2025 Translational Bioinformatics Year-in-Review

@proftatonetti @tatonetti.bsky.social

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

2025 AMIA Summits - Pittsburgh

Disclosures

- NIH, FDA, DoD, Pfizer, AstraZeneca, Janssen, Amgen, PhARMA Foundation, CARI Health
- Co-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 2025)
- 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 2024
  - 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 the last 14 months (Jan 2024 - Mar 2025)

- Focused on human biology

- **What's NOT included:**

  - Amazing biology with straightforward informatics (PRS, looking at you 👀)

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

  - Perspectives, reviews (for the most part)

This is all thanks to…

The 2025 AMIA Year-in-Review Team!

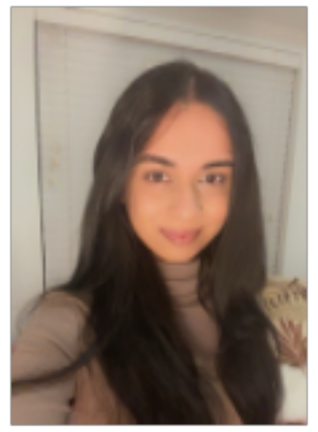
AYA AWWAD
MASS GENERAL HOSPITAL

HARSHAL SINGH CHAUHAN
UMBC

IFEANYI PETER ONYEKA
NNAMDI AZIKIWE UNIVERSITY

AKINCHAN BHARDWAJ
JIPMER, INDIA

JADA BATTEN
USF



PRATISTHA GUCKHOOL
DREXEL UNIVERSITY

MOHAMMED MUSTAFA KHAN
DEPAUL UNIVERSITY

RAGHAD MALKAWI
INDIANA UNIVERSITY

RIA TREESA RAJU
INDIANA UNIVERSITY

ASMA PATHAN
UNC CHARLOTTE

MIRIAM SAFFERN
ICAHN SCHOOL OF MEDICINE MOUNT SINAI, NY

ISHITA VASUDEV
BRONX CARE MOUNT SINAI, NY

ASMA SULTANA
UNIVERSITY OF ASIA PACIFI

NICK REID
UW, SEATTLE

SRI VARSHA REDDY CHANDIRI
UMBC

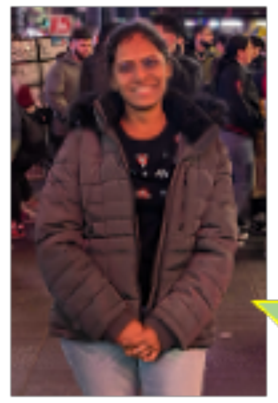
VEDA PRIYA PULIGUNDLA
UNIVERSITY OF MINNESOTA

FARHAN QUADIR
UNIVERSITY OF CHICAGO

MD. ASHRAFUL ISLAM
UNIVERSITY OF DHAKA

FLORA KIBONG
UNIVERSITY OF MICHIGAN

JUNNATUL MAWA
UNIVERSITY OF ASIA PACIFIC



ROOPA SANTOSHI GATTA
UMBC

PUSHKALA JAYARAMAN
ICAHN SCHOOL OF MEDICINE MOUNT SINAI, NY

2025 Translational Bioinformatics Year-in-Review Committee

Final List

- 822 papers triaged; 221 reviewed and scored by the committee
- 25 presented here + 9 shout outs + 3 pieces of brain candy
  - Apologies for those I missed, misunderstood, or misjudged, biases/mistakes are all mine
- 6 TBI topics:
  - ***Lose Control*** *- Drug Discovery & Repurposing*
  - ***Cruel Summer*** *- Taylor-ed for you - Precision Medicine in Action*
  - ***Good Luck, Babe*** *- Bio Euphoria - Integrating Clinical & Molecular Data*
  - ***What Was I Made For?*** *- Biomarker Discovery & Validation*
  - ***I Remember Everything*** *- EHR, Real-world Evidence, & Epidemiology*
  - ***Houdini*** *- Emerging Therapeutics & Technologies*
- 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…

“Lose Control”

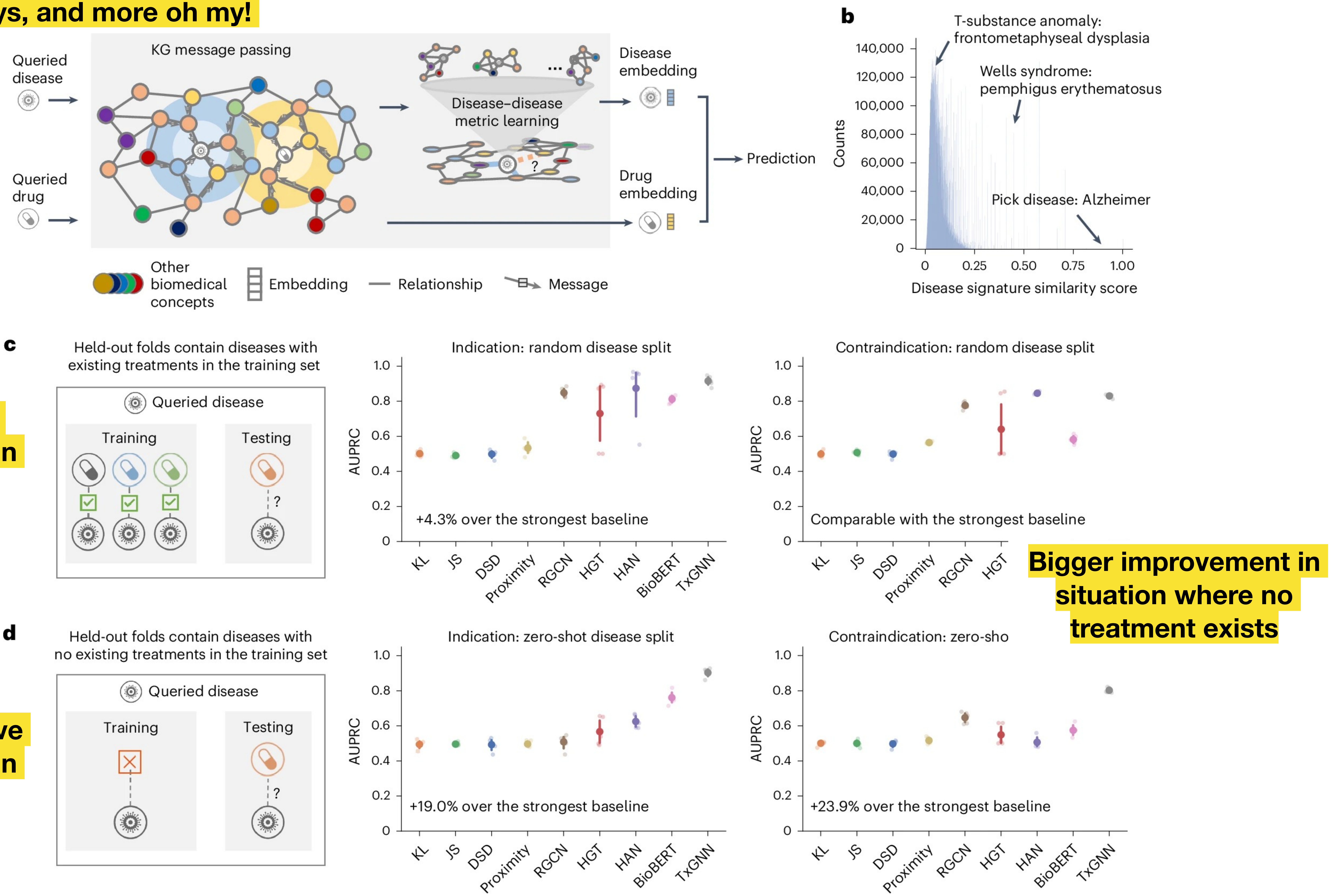
Drug Discovery & Repurposing

A foundation model for clinician-centered drug repurposing (Huang et al, *Nature Medicine*)

- Goal:
  - Previous drug repurposing strategies require some drugs for a disease to copy/learning from
  - Build a zero-shot algorithm that can make predictions even when no previous examples of drugs exist
- Method:
  - Embed diseases based on a knowledge graph of relationships between diseases, drugs, proteins, pathways, and clinical phenotypes — allows diseases to share a embedding space and thus share information
  - Use Graphical Neural Network (TxGNN) with message passing to produce embeddings for drugs or diseases (and for any of the other concepts for that matter)
  - Drugs and diseases are embedded in same space allowing them to be directly compared
- Result:
  - 49.2% improvement in drug indication prediction accuracy and 35.1% improvement in contraindication predictions over competing methods
  - Predictions aligned with real-world off-label use drug prescriptions
- Conclusion: Beautiful example of the power of building foundation models to improve performance for situations with little data available.

https://doi.org/10.1038/s41591-024-03233-x

GNN built on KG of drugs, disease, pathways, and more oh my!



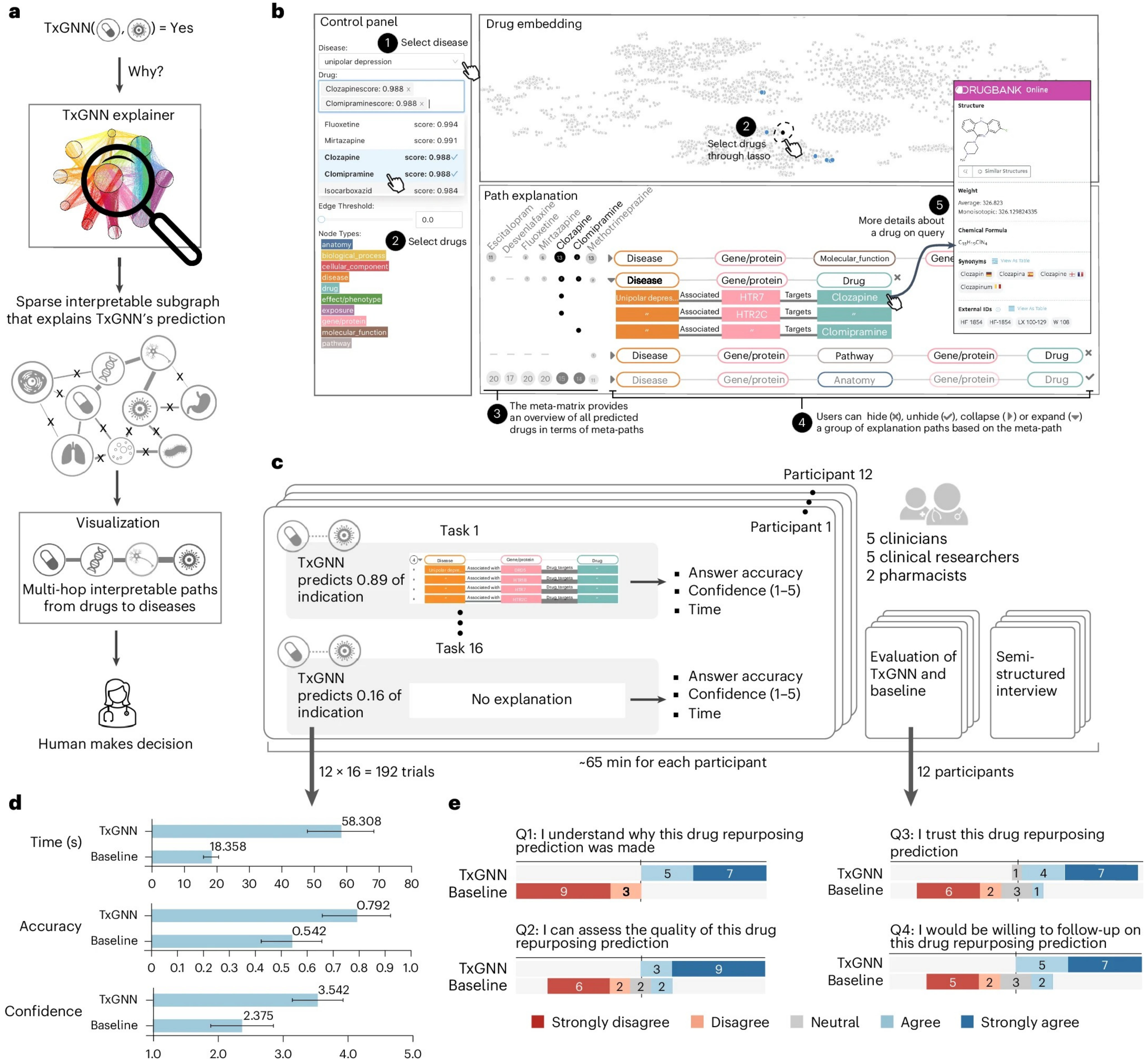
Two held-out situations:

Disease *does* have existing treatments in training

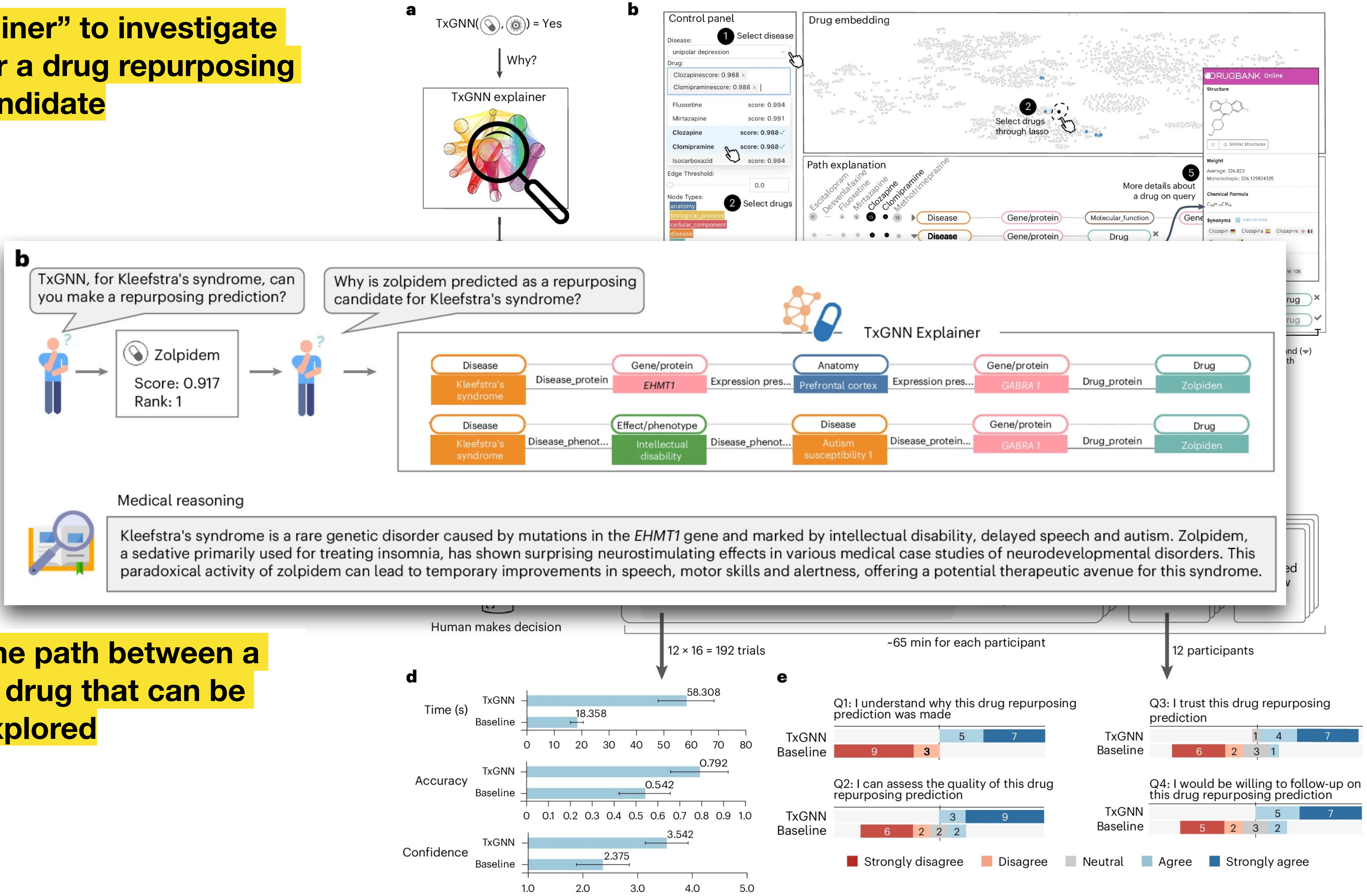
Disease *does not* have existing treatments in training

Bigger improvement in situation where no treatment exists

Built an “explainer” to investigate the evidence for a drug repurposing candidate



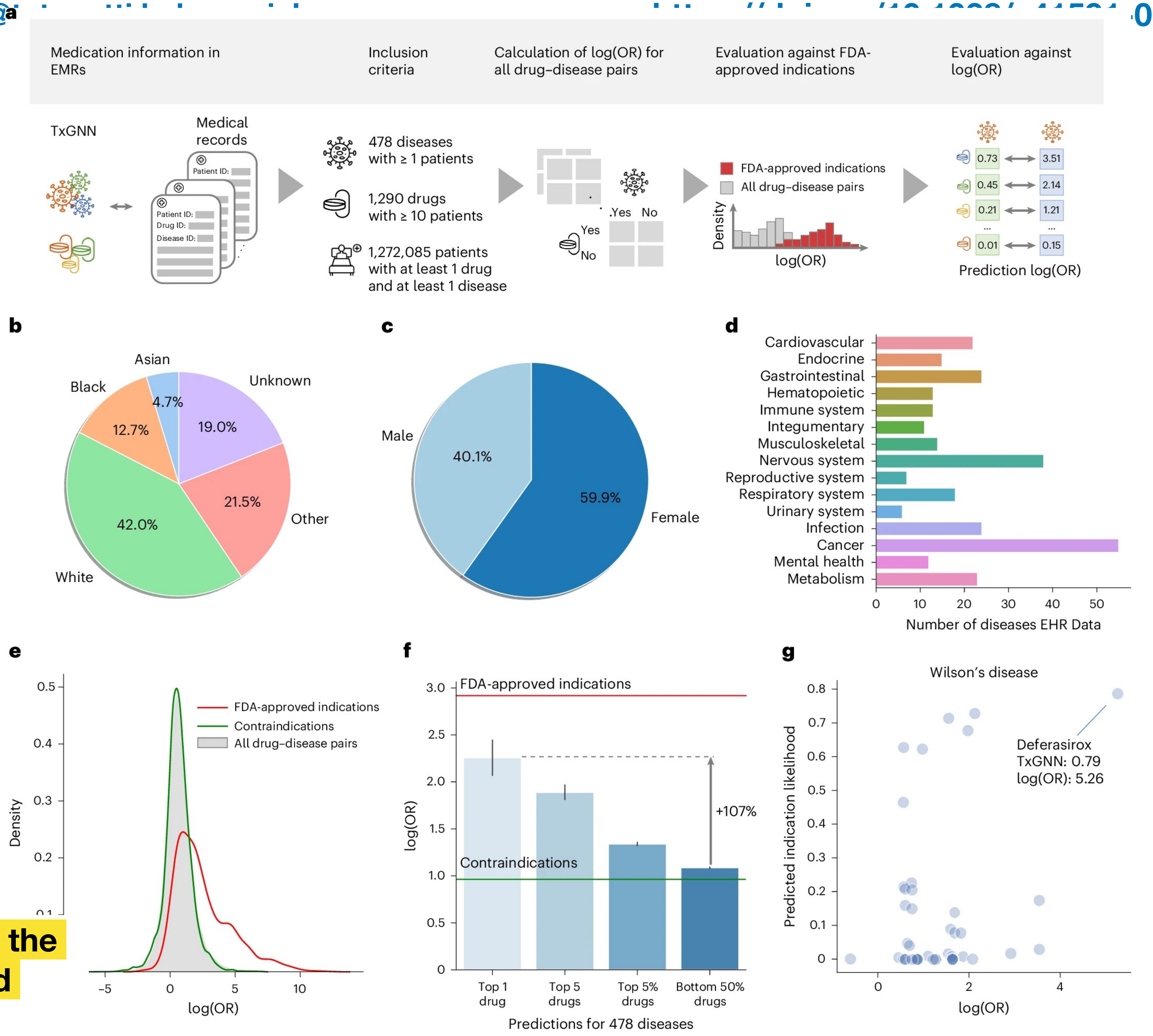
Built an “explainer” to investigate the evidence for a drug repurposing candidate



Kleefstra's syndrome is a rare genetic disorder caused by mutations in the *EHMT1* gene and marked by intellectual disability, delayed speech and autism. Zolpidem, a sedative primarily used for treating insomnia, has shown surprising neurostimulating effects in various medical case studies of neurodevelopmental disorders. This paradoxical activity of zolpidem can lead to temporary improvements in speech, motor skills and alertness, offering a potential therapeutic avenue for this syndrome.

Example of the path between a disease and drug that can be explored

Used EHR data to corroborate some drug repurposing hypotheses



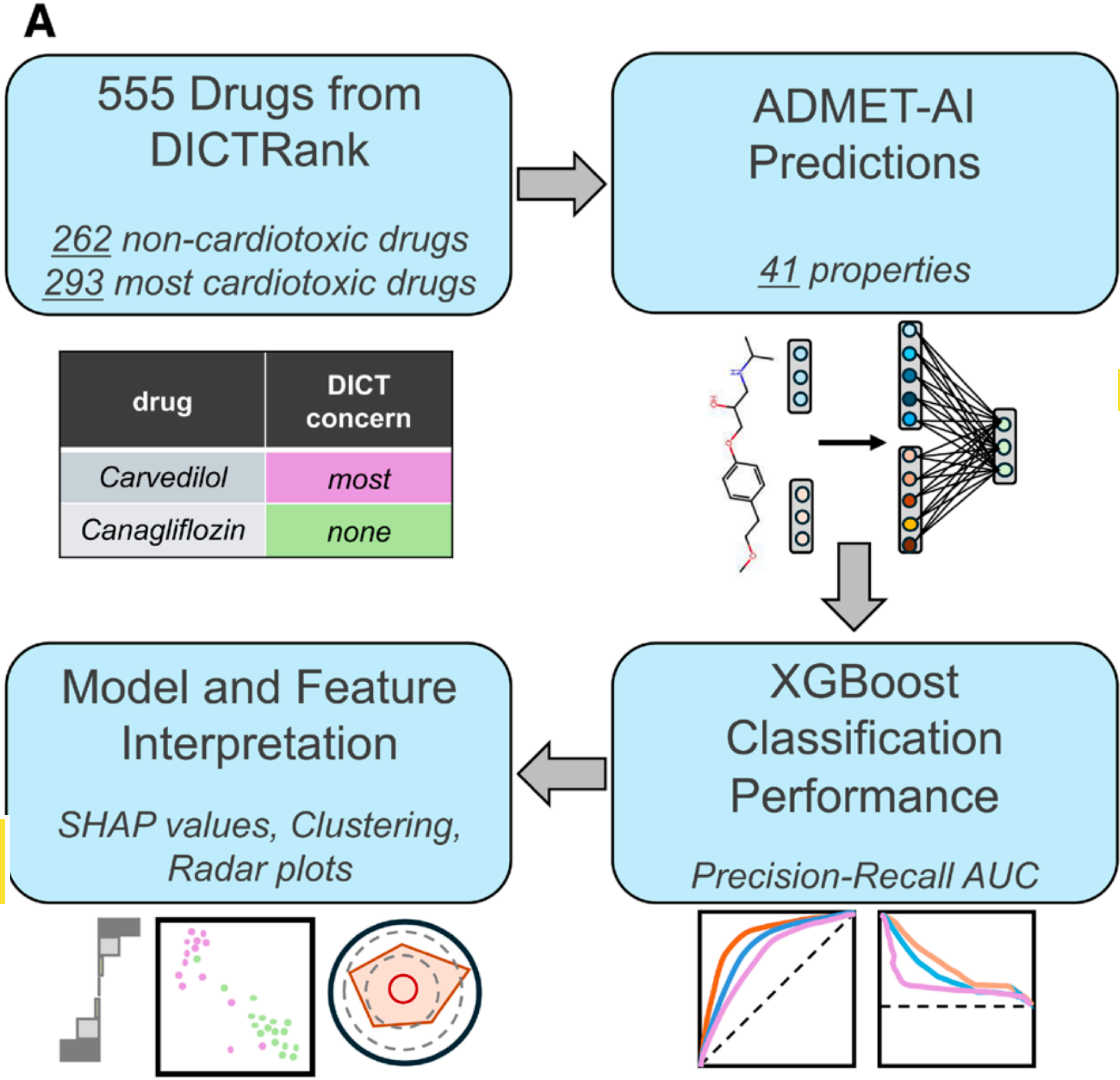
See generally a shift between the OR for approved drugs and contraindications

ADMET-AI Enables Interpretable Predictions of Drug-Induced Cardiotoxicity (Mukherjee, Swanson et al, *Circulation*)

- Goal: Predict drug-induced cardiotoxcity pre-clinically and identify causal factors
- Method:
    - Use ADME-AI to generate 41 ADMET properties
    - Feed those into Extreme Gradient Boosting to predict cardiotoxicity
- Result:
    - AUROC = 0.72 w/ top predicted features: CYP2D6 metabolism, Nrf2-antioxidant response, aromatase inhibition
    - Classify drugs into three categories: safe, high risk, withdrawn
- Conclusion: Predicting drug effects continues to be very hard.

https://doi.org/10.1161/CIRCULATIONAHA.124.070413

This DICTRank is publicly available from the FDA


A
555 Drugs from DICTRank
262 non-cardiotoxic drugs
293 most cardiotoxic drugs
ADMET-AI Predictions
41 properties
XGBoost Classification Performance
Precision-Recall AUC
Model and Feature Interpretation
SHAP values, Clustering, Radar plots

drug	DICT concern
Carvedilol	*most*
Canagliflozin	*none*

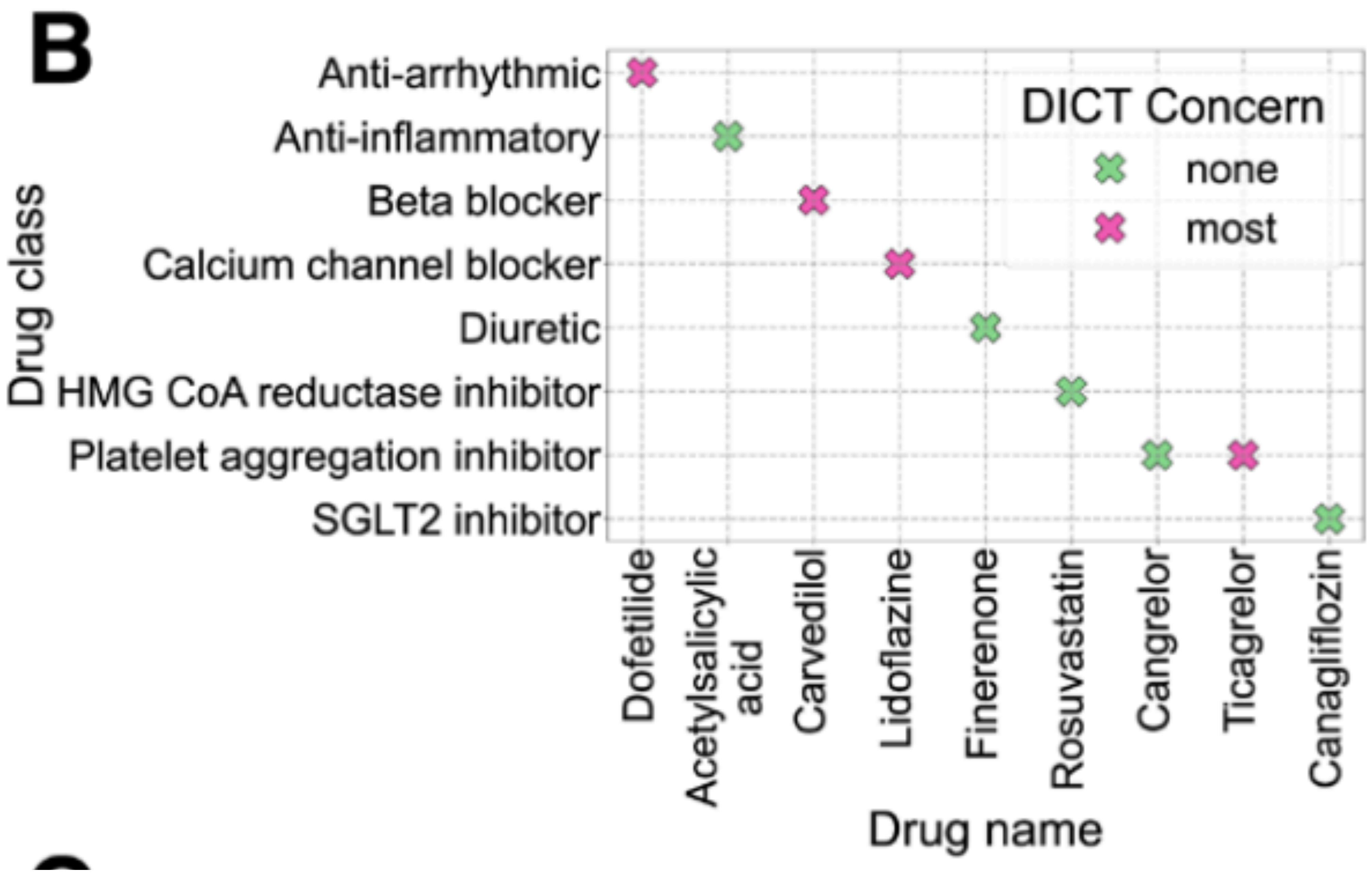
Intermediate step where labels (ADME properties) are available

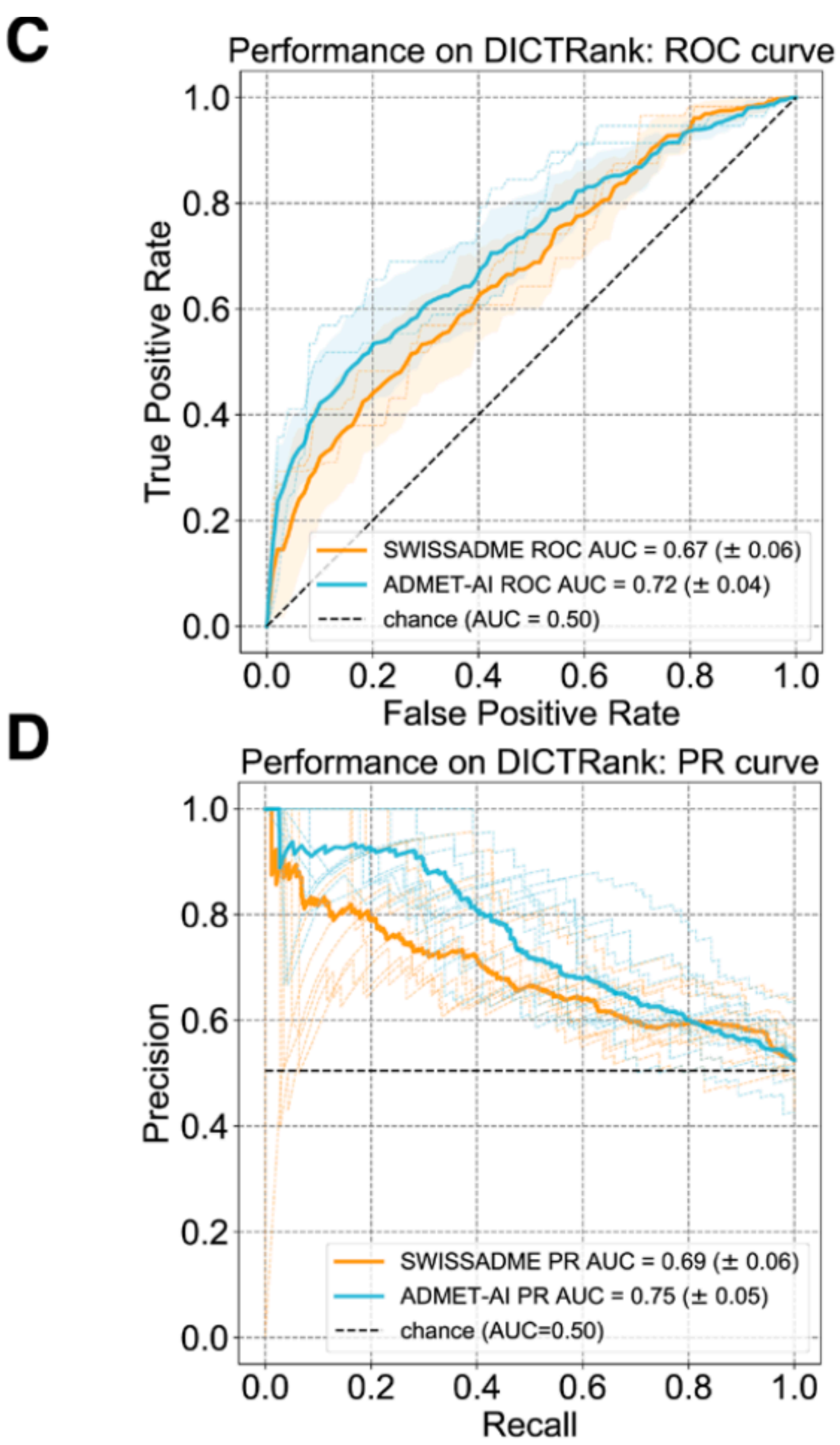
e.g. LogP, solubility, permeability, BBB, CYP activity, etc.

Intermediate step where labels (ADME properties) are available

https://doi.org/10.1161/CIRCULATIONAHA.124.070413

Drugs ranked and labels by concern in DCITRank



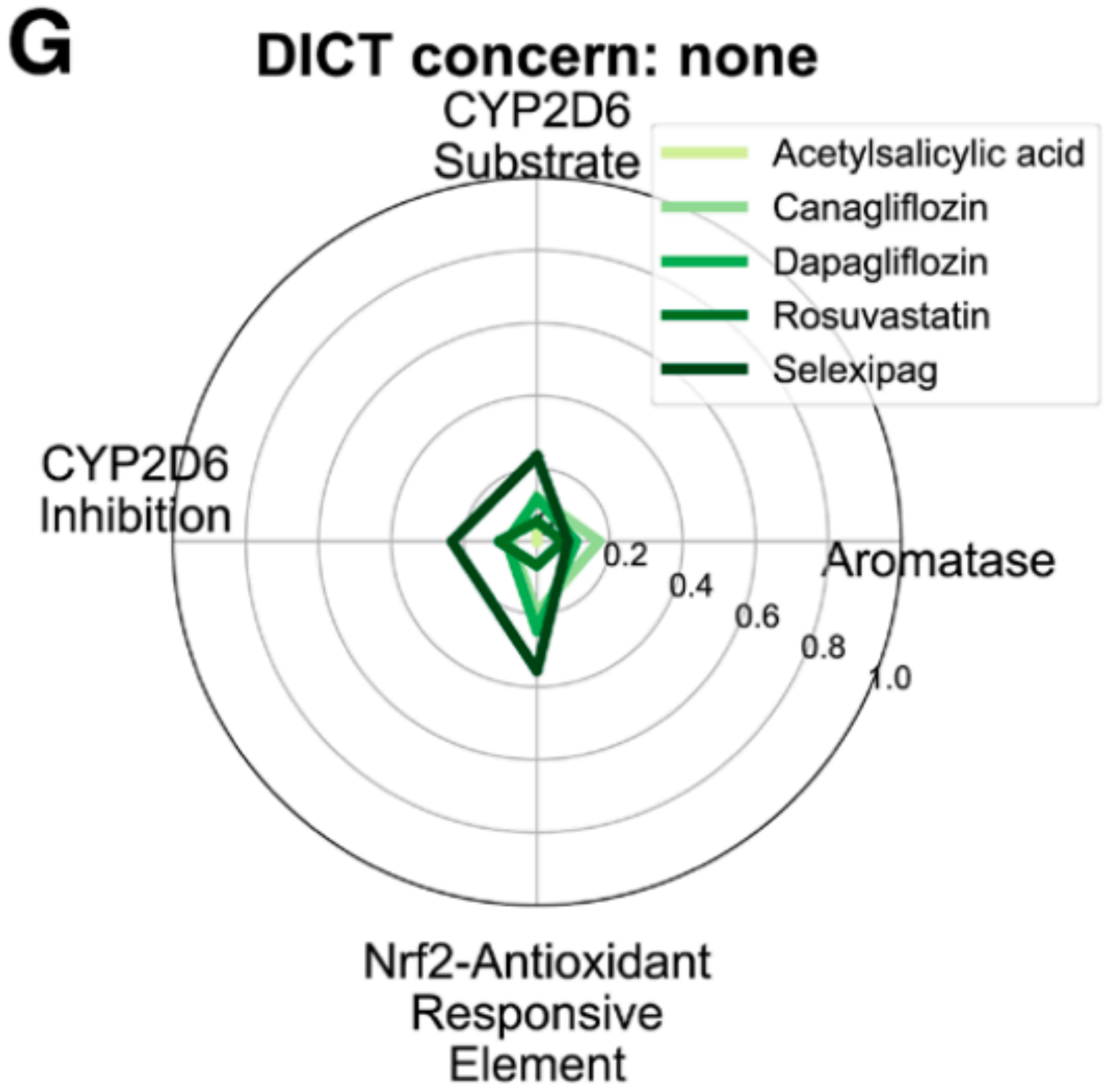


Slightly better performance

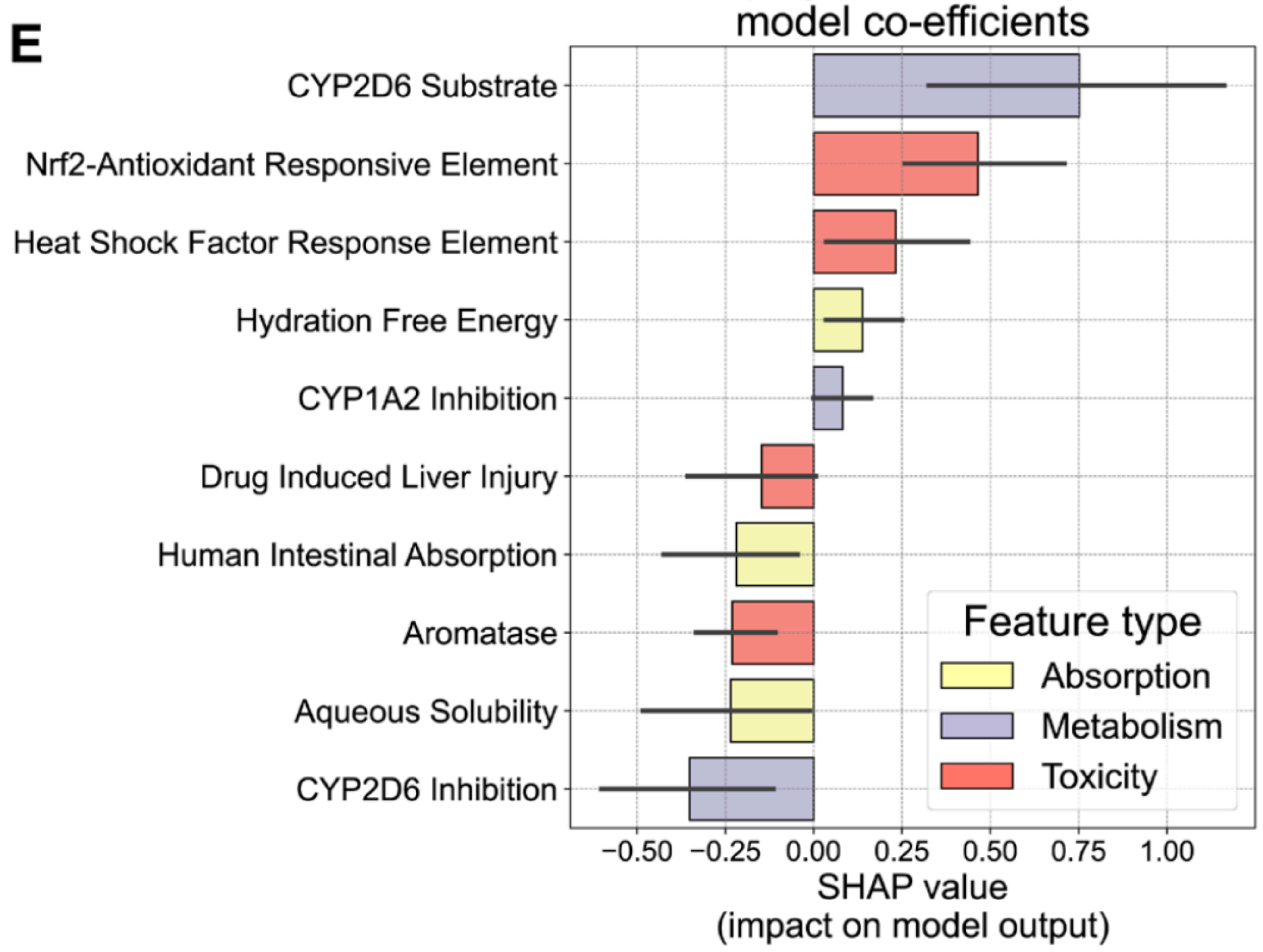
Most important features make sense

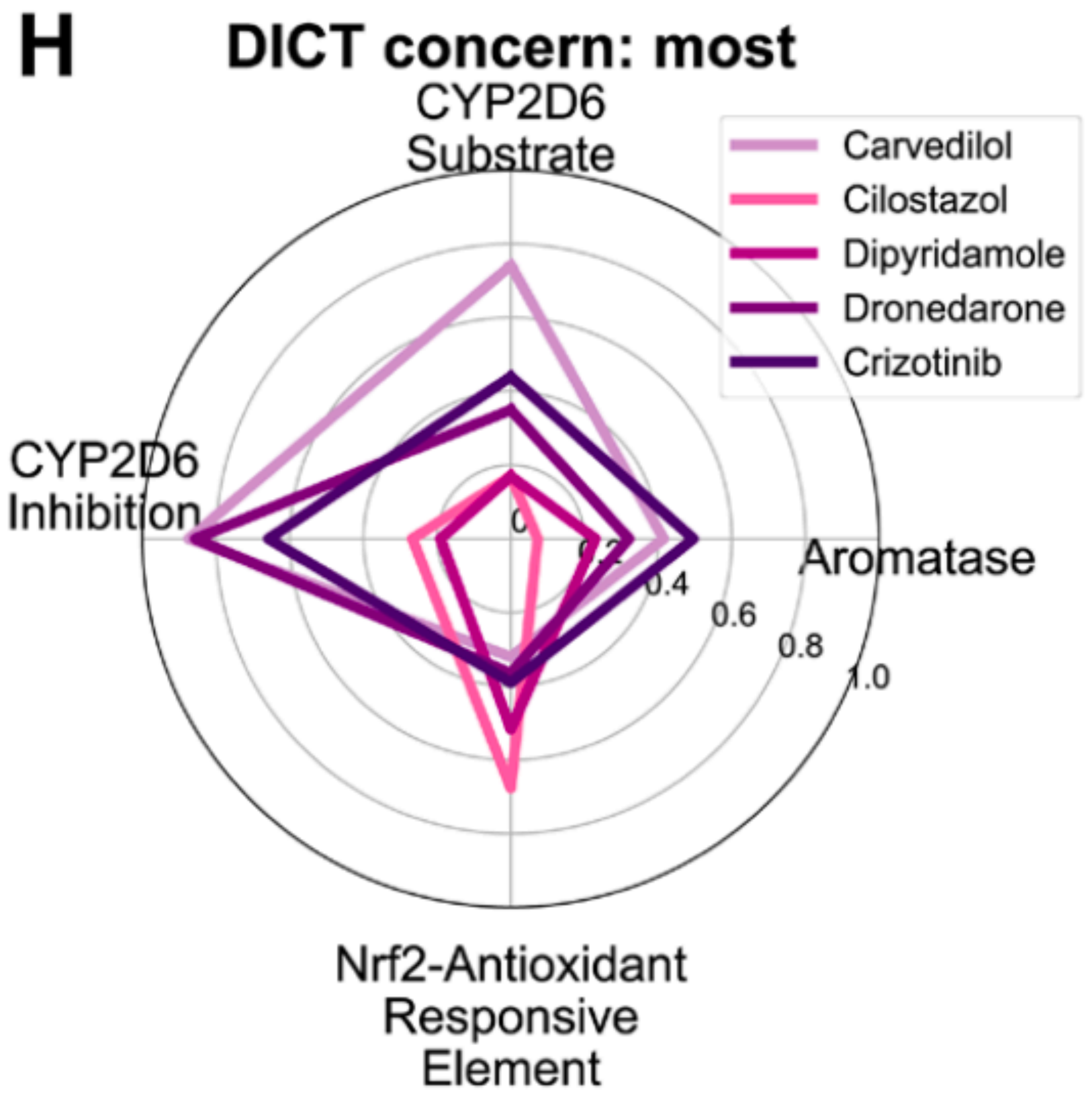
CYP2D6 metabolizes:
β-Blockers → Metoprolol, Carvedilol, Propranolol (heart failure, hypertension)
Antiarrhythmics → Flecainide, Encainide (arrhythmias)
Calcium channel blockers → Some **Dihydropyridines** (e.g., Nifedipine)
ACE inhibitors & ARBs → Some undergo minor CYP2D6 metabolism

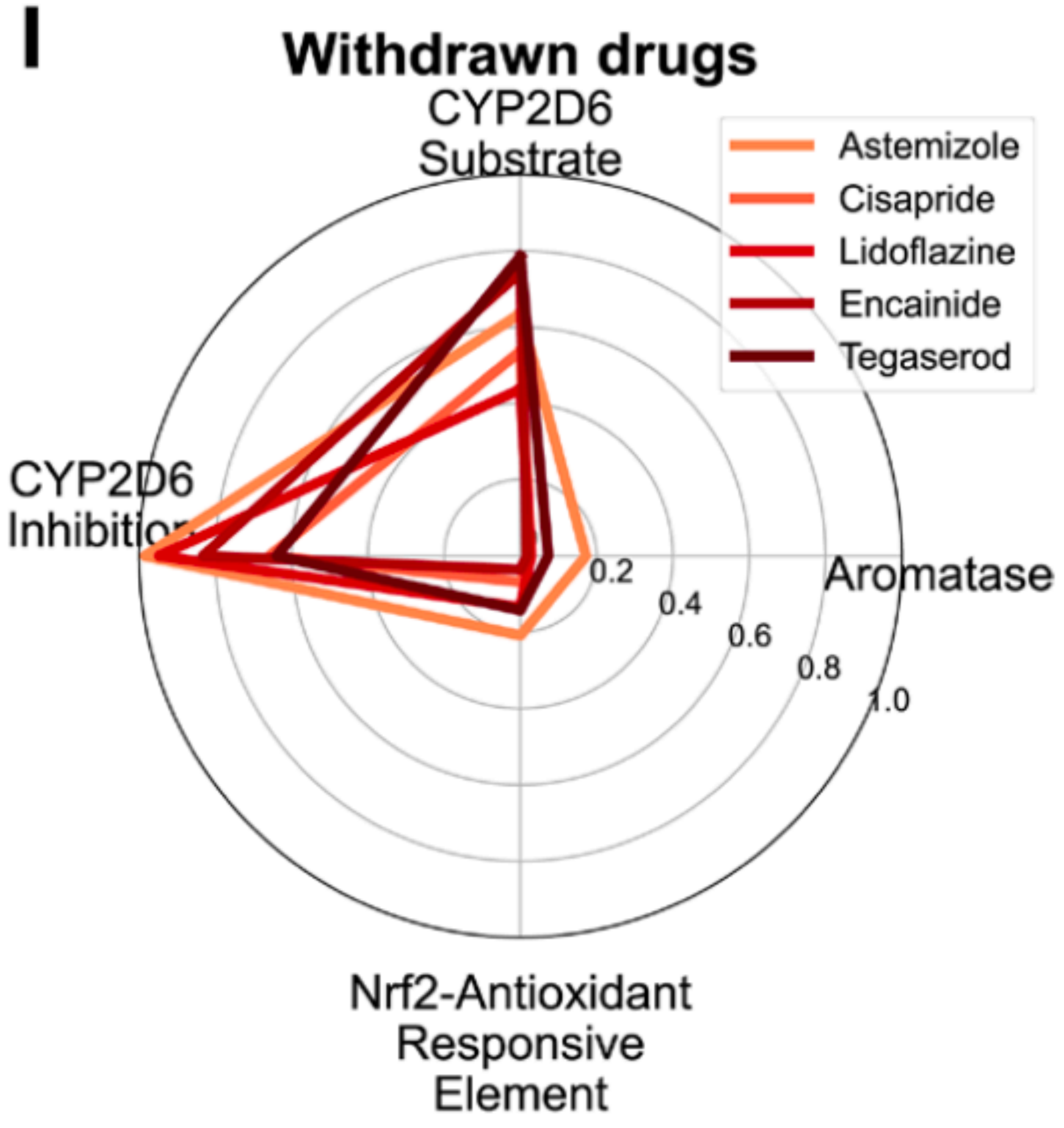
Cute way to show that these are important features



Shapley values to explain model co-efficients



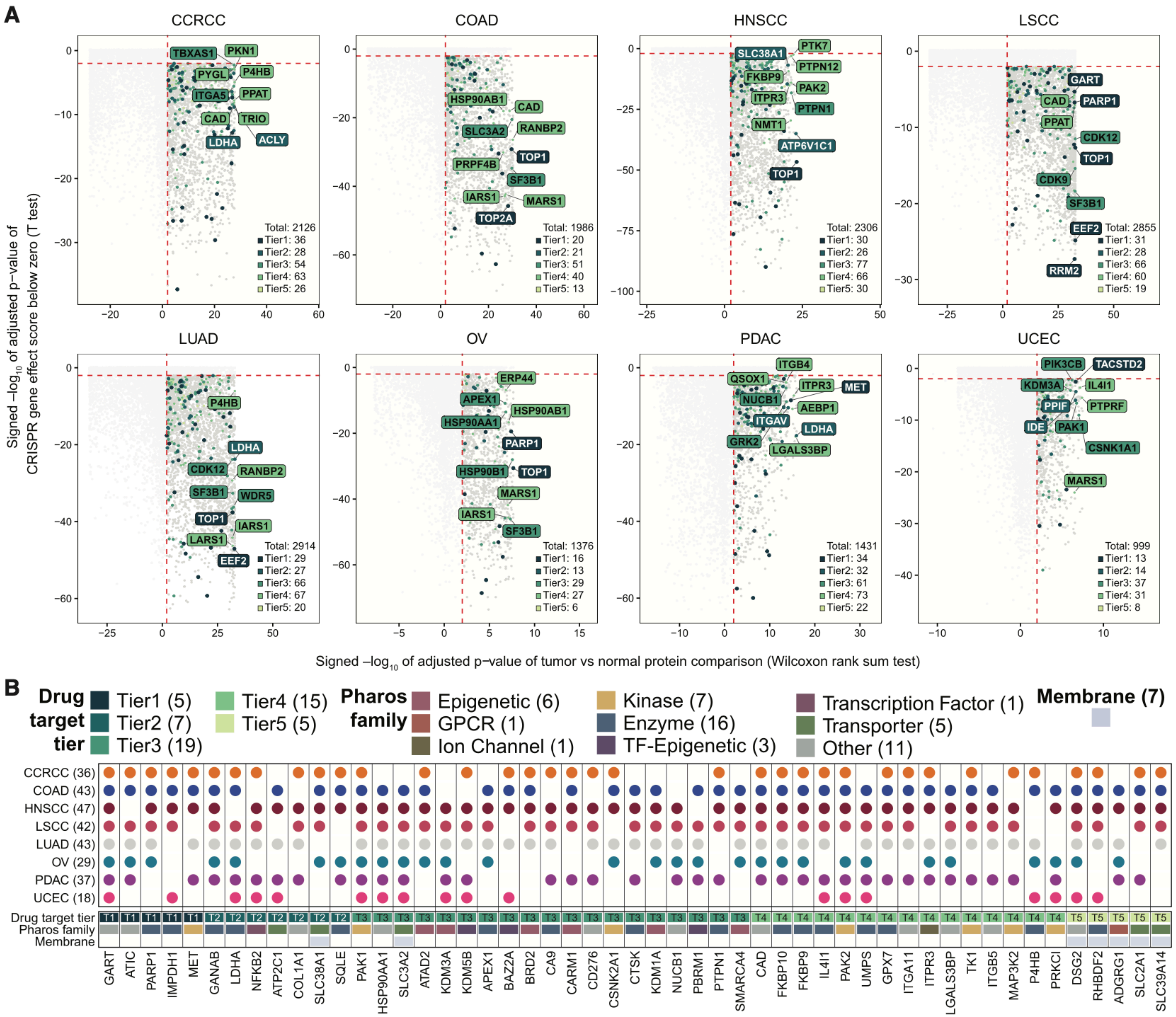




Pan-cancer proteogenomics expands the landscape of therapeutic targets (Savage et al, *Cell*)

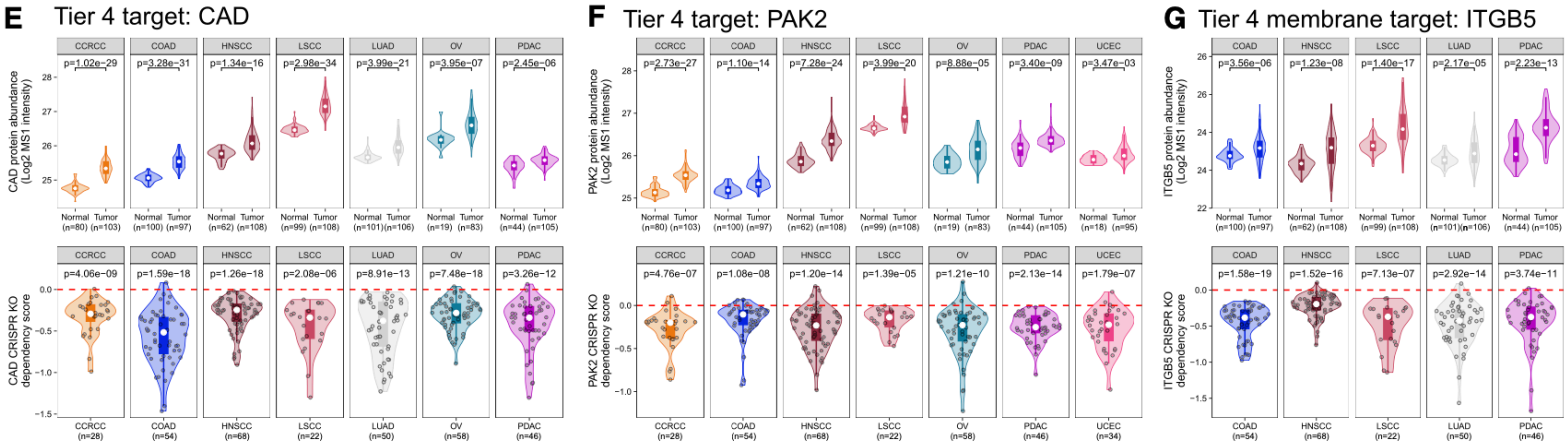
- Goal: To identify new therapeutic targets by integrating pan-cancer proteogenomic data
- Method:
  - Multifaceted and multimodal computational analysis strategy
    - Link to drug target databases; use synthetic lethal data to find pairs of cancer drivers that could be targets; identify hyper expressed or overactive proteins; prioritized neoantigens
  - Couple with experimental validation (binding affinity, SL screens, etc)
- Result: Identified and characterized 2,863 druggable proteins across five target tiers
- Conclusion: Multimodal data integration makes a stronger case for these putative targets.

Identify upregulated genes as potential targets by cancer type



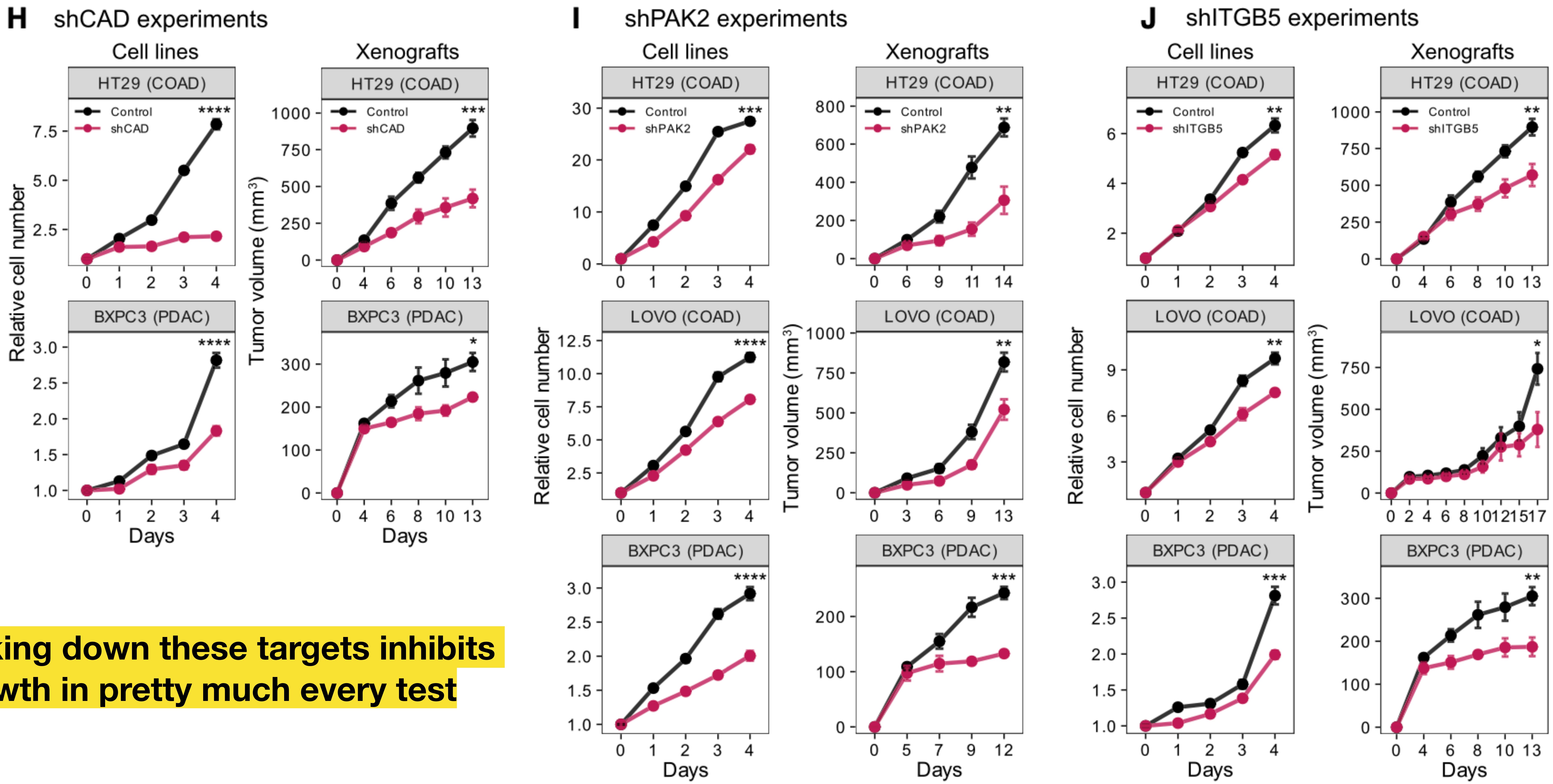
Non-pan-essential targets shared by 5+ cancer types

Confirmed putative targets are upgregulated in tumor (vs. normal) as predicted



And so with CRISPR that these cells are dependent on these genes

Knock down experiments in mouse xenografts



Knocking down these targets inhibits growth in pretty much every test

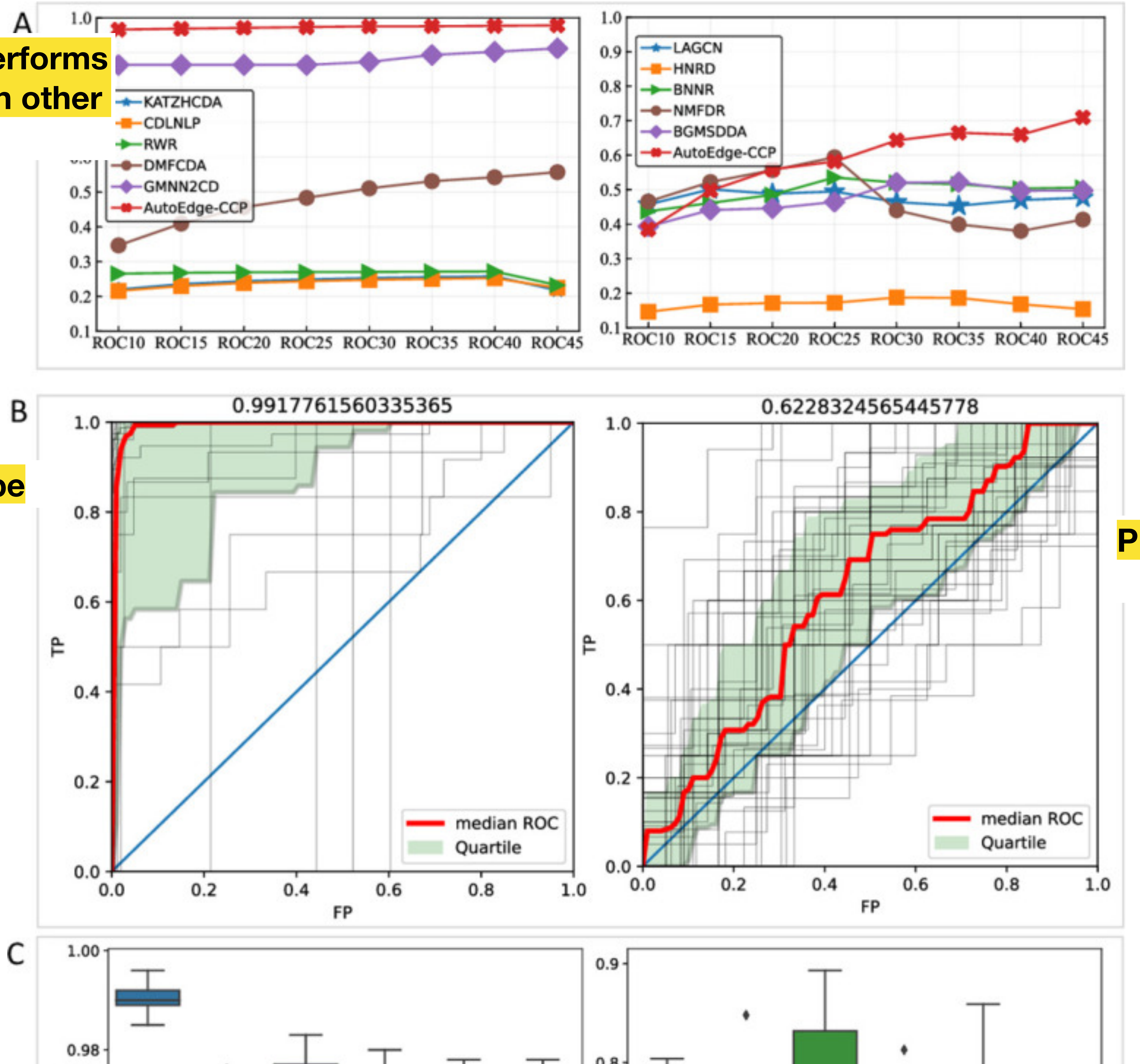
Trainee Spotlight

PUSHKALA JAYARAMAN

ICAHN SCHOOL OF MEDICINE MOUNT SINAI, NY

AutoEdge-CCP: A novel approach for predicting cancer-associated circRNAs and drugs based on automated edge embedding (Chen et al, *PLoS Comp Bio*)

- Goal: Develop a computational framework to predict associations between circular RNAs (circRNAs), drugs, and cancer
- Method:
  - Multi-source heterogenous network (circRNA, drugs, and cancer)
  - Use GNNs to build embeddings
  - Uses a Learn-to-Rank (LTR) framework to learn relationship between circRNA and cancer (and drug-cancer)
- Result:
  - Big improvements over baselines (AUROC of 0.989 vs 0.700)
  - Edge embeddings offer mechanistic interpretations
- Conclusion: The future is joint…embeddings

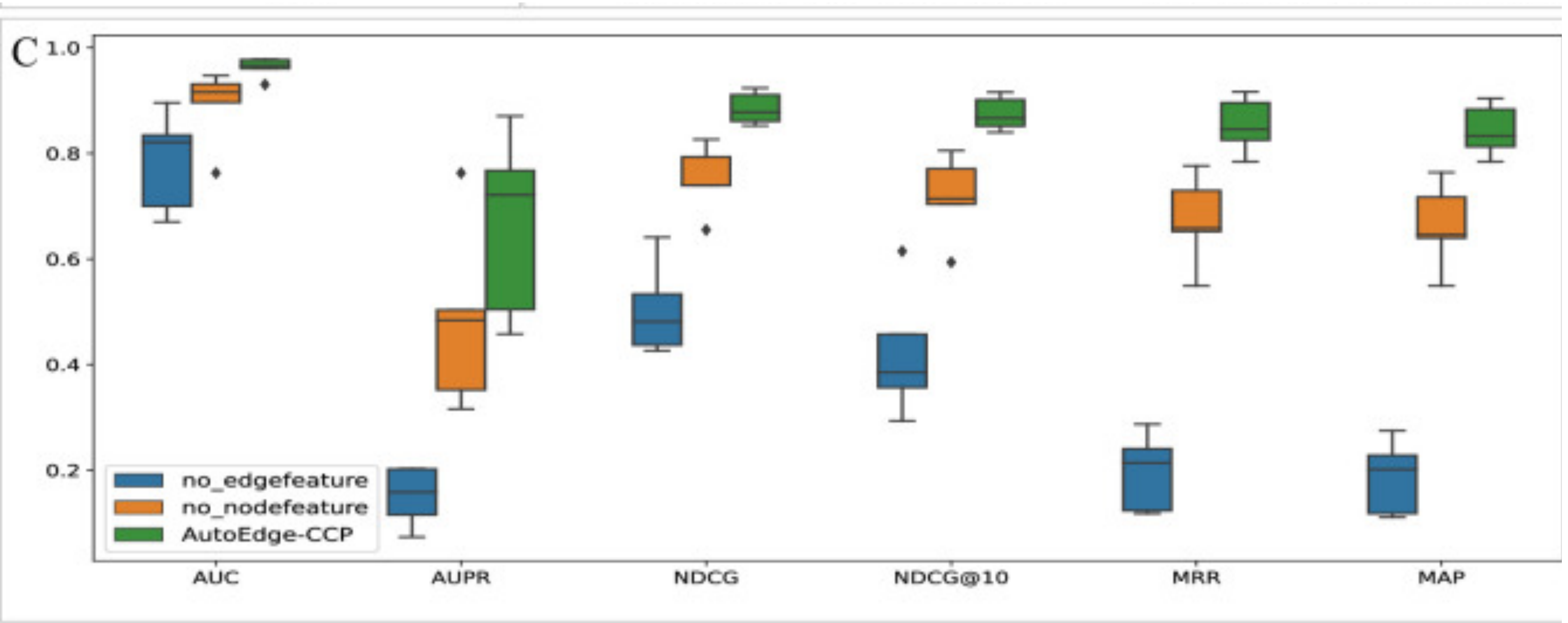


AutoEdge-CCP performs better overall than other methods

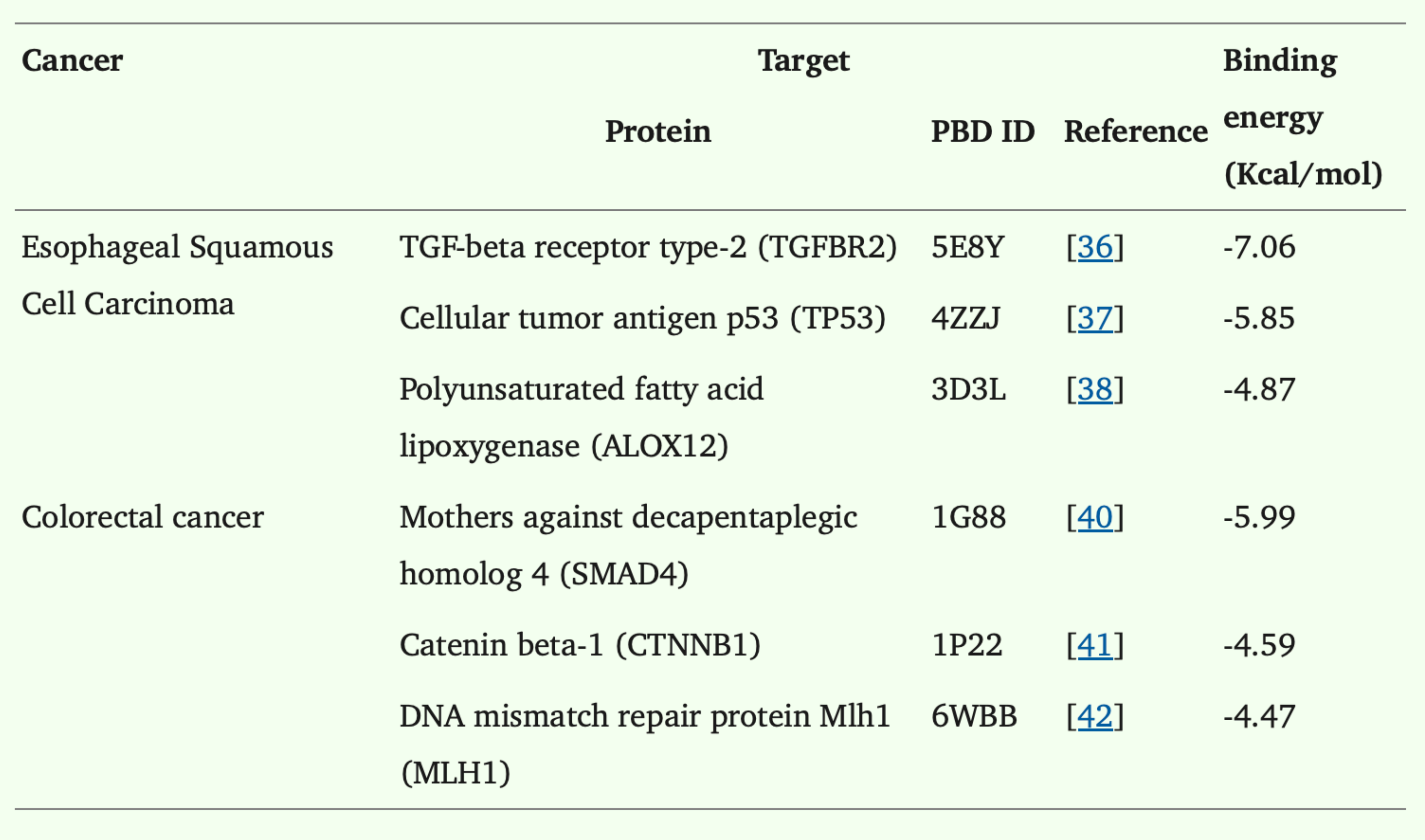
Classifying cancer type

Predicting missing (held out) edges

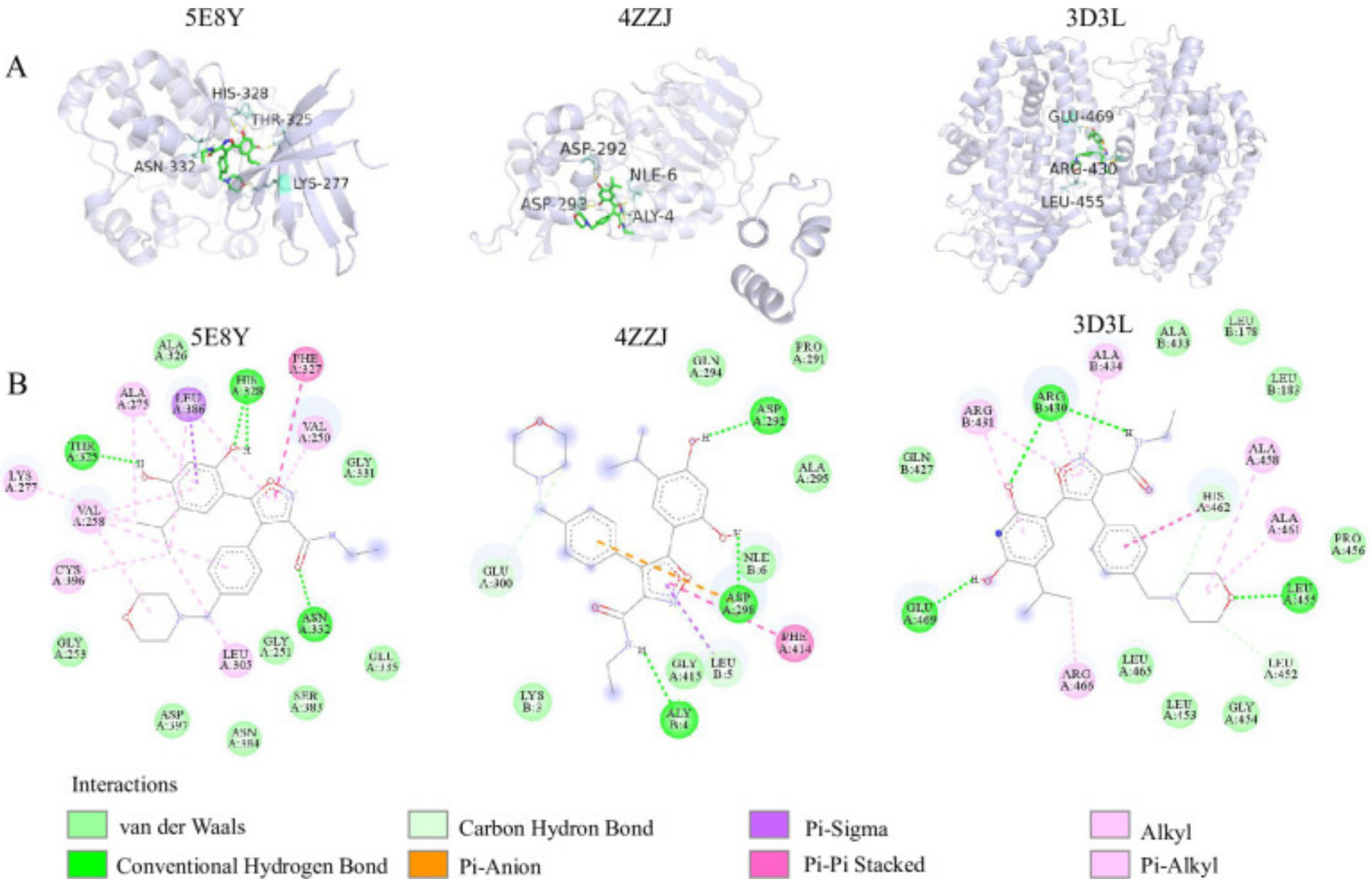
Both node and edge embeddings are needed



Novel drug-cancer predictions held up in computational docking experiments



Cancer	Target			Binding energy (Kcal/mol)
	Protein	PBD ID	Reference	
Esophageal Squamous Cell Carcinoma	TGF-beta receptor type-2 (TGFBR2)	5E8Y	[36]	-7.06
	Cellular tumor antigen p53 (TP53)	4ZZJ	[37]	-5.85
	Polyunsaturated fatty acid lipoxygenase (ALOX12)	3D3L	[38]	-4.87
Colorectal cancer	Mothers against decapentaplegic homolog 4 (SMAD4)	1G88	[40]	-5.99
	Catenin beta-1 (CTNNB1)	1P22	[41]	-4.59
	DNA mismatch repair protein Mlh1 (MLH1)	6WBB	[42]	-4.47

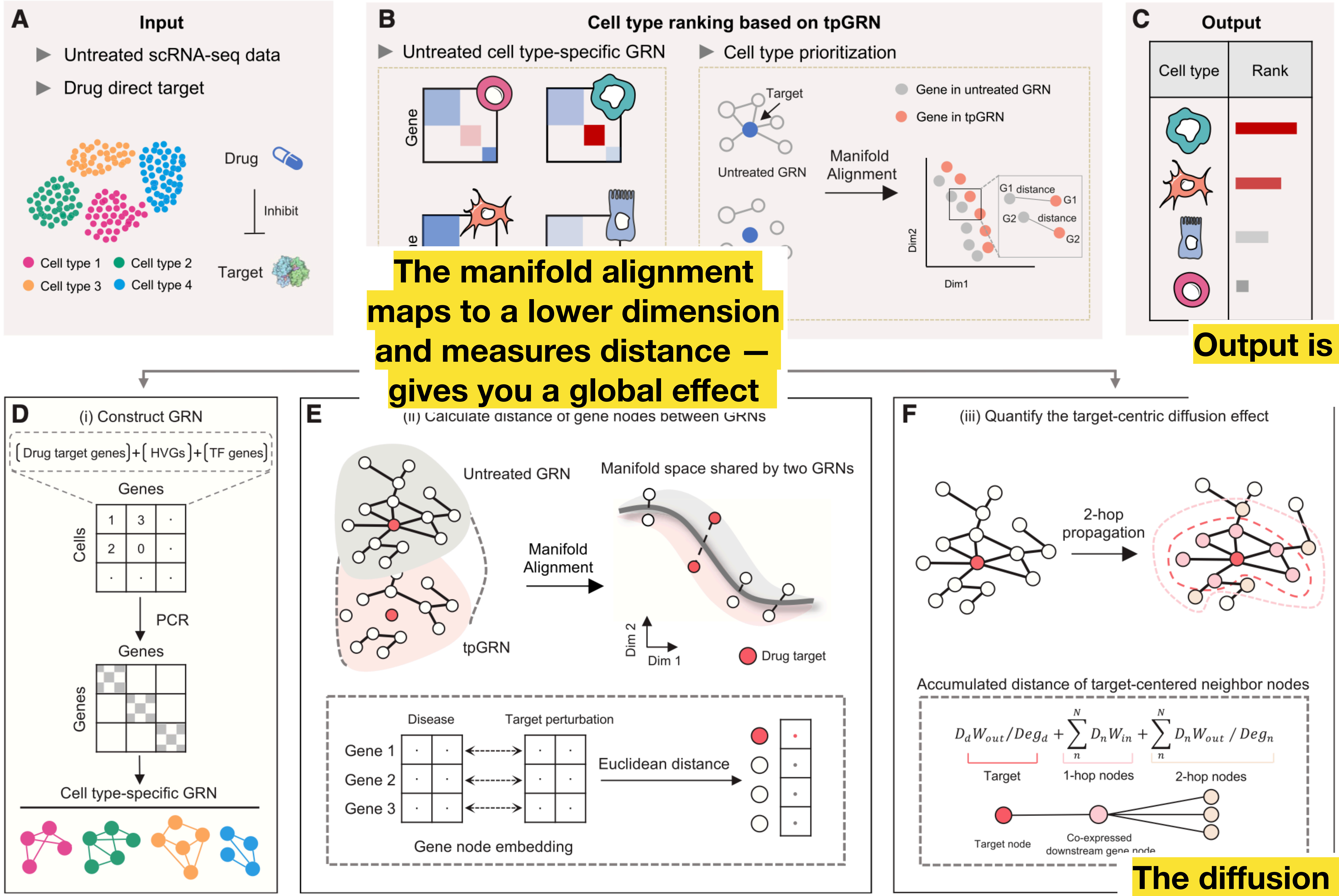


scRank infers drug-responsive cell types from untreated scRNA-seq data using a target-perturbed gene regulatory network (Li et al, *Cell Reports Medicine*)

- Goal: Identify drug-responsive cell populations without requiring post-treatment transcriptomic data
- Method:
  - Use scRNA-seq data to generate a treatment naive regulatory network
  - Simulate drug perturbation by removing the target from the network (assumes inhibition and only one target)
  - Use manifold alignment and network diffusion to measure the global and local effects on the network
- Result:
  - Outperforms existing methods (e.g. Augur and DEG) in simulations and real data
  - Identified well known cell type-drug associations from the literature
  - Experimental validation three different disease contexts
- Conclusion: I didn’t realize we could still do single cell analysis without deep learning

#IS25 #YIR25 @proftatonetti @tatonetti.bsky.social https://doi.org/10.1016/j.xcrm.2024.101568

Input is a bunch of single cell RNAseq data and knowledge on drug targets


A
Input
▶ Untreated scRNA-seq data
▶ Drug direct target
Drug
Inhibit
Target
Cell type 1
Cell type 2
Cell type 3
Cell type 4
B
Cell type ranking based on tpGRN
▶ Untreated cell type-specific GRN
Gene
▶ Cell type prioritization
Target
Untreated GRN
Gene in untreated GRN
Gene in tpGRN
Manifold Alignment
G1 distance
G1
G2
distance
G2
Dim2
Dim1
C
Output
Cell type
Rank
D
(i) Construct GRN
Drug target genes + HVGs + TF genes
Genes
Cells
1
3
2
0
PCR
Genes
Genes
Cell type-specific GRN
E
(ii) Calculate distance of gene nodes between GRNs
Untreated GRN
tpGRN
Manifold Alignment
Manifold space shared by two GRNs
Dim 2
Dim 1
Drug target
Disease
Target perturbation
Gene 1
Gene 2
Gene 3
Euclidean distance
Gene node embedding
F
(iii) Quantify the target-centric diffusion effect
2-hop propagation
Accumulated distance of target-centered neighbor nodes
$\underbrace{D_d W_{out}/Deg_d}_{\text{Target}} + \underbrace{\sum_n^N D_n W_{in}}_{\text{1-hop nodes}} + \underbrace{\sum_n^N D_n W_{out} / Deg_n}_{\text{2-hop nodes}}$
Target node
Co-expressed downstream gene node

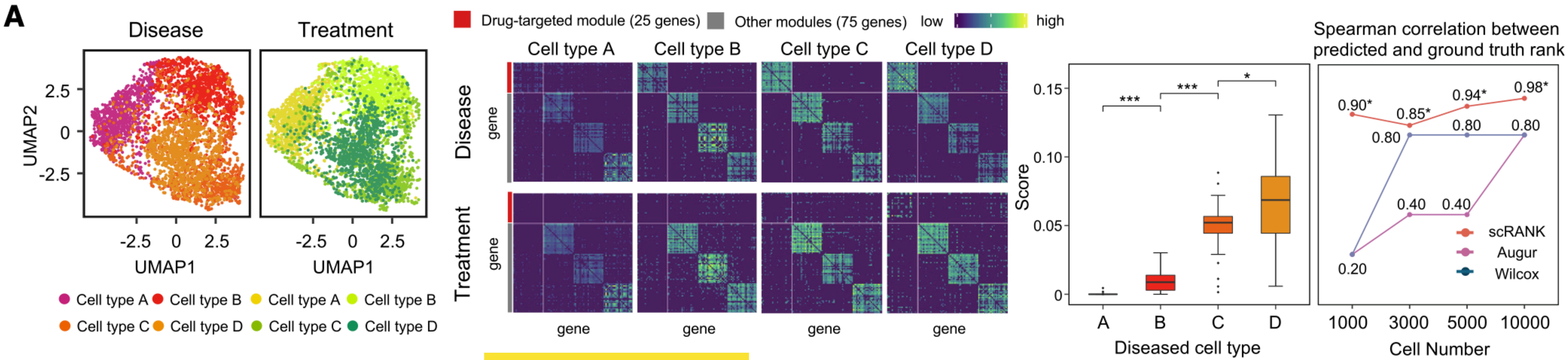
The manifold alignment maps to a lower dimension and measures distance — gives you a global effect

Output is a ranking of cell types

The diffusion metric maps out to the drugs neighbors, giving you a local effect

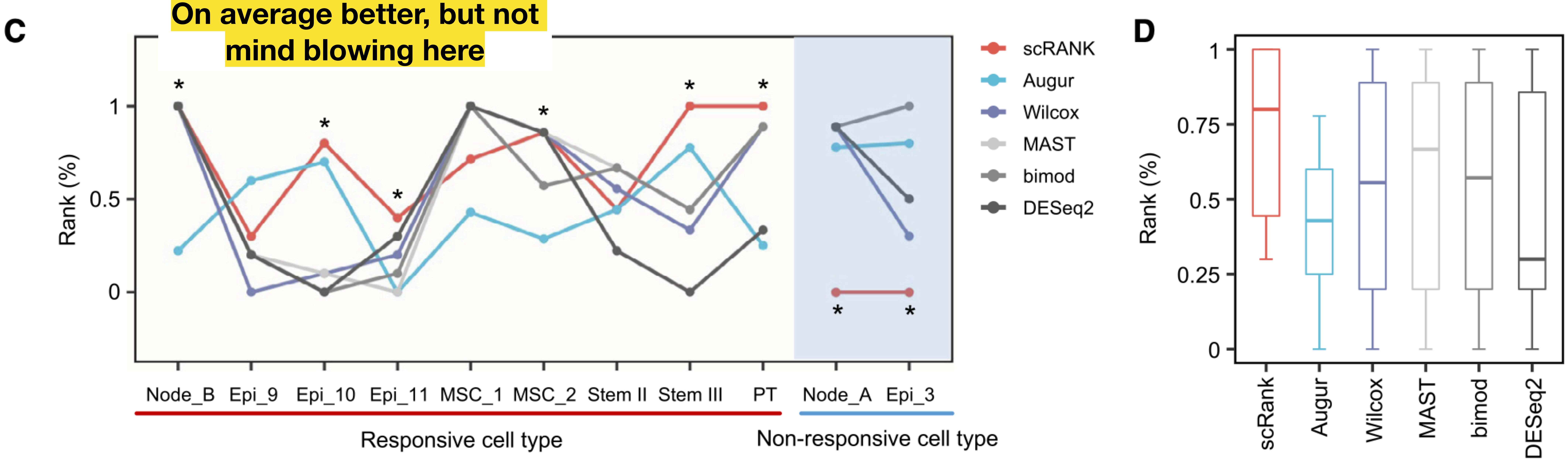
Simulations

scRANK better captures which cell types are affected



Notice modules are “knocked out”

Validation in real data



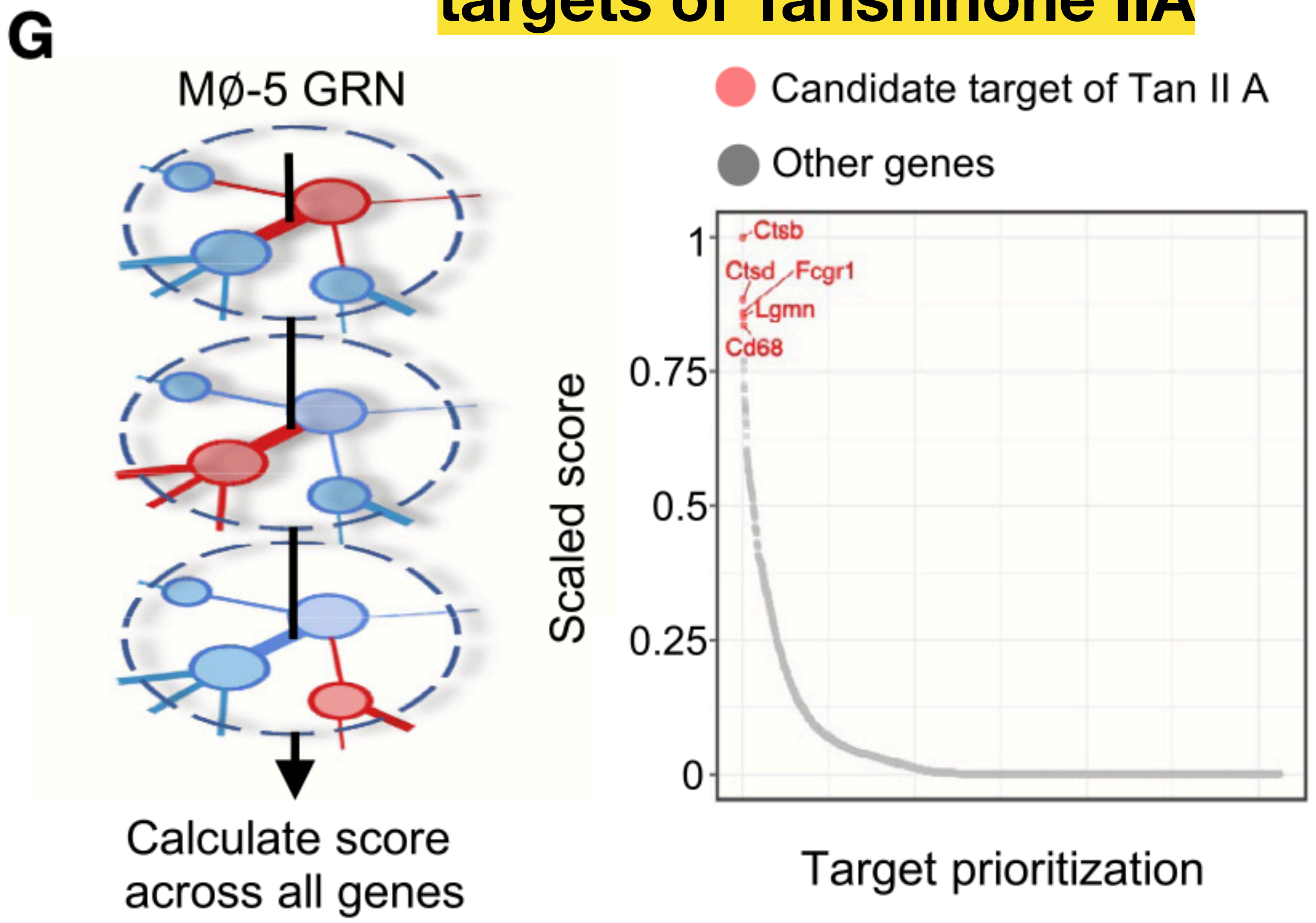
On average better, but not mind blowing here

Better on average, too

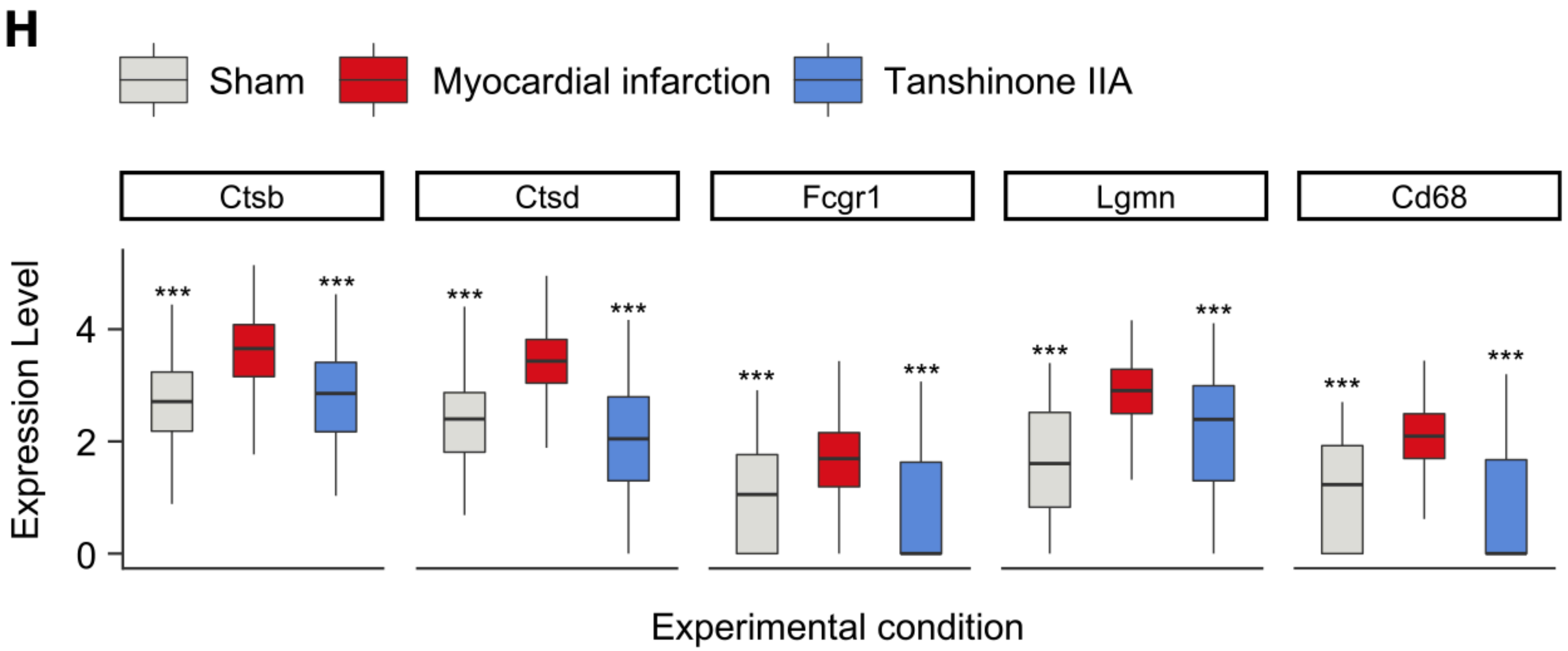
Much better and ranking non-responsive cell types lower

Diving in on myocardial infarction and tanshinone IIA

Can use scRANK to predict off-targets of Tanshinone IIA



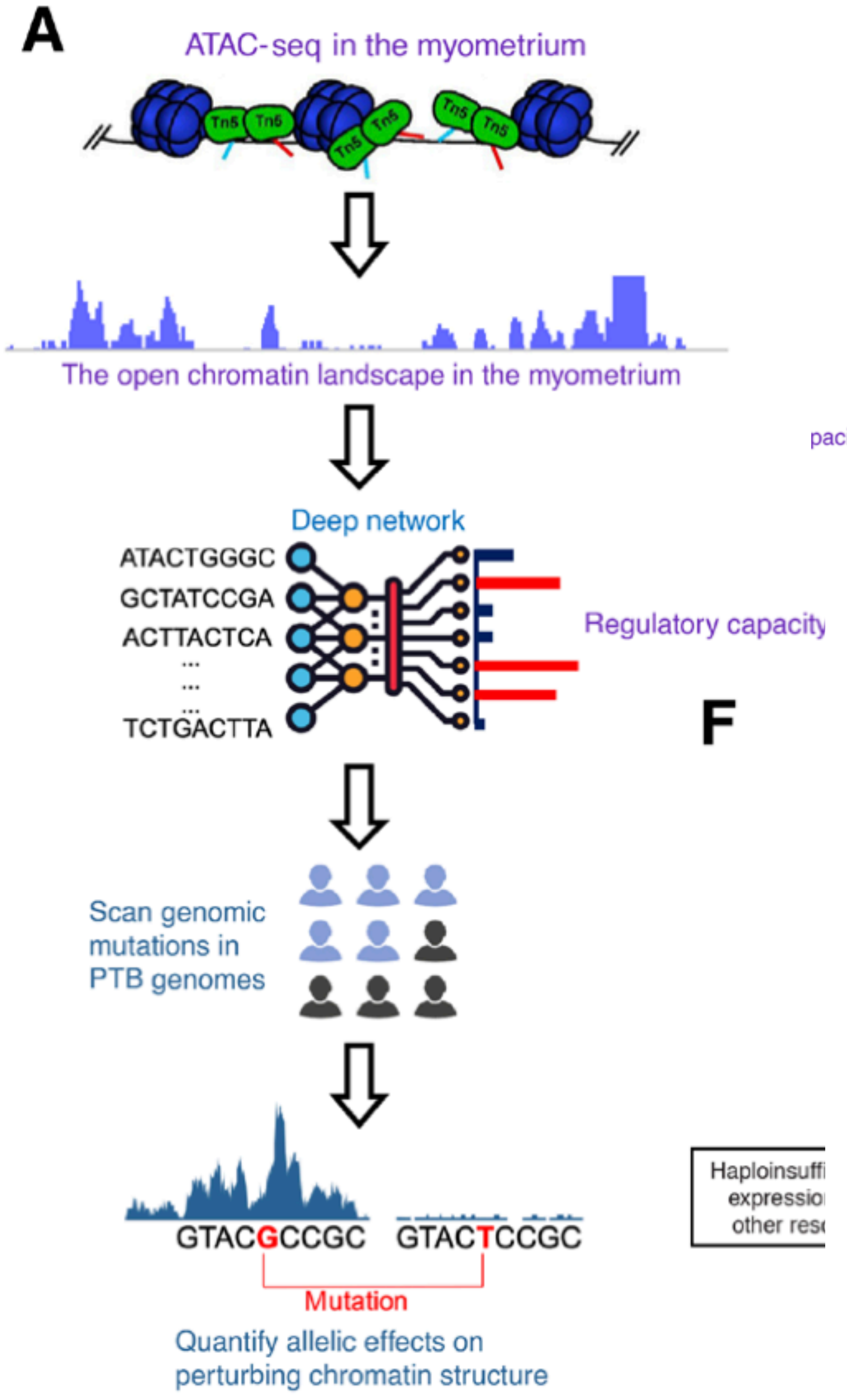
And the expression of these are inhibited by Tanshinone IIA in experiments



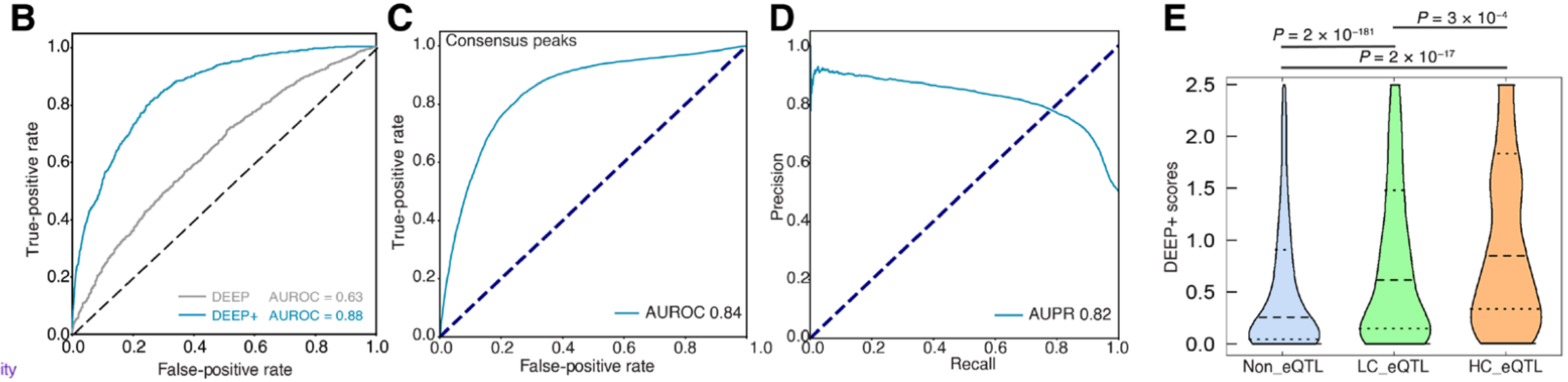
Integrative analysis of noncoding mutations identifies the druggable genome in preterm birth (Wang et al, *Science Advances*)

- Goal: Identify genetics of preterm birth; use it explain variation of response to progestin therapy; use it to identify new drug candidates
- Method:
  - Introduce DEEP+ to analyze genomic variants for effect on chromatin accessibility (genetic associations have been in noncoding regions)
  - Develop a bayesian method (BEAR) to integrate output of DEEP+ with epigenetic and transciptomic data to compute a posterior for each genomic locus
    - >High BEAR score means: strong GWAS, disrupts chromatin accessibility, affects dosage-sensitive genes, alters gene expression in the uterus
- Result:
  - Found ~1k genomic loci with high BEAR scores, including those linked to previously unidentified genes affecting muscle relaxation and inflammatory pathways
  - Mutation burden predicted response to progestin therapy
  - Discovered and validated new small molecule to treat spontaneous preterm birth
- Conclusion: 👏

DEEP+ setup


A
ATAC-seq in the myometrium
Tn5
The open chromatin landscape in the myometrium
Deep network
ATACTGGGC
GCTATCCGA
ACTTACTCA
...
TCTGACTTA
Regulatory capacity
F
Scan genomic mutations in PTB genomes
GTACGCCGC
GTACTCCGC
Mutation
Quantify allelic effects on perturbing chromatin structure
Haploinsuffi
expressio
other res
pacity

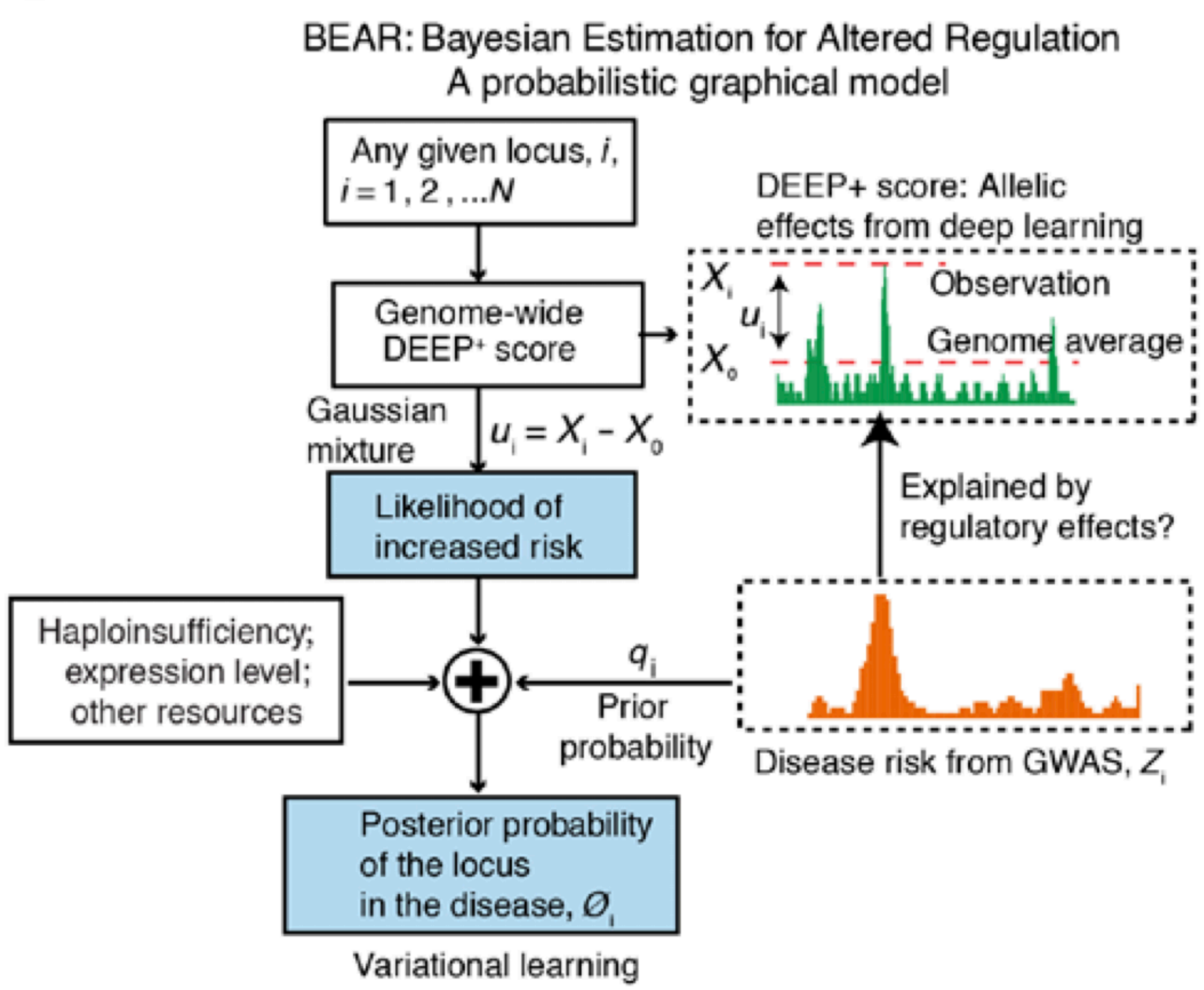
DEEP+ improves on DEEP (s/CNN/ResNet/g and some other stuff)



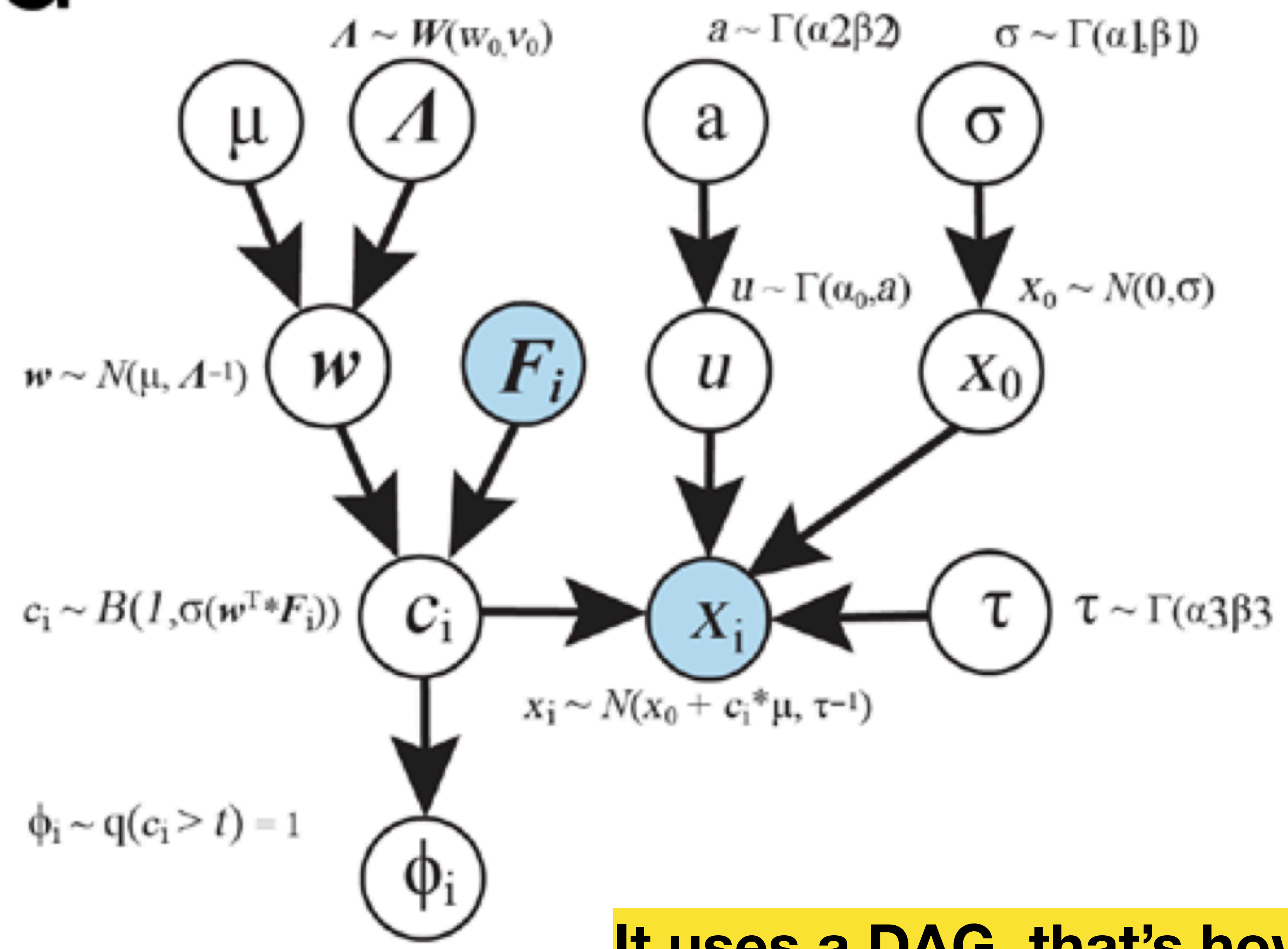
High eQTLs have higher DEEP+ scores

BEAR is a bayesian multimodal integration method

F


BEAR: Bayesian Estimation for Altered Regulation
A probabilistic graphical model
Any given locus, i, $i = 1, 2, \dots N$
Genome-wide DEEP+ score
DEEP+ score: Allelic effects from deep learning
X_i
u_i
X_0
Observation
Genome average
Gaussian mixture
$u_i = X_i - X_0$
Likelihood of increased risk
Explained by regulatory effects?
Haploinsufficiency; expression level; other resources
q_i
Prior probability
Disease risk from GWAS, Z_i
Posterior probability of the locus in the disease, $Ø_i$
Variational learning

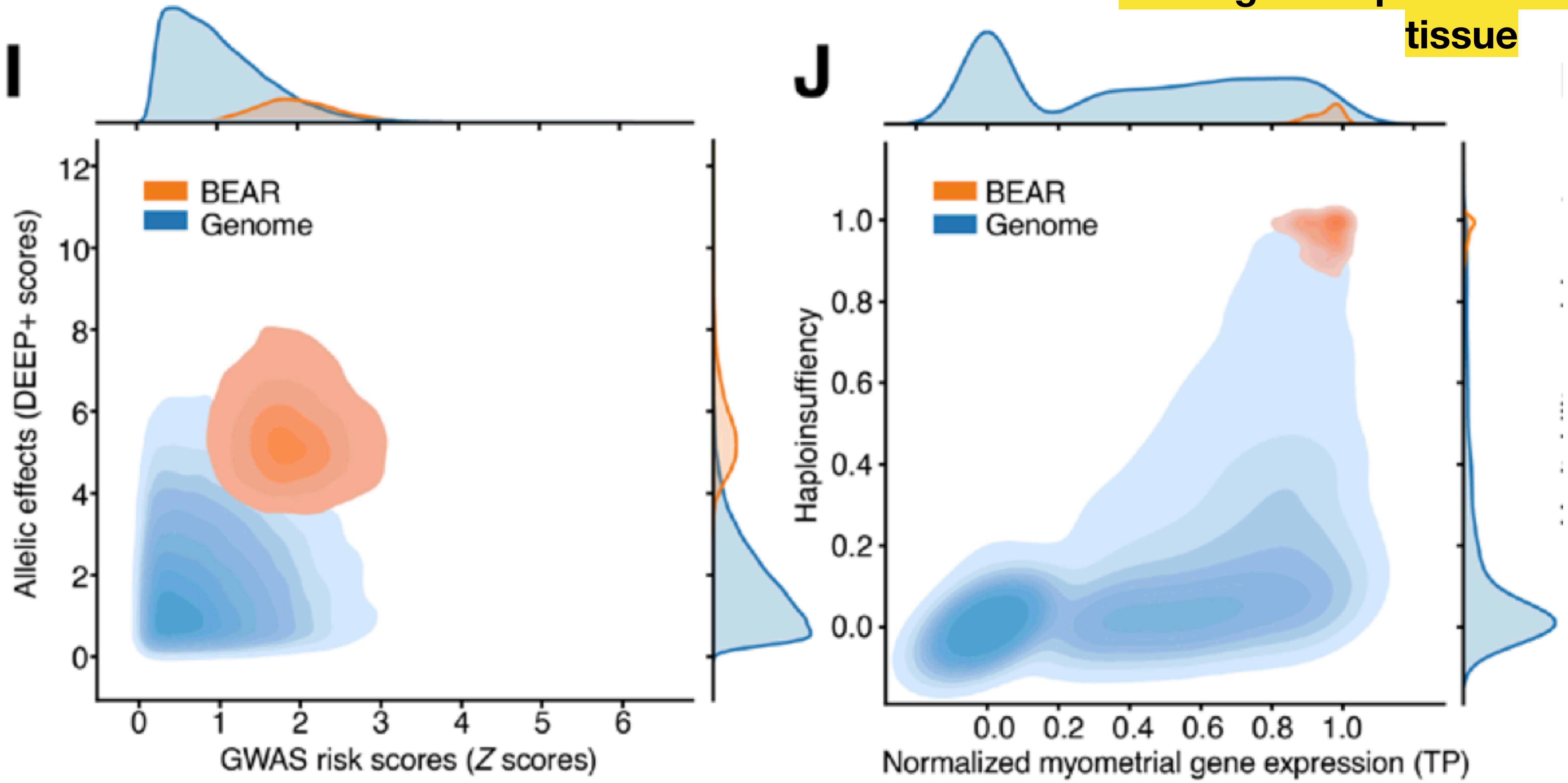
G


$\Lambda \sim W(w_0, v_0)$
$a \sim \Gamma(\alpha 2 \beta 2)$
$\sigma \sim \Gamma(\alpha 1 \beta 1)$
μ
Λ
a
σ
$u \sim \Gamma(\alpha_0, a)$
$x_0 \sim N(0, \sigma)$
$w \sim N(\mu, \Lambda^{-1})$
w
F_i
u
x_0
$c_i \sim B(1, \sigma(w^T * F_i))$
c_i
x_i
τ
$\tau \sim \Gamma(\alpha 3 \beta 3)$
$x_i \sim N(x_0 + c_i * \mu, \tau^{-1})$
$\phi_i \sim q(c_i > t) = 1$
ϕ_i

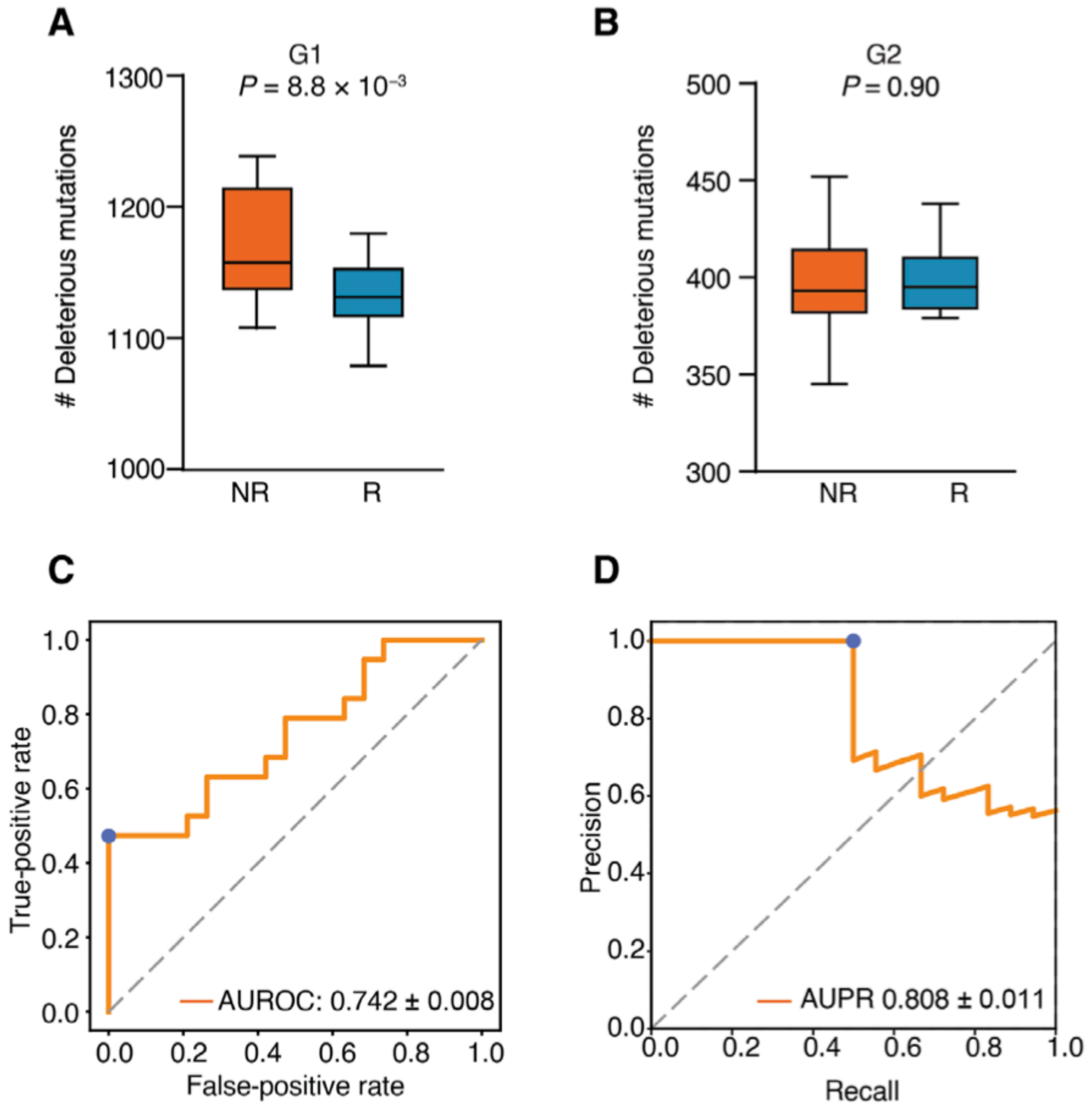
It uses a DAG, that's how you known it's Bayesian

BEAR loci have higher DEEP+ scores and strong GWAS risk associations

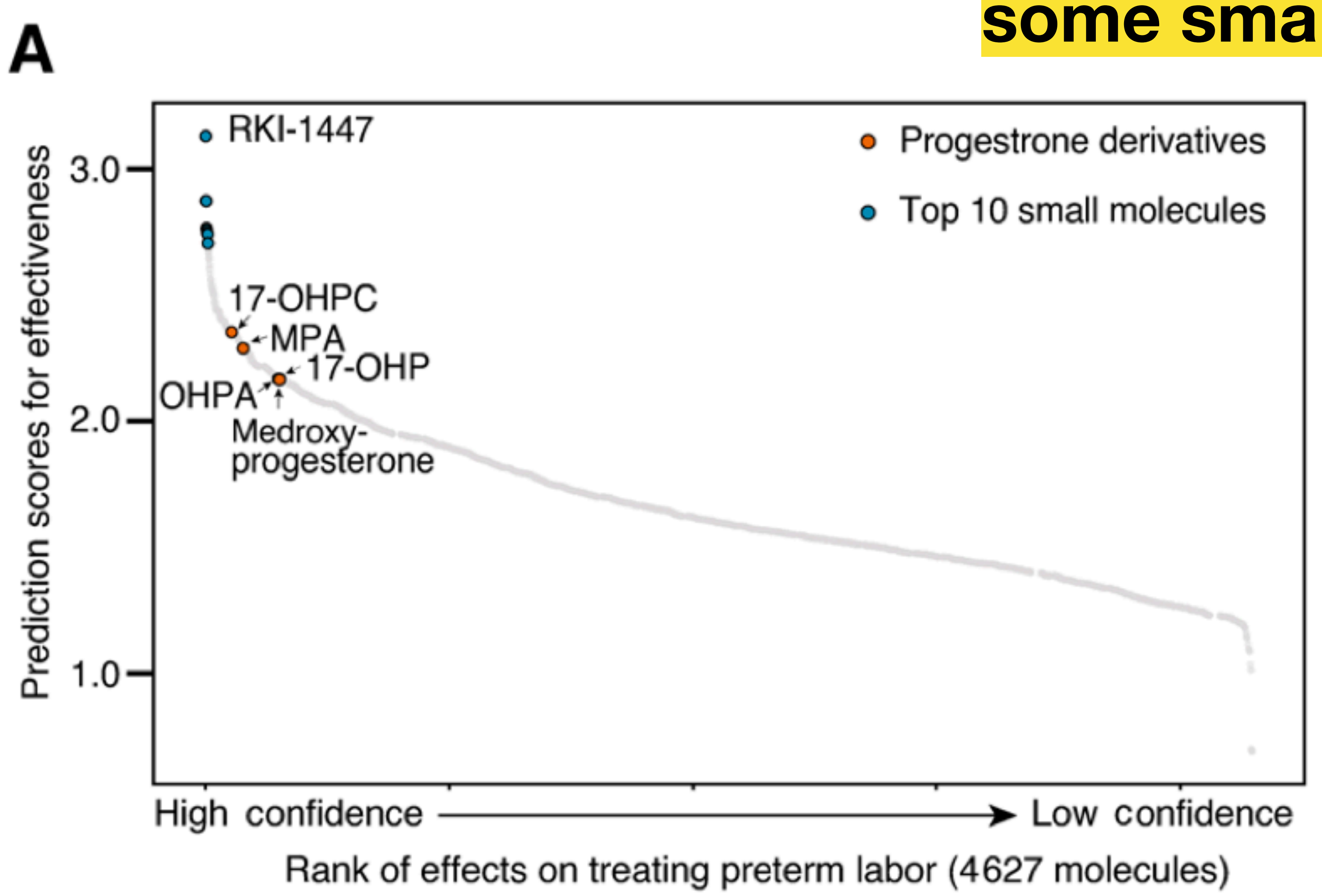
… and high haploinsufficiency scores and higher expression in myometrial tissue



Chromatin accessibility mutation burden is predictive of progestin response



Drug discovery rediscovered progesterone derivatives and also some small molecules



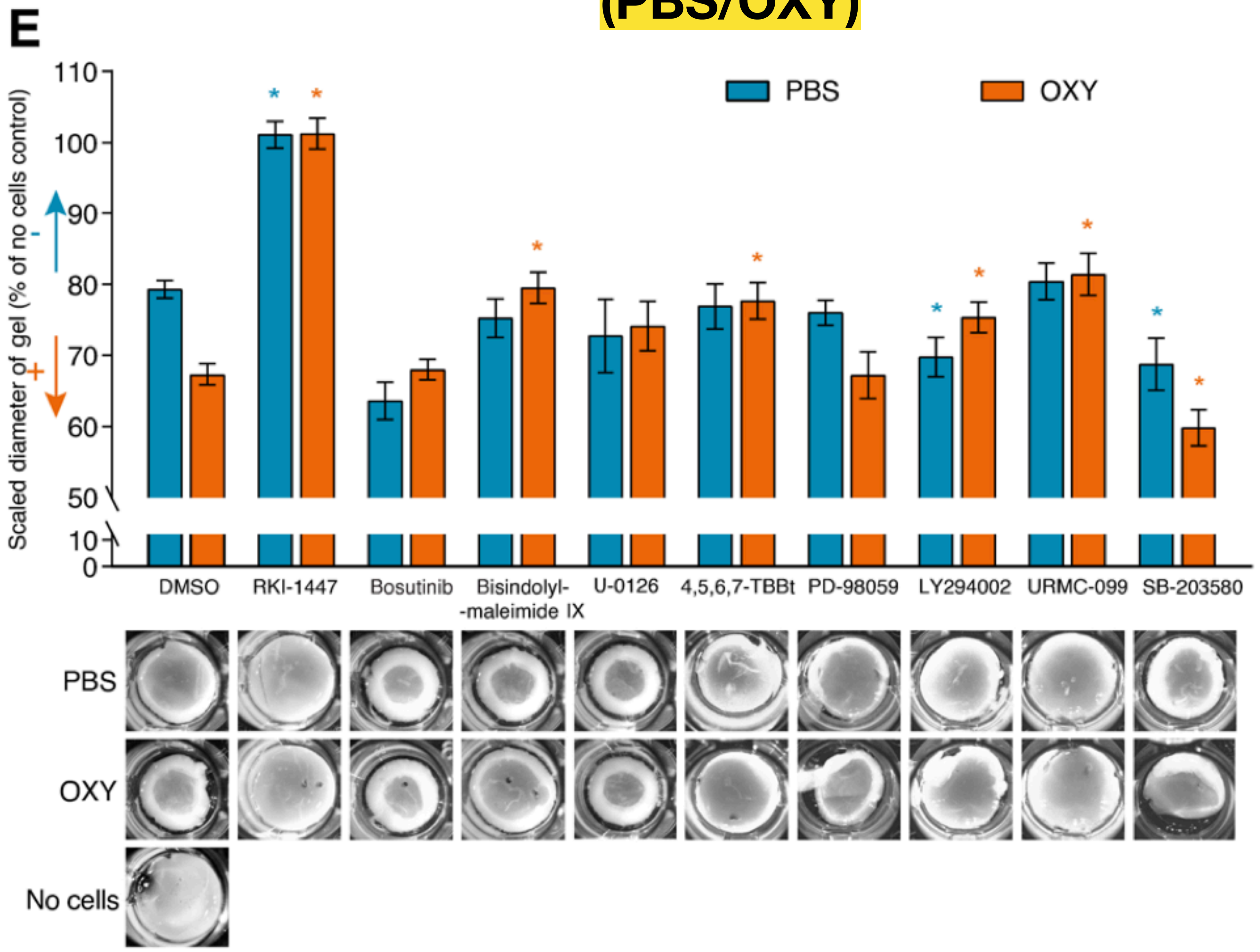
B

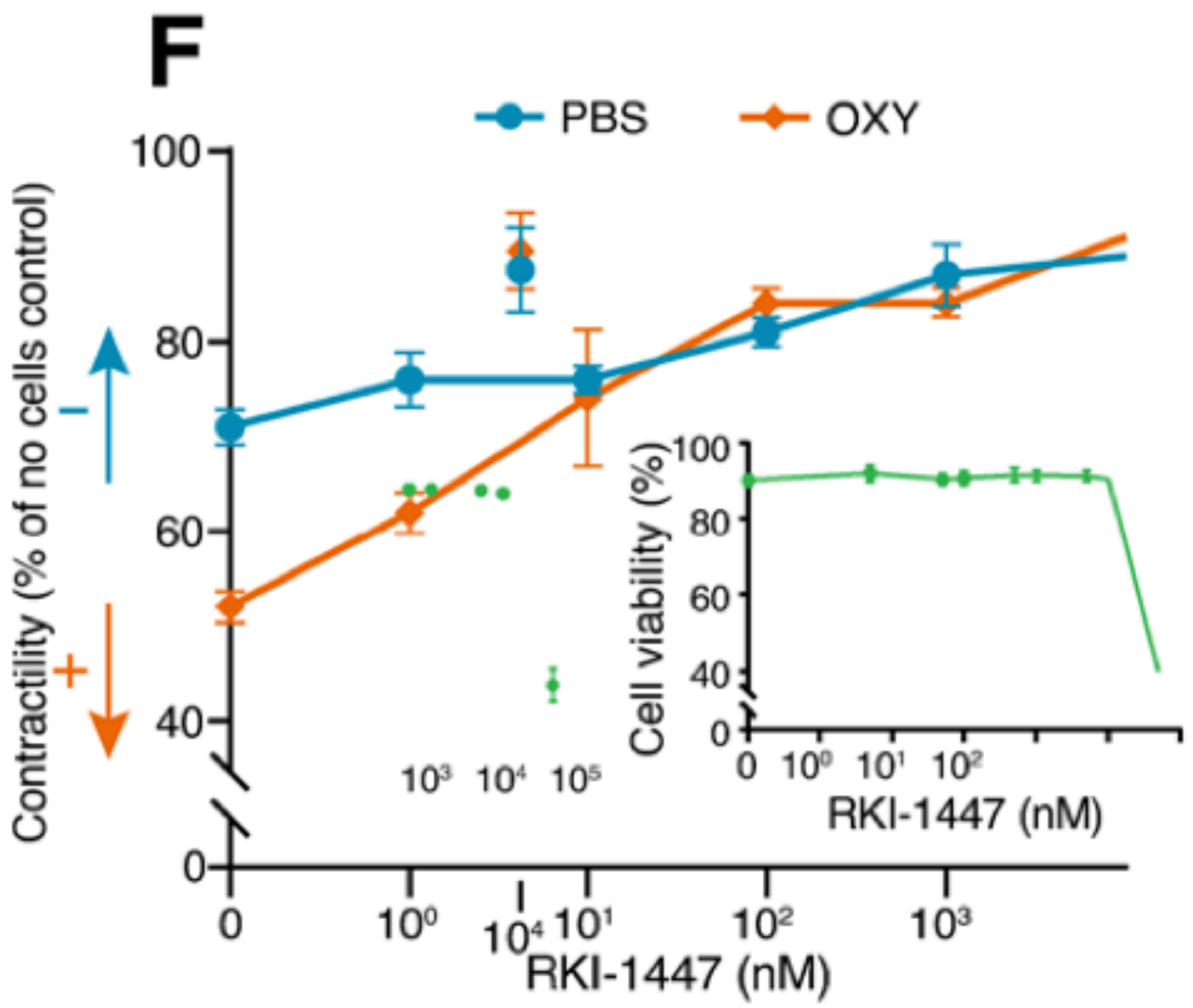
Drug name	Score	Percentile
Aspirin	2.529	0.61%
17-OHPC	2.332	3.00%
17-OHP	2.11	6.92%
Nifedipine	1.814	27.93%
Terbutaline	1.689	37.71%
Ibuprofen	1.45	69.28%
Naproxen	1.388	79.43%
Atosiban	1.129	99.58%

D

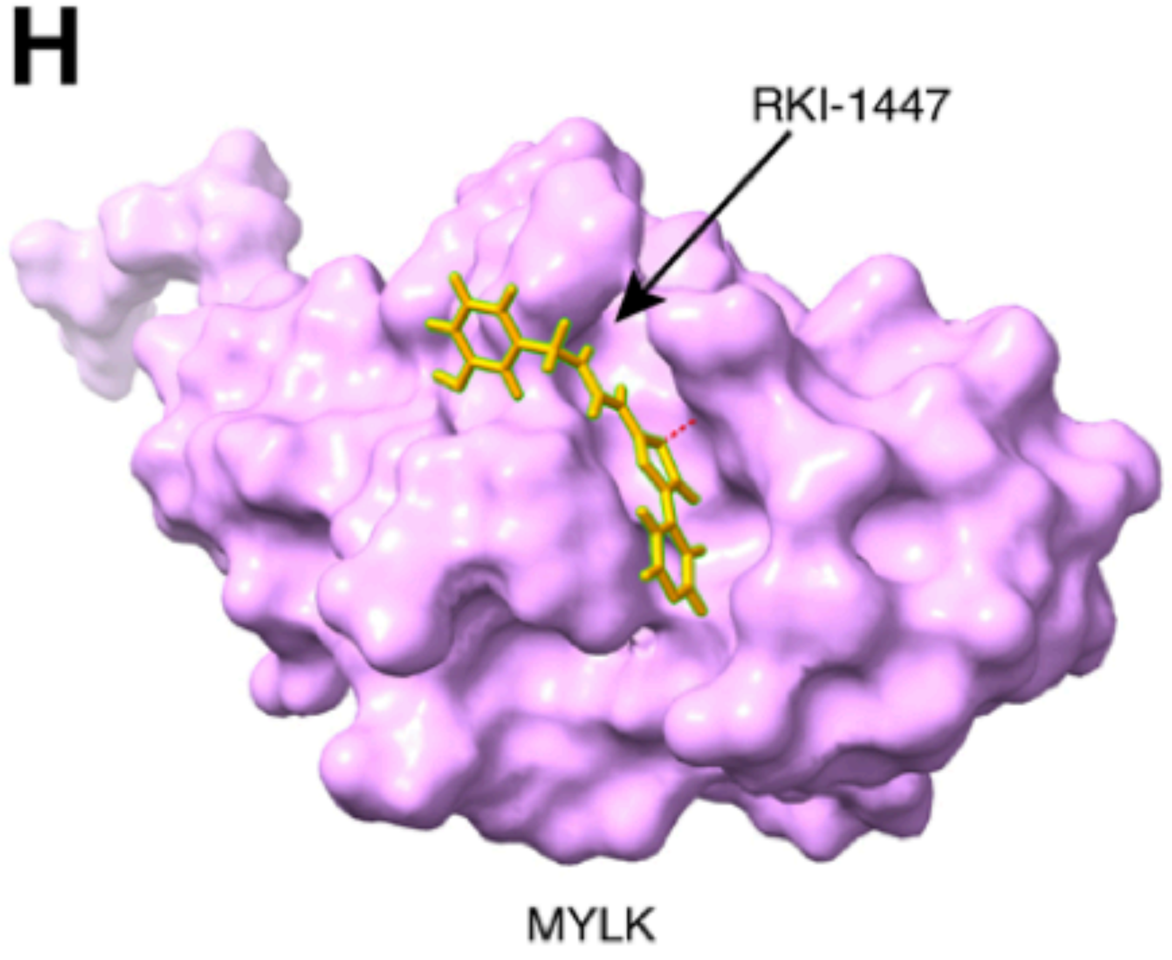
Drug name	Function
RKI-1447	ROCK kinase inhibitor
Ephedrine-HCL	Adrenoceptor agonist
Bosutinib	Bcr-Abl kinase inhibitor
Bisindolylmaleimide IX	Pan-PKC inhibitor
U-0126	MEK inhibitor
4,5,6,7-TBBt	CK2 inhibitor
PD-98059	MEK inhibitor
LY294002	PI3K inhibitor
URMC-099	MLK inhibitor
SB-203580	p38 MAPK inhibitor

Small molecule candidates reduced contracts in both physiological states (PBS/OXY)





RKI-1447 dose response curve (reduction of contractions is increase Y)



Fits in the pocket, so nice!

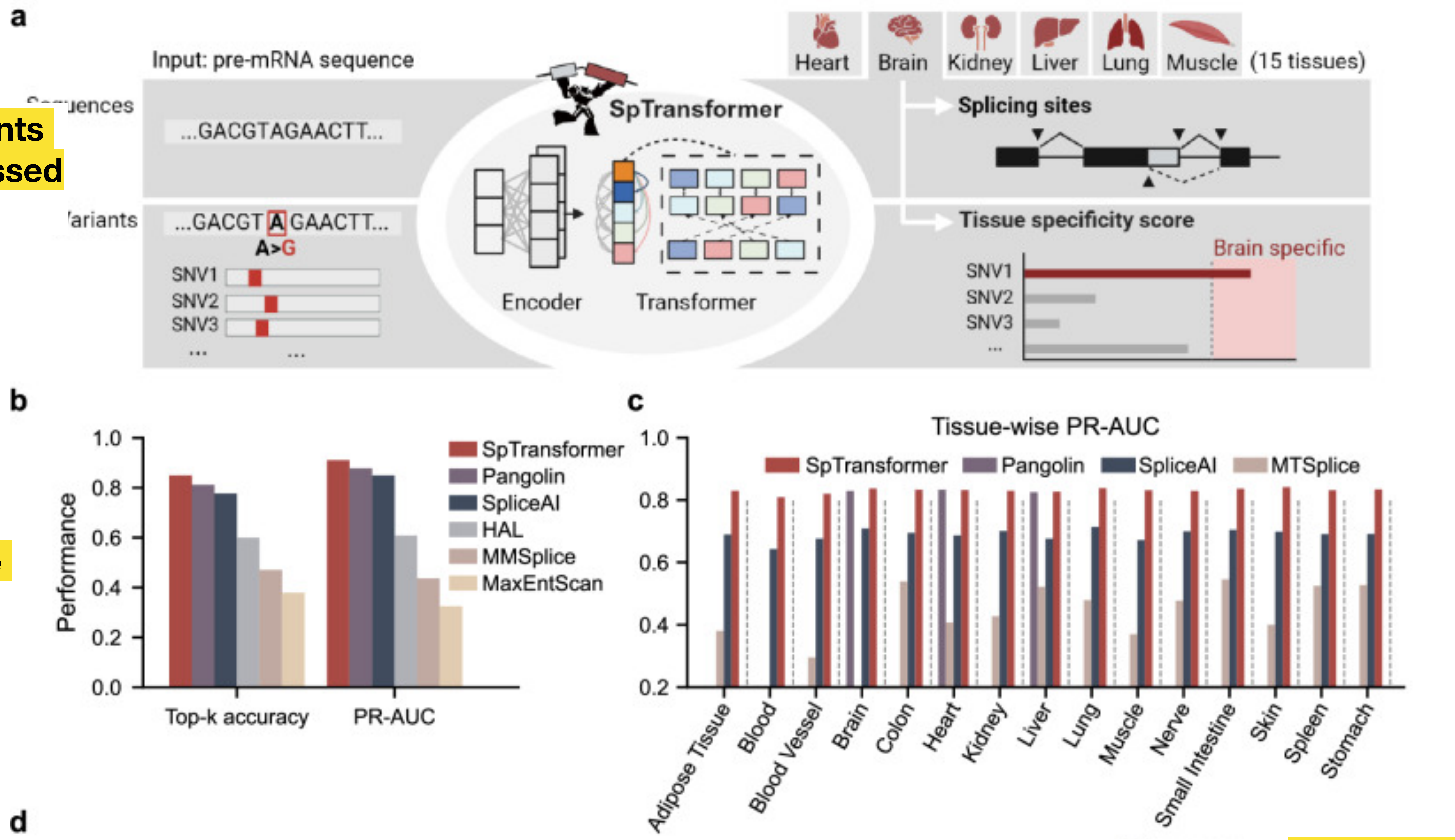
“Cruel Summer”

Taylor-ed for You—Precision Medicine in Action

SpliceTransformer predicts tissue-specific splicing linked to human diseases (You et al, *Nature Communications*)

- Goal: To predict tissue-specific RNA splicing variation
- Method:
  - Use a Sinkhorn Transformer — allowing for increased sequence context by using a structured attention mechanism
  - Integrates both GTEx (for fine-grained tissue specificity) and cross-species RNASeq data (to improve generalization)
  - Can generate a ΔSplice score for each variant to quantify its effect on splicing
- Result:
  - Found that 60% of intronic and synonymous mutations have high ΔSplice scores
  - Outperforms existing methods on splice site prediction and even more so for tissue specificity
  - Found tissue specific splicing alterations are enriched for diseases of those tissues
- Conclusion: Adding sequence and tissue context greatly improves understanding of variation

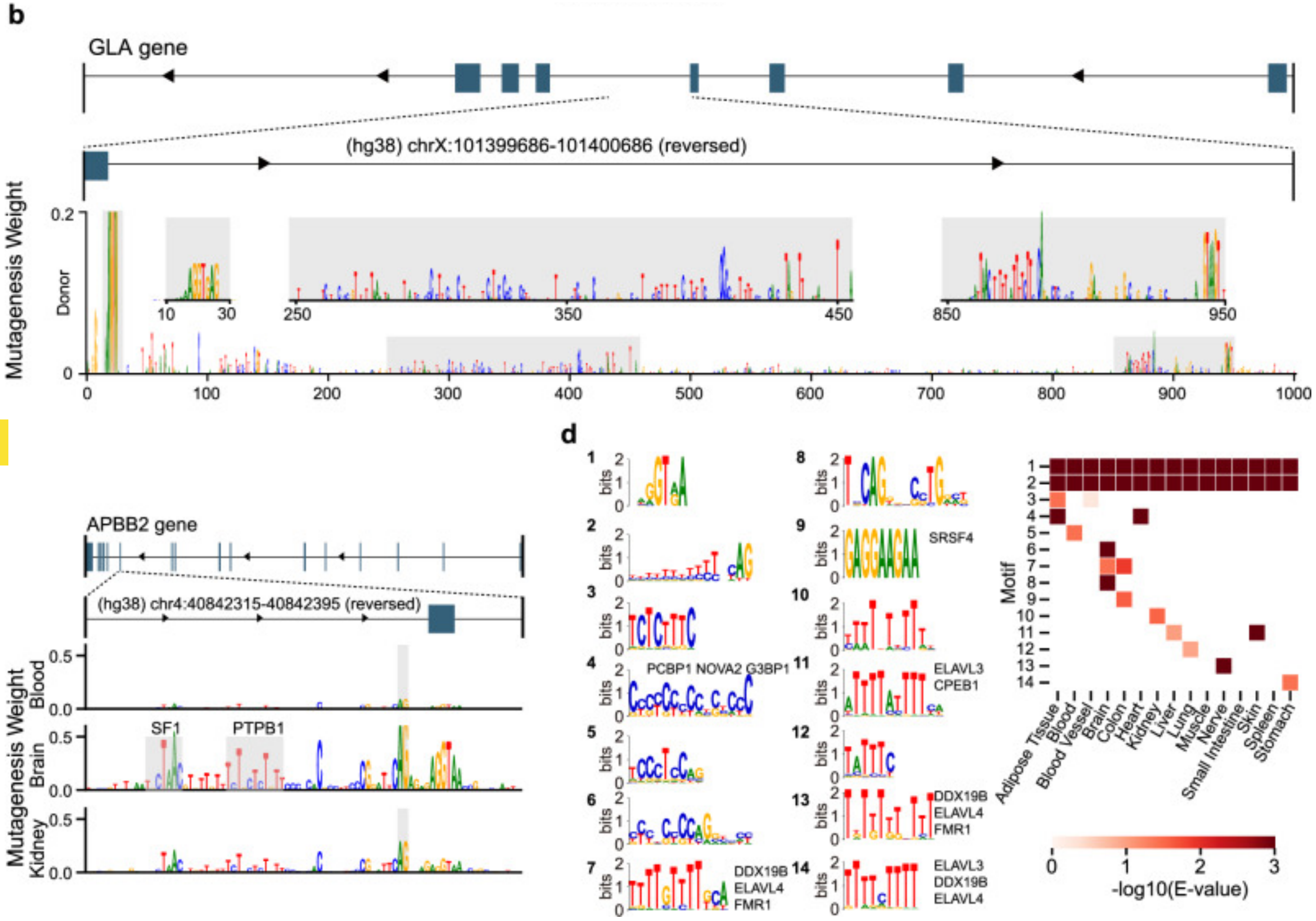
Sequence and variants are encoding and passed to transformer

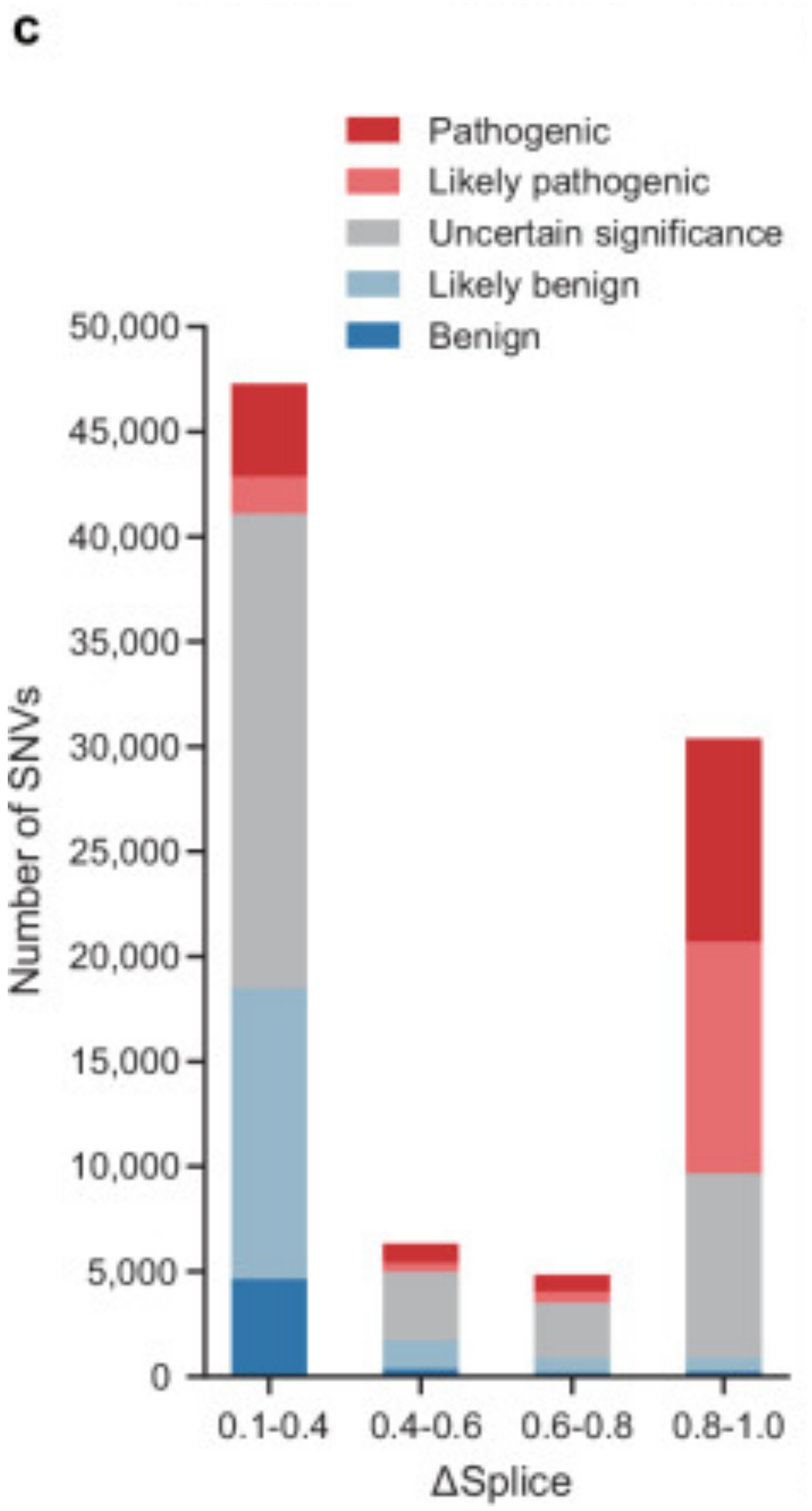


Better performance overall

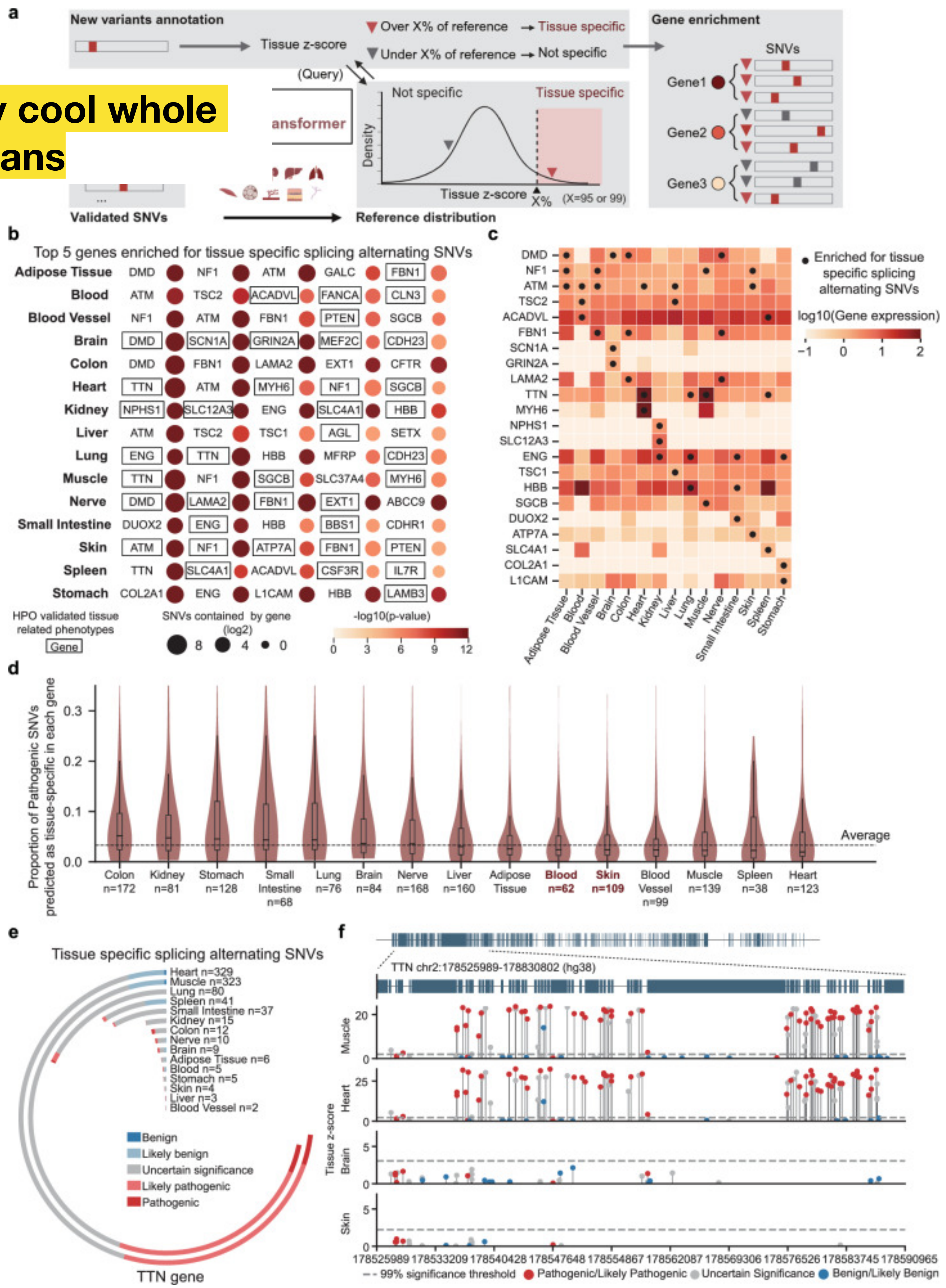
... and especially for tissue specificity

SpTransformer can score the regulatory regions for their affect on splicing





Can do some pretty cool whole genome scans



Large ΔSplice scores are more likely to be pathogenic

Integrating imaging and genomic data for the discovery of distinct glioblastoma subtypes: a joint learning approach (Guo et al, *Scientific Reports*)

- Goal: Improve power of glioblastoma studies by integrating both imaging and genomic data

- Method:

  - Joint learning model that integrates MRIs and mutations in 27 key genes

  - 571 glioblastoma patients

- Result:

  - Found three subtypes of glioblastoma with significant survival differences

  - Found image-genomic correlation between RB1 and PTEN pathways linked to diffusion metrics

- Conclusion: Multimodal data continues to empower scientific investigation

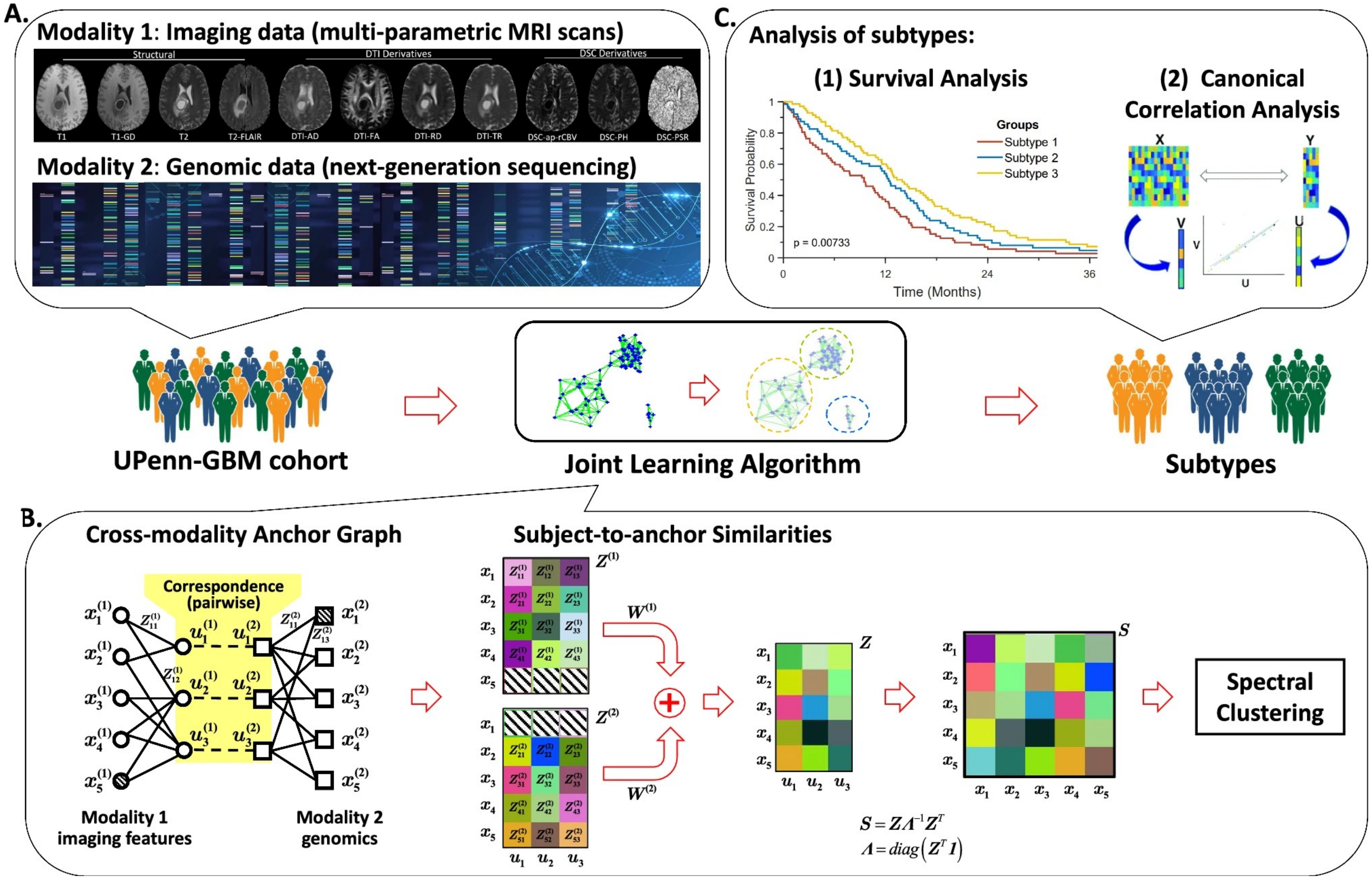
#IS25 #YIR25 @proftatonetti @tatonetti.bsky.social https://doi.org/10.1038/s41598-024-55072-y

Imaging

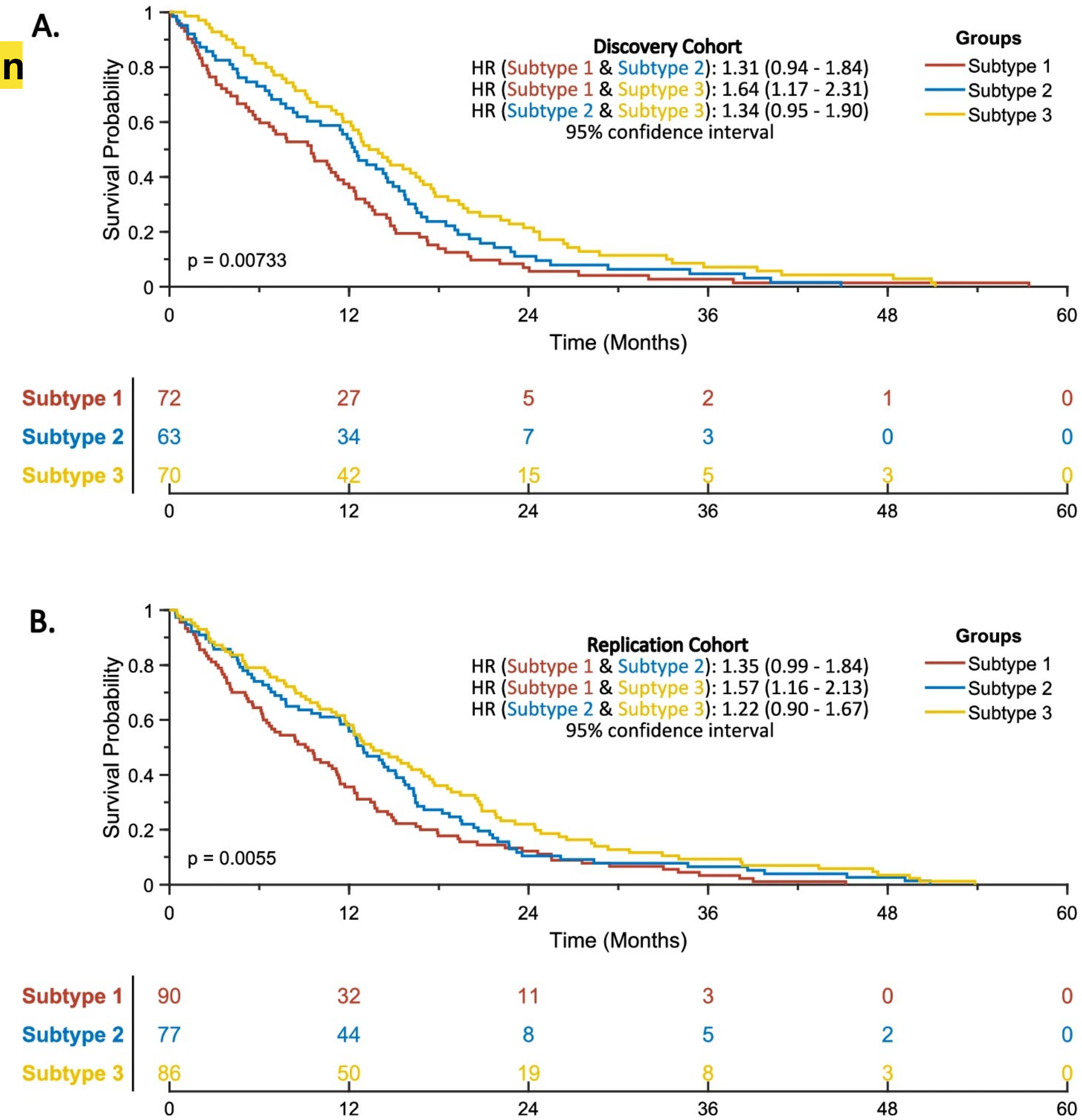
Genetics

Method is good for handling missing data

Uses Markov random walks to establish relationships



Found significant survival differences in the clusters



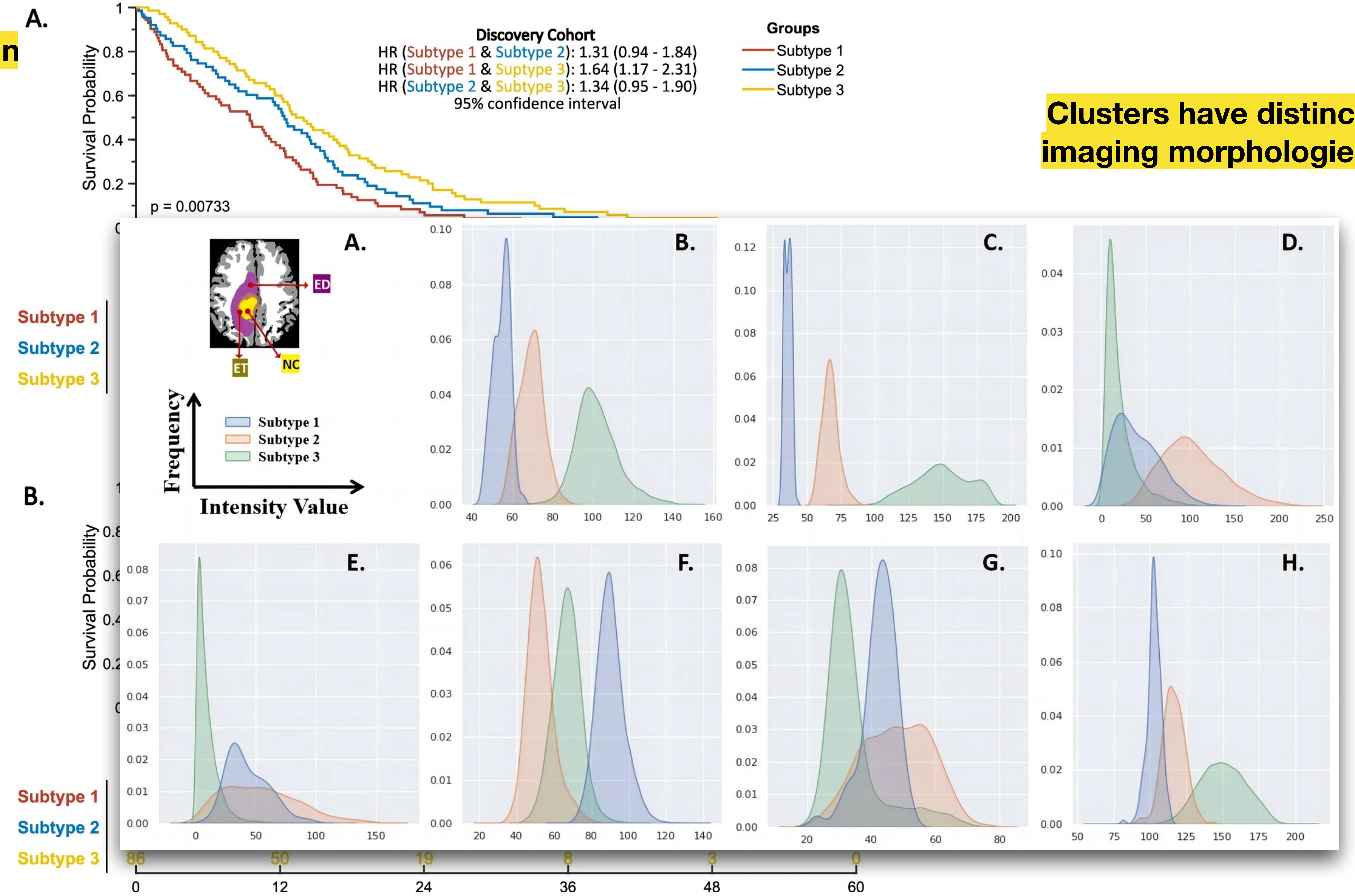
A.

	0	12	24	36	48	60
Subtype 1	72	27	5	2	1	0
Subtype 2	63	34	7	3	0	0
Subtype 3	70	42	15	5	3	0

B.

	0	12	24	36	48	60
Subtype 1	90	32	11	3	0	0
Subtype 2	77	44	8	5	2	0
Subtype 3	86	50	19	8	3	0

Found significant survival differences in the clusters



Clusters have distinct imaging morphologies

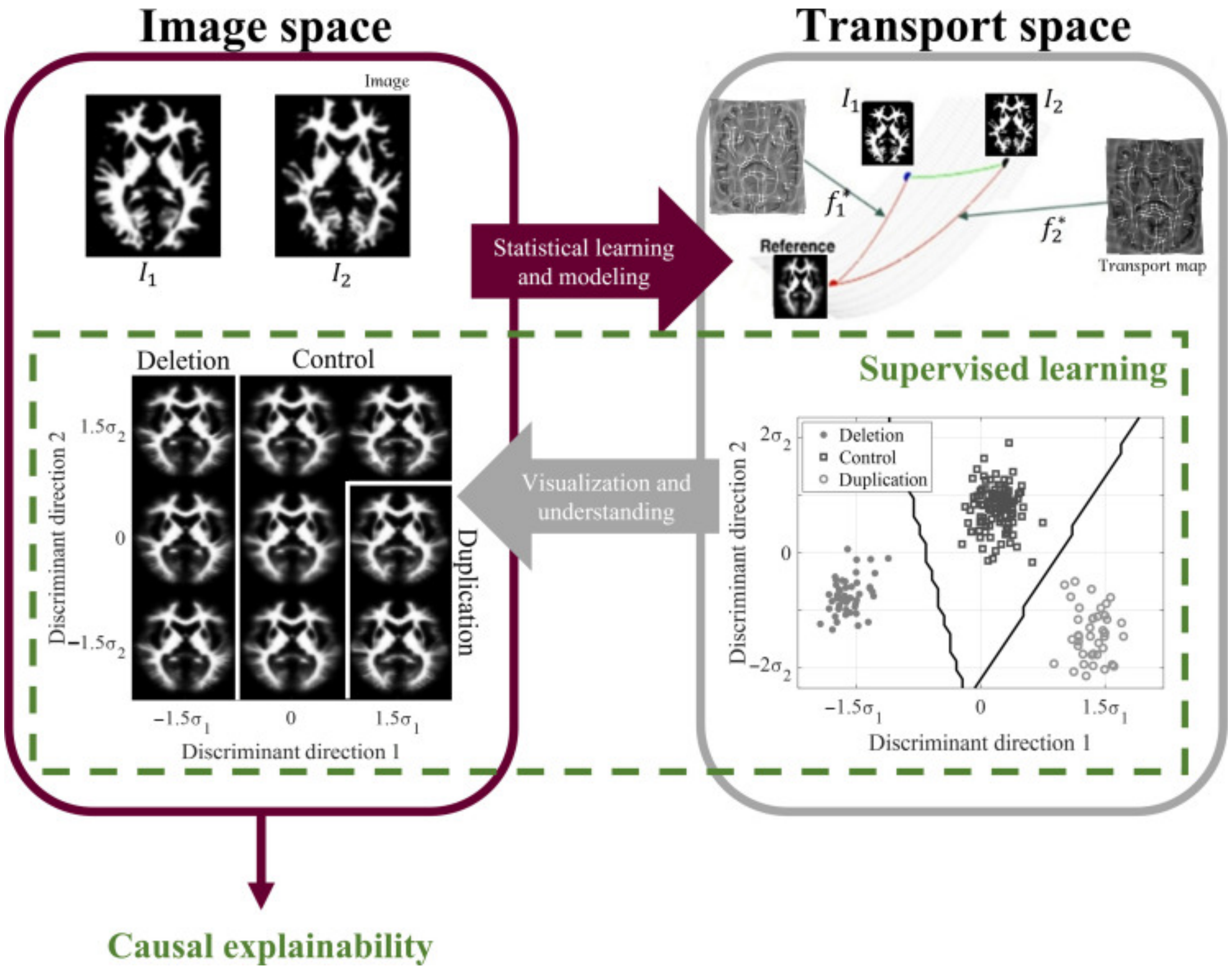
Can identify interaction between imaging and genetics

Subtype 1
Subtype 2
Subtype 3
RB1
CDKN2A
MDM4
TP53
BRAF
NF1
PTEN
PIK3R1
PIK3CA
FGFR2
MET
PDGFRA
EGFR
1
2
3
4
5
6
7
8
9
10
11
12
Imaging Feature
r
1.0
0.5
0.0
−0.5
−1.0
−log10(p)
0
1
2
3
4

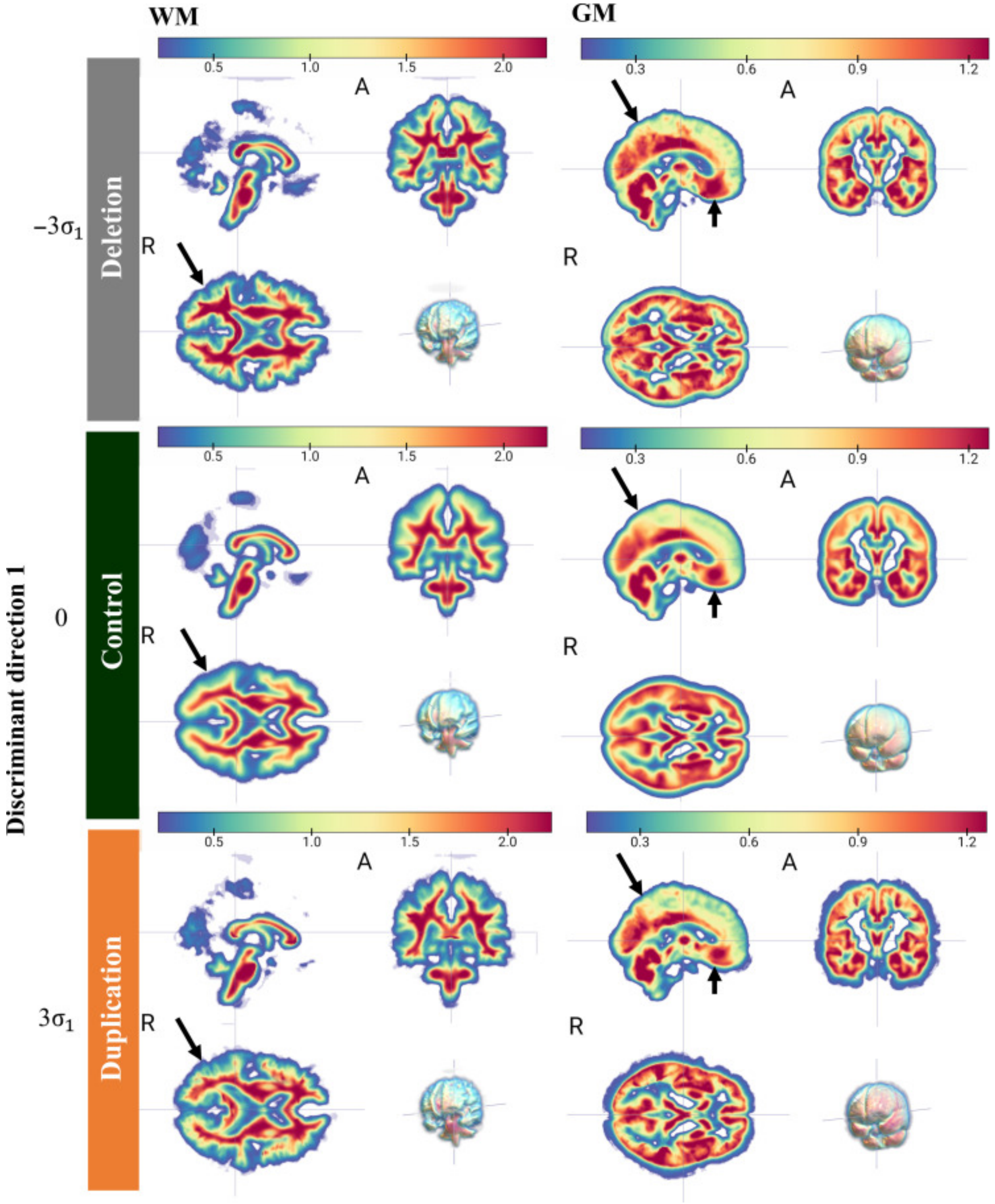
Imaging features are things like: Perfusion-related features, contrast enhancement patterns, diffusion measurements, etc.

Discovering the gene-brain-behavior link in autism via generative machine learning (Kundu et al, *Science Advances*)

- Goal: Identify brain structural changes linked to 16p11.2 CNVs (deletions & duplications) — most important genetic risk factor for Autism
- Method:
  - 206 individuals with MRIs, genetics, and behavioral assessments
  - Generative model (called 3D Transport-Based Morphometry) that quantifies the distance between patient and the reference (average of all)
  - Subsequent supervised learning using penalized latent discriminant analysis to predict variant status
- Result:
  - High accuracy at CNV prediction
  - Deletion carriers -> brain overgrowth
  - Duplication carriers -> brain undergrowth
  - Strong explanation of behavioral phenotypes
- Conclusion: Good demonstration that you can have too much information. By converting raw MRIs into these morphometry distances they were able to get better signal. There's a lesson here for all of us.

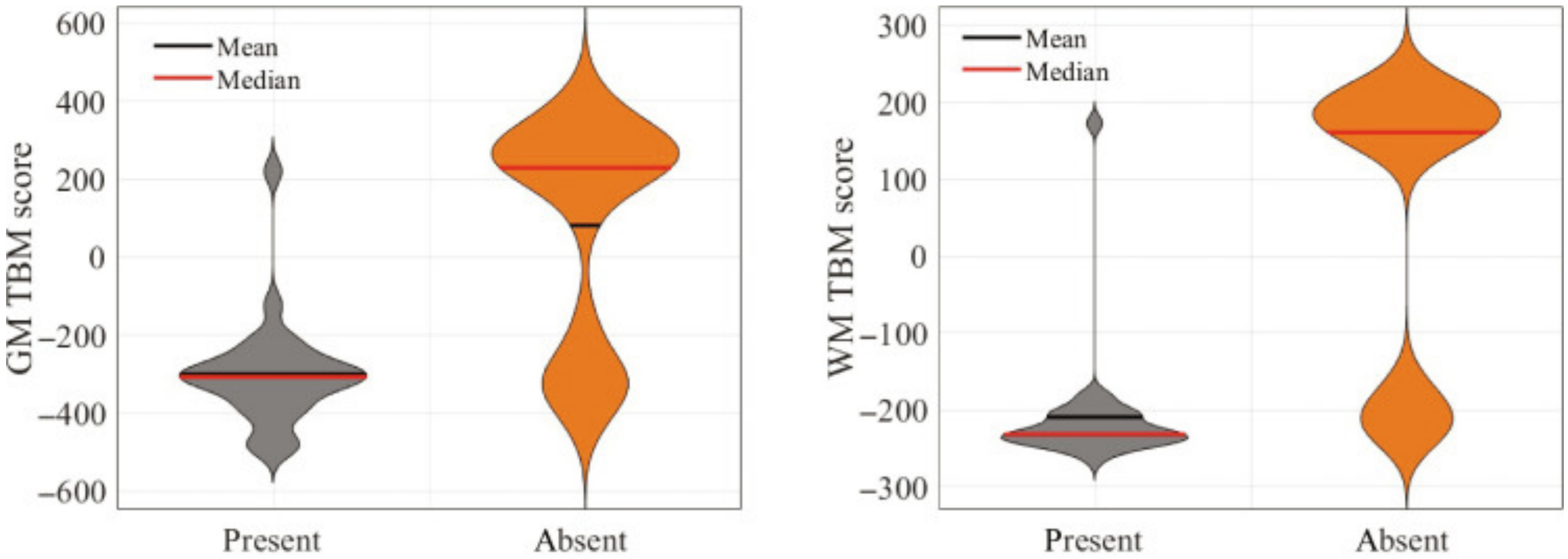


Converting to transport space highlights the differences



TBM shows what happens to the brain under different variant conditions

TBM scores are good discriminators of the presence of an articulation disorder



A gray matter

B white matter

Trainee Spotlight

VEDA PRIYA PULIGUNDLA

UNIVERSITY OF MINNESOTA

Deep learning predicts DNA methylation regulatory variants in specific brain cell types and enhances fine mapping for brain disorders (Zhou et al., 2025) https://www.science.org/doi/10.1126/sciadv.adn1870

Goal:

- Identify cell type-specific DNAm variants & Improve fine mapping of brain disorder risk loci

Method:

- Developed **INTERACT model** (CNN + Transformer) : A Deep Learning Model
- Trained on **single-nucleus DNAm data**
- **In silico mutagenesis** for variant effects
- **LD-score regression** for heritability analysis

Results:

- **High accuracy** (AUC = 0.99) in DNAm prediction
- Identified **regulatory variants for schizophrenia, depression, and Alzheimer's Disease (AD)**
- **Fine mapping** reveals cell type-specific causal loci
- **rs74504435 linked to EGFR in astrocytes** (potential AD target)

Conclusion:

- Deep learning improves genetic risk mapping

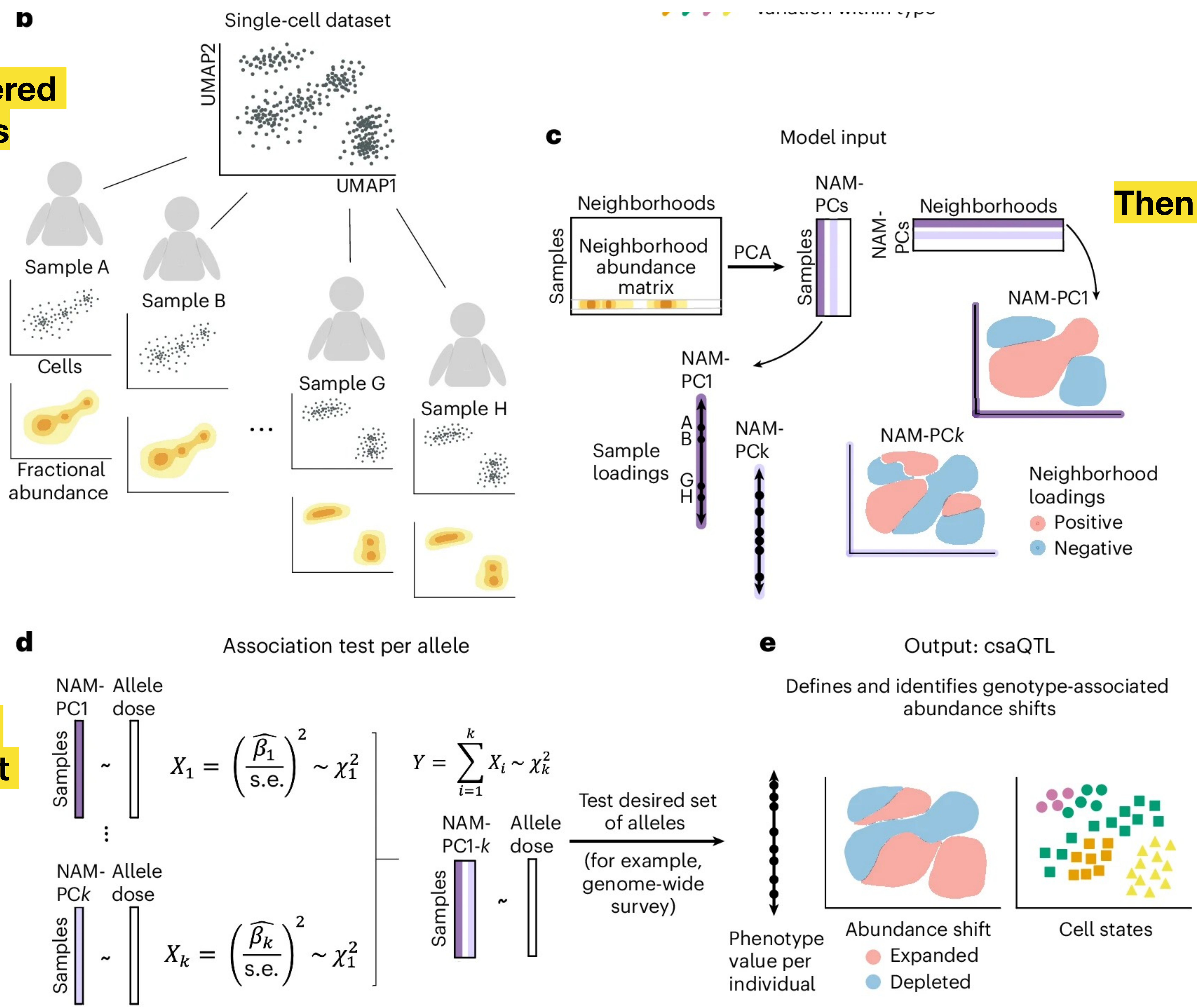
ENG 7:03 PM 3/10/2025

https://doi.org/10.1038/s41588-024-01909-1

Identifying genetic variants that influence the abundance of cell states in single-cell data (Rumker et al, *Nature Genetics*)

- Goal: Identify genetic variants that control *cell state* abundance ("csaQTLs")
- Method:
  - Introduce GeNA a statistical method to identify csaQTLs
    - Does not require states to be pre-defined
    - Uses Neighborhood Abundance Matrix (NAM) and PCA
  - Apply to dataset of ~1k individuals (800k peripheral blood mononuclear cells)
- Result:
  - Five significant csaQTLs identified, primarily affecting natural killer (NK) and myeloid cell states
  - rs3003-T was linked to increased NK cells expressing tumor necrosis factor
- Conclusion: I'll admit — not a "QTL" I had thought of before

First cells are clustered in neighborhoods

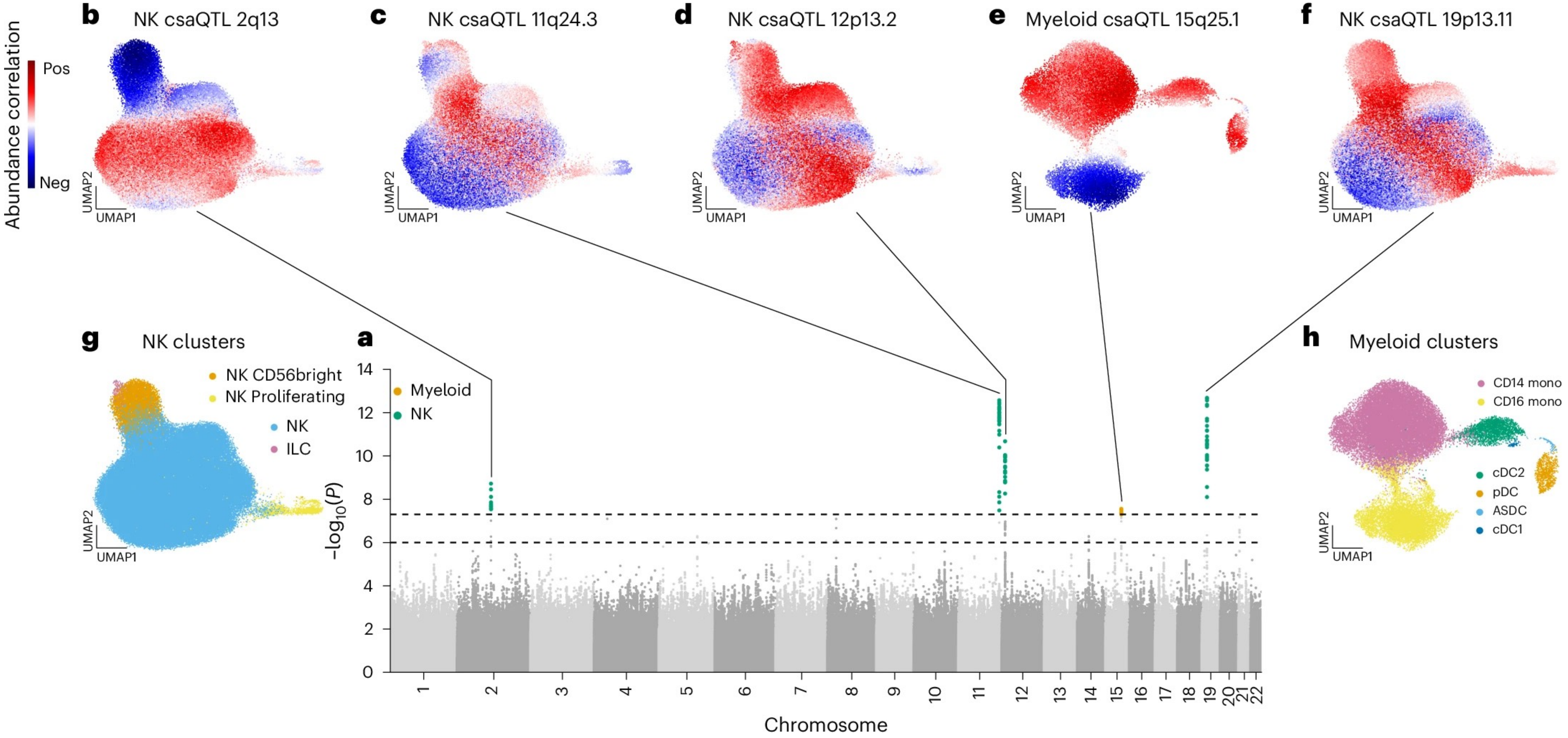

b
Single-cell dataset
UMAP2
UMAP1
Sample A
Sample B
Sample G
Sample H
Cells
Fractional abundance
...
c
Model input
Neighborhoods
Samples
Neighborhood abundance matrix
PCA
NAM-PCs
Samples
Neighborhoods
NAM-PCs
NAM-PC1
NAM-PCk
Sample loadings
A
B
G
H
NAM-PC1
NAM-PCk
Neighborhood loadings
Positive
Negative
d
Association test per allele
NAM-PC1
Allele dose
Samples
~
$X_1 = \left(\frac{\widehat{\beta_1}}{\text{s.e.}}\right)^2 \sim \chi_1^2$
NAM-PCk
Allele dose
Samples
~
$X_k = \left(\frac{\widehat{\beta_k}}{\text{s.e.}}\right)^2 \sim \chi_1^2$
$Y = \sum_{i=1}^{k} X_i \sim \chi_k^2$
NAM-PC1-k
Allele dose
Samples
~
Test desired set of alleles
(for example, genome-wide survey)
e
Output: csaQTL
Defines and identifies genotype-associated abundance shifts
Phenotype value per individual
Abundance shift
Expanded
Depleted
Cell states

Then do PCA to reduce dimension

Use a chi-squared test to look for any deviations that correlate with allelic dose

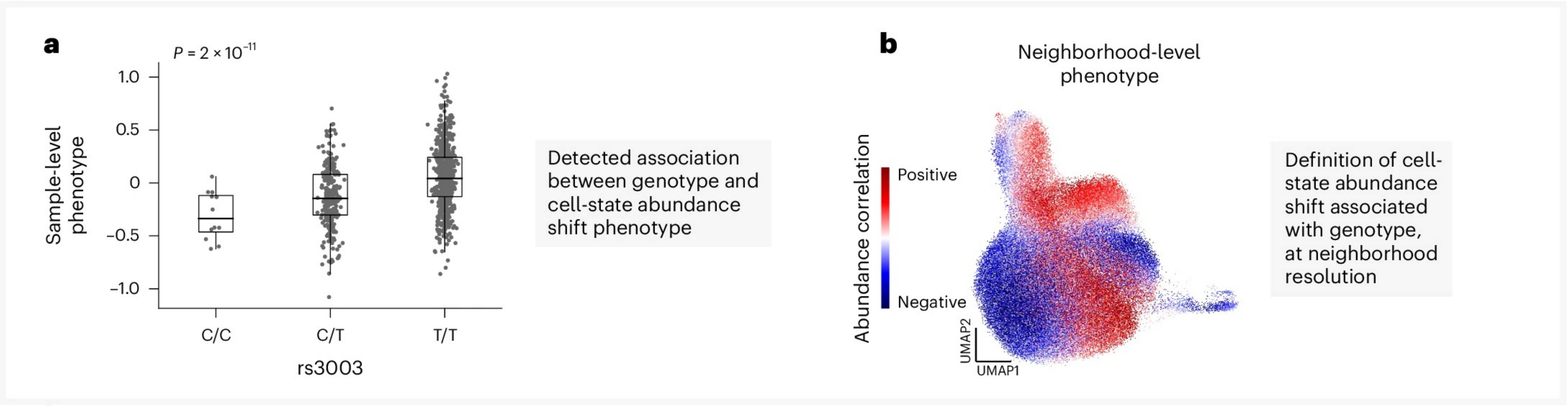
Discovered csaQTLs

Correlation between allelic dose and abundance of given cell state



Deeper dive on rs3003

GeNA output



Detected association between genotype and cell-state abundance shift phenotype

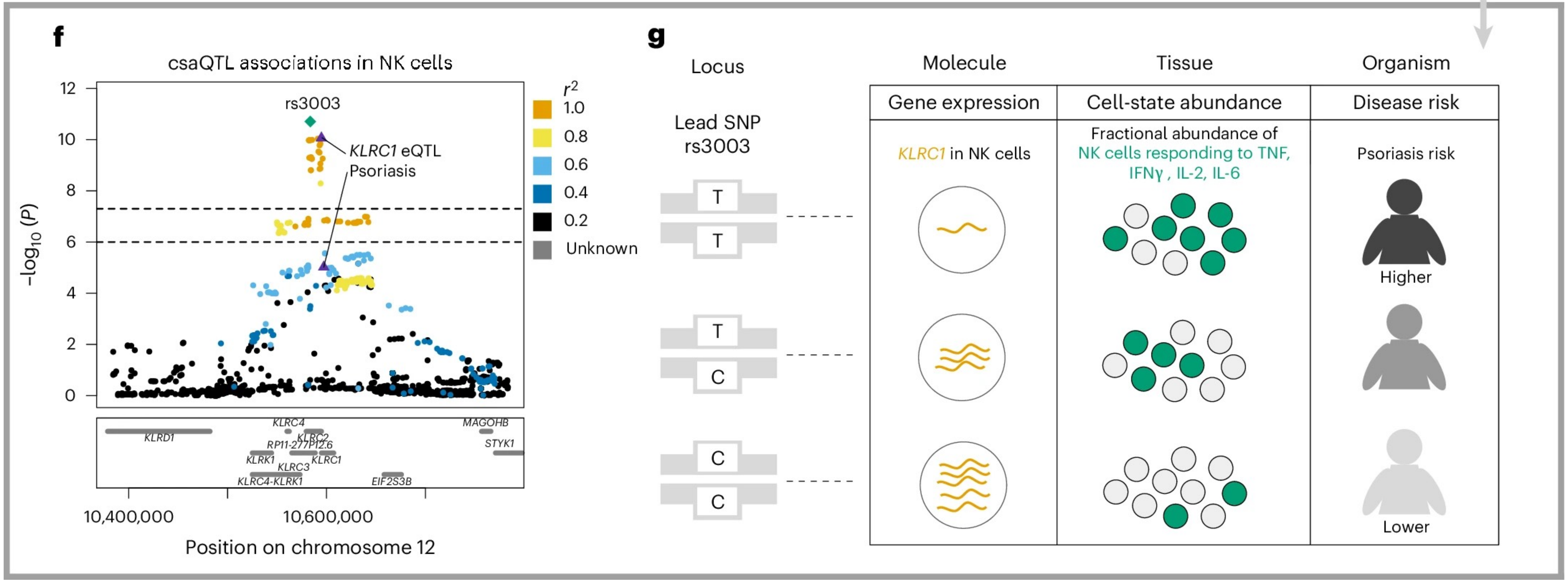
Definition of cell-state abundance shift associated with genotype, at neighborhood resolution

Does/Abundance relationships

Deeper dive on rs3003

Potential psoriasis mechanism

Colocalization of molecular, tissue and organism-level traits



Hit overlaps with known psoriasis hits

“Good Luck, Babe”

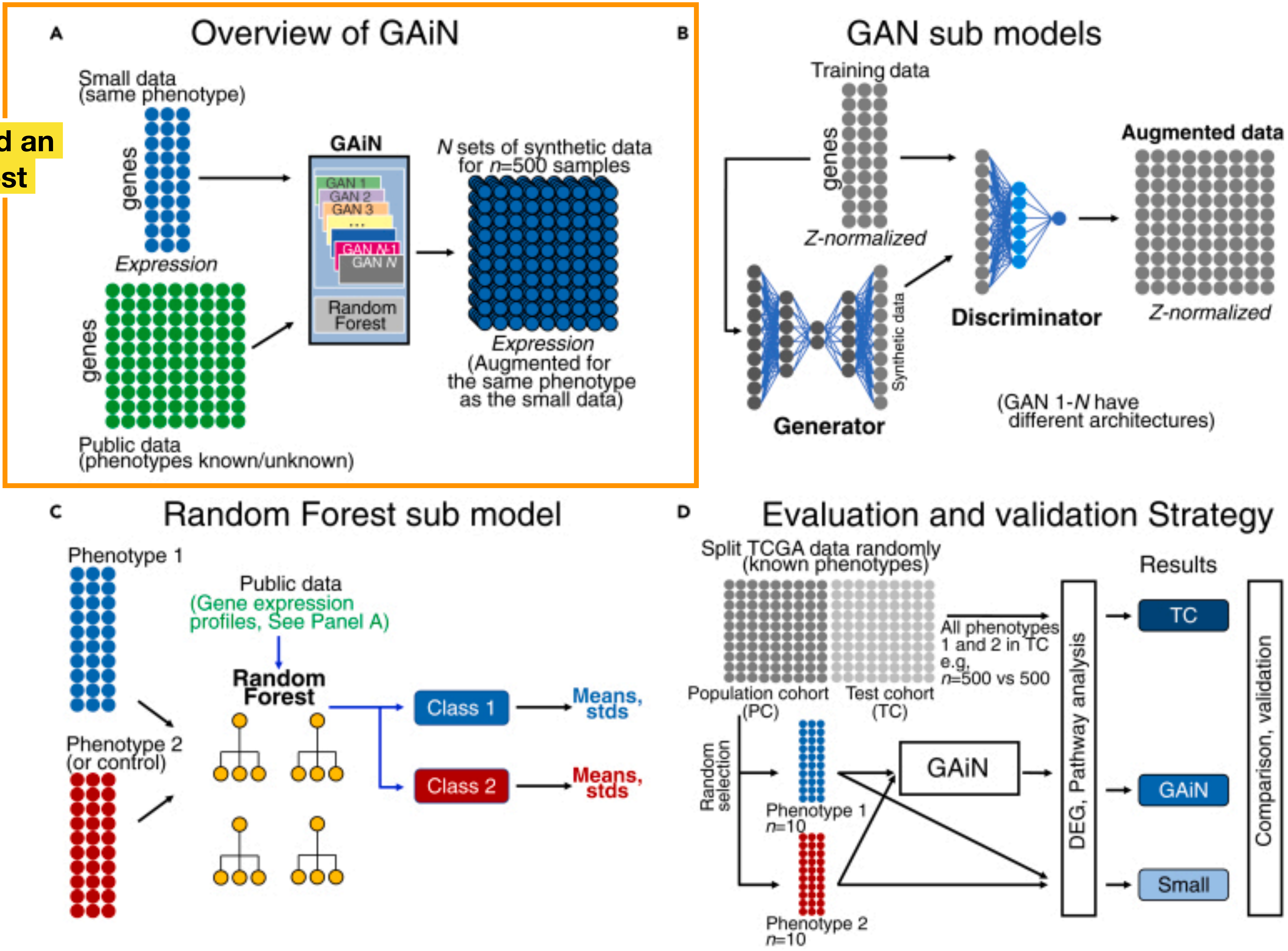
Bio Euphoria — Integrating Clinical & Molecular Data

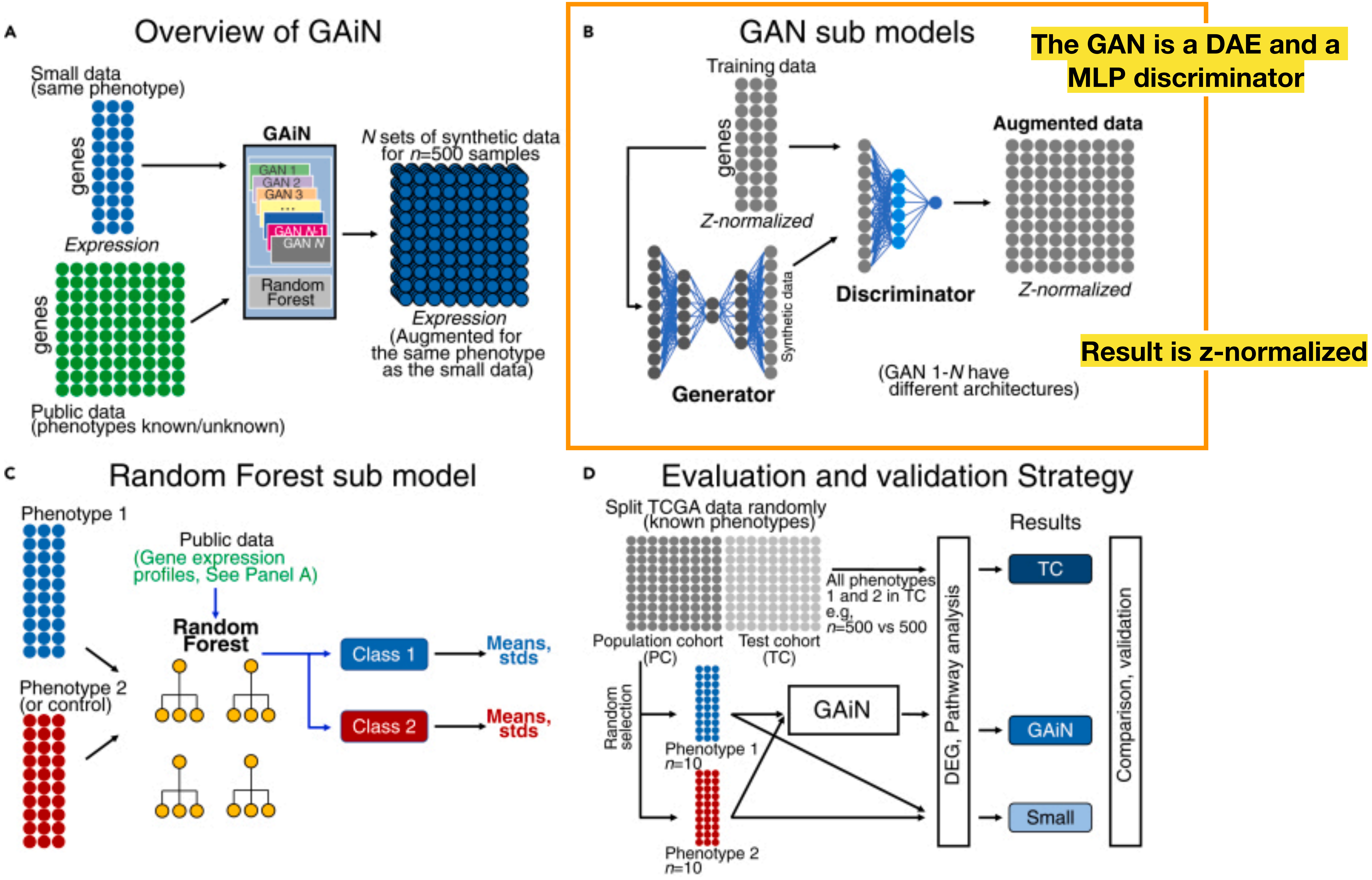
https://doi.org/10.1016/j.patter.2023.100910

GAiN: An integrative tool utilizing generative adversarial neural networks for augmented gene expression analysis (Waters et al, *Patterns*)

- Goal: Boost the power of expression experiments with small sample sizes
- Method:
  - Use a Generative Adversarial Network (GAN) comprised of a Denoising Autoencoder (DAE) and an MLP
  - H: Removing the noise in small sample experiments will boost their power
- Result: Able to recapitulate an experiment of N=533 with just 10 samples!
- Conclusion: This is either very clever or they are violating the laws of information theory

Uses a GAN and an Random Forest

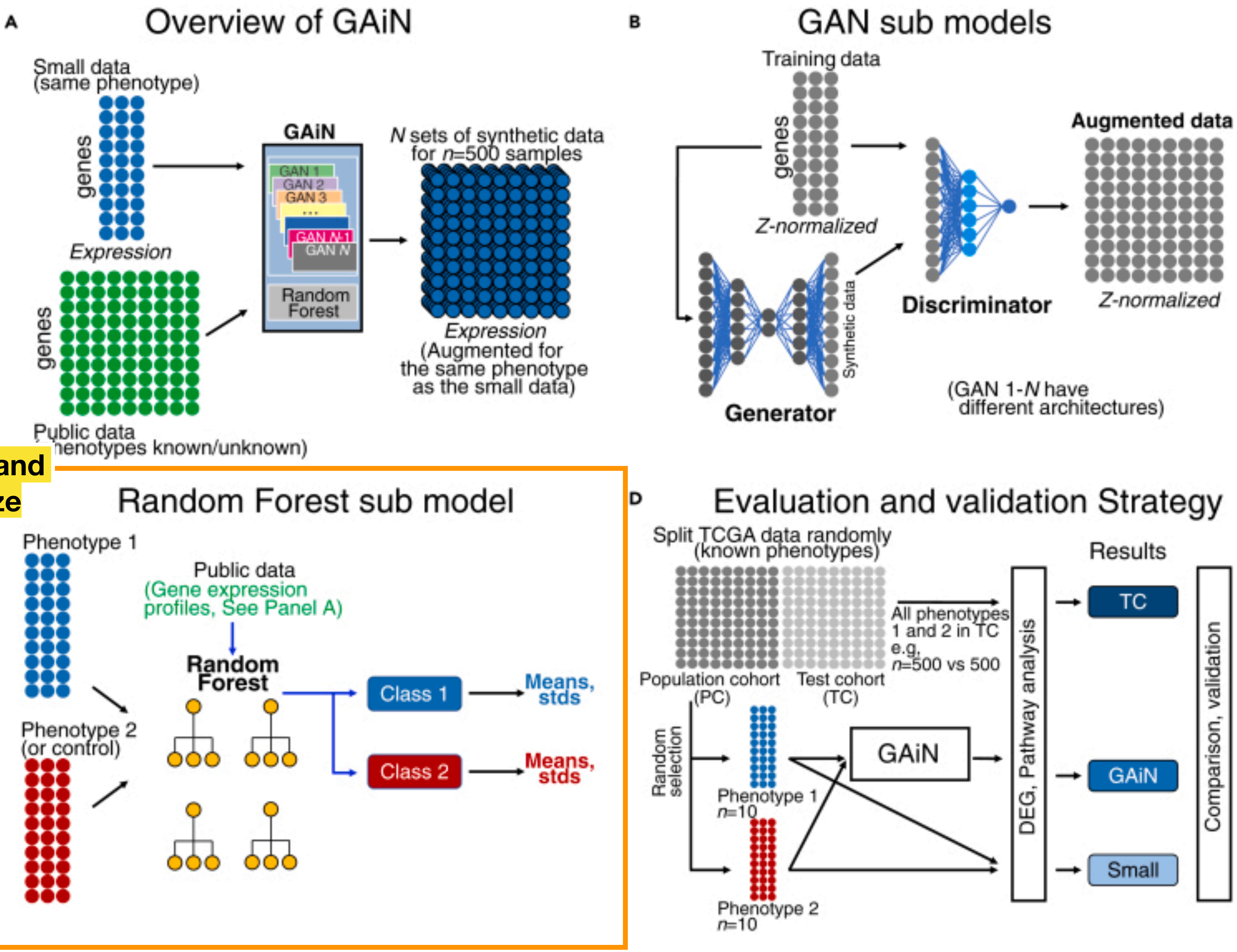




The GAN is a DAE and a MLP discriminator

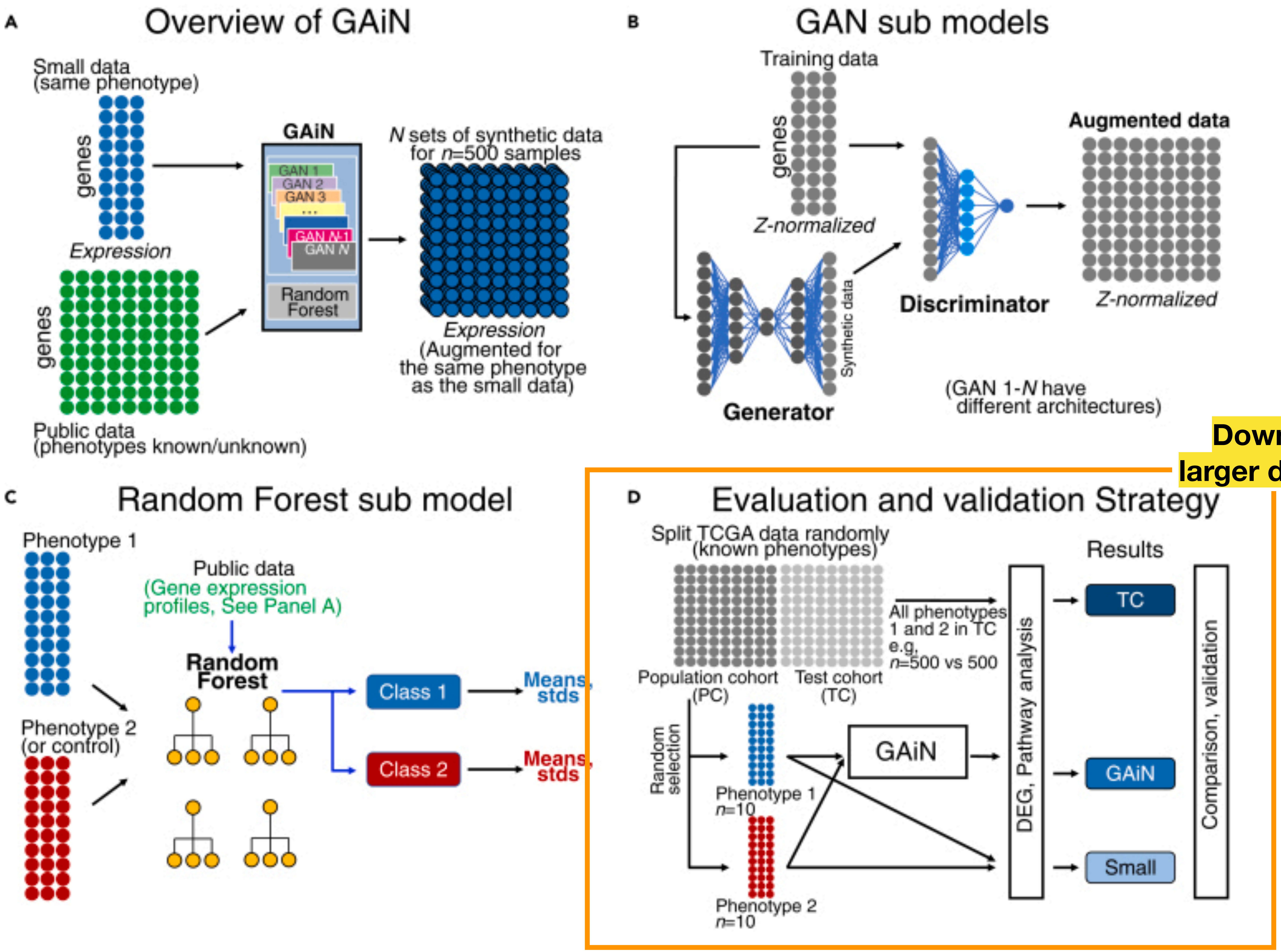
Result is z-normalized

https://doi.org/10.1016/j.patter.2023.100910



Use population data and RF to un-z-normalize

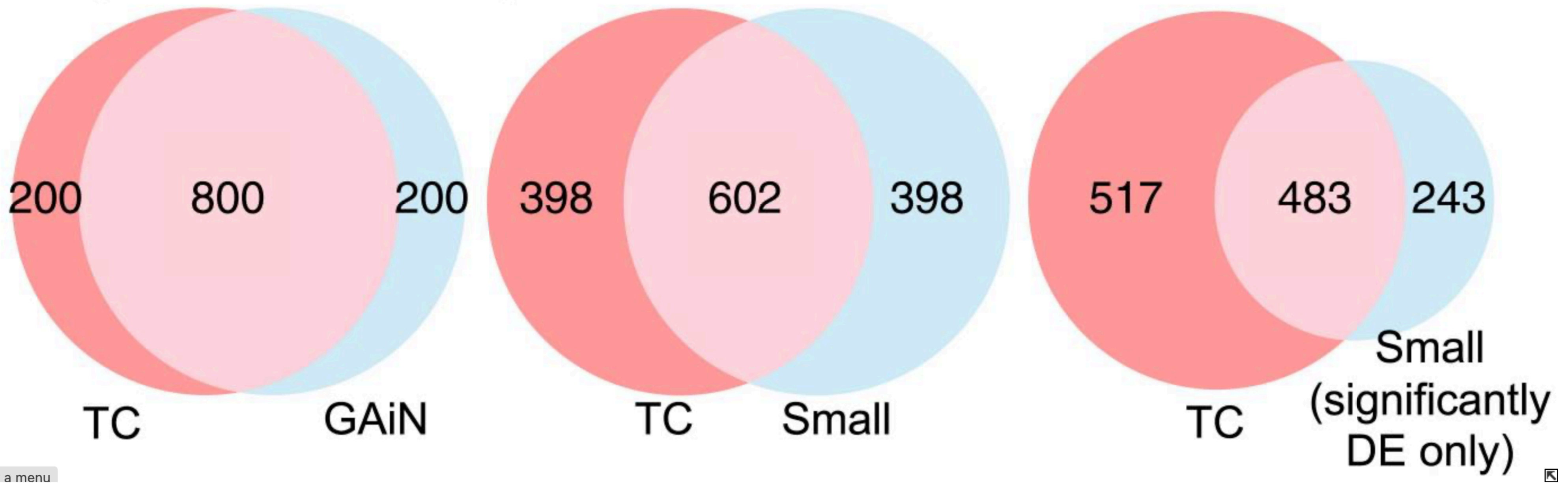
https://doi.org/10.1016/j.patter.2023.100910



Downsample from a larger dataset to validate

#IS25 #YIR25 @proftatonetti @tatonetti.bsky.social https://doi.org/10.1016/j.patter.2023.100910

DE genes from using TC, GAiN, and Small data



Better Overlap with real data than equal amount of small data

Better prediction of real data than small data even if duplicated

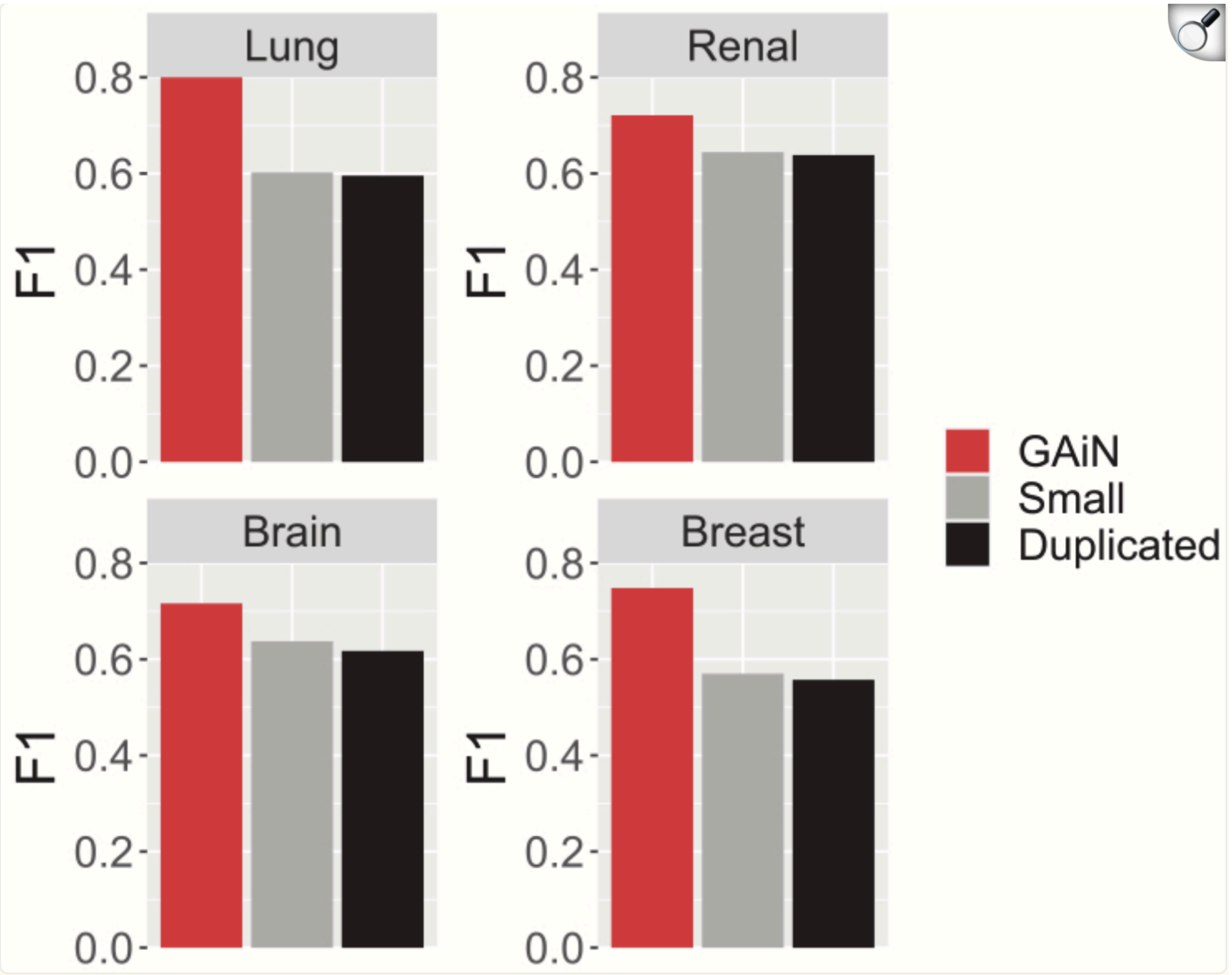
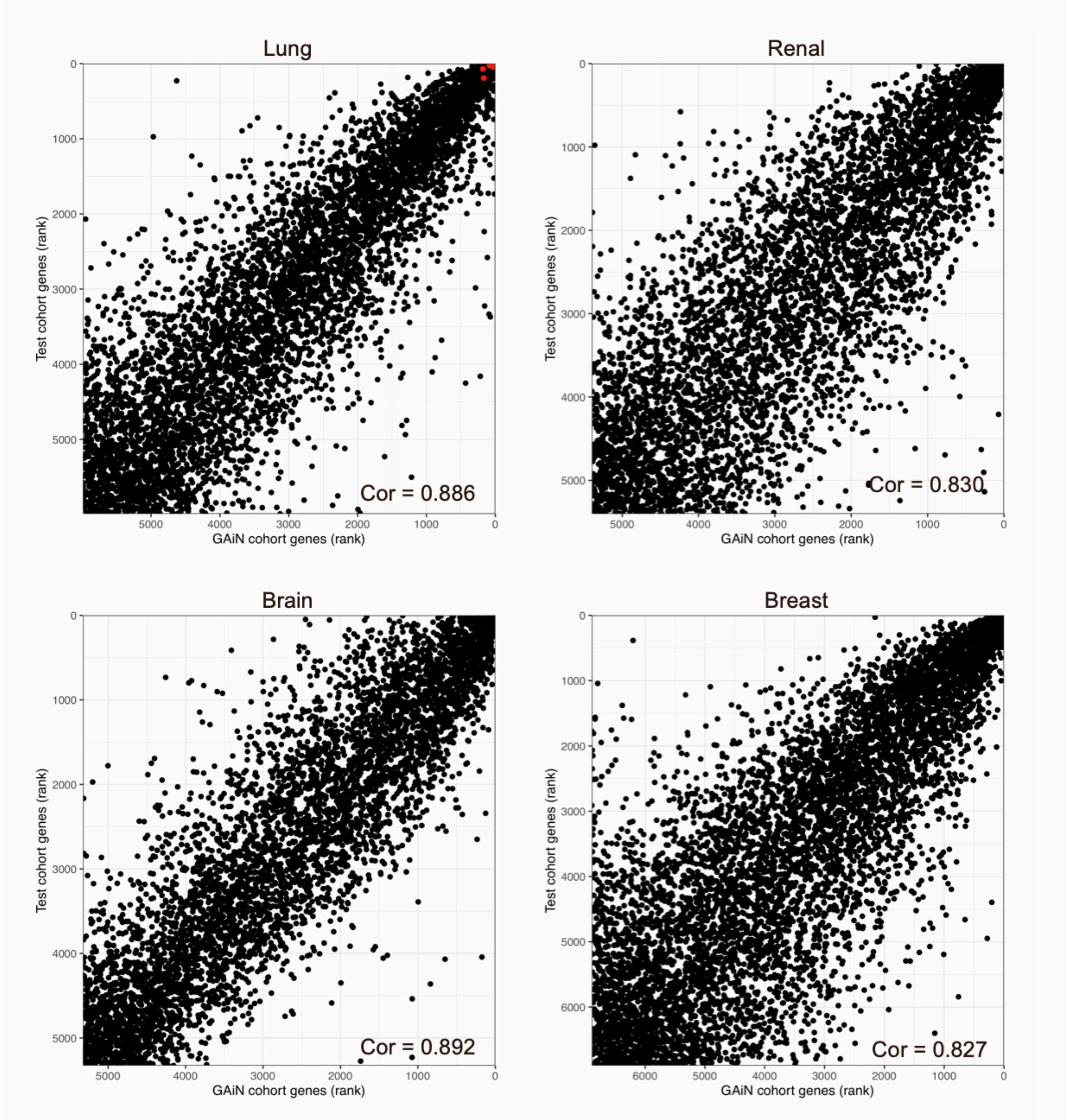
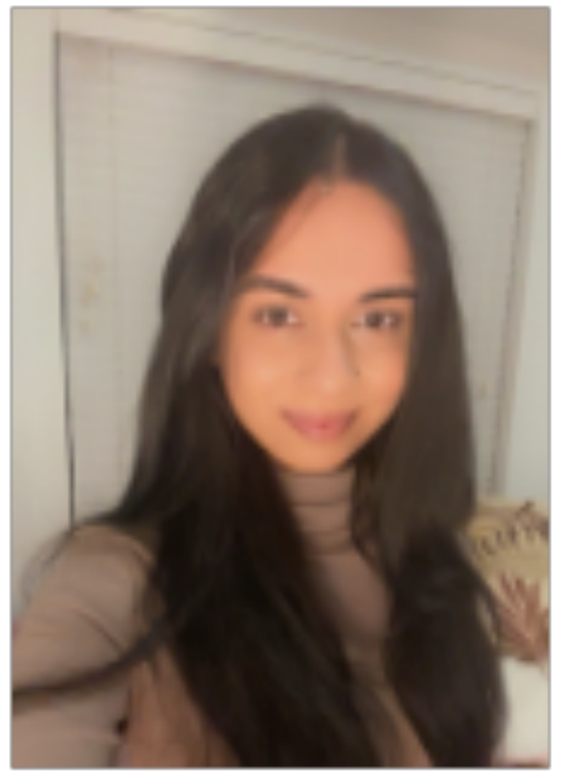


Figure S6. Rank comparison of DE genes for GAiN in phenotypic comparisons, related to Figure 3.



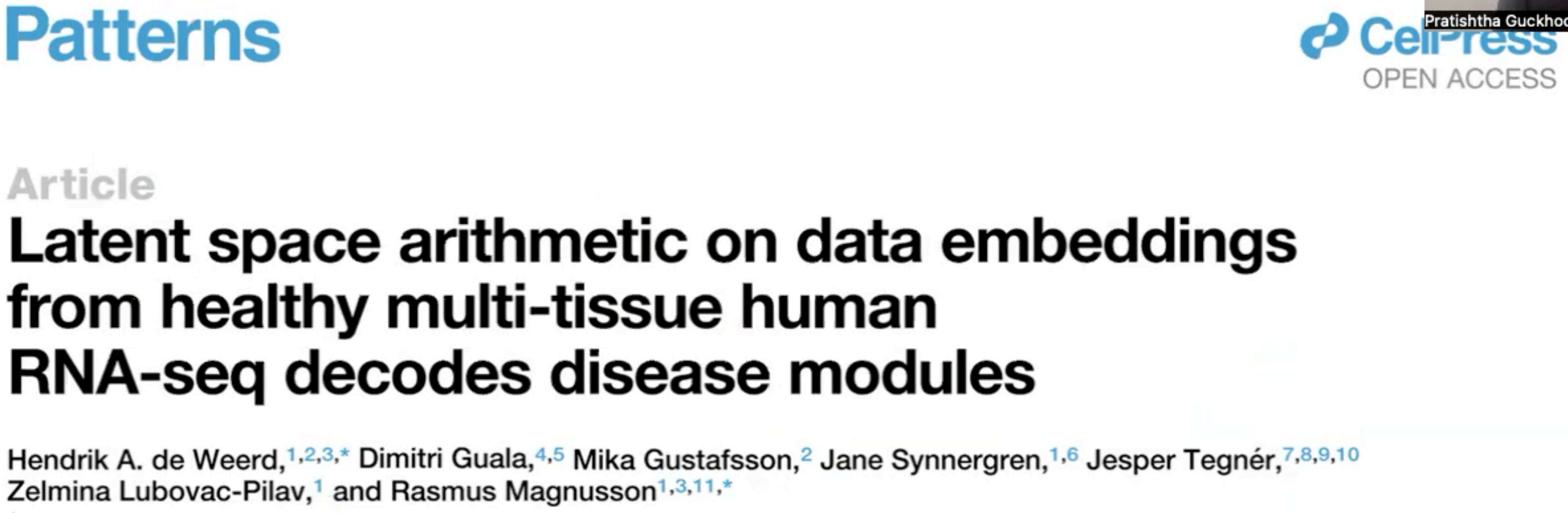
Ranks are pretty darn close to real data ranks

Trainee Spotlight



PRATISTHA GUCKHOOL

DREXEL UNIVERSITY



Patterns

CellPress OPEN ACCESS

Article

Latent space arithmetic on data embeddings from healthy multi-tissue human RNA-seq decodes disease modules

Hendrik A. de Weerd,[1,2,3,*] Dimitri Guala,[4,5] Mika Gustafsson,[2] Jane Synnergren,[1,6] Jesper Tegnér,[7,8,9,10] Zelmina Lubovac-Pilav,[1] and Rasmus Magnusson[1,3,11,*]

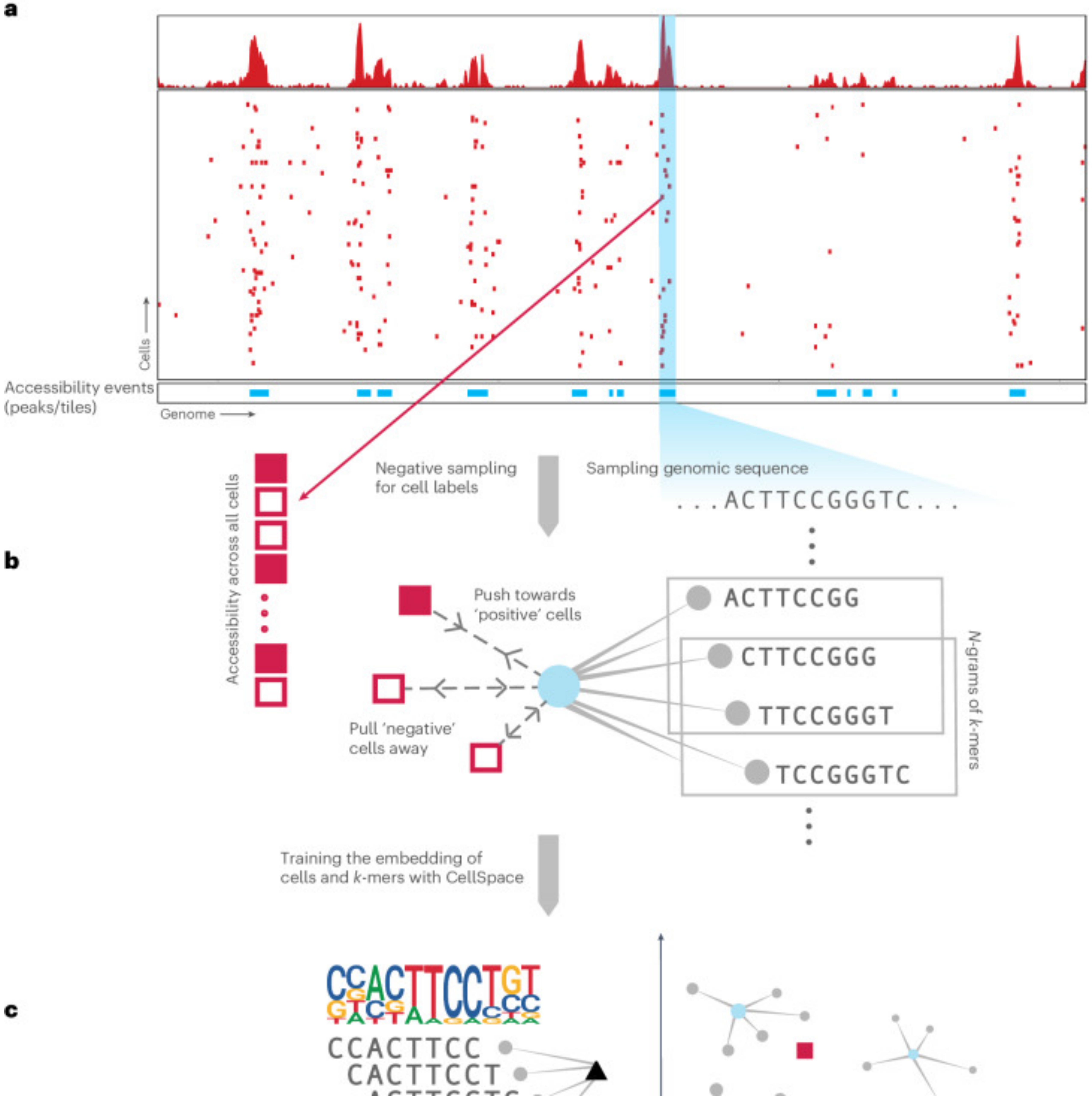
The Big Research Question:

How can deep learning models help identify disease-relevant genes from healthy transcriptomics data?

Scalable and unbiased sequence-informed embedding of single-cell ATAC-seq data with CellSpace (Tayyebi et al, *Nature Methods*)

- Goal: Improve ATAC-seq analysis by integrating a genome sequence context
- Method:
  - Train a new embedding model that learns from the k-mers of “open” events (i.e. regions of the genome with enough reads)
  - Jointly learn with the cells to produce cell-sequence embeddings
- Result:
  - Embeddings naturally control for batch/donor effects
  - Can predict motif-activity on cell-by-cell basis
  - Creates a denser cell-by-event matrix for use with other analytical methods
- Conclusion: Sequence context is information that can be captured and integrated into our datasets

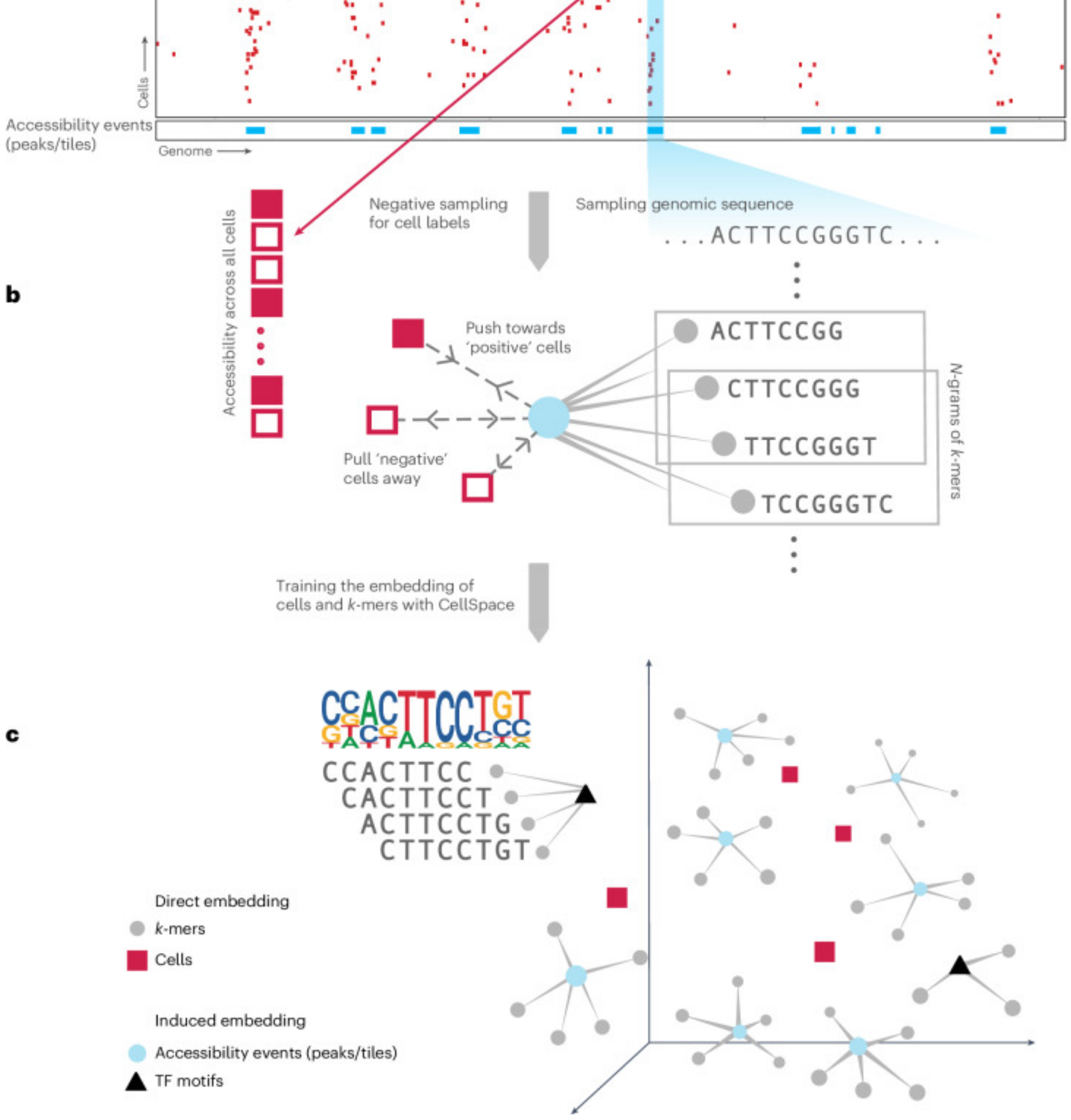
#IS25 #YIR25 @proftatonetti @tatonetti.bsky.social https://doi.org/10.1038/s41592-024-02274-x



a
Cells
Accessibility events (peaks/tiles)
Genome
b
Accessibility across all cells
Negative sampling for cell labels
Sampling genomic sequence
...ACTTCCGGGTC...
ACTTCCGG
CTTCCGGG
TTCCGGGT
TCCGGGTC
N-grams of k-mers
Push towards 'positive' cells
Pull 'negative' cells away
Training the embedding of cells and k-mers with CellSpace
c
CCACTTCCTGT
CCACTTCC
CACTTCCT

Each k-mer that is "open" is used to generate a training set

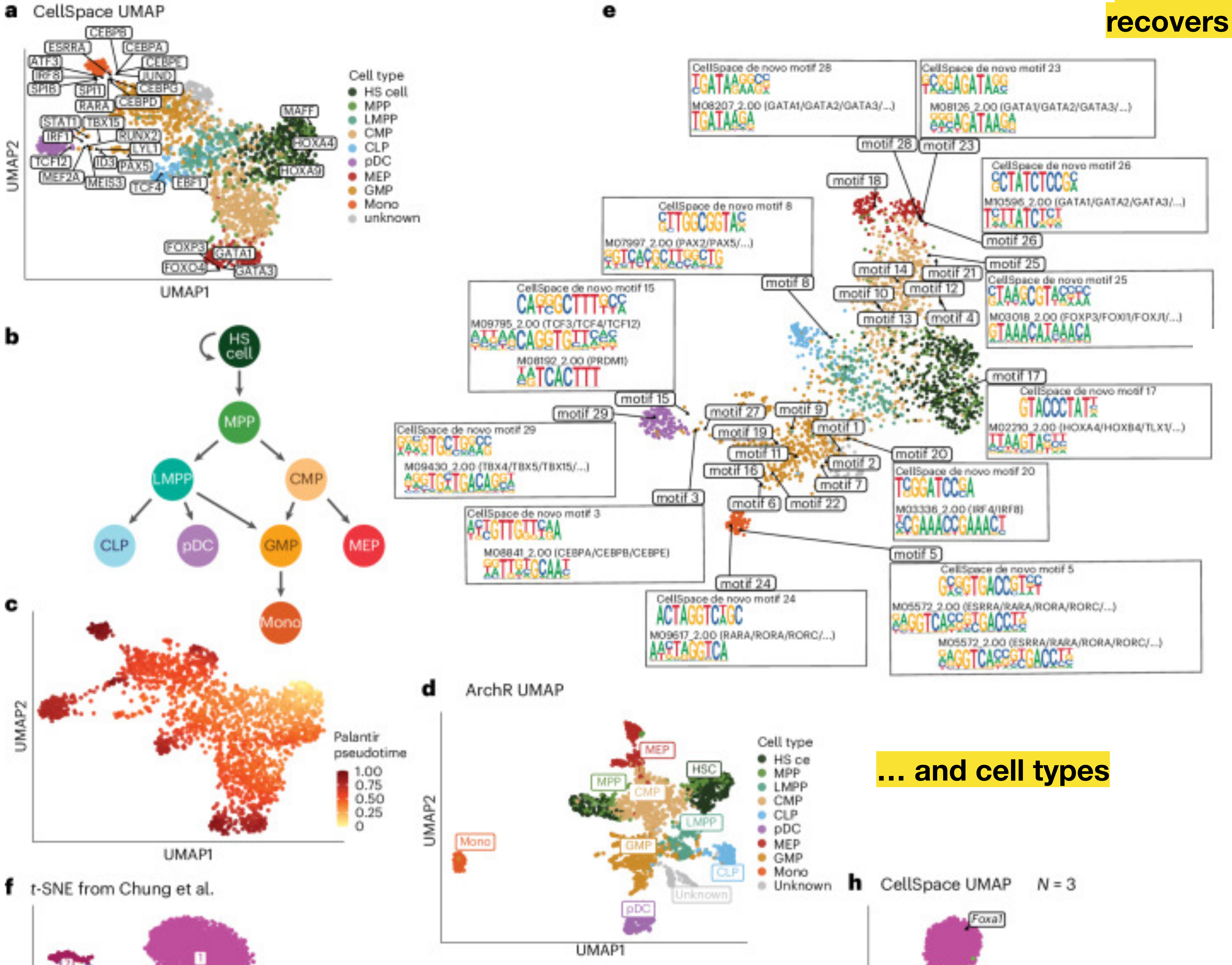
https://doi.org/10.1038/s41592-024-02274-x



Producing a sequence-based embedding space

#IS25 #YIR25 @proftatonetti @tatonetti.bsky.social https://doi.org/10.1038/s41592-024-02274-x

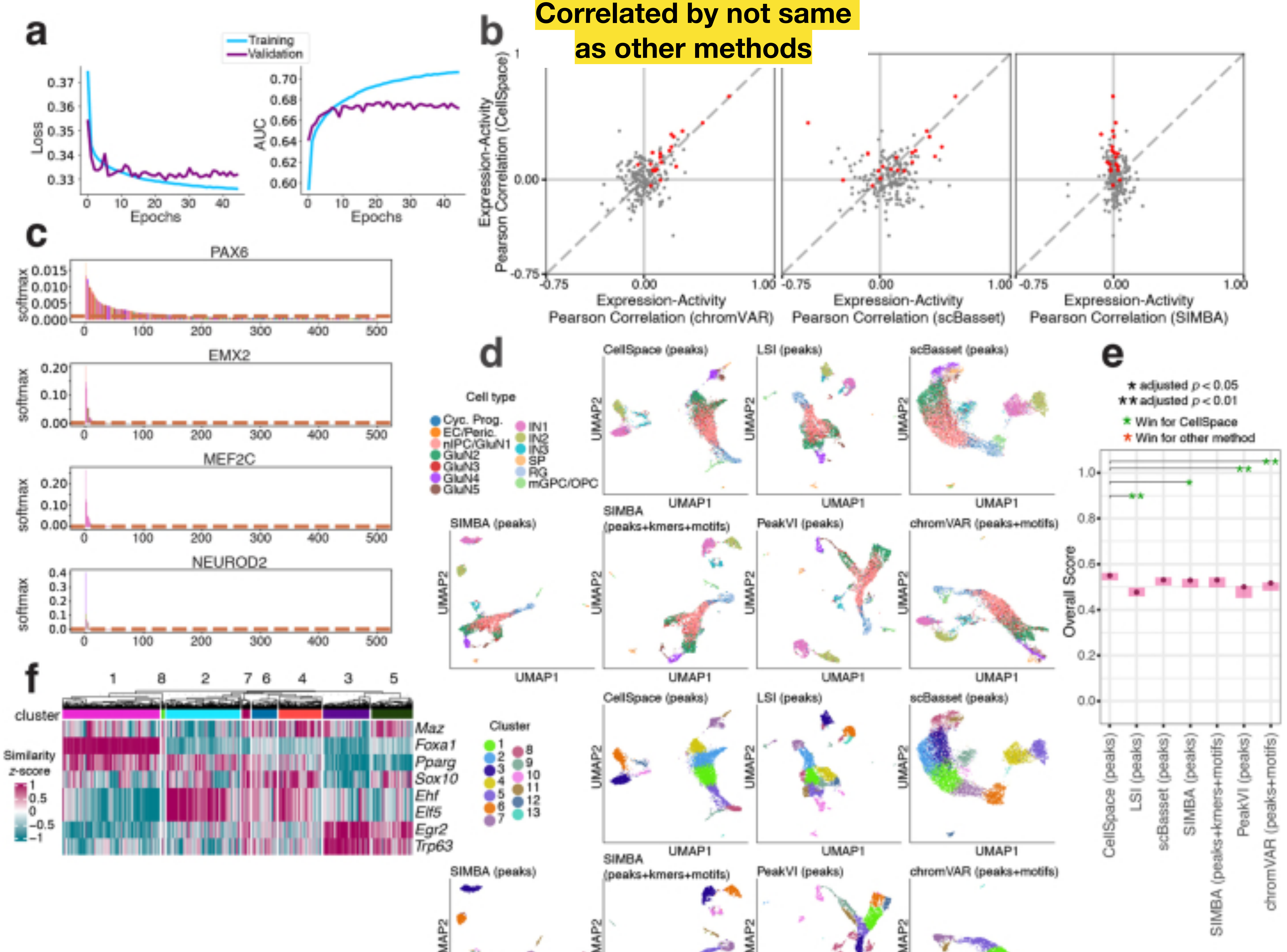
This embedding space recovers known biology



… like TF motifs

… and cell types

Correlated by not same as other methods

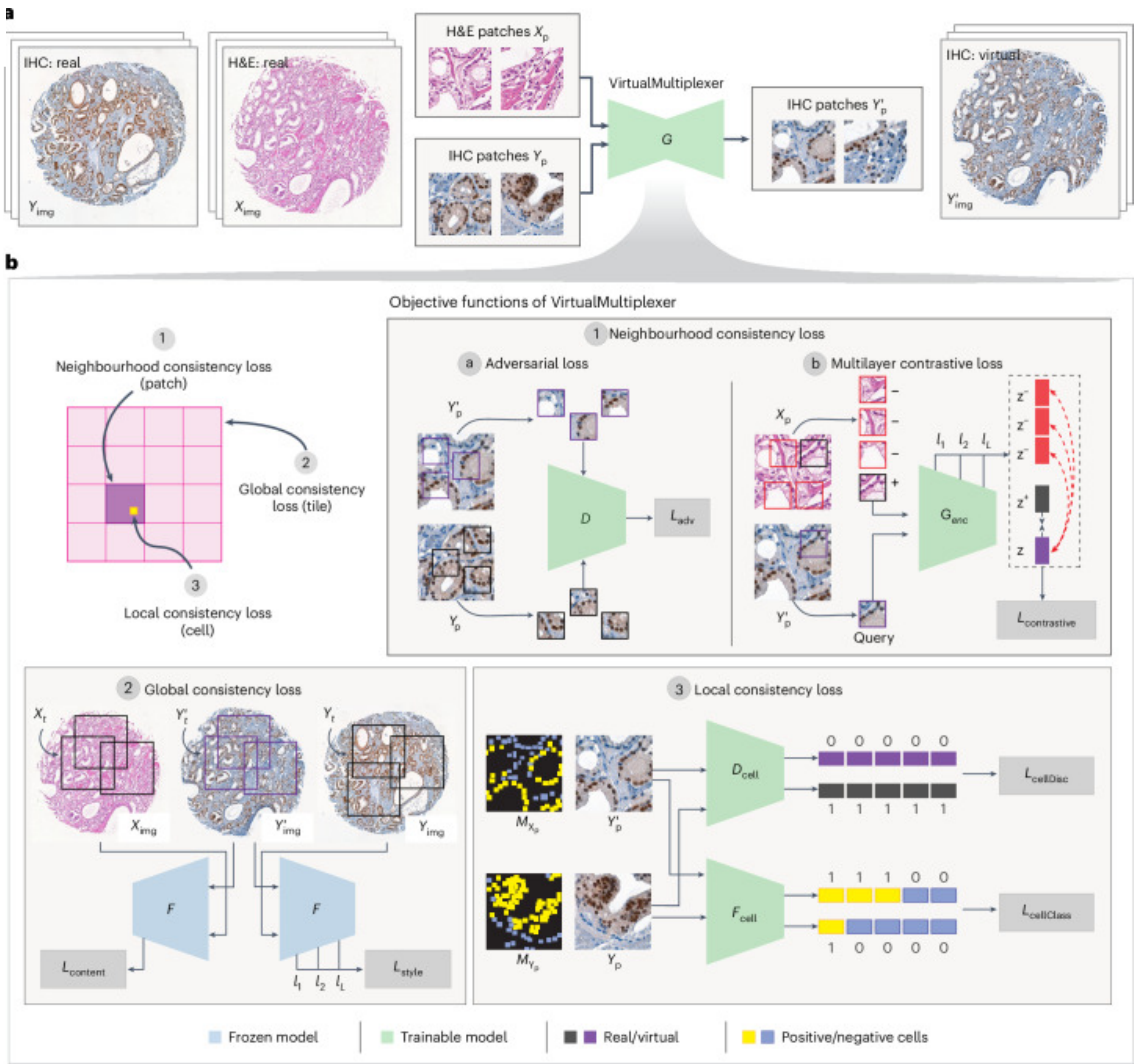


Significantly outperforms other methods

Accelerating histopathology workflows with generative AI-based virtually multiplexed tumour profiling (Pati et al, *Nature Machine Intelligence*)

- Goal: Synthezsize multiplexed immunohistochemistry (IHC) images from standard H&E slides
- Method:
  - VirtualMultiplexer trained on *unpaired* H&E and IHC data to predict IHC markers (AR, NKX3.1, CD44, CD146, p53, ERG) from H&E alone
  - Graph Transformer to predict clinically relevant endpoints
- Result:
  - model-generated IHC images were indistinguishable from real ones
  - Improve time to get IHC from weeks to seconds
- Conclusion: Pretty incredible results and love their “Visual Turing Test” evaluation metric

Unpaired H&E and IHC as input



Multiple loss functions

- neighborhood consistency loss (path level)
- Global consistency loss
- Local consistency loss (cell type priors)

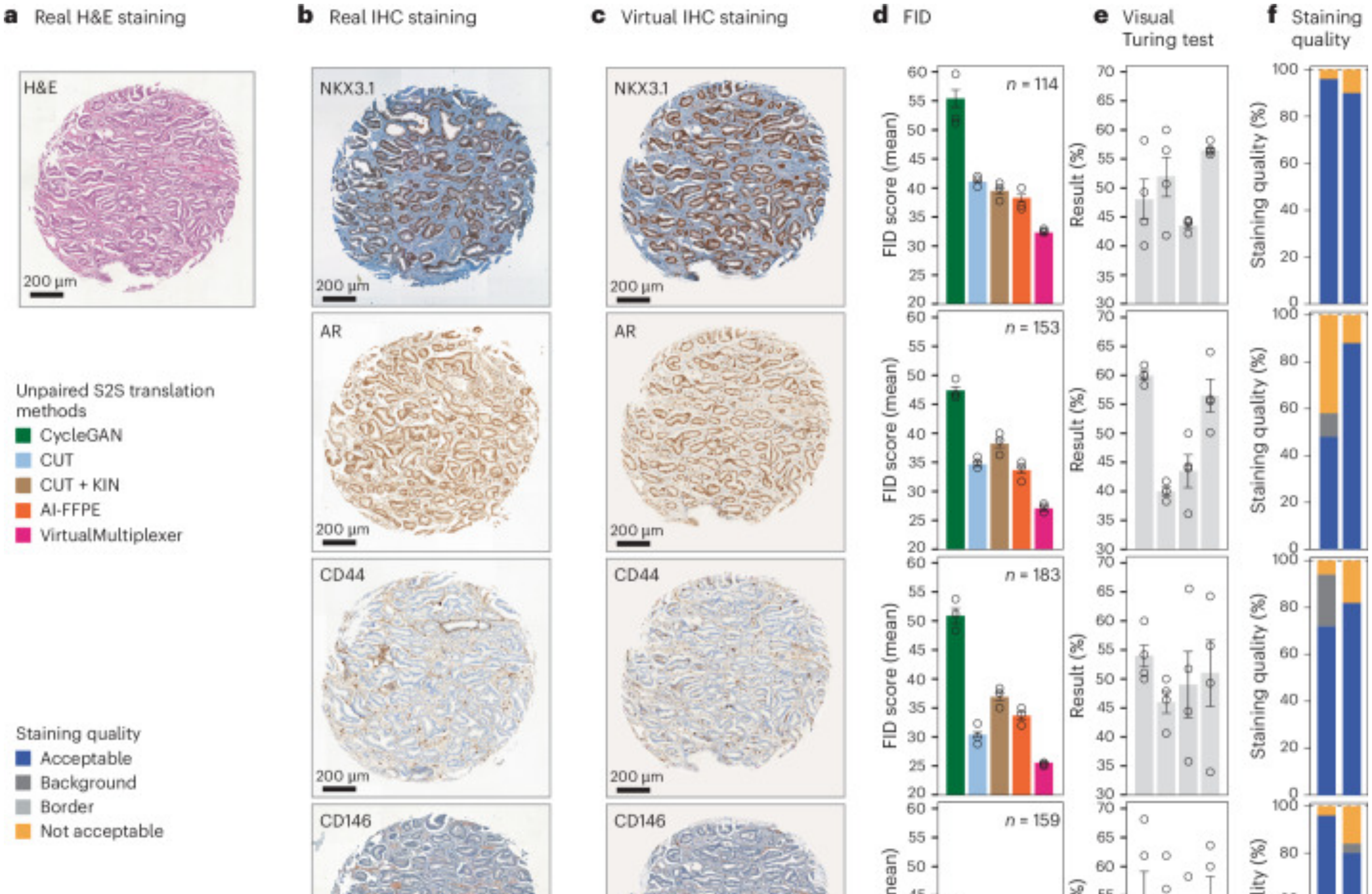
https://doi.org/10.1038/s42256-024-00889-5

Real Input

Real Reference

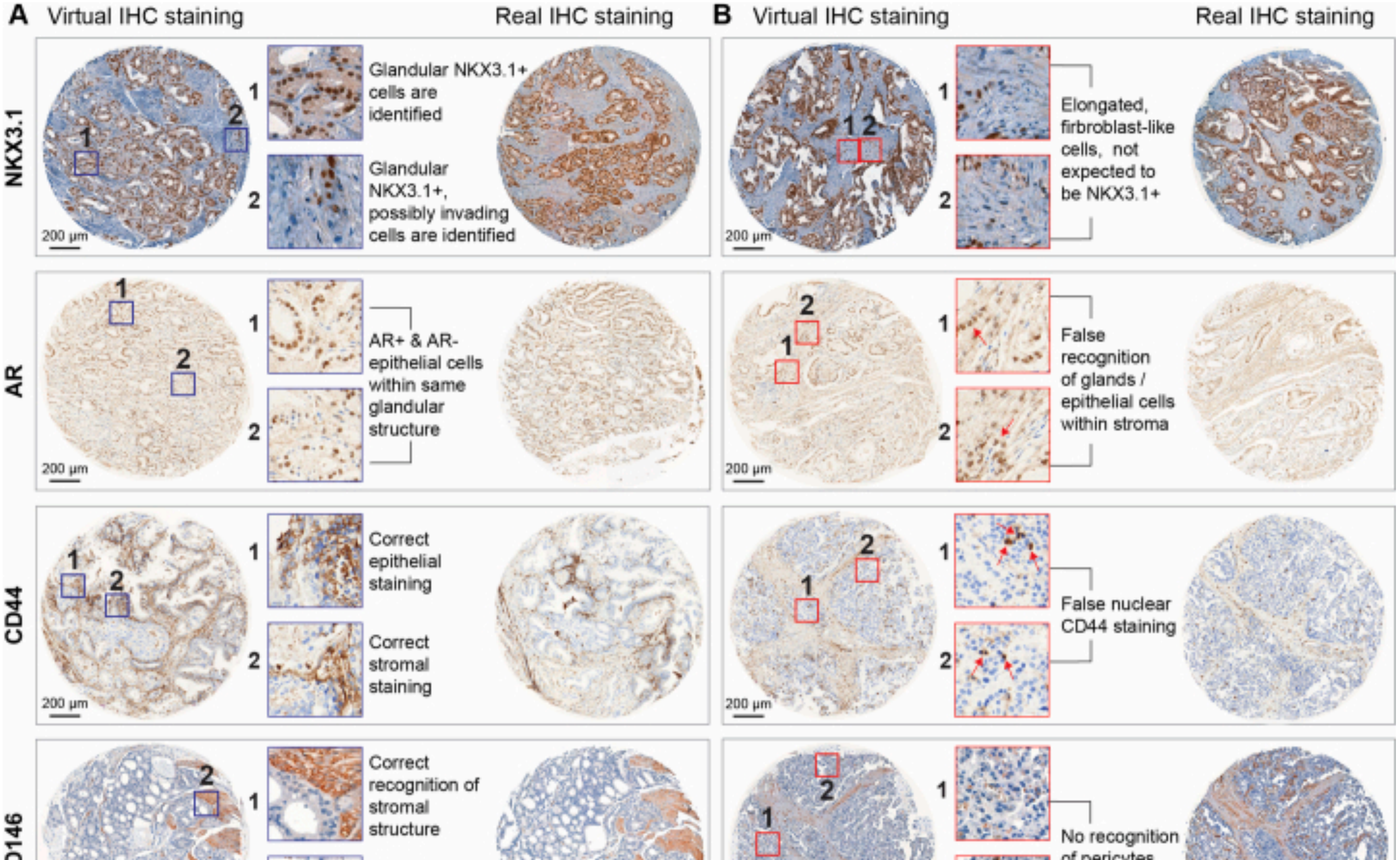
Virtual Staining

Indistinguishable by experts



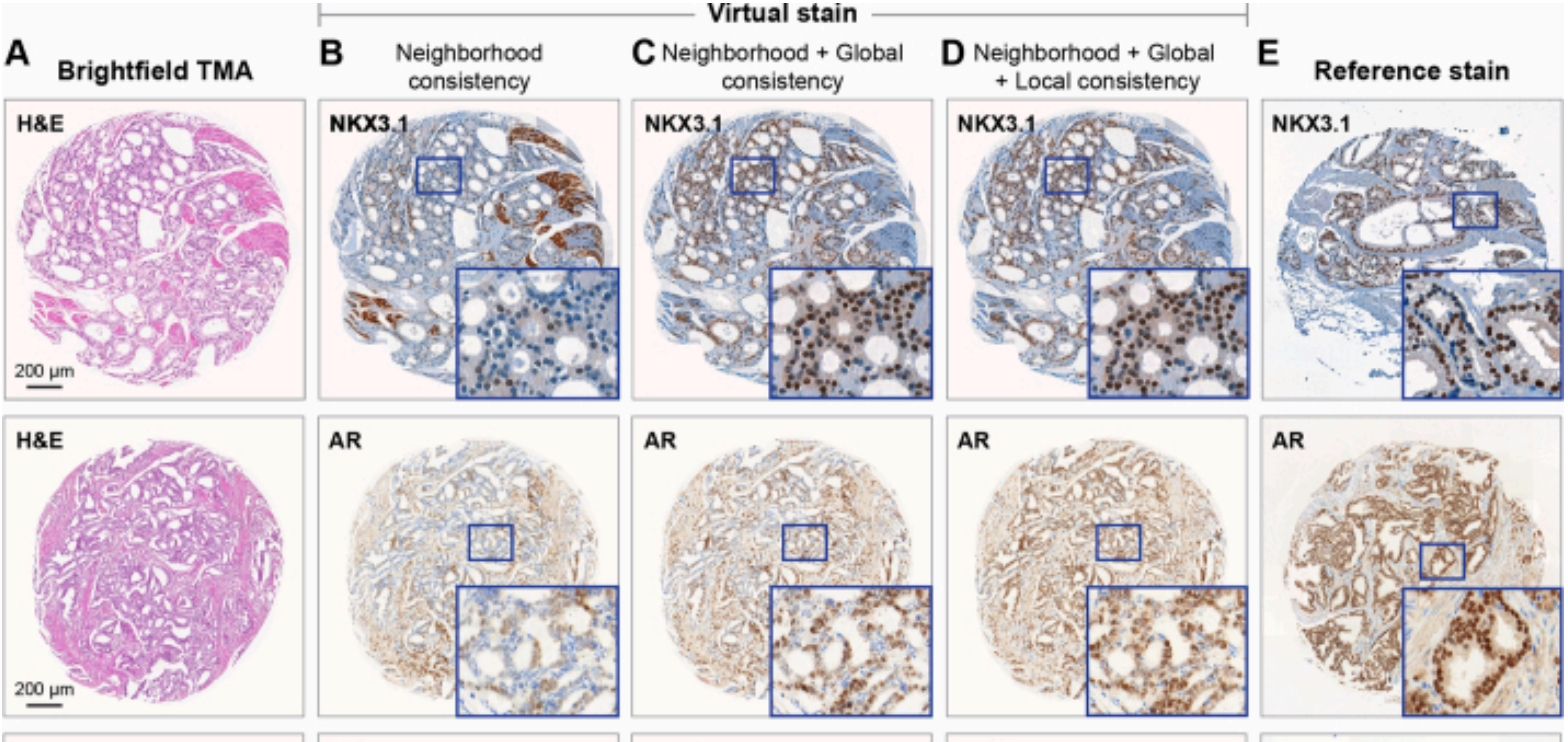
Molecular morphology that the model gets right

... and some it doesn't


A
Virtual IHC staining
Real IHC staining
NKX3.1
Glandular NKX3.1+ cells are identified
Glandular NKX3.1+, possibly invading cells are identified
AR
AR+ & AR- epithelial cells within same glandular structure
CD44
Correct epithelial staining
Correct stromal staining
D146
Correct recognition of stromal structure
200 μm
1
2
B
Virtual IHC staining
Real IHC staining
Elongated, firbroblast-like cells, not expected to be NKX3.1+
False recognition of glands / epithelial cells within stroma
False nuclear CD44 staining
No recognition of pericytes

Visual ablation study

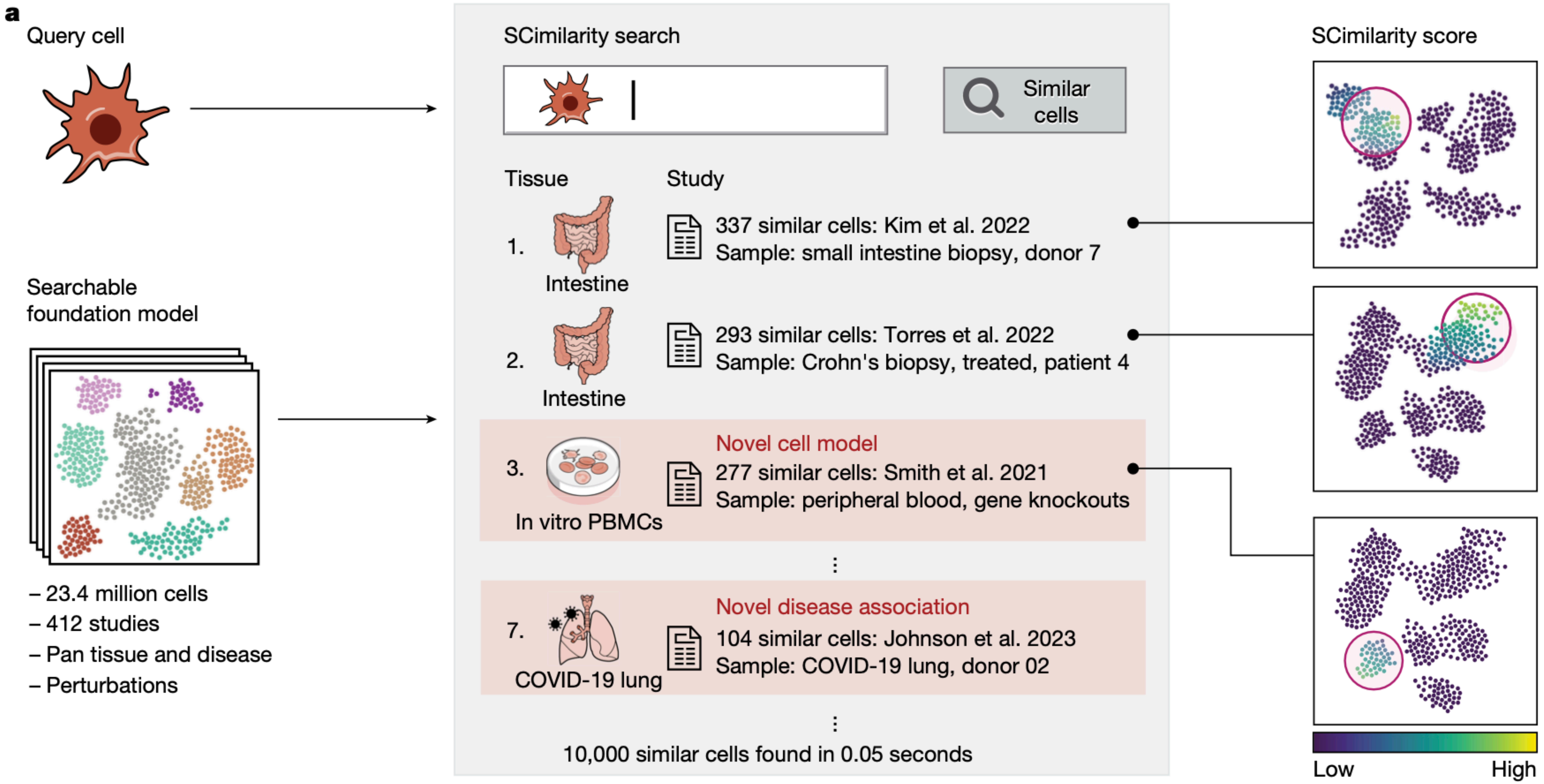
Evaluating each of those losses and their effects on the virtual stain



A cell atlas foundation model for scalable search of similar human cells (Heimberg et al, *Nature*)

- Goal: Build a scRNA-Seq foundation model that enables super fast searches
- Method:
  - Introduce SCimilarity which embeds cells into low-dimensional space while preserving overall representation of the profile
  - Trained on 23.4 million cells from 412 scRNA-seq studies
    - Use **triplet loss** to force similar cells together and **reconstruction loss** to maintain subtle expression variation
  - Precompute an Approximate Nearest Neighbor Index using hnswlib
- Result:
  - Performs great on benchmarks and is super fast:
    - 10,000 similar cells retrieved in 0.05 seconds from a 23.4-million-cell reference dataset
- Conclusion: Allows for rapid scalable searches through massive single cell datasets

Cute graphic showing you query with your cell and get similar results fast



1. Intestine — 337 similar cells: Kim et al. 2022
Sample: small intestine biopsy, donor 7

2. Intestine — 293 similar cells: Torres et al. 2022
Sample: Crohn's biopsy, treated, patient 4

3. In vitro PBMCs — Novel cell model
277 similar cells: Smith et al. 2021
Sample: peripheral blood, gene knockouts

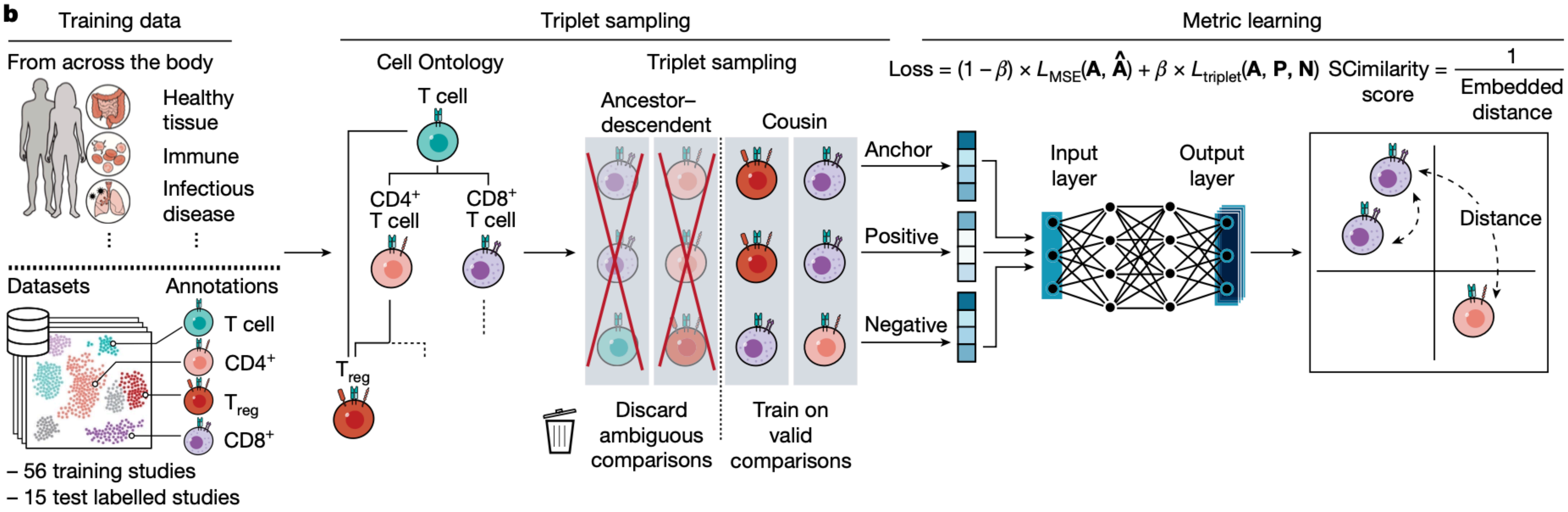
⋮

7. COVID-19 lung — Novel disease association
104 similar cells: Johnson et al. 2023
Sample: COVID-19 lung, donor 02

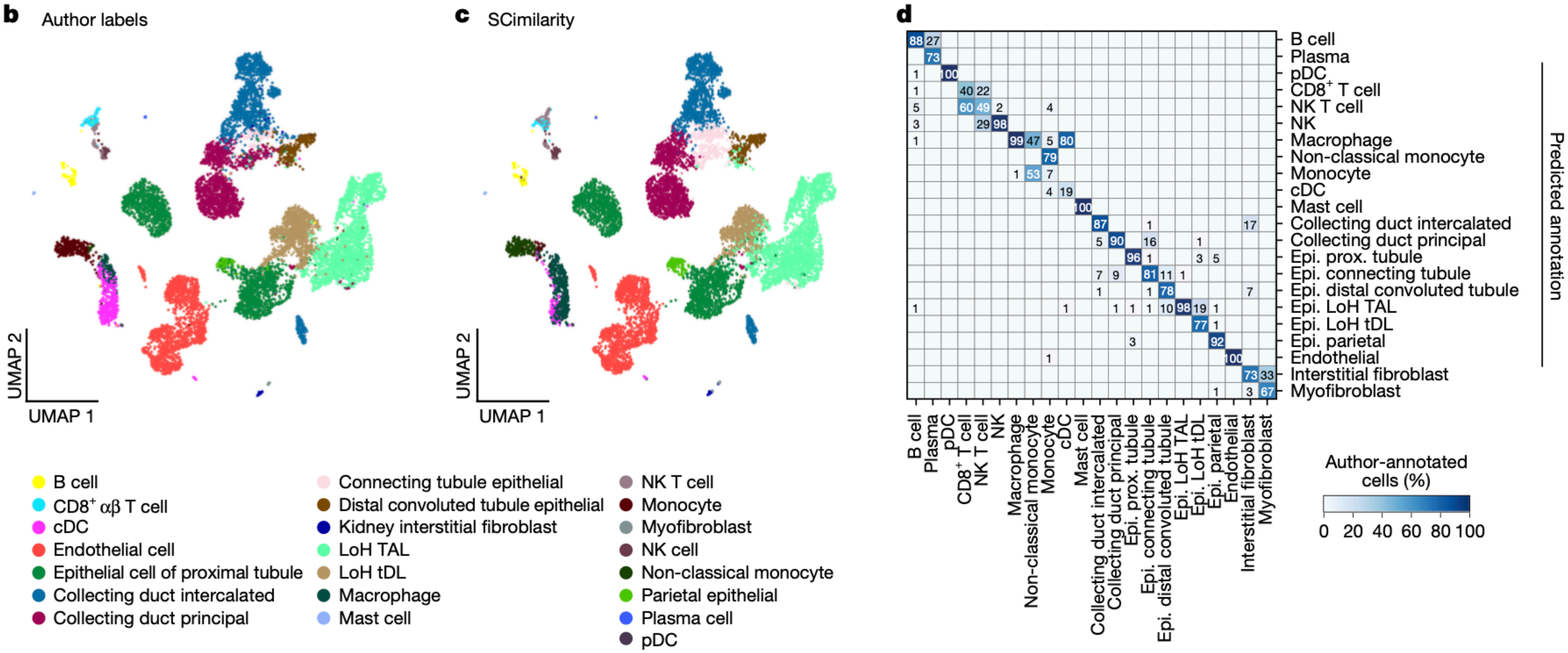
⋮

10,000 similar cells found in 0.05 seconds

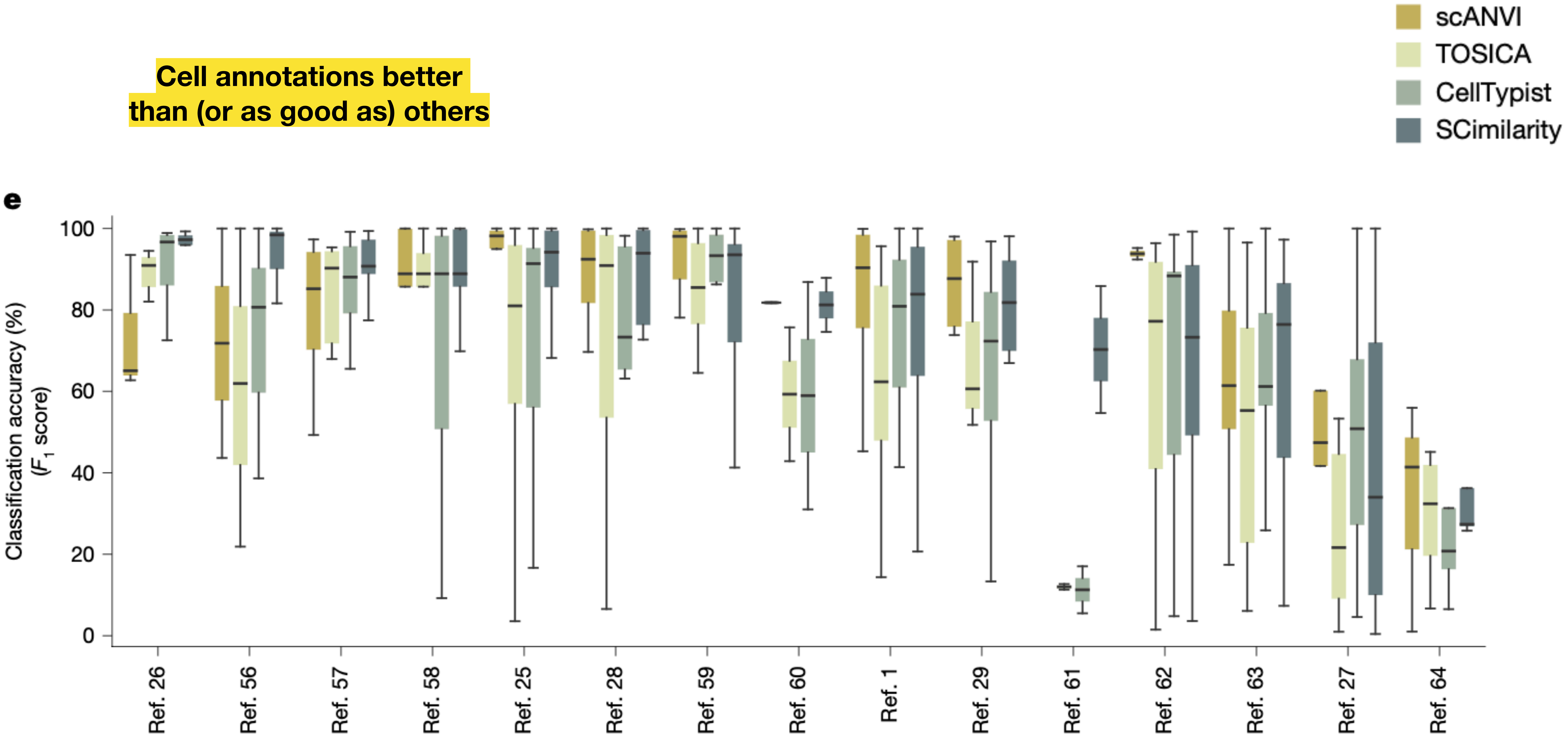
Triplet loss not only pushes similar samples together, it pushes dissimilar ones apart



Accurately annotates cells (and does it in 0.02s)



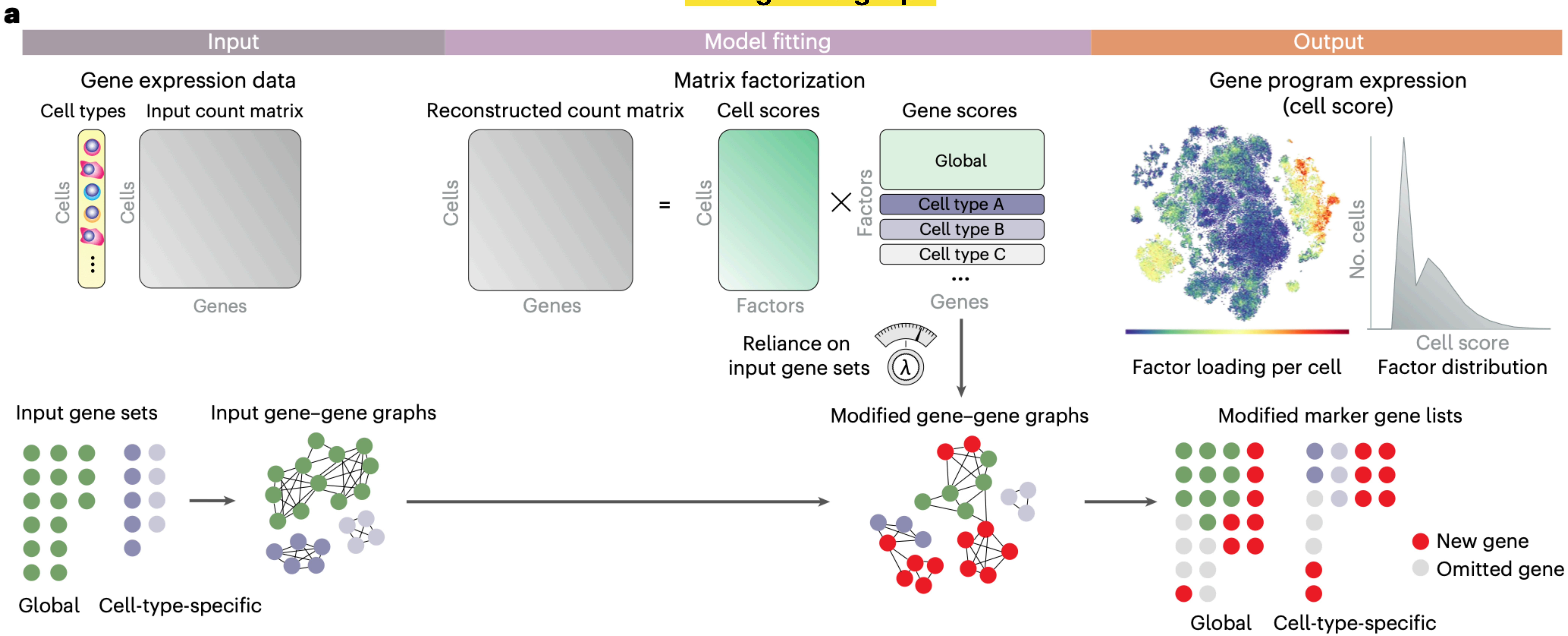
Cell annotations better than (or as good as) others



Supervised discovery of interpretable gene programs from single-cell data (Kunes, Walle et al, *Nature Biotechnology*)

- Goal: Use prior knowledge to extract biologically meaningful gene programs from single-cell RNA-seq data
- Method: Introduce the Spectra algorithm, a matrix factorization method constrained by existing biological knowledge
- Result:
  - Outperforms existing methods (expiMap, Slalom, NMF)
    - Finds immune checkpoint therapy (ICT) response factors in CD8+ T cells
    - Identifies tumor-reactive vs. exhausted T cell programs (prev. methods struggled here)
    - Predicts patient response to anti-PD-1 therapy
- Conclusion: Biological constraints on an unsupervised approach can get you a good balance between discovery and realism

Penalty function forces the matrix factorization to align the graph

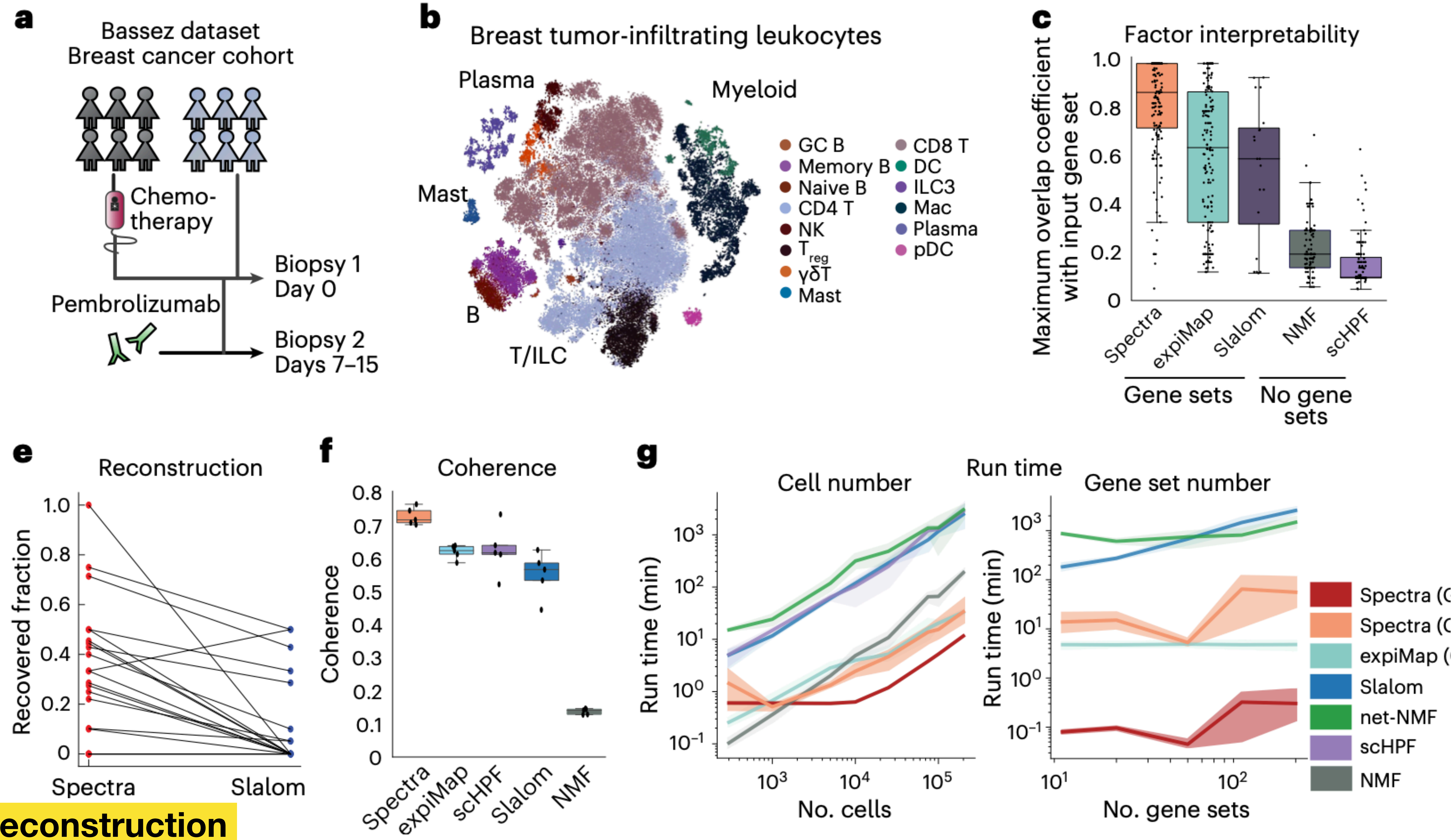


Build gene-gene graphs from existing knowledge

Ready for downstream analysis

Eval on breast cancer dataset

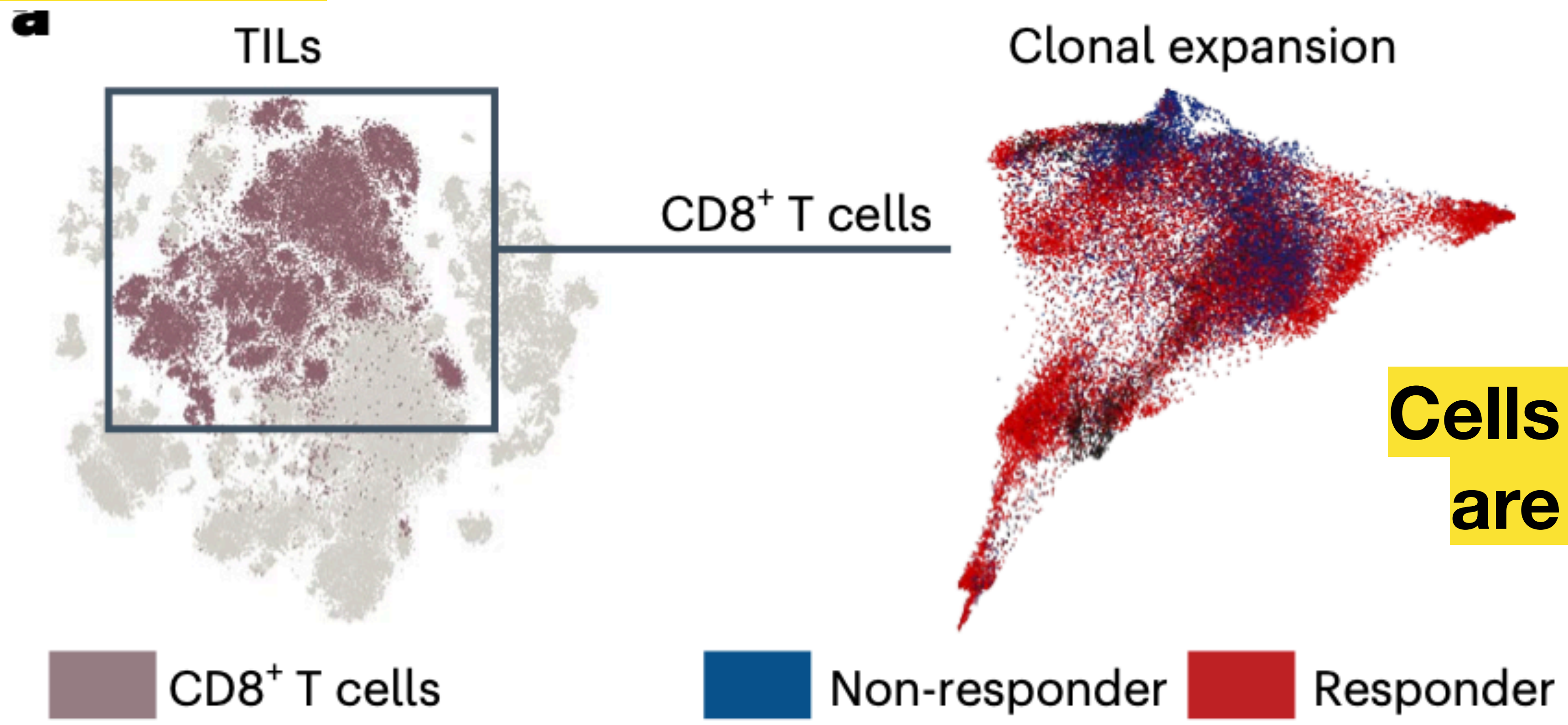
Better overlap with expected genes



Better reconstruction of held out genes

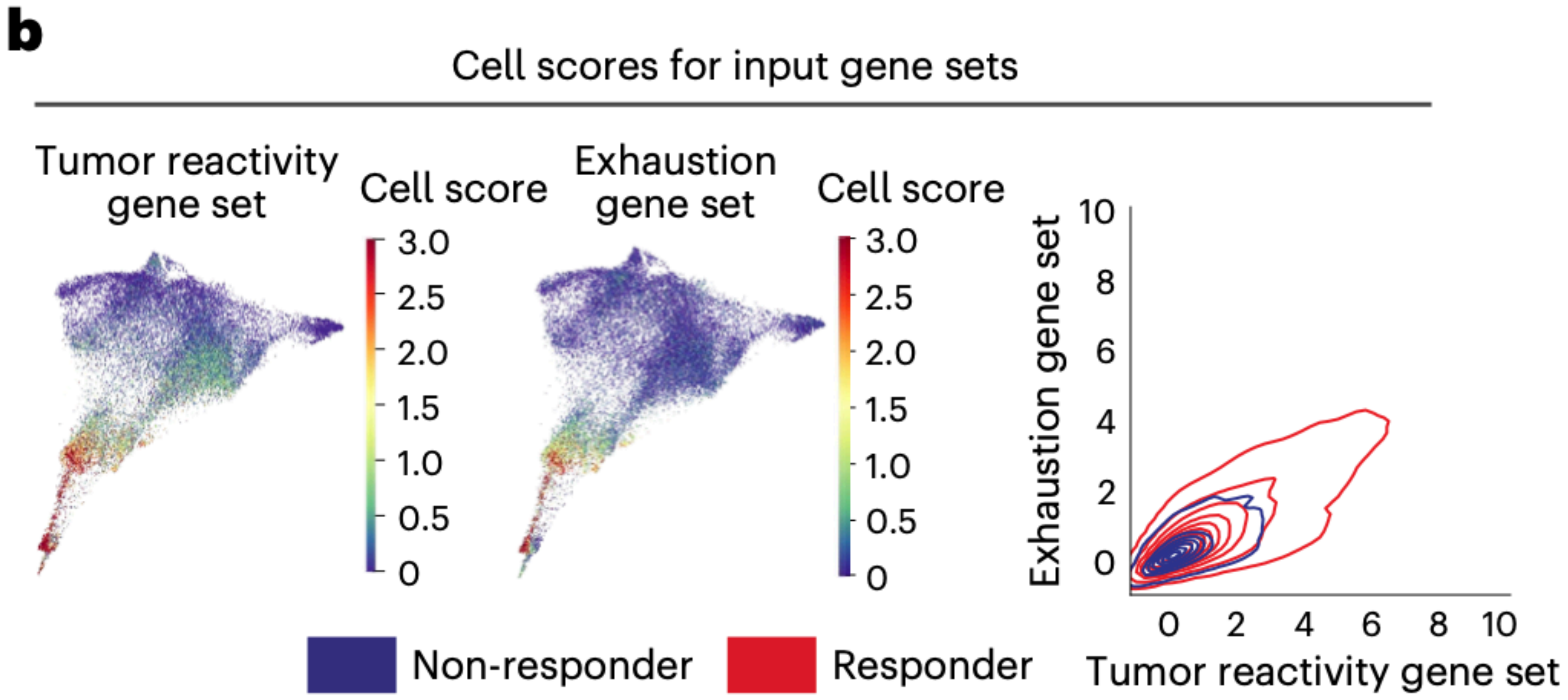
Faster!

Differentiating tumor-reactive CD8 cells from exhausted CD8 cells is difficult

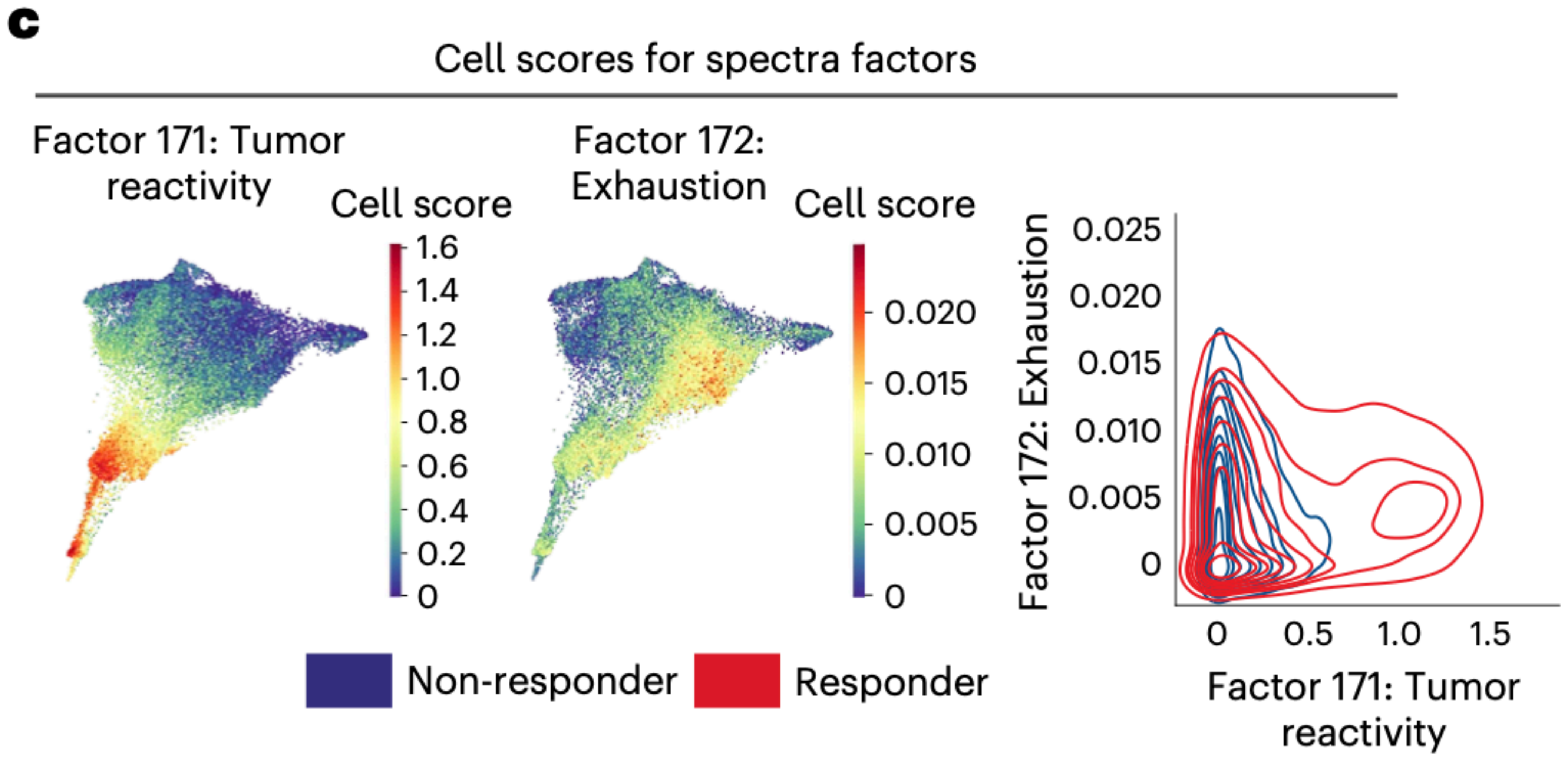


Cells that expand are still active

Previous methods can't separate these cells



Spectra can separate these cells



"What Was I Made For?"

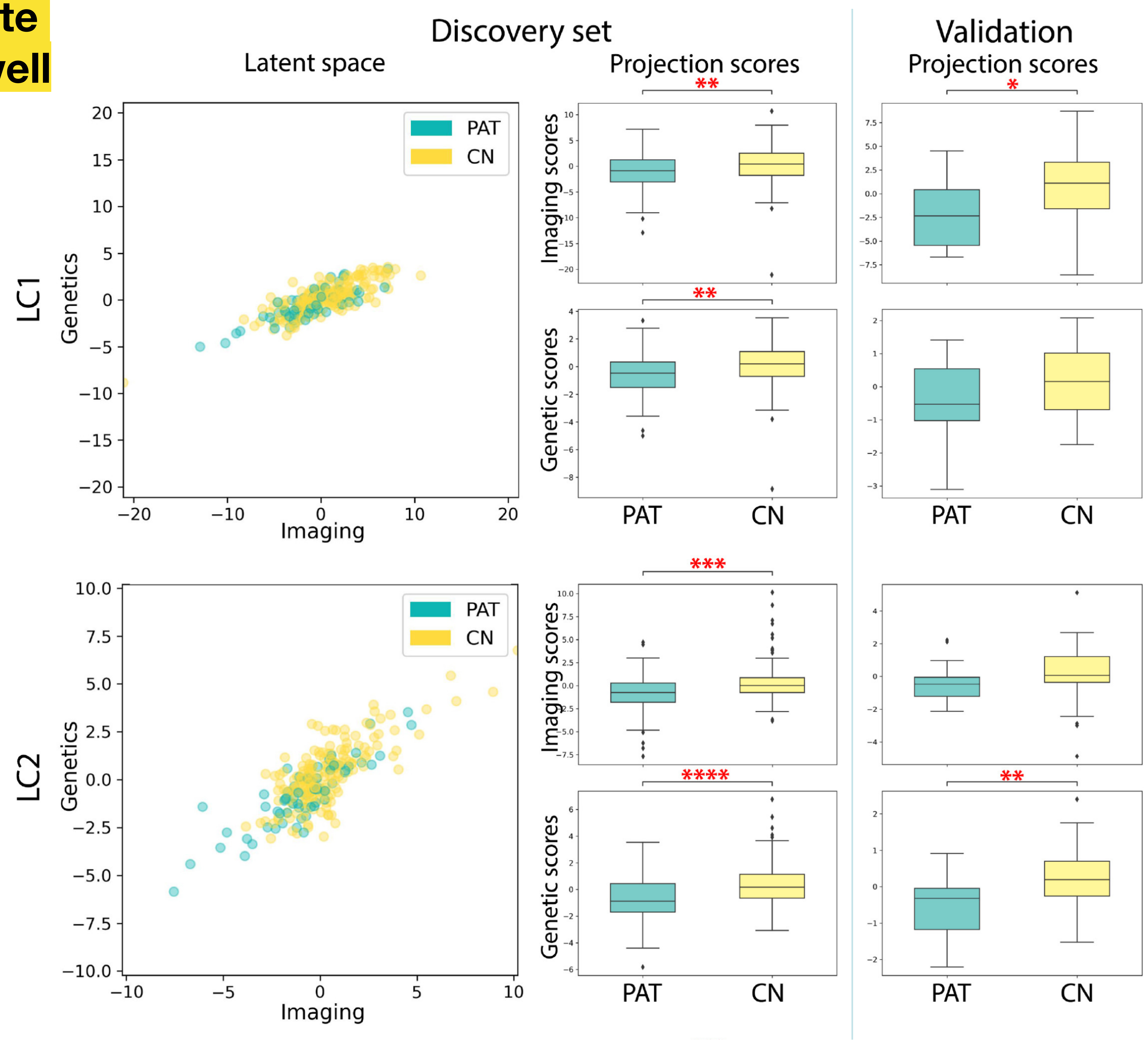
Biomarker Discovery & Validation

https://doi.org/10.1016/j.jbi.2023.104569

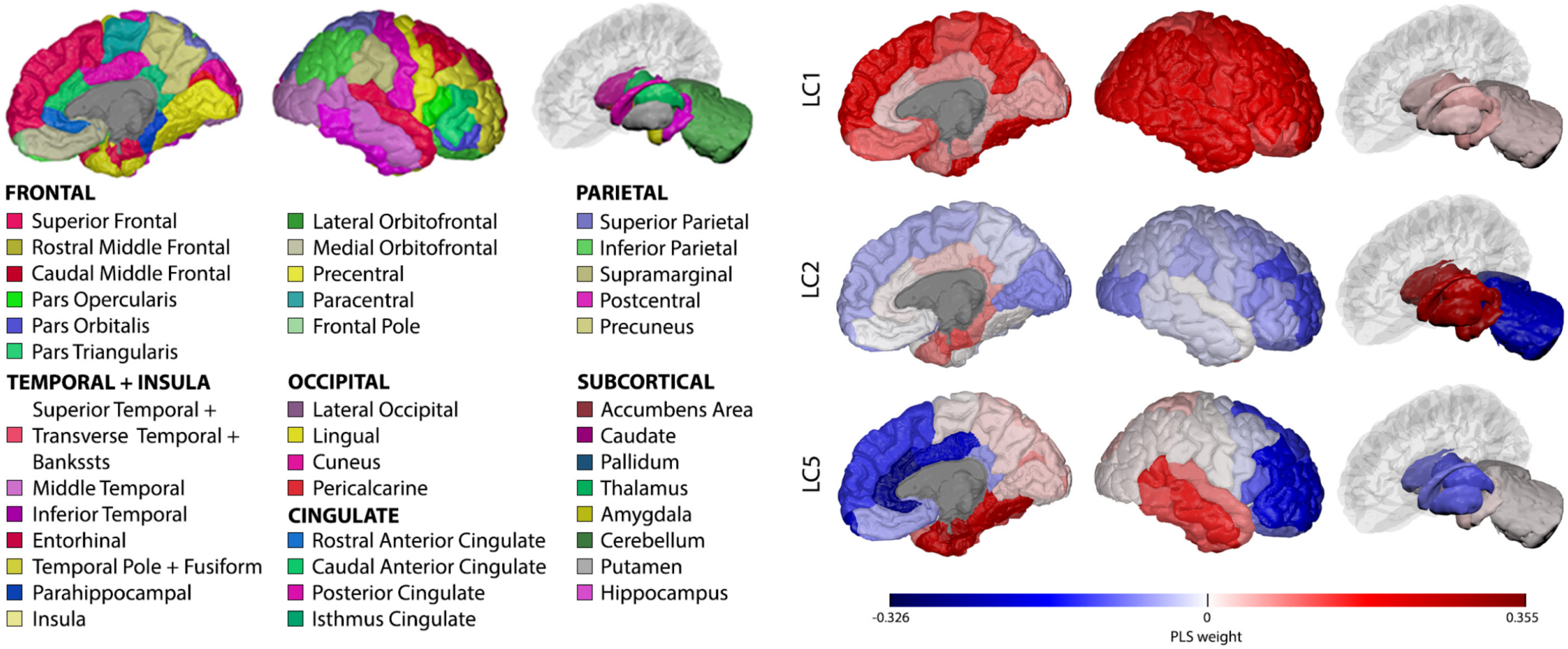
Identifying the joint signature of brain atrophy and gene variant scores in Alzheimer’s Disease (Cruciani et al, *JBI*)

- Goal: Move beyond univariate modeling in AD
- Method:
    - Derive individual matrices to represent imaging and genetics separately
    - Using partial least squares to identify joint latent space
    - Validate using permutation testing and subsequent transcriptional analysis
- Result:
    - EPHX1 (Biological oxidation pathway) → Linked to subcortical atrophy
    - BCAS1 (Myelination process) → Temporal lobe atrophy (especially dentate gyrus)
- Conclusion: Statistical (not AI) based approaches are still relevant

Latent spaces separate patients fro controls well



Partial least squares components map to relevant brain structures



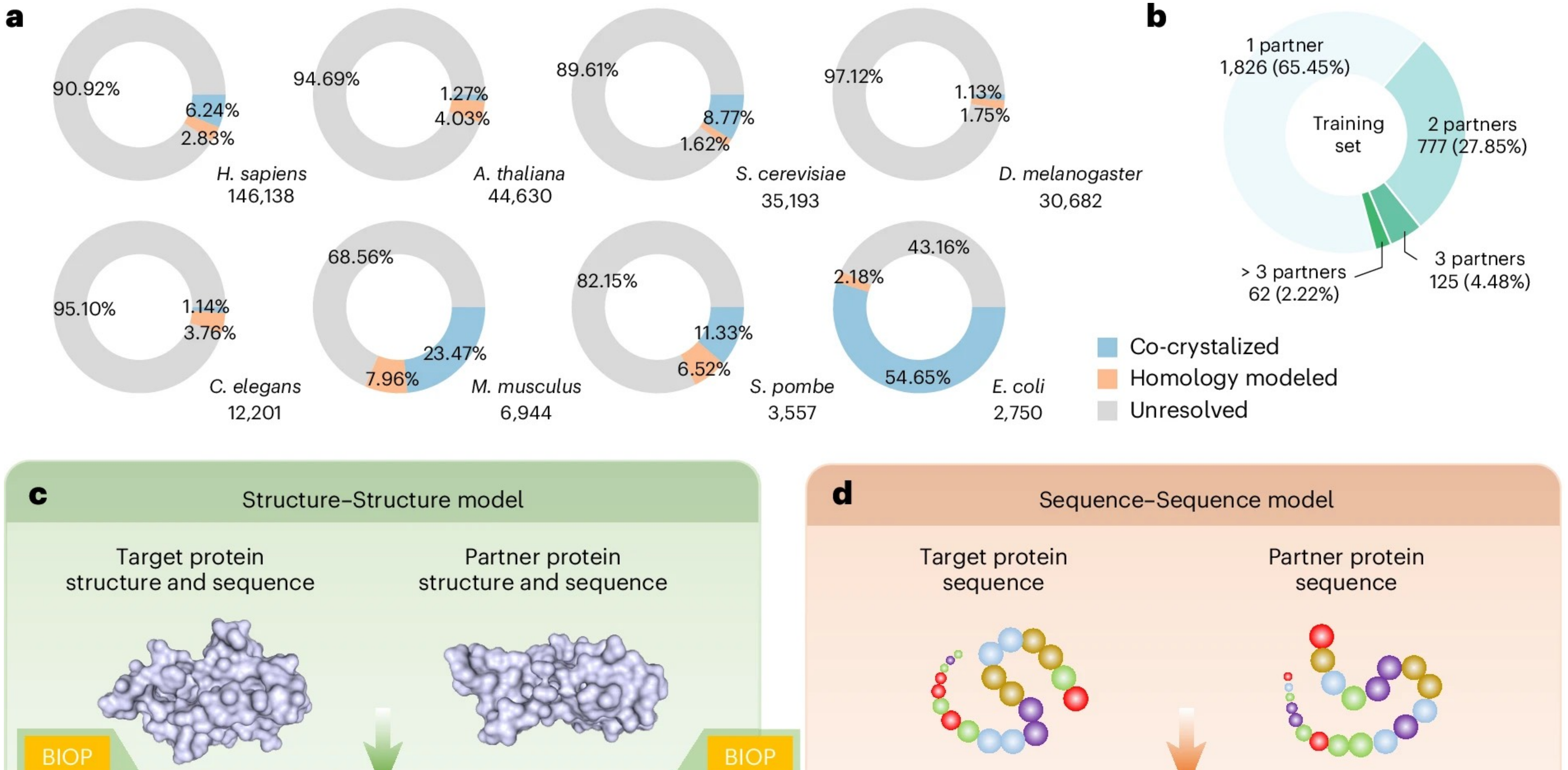
A structurally informed human protein–protein interactome reveals proteome-wide perturbations caused by disease mutations (Xiong et al, *Nature Biotechnology*)

- Goal: To predict protein-protein interaction interfaces without direct structural information
- Method:
  - An ensemble of four deep learning methods (structure-structure, sequence-sequence, structure-sequence, and sequence-structure)
  - Graph Convolutional Networks and RNNs
- Result:
  - Predicts alleles enriched for disease-associated mutations
  - Applied to 11k cancer genomes and found 586 “oncoPPIs”
- Conclusion: Adds a critical level of understanding to PPIs; goes way beyond what we’ve seen from computational methods in this space before.

The training set

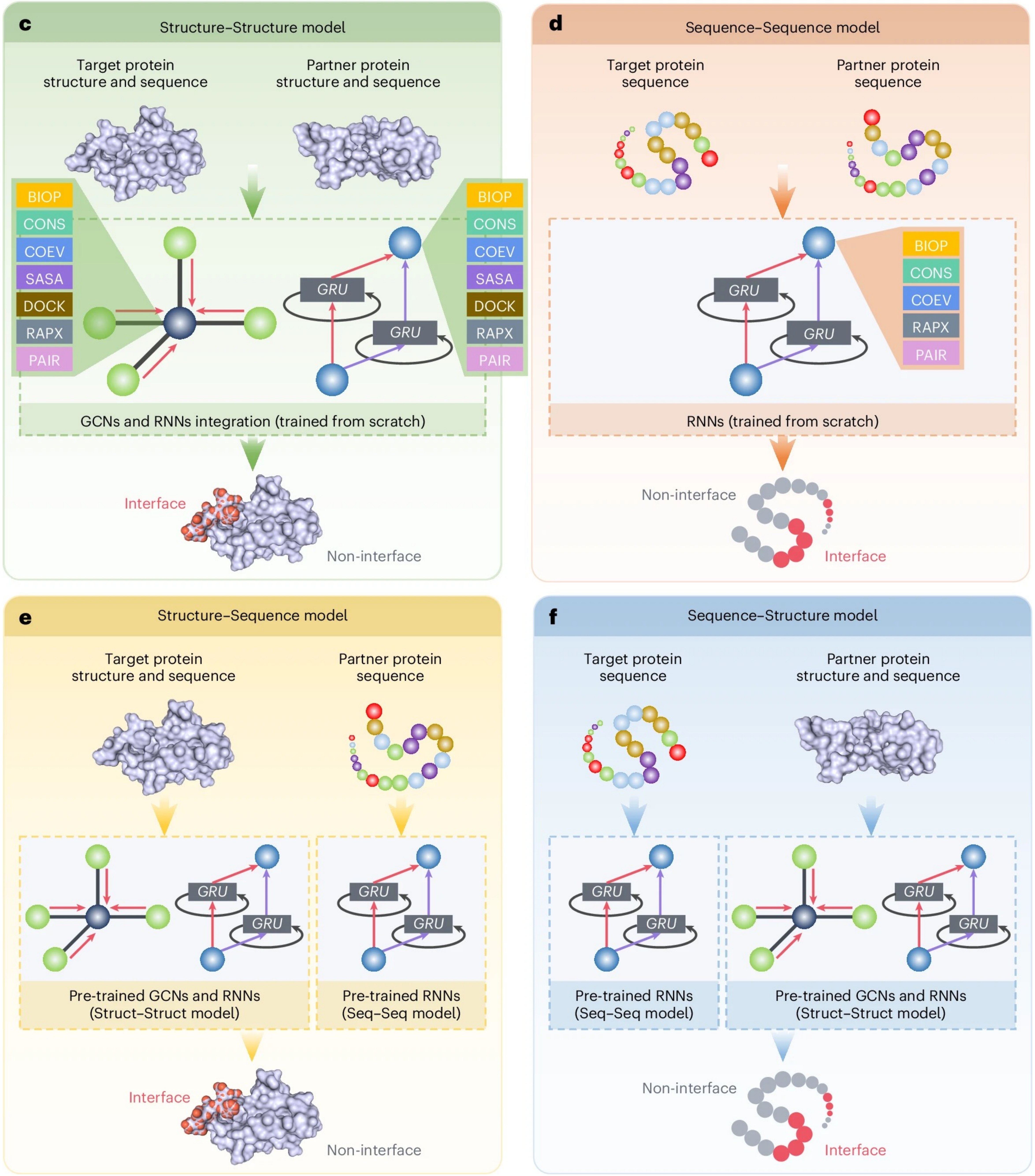
A small amount of PPIs are resolved structurally

Using Proteins with >1 partner allows for predicting “partner-specific” interfaces



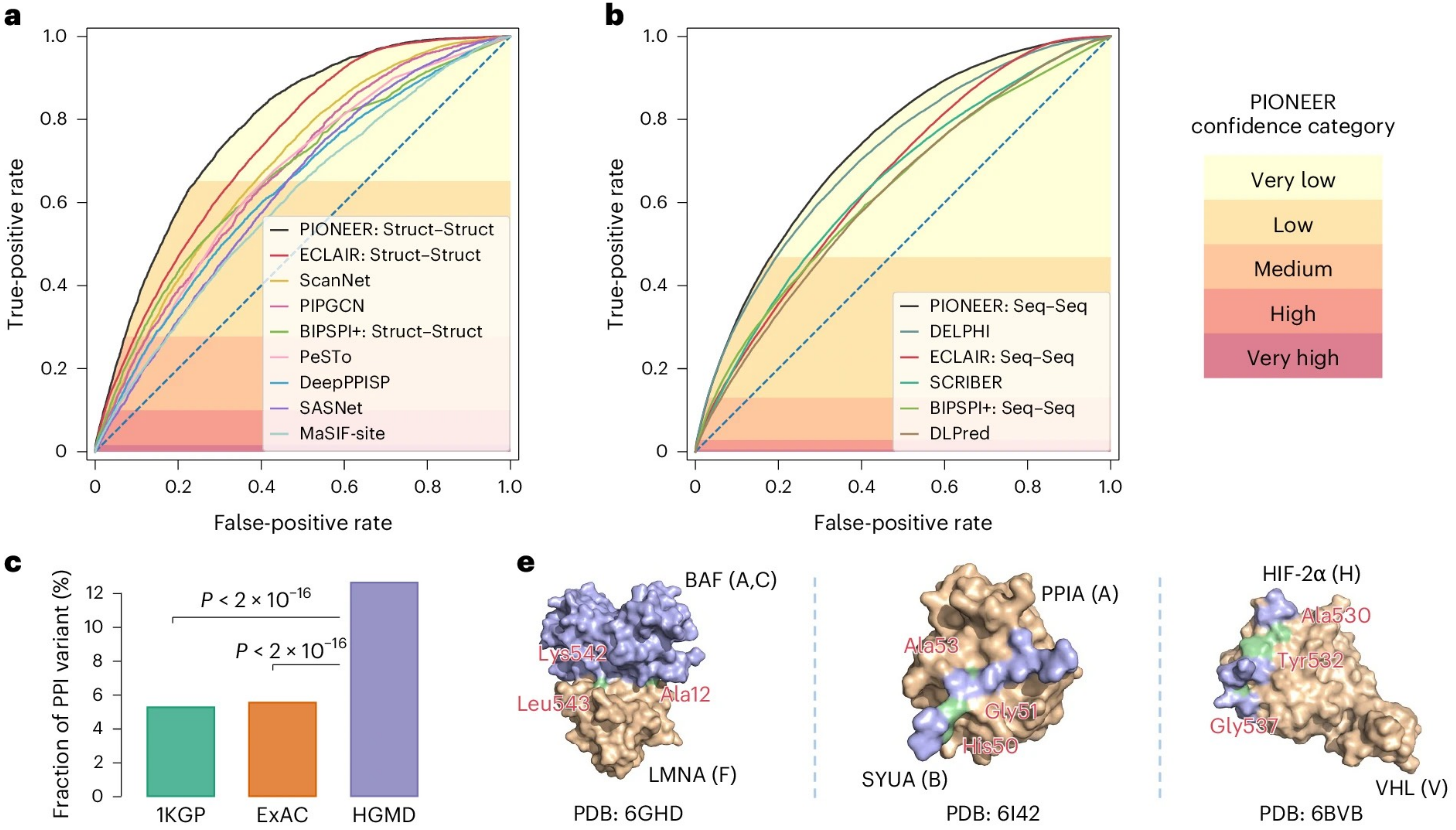
https://doi.org/10.1038/s41587-024-02428-4

The ensemble


c
Structure–Structure model
Target protein structure and sequence
Partner protein structure and sequence
BIOP
CONS
COEV
SASA
DOCK
RAPX
PAIR
GRU
GRU
GCNs and RNNs integration (trained from scratch)
Interface
Non-interface
d
Sequence–Sequence model
Target protein sequence
Partner protein sequence
BIOP
CONS
COEV
RAPX
PAIR
GRU
GRU
RNNs (trained from scratch)
Non-interface
Interface
e
Structure–Sequence model
Target protein structure and sequence
Partner protein sequence
GRU
GRU
Pre-trained GCNs and RNNs (Struct–Struct model)
Pre-trained RNNs (Seq–Seq model)
Interface
Non-interface
f
Sequence–Structure model
Target protein sequence
Partner protein structure and sequence
GRU
GRU
Pre-trained RNNs (Seq–Seq model)
Pre-trained GCNs and RNNs (Struct–Struct model)
Non-interface
Interface

https://doi.org/10.1038/s41587-024-02428-4

Outperforms both state-of-the-art Structure-based and Sequence-based methods



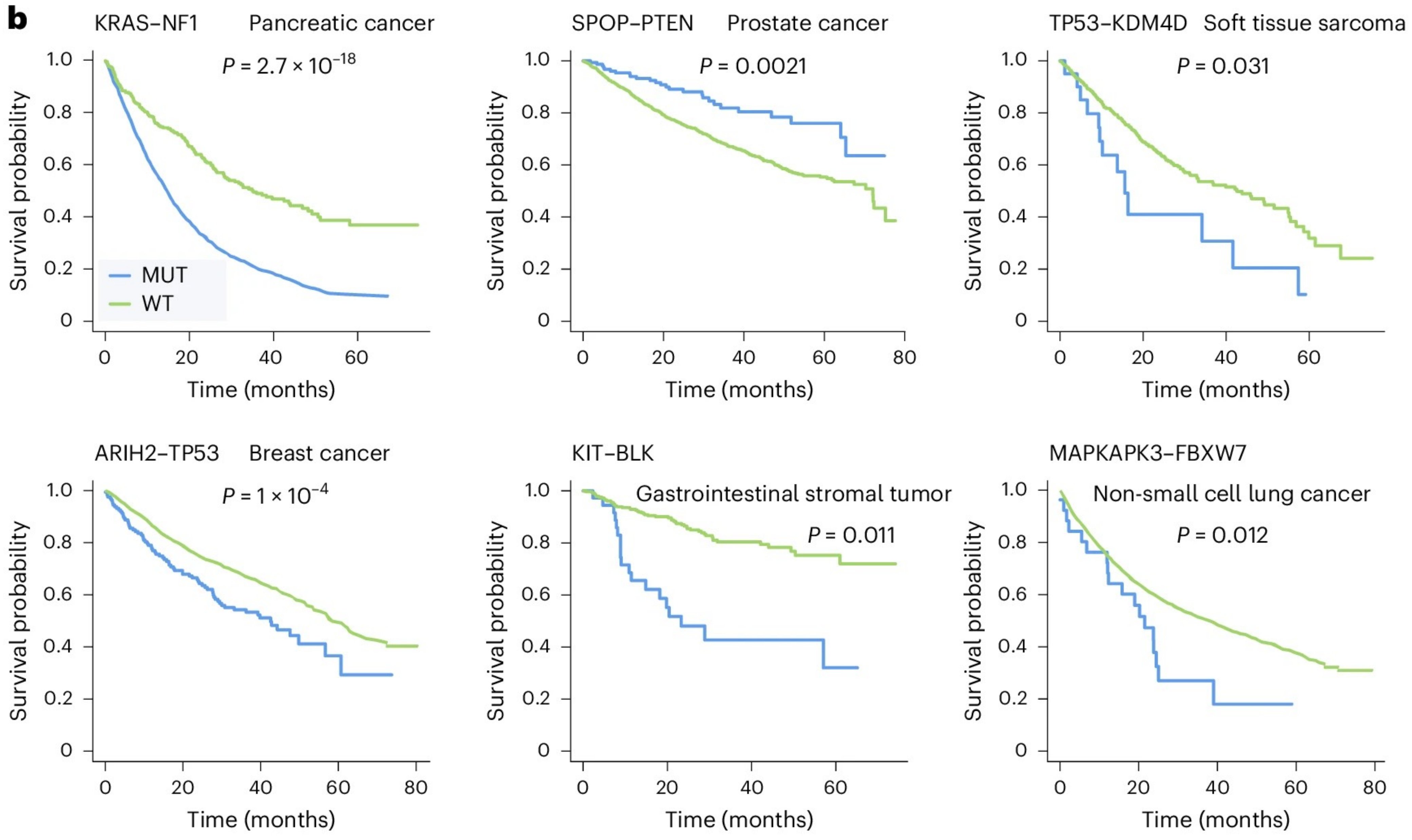
Enrichment of mutation burden

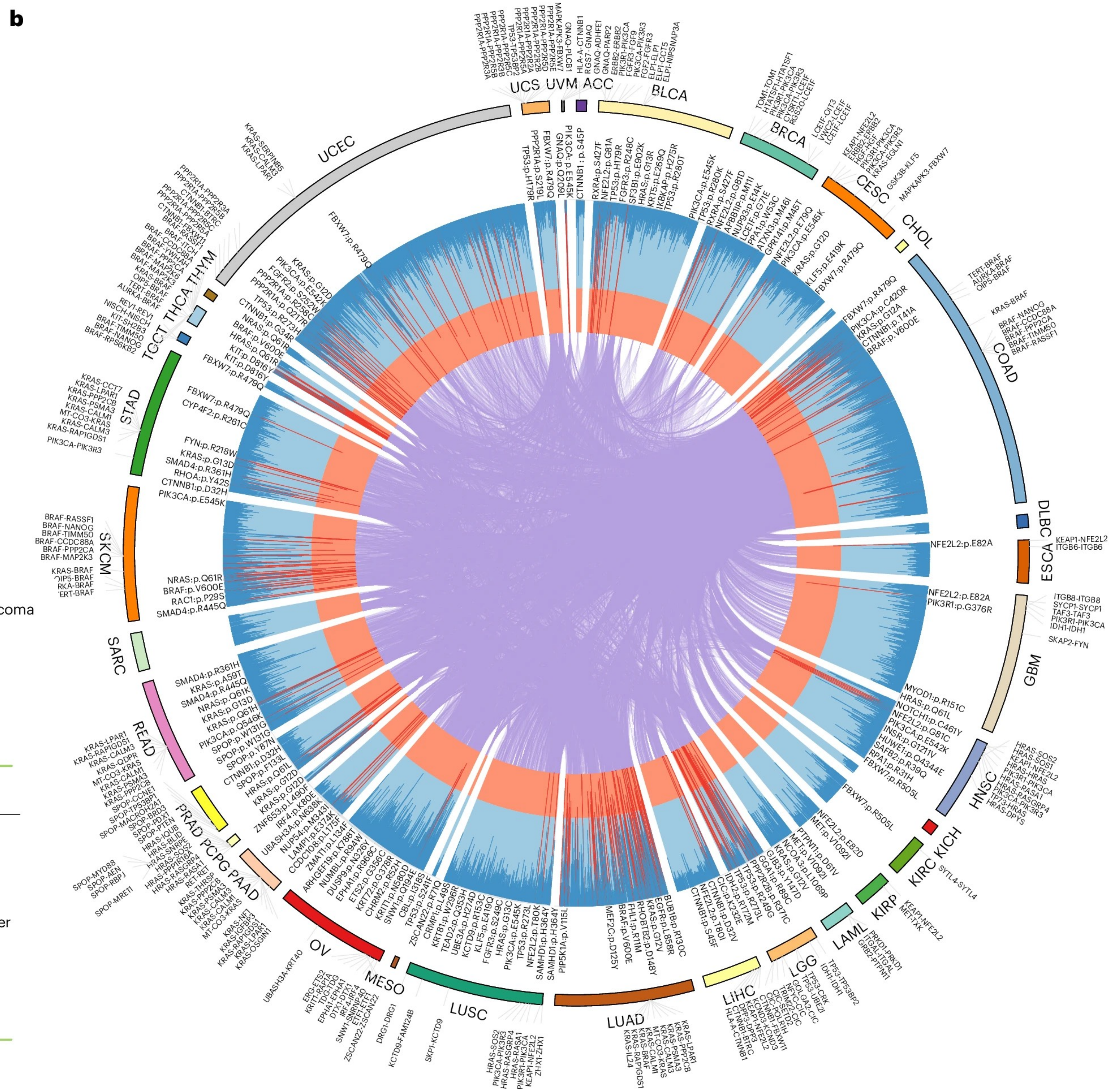
Example of predicted PPI interfaces

https://doi.org/10.1038/s41587-024-02428-4

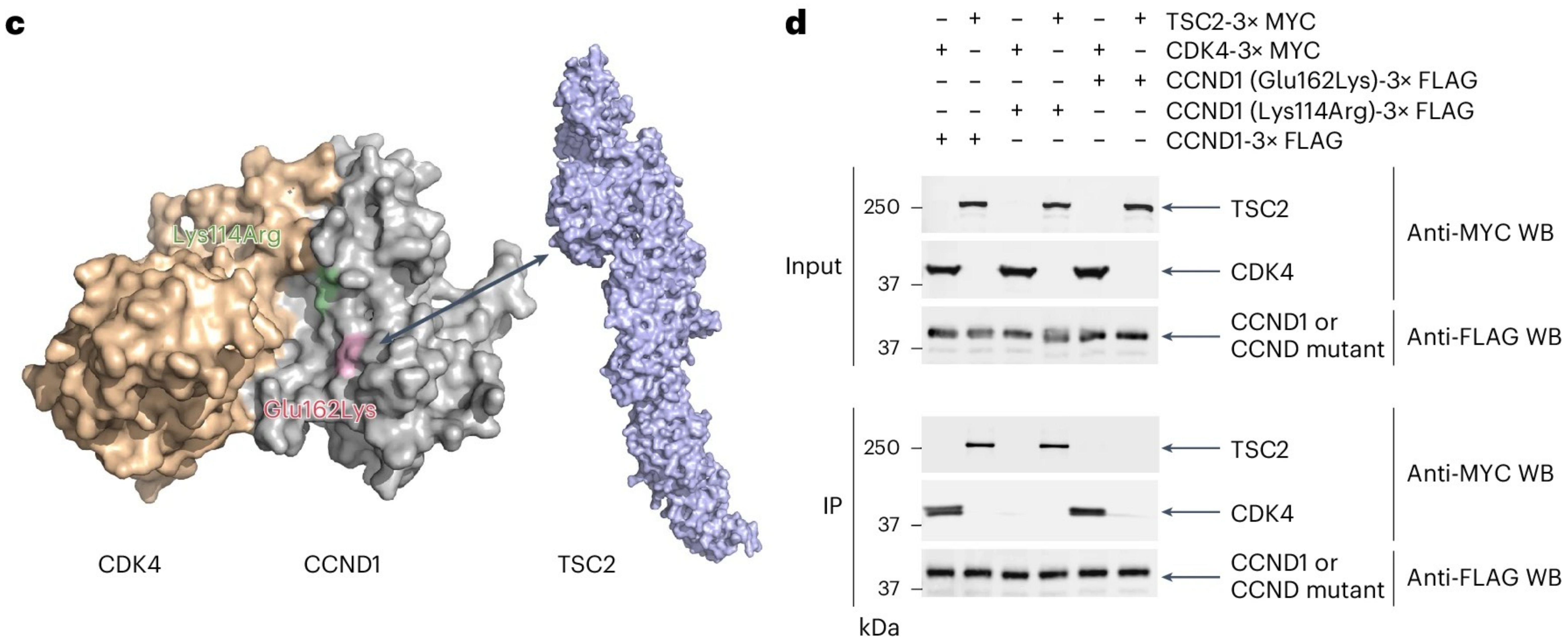
The cancer protein-protein-interactome

… is predictive of survival in TCGA





I’m told this is experimental validation of the predicted interface 🤷‍♂️



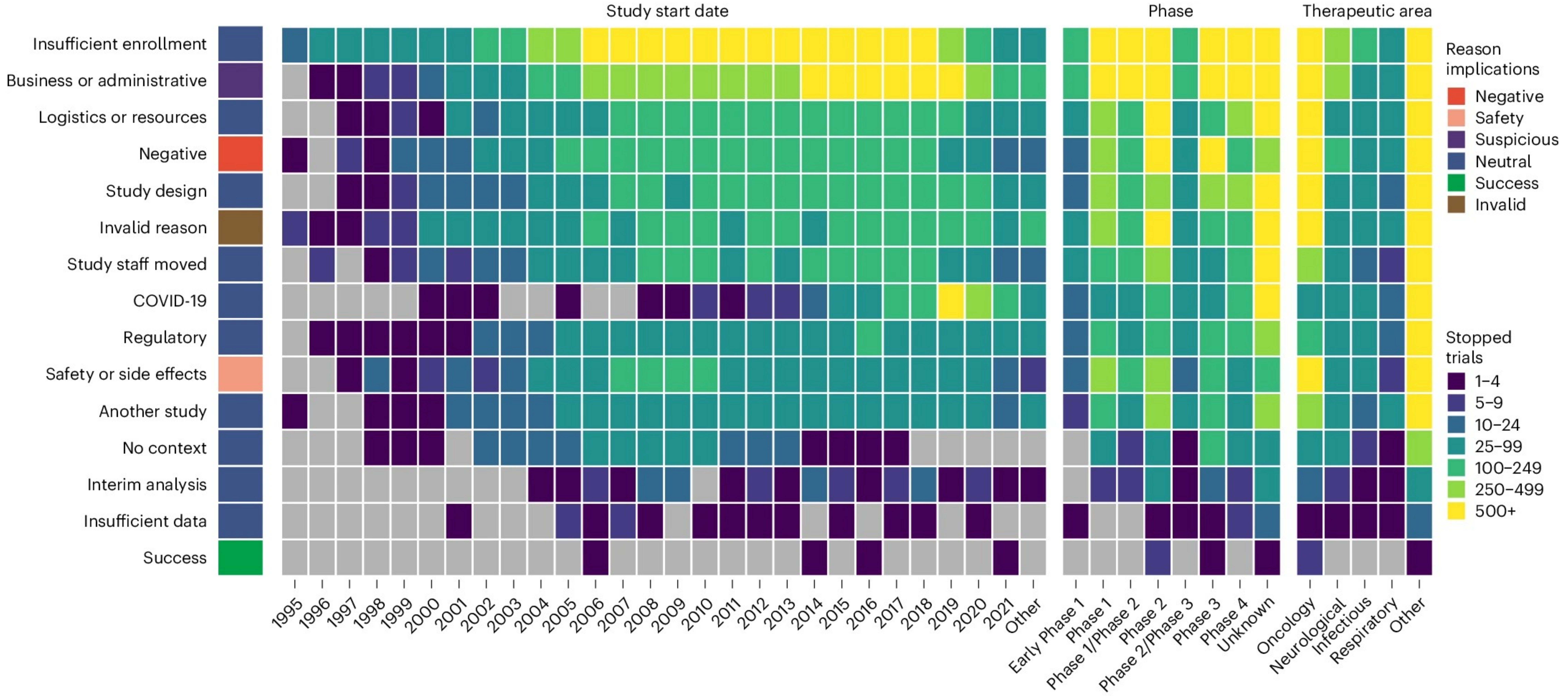
"I Remember Everything"

EHR, Real-World Evidence, & Epidemiology

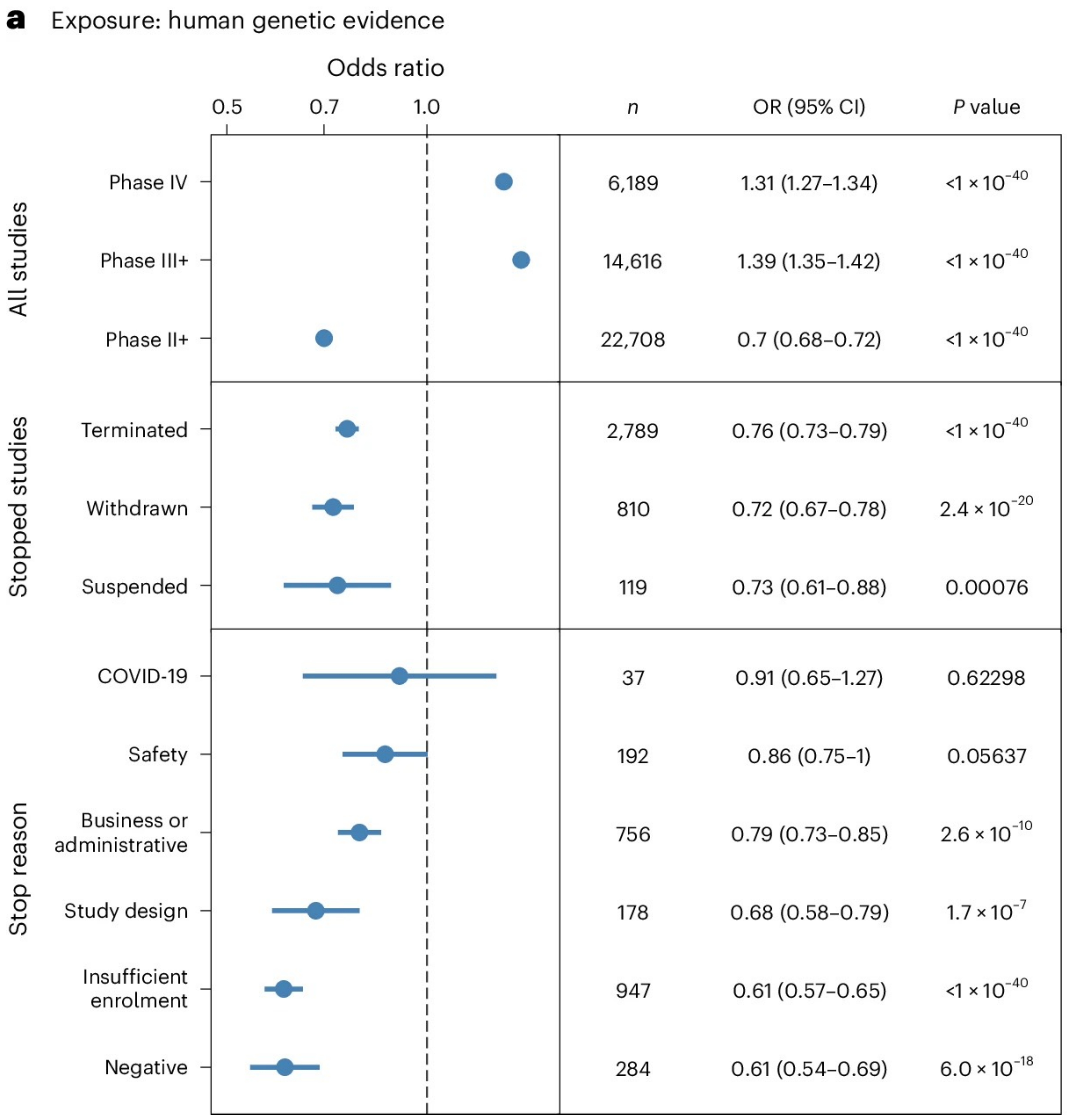
Genetic factors associated with reasons for clinical trial stoppage (Razuvayevskaya et al, *Nature Genetics*)

- Goal: Investigate if genetic factors play a role in clinical trial failures
- Method:
  - Fine-tuned a BERT model to analyze free-text reasons for stoppage in 28k clinical trials (CT.gov)
  - Link trials to target genetics
  - Stratify trials by the amount of genetic evidence
- Result:
  - Trials with weak genetic evidence are more likely to stop due to lack of efficacy (OR=0.61, p<0.0000…001)
  - Highly constrained genes (i.e. loss intolerant) are more likely to result in safety issues if targeted
- Conclusion: SOTR NLP/LLMs continue to make unstructured data available for research!

Reasons for stopping trials



Studies that have stopped are less likely to have genetic evidence

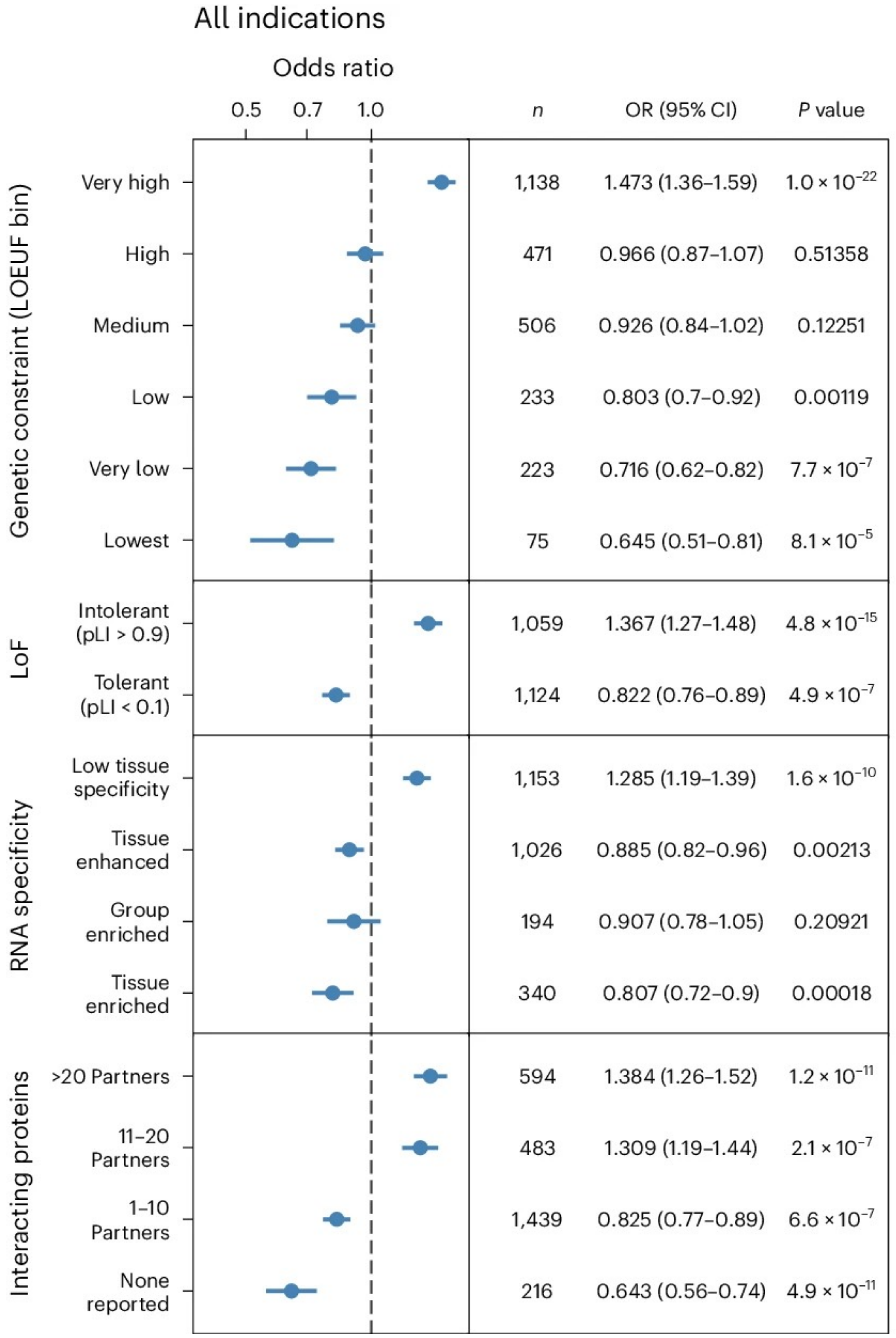


		n	OR (95% CI)	*P* value
All studies	Phase IV	6,189	1.31 (1.27–1.34)	$<1 \times 10^{-40}$
	Phase III+	14,616	1.39 (1.35–1.42)	$<1 \times 10^{-40}$
	Phase II+	22,708	0.7 (0.68–0.72)	$<1 \times 10^{-40}$
Stopped studies	Terminated	2,789	0.76 (0.73–0.79)	$<1 \times 10^{-40}$
	Withdrawn	810	0.72 (0.67–0.78)	2.4×10^{-20}
	Suspended	119	0.73 (0.61–0.88)	0.00076
Stop reason	COVID-19	37	0.91 (0.65–1.27)	0.62298
	Safety	192	0.86 (0.75–1)	0.05637
	Business or administrative	756	0.79 (0.73–0.85)	2.6×10^{-10}
	Study design	178	0.68 (0.58–0.79)	1.7×10^{-7}
	Insufficient enrolment	947	0.61 (0.57–0.65)	$<1 \times 10^{-40}$
	Negative	284	0.61 (0.54–0.69)	6.0×10^{-18}

If the Target is constrained, more side effects

Low tissue-specific —> more side effects

More interacting partners —> more side effects



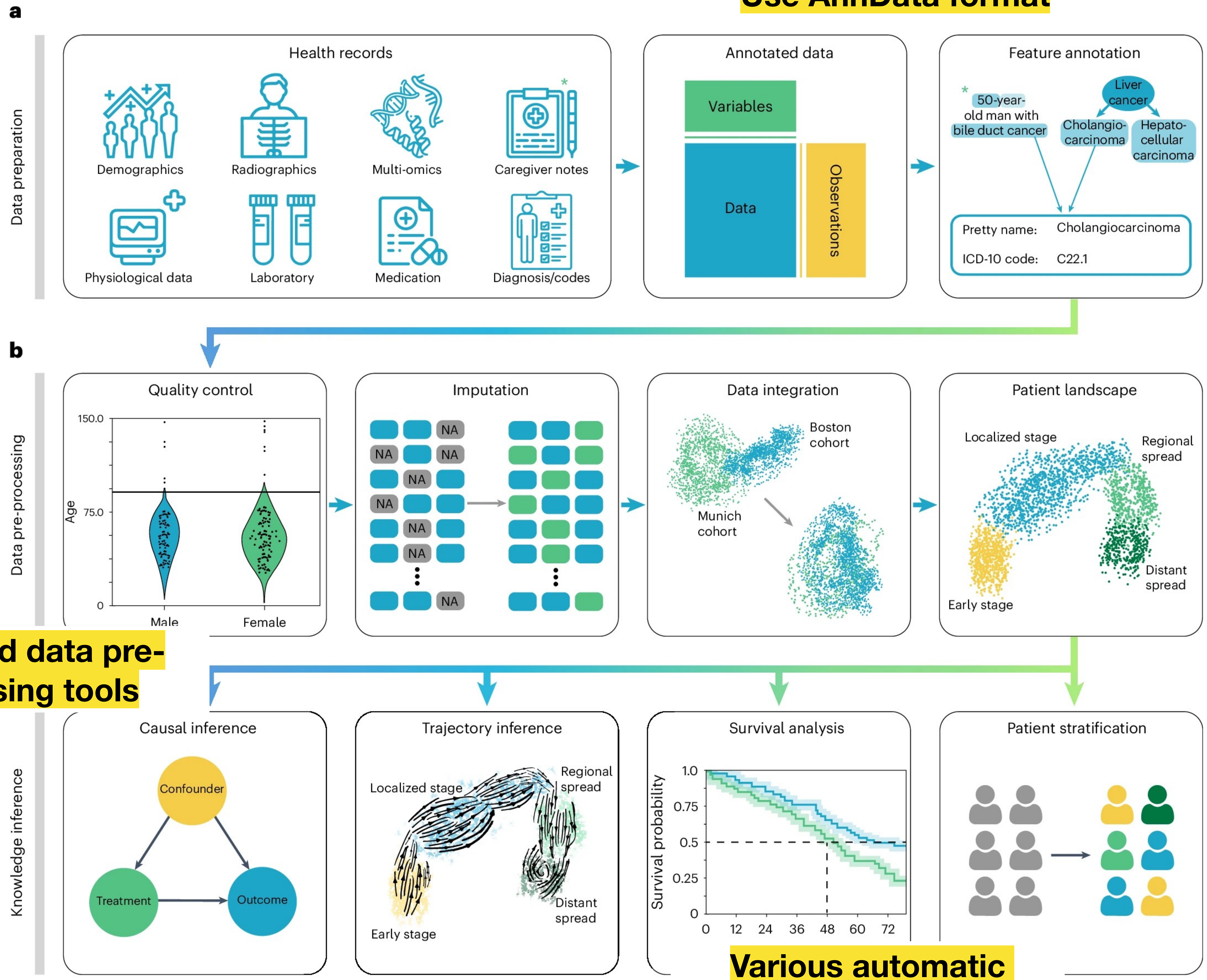
		n	OR (95% CI)	*P* value
Genetic constraint (LOEUF bin)	Very high	1,138	1.473 (1.36–1.59)	1.0×10^{-22}
	High	471	0.966 (0.87–1.07)	0.51358
	Medium	506	0.926 (0.84–1.02)	0.12251
	Low	233	0.803 (0.7–0.92)	0.00119
	Very low	223	0.716 (0.62–0.82)	7.7×10^{-7}
	Lowest	75	0.645 (0.51–0.81)	8.1×10^{-5}
LoF	Intolerant (pLI > 0.9)	1,059	1.367 (1.27–1.48)	4.8×10^{-15}
	Tolerant (pLI < 0.1)	1,124	0.822 (0.76–0.89)	4.9×10^{-7}
RNA specificity	Low tissue specificity	1,153	1.285 (1.19–1.39)	1.6×10^{-10}
	Tissue enhanced	1,026	0.885 (0.82–0.96)	0.00213
	Group enriched	194	0.907 (0.78–1.05)	0.20921
	Tissue enriched	340	0.807 (0.72–0.9)	0.00018
Interacting proteins	>20 Partners	594	1.384 (1.26–1.52)	1.2×10^{-11}
	11–20 Partners	483	1.309 (1.19–1.44)	2.1×10^{-7}
	1–10 Partners	1,439	0.825 (0.77–0.89)	6.6×10^{-7}
	None reported	216	0.643 (0.56–0.74)	4.9×10^{-11}

https://doi.org/10.1038/s41591-024-03214-0

An open-source framework for end-to-end analysis of electronic health record data (Huemos et al, *Nature Medicine*)

- Goal: Democratize EHR data analysis
- Method:
  - Develop open source modular tools based on existing data analysis standards
  - Implement a suite of commonly used analytical methods and enable multi-modal data integration
- Result:
  - Demonstrate on 4 different datasets with EHR data (e.g. UKB)
    - Phenotype stratification; biomarker discovery; causal inference; risk prediction
- Conclusion: EHRs are quickly becoming one of our most valuable research assets; the more people with access, the better

Use AnnData format

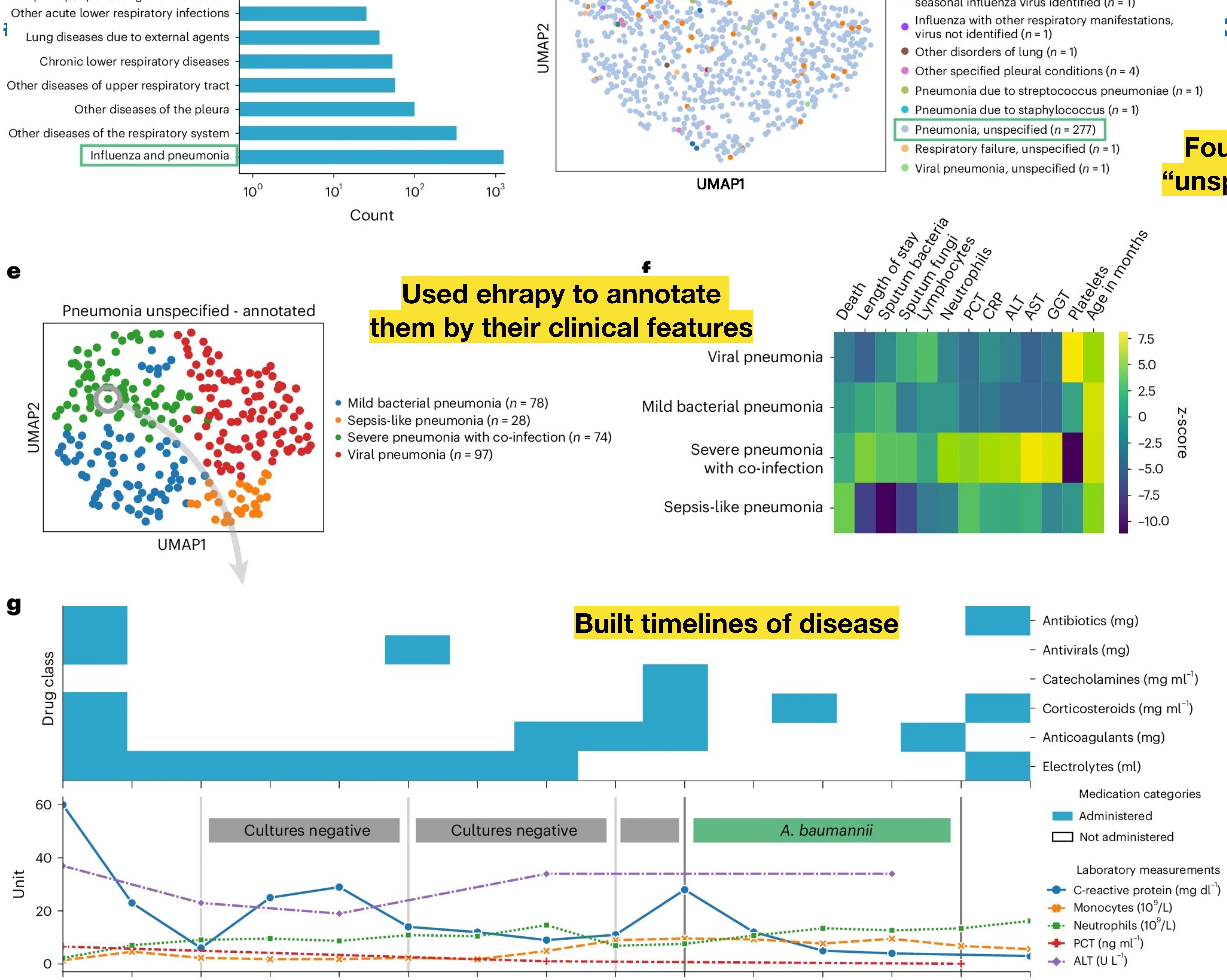


Integrated data pre-processing tools

Various automatic modeling capabilities

#IS25

38/s41591-024-03214-0

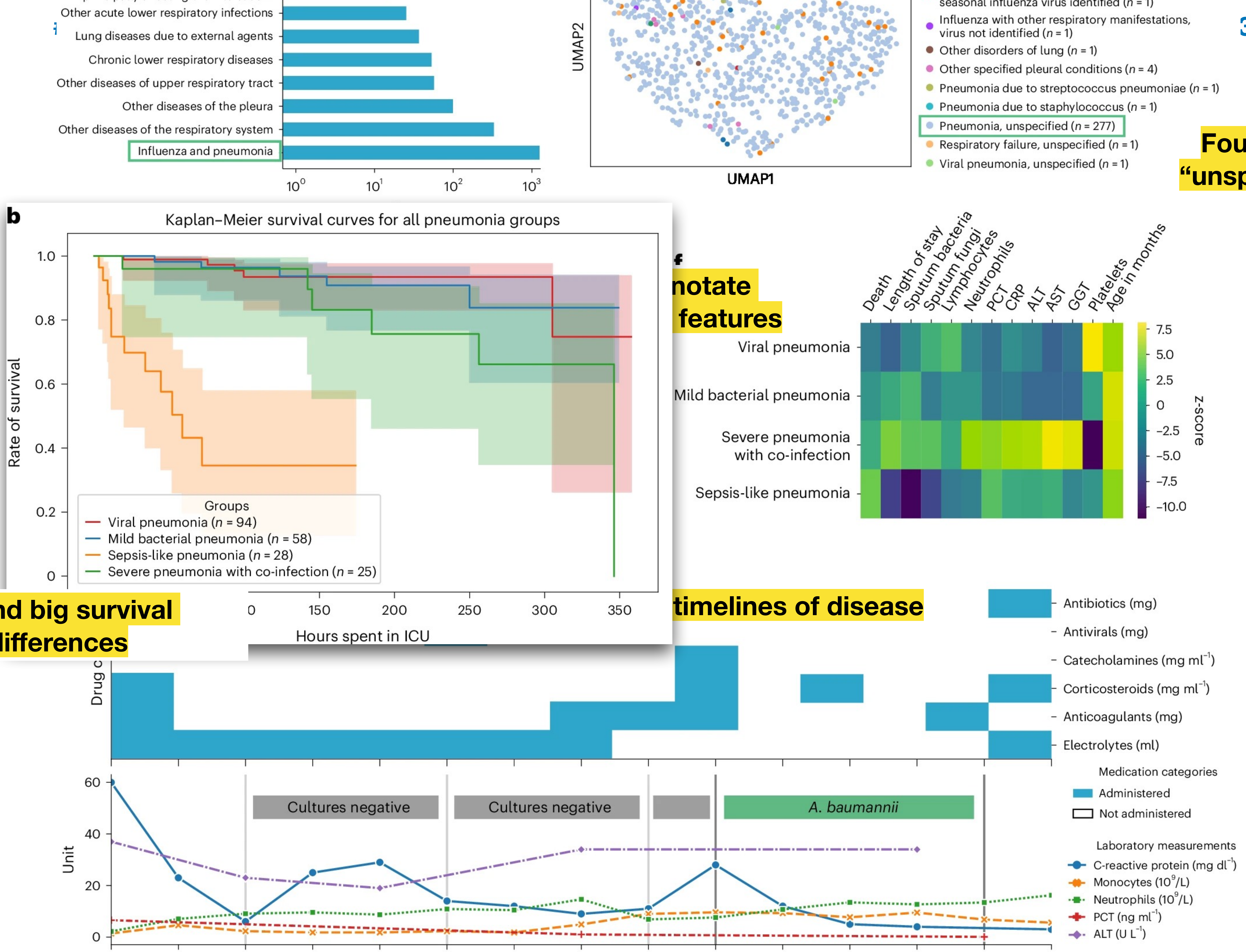


Found ICD10 code for “unspecified pneumonia”

Used ehrapy to annotate them by their clinical features

Built timelines of disease

38/s41591-024-03214-0



Found ICD10 code for “unspecified pneumonia”

notate features

Found big survival differences

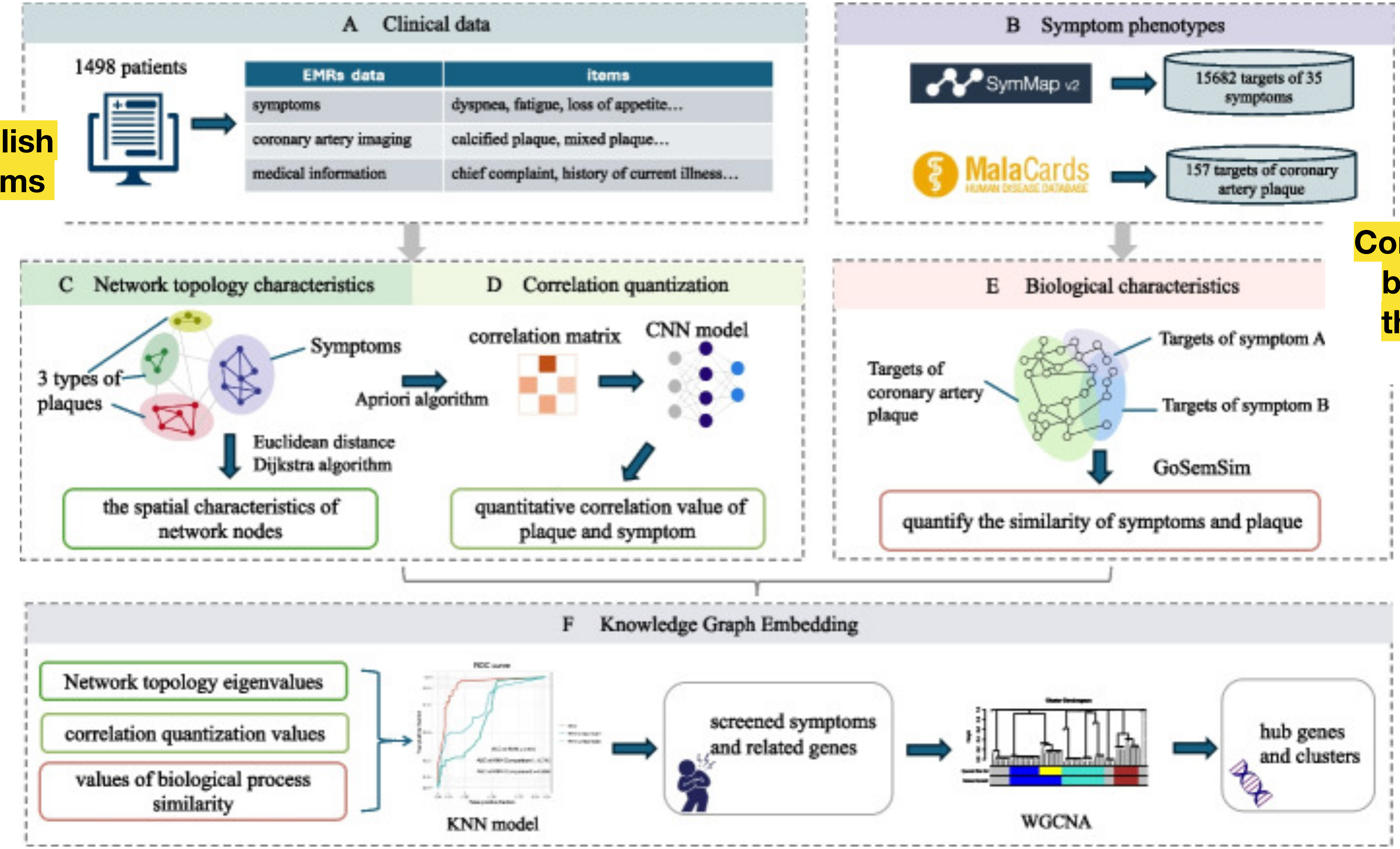
timelines of disease

The biomedical knowledge graph of symptom phenotype in coronary artery plaque: machine learning-based analysis of real-world clinical data (Huan et al, *BioData Mining*)

- Goal: Better understand the precursors of heart disease and you'll better understand heart disease
    - i.e. Study the multifactorial nature of coronary artery plaque
- Method:
    - Use ~1,500 patients EHRs to get a diverse presentation of plaques and use for training/evaluation
    - Integrate genetic and pathway knowledge (e.g. STRING) to add mechanistic layer and to identify hub genes which could explain differences
- Result:
    - Identified 23 symptom phenotypes, 41 association rules, and 61 hub genes
    - Lipid metabolism and inflammation pathways are key drivers of symptom differences
- Conclusion: The EHR is a rich resource that's underutilized. Leveraging the symptomatic variances of how disease present is a great new avenue to study human disease.

Use the EHR to establish link between symptoms and plaque

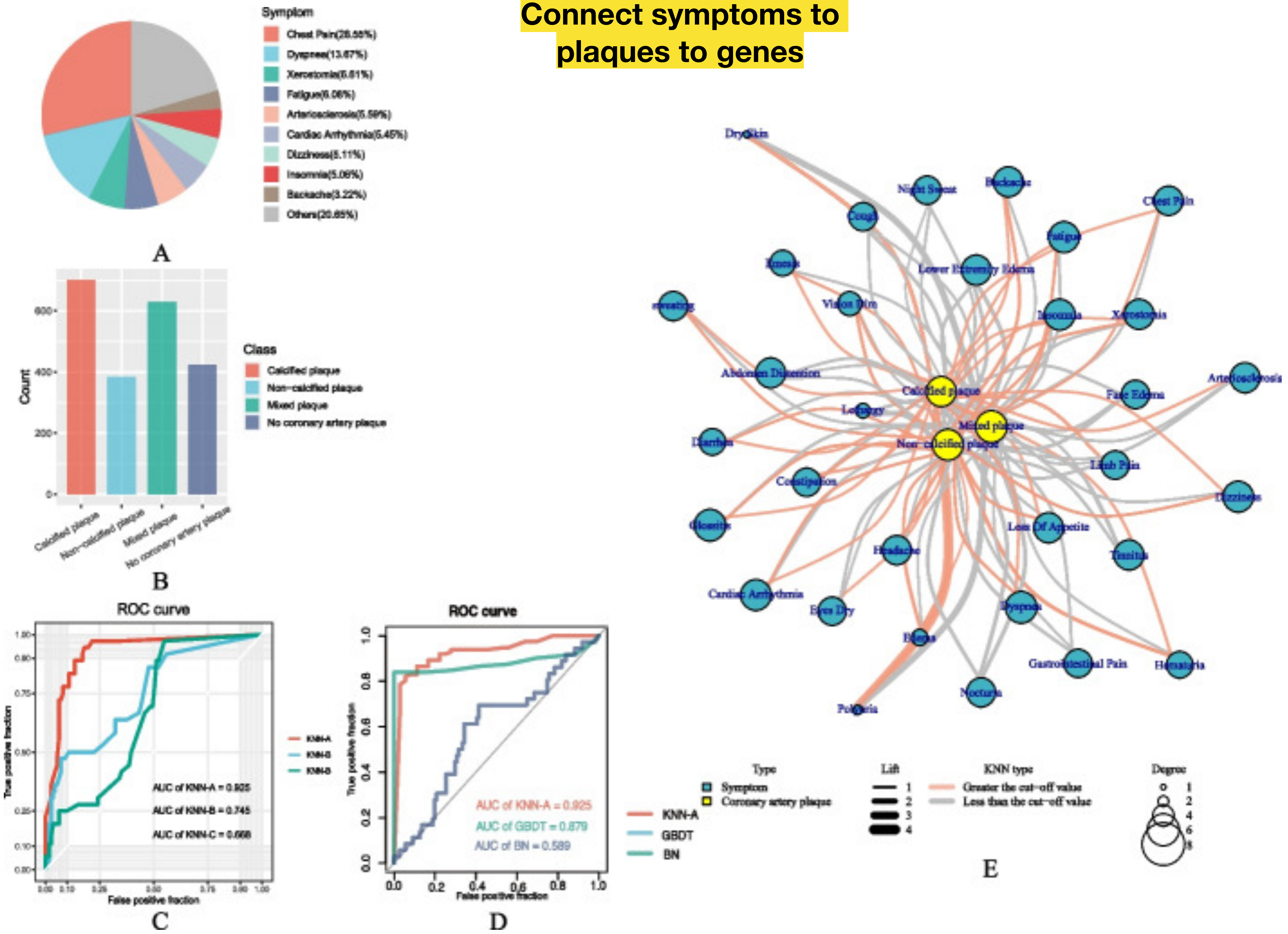
Connect symptoms to biological network through annotated targets


A Clinical data
1498 patients
B Symptom phenotypes
SymMap v2
15682 targets of 35 symptoms
MalaCards HUMAN DISEASE DATABASE
157 targets of coronary artery plaque
C Network topology characteristics
3 types of plaques
Symptoms
Euclidean distance
Dijkstra algorithm
the spatial characteristics of network nodes
D Correlation quantization
Apriori algorithm
correlation matrix
CNN model
quantitative correlation value of plaque and symptom
E Biological characteristics
Targets of coronary artery plaque
Targets of symptom A
Targets of symptom B
GoSemSim
quantify the similarity of symptoms and plaque
F Knowledge Graph Embedding
Network topology eigenvalues
correlation quantization values
values of biological process similarity
ROC curve
KNN model
screened symptoms and related genes
WGCNA
hub genes and clusters

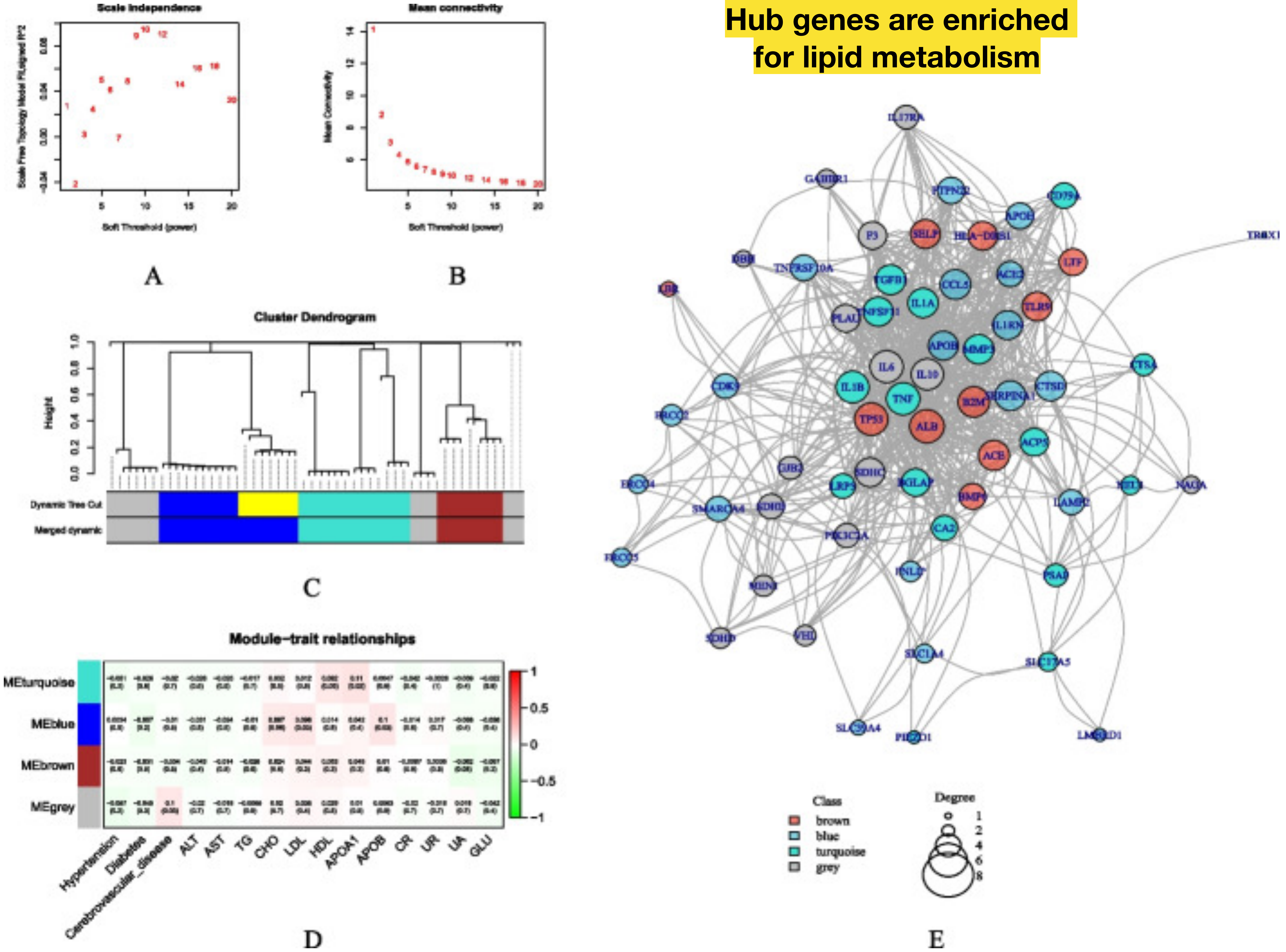
EMRs data	items
symptoms	dyspnea, fatigue, loss of appetite…
coronary artery imaging	calcified plaque, mixed plaque…
medical information	chief complaint, history of current illness…

Use classic ML approaches to build predictive models

Connect symptoms to plaques to genes



Hub genes are enriched for lipid metabolism



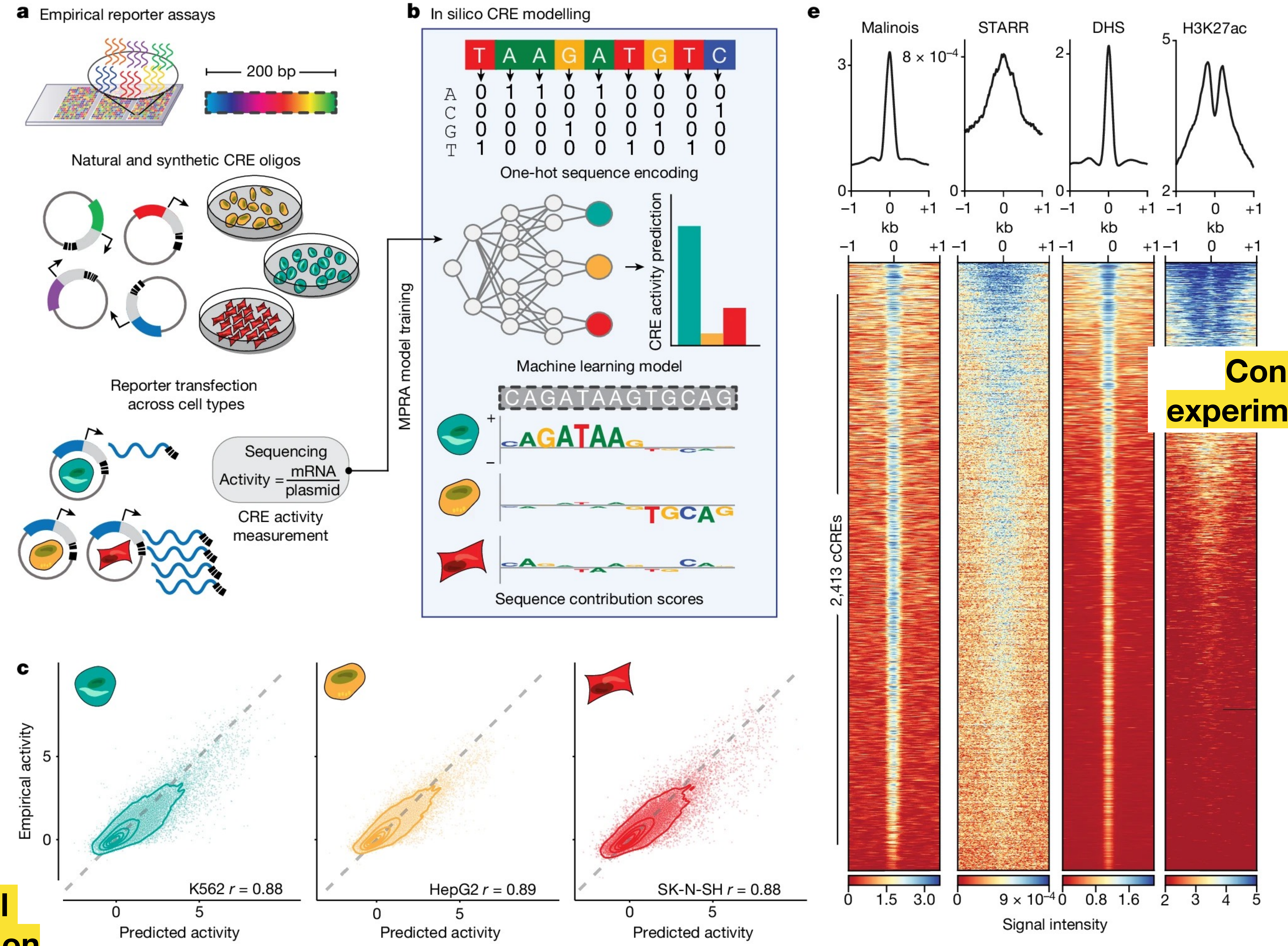
"Houdini"

Emerging Therapeutics & Technologies

Machine-guided design of cell-type-targeting cis-regulatory elements (Gosai, Castro et al, *Nature*)

- Goal: Design and validate synthetic cis-regulatory elements (CREs) that drive cell-type-specific gene expression
- Method:
    - A CNN trained on massively parallel reporter assay (MPRA) data to predict effect of a given sequence on cell-type-specific expression (~776 million sequences assayed!)
    - Since there are >2E120 possible sequences, need a strategy for selecting them — introduce CODA which is an optimization strategy
- Result:
    - Synthetic CREs outperformed natural sequences in driving cell-type-specific expression; confirmed with extensive *in vitro* testing
    - Regulatory grammar of synthetic CREs showed a distinct motif vocabulary
- Conclusion: Award for most mind-blowing paper this year. 🤯

Architecture of Manilois



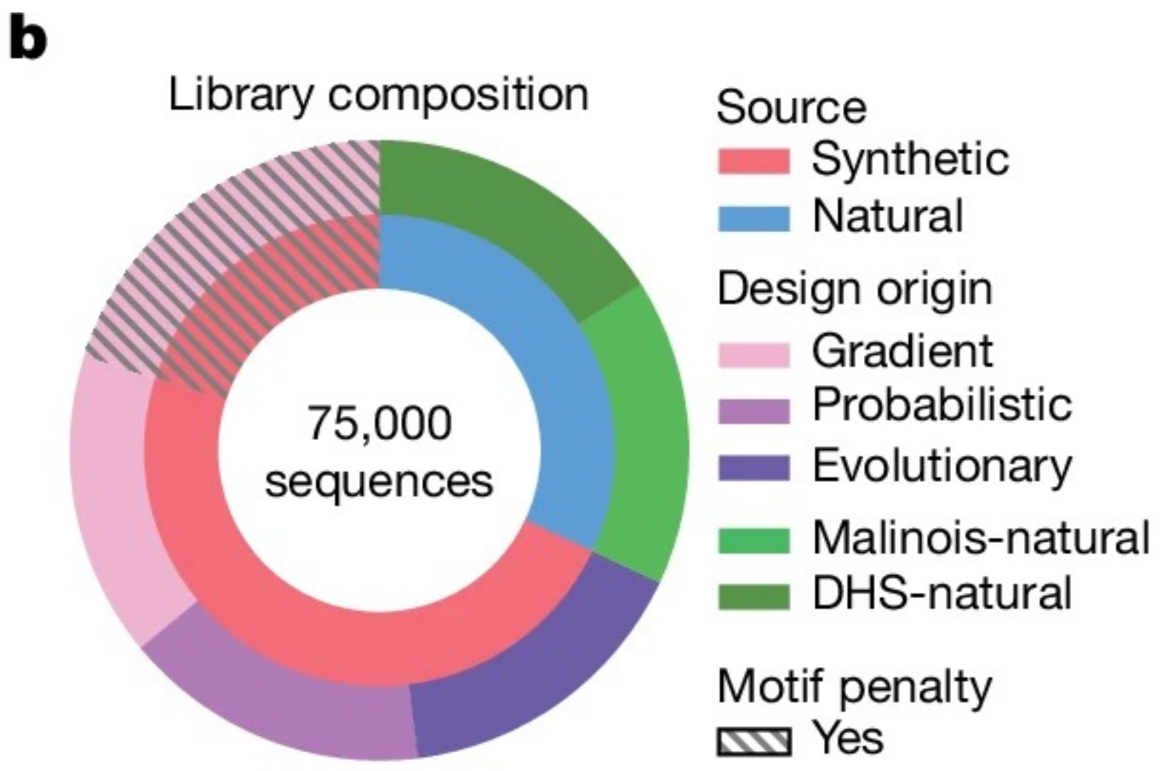
	T	A	A	G	A	T	G	T	C
A	0	1	1	0	1	0	0	0	0
C	0	0	0	0	0	0	0	0	1
G	0	0	0	1	0	0	1	0	0
T	1	0	0	0	0	1	0	1	0

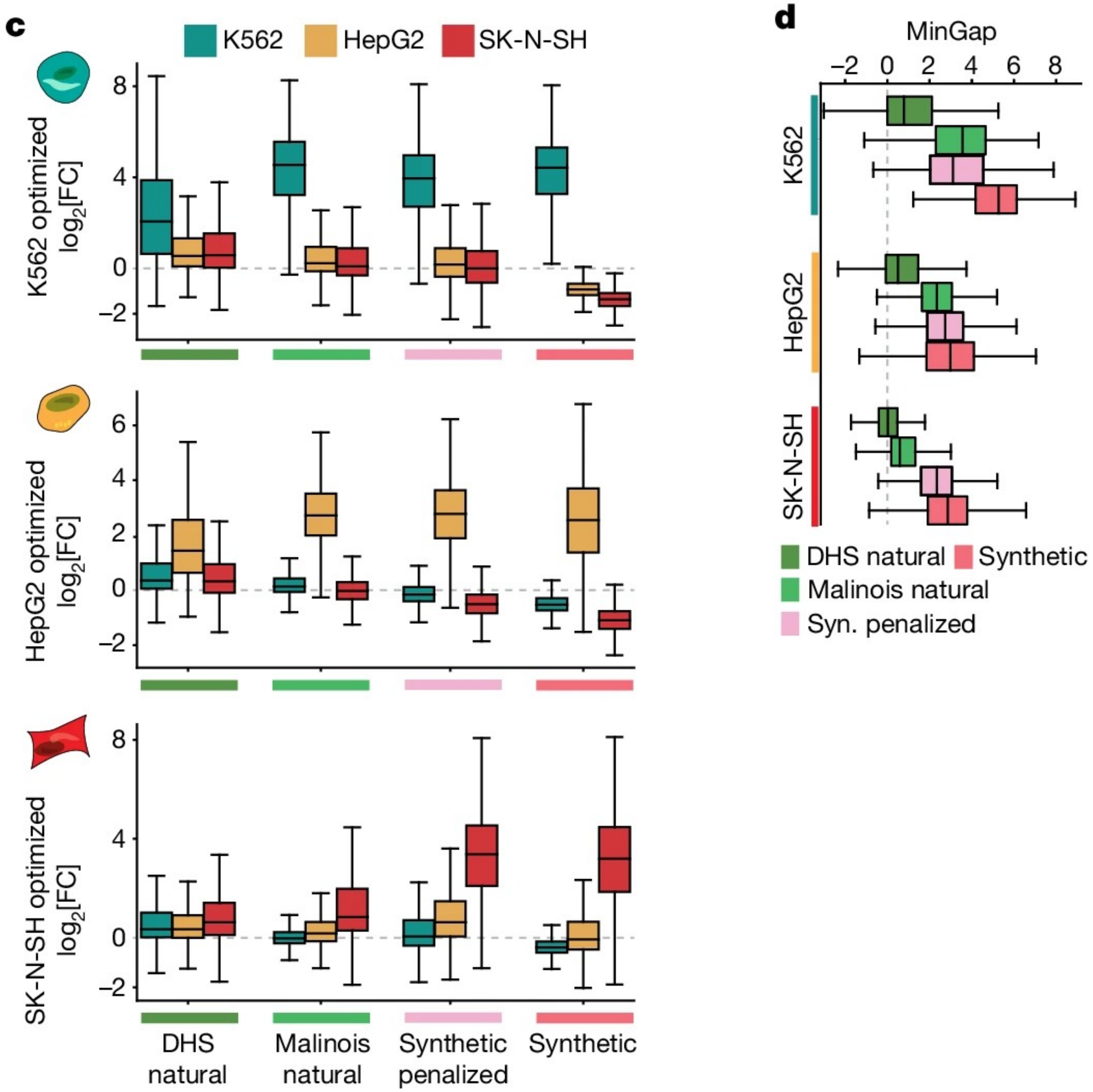
Data comes from these massively parallel reporter assays (MPRA)

Consistent with experimental measures

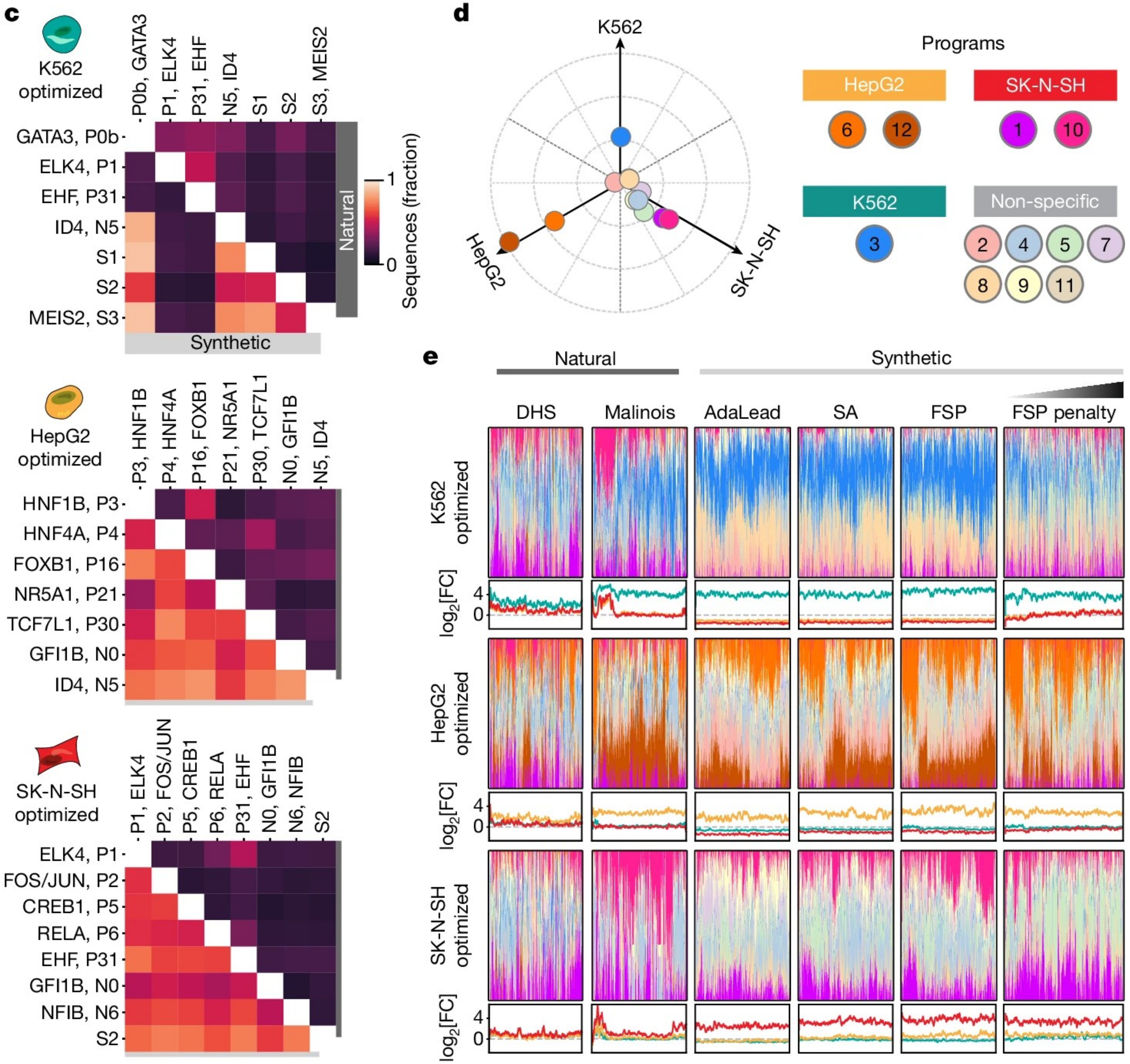
Predicted v. Actual shows high correlation

Experimental results show synthetic CREs are stronger than naturals



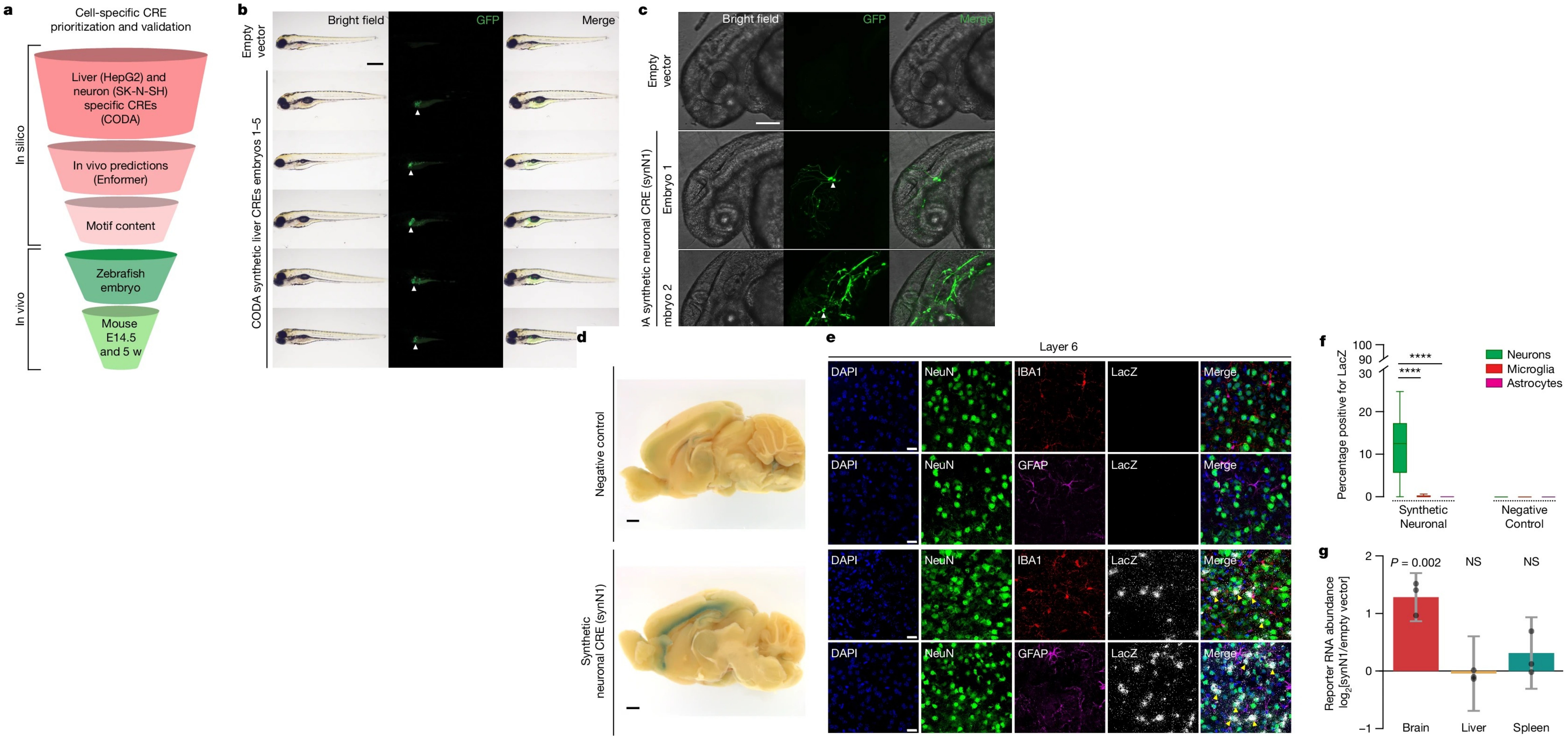


Synthetic and naturals are quite a bit different



Synthetic and naturals are quite a bit different

Show striking specificity in both zebrafish and mice



mRNA-LM: full-length integrated SLM for mRNA analysis (Li et al, *Nucleic Acids Research*)

- Goal: Develop a small language model for mRNA that includes the coding region PLUS the 5’ and 3’ untranslated regions (UTR)

- Method:

  - Use contrastive learning (CLIP) but instead of text + image, it’s 3 different mRNA sequences to construct a joint language model for all three regions

  - Pretrained BERT-based models on each of the 3, then combined using CLIP to create mRNA-LM

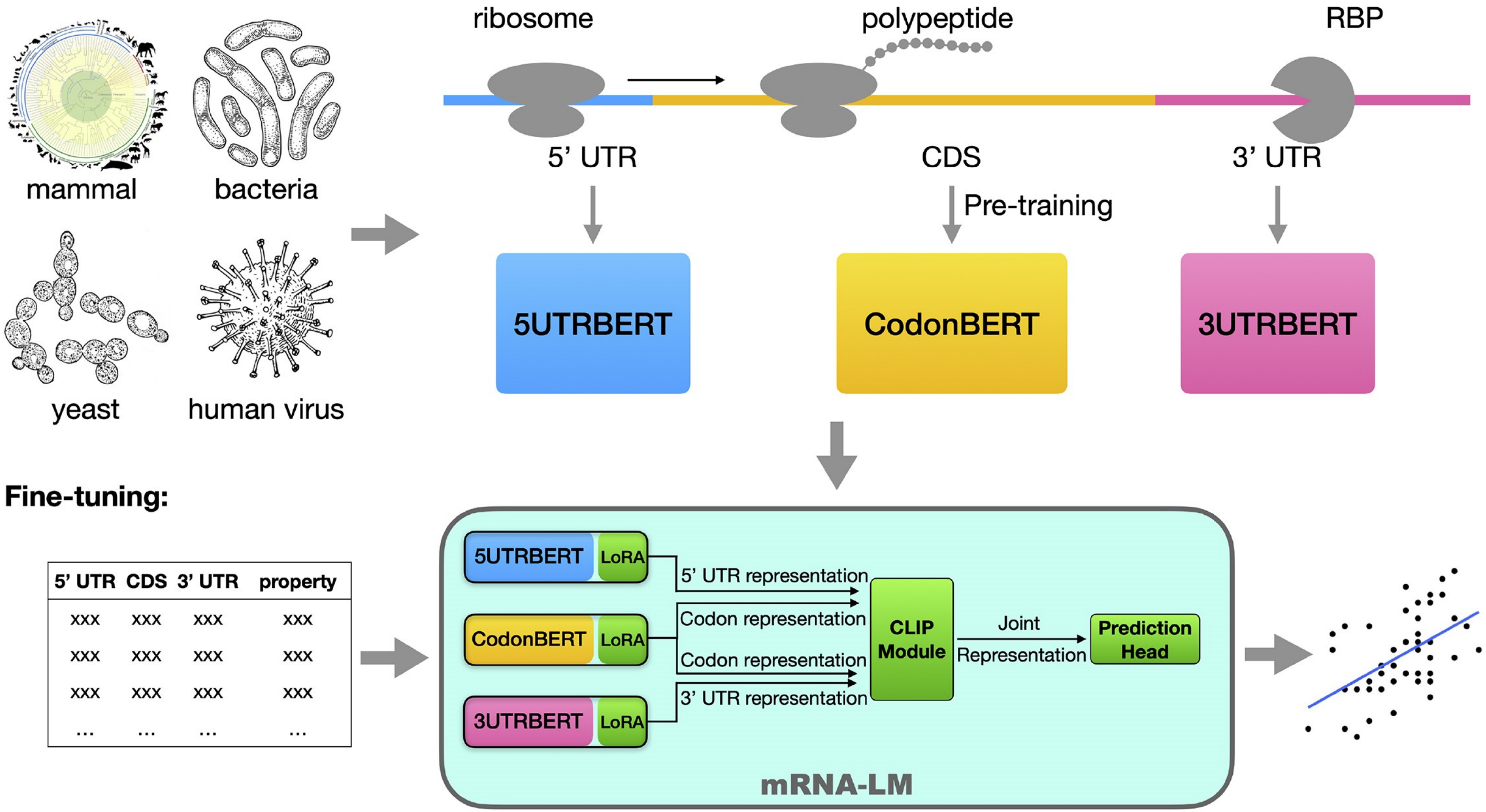
  - Fine-tuned mRNA-LM on mRNA half-life, translation rate, transcript expression, and protein expression

- Result: Significantly outperformed other methods on the trained tasks; Performed well at zero shot tasks

- Conclusion: Nice clear methods made this a joy to read. Recommended for your JCs!

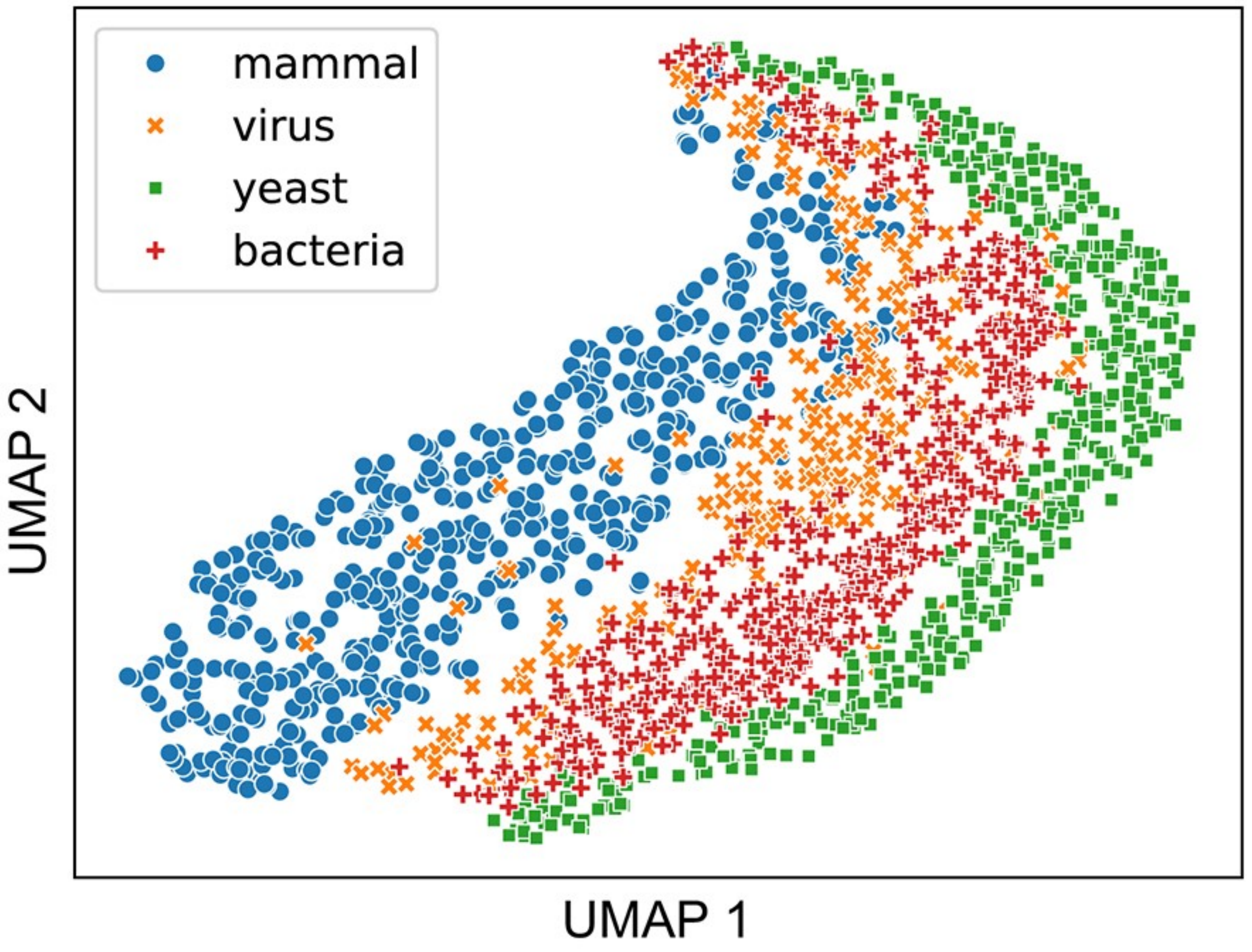
Trained on >10 million RNA sequences

3 BERT models



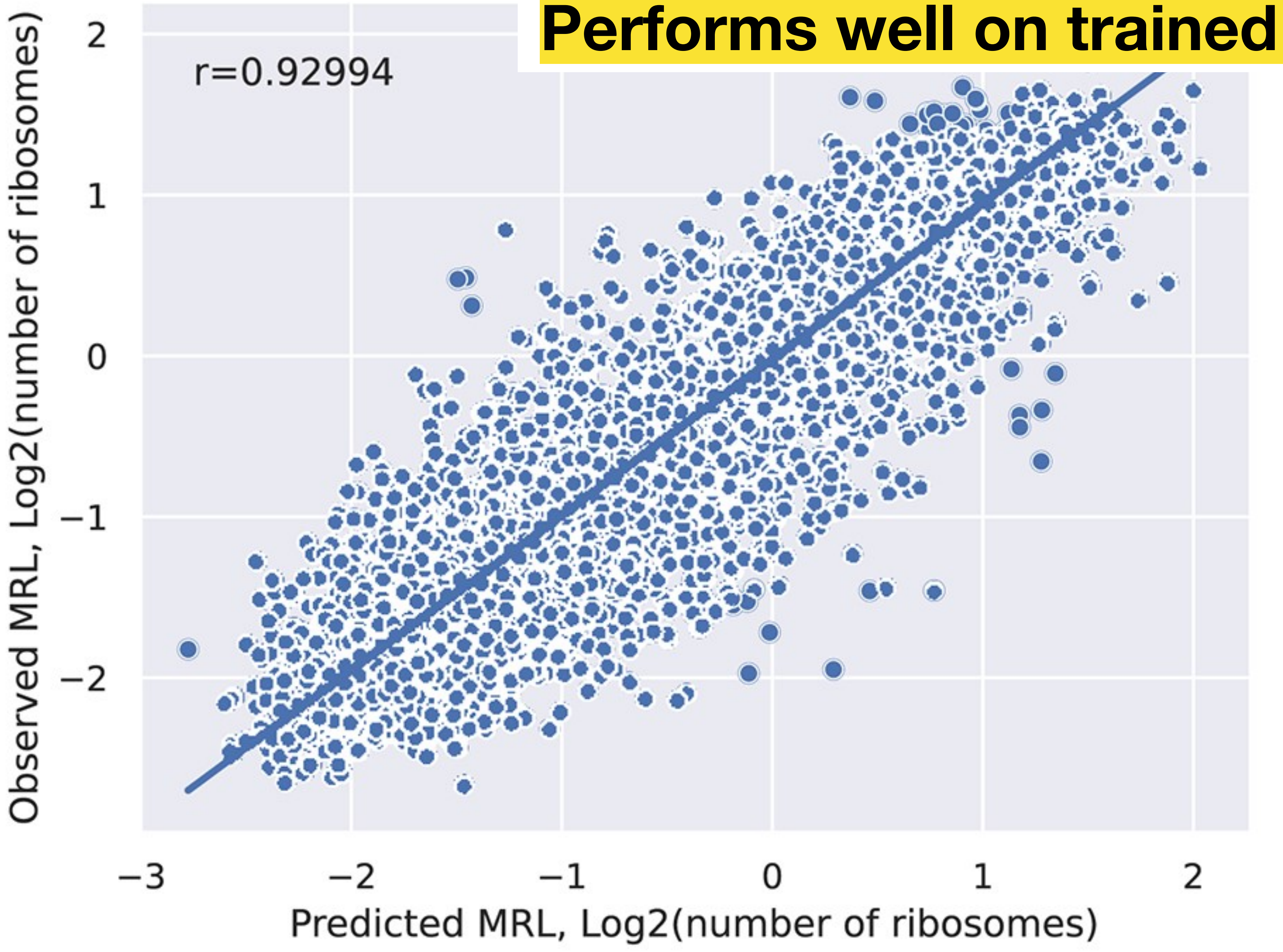
5' UTR	CDS	3' UTR	property
xxx	xxx	xxx	xxx
xxx	xxx	xxx	xxx
xxx	xxx	xxx	xxx
…	…	…	…

Combined using CLIP

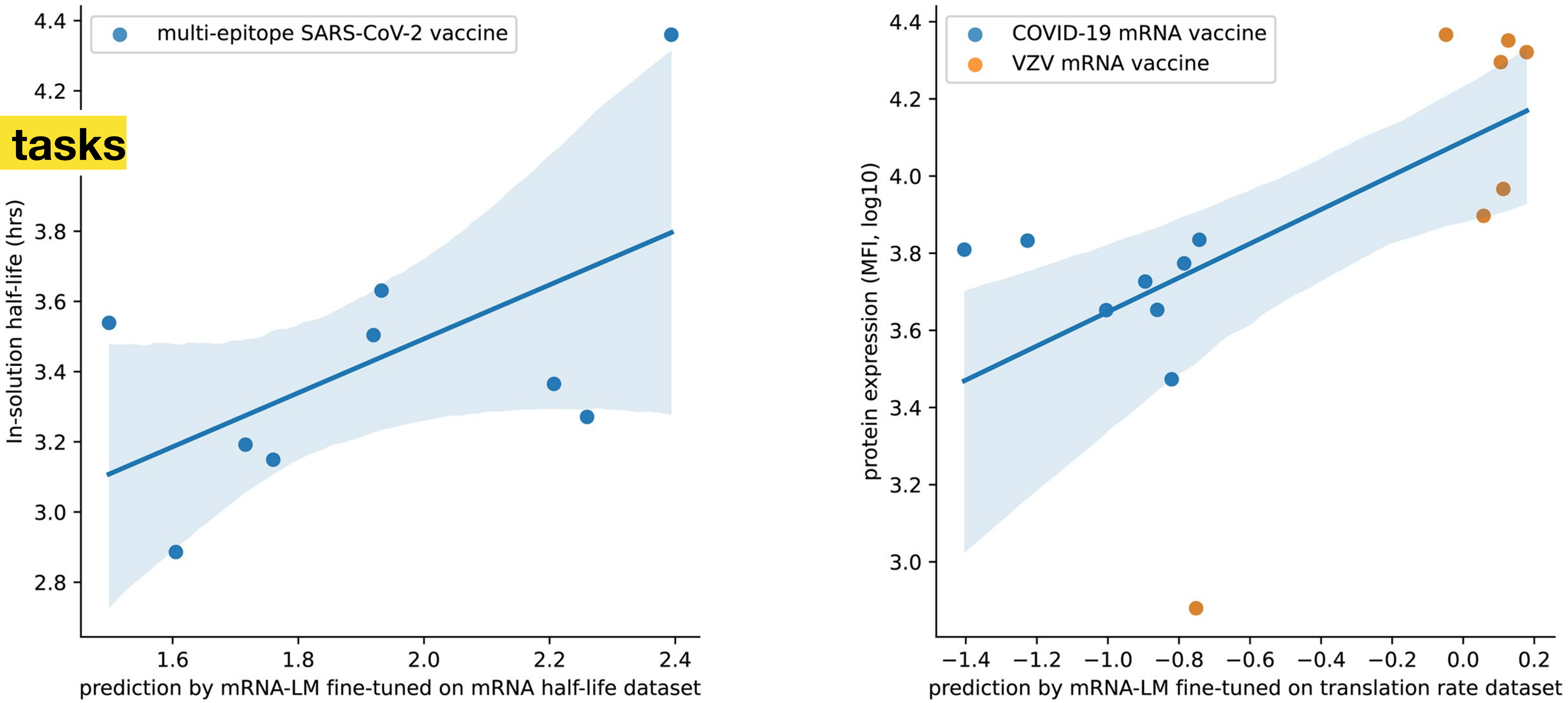


A

Performs well on trained tasks



And on zero-shot tasks



A

B

"Feather"

A Feather in Your Cap — Shout-Outs & Honorable Mentions

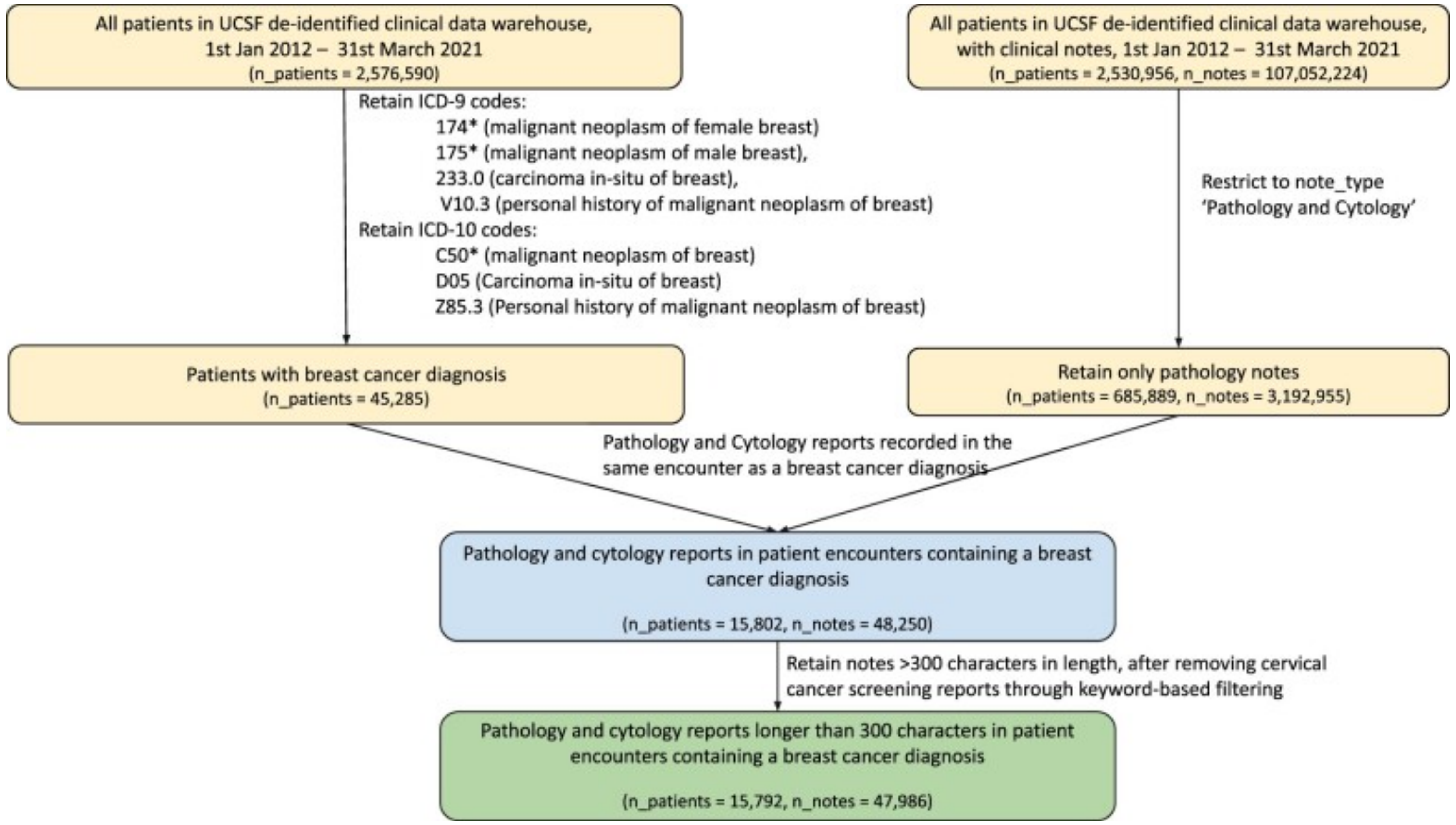
https://doi.org/10.1093/jamia/ocae146

A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports (Sushil et al, *JAMIA*)

- Goal: Compare state-of-the-art LLMs to supervised ML techniques for extracting important information from pathology reports
- Method:
  - 769 manually annotated breast cancer pathology reports from UCSF
  - LLMs (zero-shot inference): GPT-4, GPT-3.5, Starling-7B-beta, ClinicalCamel-70B
  - Supervised ML models: Random Forests, LSTM-Att, UCSF-BERT
- Result:
  - GPT-4 performed better or as well as the best supervised method, in particular on tasks with high label imbalance
  - Supervised models struggled with generalizability
- Conclusion: For information extraction, LLMs are great

#IS25 #YIR25 𝕏@proftatonetti @tatonetti.bsky.social https://doi.org/10.1093/jamia/ocae146

~15k patients and ~50k reports available



All patients in UCSF de-identified clinical data warehouse, 1st Jan 2012 – 31st March 2021 (n_patients = 2,576,590)

Retain ICD-9 codes:
174* (malignant neoplasm of female breast)
175* (malignant neoplasm of male breast),
233.0 (carcinoma in-situ of breast),
V10.3 (personal history of malignant neoplasm of breast)
Retain ICD-10 codes:
C50* (malignant neoplasm of breast)
D05 (Carcinoma in-situ of breast)
Z85.3 (Personal history of malignant neoplasm of breast)

Patients with breast cancer diagnosis (n_patients = 45,285)

All patients in UCSF de-identified clinical data warehouse, with clinical notes, 1st Jan 2012 – 31st March 2021 (n_patients = 2,530,956, n_notes = 107,052,224)

Restrict to note_type 'Pathology and Cytology'

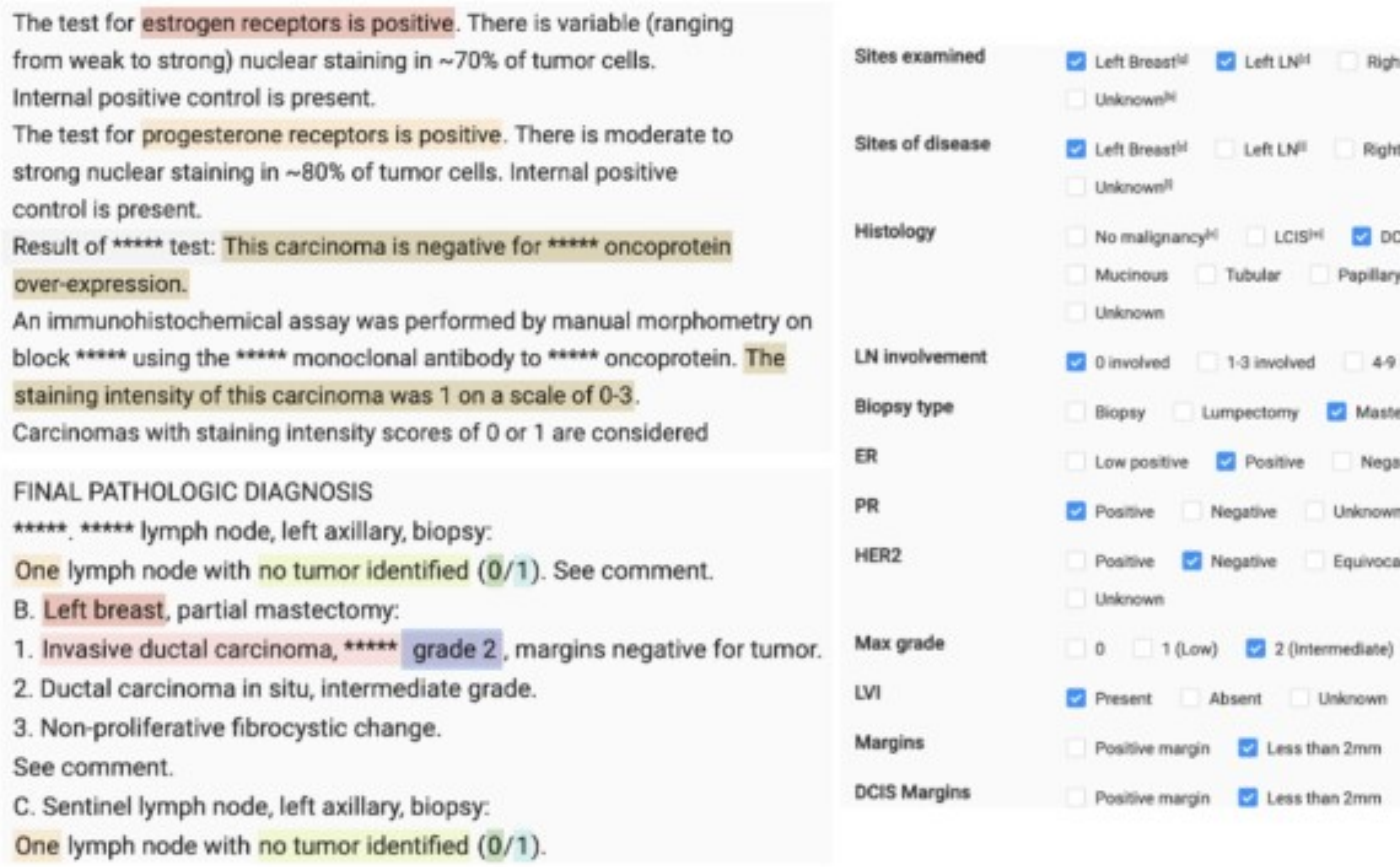
Retain only pathology notes (n_patients = 685,889, n_notes = 3,192,955)

Pathology and Cytology reports recorded in the same encounter as a breast cancer diagnosis

Pathology and cytology reports in patient encounters containing a breast cancer diagnosis (n_patients = 15,802, n_notes = 48,250)

Retain notes >300 characters in length, after removing cervical cancer screening reports through keyword-based filtering

Pathology and cytology reports longer than 300 characters in patient encounters containing a breast cancer diagnosis (n_patients = 15,792, n_notes = 47,986)



The test for estrogen receptors is positive. There is variable (ranging from weak to strong) nuclear staining in ~70% of tumor cells. Internal positive control is present.
The test for progesterone receptors is positive. There is moderate to strong nuclear staining in ~80% of tumor cells. Internal positive control is present.
Result of ***** test: This carcinoma is negative for ***** oncoprotein over-expression.
An immunohistochemical assay was performed by manual morphometry on block ***** using the ***** monoclonal antibody to ***** oncoprotein. The staining intensity of this carcinoma was 1 on a scale of 0-3.
Carcinomas with staining intensity scores of 0 or 1 are considered

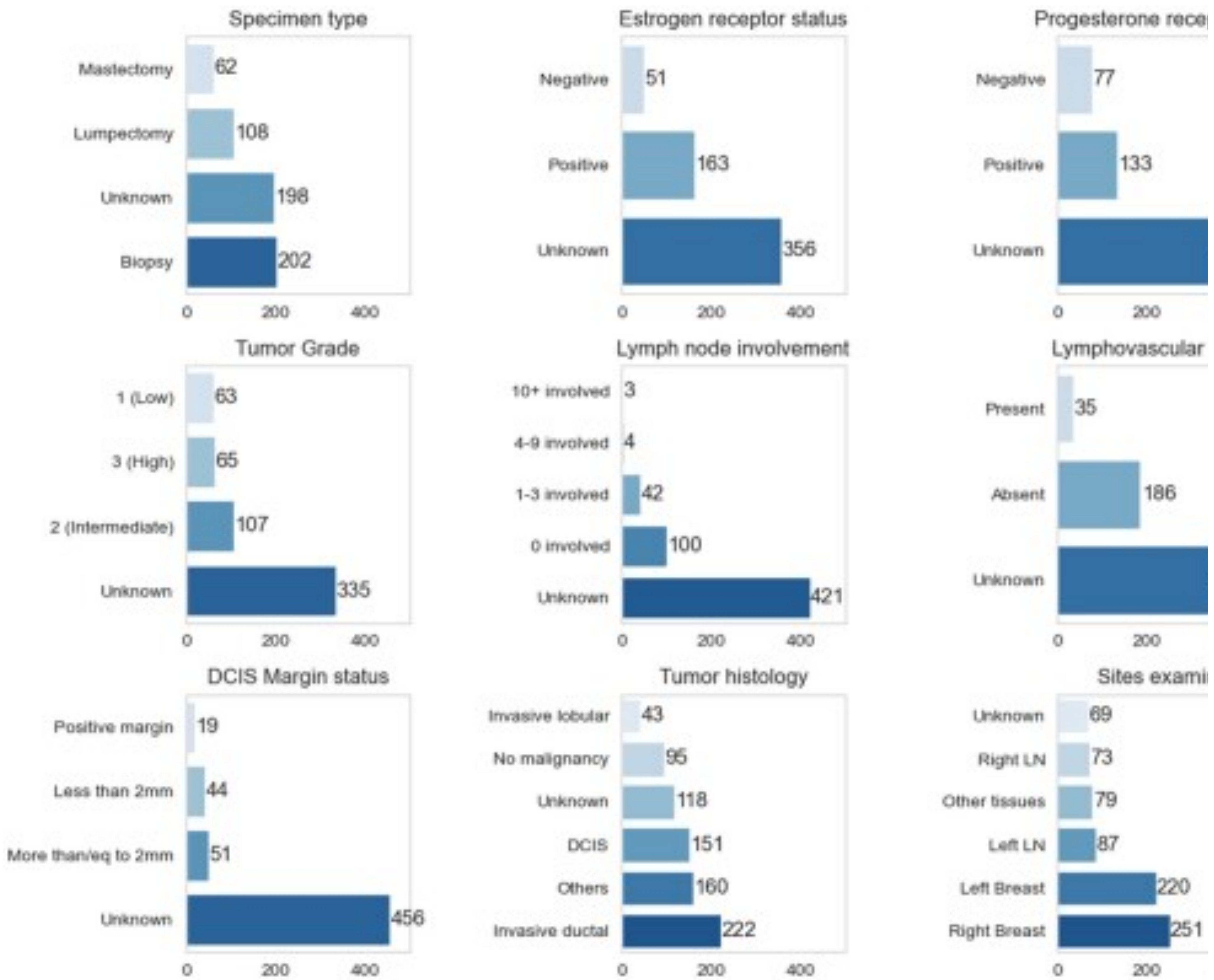
FINAL PATHOLOGIC DIAGNOSIS
*****. ***** lymph node, left axillary, biopsy:
One lymph node with no tumor identified (0/1). See comment.
B. Left breast, partial mastectomy:
1. Invasive ductal carcinoma, ***** grade 2, margins negative for tumor.
2. Ductal carcinoma in situ, intermediate grade.
3. Non-proliferative fibrocystic change.
See comment.
C. Sentinel lymph node, left axillary, biopsy:
One lymph node with no tumor identified (0/1).

Sites examined	☑ Left Breast ☑ Left LN □ Righ □ Unknown
Sites of disease	☑ Left Breast □ Left LN □ Right □ Unknown
Histology	□ No malignancy □ LCIS ☑ DC □ Mucinous □ Tubular □ Papillary □ Unknown
LN involvement	☑ 0 involved □ 1-3 involved □ 4-9
Biopsy type	□ Biopsy □ Lumpectomy ☑ Maste
ER	□ Low positive ☑ Positive □ Nega
PR	☑ Positive □ Negative □ Unknown
HER2	□ Positive ☑ Negative □ Equivoca □ Unknown
Max grade	□ 0 □ 1 (Low) ☑ 2 (Intermediate)
LVI	☑ Present □ Absent □ Unknown
Margins	□ Positive margin ☑ Less than 2mm
DCIS Margins	□ Positive margin ☑ Less than 2mm

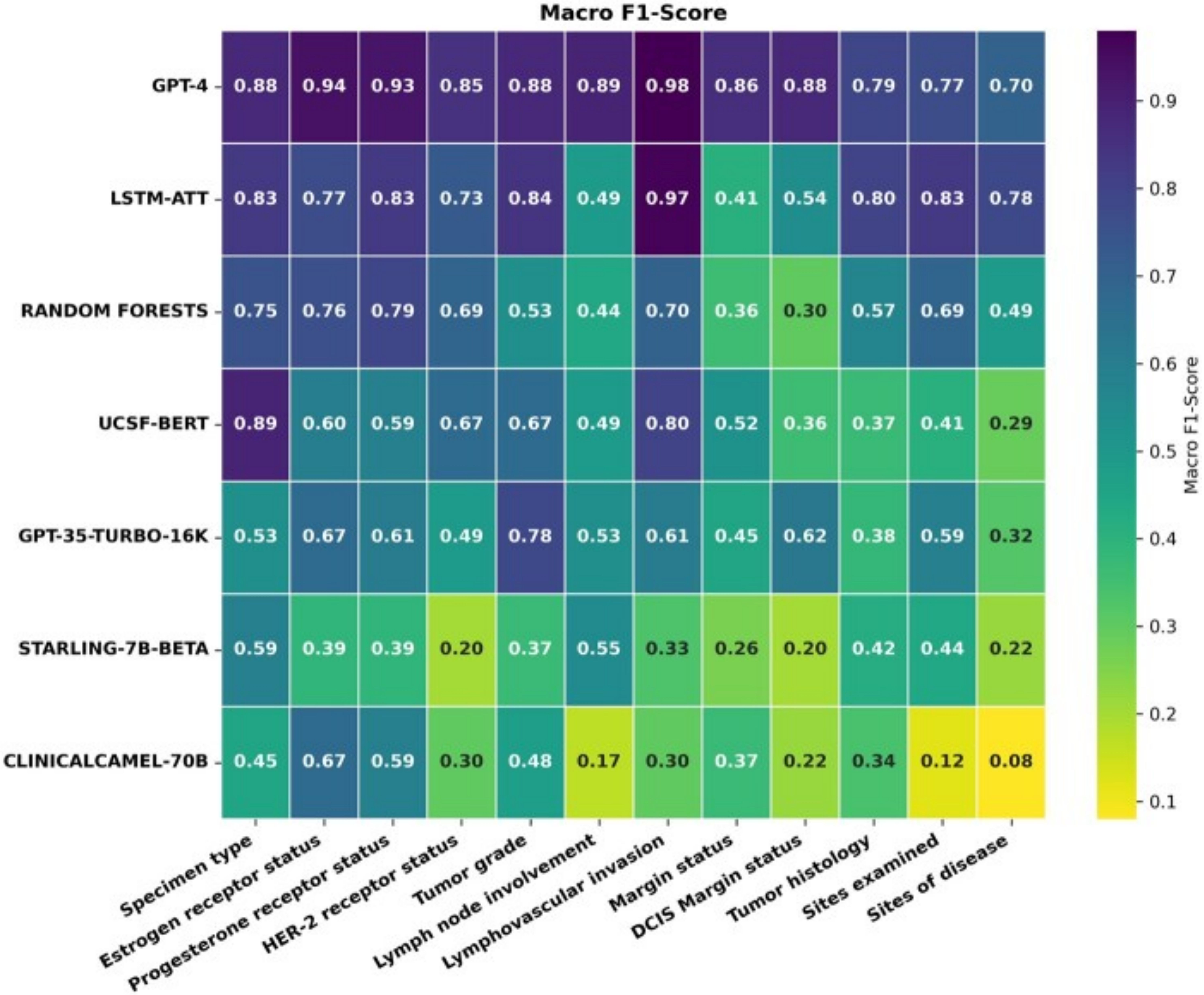
Example of a report annotation

https://doi.org/10.1093/jamia/ocae146

Label distributions



Performances

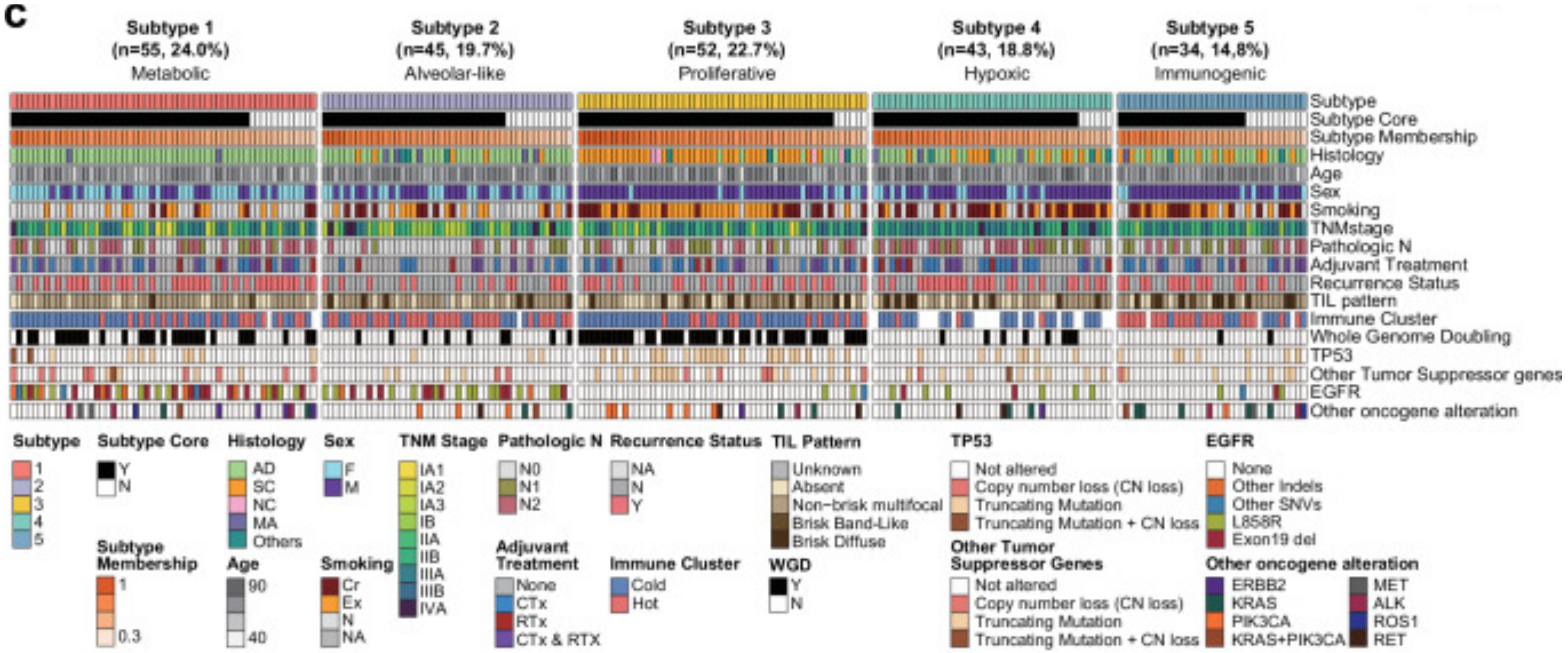


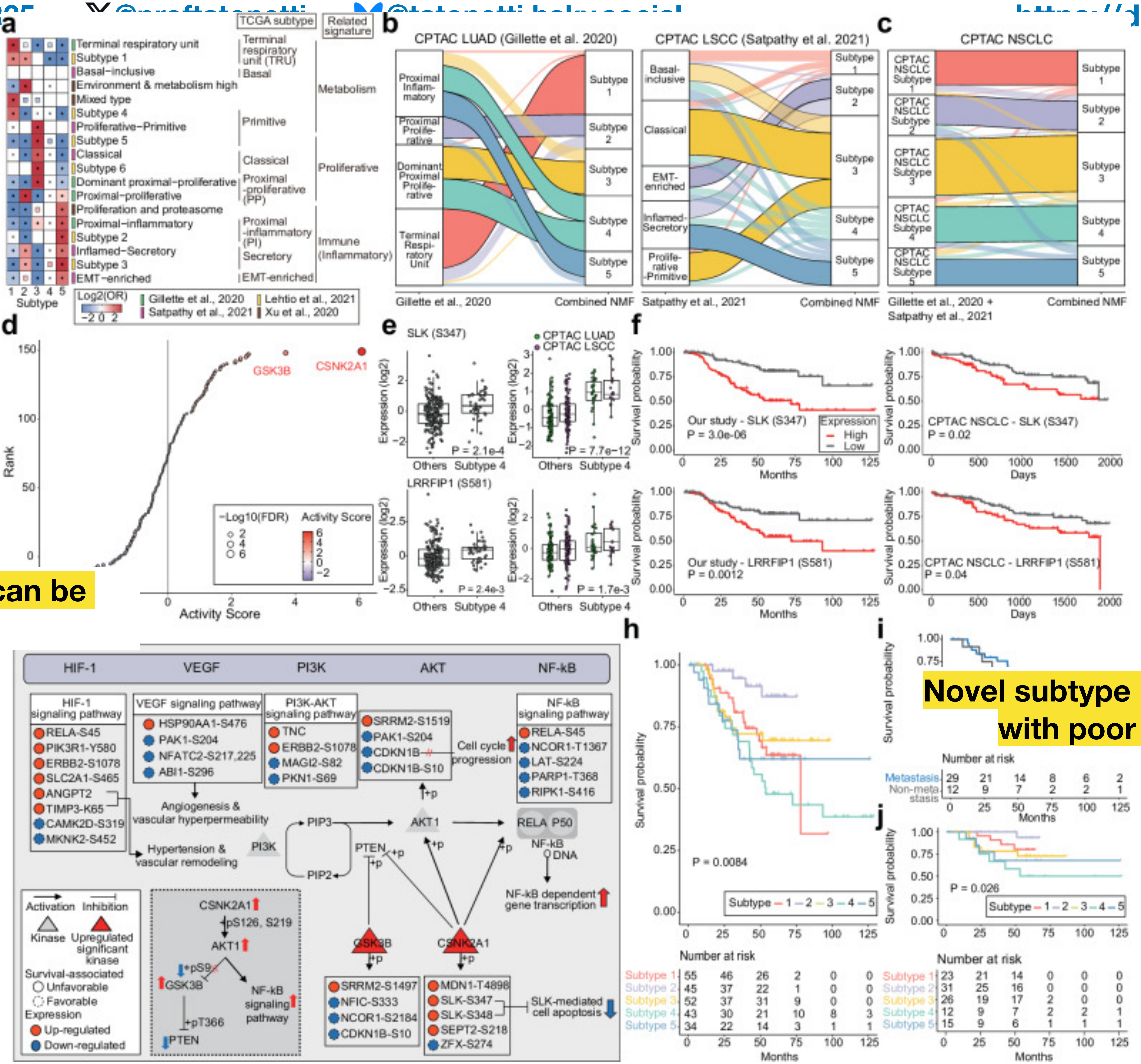
Proteogenomic analysis reveals non-small cell lung cancer subtypes predicting chromosome instability, and tumor microenvironment (Song, Choi, Kim, Hwang et al, *Journal*)

- Goal: To redefine NSCLC subtypes based on **proteogenomic and multiomics analysis**; and link these subtypes to clinical outcomes
- Method:
  - Multiomics integration of genomic (WES), transcriptomic (RNA-seq), proteomic (TMT-labeling), phosphoproteomic, and acetylproteomic data
  - Non-negative matrix factorization (NMF) clustering to identify molecular subtypes
    - Good for heterogenous data distribution, scales, and sparseness; allows for “soft-clustering”
- Result:
  - Found 5 molecular subtypes
  - 1 of which is new (hypoxic subtype) and associated with poor outcomes
- Conclusion: Right method with the right data has the right result.

https://doi.org/10.1038/s41467-024-54434-4

I’m a sucker for this style of a Table 1





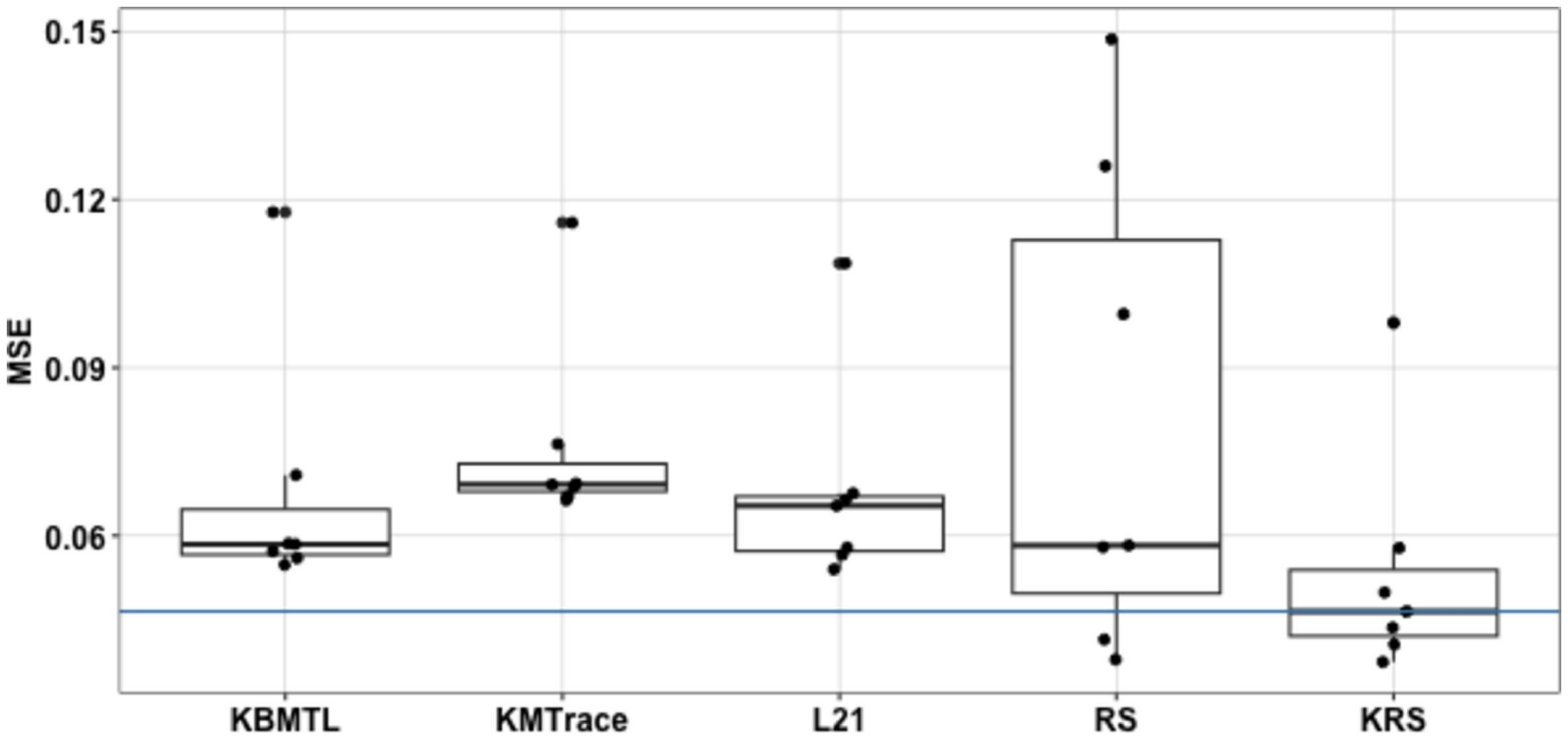
How these subtypes relate to those found by others

Driving factors can be identified

Novel subtype is associated with poor survival

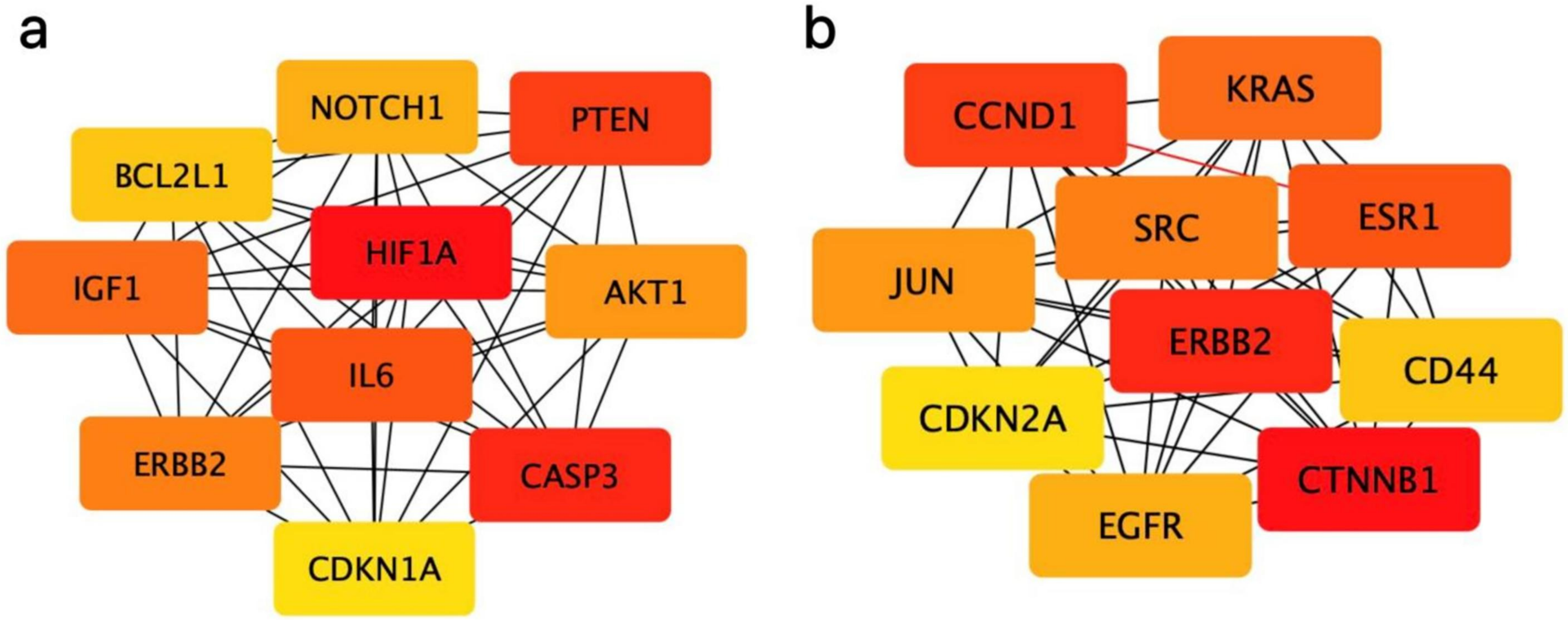
Novel machine learning model for predicting cancer drugs' susceptibilities and discovering novel treatments (Cao et al, *JBI*)

- Goal: Predict drug effect on cancer cells using genetics (a tale as old as time)
- Method:
  - Introduce Kernalized Residual Stacking (KRS)
    - Multi-task learning method (each drug's effect on the line is a task); tasks can share information through residual correction
    - Radial Basis Function (RBF) kernel reduces dimensionality
    - Allows for feature importance to be quantified to facilitate interpretability
- Result:
  - Outperforms benchmarks; Identified the PI3K-Akt pathway as a key cancer drug response regulator
  - Identify 8 novel cancer drug repurposing candidates
- Conclusion: I like the information sharing aspects of this, although wish they pushed it a bit further.



Better at predicting IC50 values

Can identify drivers of performance
(e.g. ERBB2)



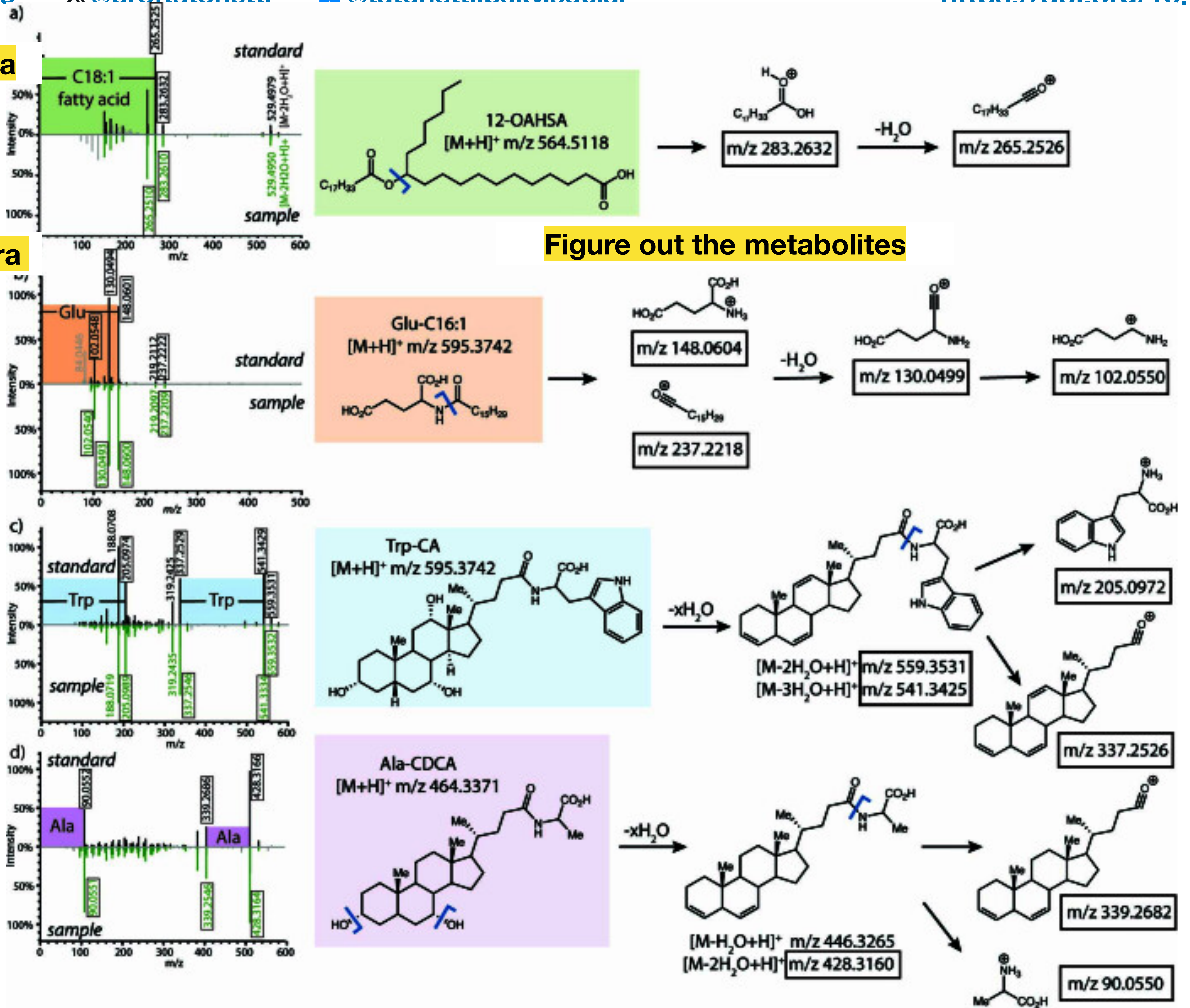
Reverse Metabolomics for Discovery of Chemical Structures from Humans (Gentry et al, *Nature*)

- Goal: Only 10% of metabolites in metabolomics studies are identified; Identify them!
- Method:
  - Generated mass spec spectra for 2,430 molecules and do a reverse database lookup on 1.2 billion metabolomics results
  - Associate the molecule with phenotypes from the metabolimcs studies
- Result:
  - Identified 139 previously unreported bile acid conjugates
  - Some were significantly associated with disease (e.g. Crohn’s disease)
- Conclusion: This is why we make our data public!

Generate new spectra

Match existing spectra

Figure out the metabolites

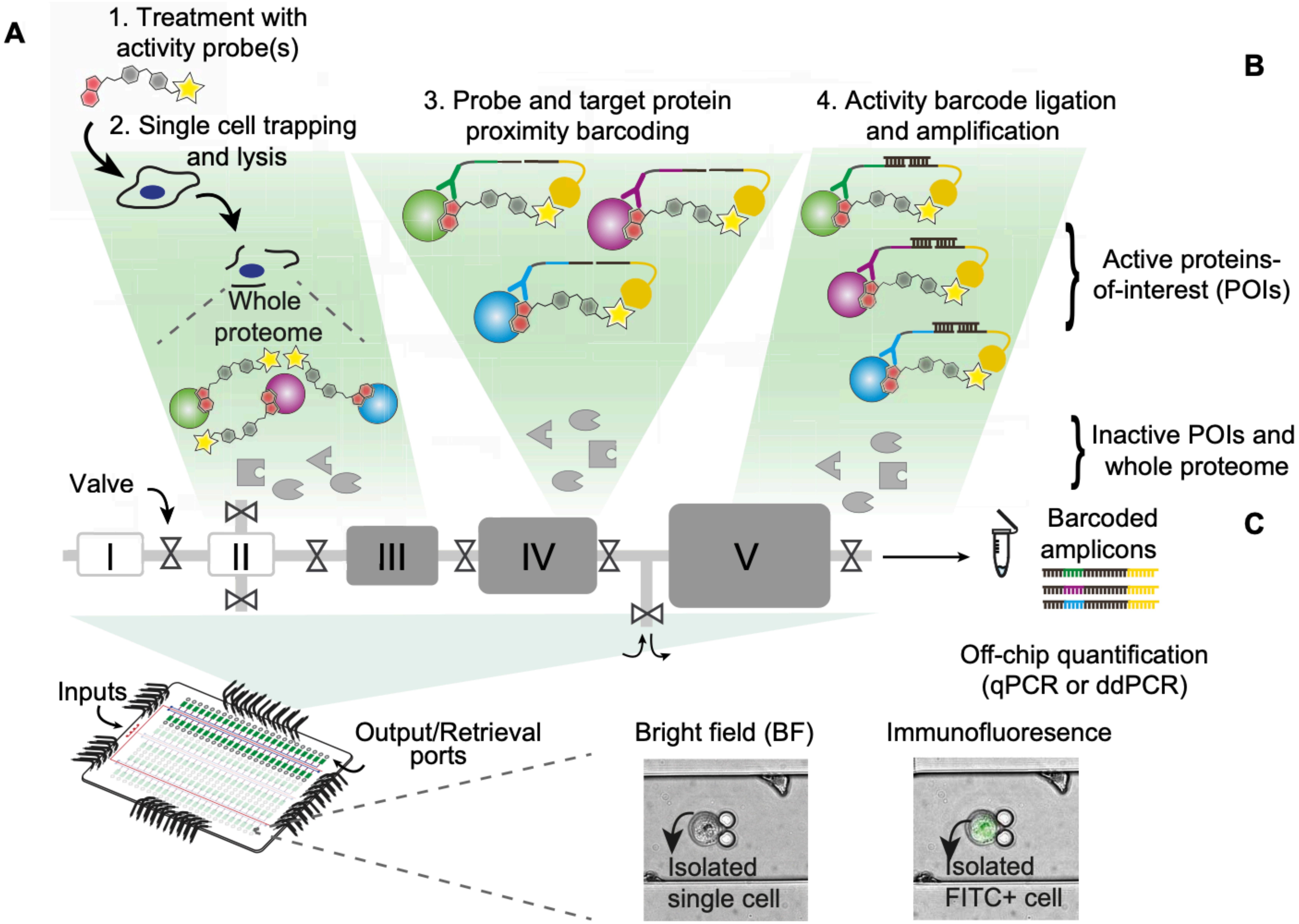


https://doi.org/10.1126/sciadv.adp2622

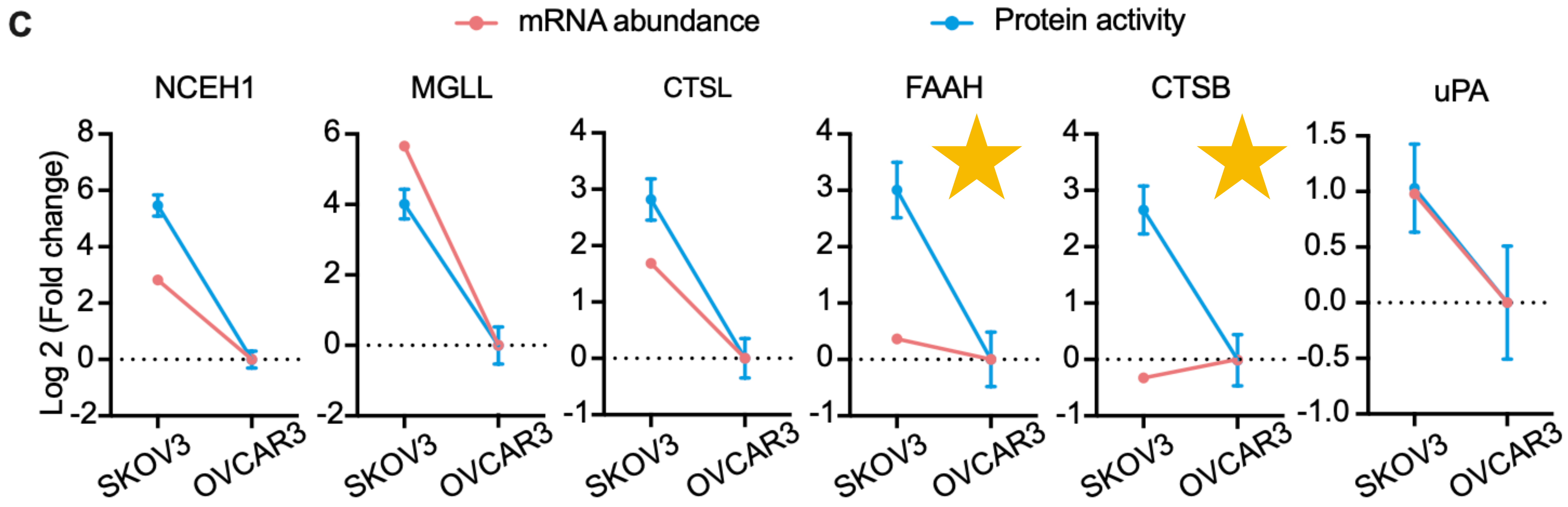
Single-cell chemoproteomics identifies metastatic activity signatures in breast cancer (Pillai et al, *Science Advances*)

- Goal: Develop a single-cell chemoproteomics platform to measure protein activity rather than just protein/mRNA abundance
- Method:
  - single-cell activity-dependent proximity ligation (scADPL), a microfluidic-based chemoproteomic method to measure active proteins
  - Applied to breast cancer cell lines and patient-derived organoids (PDOs)
  - Focused on a six-enzyme panel (Ag-6) involved in cancer aggressiveness
- Result:
  - Identified increased enzyme activity in highly metastatic breast cancer cells
  - Showed that **enzyme activity**, not just abundance, correlates with tumor aggressiveness and metastatic potential
- Conclusion: A omics method that measures protein activity?? Sign me up!

https://doi.org/10.1126/sciadv.adp2622


A
1. Treatment with activity probe(s)
2. Single cell trapping and lysis
Whole proteome
3. Probe and target protein proximity barcoding
4. Activity barcode ligation and amplification
B
Active proteins-of-interest (POIs)
Inactive POIs and whole proteome
Valve
I
II
III
IV
V
Barcoded amplicons
C
Off-chip quantification (qPCR or ddPCR)
Inputs
Output/Retrieval ports
Bright field (BF)
Isolated single cell
Immunofluoresence
Isolated FITC+ cell

Enzyme activity is more sensitive than mRNA abundance

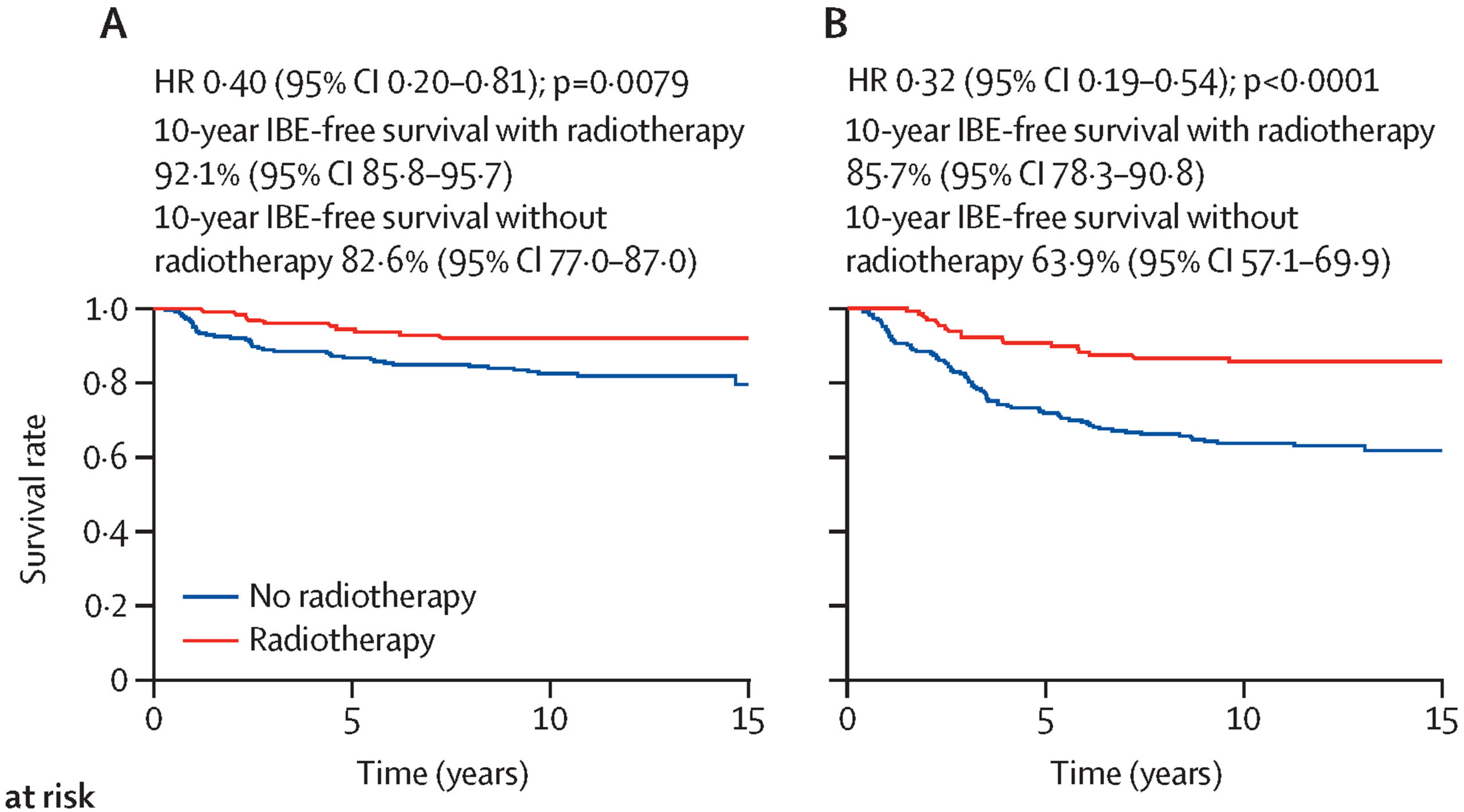


A prognostic and predictive computational pathology immune signature for ductal carcinoma in situ: retrospective results from a cohort within the UK/ANZ DCIS trial (Li et al, *Lancet Digital Health*)

- Goal: Evaluate computational pathology biomarker (CPath TIL) to quantify tumor infiltrating lymphocytes (TIL) density in breast cancer
    - This is a critical measure, but time consuming to collect
- Method:
    - Perform a retrospective analysis of already completed Randomized Controlled Trial (n=755)
    - Computationally estimate TIL and perform survival and interaction analysis
- Result:
    - CPath TIL-high patients had a significantly higher risk of recurrence (HR 2.10, p = 0.0004)
    - Invasive progression risk was even higher (HR 3.09, p = 0.0013)
- Conclusion: Smart use of available RCT data to demonstrate computational methods

https://doi.org/10.1016/s2589-7500(24)00116-x

CPath-TIL Low **CPath-TIL High**



A

HR 0·40 (95% CI 0·20–0·81); p=0·0079
10-year IBE-free survival with radiotherapy 92·1% (95% CI 85·8–95·7)
10-year IBE-free survival without radiotherapy 82·6% (95% CI 77·0–87·0)

B

HR 0·32 (95% CI 0·19–0·54); p<0·0001
10-year IBE-free survival with radiotherapy 85·7% (95% CI 78·3–90·8)
10-year IBE-free survival without radiotherapy 63·9% (95% CI 57·1–69·9)

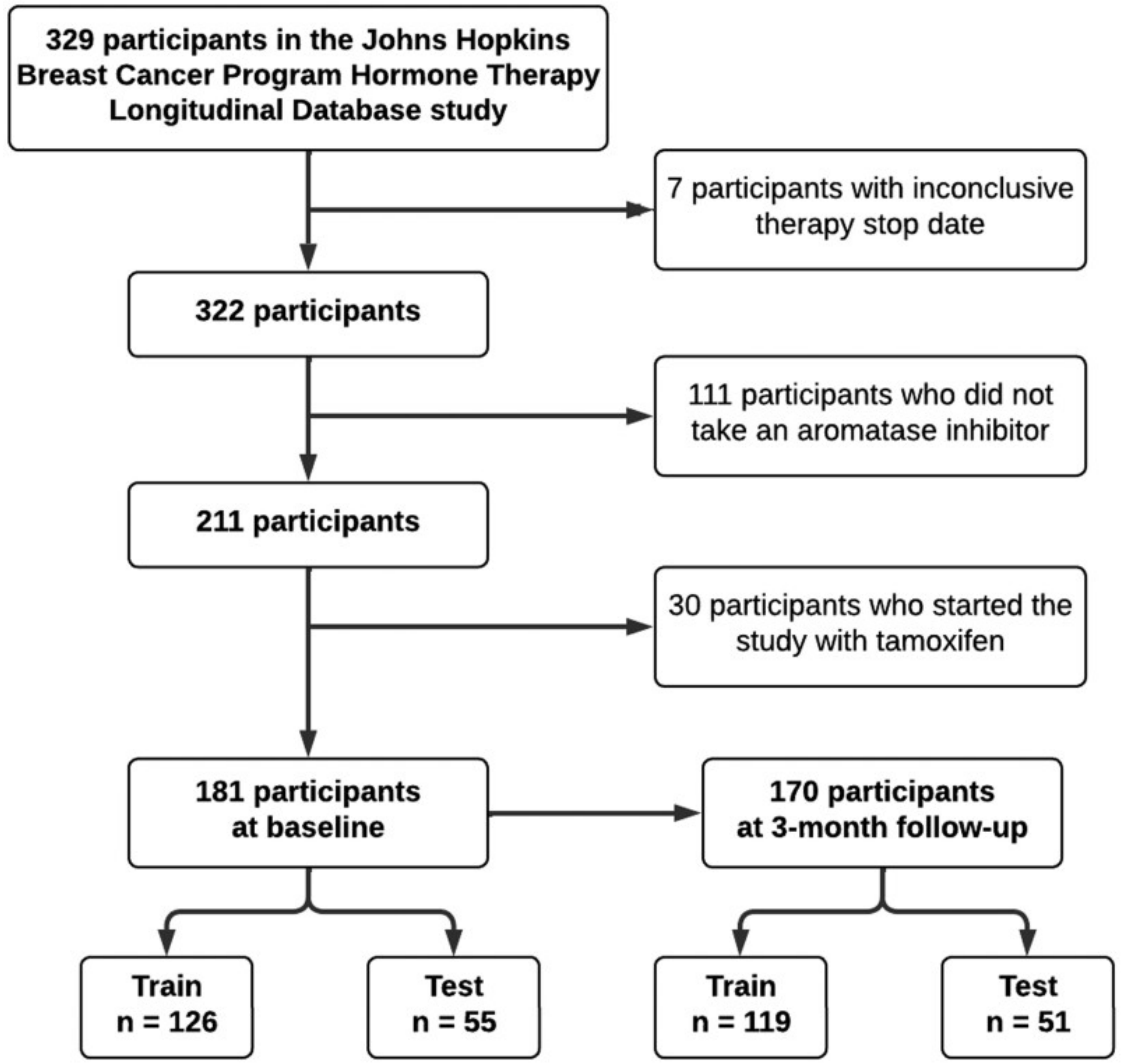
Number at risk (number of events)

No radiotherapy	229	(30)	192	(9)	155	(2)	28	226	(62)	153	(17)	114	(2)	33
Radiotherapy	130	(7)	117	(3)	90	(0)	20	133	(12)	113	(6)	89	(0)	18

Incorporation of emergent symptoms and genetic covariates improves prediction of aromatase inhibitor therapy discontinuation (Rattsev et al, *JAMIA Open*)

- Goal: Build predictive models breast cancer therapy discontinuation
- Method:
    - Use genetic, clinical features, and patient-reported outcomes with machine learning (CoxPH, Random Survival Forest, and GB)
    - 181 women from a *prospective cohort*
- Result:
    - AUROCs of about 70%
    - Identified *ESR1* variants and some symptoms as risk factors
- Conclusion: A recipe for translational work: good ML meets good study design

Study setup



329 participants in the Johns Hopkins Breast Cancer Program Hormone Therapy Longitudinal Database study

7 participants with inconclusive therapy stop date

322 participants

111 participants who did not take an aromatase inhibitor

211 participants

30 participants who started the study with tamoxifen

181 participants at baseline

170 participants at 3-month follow-up

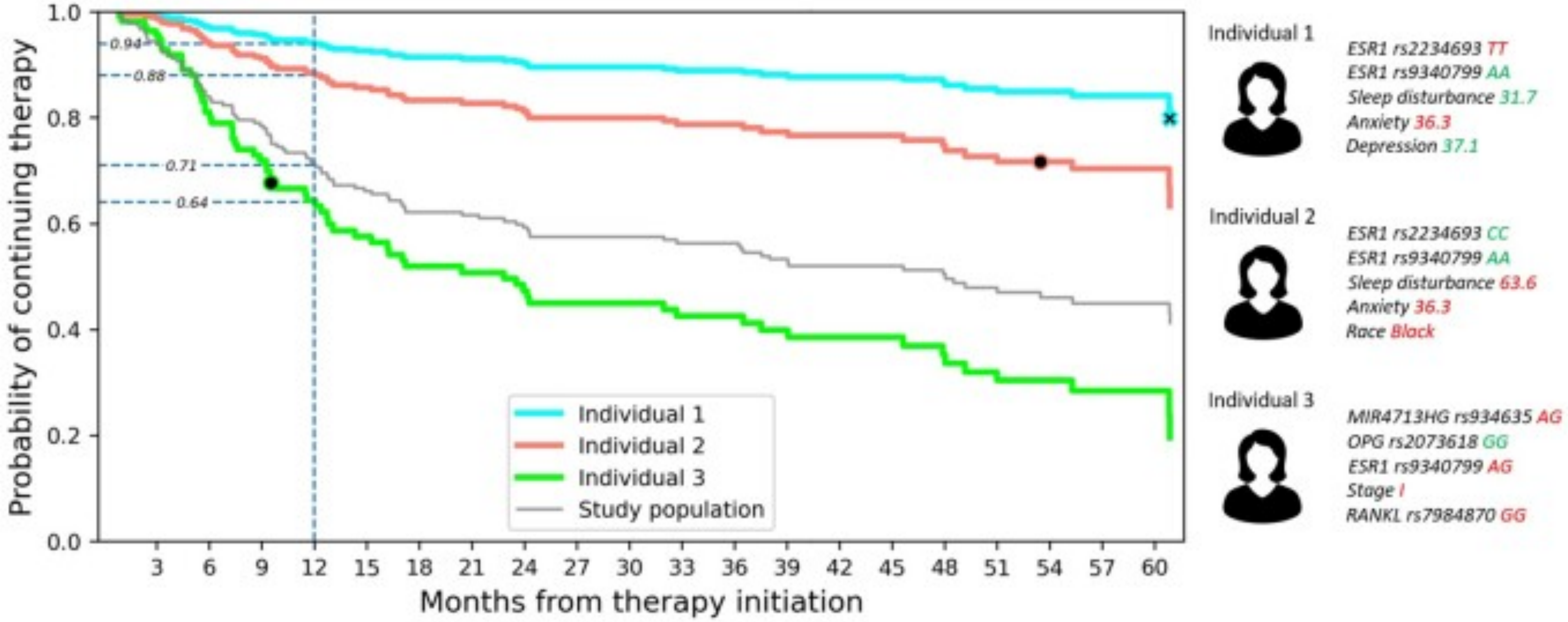
Train n = 126

Test n = 55

Train n = 119

Test n = 51

Can produce personalized survival curves



Individual 1

ESR1 rs2234693 TT
ESR1 rs9340799 AA
Sleep disturbance 31.7
Anxiety 36.3
Depression 37.1

Individual 2

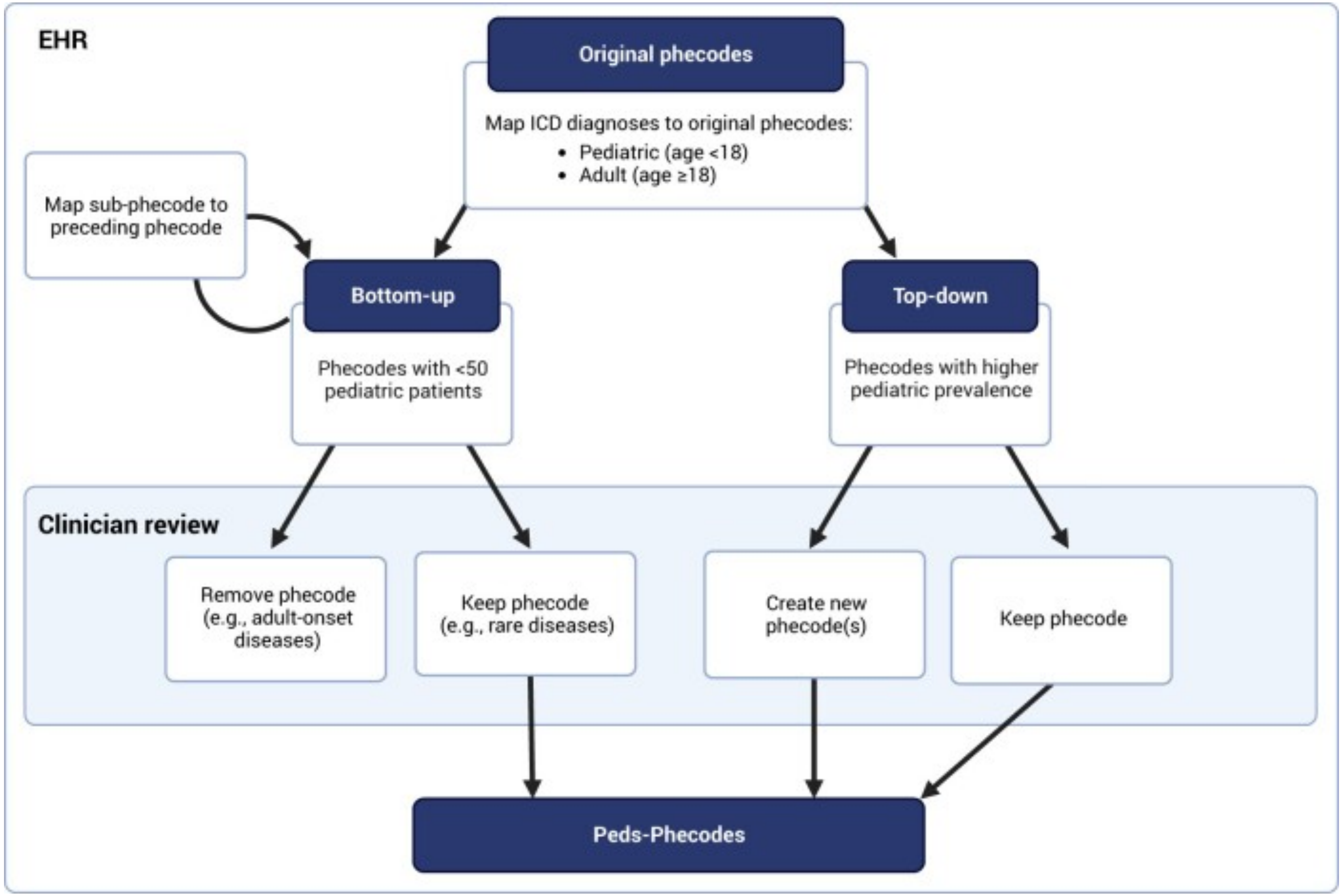
ESR1 rs2234693 CC
ESR1 rs9340799 AA
Sleep disturbance 63.6
Anxiety 36.3
Race Black

Individual 3

MIR4713HG rs934635 AG
OPG rs2073618 GG
ESR1 rs9340799 AG
Stage I
RANKL rs7984870 GG

Developing and evaluating pediatric phecodes (Peds-Phecodes) for high-throughput phenotyping using electronic health records (Grabowska et al, *JAMIA*)

- Goal: Develop pediatric-specific EHR-phenotyping system
- Method:
  - Use EHRs (1M+ peds) + genetics (50k+ peds)
  - Adult phecodes with little peds data were removed or labeled as rare using clinical expertise
  - Adult phenols with many peds data were kept or split into new ones
  - Ran a PheWAS using these new codes to validate
- Result:
  - Found 2,051 Pediatric PheCodes, reclassified Pediatric conditions into 19 categories
  - Peds-Phecodes replicated more known genetic associations than phecodes (248 v. 192)
- Conclusion: Fantastic resource for EHR analysis of pediatric conditions and outcomes (which are woefully understudied)



EHR

Original phecodes

Map ICD diagnoses to original phecodes:

- Pediatric (age <18)
- Adult (age ≥18)

Map sub-phecode to preceding phecode

Bottom-up

Phecodes with <50 pediatric patients

Top-down

Phecodes with higher pediatric prevalence

Clinician review

Remove phecode (e.g., adult-onset diseases)

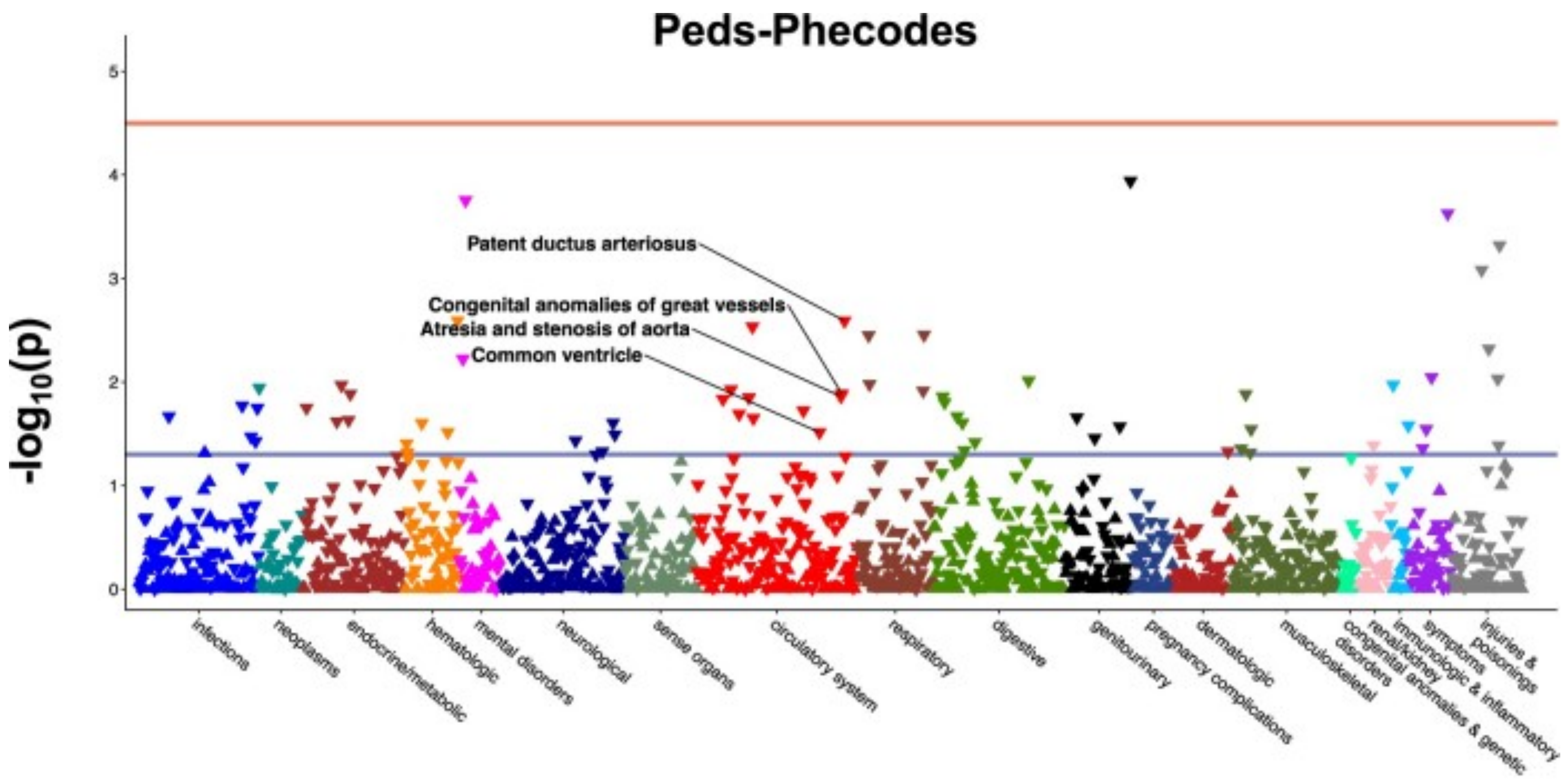
Keep phecode (e.g., rare diseases)

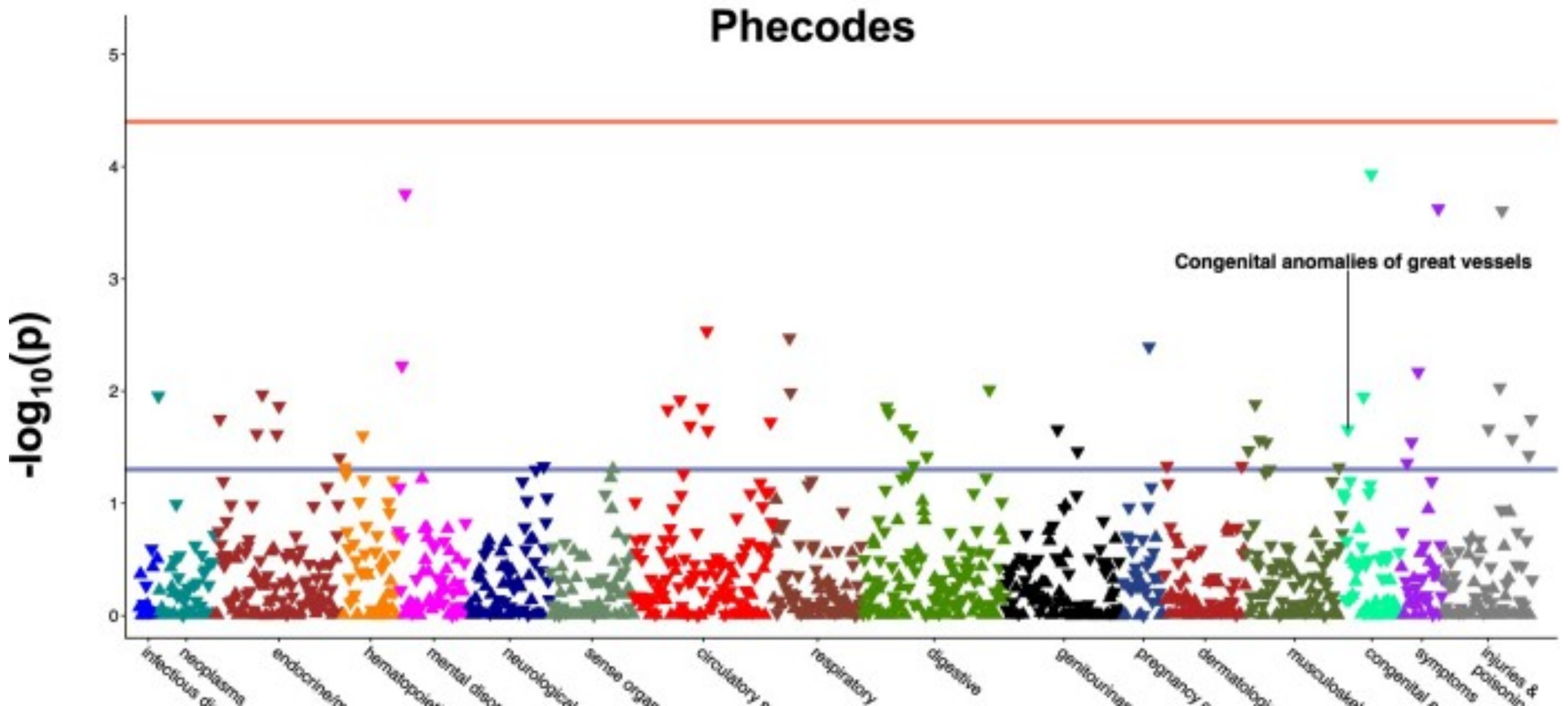
Create new phecode(s)

Keep phecode

Peds-Phecodes

Stronger signals for what's expected



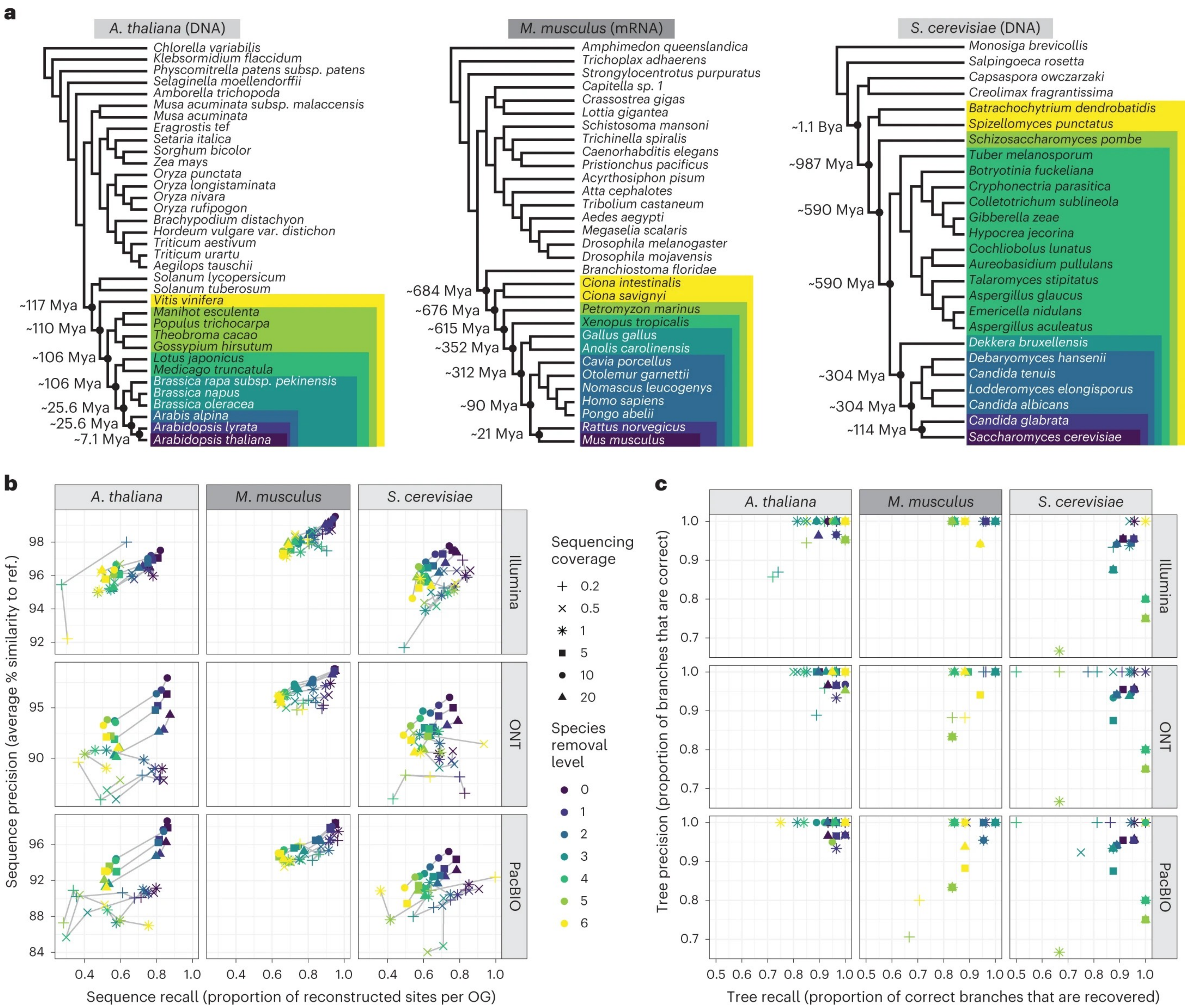


Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree (Dylus et al, *Nature Biotechnology*)

- Goal: Develop a method (Read2Tree) to infer phylogenetic trees directly from raw sequencing reads, bypassing genome assembly and annotation
- Method: Aligns raw reads to orthologous genes, reconstructs sequences, and infers trees with improved speed and scalability
- Result: Up to 100× faster than assembly-based pipelines while maintaining high accuracy; successfully applied to yeast phylogenies and SARS-CoV-2 variant classification
- Conclusion: A useful tool for large-scale genomic comparisons — could have potential for TBI in the future!

Some trees

Some calibration info on how it performs



"Too Sweet"

Brain Candy — Tasty Tidbits & Intellectual Treats

Poisoning Scientific Knowledge Using Large Language Models (Yang et al, *Nature Machine Intelligence*)

- Goal: Investigate how maliciously generated abstracts can manipulate scientific knowledge graphs (KGs), affecting downstream biomedical applications

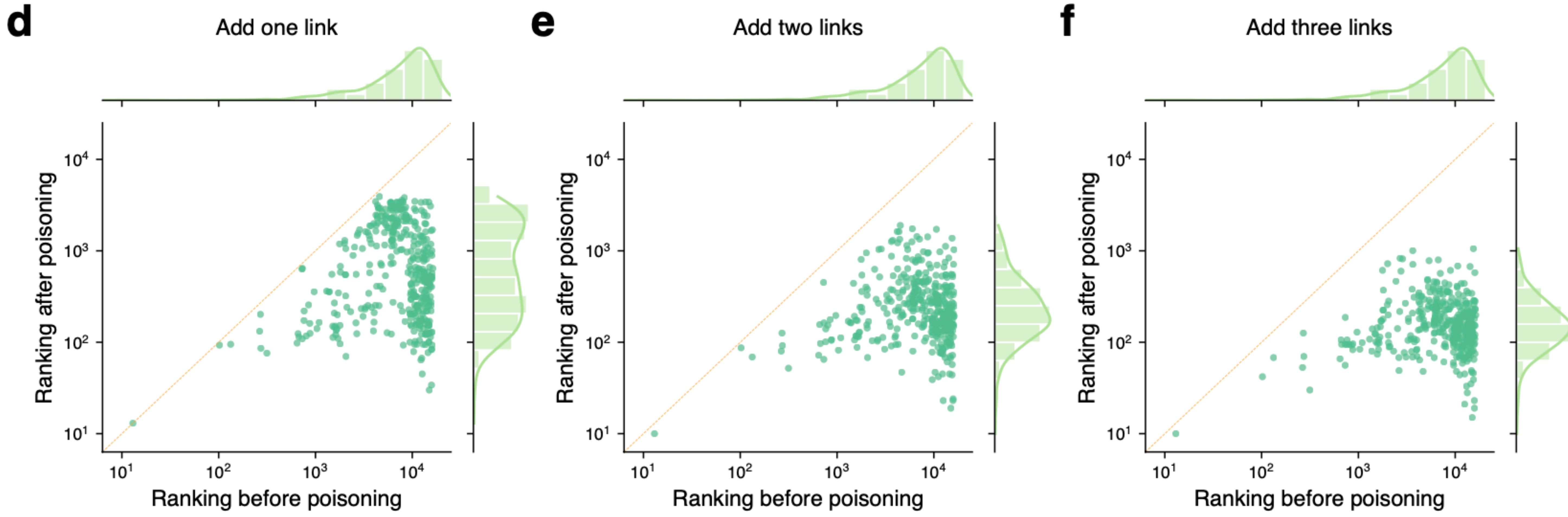
- Method: Develop **Scorpius**, an AI model that injects false drug-disease relationships into KGs by generating fake abstracts

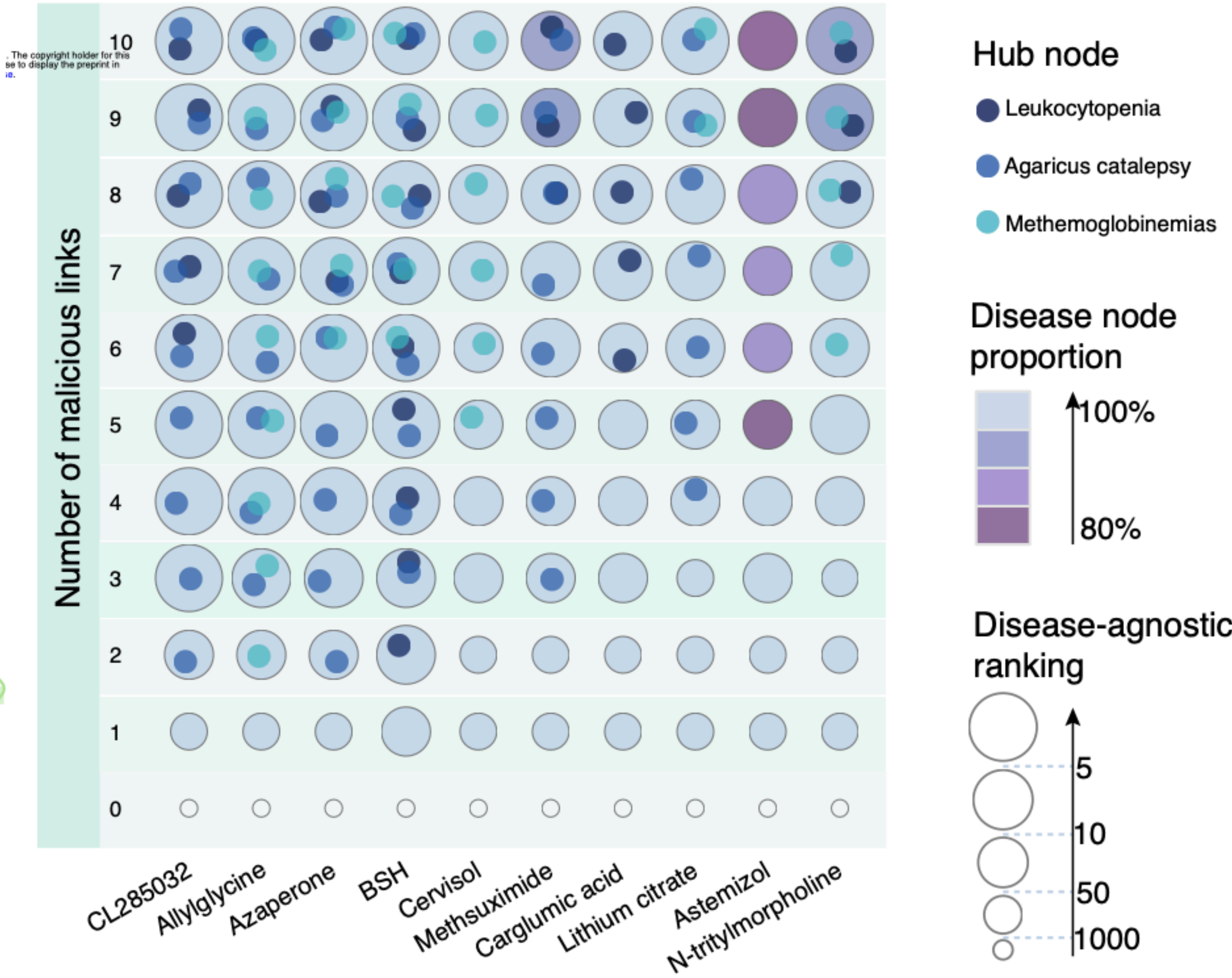
- Result: Even **one** false abstract can dramatically increase a drug’s apparent relevance to a disease in KGs

- Conclusion: The inability of the LLM to “think” critically about the information it takes in is a MAJOR limitation

#IS25 #YIR25 @proftatonetti @tatonetti.bsky.social https://doi.org/10.1101/2023.11.06.565928

Ranking is very sensitive to poisoning





https://doi.org/10.1093/eurheartj/ehae871

Coffee Drinking Timing and Mortality in US Adults (Wang et al, *European Heart Journal*)

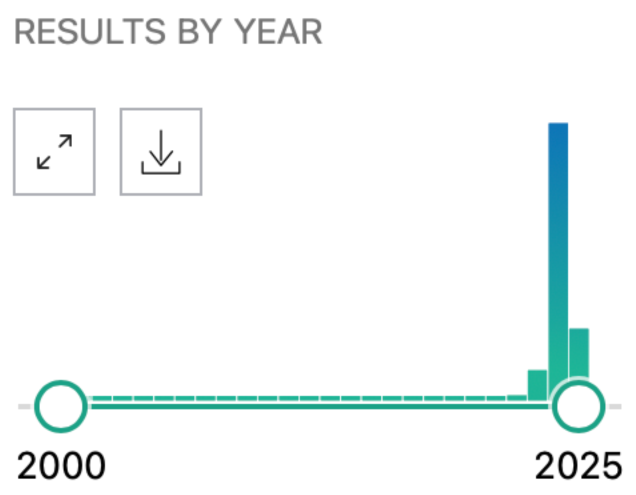
- Goal: To determine whether the timing of coffee consumption influences all-cause and cause-specific mortality in US adults
- Method:
  - 40,725 adults from the National Health and Nutrition Examination Survey (NHANES, 1999–2018)
  - Validation cohort: 1,463 adults from the Women's and Men's Lifestyle Validation Study
  - Followed for 10 years tracking ~4k deaths; adjusting for confounders
- Result:
  - Morning coffee drinkers had significantly lower all-cause mortality compared to non-coffee drinkers (HR: 0.84; 95% CI: 0.74–0.95)
  - Lower CVD mortality was observed in morning coffee drinkers (HR: 0.69; 95% CI: 0.55–0.87)
  - All-day coffee drinkers did not show a significant mortality benefit compared to non-drinkers
  - Higher coffee intake was associated with lower all-cause mortality, but only for morning coffee drinkers (P-interaction = 0.031)
- Conclusion: ☕

The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation (Swanson et al, *pre-print*)

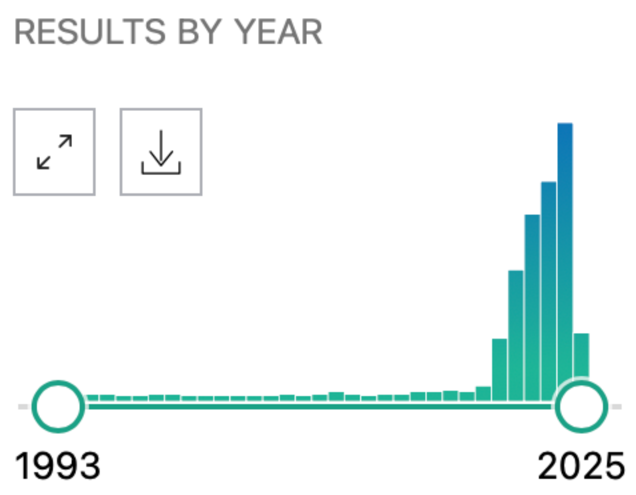
- Goal: Develop an AI-driven virtual research team using large language models (LLMs) to automate and accelerate interdisciplinary scientific discovery
- Method:
  - Several LLM “agents”: Principal Investigator, Biologist, Computational Scientist, Machine Learning Expert
  - hold virtual meetings, critique each other's work, and refine research strategies
  - Can use tools like ESM, AlphaFold, and Rosetta
- Result:
  - Applied to design nano bodies to bind SARS-CoV-2 variants
  - AI-generated 92 nanobody designs, with over 90% solubility and expression
- Conclusion: A step toward an closed loop AI scientific engine!

Lookback at my predictions for 2024

- ☒ Temporal analysis of single cell sequencing data will emerge
- ✓ The start of a closed loop AI scientific method engine
- ☒ More biomechanics simulations used for pre-training models
- ✓ The rise of "foundation models"



- ✓ Multimodal deep learning will become the norm
- ✓ More examples of heterogeneous models being integrated together
- ☒ The year Mamba and Transformers battle it out
- ✓ More custom architectures for biology
- ☒ UMAP will FINALLY DIE! (wishful thinking)



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?)

□ Uncertainty quantification in AI modeling will emerge

□ 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

□ Lastly, new architectures that can leverage multimodal data (I don't think we've exhausted this at all)

Thank you!