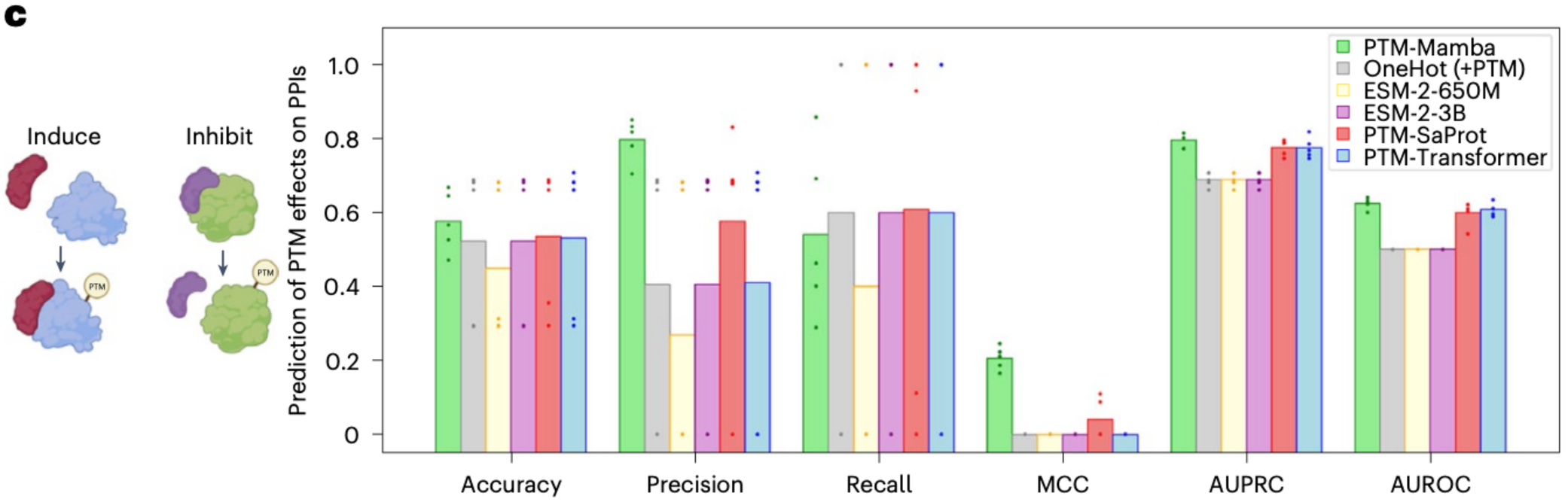
Improvement in disease-association prediction



Improvement in druggability prediction

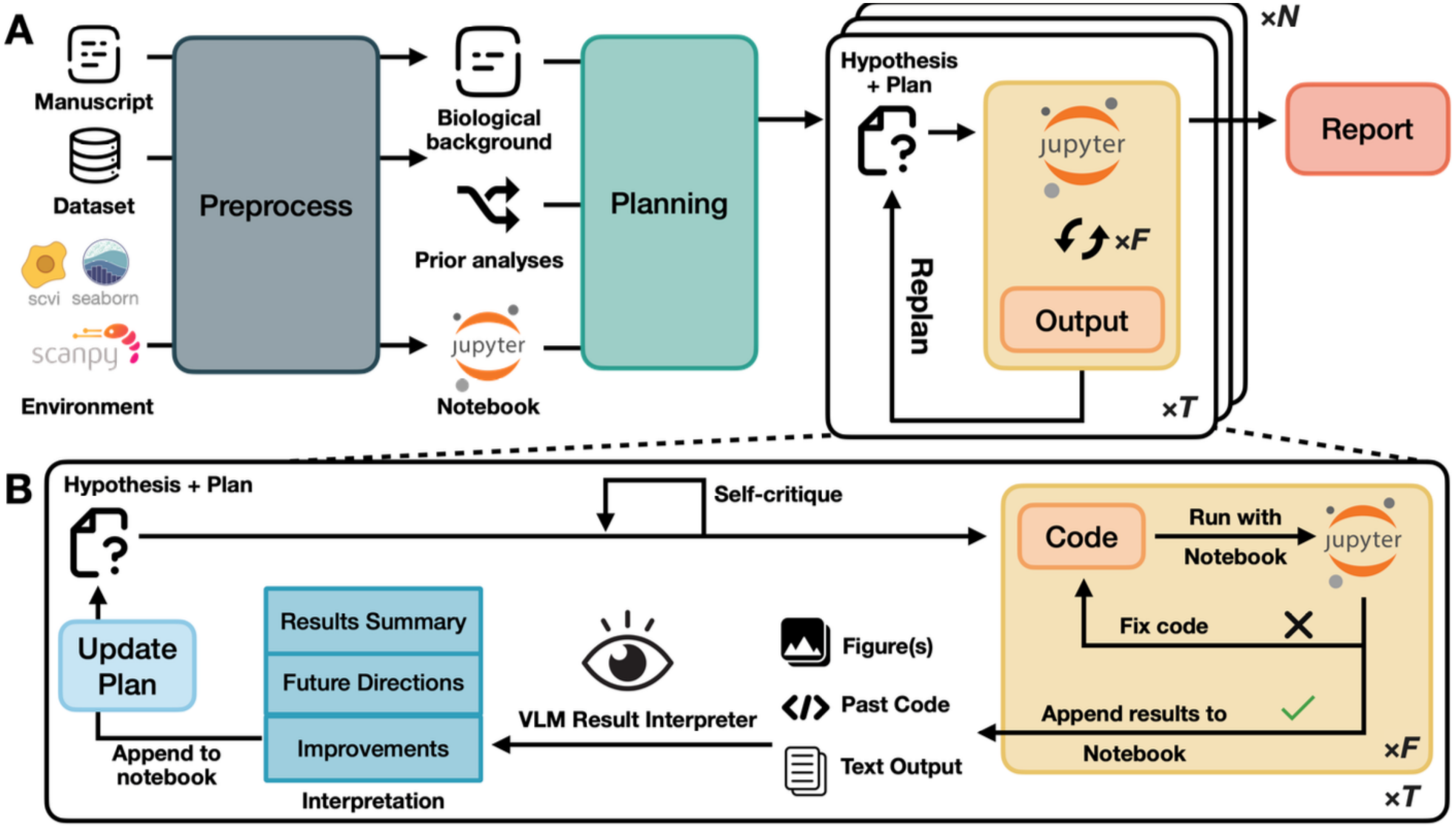


And protein-protein interaction prediction

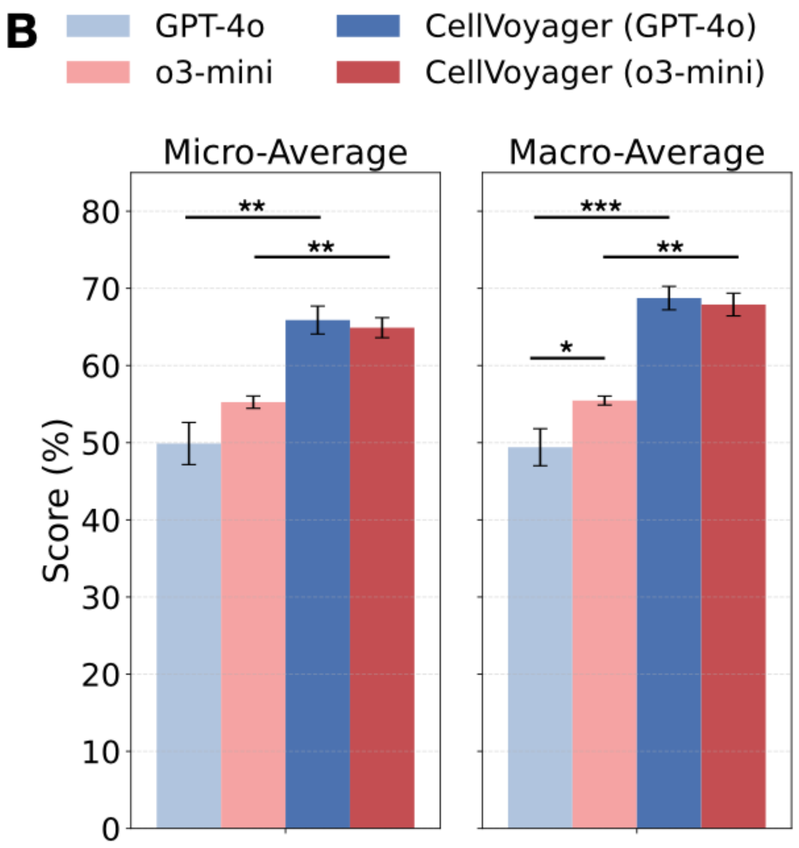


CellVoyager: AI CompBio Agent Generates New Insights by Autonomously Analyzing Biological Data (Alber, Chen, Sun et al, *Nature Methods*)

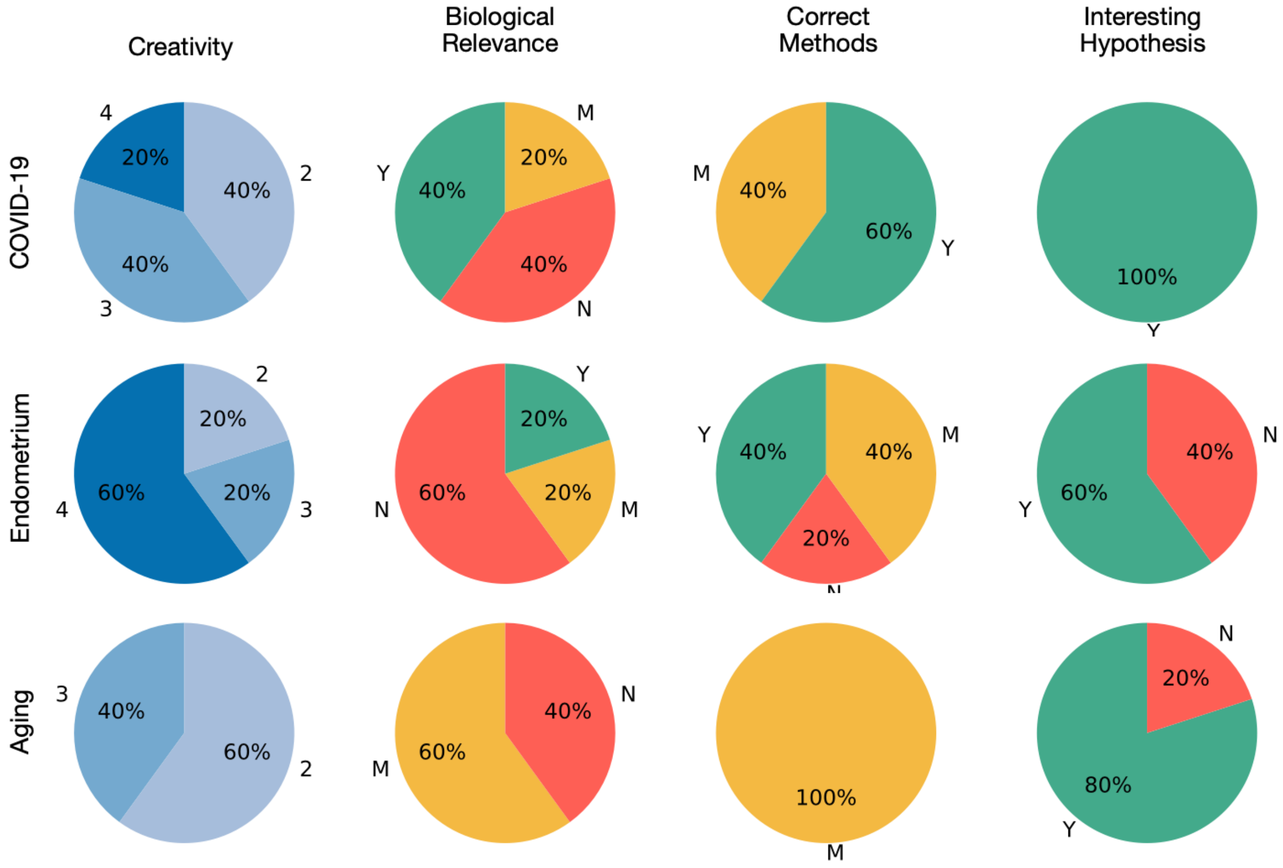
- **Goal**: Build an AI agent that does more than execute user requests — it autonomously explores single-cell datasets for new biological hypotheses
- **Method**: LLM agent + scRNA-seq dataset + manuscript context + prior analyses → generates "exploration blueprints," writes/runs code in Jupyter, self-critiques, fixes errors, interprets figures, replans, and summarizes findings
- **Result**: On CellBench, CellVoyager outperformed base GPT-4o and o3-mini at predicting analyses from 50 published scRNA-seq papers; in three real case studies, original authors judged many agent-generated analyses as creative, biologically relevant, and worth follow-up
- **Conclusion**: We may be entering the era where AI does not just help us run the analysis we asked for — it helps us find the analyses we did not think to ask


A
Manuscript
Dataset
scvi
seaborn
scanpy
Environment
Preprocess
Biological background
Prior analyses
jupyter
Notebook
Planning
×N
Hypothesis + Plan
jupyter
×F
Output
Replan
×T
Report
B
Hypothesis + Plan
Self-critique
Code
Run with
Notebook
jupyter
Fix code
Append results to
Notebook
×F
×T
Update Plan
Append to notebook
Results Summary
Future Directions
Improvements
Interpretation
VLM Result Interpreter
Figure(s)
Past Code
Text Output

Worked better on benchmarks

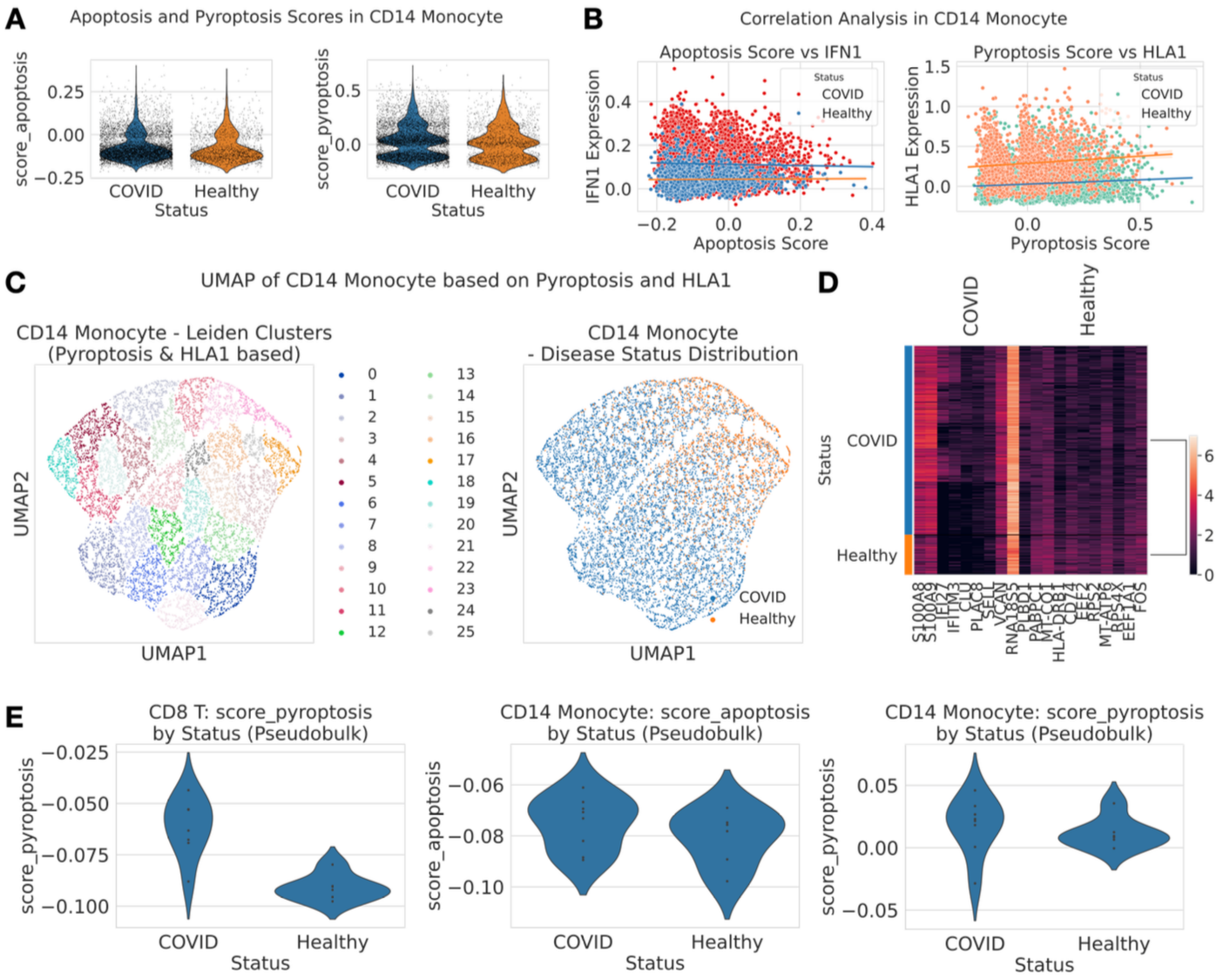


And humans thought it was pretty good too



The rest of the paper is agent-generated analyses and why they're reasonable

E.g.



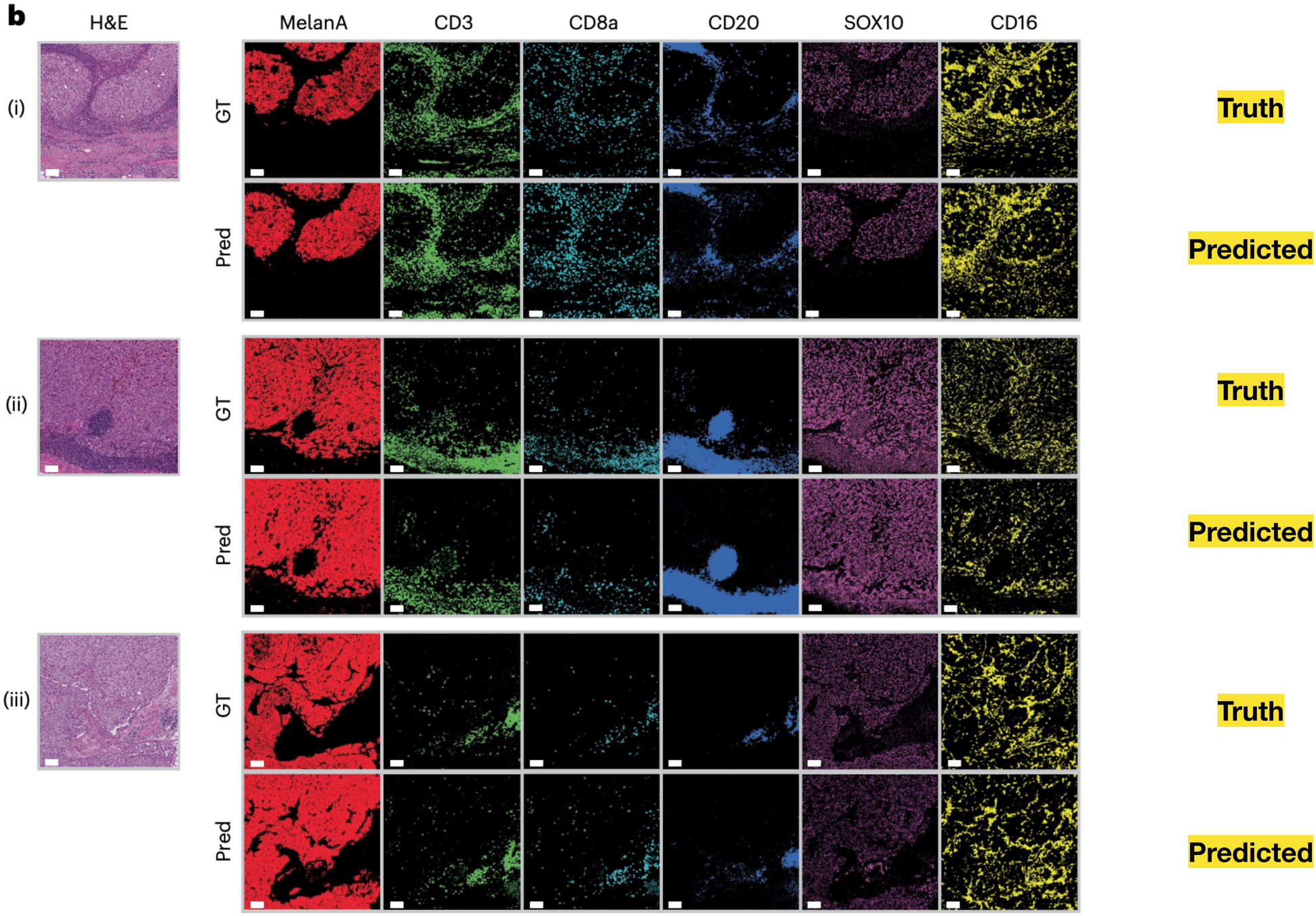
Call Me Maybe - *Carly Rae Jepsen*

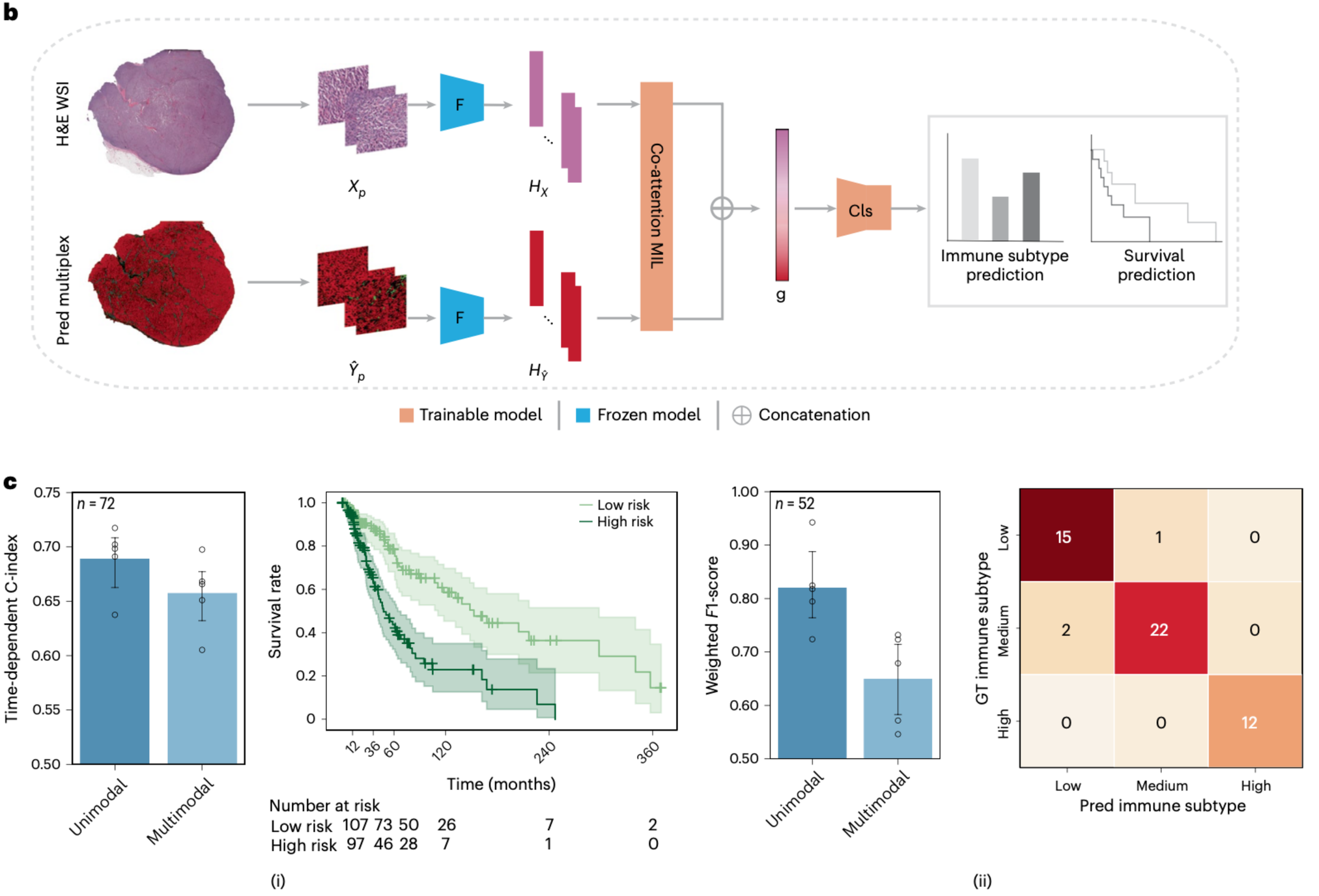
Shout Outs

Cool papers that deserve more than they got because it's a lot of papers to read

Histopathology-based protein multiplex generation using deep learning (Andani, Chen, Ficek-Pascual et al, *Nature Machine Intelligence*)

- **Goal**: Multiplexed protein imaging can characterize tumor–immune interactions, but cost, time, and tissue requirements limit clinical-scale use
- **Method**: Train HistoPlexer, a conditional GAN, to generate 11-channel protein multiplexes from routine H&E images, using losses that tolerate serial-section mismatch and preserve patch-level structure
- **Result**: Generated protein maps resembled real IMC, preserved spatial co-localization patterns, stratified melanoma into immune-hot/cold subtypes, and improved survival and immune subtype prediction over H&E alone
- **Conclusion**: H&E slides can be computationally augmented with virtual spatial proteomics to support tumor microenvironment phenotyping

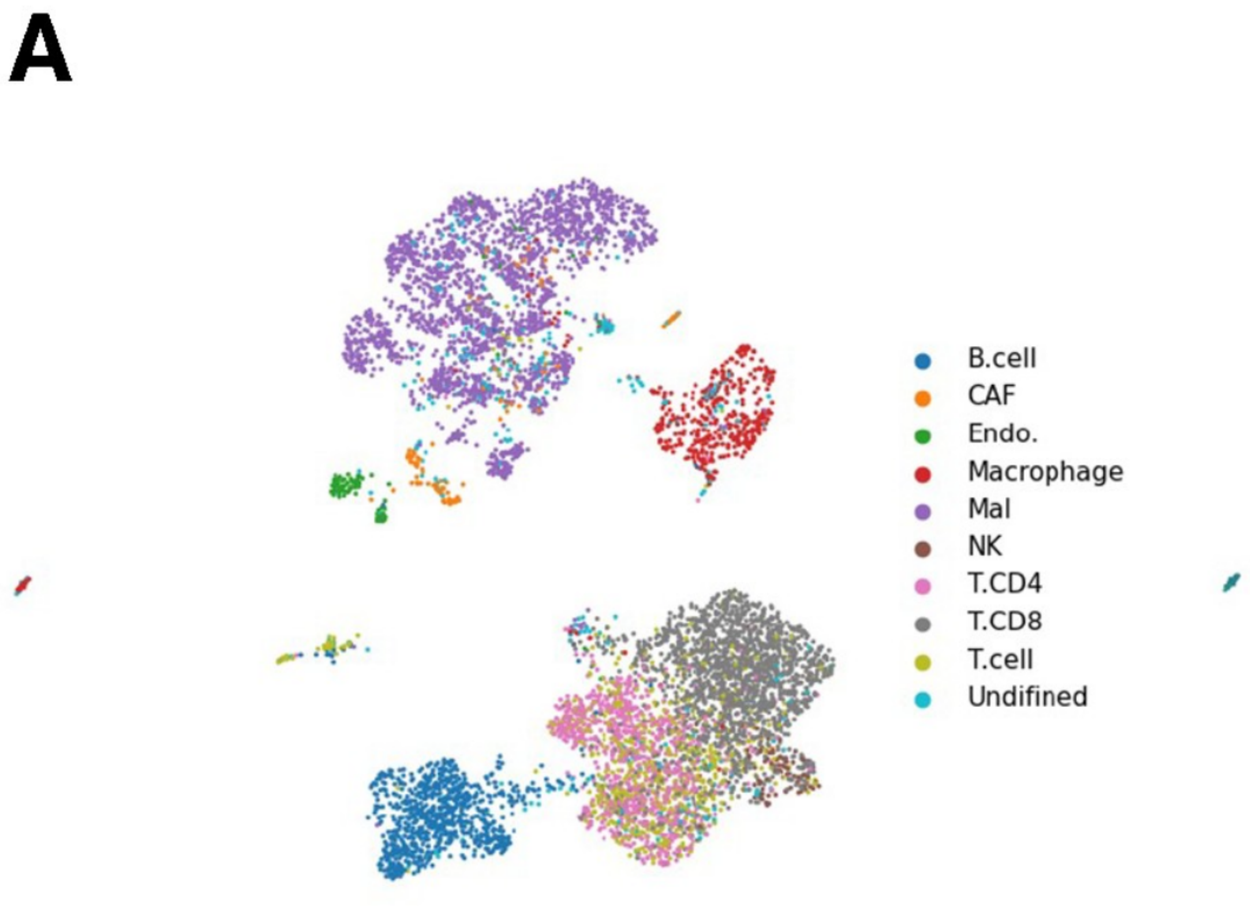

b
H&E
MelanA
CD3
CD8a
CD20
SOX10
CD16
(i)
(ii)
(iii)
GT
Pred
Truth
Predicted
Truth
Predicted
Truth
Predicted

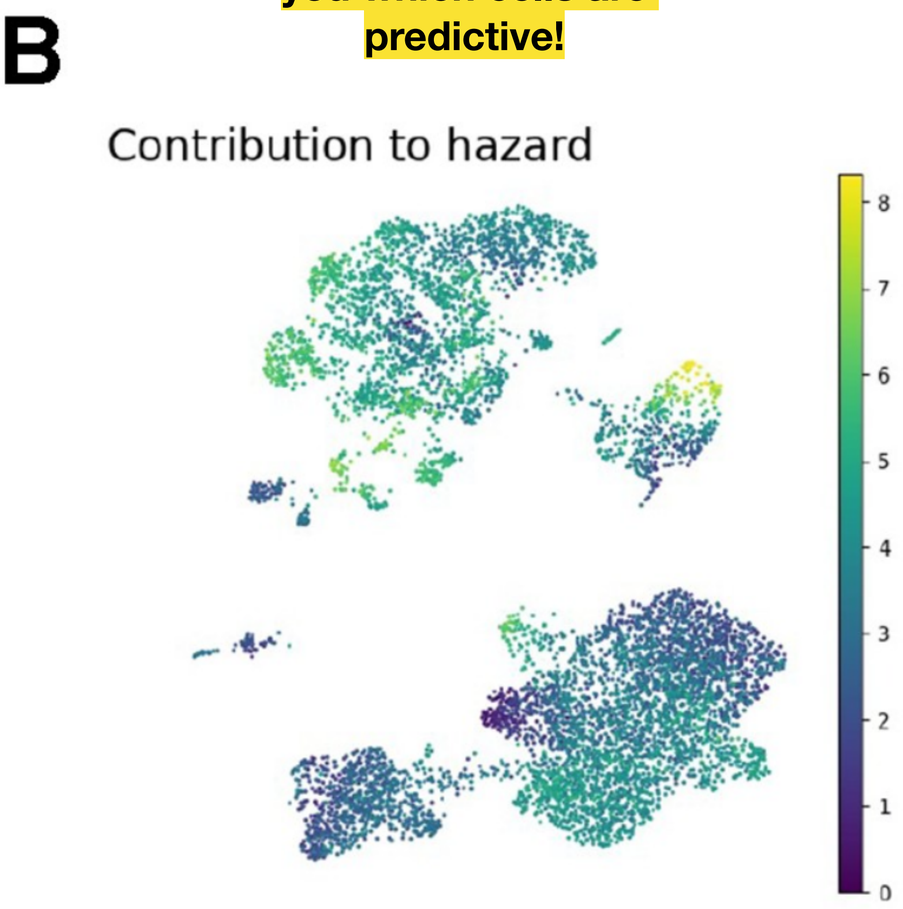

b
H&E WSI
Pred multiplex
F
F
Co-attention MIL
Cls
g
Immune subtype prediction
Survival prediction
Trainable model
Frozen model
Concatenation
c
Time-dependent C-index
n = 72
Unimodal
Multimodal
Survival rate
Low risk
High risk
Time (months)
Number at risk
Low risk 107 73 50 26 7 2
High risk 97 46 28 7 1 0
(i)
Weighted F1-score
n = 52
Unimodal
Multimodal
GT immune subtype
Pred immune subtype
Low
Medium
High
15 1 0
2 22 0
0 0 12
(ii)

scSurv: a deep generative model for single-cell survival analysis (Mizukoshi, Kojima, Hayashi et al, *Bioinformatics*)

- **Goal**: Figure out which individual cells contribute to patient survival, even when large outcome-linked cohorts only have bulk RNA-seq.
- **Method**: Train a VAE on scRNA-seq to learn latent cell states, deconvolve bulk RNA-seq into single-cell proportions, then extend a Cox proportional hazards model to estimate each cell's contribution to risk
- **Result**: scSurv outperformed cluster-level deconvolution + Cox models in simulations and showed predictive signal across several TCGA cancers — especially in melanoma
- **Conclusion**: Clever way to borrow the scale of bulk survival cohorts and the resolution of single-cell data to turn cell states into clinical risk factors

Cool plots like this that show you which cells are predictive!



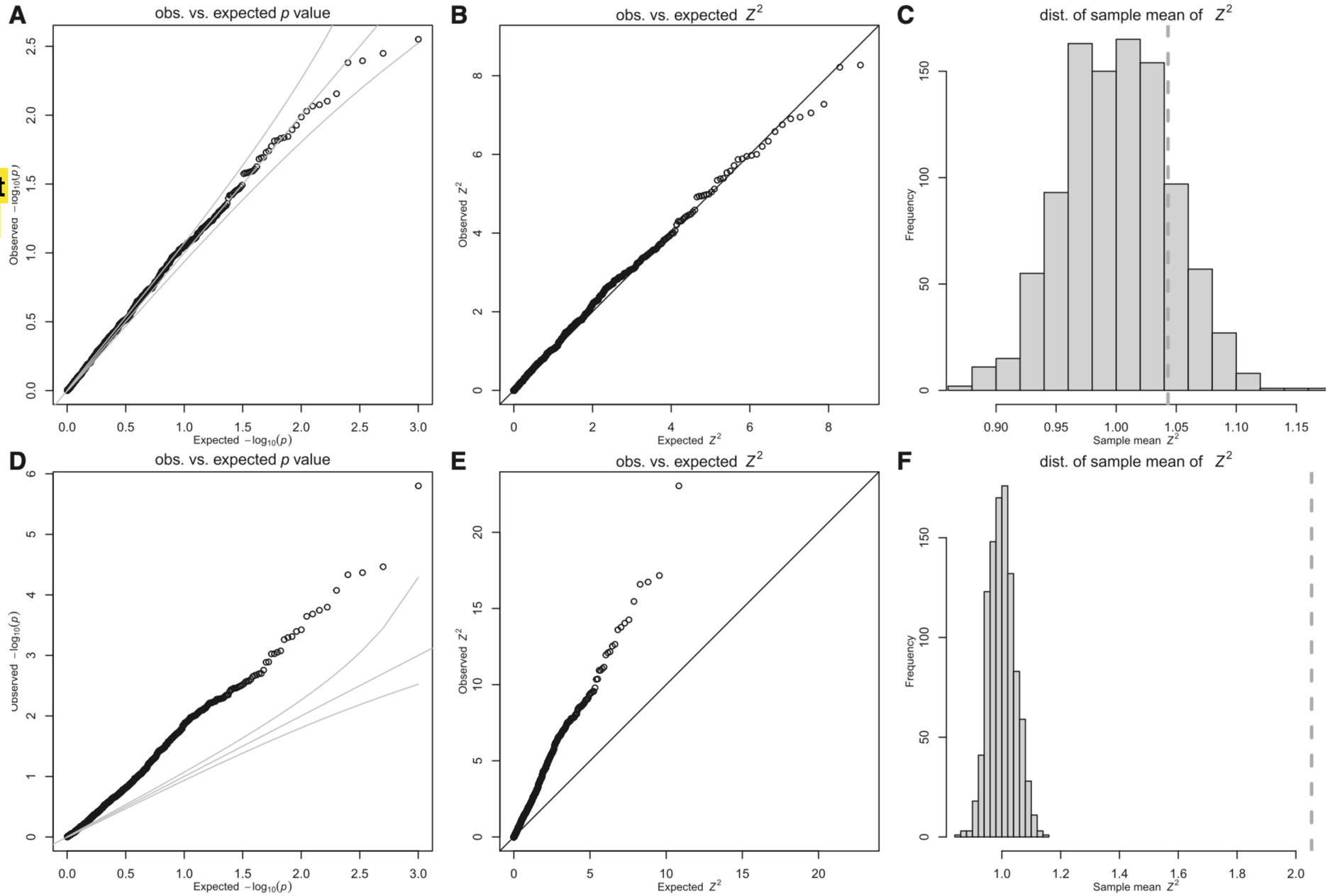


A gene-specific variance-control approach corrects polygenicity-driven inflation observed in transcriptome-wide association studies (Liang, Nyasimi, and Im, *AJHG*)

- **Goal**: Determine whether TWAS/xWAS association tests remain well calibrated for highly polygenic complex traits
- **Method**: Simulate polygenic null traits, test predicted molecular traits in UK Biobank, derive gene-specific inflation factors, and correct each TWAS Z-score using variance control
- **Result**: Standard TWAS shows inflated false positives that grow with GWAS sample size and trait heritability; variance control restores calibration better than existing approaches
- **Conclusion**: Polygenicity makes TWAS p-values overconfident; variance control gives them a much-needed reality check

Non-polygenic null (what stats were designed for)

polygenic null



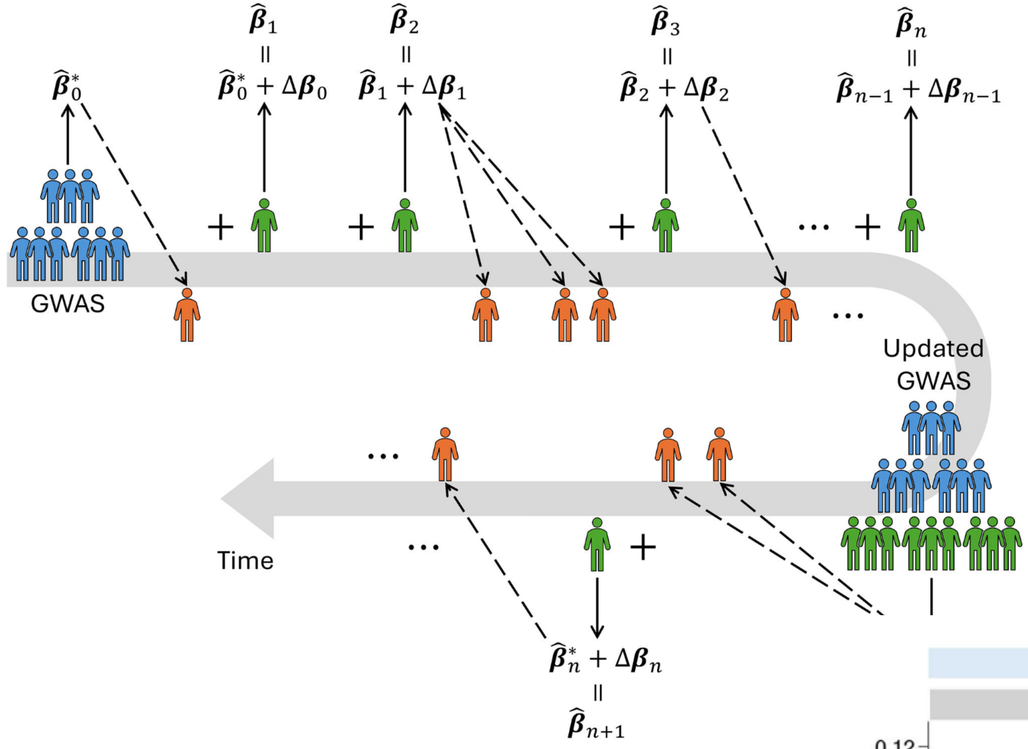
Real-time Dynamic Polygenic Prediction for Streaming Data (Tubbs, Chen, Duan, Huang, Ge, *Nature Genetics*)

- **Goal**: Existing PRSs are trained on static GWAS summary statistics, but health systems and biobanks keep generating new genotype–phenotype data; can PRS models update continuously as each new sample arrives?

- **Method**: Extend PRS-CS into rtPRS-CS, using stochastic gradient descent to refine SNP weights online from each incoming genotyped and phenotyped individual, with dynamic ancestry adjustment and standardization

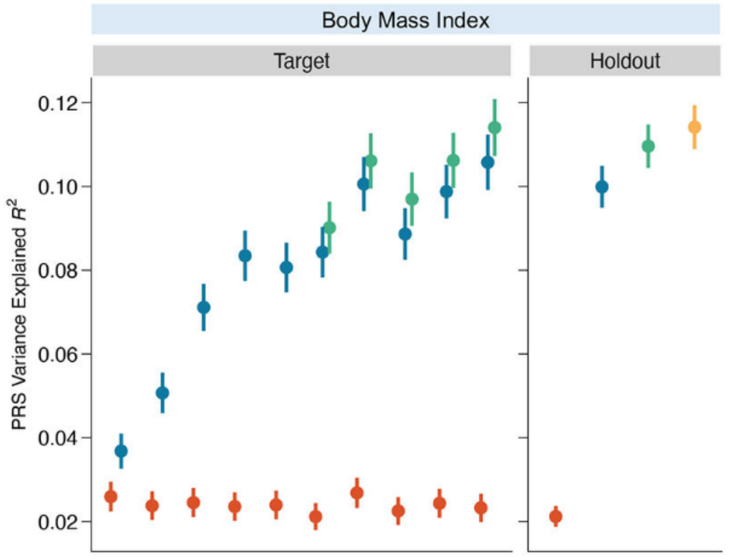
- **Result**: In simulations and biobank analyses, rtPRS-CS steadily improved prediction over time and approached the performance of full GWAS retraining; with an intermediate update, performance became statistically indistinguishable from the theoretical upper bound for most quantitative traits

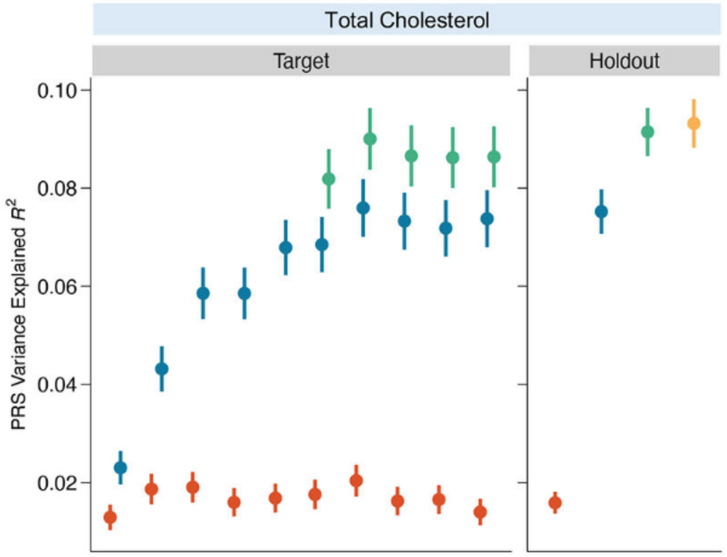
- **Conclusion**: A static PRS is starting to look like a software version problem; polygenic prediction can now be updated in real time as clinical and biobank data stream in



Update as you go

Performance pretty close to refitting





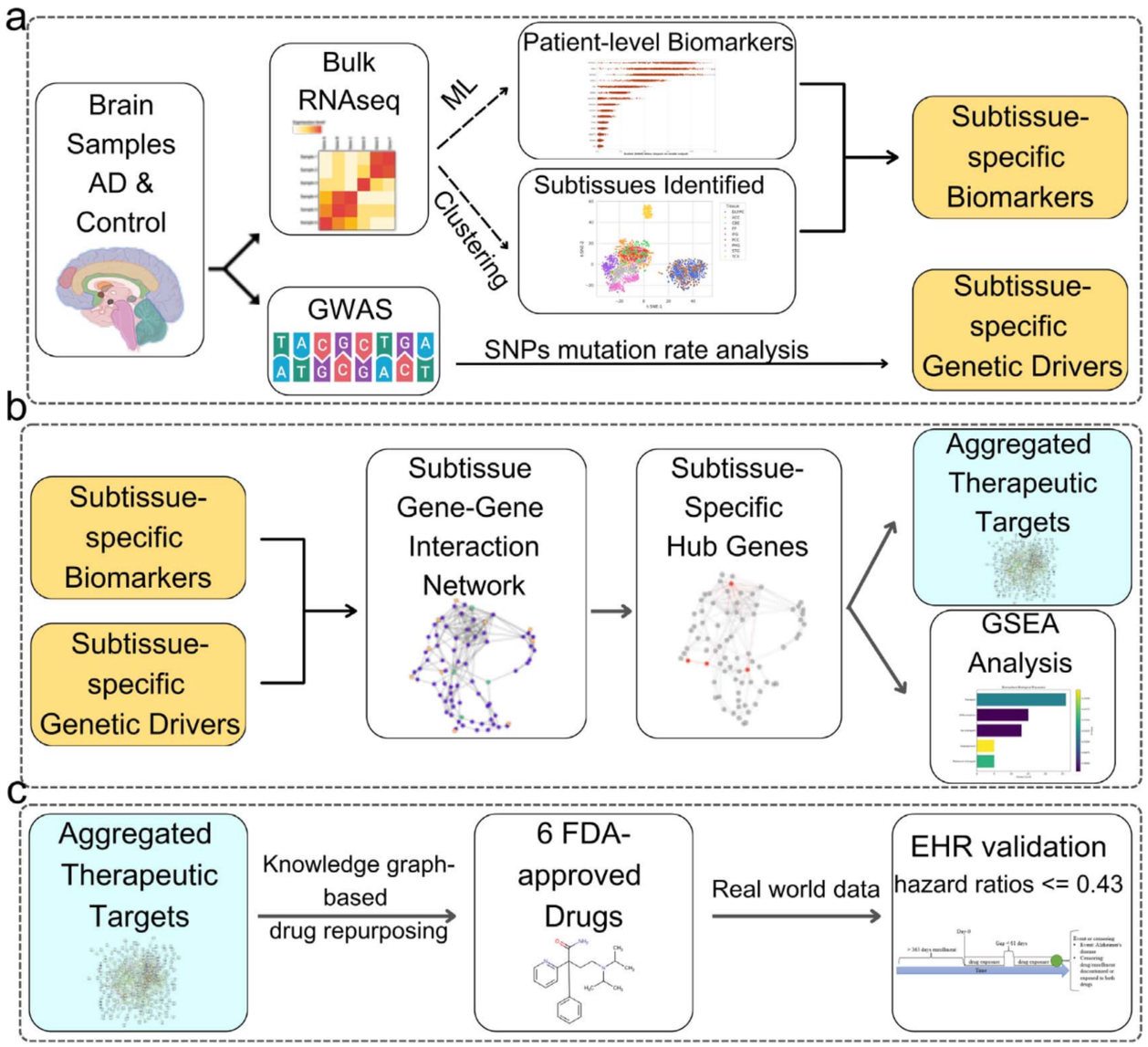
Integrating explainable artificial intelligence with multiomics systems biology and electronic health record data mining for personalized drug repurposing in Alzheimer's disease (Mottaqi, Zhang, Xie et al, *Briefings in Bioinformatics*)

- **Goal**: Identify patient- and brain-region-specific molecular drivers of Alzheimer's disease and use them to nominate repurposable drugs

- **Method**: PRISM-ML combines matched brain RNA-seq/GWAS from 2,105 samples, Random Forest + SHAP for sample-level biomarkers, subtissue clustering, coexpression-network bottleneck genes, knowledge-graph drug matching, and claims-data validation

- **Result**: Identified 36 molecular subtissues, 262 bottleneck genes, and six FDA-approved candidate drugs; promethazine exposure was associated with lower AD incidence versus an antihistamine comparator

- **Conclusion**: A nice full-stack precision-repurposing pipeline, though the excitement is mostly in the integration and validation rather than any one algorithmic leap

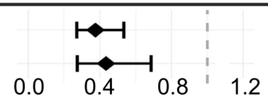


Thoughtful integrative pipeline

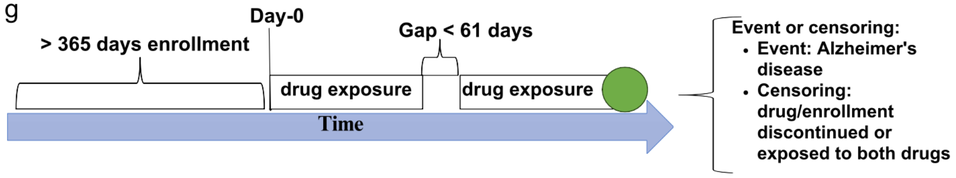
Validation in EHR

f

Model	Alzheimer's disease hazard ratio (95% CL)	P-value
Covariate adjustment	0.38 (0.27-0.53)	< 0.001
Inverse probability weighting	0.43 (0.27-0.69)	< 0.001



g



Multimodal large language models and mechanistic modeling for glucose forecasting in type 1 diabetes patients (Wolber, Samadi, Sellin, Schuppert, *Journal of Biomedical Informatics*)

- **Goal**: Improve glucose forecasting for type 1 diabetes by using meal images instead of tedious manual food logging

- **Method**: Multimodal LLM estimates macronutrients from meal photos → mechanistic Bézier curves model patient-specific temporal effects of food and insulin → LightGBM predicts future glucose

- **Result**: Bézier approach performed best across datasets, with 30-min RMSE ≈15–17 mg/dL and 60-min RMSE ≈25–28 mg/dL; patient-specific curves showed distinct metabolic response patterns

- **Conclusion**: Not quite TBI, but a fun glimpse of where personal health forecasting is going: take a picture of lunch, infer the physiology, predict the future

Human interpretable grammar encodes multicellular systems biology models to democratize virtual cell laboratories (Johnson, Bergman, Rocha, et al., *Cell*)

- **Goal**: Make agent-based models of multicellular systems easier to build, interpret, reproduce, and connect to multi-omics data

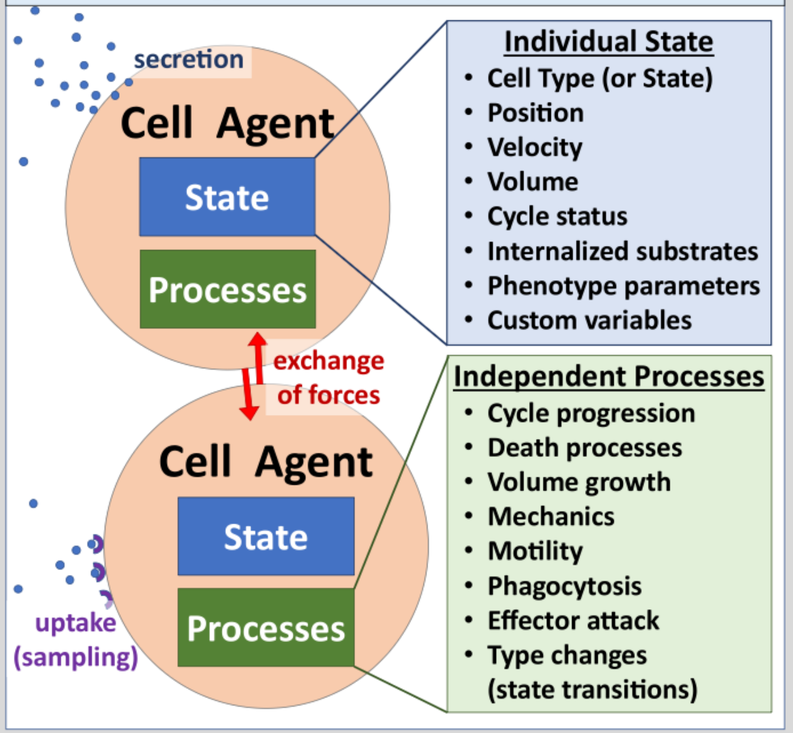
- **Method**: Define a “cell behavior hypothesis grammar” that converts plain-language rules about cell signals and behaviors into executable PhysiCell models

- **Result**: Demonstrated virtual experiments across tumor growth, PDAC invasion, immunotherapy response, and cortical layer formation using spatial and single-cell data.

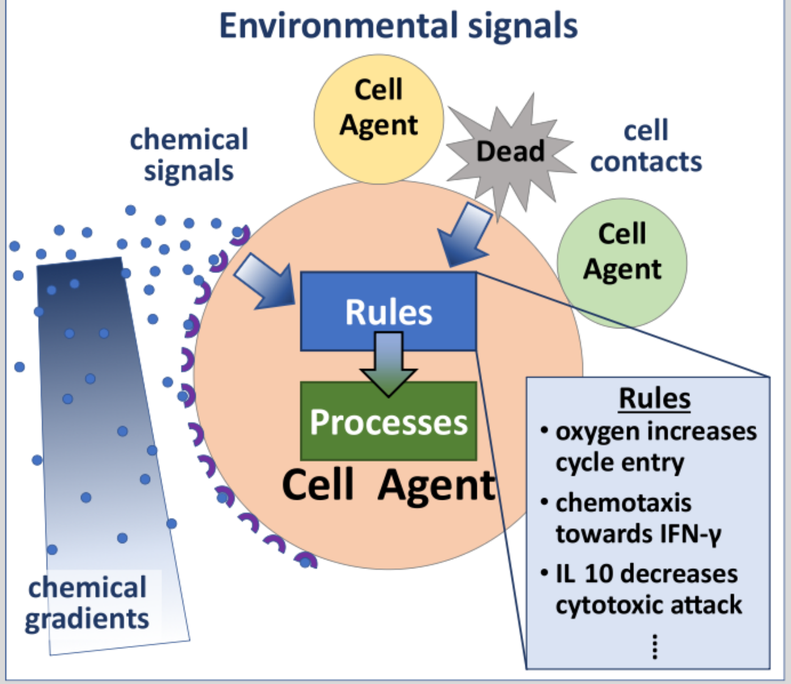
- **Conclusion**: A practical step toward virtual cell laboratories where biological hypotheses can be written, simulated, parameterized, and tested

#IS25 #YIR25 @proftatonetti @tatonetti.bsky.social

A Cell agents as independent actors



B Cell agents as signal processors



C From rules to mathematics

(1) Dictionary of Signals (S)
intra/extracellular substrates, pressure, volume, cell contacts, death status, damage

(2) Dictionary of Behaviors (B)
cycling, death, secretion/uptake, adhesion mechanics, migration/chemotaxis, transition, asymmetric division, phagocytosis, fusion, immunogenicity, cell attack, damage/repair

(3) Hypothesis Statements (ABM rules or logical model)
For [cell type *T*], [*S*] increases [behavior *b*]
For [cell type *T*], [*S*] decreases [behavior *b*]

For MCF-7 cells, cisplatin increases apoptosis

For naïve T cells, IL-10 decreases transition to CD8+ T cells

(4) Auto-Generated Mathematics

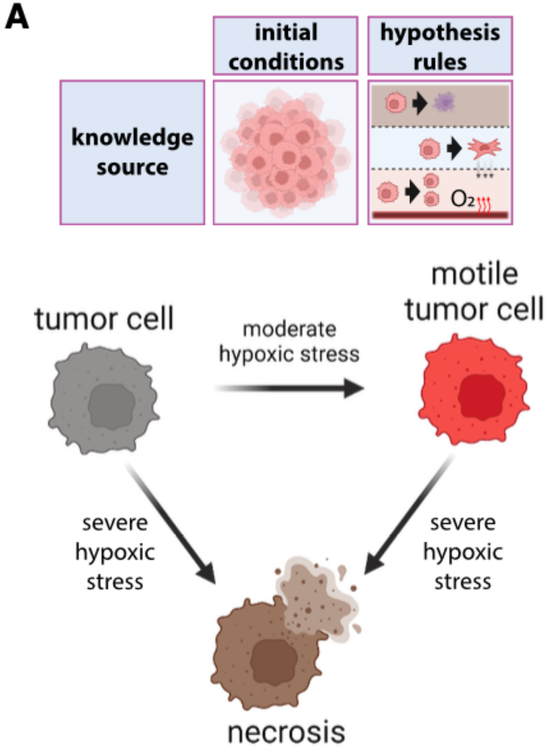
MCF-7:

$$d_A = 0 + 0.005 \cdot \frac{\left(\frac{\text{cisplatin}}{0.5}\right)^4}{1 + \left(\frac{\text{cisplatin}}{0.5}\right)^4}$$

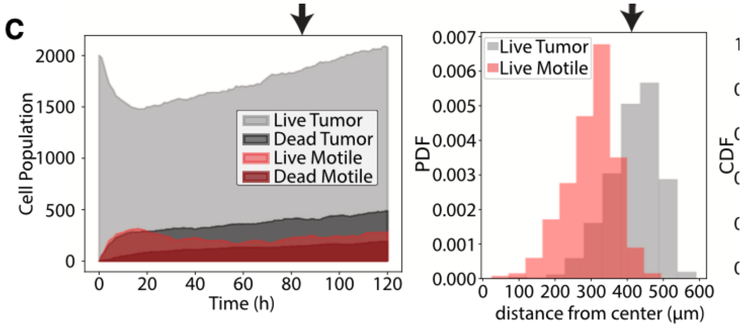
naïve T cells:

$$r_{T,\text{naïve}\rightarrow\text{effector}} = 0.001 \cdot \left(1 - \frac{\left(\frac{\text{IL10}}{0.25}\right)^2}{1 + \left(\frac{\text{IL10}}{0.25}\right)^2}\right)$$

A



C



Sweet Dreams (Are Made of This) - *Eurhythmics*

Brain Candy

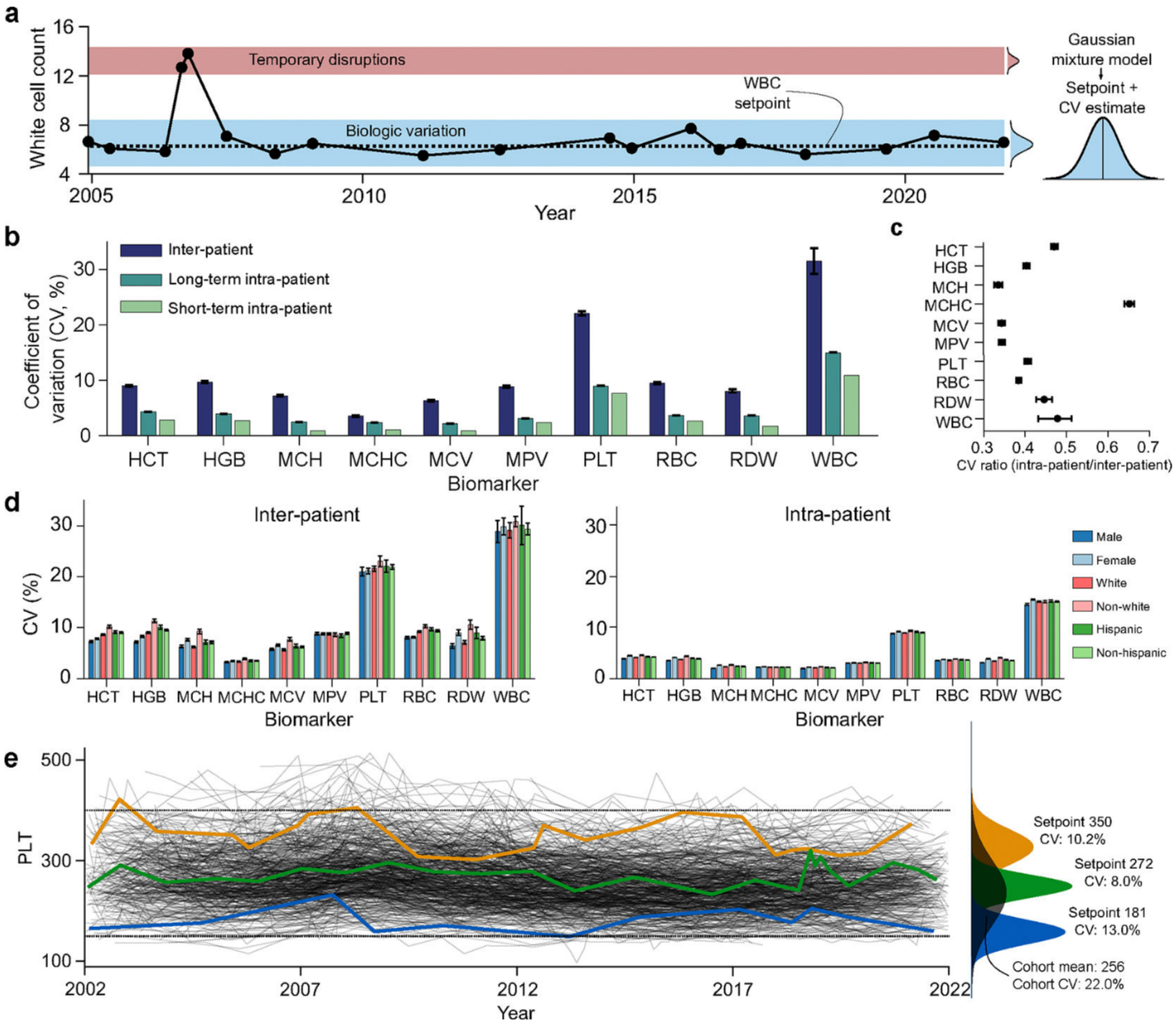
Hematologic setpoints are a stable and patient-specific deep phenotype (Foy, Petherbridge, Roth et al, *Nature*)

- **Goal**: Determine whether routine CBC values define stable, patient-specific physiologic baselines useful for precision medicine
- **Method**: Analyze longitudinal EHR CBCs from thousands of healthy adults over decades; estimate patient-specific hematologic “setpoints” and test associations with genetics, mortality, disease risk, and diagnostic interpretation
- **Result**: CBC setpoints were stable for ≥20 years, distinguished a typical healthy adult from 98% of others, improved GWAS power, and stratified mortality and disease risk even within normal reference ranges
- **Conclusion**: your normal CBC is not my normal CBC

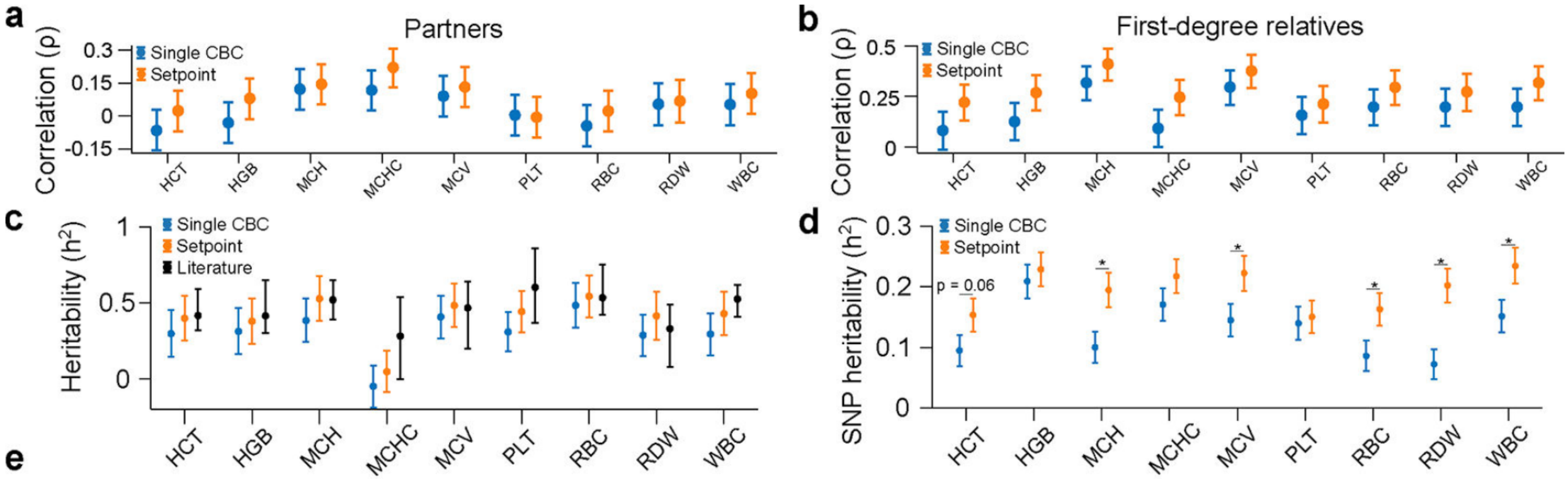
How the set point is defined

Inter-patient has more variance than intra-patient

Set points are stable for 20+ years



You share your set points with your family



Not your partner (or maybe a little?)

Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance (Van Calster, Collins, Vickers et al, *Lancet Digital Health*)

- **Goal**: Clarify which performance measures should be used when validating predictive AI models for medical decision-making
- **Method**: Review 32 measures across discrimination, calibration, overall performance, classification, and clinical utility; test concepts using the ADNEX ovarian malignancy model
- **Result**: Many common metrics are problematic: classification measures are improper at most clinically relevant thresholds, and F1 score fails both key criteria
- **Conclusion**: Stop asking whether the model is “accurate”; ask whether its probabilities are calibrated and whether acting on them helps patients

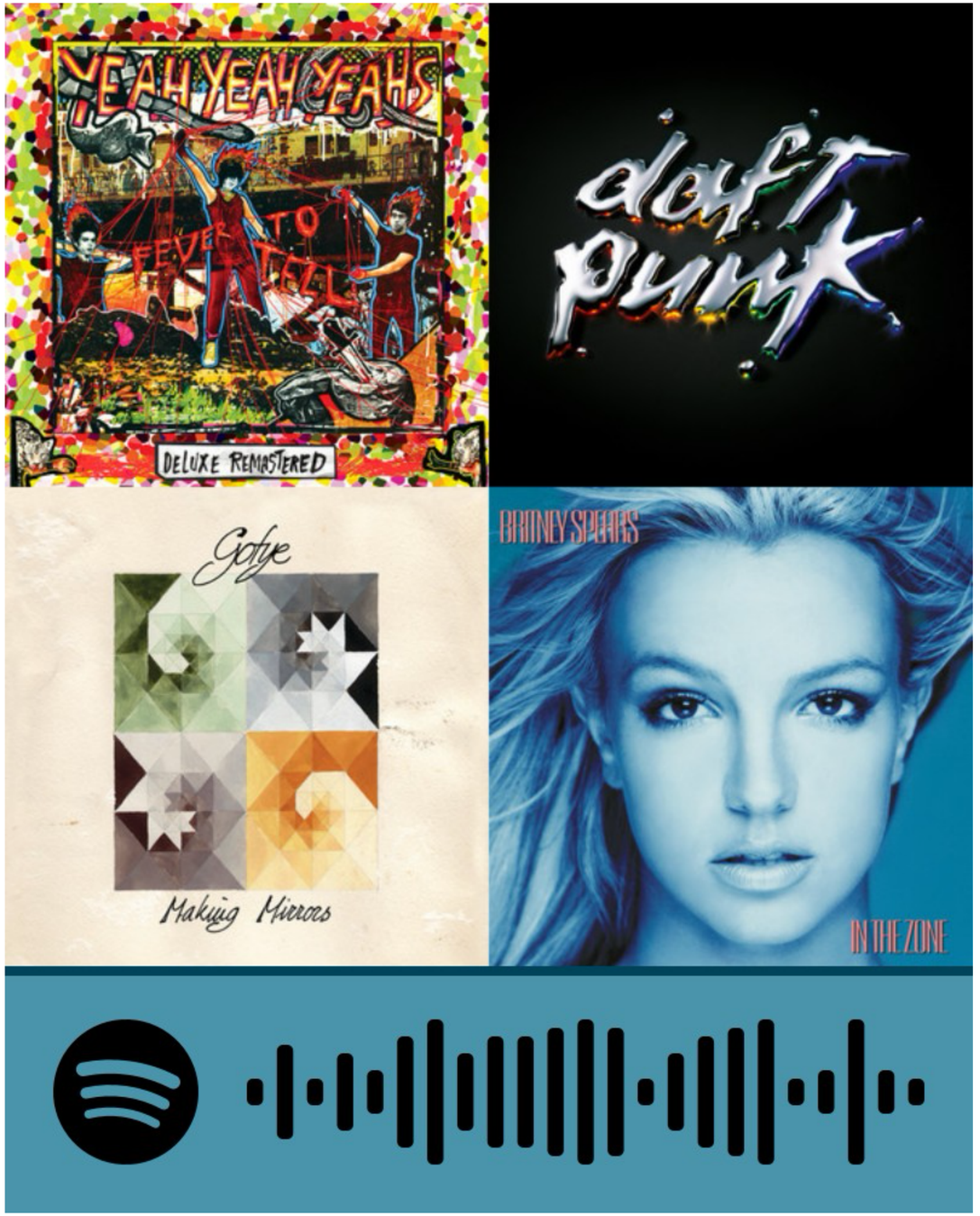
Look back at my predictions for 2025 ✨

- ☑ Emergence of multimodal CLIP models for TBI, particularly with language
- ☑ Synthetic data will find some compelling use cases (I haven't seen one yet)
- ☑ Foundation models will have big impact on rare disease work
- ☑ Diffusion models for novel drug discovery coupled with experimental validation (and trials?)
- ☑ AI efficiency boosts will lead to real time/streaming applications in TBI
- ☑ New explainable AI techniques (e.g. SAE) will begin to get used in TBI
- ☑ An initial biomedical application of quantum computing
- ☒ Uncertainty quantification in AI modeling will emerge
- ☑ Lastly, new architectures that can leverage multimodal data (I don't think we've exhausted this at all)

Predictions for 2025 ✨

- [ ] **Synthetic cohorts get a job -** virtual data improves real analyses, not just figures
- [ ] **Foundation models go where labels are scarce -** rare disease, immune repertoires, small cohorts
- [ ] **Drug discovery becomes closed-loop -** predict, test, update, repeat
- [ ] **Agents enter the wet/dry lab workflow -** autonomous analysis with testable hypotheses
- [ ] **Multimodal alignment beats multimodal concatenation -** models bridges data types
- [ ] **Trajectory models meet mechanism -** disease progression models connect to biology
- [ ] **Uncertainty becomes a first-class output -** especially for clinical-facing AI
- [ ] **Benchmarks get serious -** evaluation becomes as important as architecture

Thank you!


YEAH YEAH YEAHS
FEVER TO TELL
DELUXE REMASTERED
daft punk
Goldfrapp
Making Mirrors
BRITNEY SPEARS
IN THE ZONE