2024 Translational Bioinformatics Year-in-Review

@proftatonetti

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

2024 AMIA Summits - Boston

Disclosures

- NIH, FDA, Pfizer, AstraZeneca, Janssen, Amgen, PhARMA Foundation, CARI Health
- 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 2024)
- Recognize innovative work and identify opportunities for the future

Process

- Follow the literature throughout the year
- Work in collaboration with the talented and generous **AMIA Year-in-Review Committee**
- Triage all papers from a set of relevant journals since Jan 2023
  - 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 1-3 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 2023 - Mar 2024)

- Focused on human biology

- **What's NOT included:**

  - Amazing biology with straightforward informations

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

  - Perspectives, reviews (for the most part)

This is all thanks to…

The 2024 AMIA Year-in-Review Team!



Nick Reid
PhD Candidate
University of Washington
Chair-Elect, AMIA Student Working Group



Humayera Islam
PhD Candidate in Health Informatics
University of Missouri-Columbia
Chair, AMIA Student Working Group

Abrianna Onwuemene
University of Michigan School of Information-MHI

Aziz Zafar
Department of Biomedical Informatics, Ph.D. student, Columbia University



Burhan Sabuwala
Computational Biology and Bioinformatics, PhD program, Yale University



Christopher Allsman
Biomedical Informatics MA student, Columbia University

Farhan Quadir
University of Missouri-Columbia/PhD in Computer Science



Flora Kibong
University of Michigan School of Information-MHI



Kae Nurge
Columbia University, Masters in Biomedical Informatics



Kevin Jin
PhD Computational Biology and Bioinformatics, Yale University



Pushkala Jayaraman
Icahn School of Medicie at Mount Sinai New York/ PhD in Biomedical Sciences (AI and Emerging Tech)



Shashank Yadav
University of Arizona / Biomedical Engineering



Susannah Kisvarday
Boston Children's Hospital



Tra Nguyen
Boston Medical Center



Venessa Olanrewaju
Informatics Fellow, Association of Public Health Laboratories

Vishala Mishra
Informatics, Duke

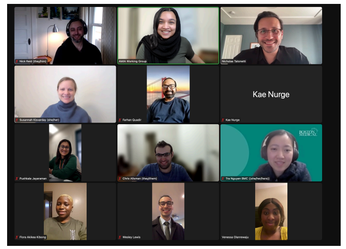


Wesley (Wes) Lewis
Computational Biology and Bioinformatics
PhD at Yale

Xueting (Leona) Wang
Computational Biology and Bioinformatics
PhD at Yale



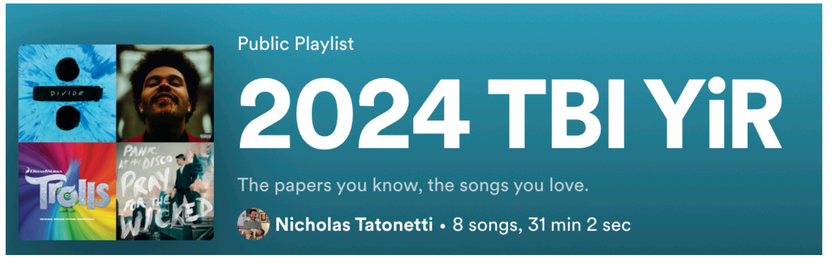
Zhiyu Wan
Postdoc at Department of Biomedical Informatics, Vanderbilt University Medical Center



2024 Translational Bioinformatics Year-in-Review Committee

Final List

- 1,085 papers triaged; 462 reviewed and scored by the committee
- 29 presented here + 9 shout outs
  - Apologies for those I missed, misunderstood, or misjudged, mistakes are all mine
- 6 TBI topics:
  - ***Shape of you*** - *Unveiling the Structure in Biology*
  - ***Blinding Lights*** - *Shining a Light on Genetics and Genomics*
  - ***Halo*** - *Advancing Prediction and Classification*
  - ***Photograph*** - *Capturing Insights through Imaging*
  - ***You Belong With Me*** - *Exploring interactions and Networks*
  - ***High Hopes*** - *Pushing the Boundaries in Drug Discovery*



https://spoti.fi/3TGtWmT

- All authors are mentioned if ≤3, all first authors otherwise
- Slides will be posted to www.tatonettlab.org and linked to my twitter (X, whatever) account

Here we go…

breathe, Nicholas



"Shape of You"

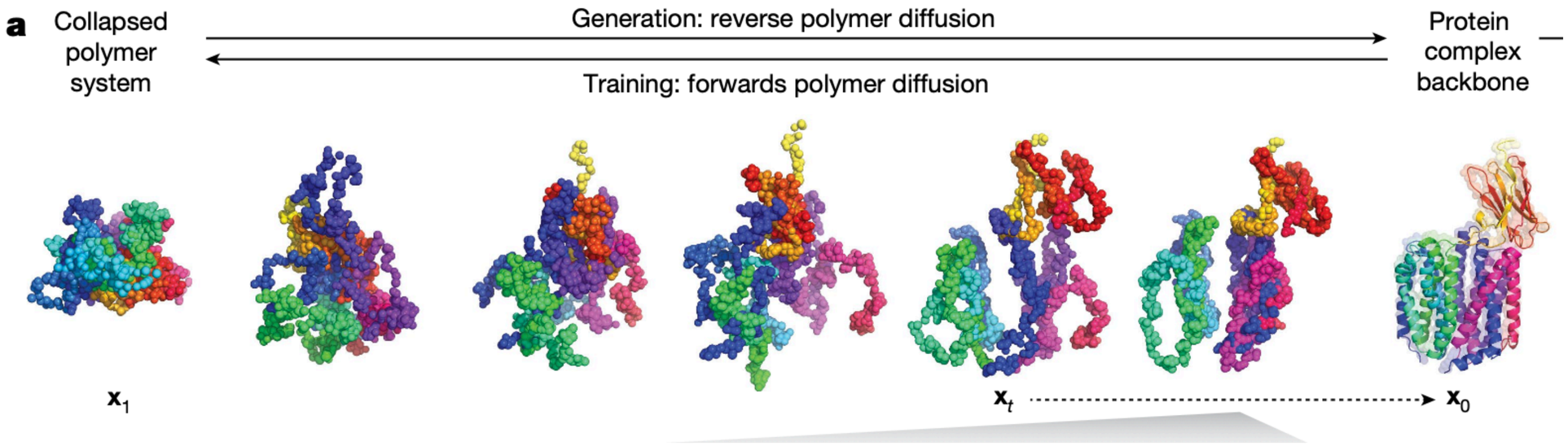
Unveiling the Structures in Biology

 https://doi.org/10.1038/s41586-023-06728-8

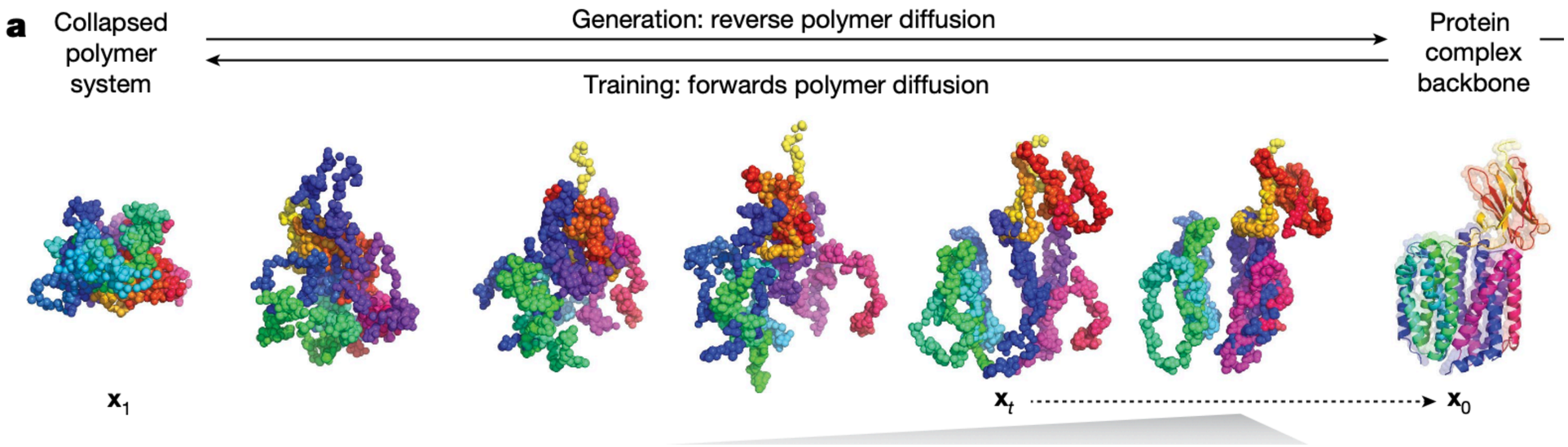
Illuminating protein space with a programmable generative model (Ingraham et al, *Nature*)

- Goal: Build a generative AI model of proteins
- Method: Graph neural networks and diffusion models
- Result:
  - Chroma produces structures that satisfy energetics constraints and re-fold correctly
  - Experimentally validated a handful of proteins to prove it
- Conclusion: Synthetic biology is about to get really interesting

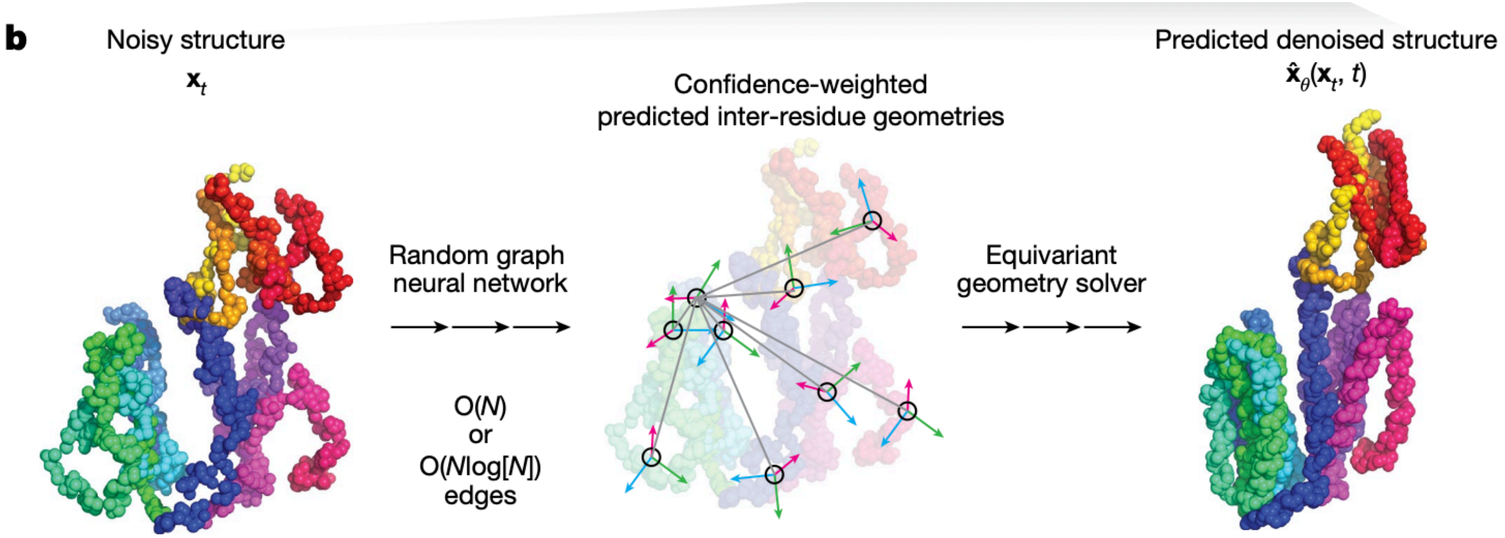
Take real backbones and learn how to turn them into mush



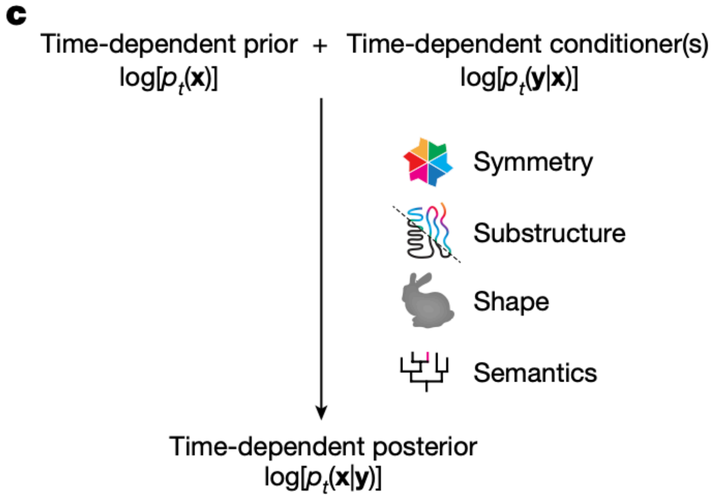
Run it in reverse to generate backbones



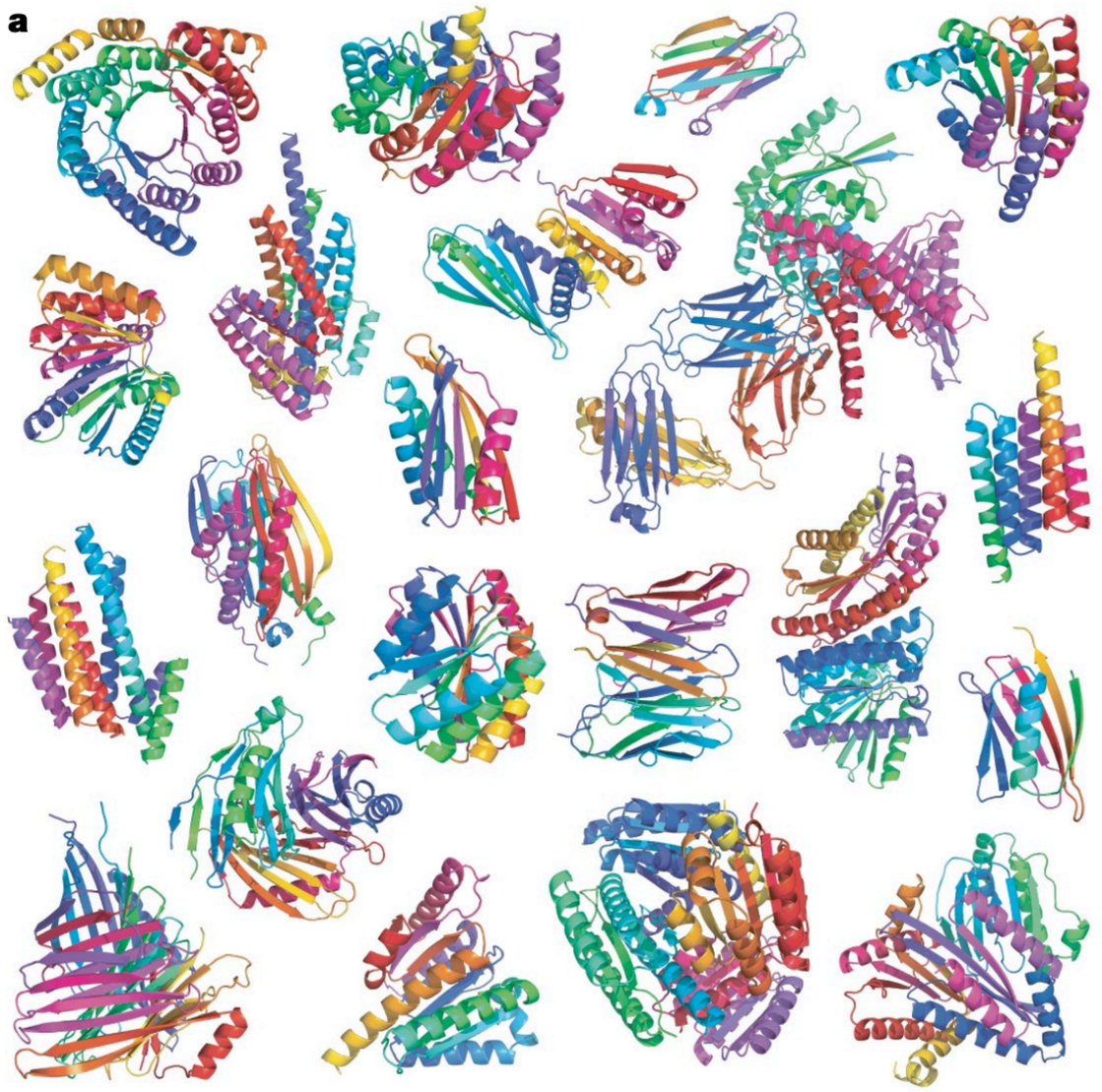
Can use the model for denoising



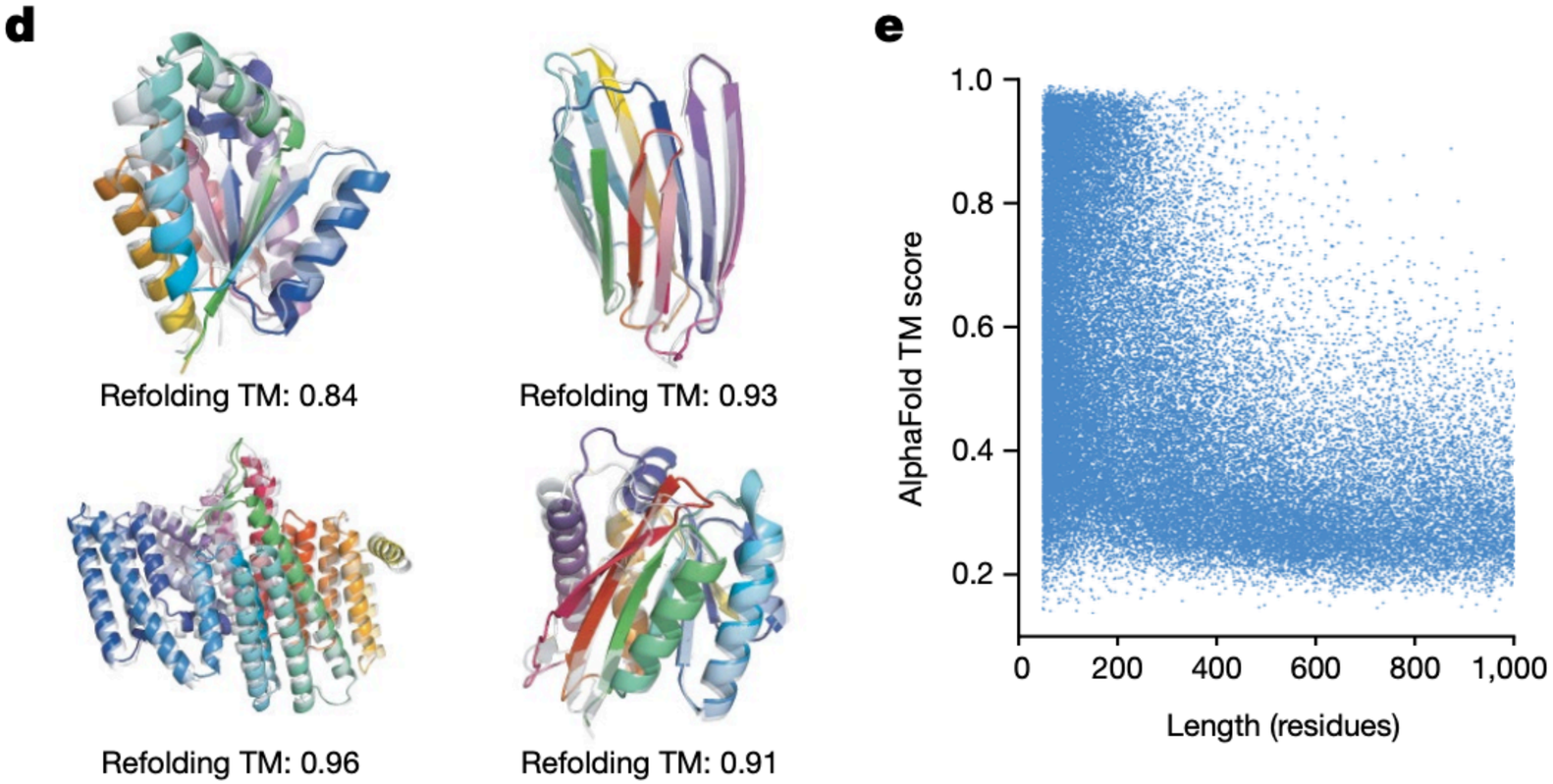
Can set priors on important protein structure constraints



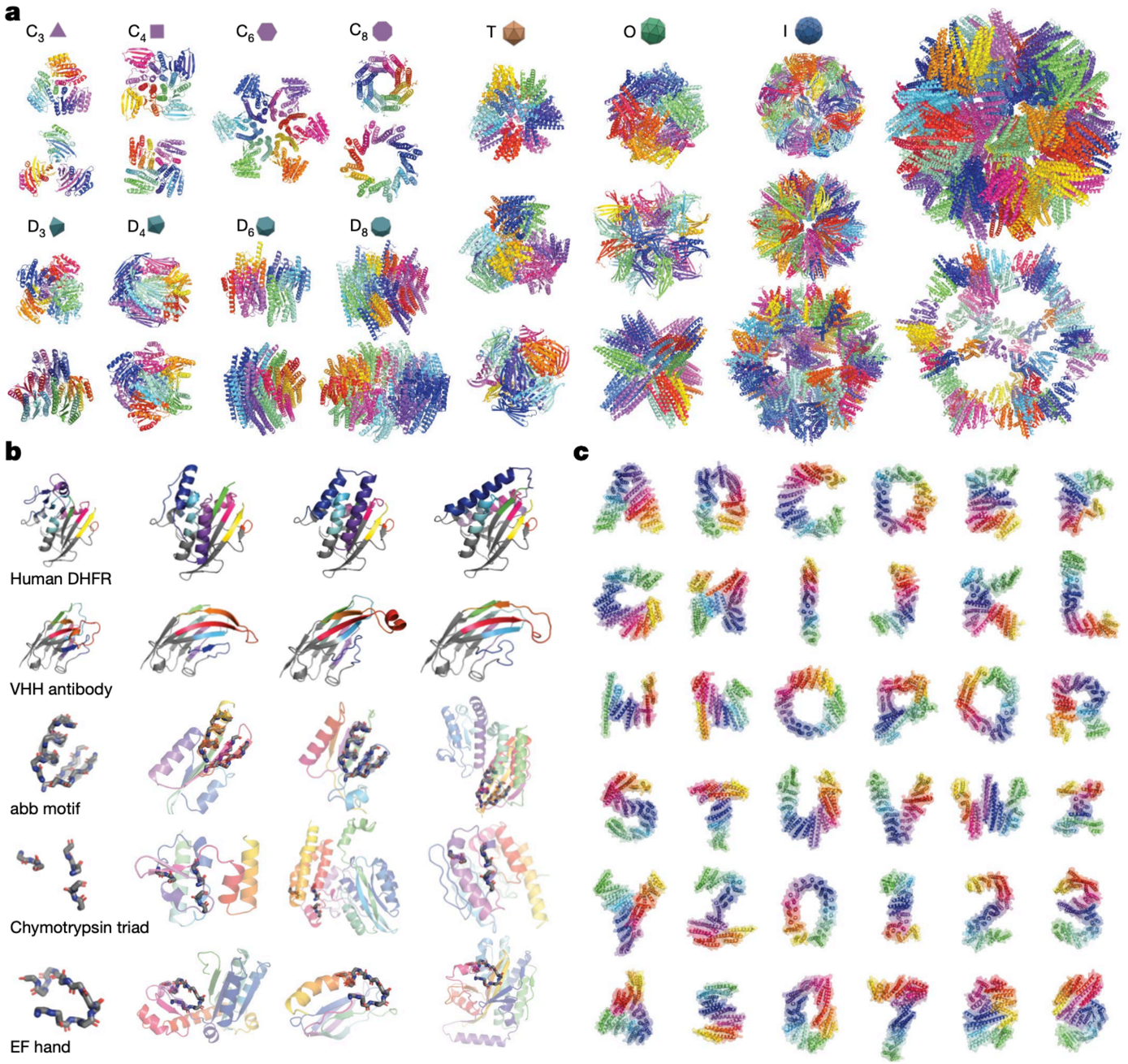
Pretty, novel proteins



Have AlphaFold re-fold them and see if you get the same structure…



#IS24 #YIR24 @proftatonetti



Supposedly you could express this protein and it would fold like this

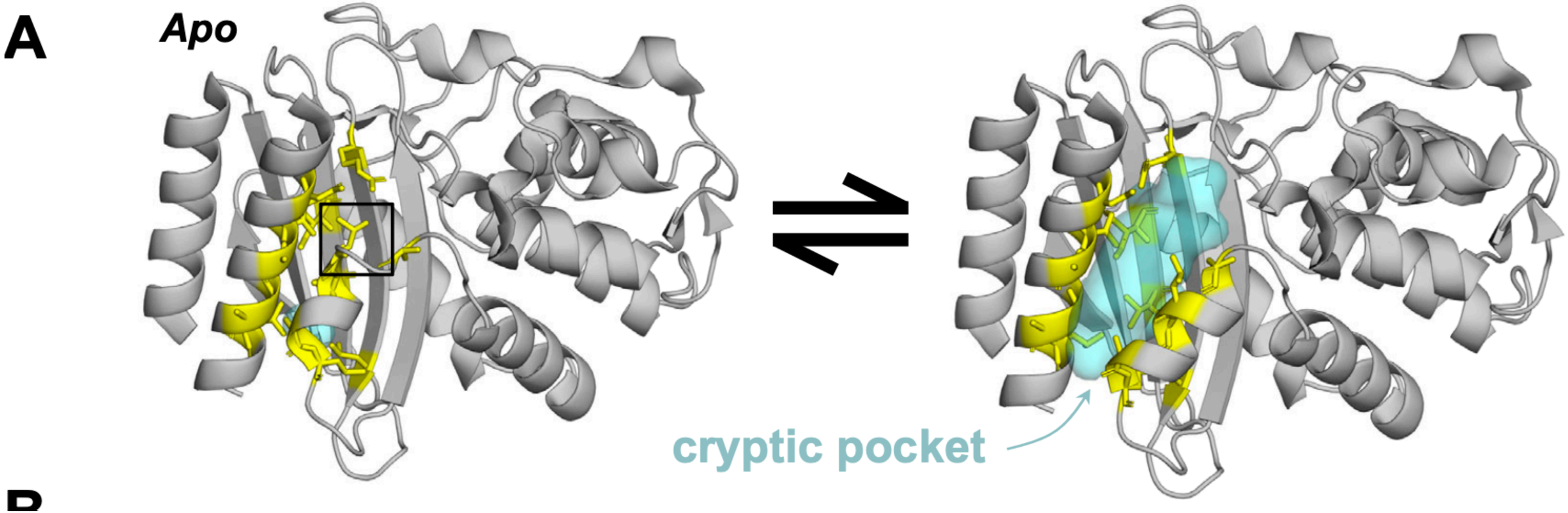
Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network (Meller et al, *Nature Communications*)

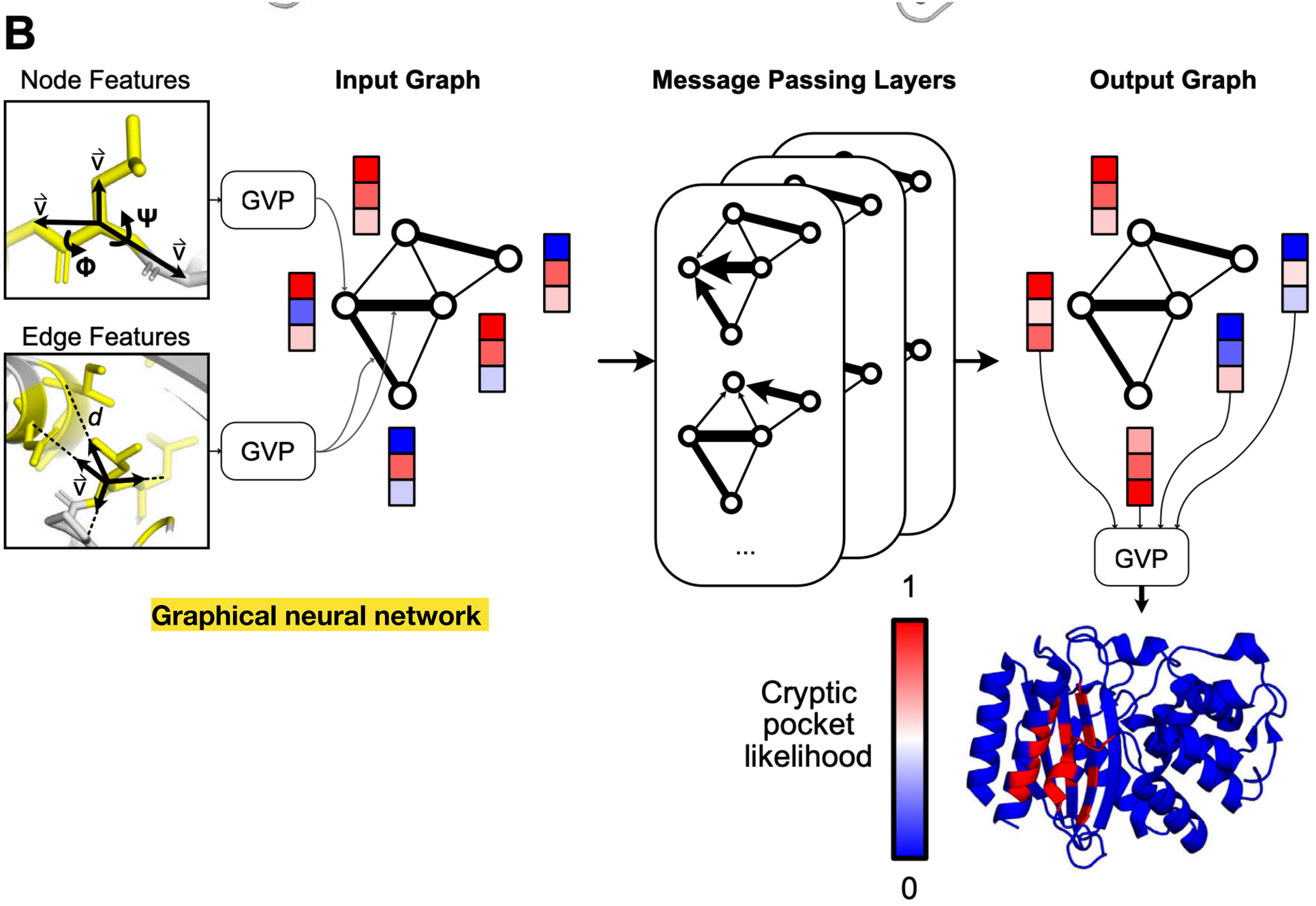


- Goal: Predict cryptic binding pockets (only formed in dynamics)
- Method: Graphical neural network + {DATA}
- Result:
  - AUROC of 0.87 on cryptic pocket prediction (against benchmark of 39 proteins)
  - 1,000X faster than previous methods
  - Apply to whole proteome, estimate that more than half of “undruggable” proteins have cryptic pockets!
- Conclusion: The druggable genome just got a whole lot bigger…

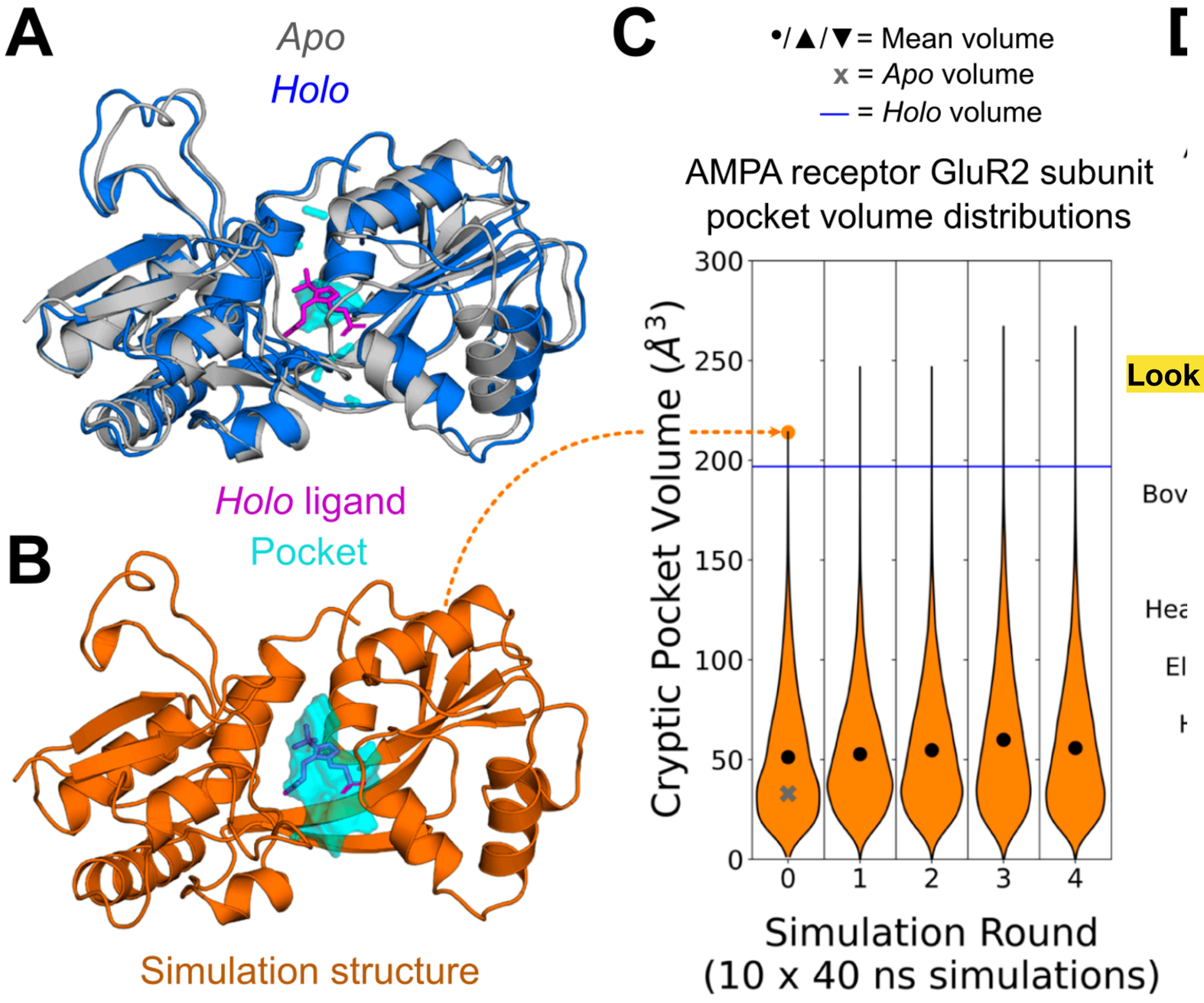
Structure from crystallography

Pocket opens up during molecular dynamics simulation (and real life)

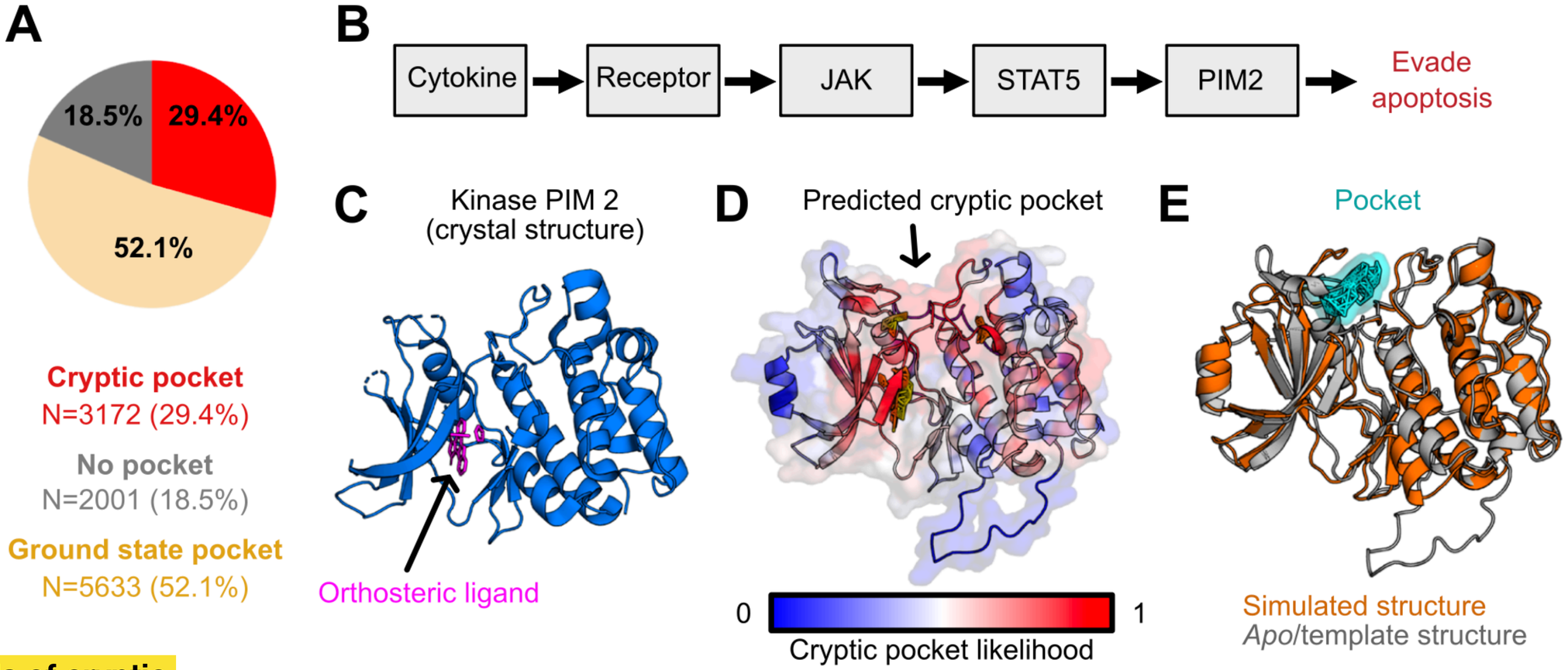




Graphical neural network


A
Apo
Holo
Holo ligand
Pocket
B
Simulation structure
C
•/▲/▼= Mean volume
x = Apo volume
— = Holo volume
AMPA receptor GluR2 subunit
pocket volume distributions
Cryptic Pocket Volume (Å3)
300
250
200
150
100
50
0
0
1
2
3
4
Simulation Round
(10 x 40 ns simulations)
Look how rare this pocket is!
Example prediction

Drug development example

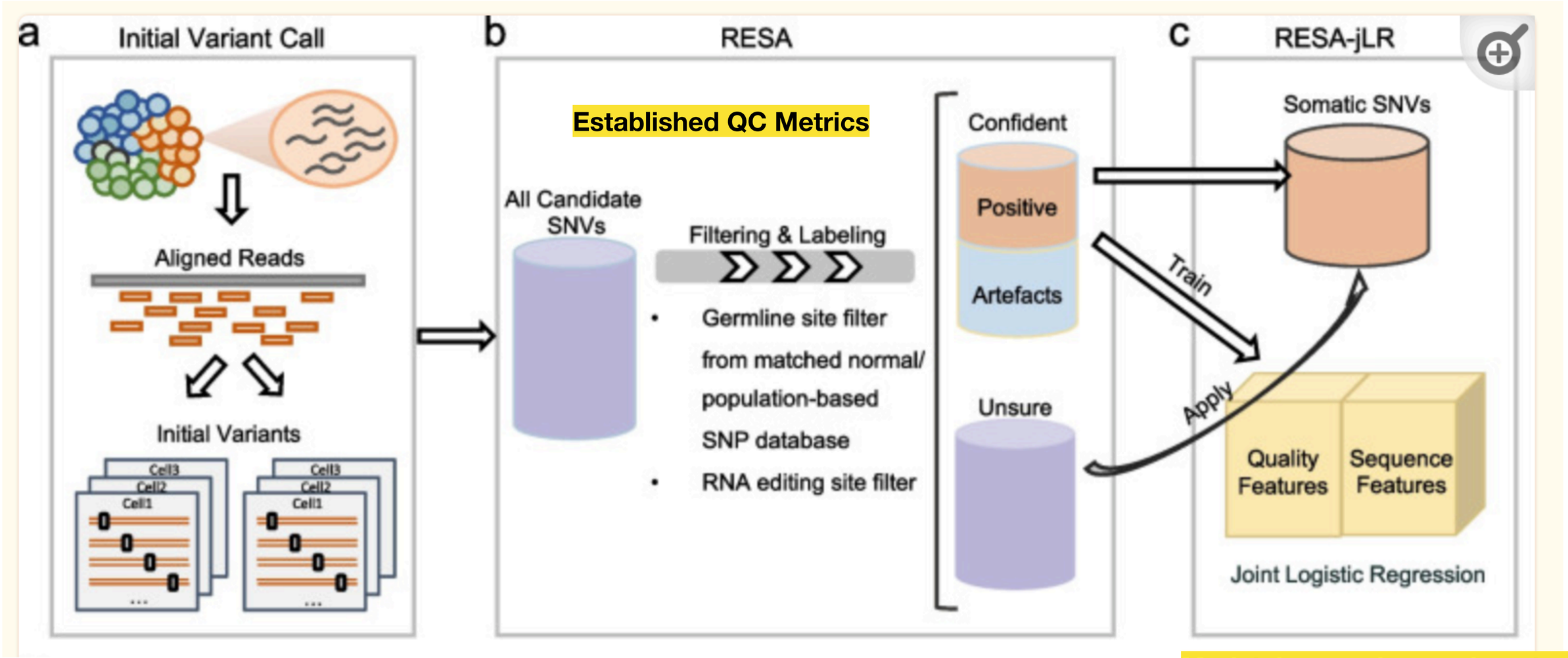


Thousands of cryptic pocket predictions

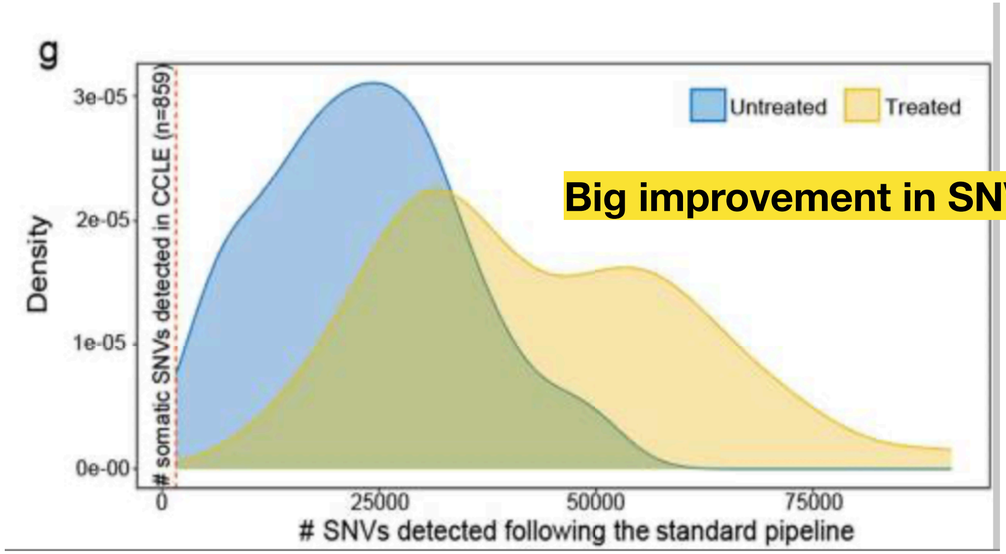
 https://doi.org/10.1186/s13073-023-01269-1

De novo identification of expressed cancer somatic mutations from single-cell RNA sequencing data (Zhang et al, *Genome Medicine*)

- Goal: Make SNV calling from scRNA-Seq possible and reliable
  - Highly desirable to know which somatic variants exist in a cell from which you have transcriptomics data
  - Current methods either integrate scWES (difficult) or were designed for bulk RNA (noisy)
- Method:
  - Use high quality filters to identify high confidence positives and artifacts (ie. negatives)
  - Train a model to differentiate and apply to the low-confidence/unsure variants
- Result: Massive improvement in scSNVs detected (from reference dataset of paired WES)
- Conclusion: Single cell data is so rich we've really only just begun, lots of work to do!

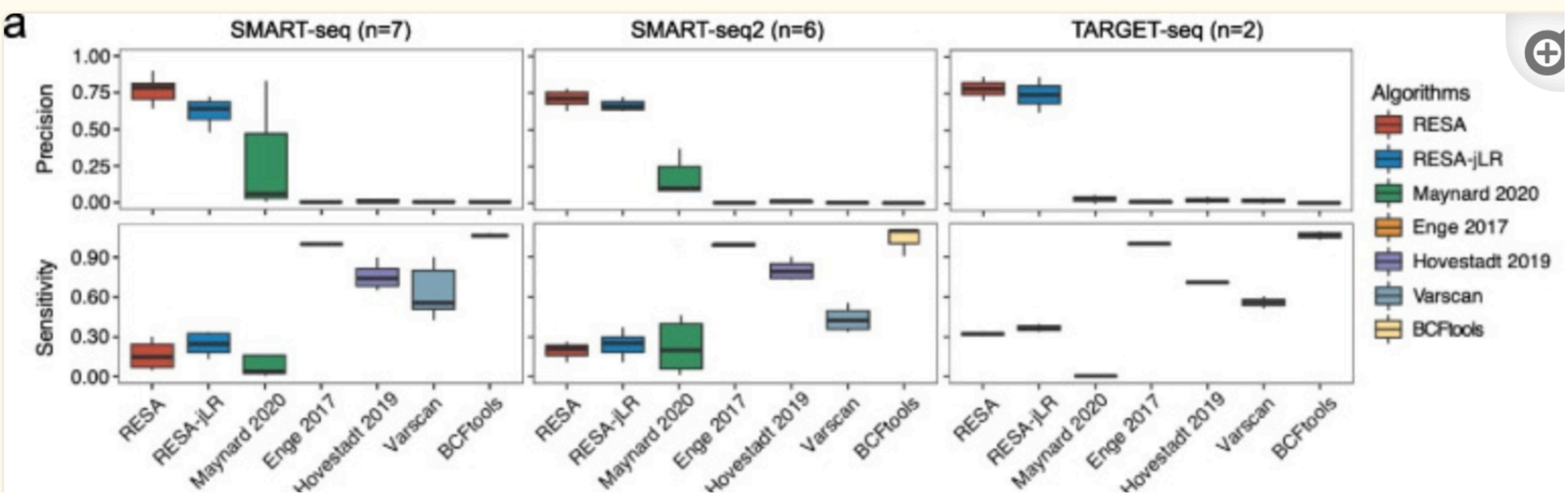


Model that can be used to classify



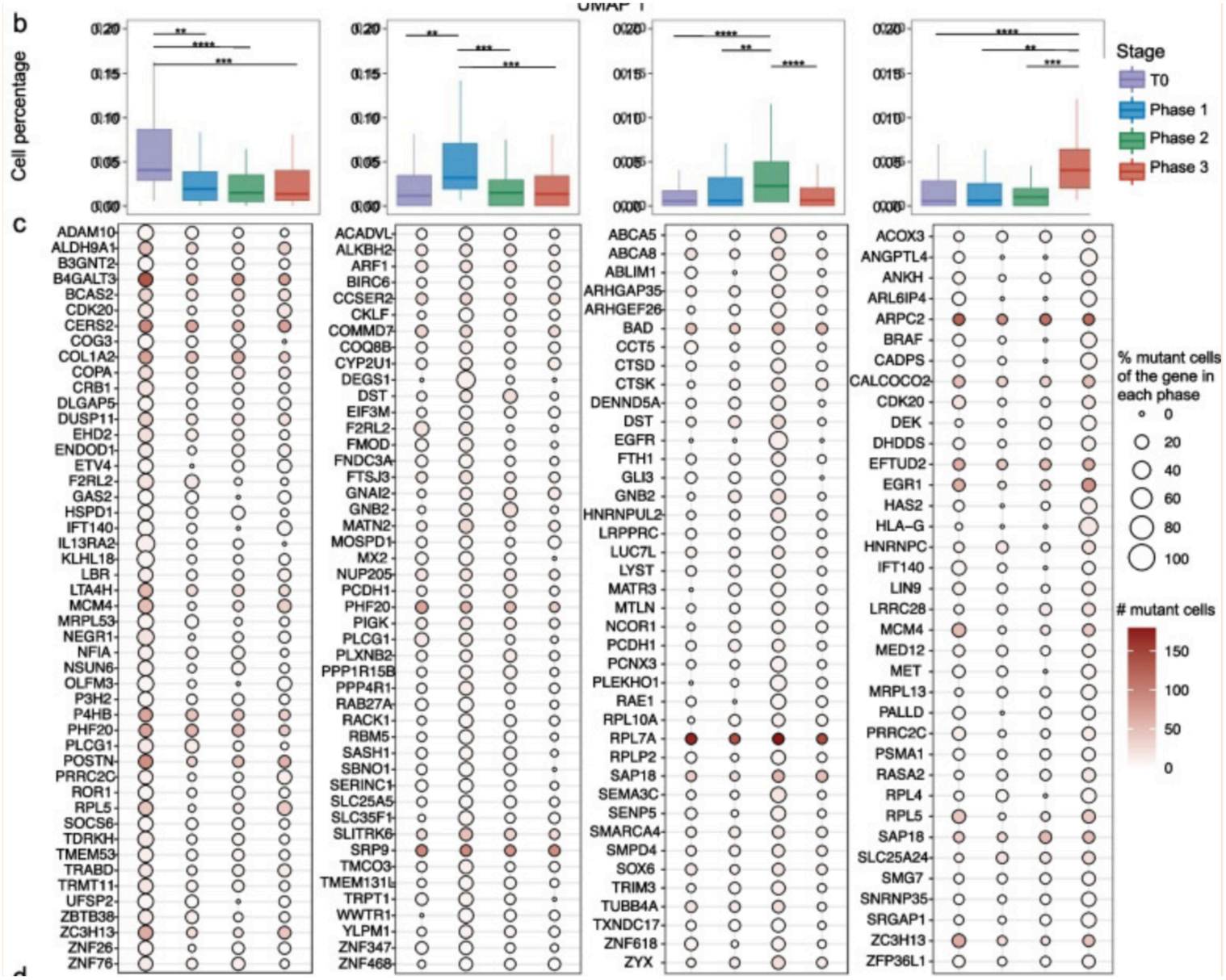
Big improvement in SNV detection (*in silico*)

Much better than other methods on real data



Shock and Awe Figure

Shows they can rediscover SNVs known to be associated with different cancer stages

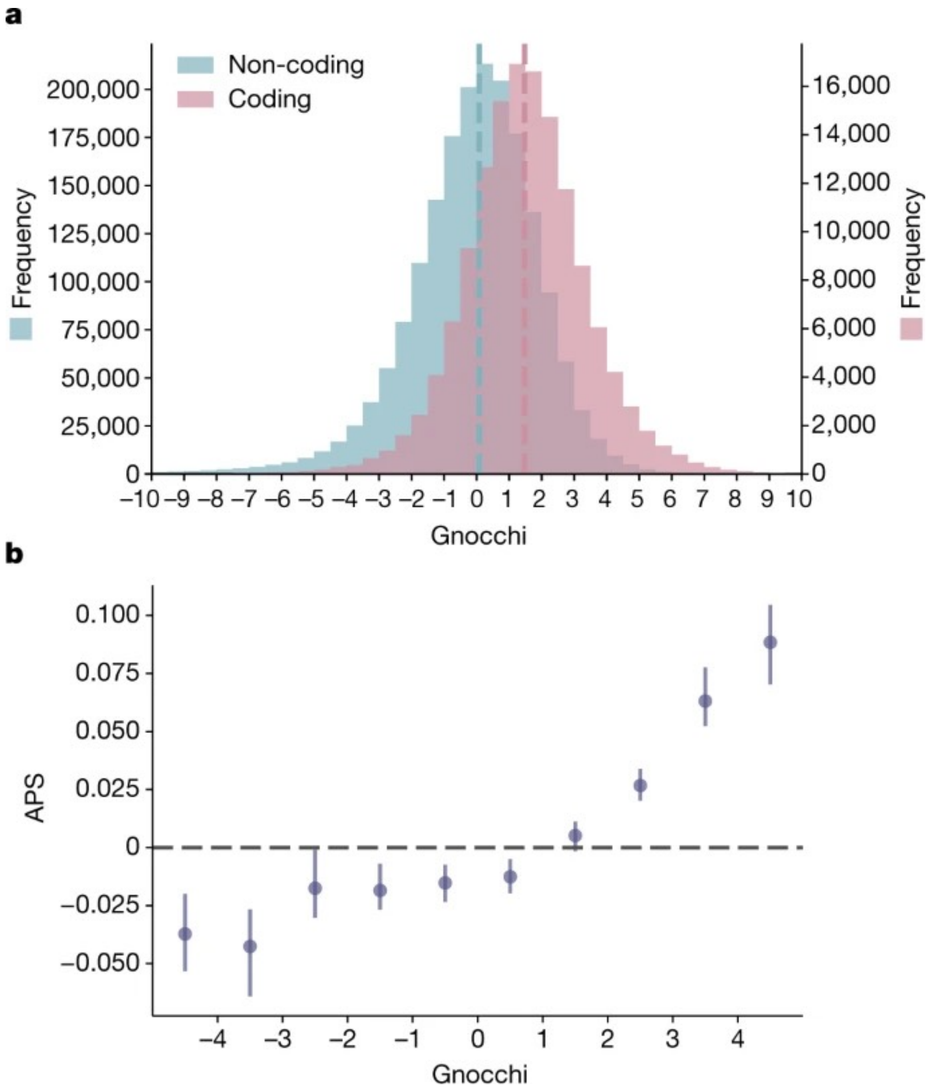


 https://doi.org/10.1038/s41586-023-06045-0

A genomic mutational constraint map using variation in 76,156 human genomes (Chen et al, *Nature*)

- Goal: Build a constraint model describing non-coding human genetic variation
- Method:
  - Use 76,156 human genomes
  - Divide genome into continuous non-overlapping 1kb sections
  - Quantify constraint (lack of variation) for each of these windows and z-score normalize (== Gnocchi score)
- Result:
  - Gnocchi scores are higher for coding than non-coding (gut check) and correlated with known structural variation constraints
  - State-of-the-art performance at classifying function of non-coding variants
- Conclusion: Still a place for non-deep learning based models in genetics

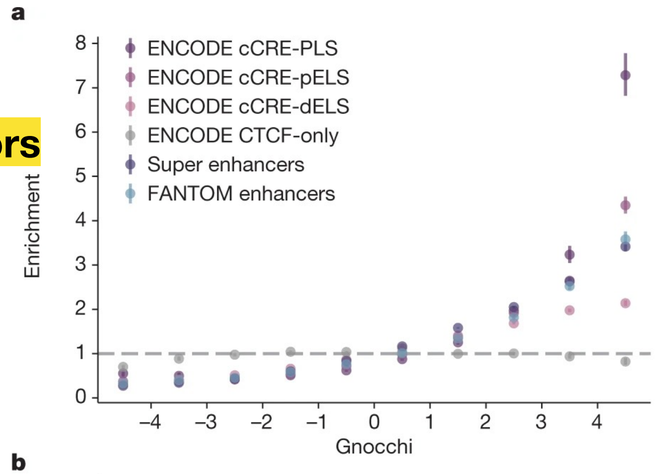
Gut checks

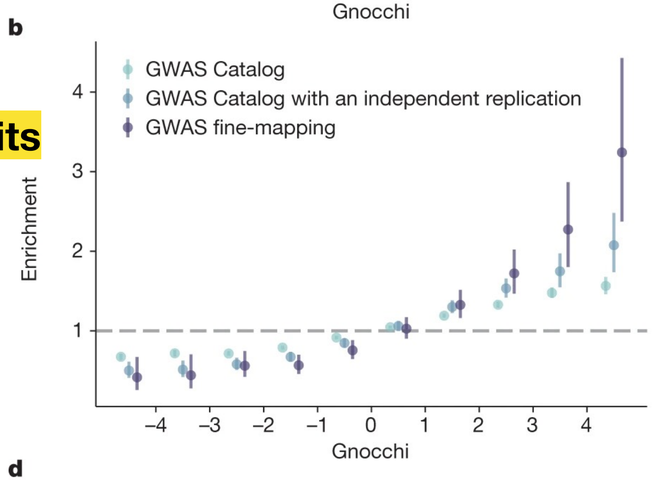


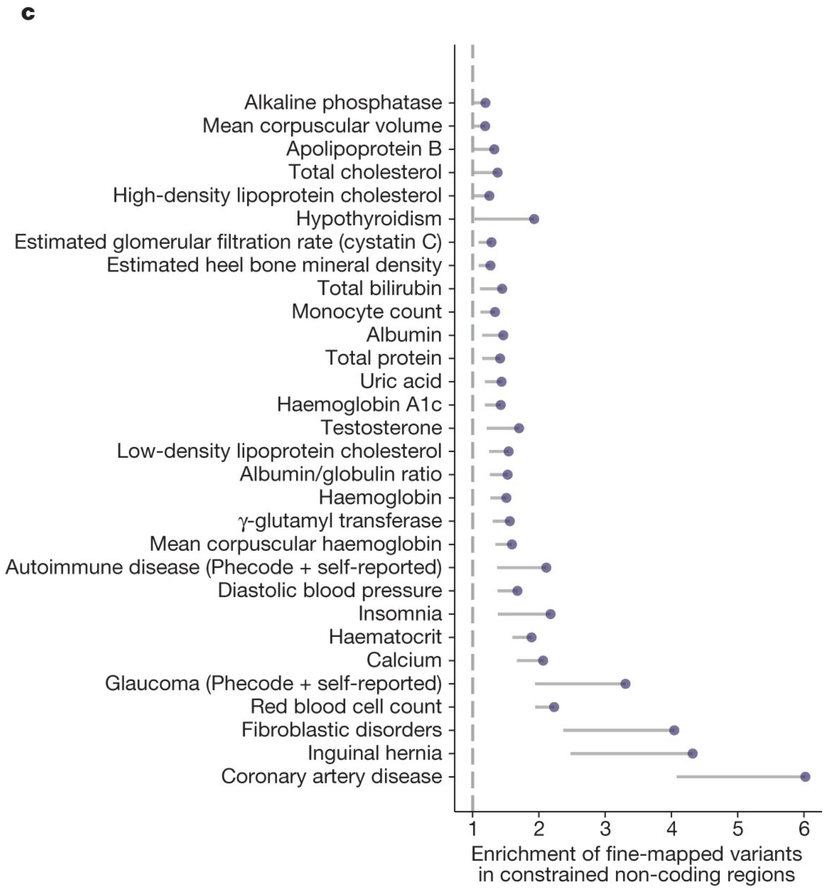
This is a structural variation constraint metric

Correlates with regulators

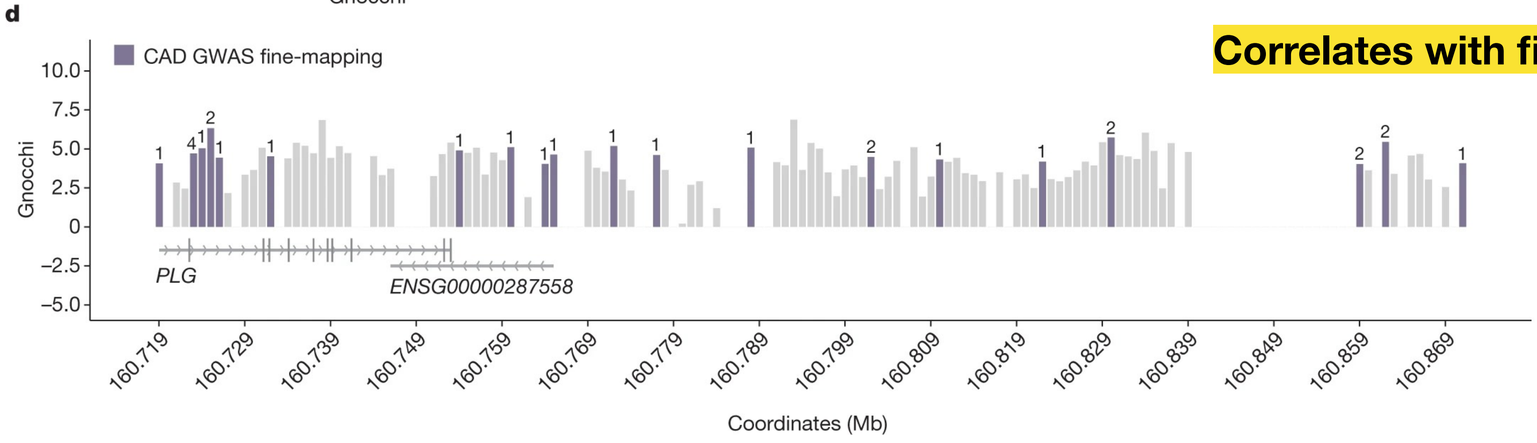
Correlates with GWAS hits



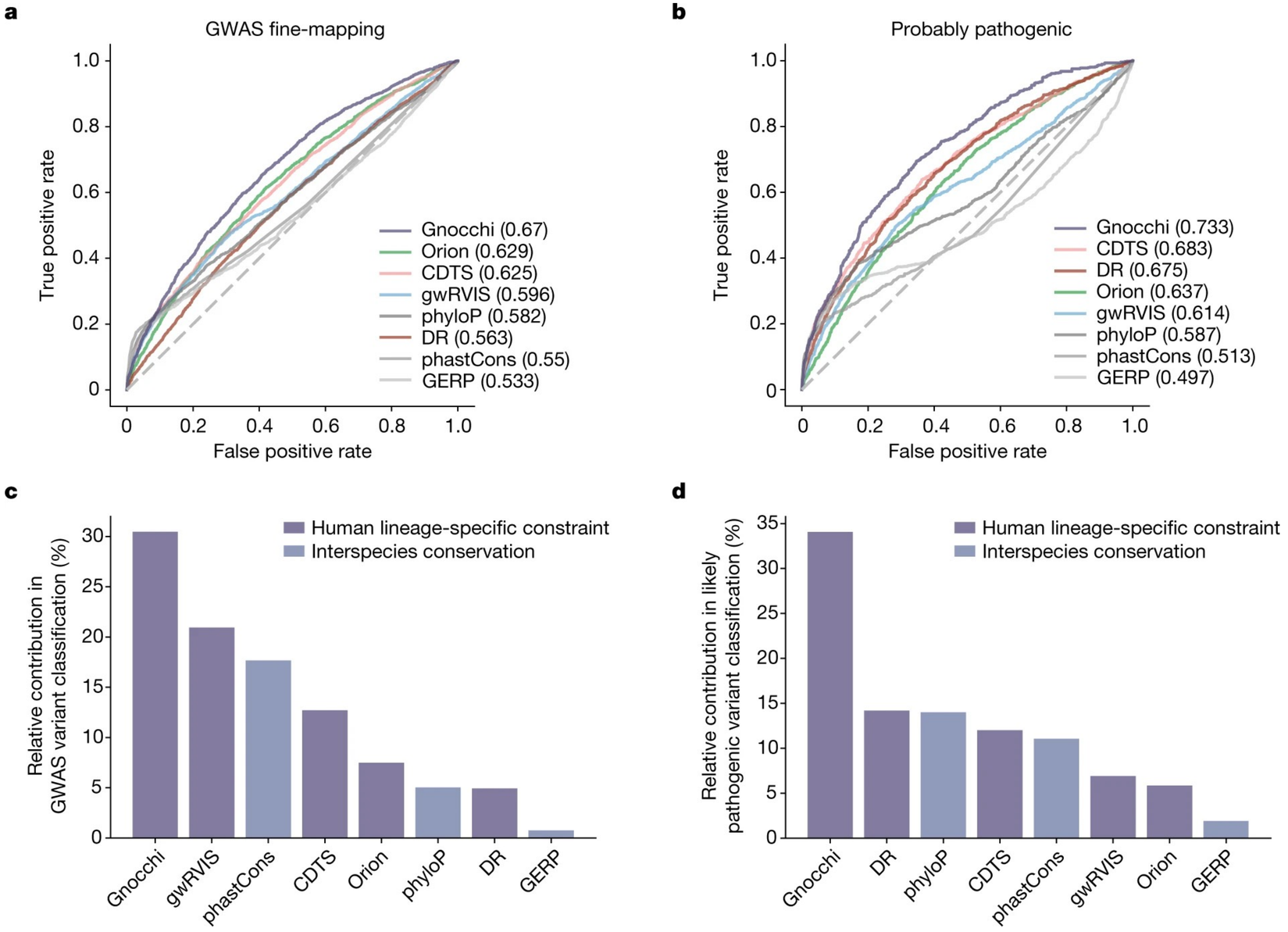




Correlates with fine mapping results



Best in class performance at functional prediction



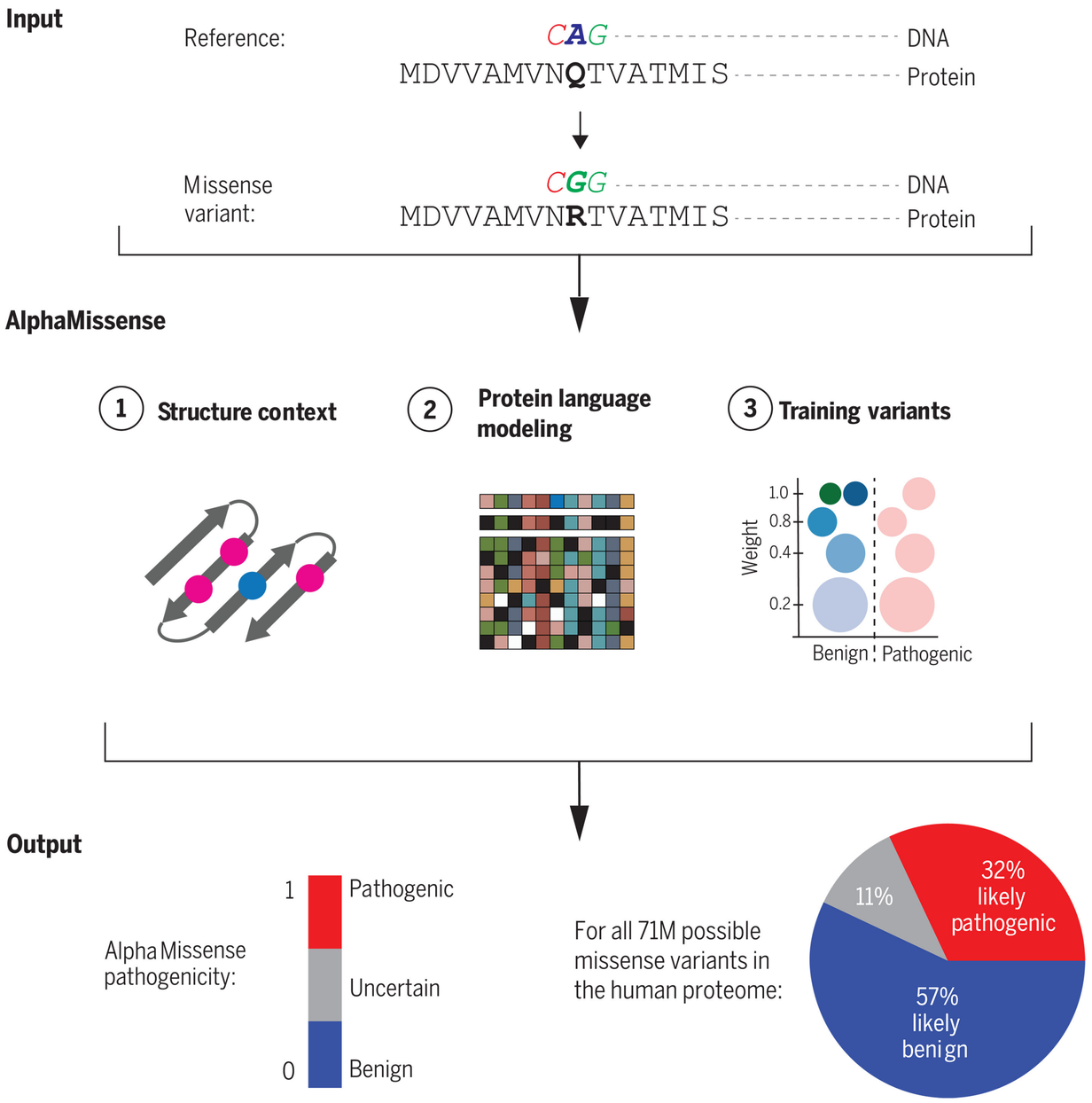
Accurate proteome-wide missense variant effect prediction with AlphaMissense (Cheng et al, *Science*)

- Goal: Classify all missense variants and pathogenic or benign
- Method:
  - Unsupervised protein language modeling
  - Incorporate structural information from AlphaFold
  - Fine-tuning on weak labels form pop frequency data
- Result:
  - State-of-the-art missense pathogenicity classification
  - 32% of all missense variants are likely pathogenic, 57% benign
- Conclusion: Huge resource for the genetics and molbio community if predictions hold up to prospective testing

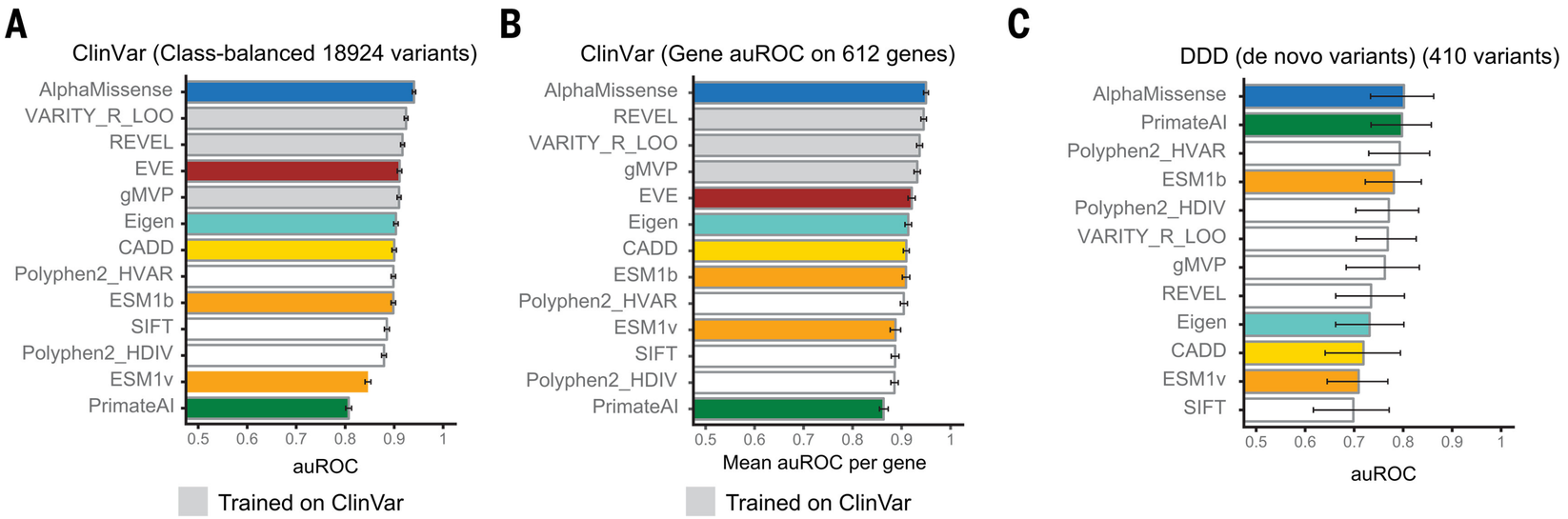
Protein sequences as input

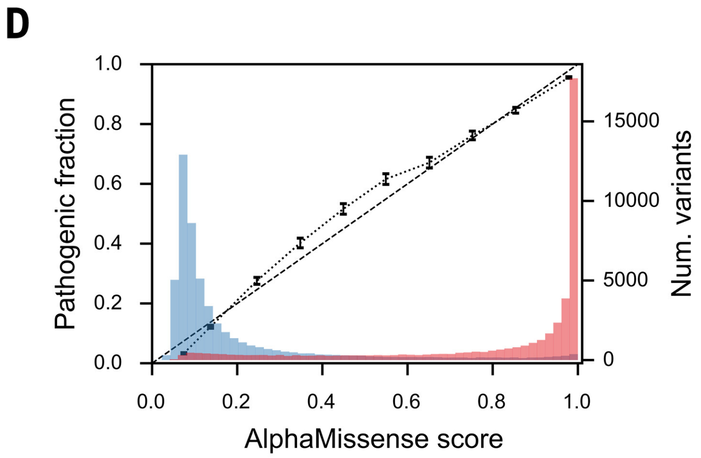
- **Incorporate**
- **Structure**
- **Language modeling**
- **Weak labeling**

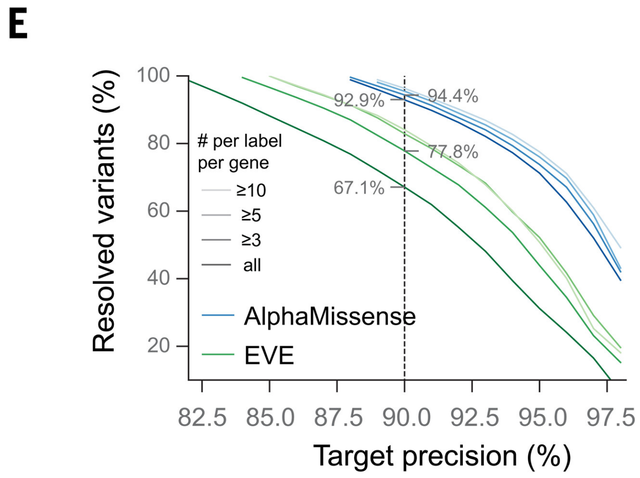
Make predictions



Best in class performance



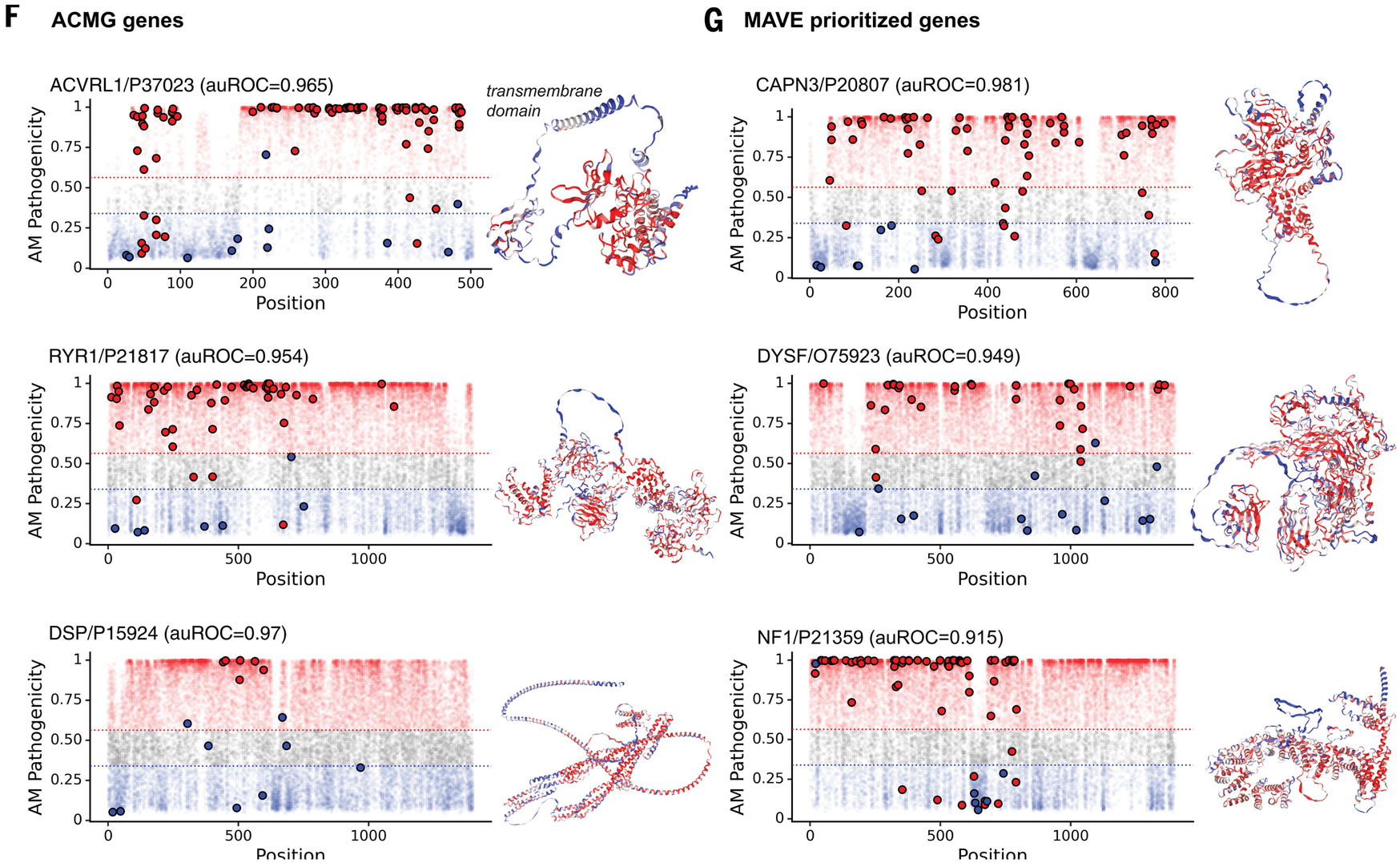




Only “call” variants with expected precision of ≥90%

Best in class performance

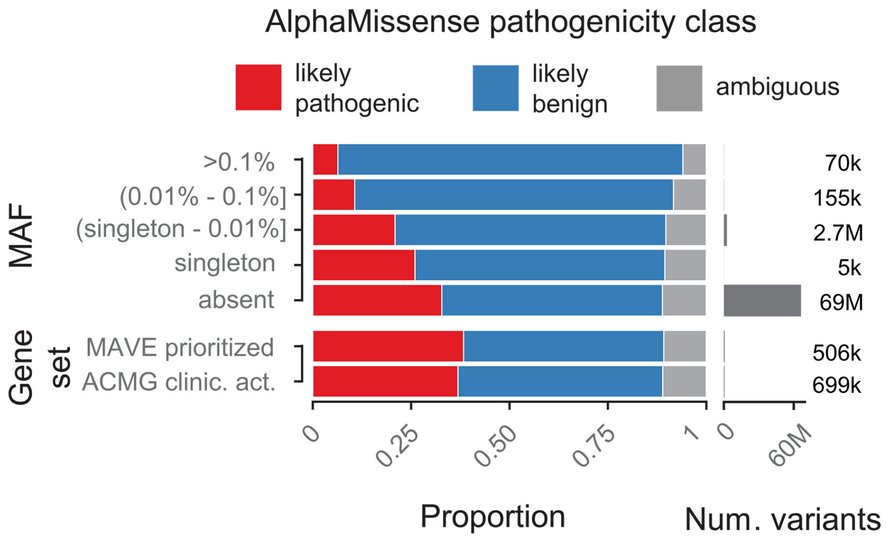
Residue level pathogenicity predictions



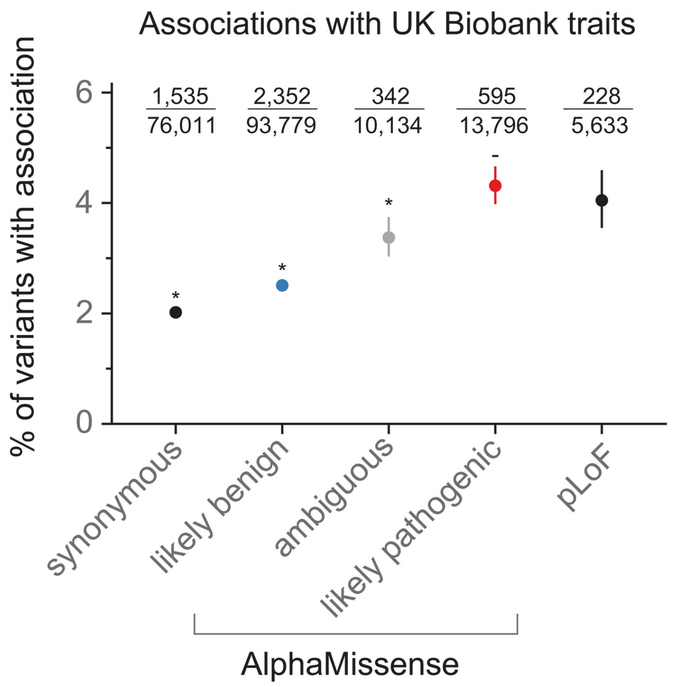
A

CHROM	POS	REF	ALT	genome	uniprot_id	transcript_id	protein_variant	am_pathogenicity	am_class
chr7	44149825	T	C	hg38	P35557	ENST00000403799	D205G	0.999516	likely_pathogenic
...	...	...	...	...	...	...	...	...	...

B



C

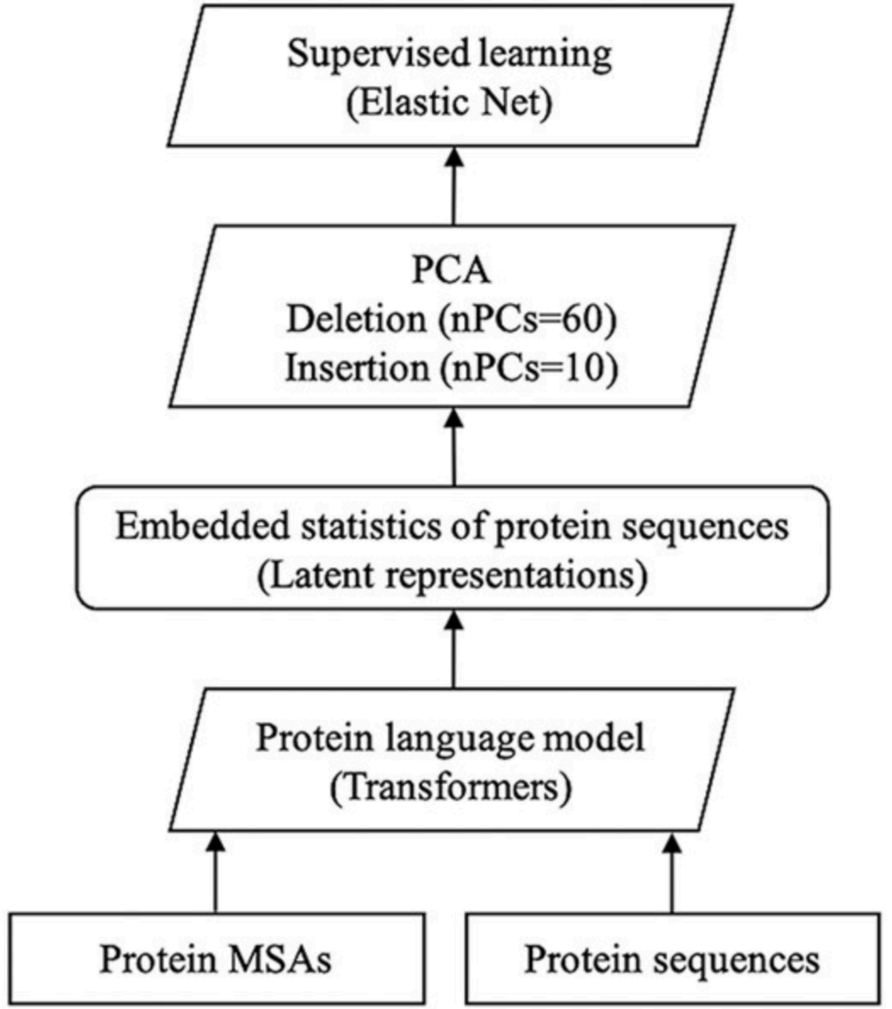


Likely pathogenic AlphaMissense variants are more likely to have GWAS associations

SHINE: protein language model-based pathogenicity prediction for short inframe insertion and deletion variants (Fan et al, *Briefings in Bioinformatics*)

- Goal: Accurate classify pathogenicity of inframe indels
  - *In-frame* indels are way more difficult to annotate for deleteriousness because they not as obviously damaging as frameshifts
  - However, they're still a major source of pathogenicity
- Method: transfer learning using protein-language models to perform few-shot classification
- Result: New state-of-the-art accuracy
- Conclusion: large protein-language models are powerful zero-shot and few-shot learning techniques for bio domains with little to no data!

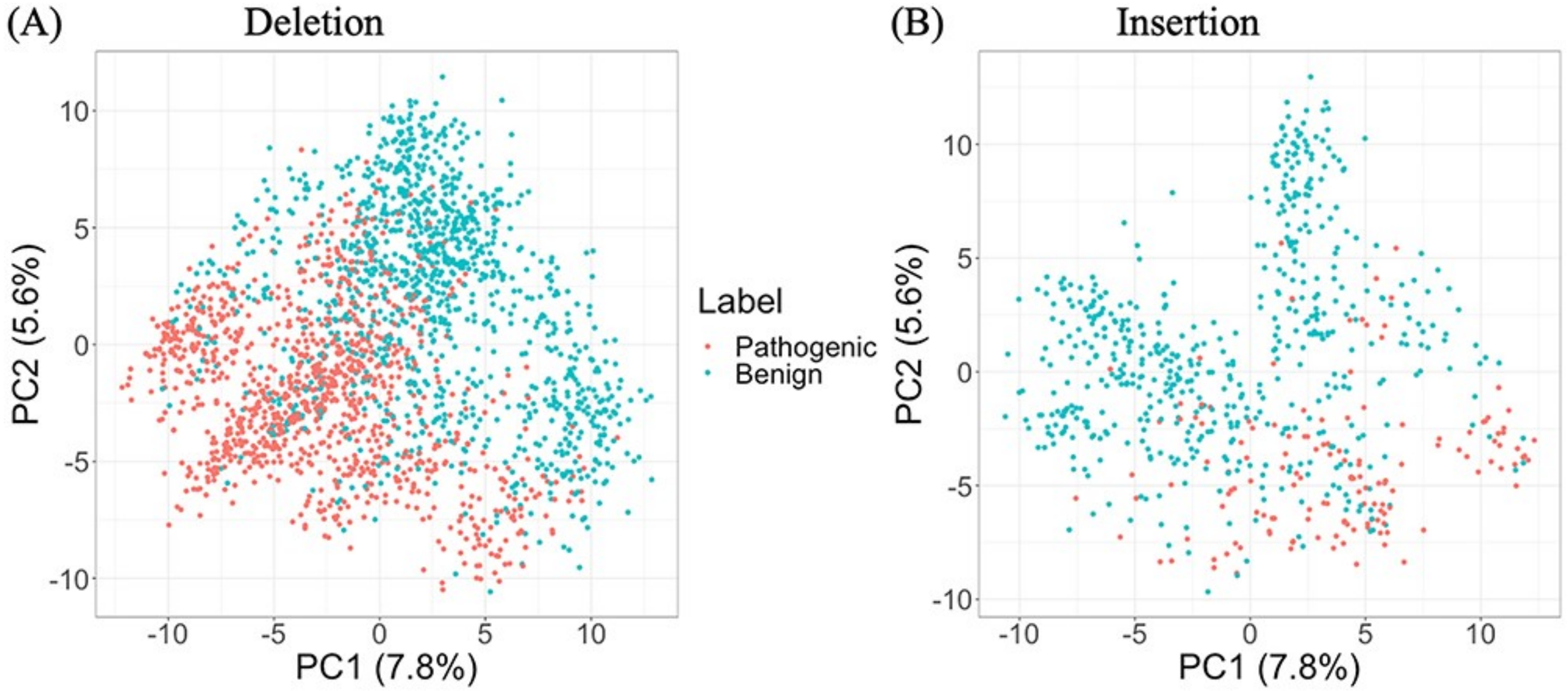
Figure 1

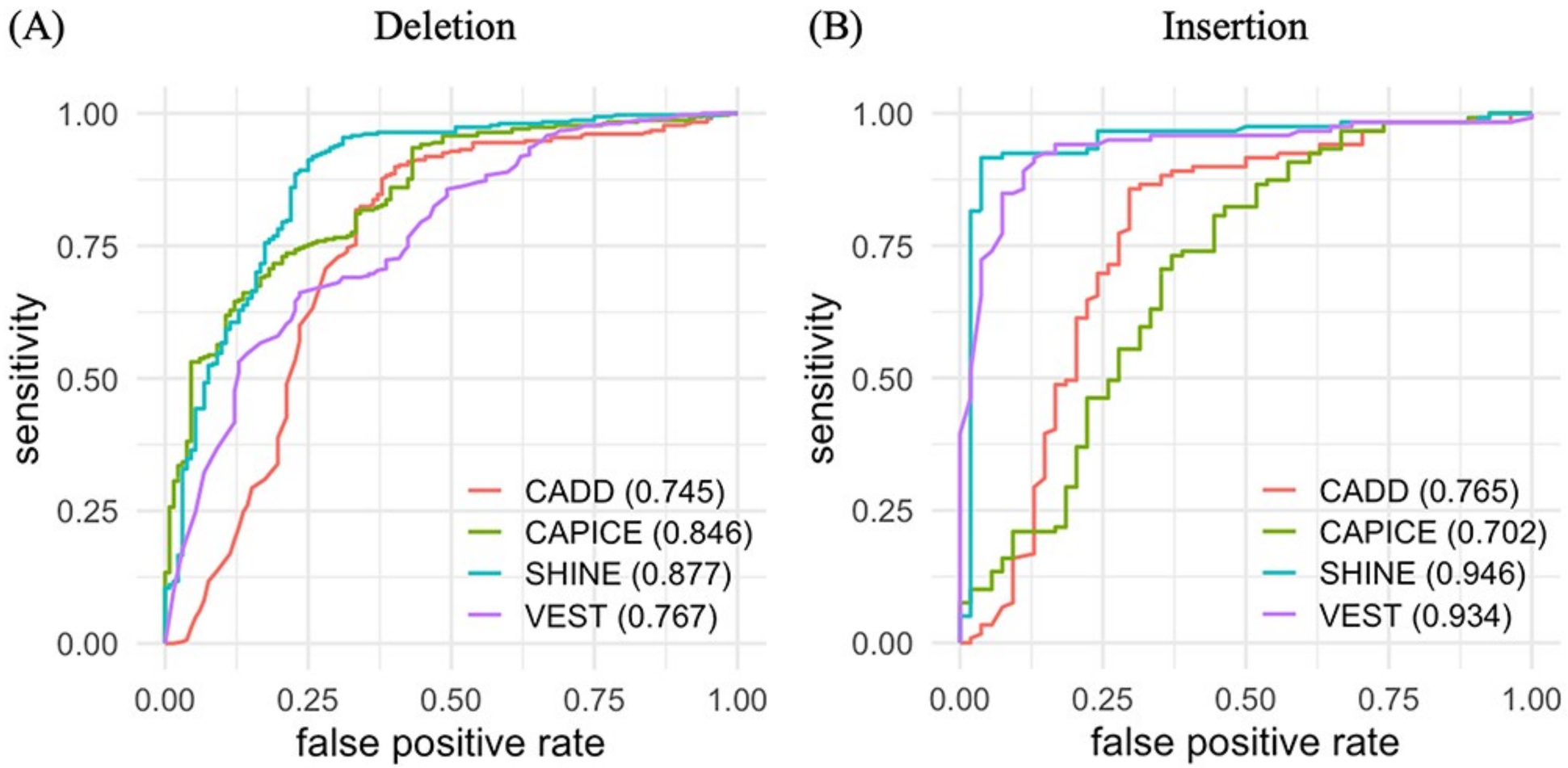


Small amount of training data —>

ESM-1b and MSA —>

PCA of embeddings show evidence of separation of pathogenic variants

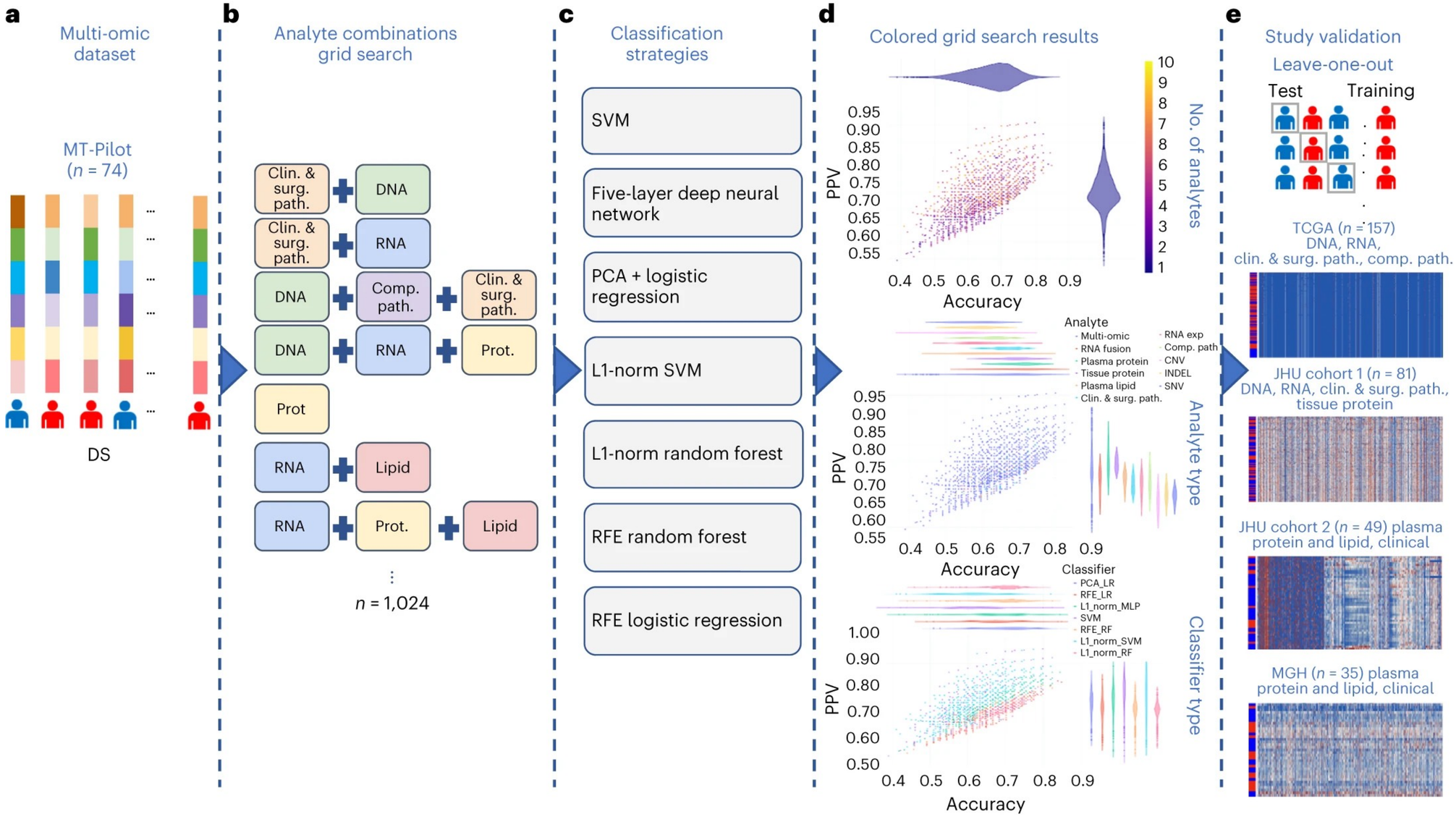



(A)
Deletion
sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
false positive rate
CADD (0.745)
CAPICE (0.846)
SHINE (0.877)
VEST (0.767)
(B)
Insertion
sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
false positive rate
CADD (0.765)
CAPICE (0.702)
SHINE (0.946)
VEST (0.934)

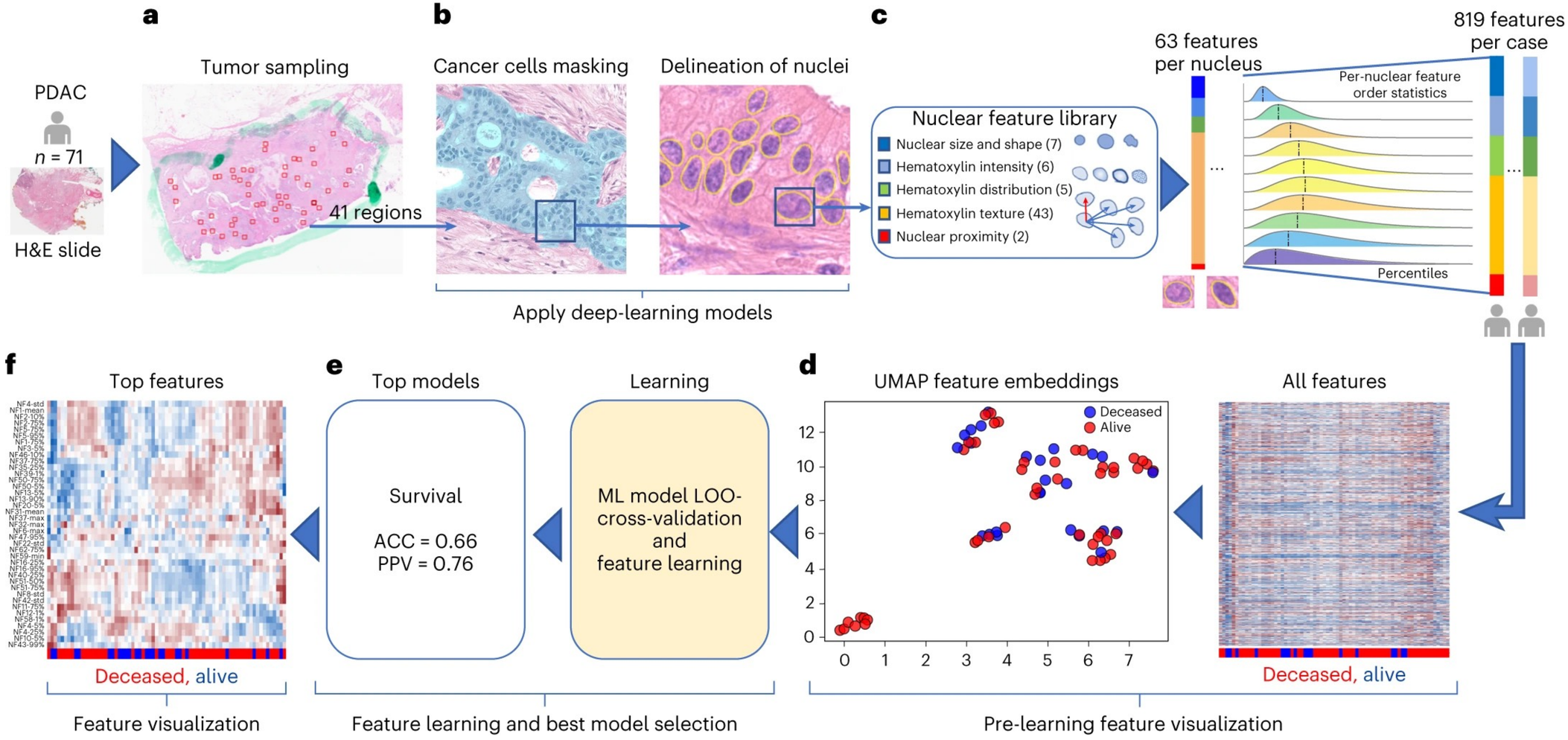
The Molecular Twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients (Osipov and Nikolic et al, *Nature Cancer*)

- Goal: Build precision medicine platform for predicting survival and treatment outcomes in pancreatic cancer
- Method: Integrative multiomics + ML
  - DNA, RNA, Proteomics, Lipidomics, Path Imaging, and Clinical Features
- Result:
  - Using all ~7k features performed best
  - A “most parsimonious” model of ~600 performed almost as well
- Conclusion:

Systematic exploring of feature and modeling space

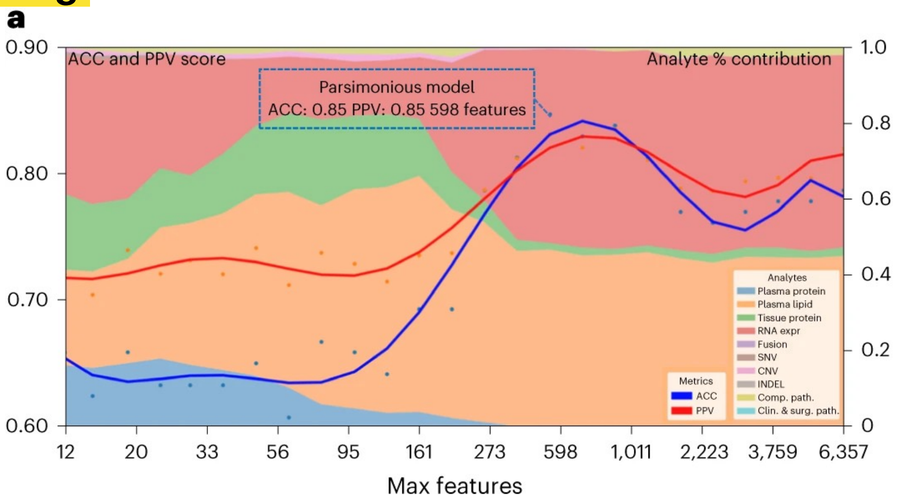


Integrated imaging modalities in addition to molecular and clinical

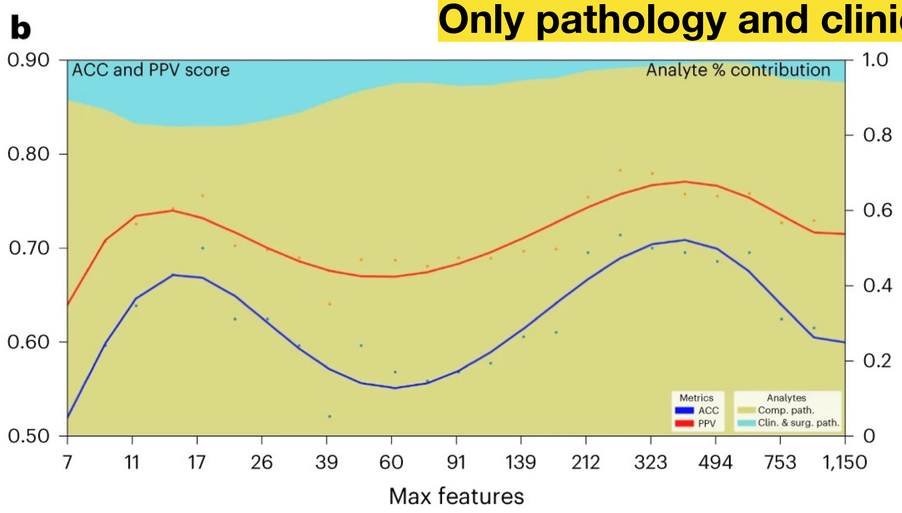


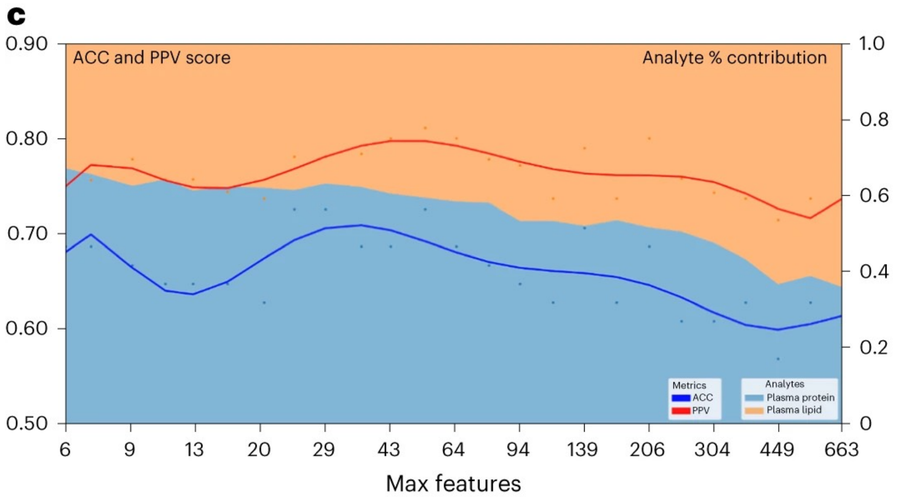
Explored different ways of finding high performing parsimonious model

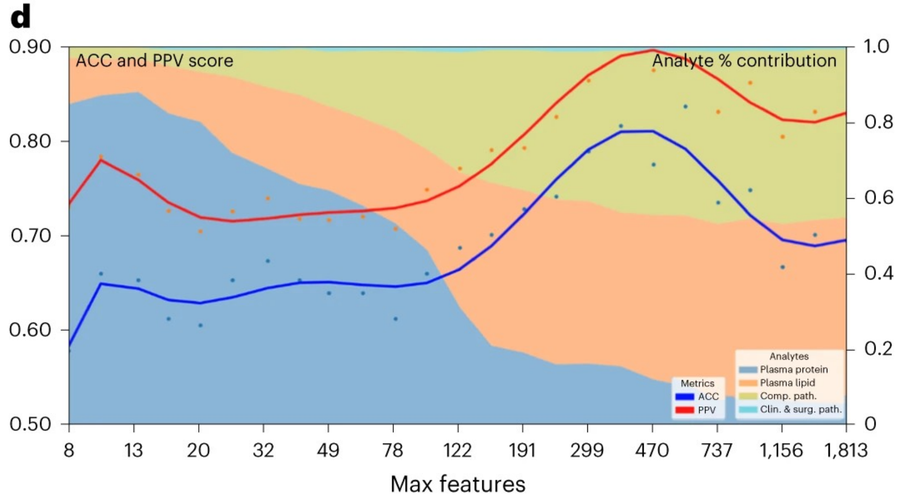
Everything



Only pathology and clinical data







Only protein and lipids

Protein, Lipids, path, and clinical



"Blinding Lights"

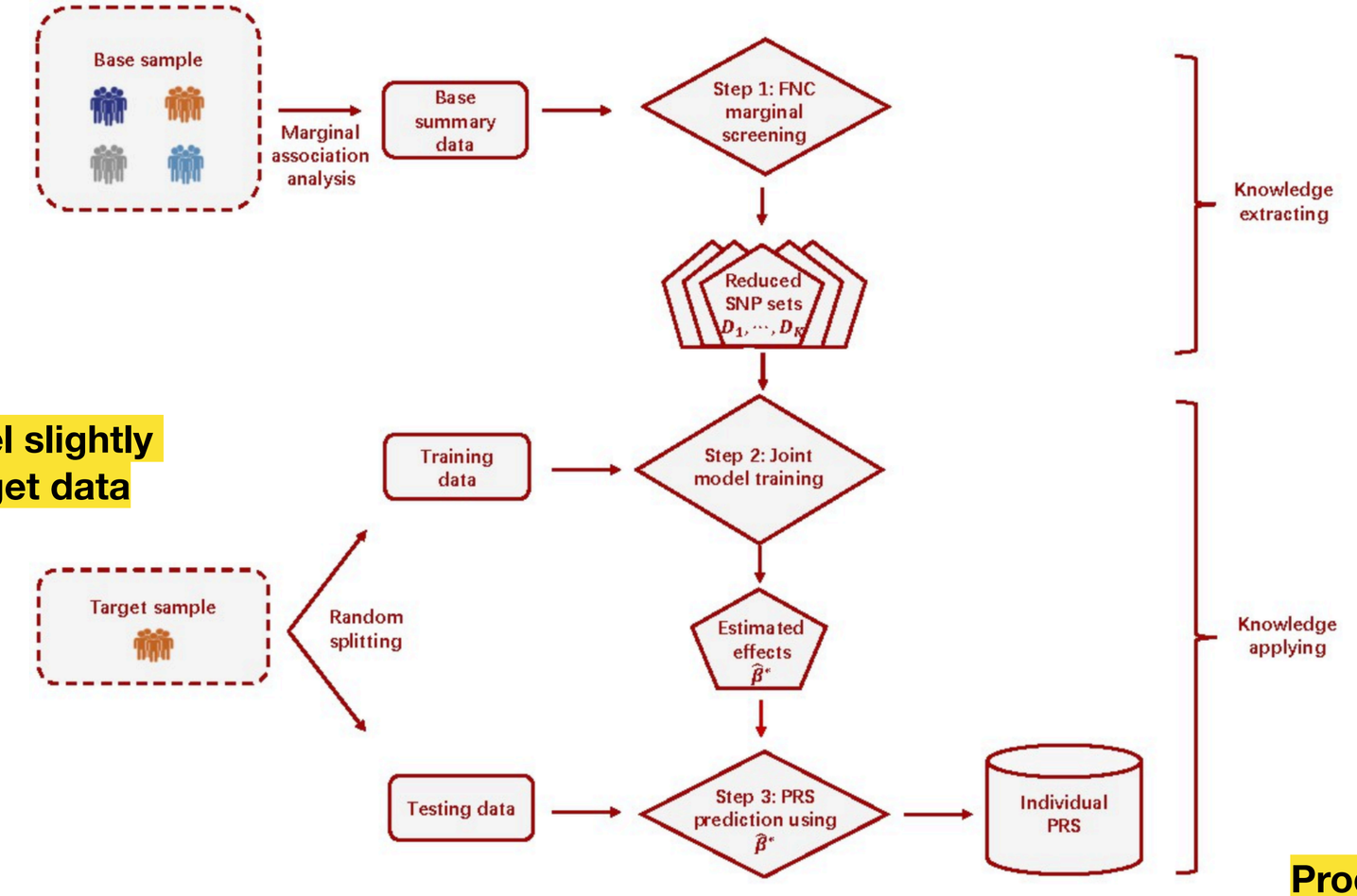
Shining a Light on Genomics and Genetics

 https://doi.org/10.1371/journal.pgen.1010597

Transfer learning with false negative control improves polygenic risk prediction (Jeng et al, *PLOS Genetics*)

- Goal: address poor generalizability of PRS when they're used outside of the population the GWAS
- Method:
  - Transfer learning with "false negative control" marginal screening
  - Basically use a little of the target population to customize the PRS
- Result:
  - Works great in simulations; pretty good in actual data
- Conclusion:

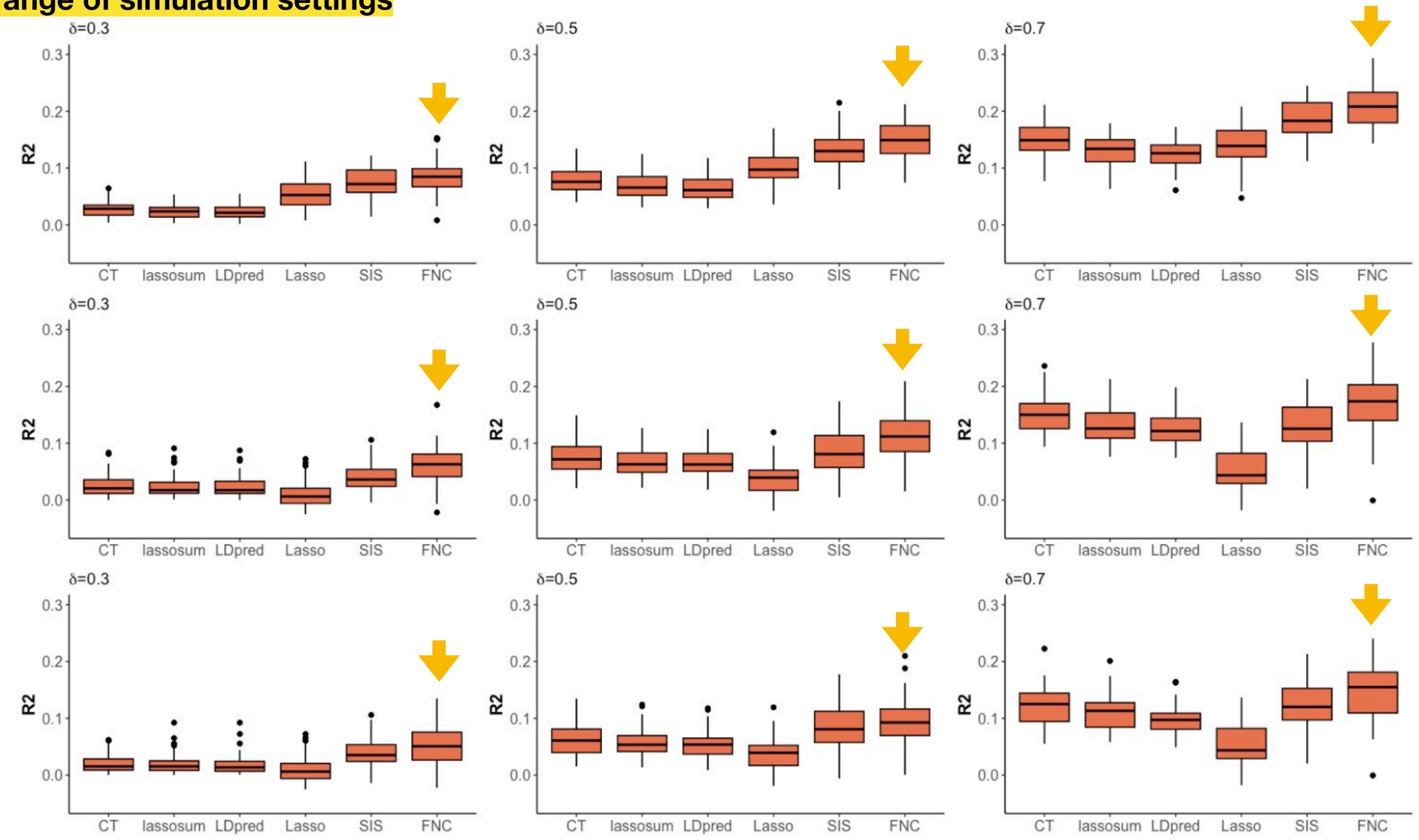
GWAS summary stats from large resource (e.g. UKB)



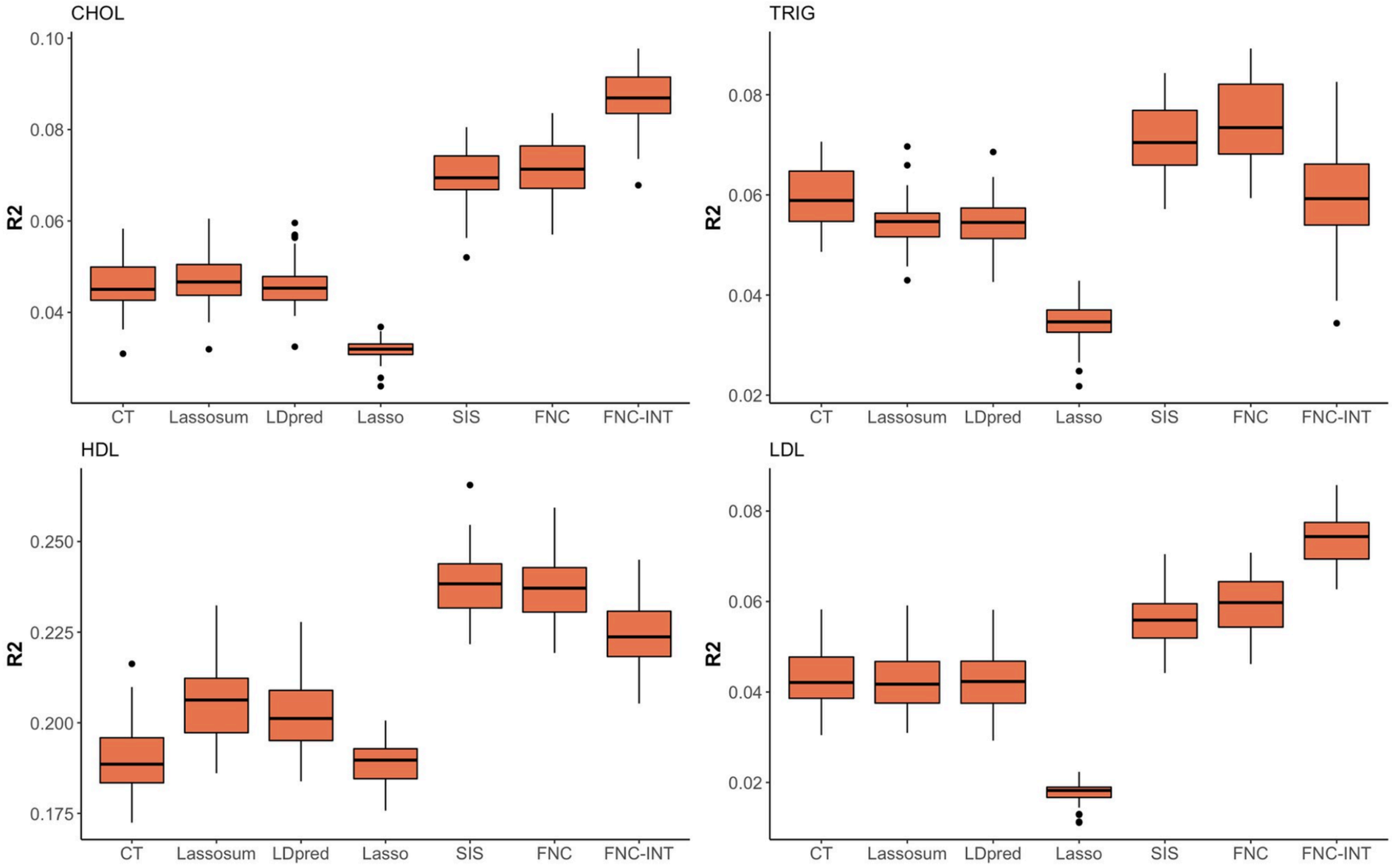
Update the model slightly using some target data

Produce a better PRS for the target pop

Great performance across a wide range of simulation settings



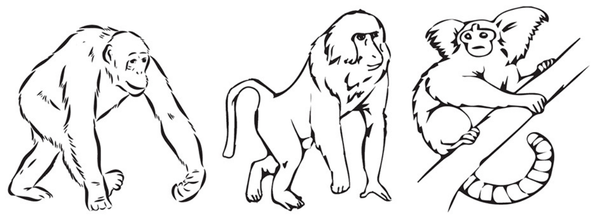
Pretty good for real data, too



Rare penetrant mutations confer severe risk of common diseases (Fizzier and McRae et al, *Science*)

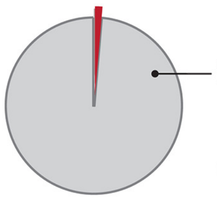
- Goal: Evaluate the risk of rare variants in genome-wide association studies
- Method:
  - Developed PrimateAI-3D - 3D CNN trained on sequencing data from 233 primate species
  - Apply to 450k UKB invidiuals to improve rare-variant association tests for 90 traits
- Result:
  - Improved gene discovery (73% more gene-phenotype associations)
  - Associates better supported by orthogonal genetic evidence
  - Developed some a rare-variant PRS
- Conclusion: The key to understanding human variation may lie within the diversity of our primate cousins

From their other Science paper…

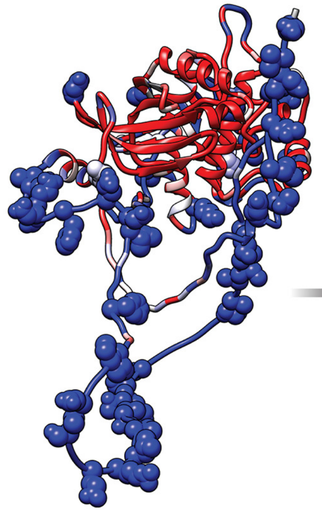


4.3 million common benign variants from 233 primate species

Validation of primate variants in human clinical variant database

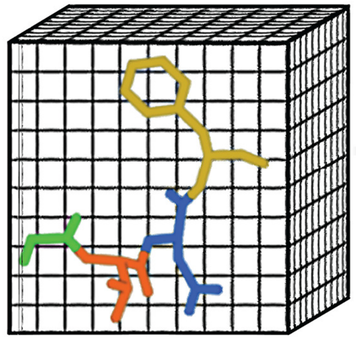


98.7% of common primate variants in ClinVar are benign

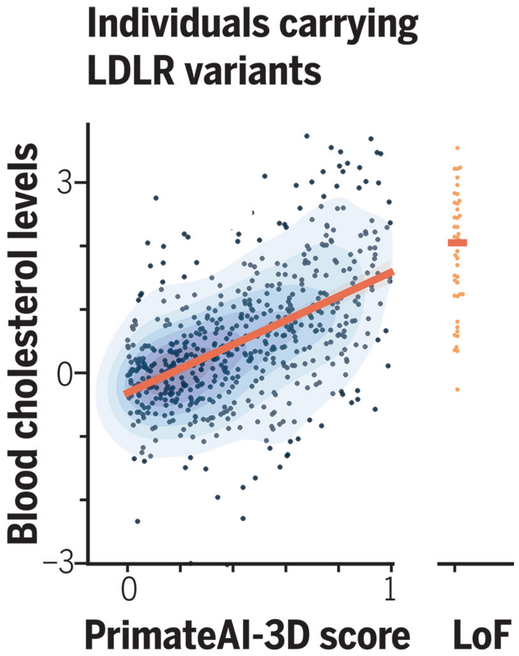


Benign primate variants superimposed on 3D protein structures

Train on 3D protein structures (CNNs)



3D convolutions + deep learning protein language models

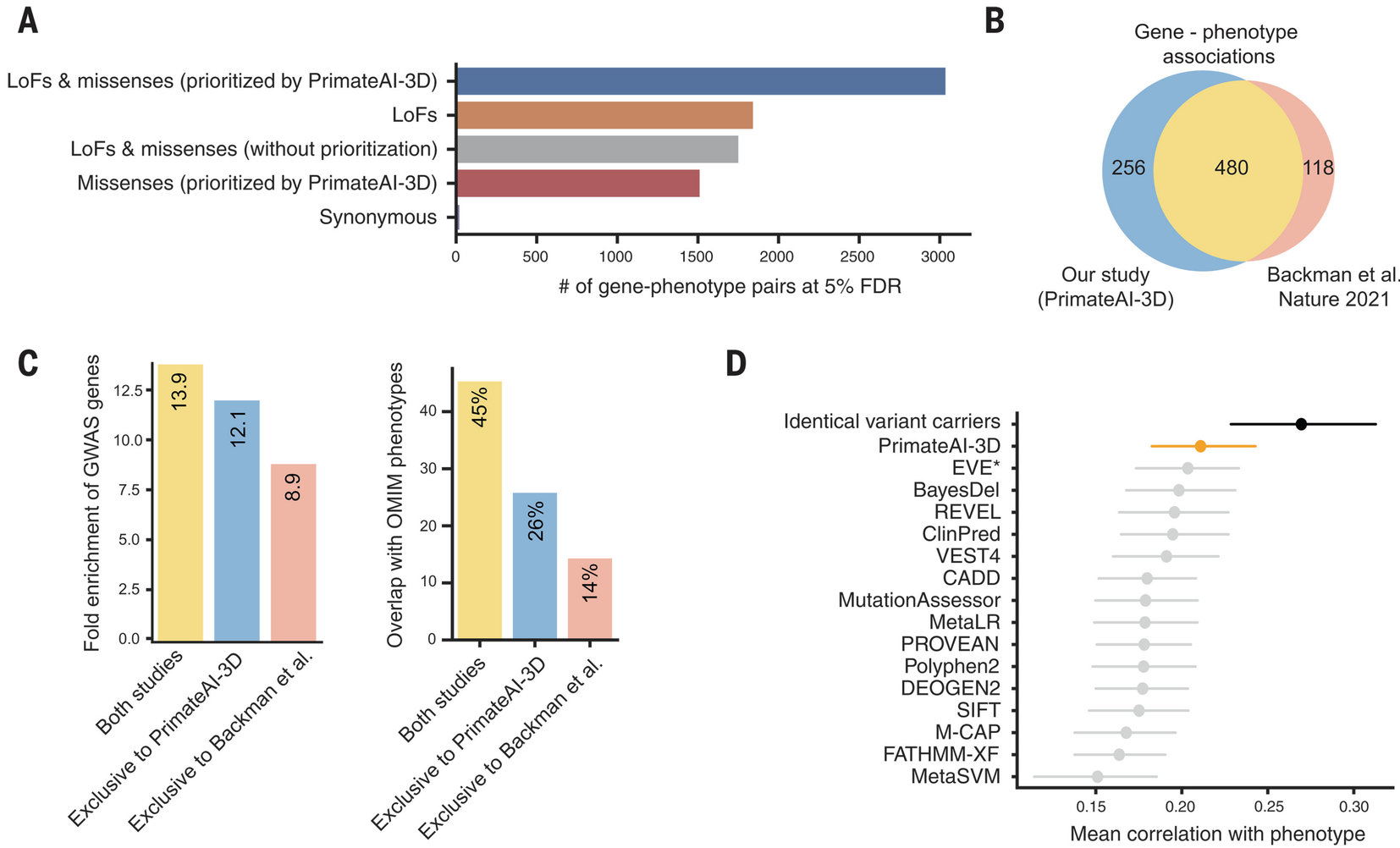


Validation of variant effect predictions in clinical cohorts

Observed variation in primates is benign in humans

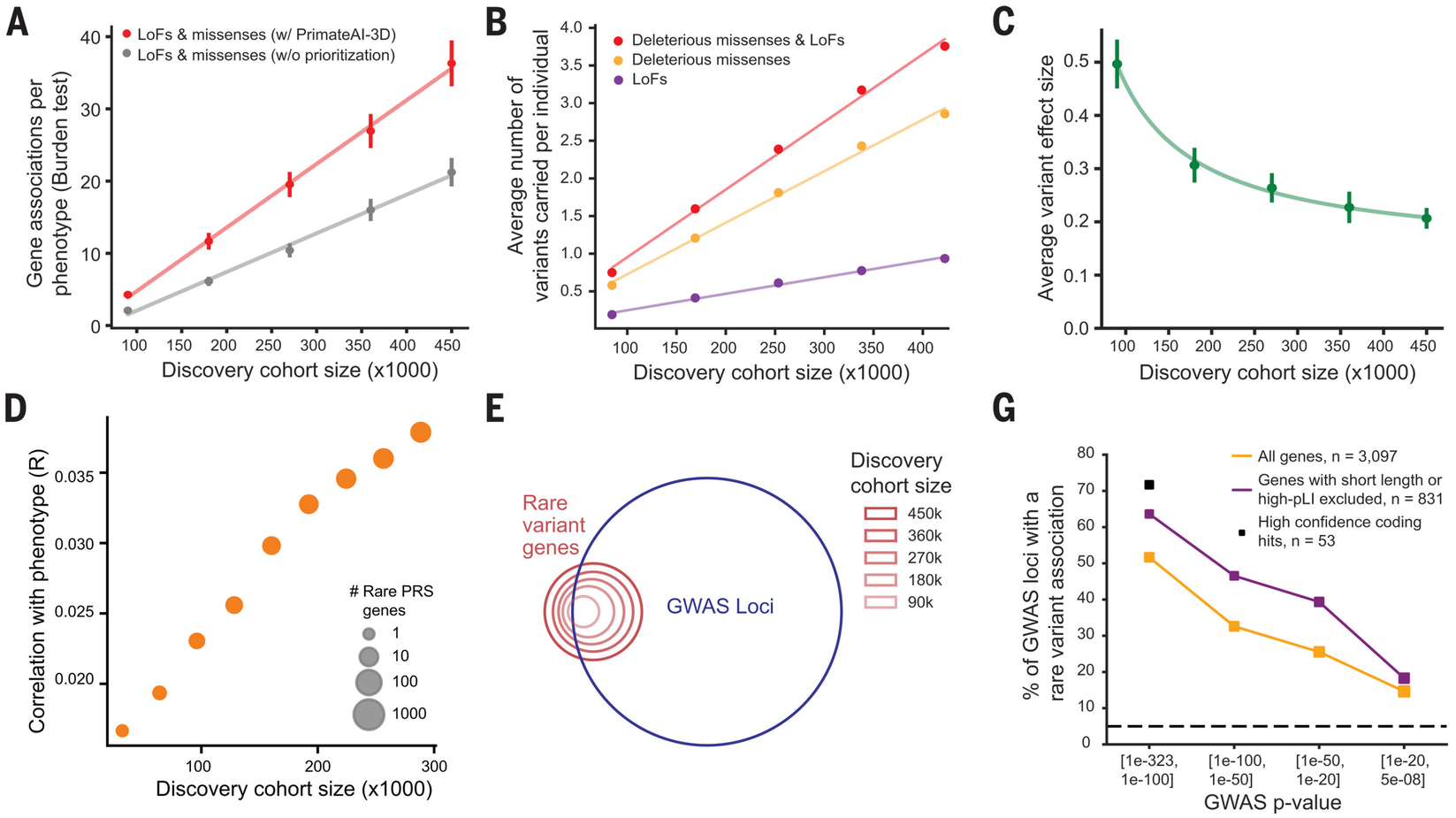
https://www.science.org/doi/full/10.1126/science.abn8197

PrimateAI-3D prioritizes LoFs better



PrimateAI-3D correlates with phenotypes better

Can forecast expected rare-variant discoveries as cohort sizes grow



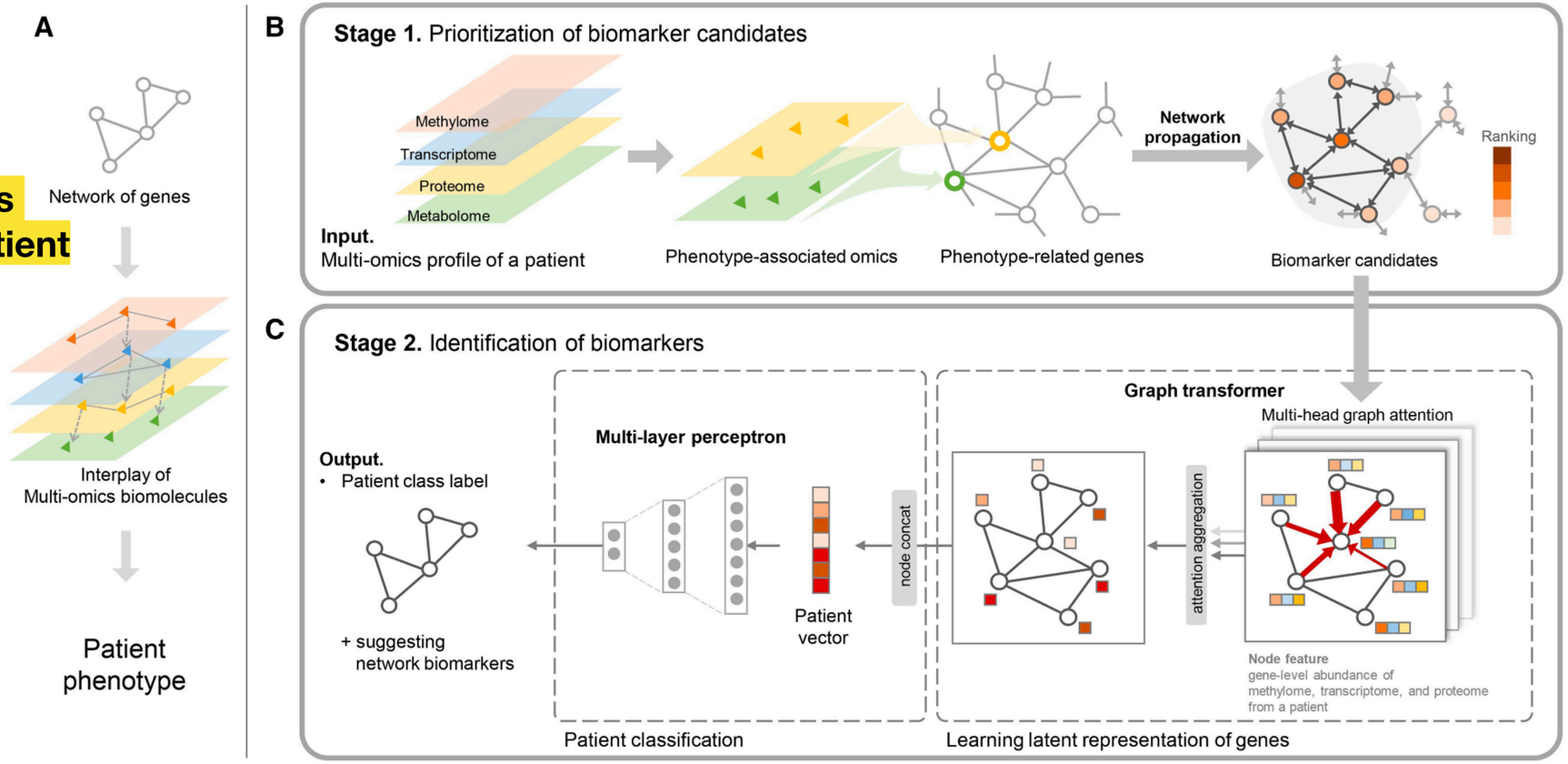
Will help to design future GWAS

GOAT: Gene-level biomarker discovery from multi-Omics data using graph ATtention neural network for eosinophilic asthma subtype (Jeong et al, *Bioinformatics*)

- Motivation:
  - Multiomics analysis is especially powerful for heterogenous diseases, like asthma
  - Integrating these data to find new biomarkers remains a challenge
- Method: Deep attention model named GOAT that can produce mechanistic hypotheses
- Result: Found genes associated with asthma that wouldn't be found using individual omics analysis
- Conclusion: Lots of strategies for multiomics data integration, loving what I'm seeing from this field.

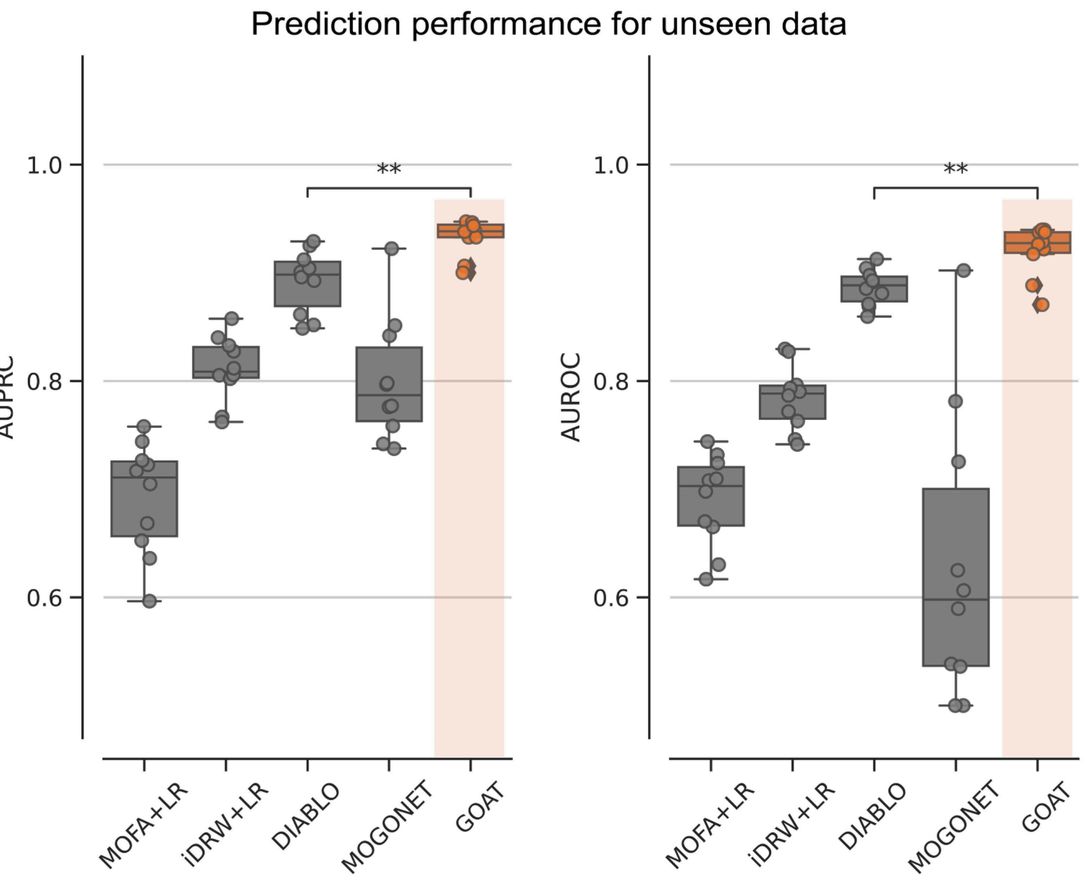
Unsupervised stage

Model of what's happening in a patient

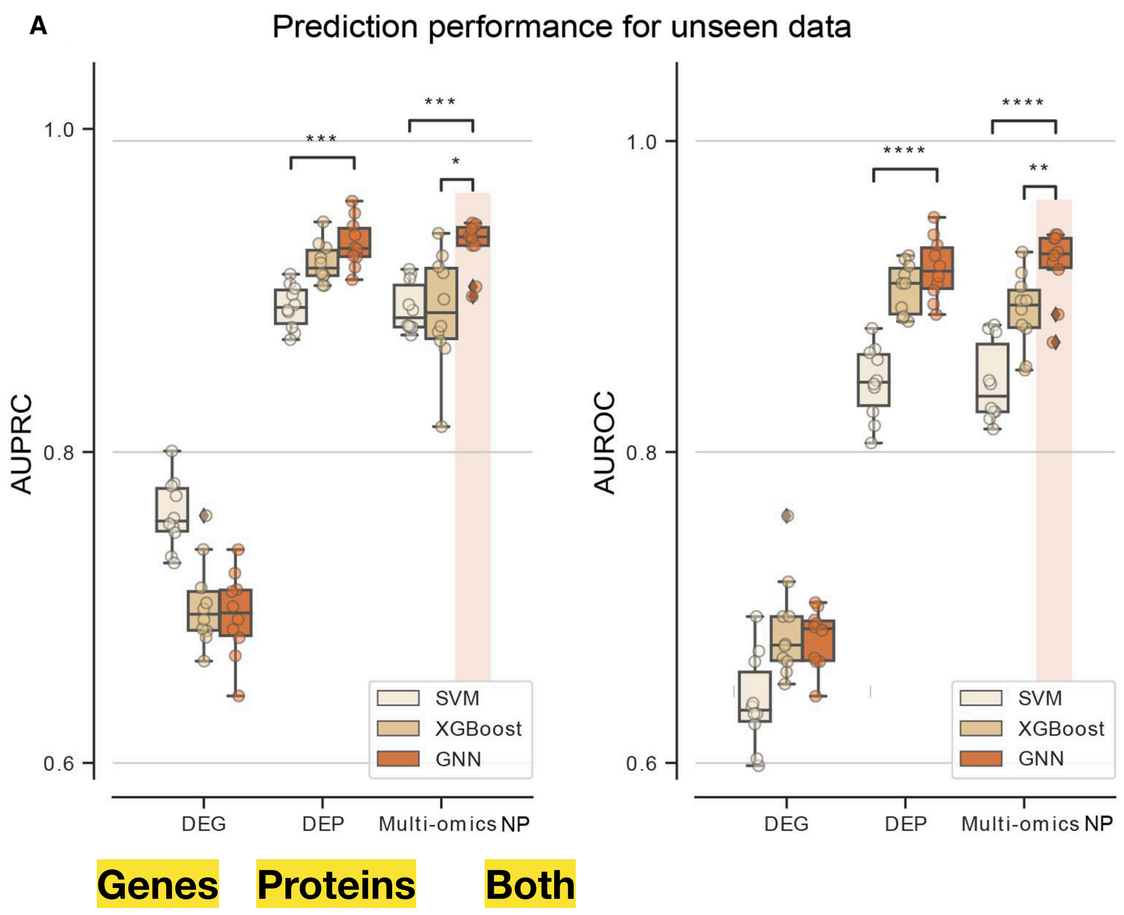


Supervised stage

Outperforms other methods

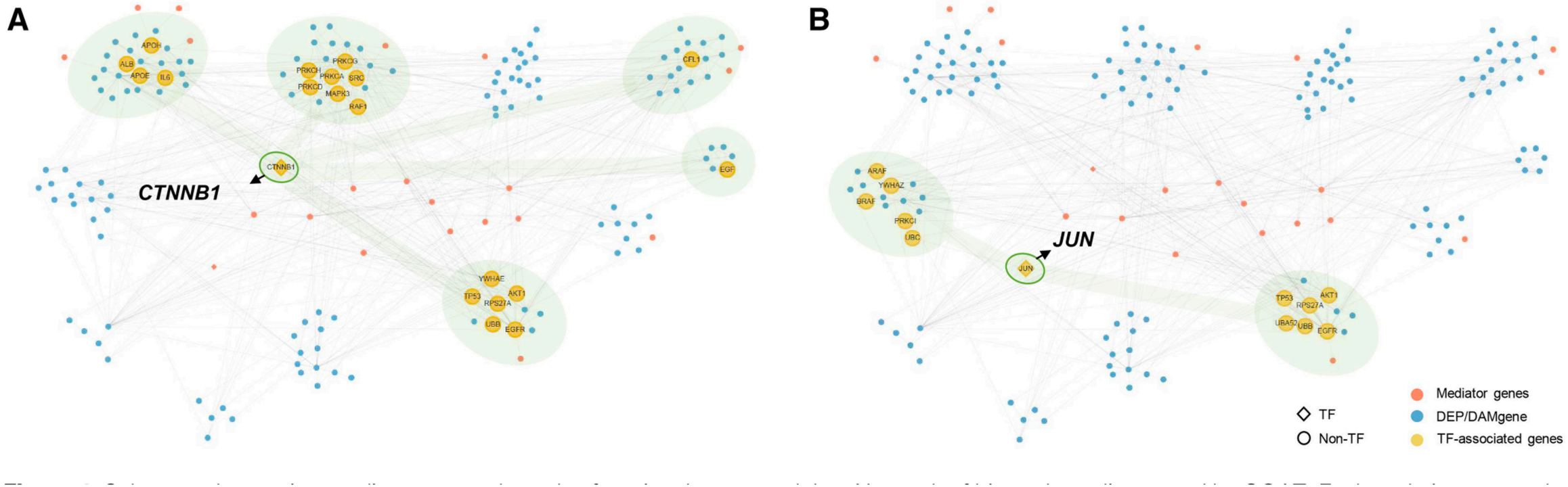


GOAT integrates multimodal data best



Genes **Proteins** **Both**

Patient-specific biomarker hypotheses

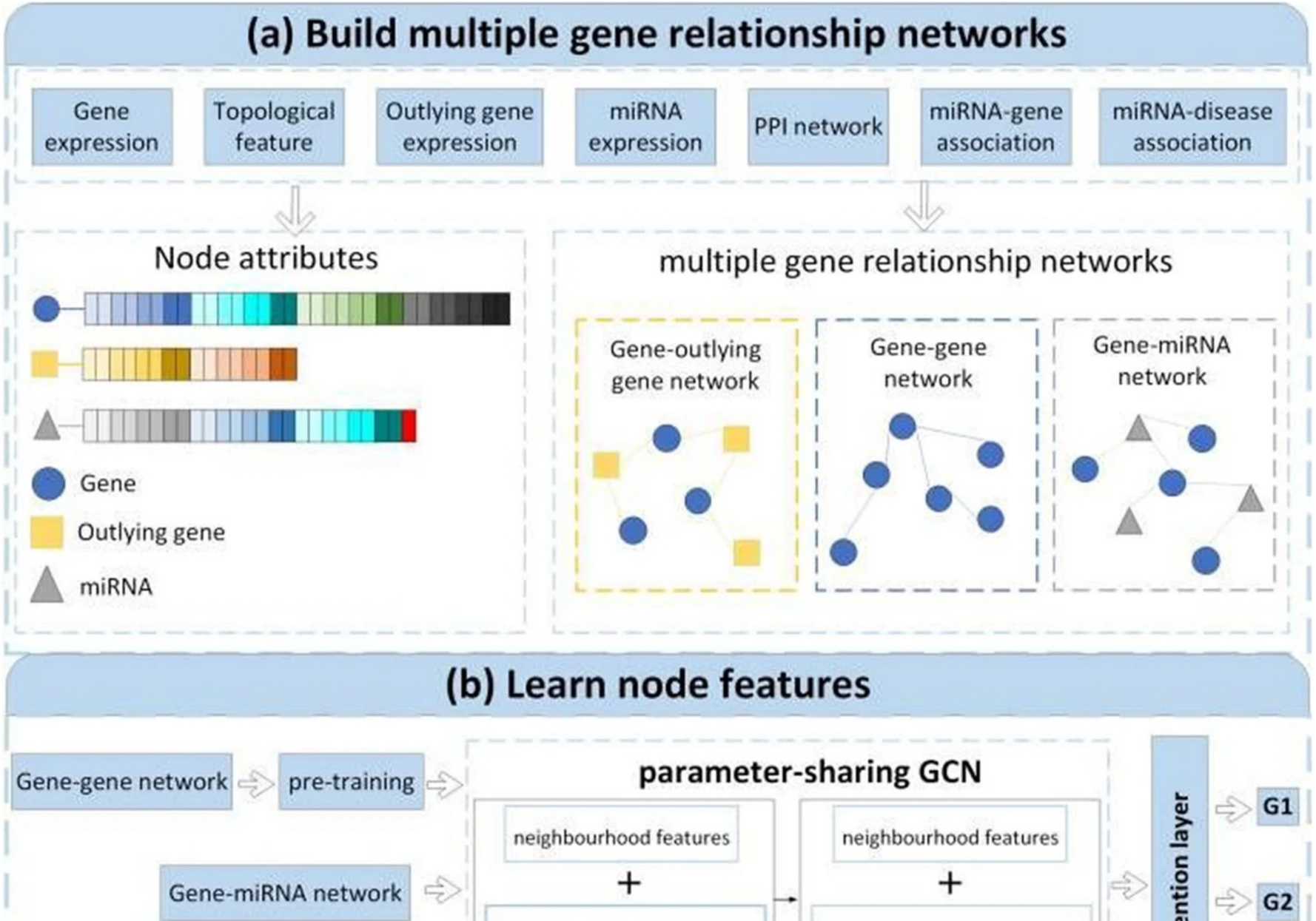


 https://doi.org/10.1186/s12859-023-05140-3

Identifying cancer driver genes based on multi-view heterogeneous graph convolutional network and self-attention mechanism (Peng et al, *BMC Bioinformatics*)

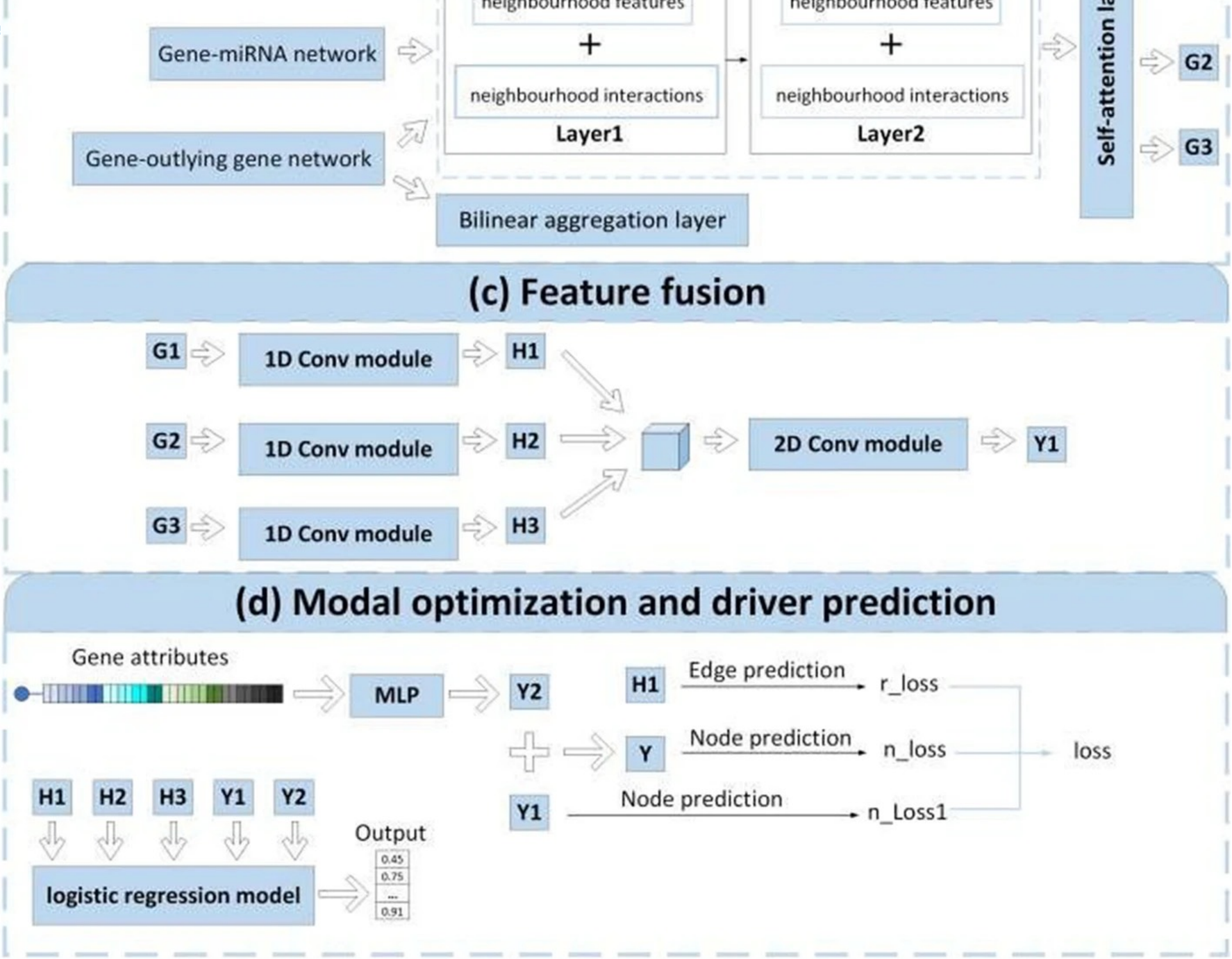
- Goal: More accurately identify cancer driver genes from heterogenous data
- Method:
  - Graph convolutional networks with self-attention mechanism on:
    - gene-gene, gene-outlying gene (i.e. regulation), and gene-miRNA networks
  - Use derived and original features together in a LR to predict driver genes
- Result:
  - Good performance on AUROC and AUPR and good robustness
  - Self-attention mechanisms allowed for structures, like protein complexes, to be identified automatically
- Conclusion: Self-attention (and other deep learning strategies) can be leveraged to discover biological complexity

Multimodal data represented as graphs



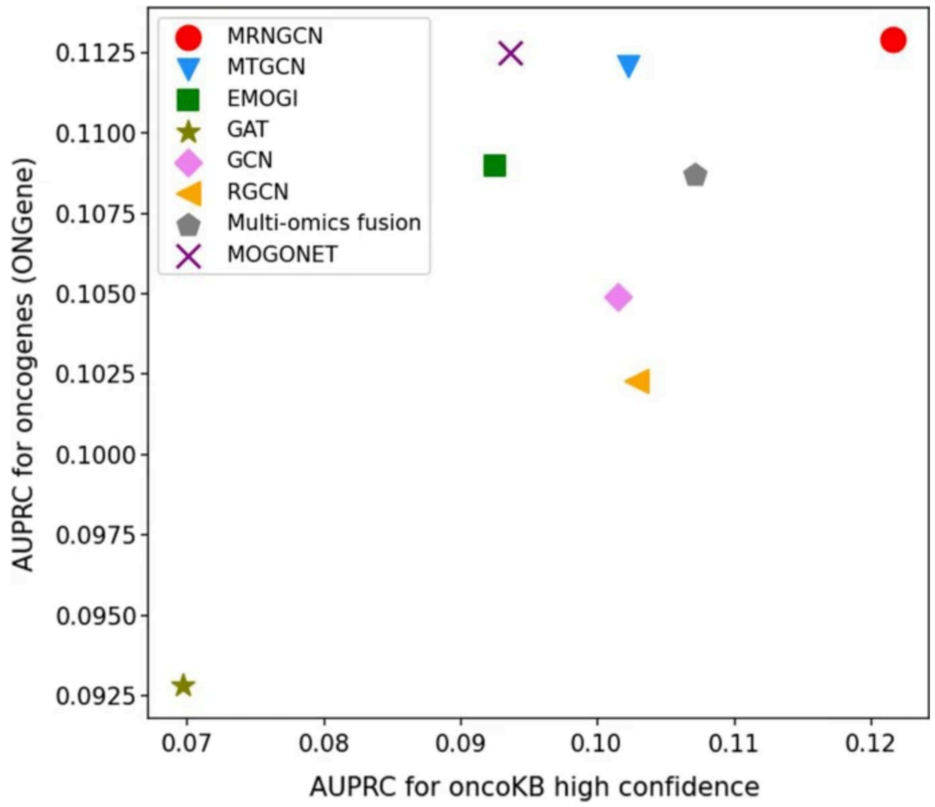
#IS24 #YIR24 @proftatone

Graph representations integrated to identify drivers



Best in class performance based on two different benchmarks

Fig. 2



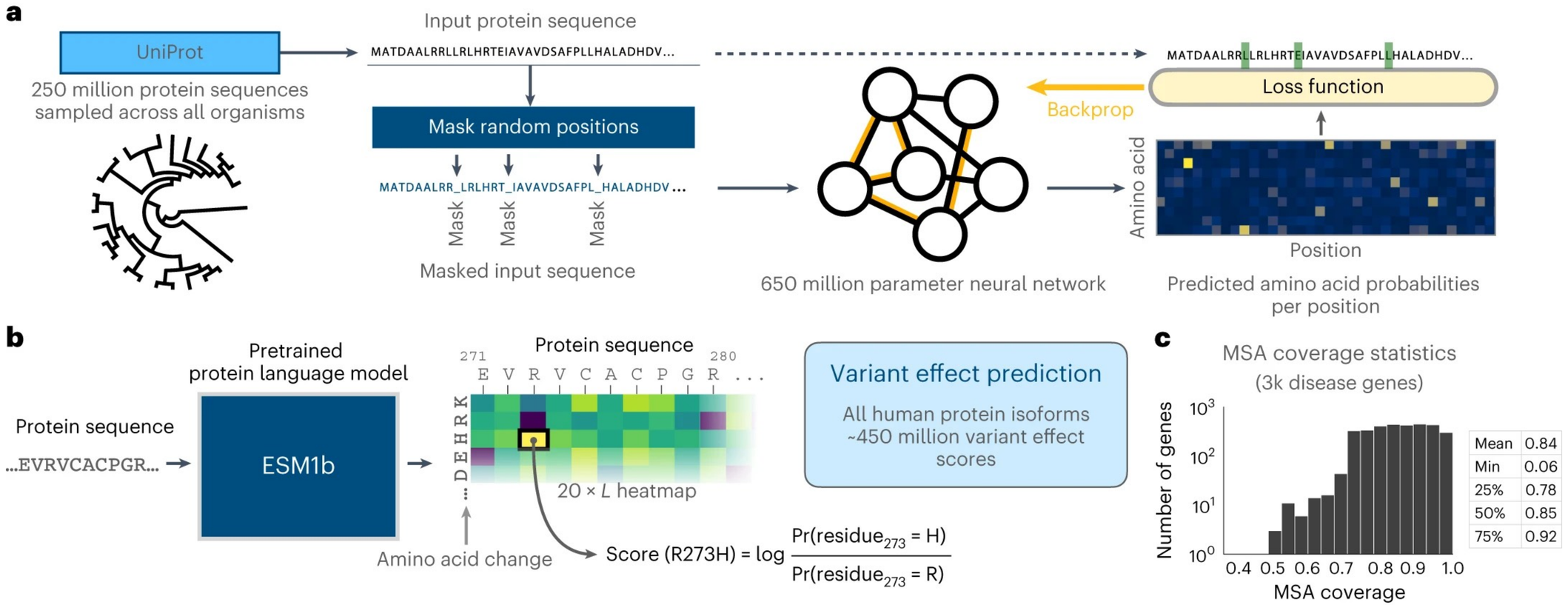
 https://doi.org/10.1038/s41588-023-01465-0

Genome-wide prediction of disease variant effects with a deep protein language model (Brandes et al, *Nature Genetics*)

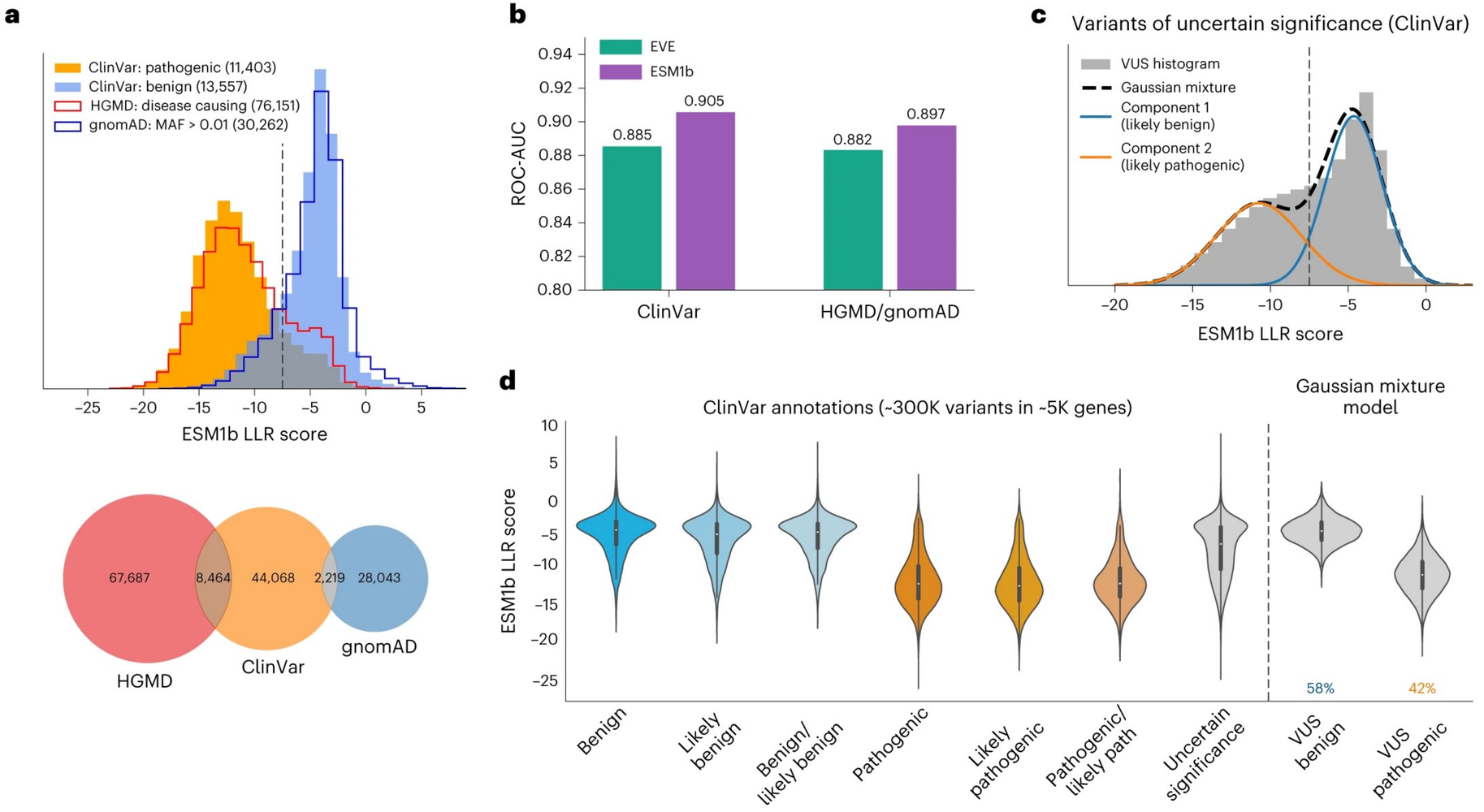
- Goal: Systematically predict clinical effects of 450 million missense variants
  - ESM-1b (trained on 250mm protein sequences) is a protein language model shown to implicitly learn structure and function, including long distance interactions and binding sites
- Method: Use ESM-1b Log Likelihood Ratios between WT and observe residue
- Result:
  - Big separation between benign and pathogenic clinvar variants in LLR values
  - Can even different when one isoform is pathogenic but another is not
- Conclusion: Evolution is a powerful animal.

https://github.com/facebookresearch/esm

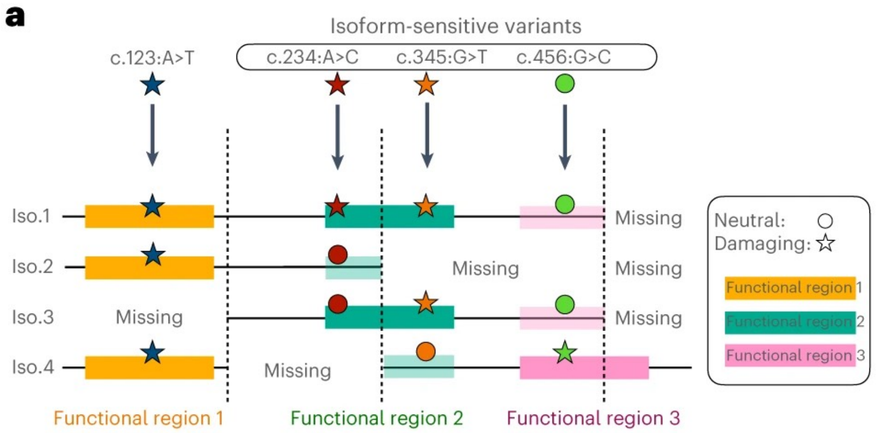
Recap of ESM-1b

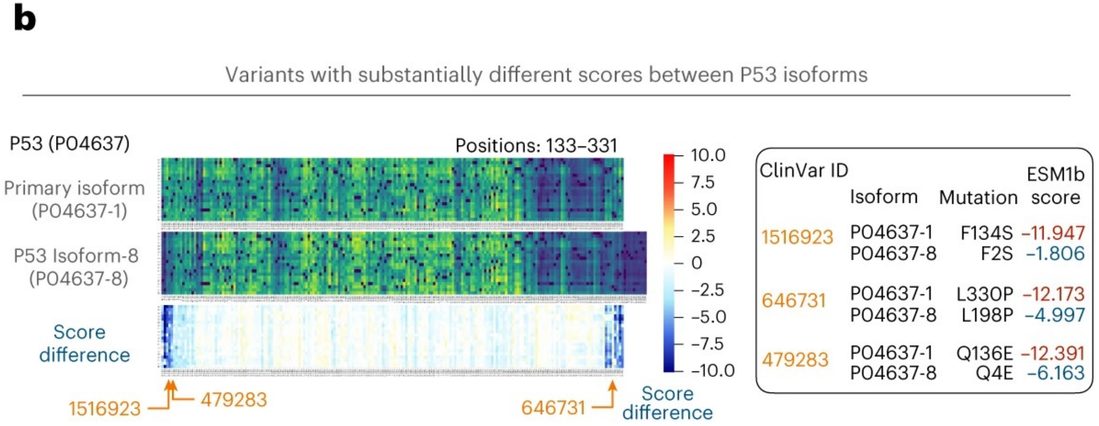


Impressive zero-shot performance

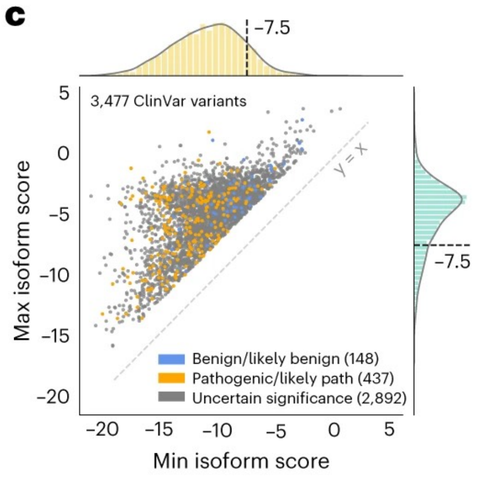


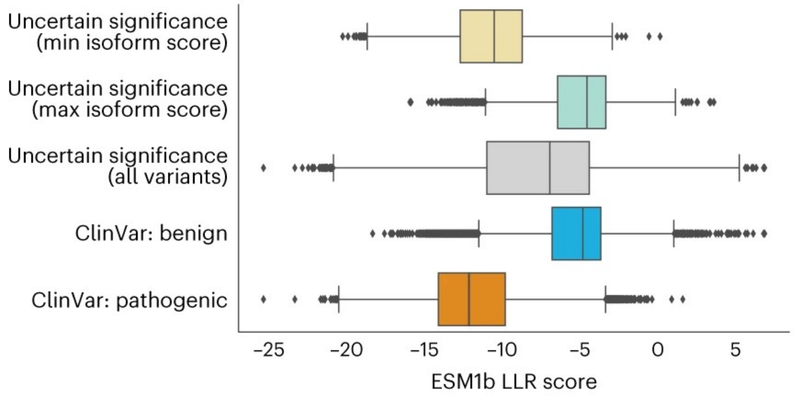
Find the variants that are isoform-sensitive

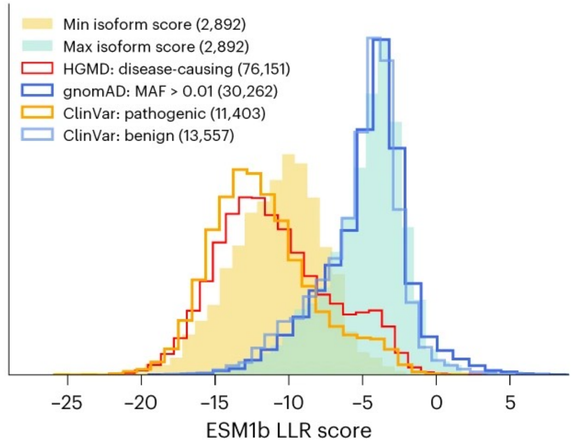




One isoform is pathogenic and another is not!



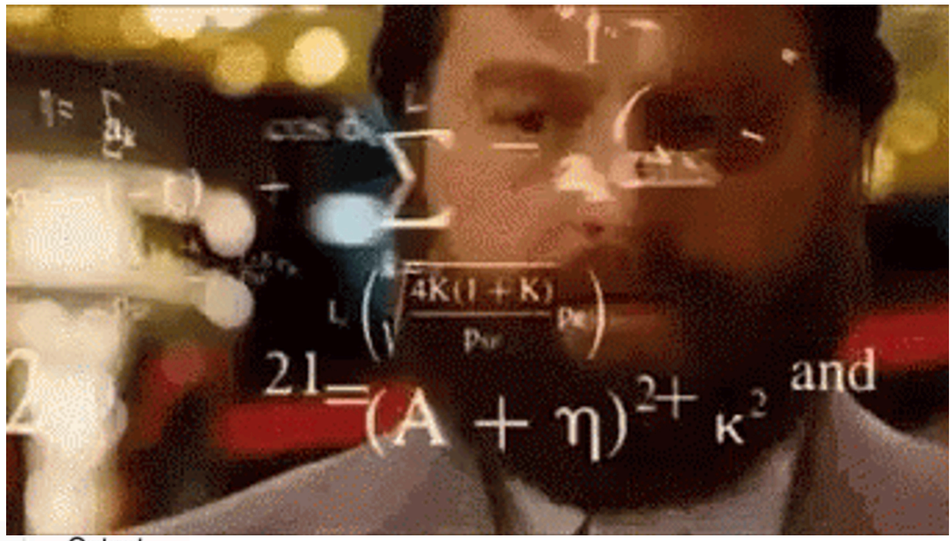


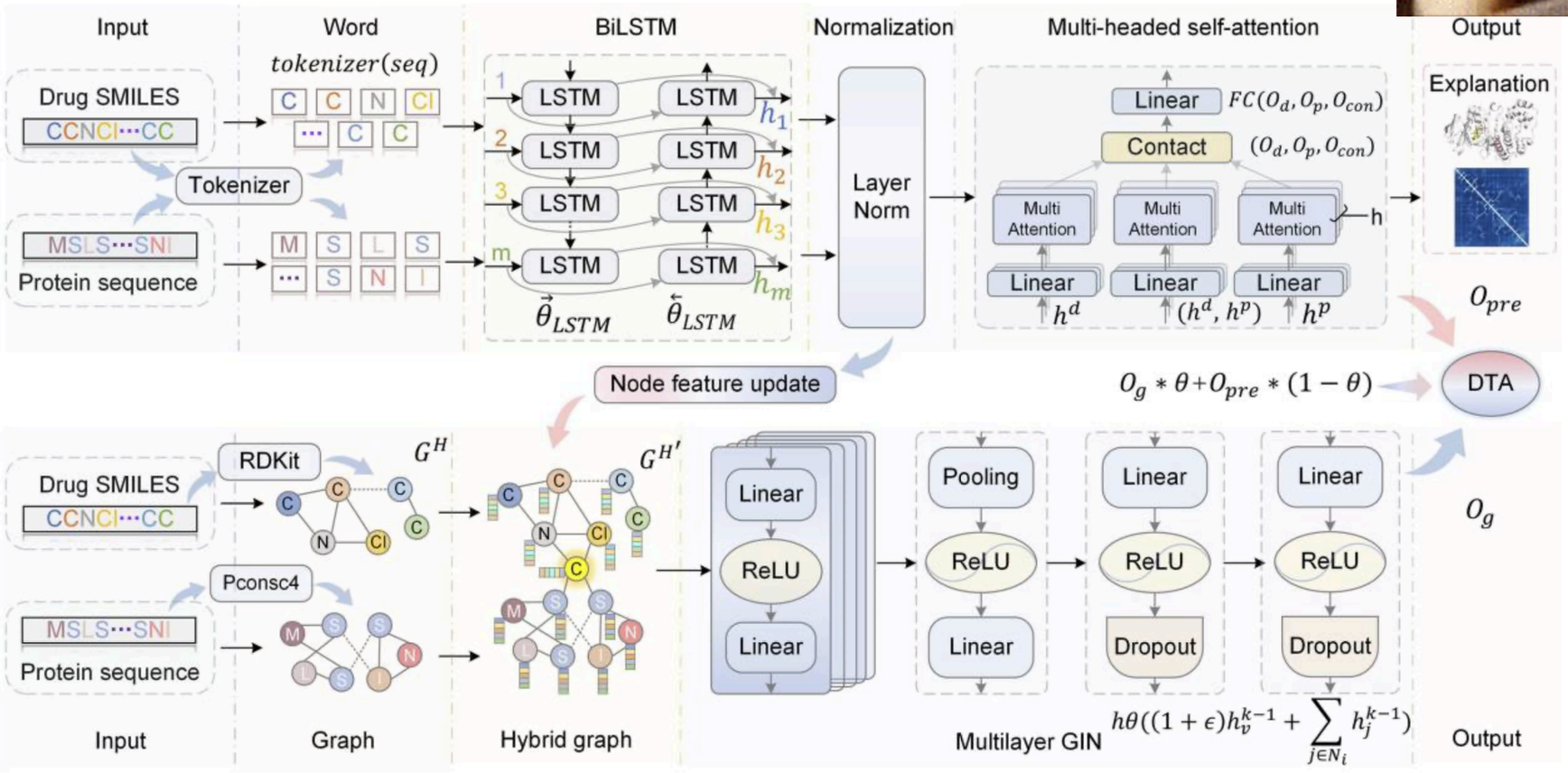


 https://doi.org/10.1093/bioinformatics/btad355

NHGNN-DTA: a node-adaptive hybrid graph neural network for interpretable drug-target binding affinity prediction (He, Chen, and Chen, *Bioinformatics*)

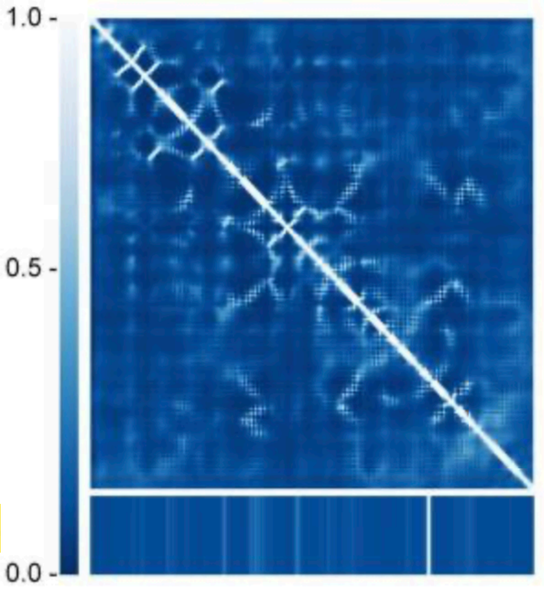
- Goal: Predict drug-target affinities
- Method: Node-adpative hybrid neural network
  - Encodes drugs and targets as both graphs and text sequences
  - Produces a secondary graph that is then fed into multilayer graphical isomorphism network (GIN)
- Result: Establish a new state-of-the-art on well established industry benchmarks
- Conclusion: Enormous application if these new performances mean *in silico* DTA predictions will hold up to prospective experimental validation

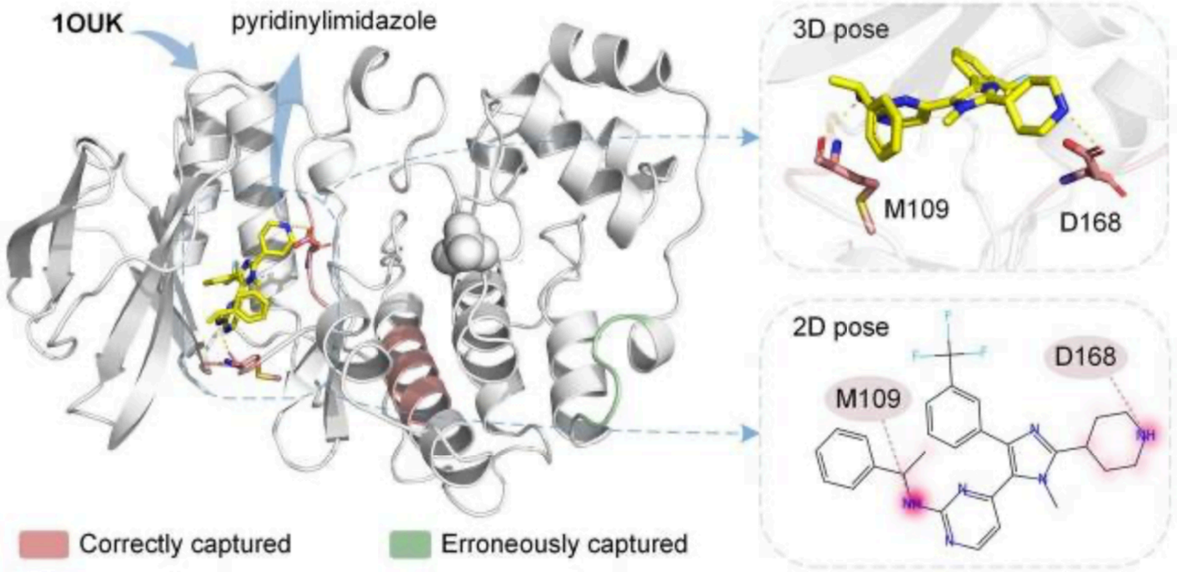

21=(A + η)²+ κ² and


Input
Word
BiLSTM
Normalization
Multi-headed self-attention
Output
Drug SMILES
CCNCl···CC
Tokenizer
MSLS···SNI
Protein sequence
tokenizer(seq)
LSTM
Layer Norm
Linear
Contact
Multi Attention
Explanation
Node feature update
DTA
RDKit
Pconsc4
Input
Graph
Hybrid graph
ReLU
Pooling
Dropout
Multilayer GIN
Output

Contact map —>

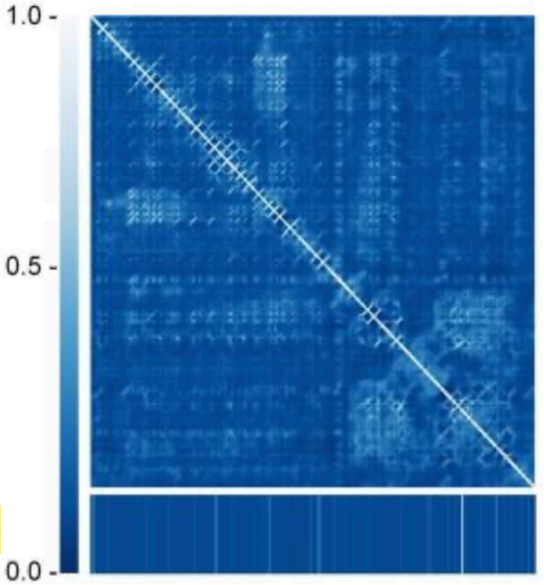
Attention values—>

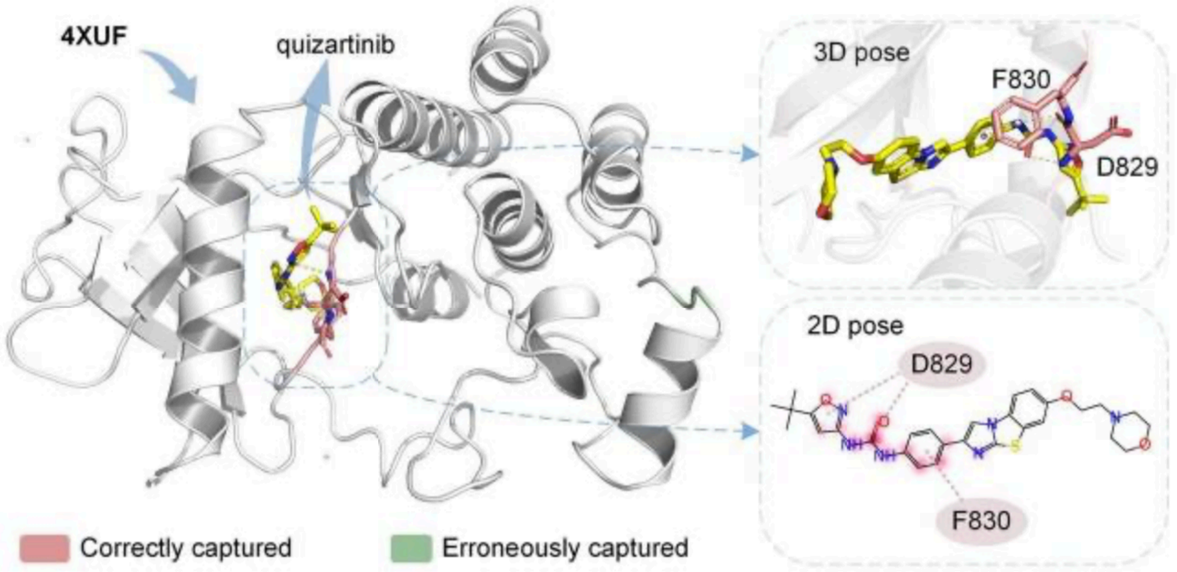




Contact map —>

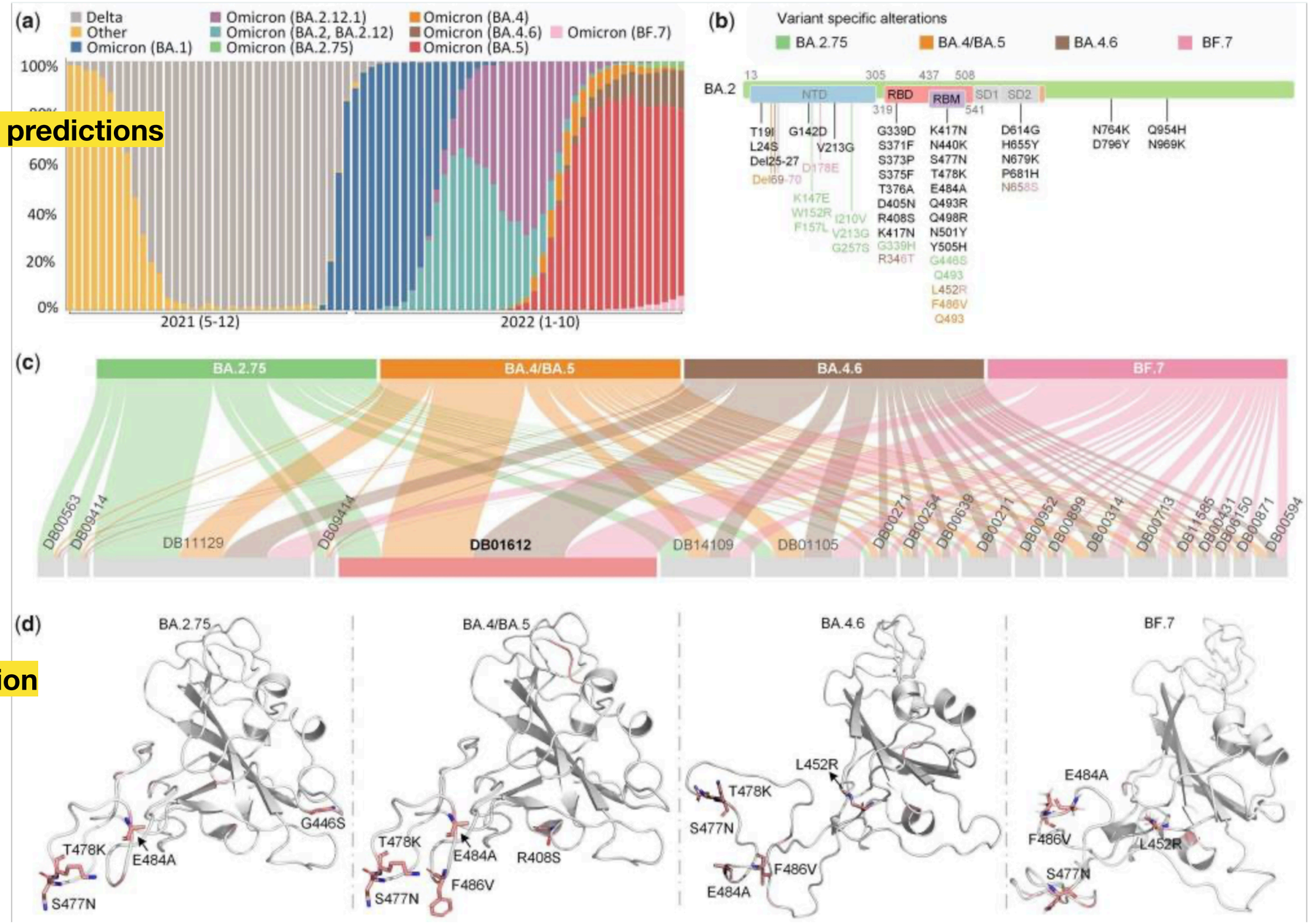
Attention values—>





Can make variant-dependent predictions

And visualize the attention





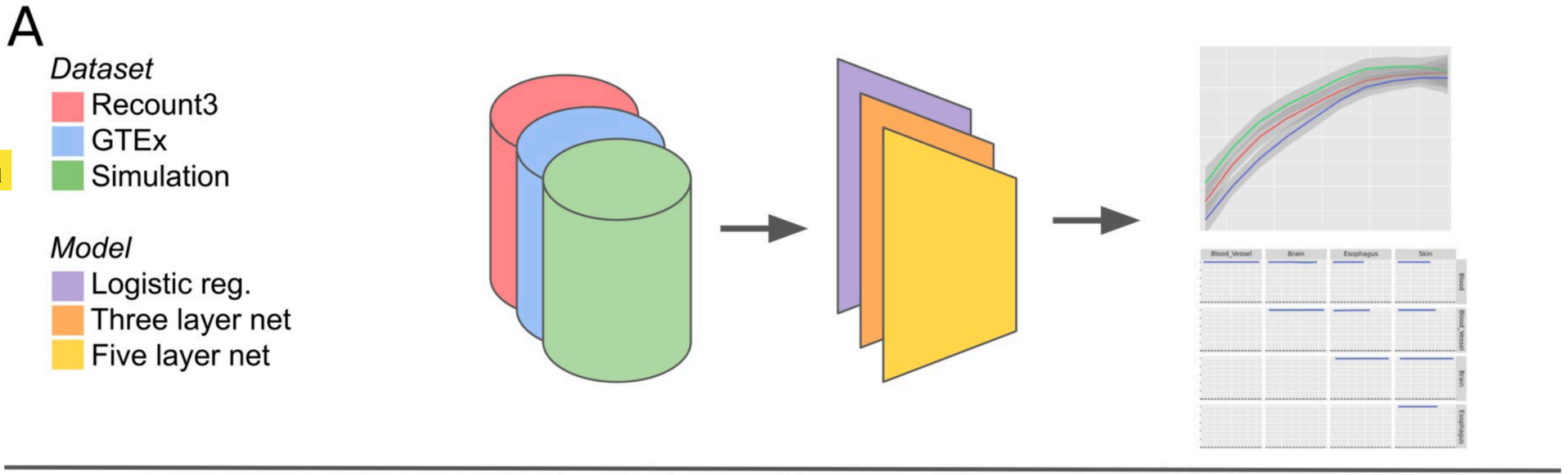
“Halo”

Advancing Prediction and Classification

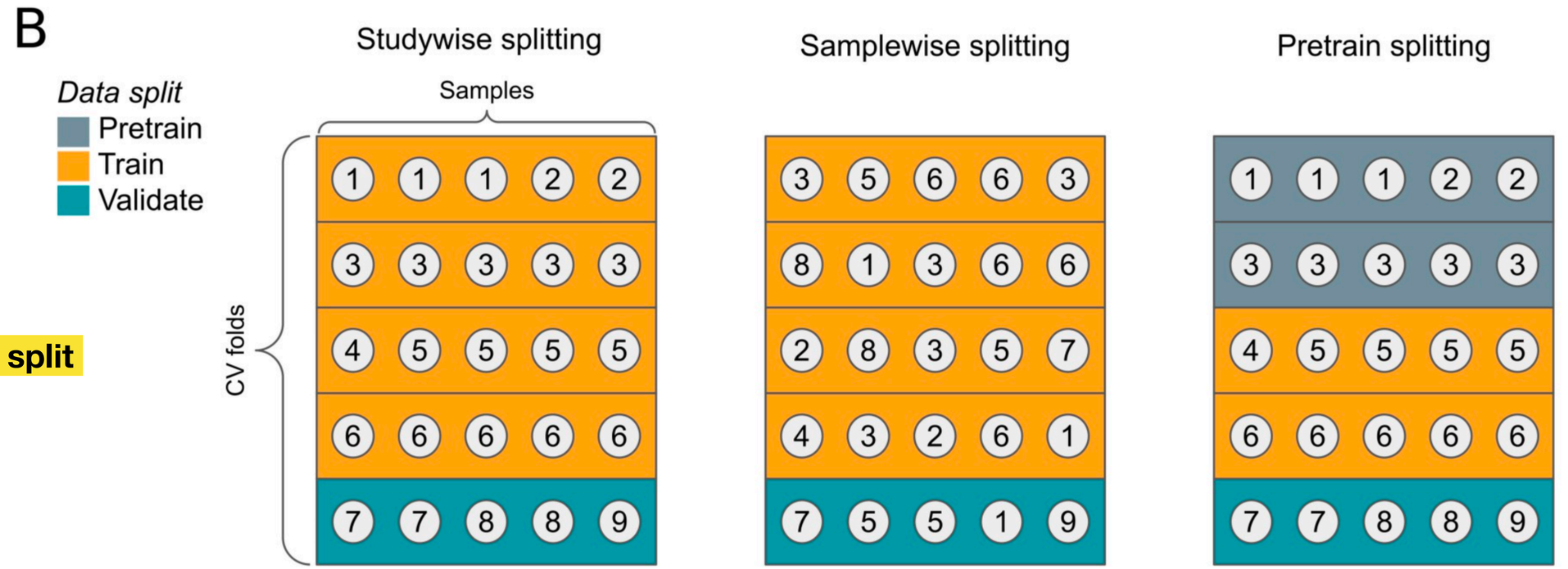
The effect of non-linear signal in classification problems using gene expression (Heil et al, *PLOS Computational Biology*)

- Goal: We know gene expression is non-linear and complex, but we love our simple linear models and the interpretability that comes with them
  - Rigorously evaluate where this tradeoff lies and when complexity/simplicity matters
- Method:
  - Model Recount3+GTEx using linear and non-linear models
  - Remove the “linear portion” of the signal and see if there’s anything left
- Result:
  - For most tasks linear performs as well as non-linear for these top-line tasks
  - However, demonstrated that there is significant non-linear signal in there too
- Conclusion: Keep doing whatever you think is best for your problem.

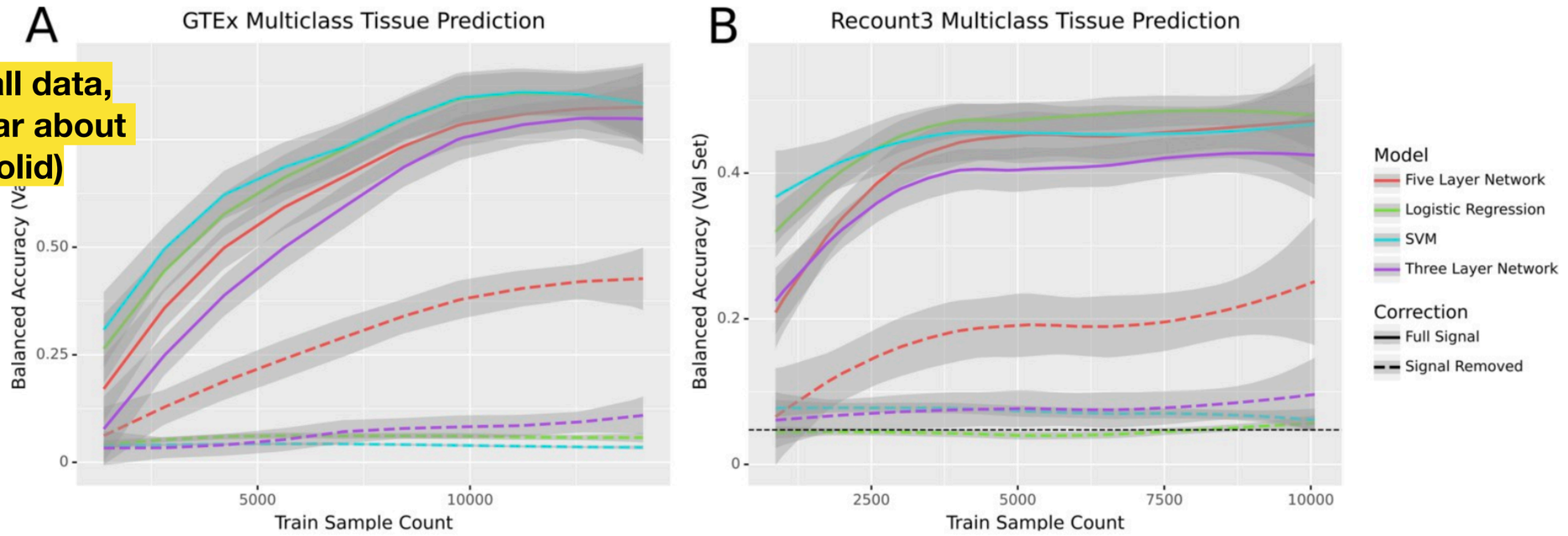
Combination of real data and simulations

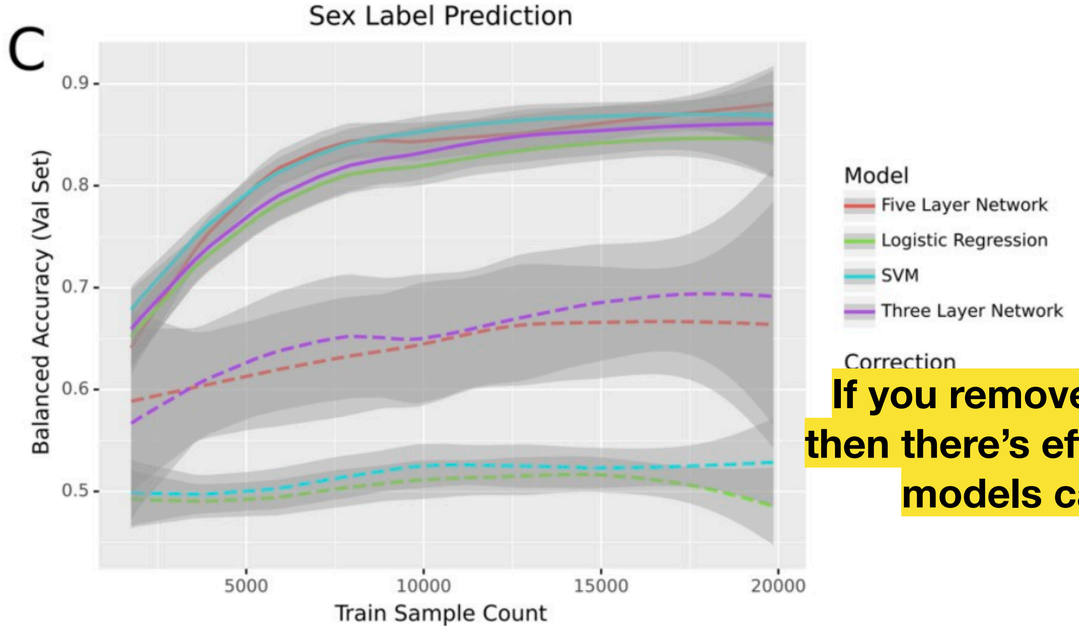


Different validation split strategies



If you include all data, Linear/non-linear about the same (solid)



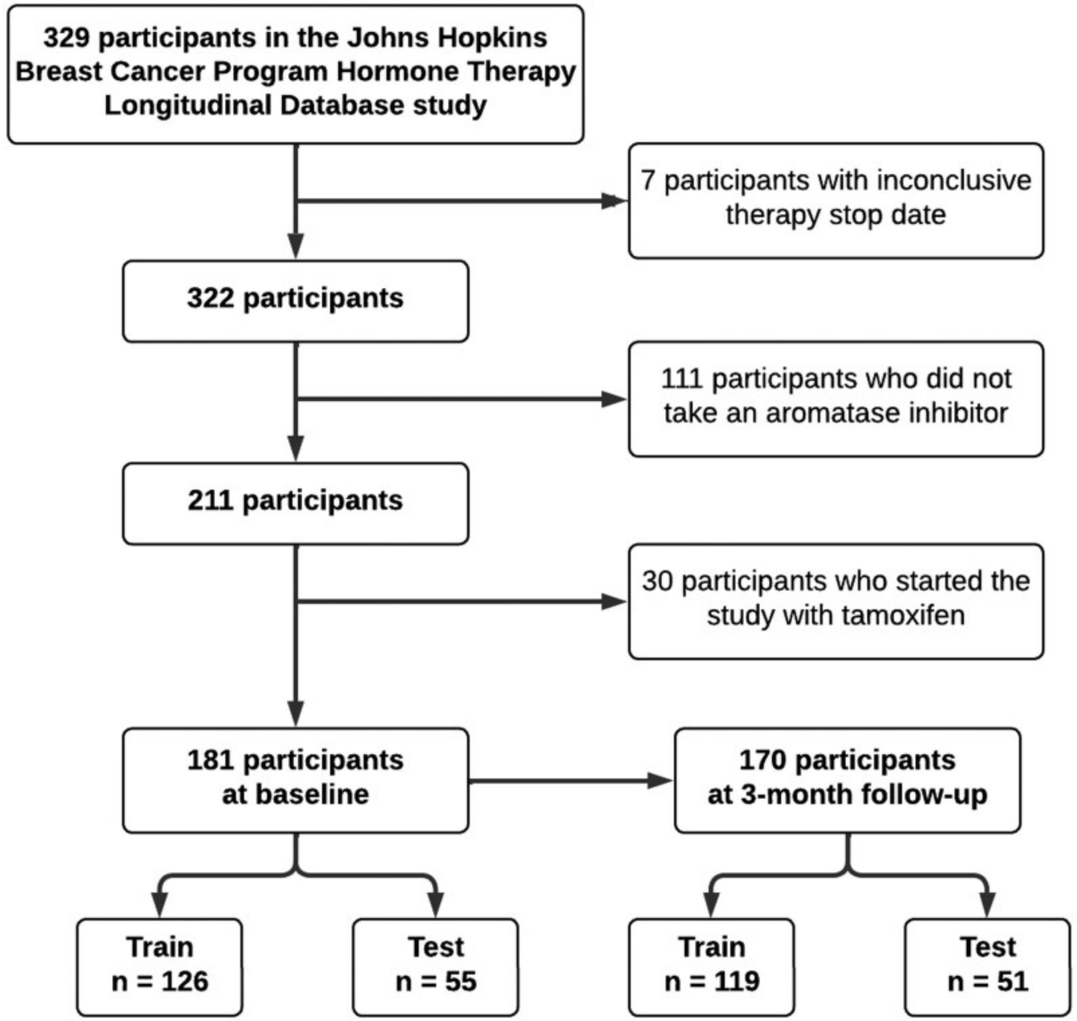


If you remove the linear portion, then there's effects only non-linear models can get (dashed)

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

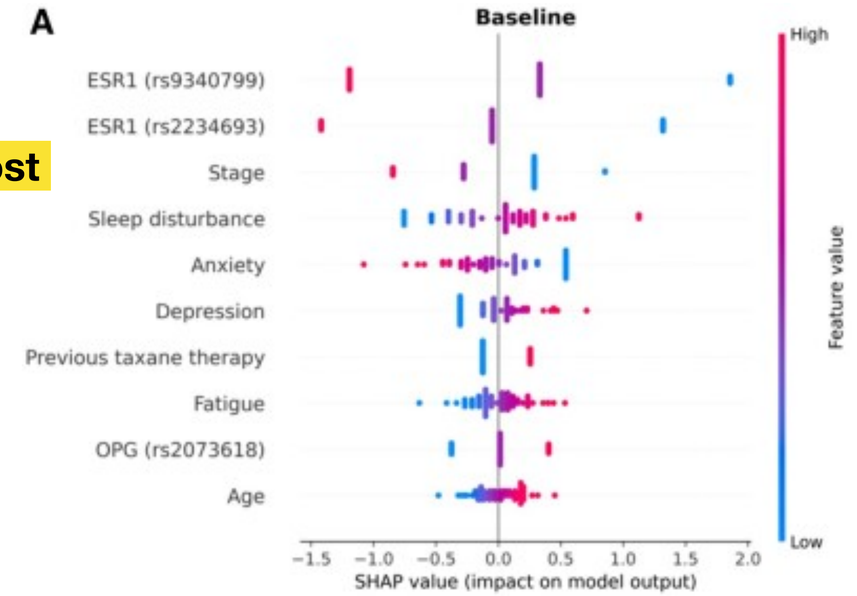
- Goal: Understand the genetic and socioenvironmental risk factors associated with discontinuation of aromatase inhibitors
- Method:
  - Prospective cohort of 181 women with breast cancer
  - Survival machine learning and time-dependent analysis
  - Personalized survival curves
- Result: Identified symptoms associated with discontinuation and variants in ESR1
- Conclusion: Longitudinal analysis reveals possible mechanisms behind drug adherence

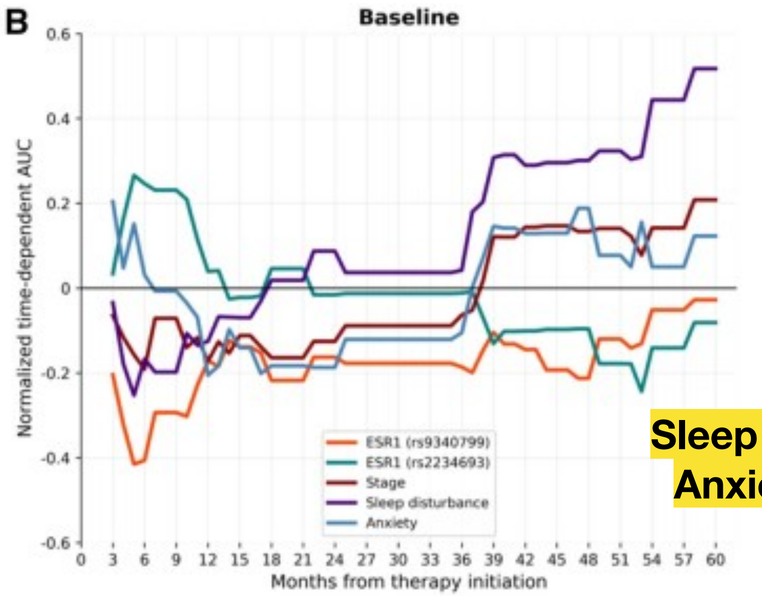
Prospective cohort study



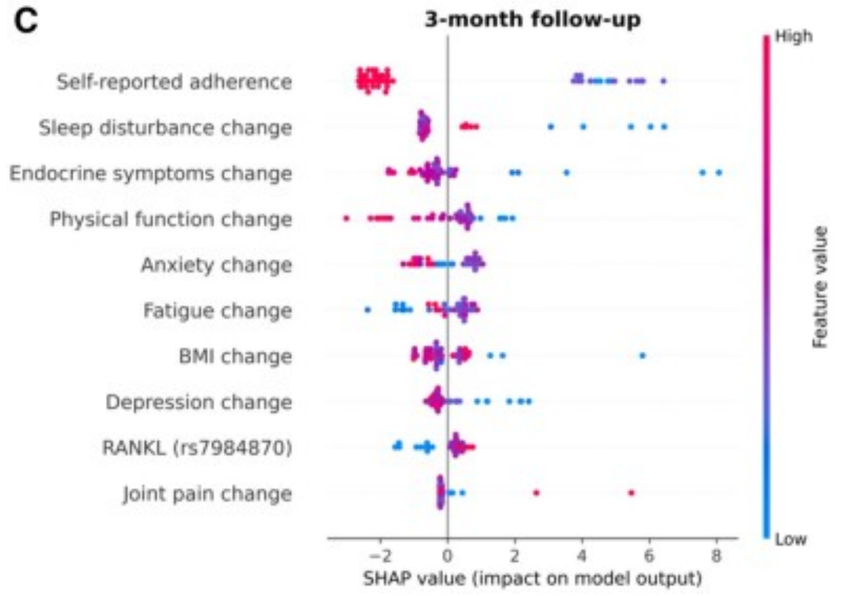
Many patients with 3 month follow up

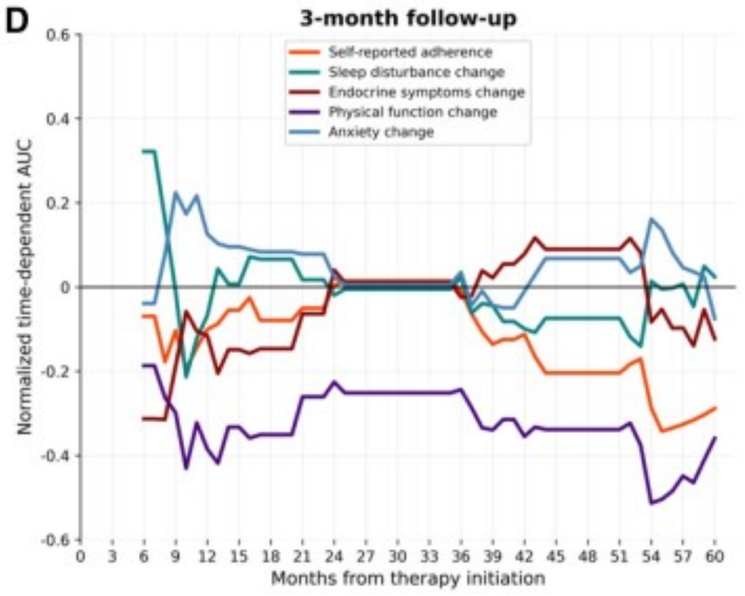
ESR1 variants most predictive



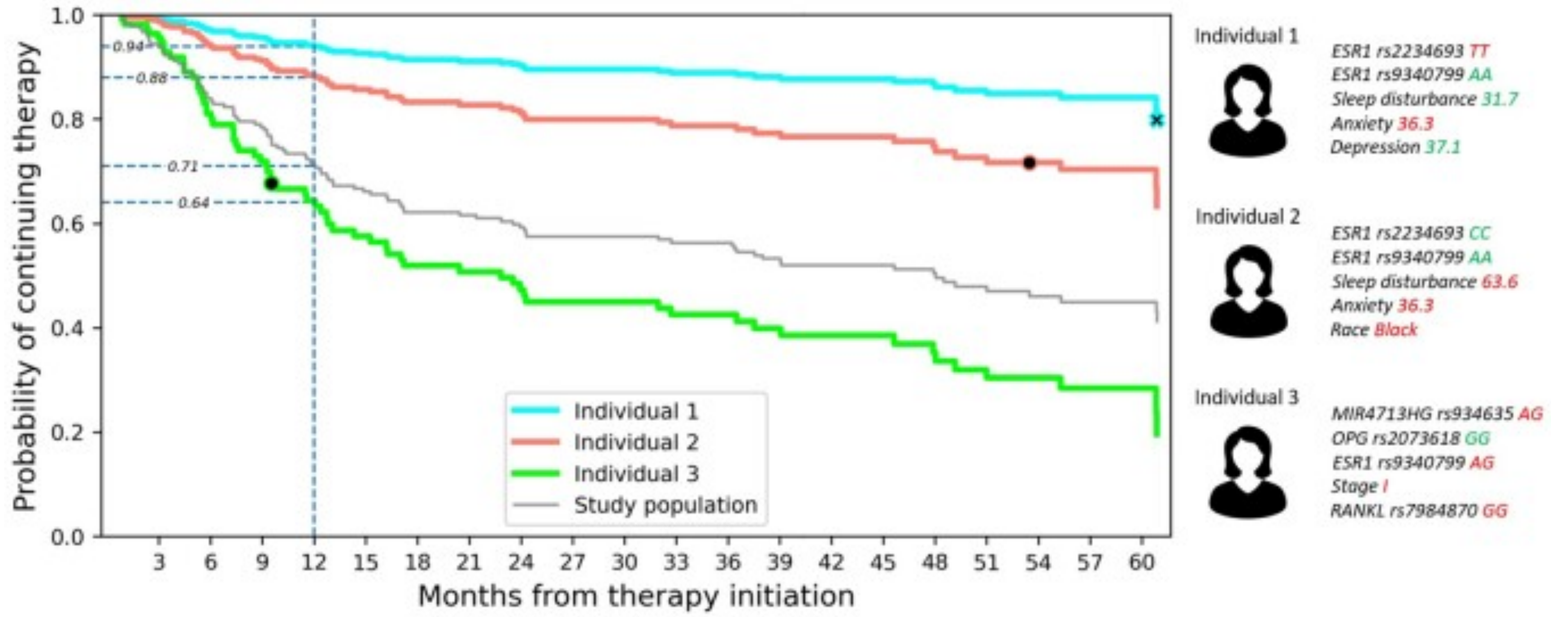


Sleep disturbance and Anxiety is important





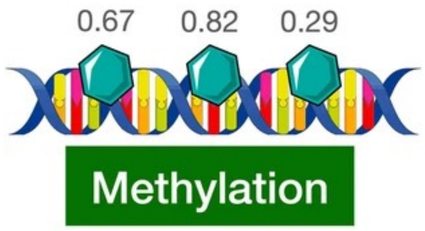
Use models to build out *personal survival curves!*

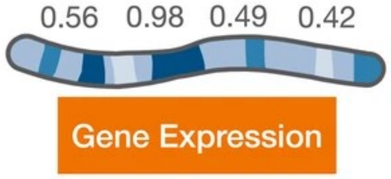


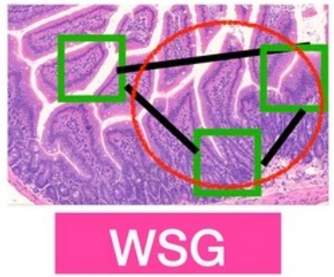
Assessment of emerging pretraining strategies in interpretable multimodal deep learning for cancer prognostication (Azher et al, *BioData Mining*)

- Goal: Evaluate different pre-training strategies rigorously
- Method: Pre-train in cancer setting and compare:
  - Kaplan Meier Curves
  - Partial likelihood ratio tests
  - pTNM staging, and others
- Result: Depends on the task, but by rigorously evaluated multiple approaches you can increase your performance by up to 12%
- Conclusion: No easy answers, evaluate your pertaining strategy rigorously, don’t assume anything

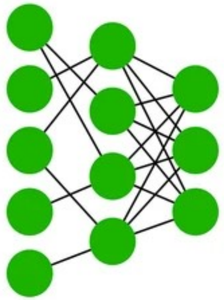
Different data modalities



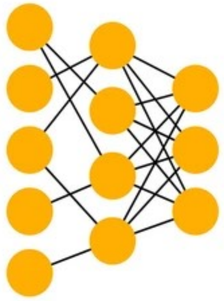




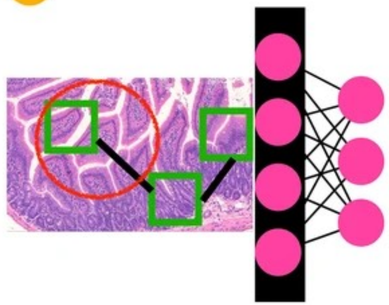
Different pre-training strategies



Crossmodal GE Pretraining
Unimodal Survival Networks
Transfer Learning

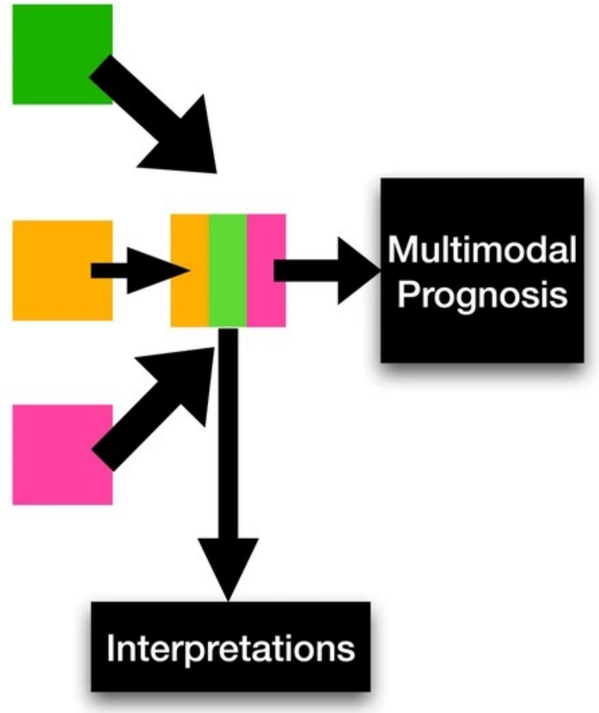


Crossmodal DNAm Pretraining
Unimodal Survival Networks
Transfer Learning

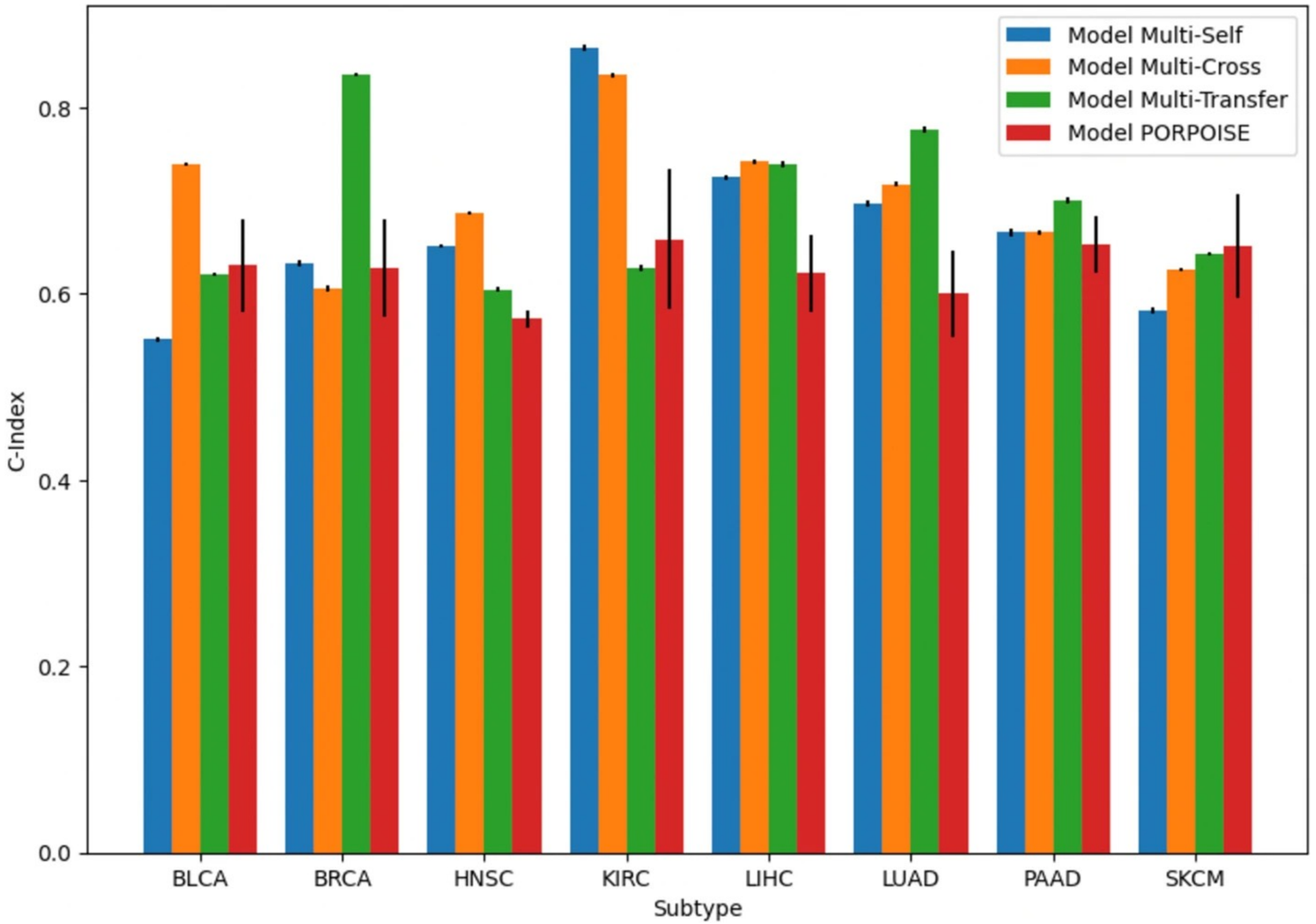


Crossmodal Pretraining
Unimodal Survival Networks
Transfer Learning
Contrastive Learning

Common evaluations

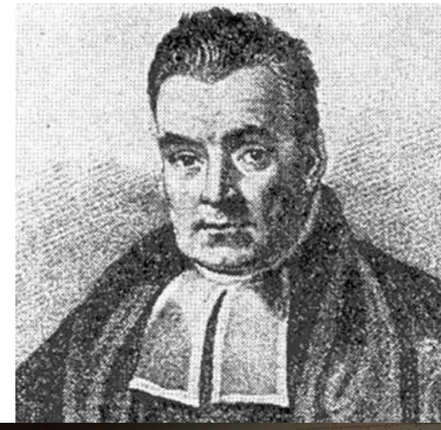


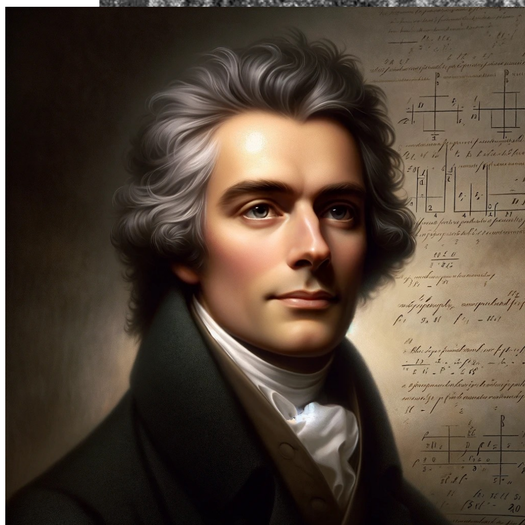
Different strategies perform best



BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability (Hoggart et al, *Nature Genetics*)

- Goal: Build a PRS that won't suck when applied across ancestry groups
- Method:
  - Bayesian method that integrates trans-ancestry GWAS summary stats
  - Designed for when you have a well-powered GWAS in a discovery population and low-powered GWAS in a target population
- Result:
  - Significant improvement in PRS R-sqaured for diverse populations (esp. African ancestry)
- Conclusion: Bayes continues to do work 💪

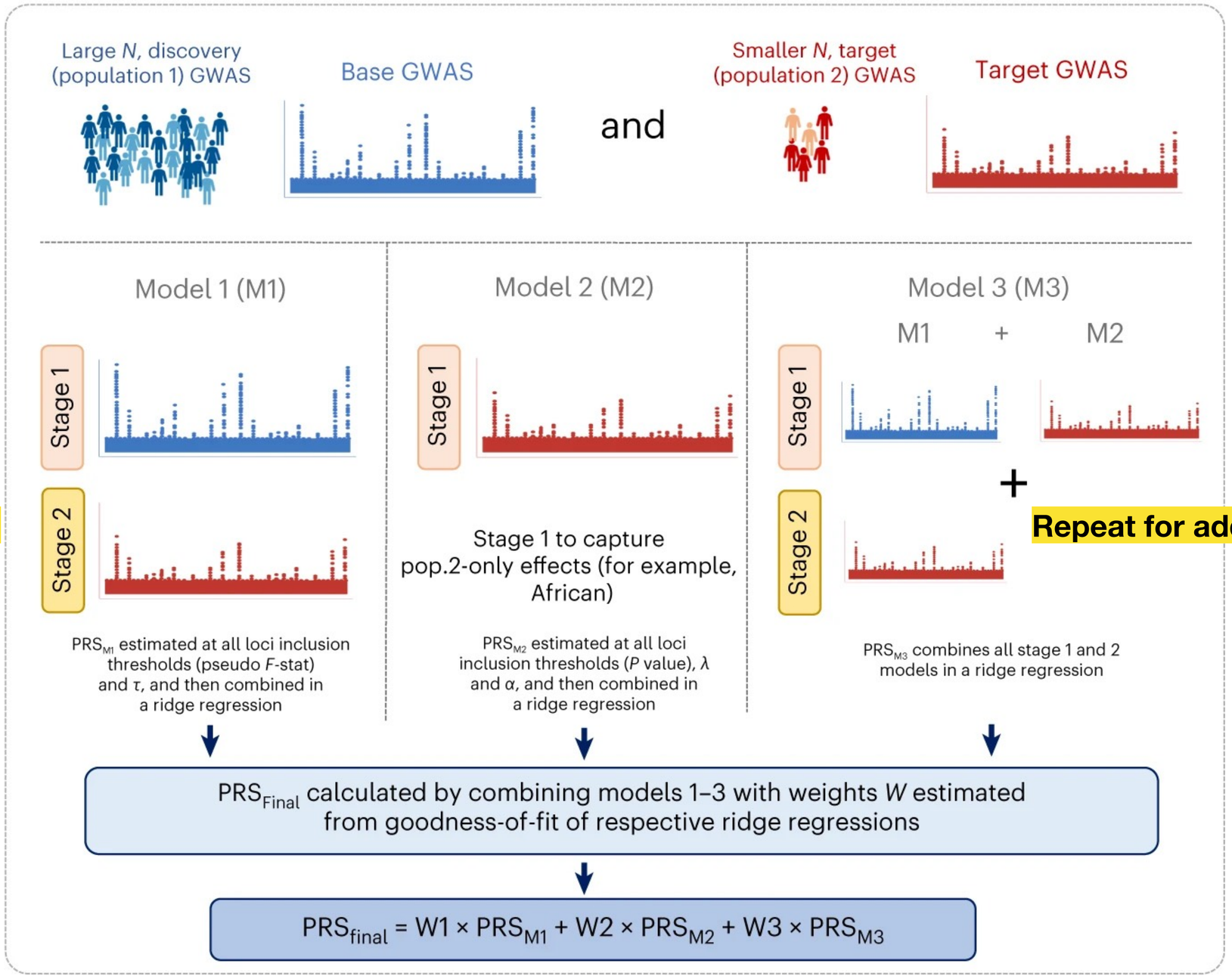




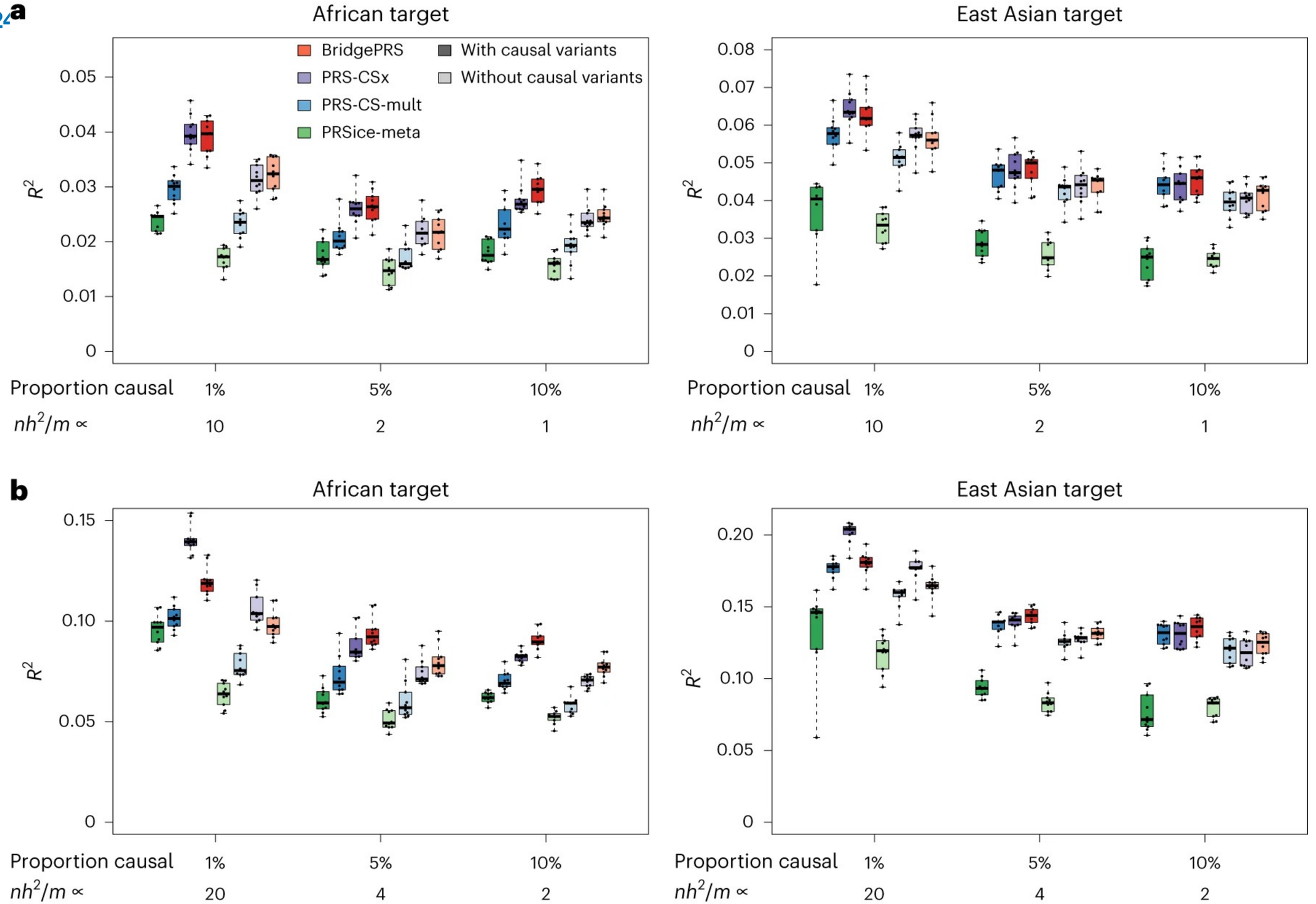
Stage 1: Train a well-powered PRS

Stage 2: Optimize for target population

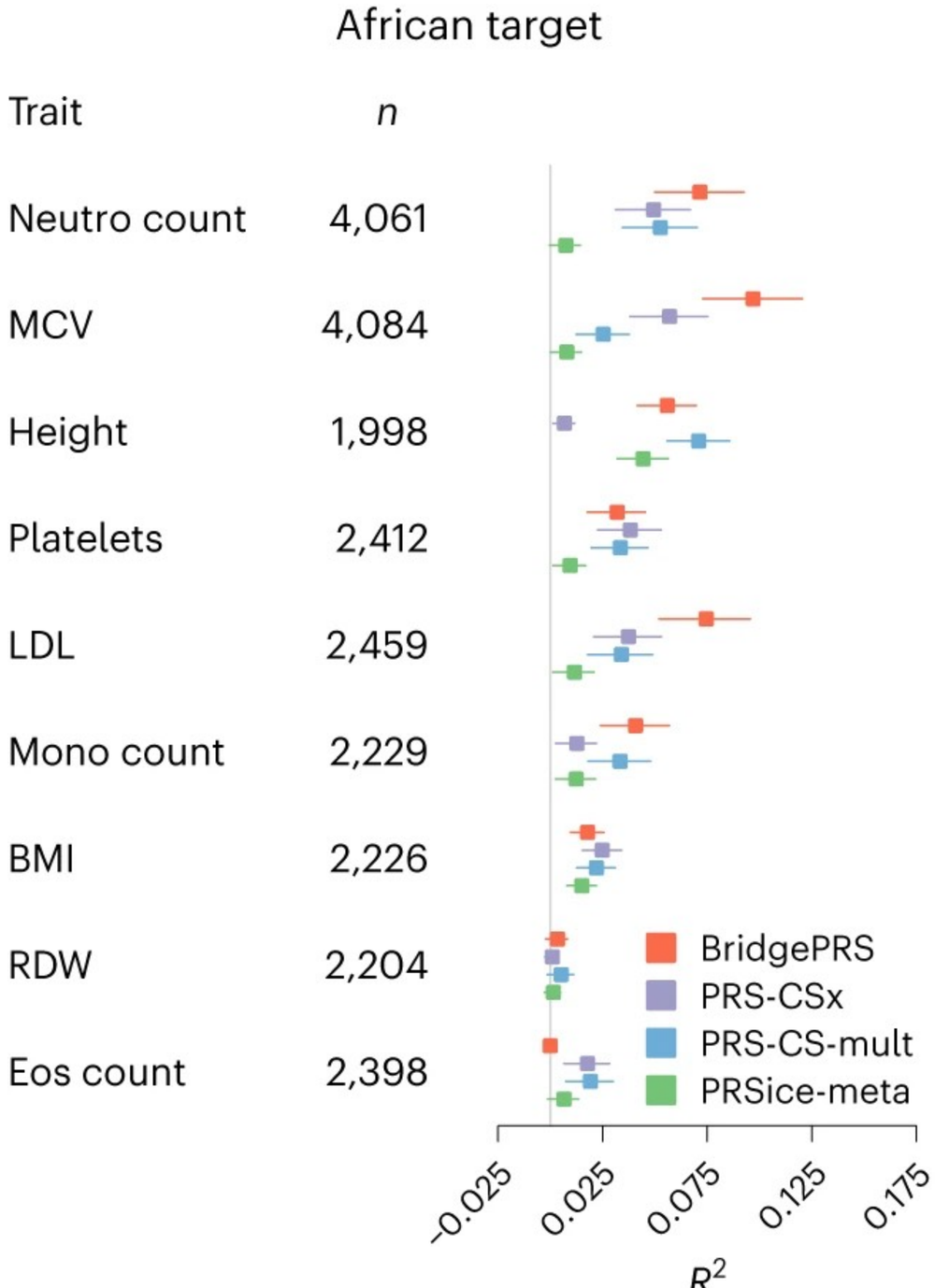
- use stage 1 posteriors as priors



Repeat for add'l pops


a
African target
East Asian target
BridgePRS
PRS-CSx
PRS-CS-mult
PRSice-meta
With causal variants
Without causal variants
R^2
0.05
0.04
0.03
0.02
0.01
0
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Proportion causal
1%
5%
10%
$nh^2/m \propto$
10
2
1
b
African target
East Asian target
R^2
0.15
0.10
0.05
0
0.20
0.15
0.10
0.05
0
Proportion causal
1%
5%
10%
$nh^2/m \propto$
20
4
2

Validates well in external data





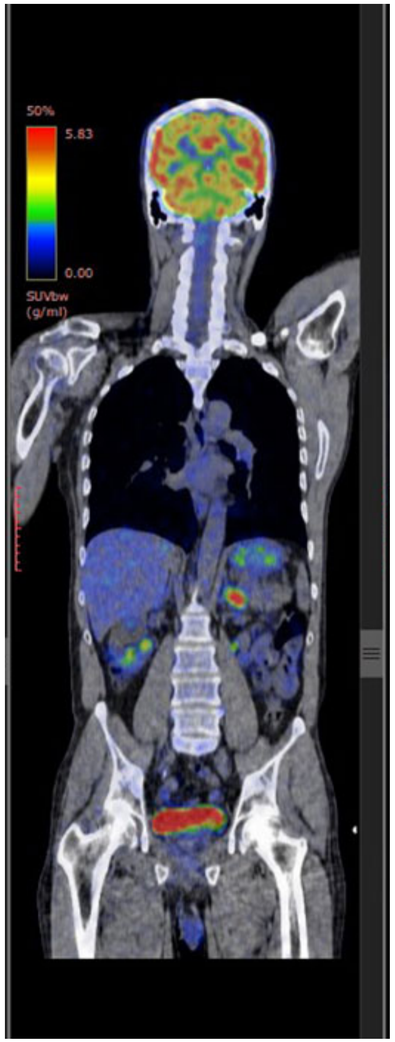
“Photograph"

Capturing Insights through Imaging

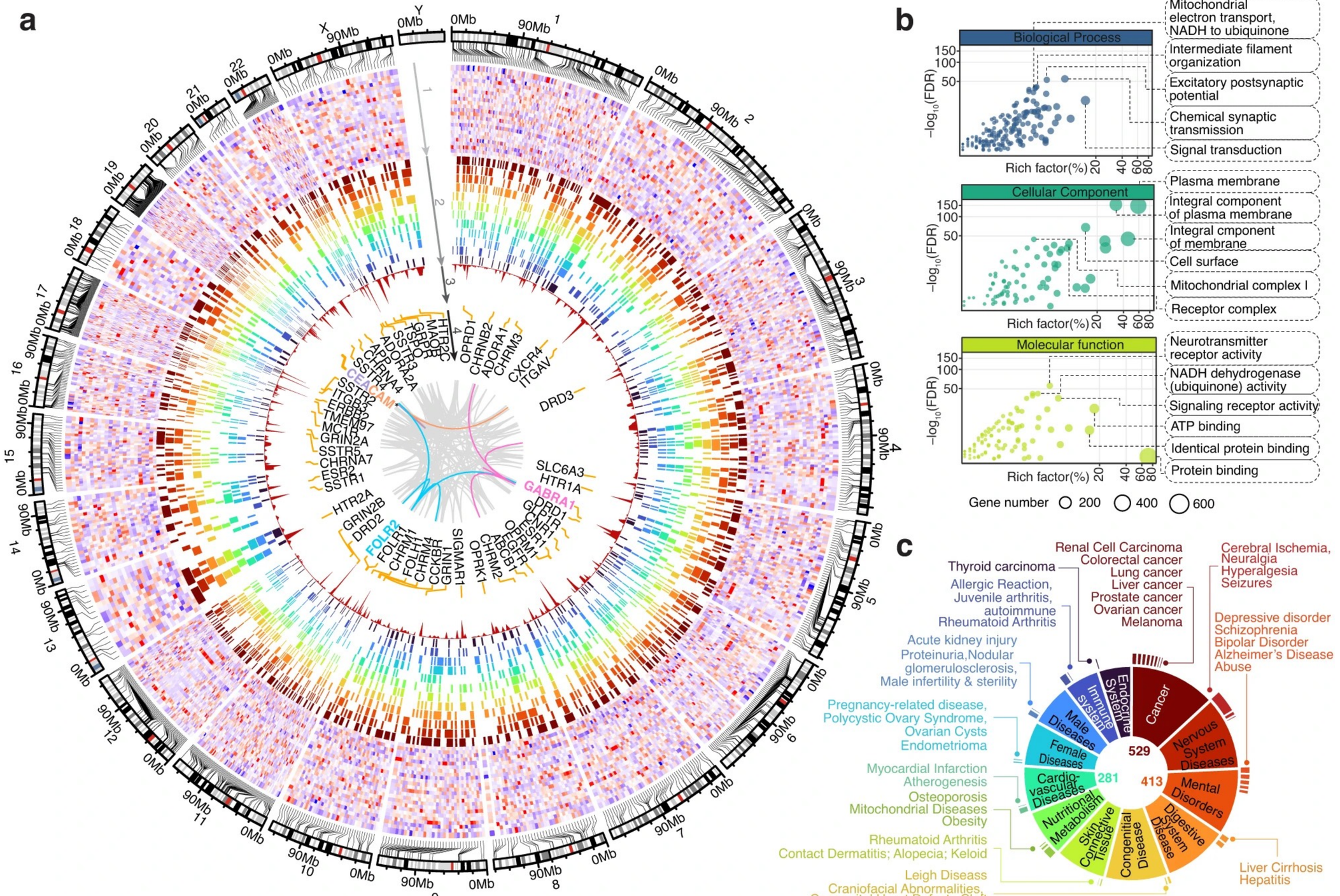
 https://doi.org/10.1038/s41467-023-43123-3

The Imageable Genome (Jané, Xu, and Taelman et al, *Nature Communications*)

- Goal: Reveal “the part of the human genome whose expression can be assessed via molecular imaging”
  - PET-CT is a non-invasive way to image cellular genetic activity
  - Update of these methods have been limited, most because of a lack of knowledge of what genes are imageable
- Method: AI-assisted review of all PubMed and several field specific databases to find what radio tracers have been used with which genes
- Result:
  - 1,173 genes across all chromosomes (except Y) and many diseases
  - Most are membrane bound signaling receptors
- Conclusion: Everyone look under your seat…
    - You get a PET-CT, you get a PET-CT!

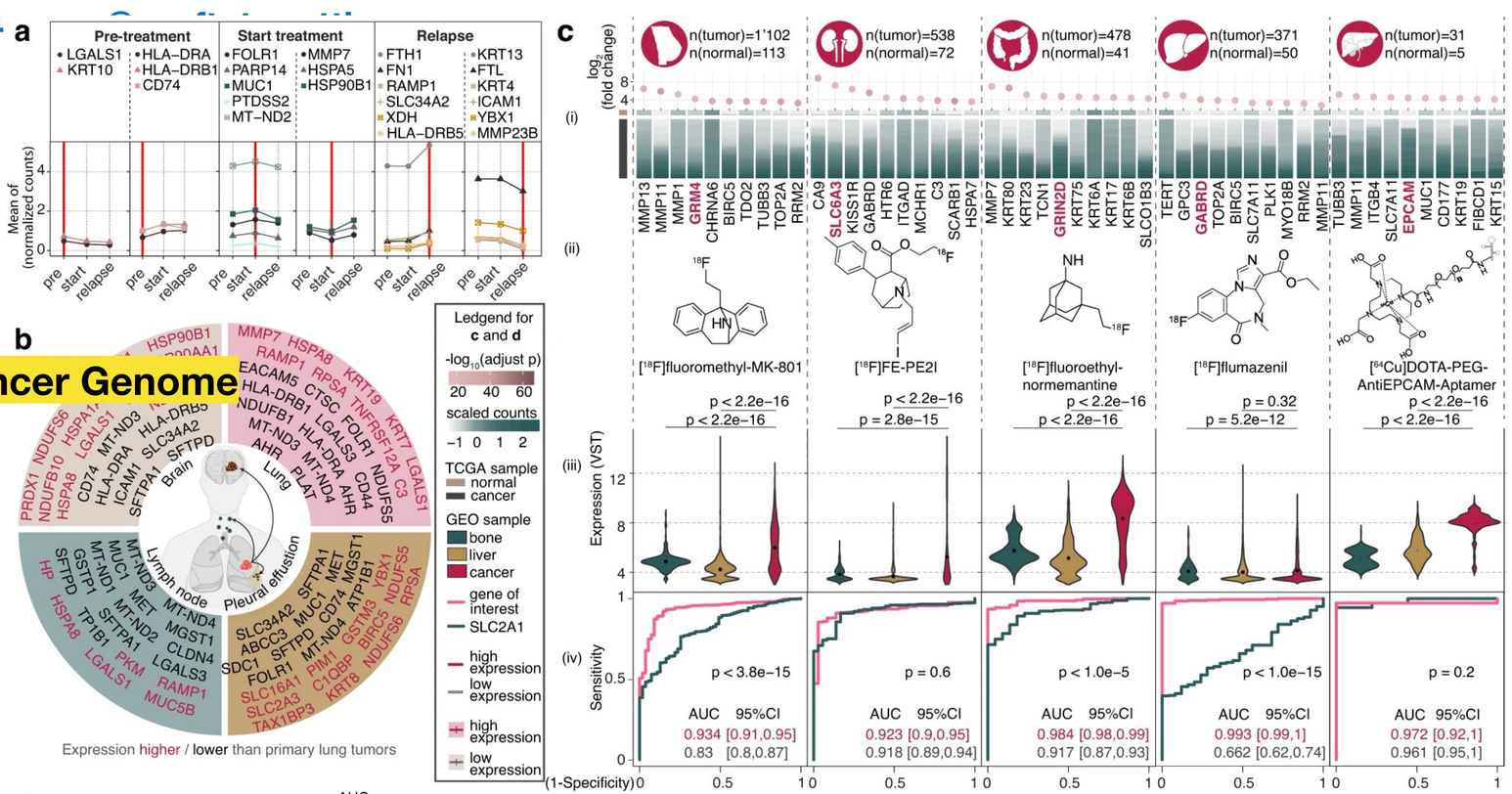





a
b
Biological Process
Mitochondrial electron transport, NADH to ubiquinone
Intermediate filament organization
Excitatory postsynaptic potential
Chemical synaptic transmission
Signal transduction
Cellular Component
Plasma membrane
Integral component of plasma membrane
Integral cmponent of membrane
Cell surface
Mitochondrial complex I
Receptor complex
Molecular function
Neurotransmitter receptor activity
NADH dehydrogenase (ubiquinone) activity
Signaling receptor activity
ATP binding
Identical protein binding
Protein binding
Rich factor(%)
−log10(FDR)
Gene number 200 400 600
c
Renal Cell Carcinoma
Colorectal cancer
Lung cancer
Liver cancer
Prostate cancer
Ovarian cancer
Melanoma
Thyroid carcinoma
Allergic Reaction, Juvenile arthritis, autoimmune Rheumatoid Arthritis
Acute kidney injury Proteinuria,Nodular glomerulosclerosis, Male infertility & sterility
Pregnancy-related disease, Polycystic Ovary Syndrome, Ovarian Cysts Endometrioma
Myocardial Infarction Atherogenesis
Osteoporosis
Mitochondrial Diseases
Obesity
Rheumatoid Arthritis
Contact Dermatitis; Alopecia; Keloid
Leigh Diseases
Craniofacial Abnormalities
Cerebral Ischemia, Neuralgia Hyperalgesia Seizures
Depressive disorder Schizophrenia Bipolar Disorder Alzheimer's Disease Abuse
Liver Cirrhosis Hepatitis
Cancer
Nervous System Diseases
Mental Disorders
Digestive System Disease
Congenital Disease
Skin Connective Tissue
Nutritional Metabolism
Cardio-vascular Diseases
Female Diseases
Male Diseases
Immune system
Endocrine System
529
413
281

#IS24 #YIR24

The Imageable Cancer Genome



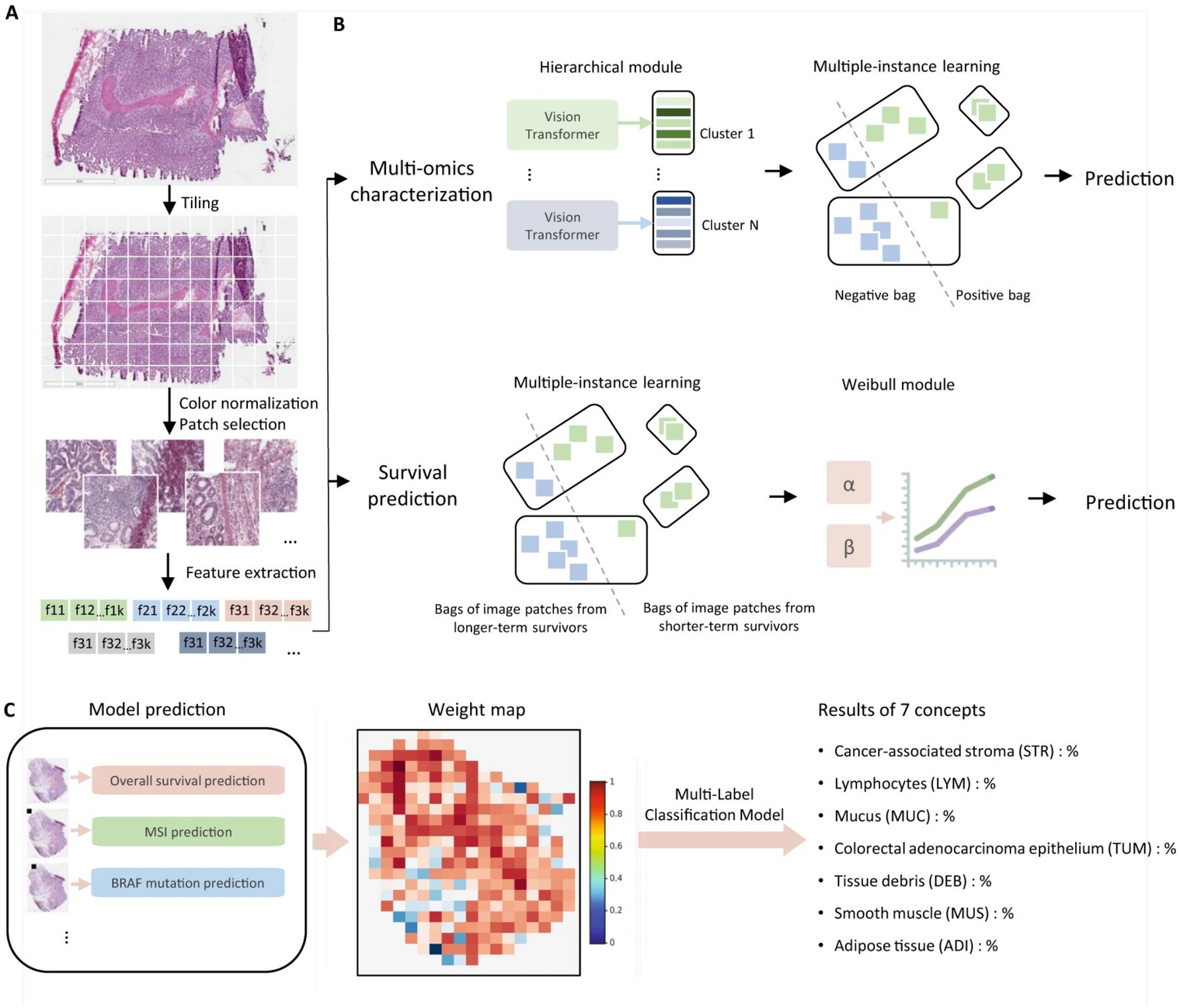
You can find your "favorite" gene

Imageable genes separate survival

 https://doi.org/10.1038/s41467-023-37179-4

Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients (Tsai et al, *Nature Communications*)

- Goal: Derive prognostic and treatment biomarkers from routinely collected histopathology slides.
- Method: Multi-task vision transformer model (identify, survival, MSI status, copy number alterations)
- Result: State-of-the-art performance across all tasks
- Conclusion: Clinical geneticists and pathologists are in a race to eliminate each other’s jobs.

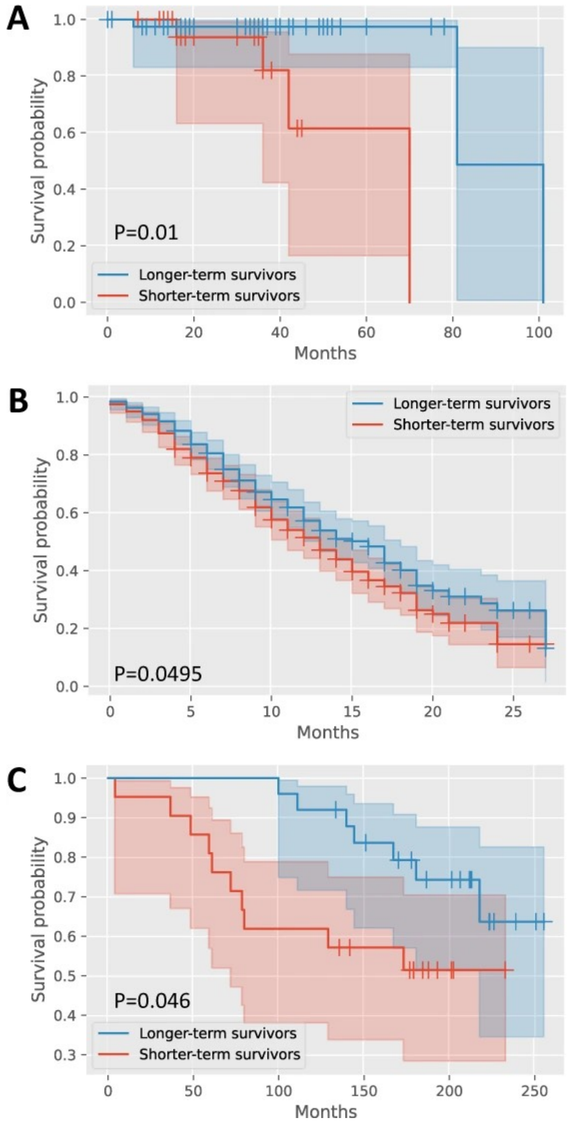

A
Tiling
Color normalization
Patch selection
Feature extraction
f11 f12..f1k f21 f22..f2k f31 f32..f3k
f31 f32..f3k f31 f32..f3k
B
Multi-omics characterization
Hierarchical module
Vision Transformer
Cluster 1
Vision Transformer
Cluster N
Multiple-instance learning
Negative bag
Positive bag
Prediction
Survival prediction
Multiple-instance learning
Bags of image patches from longer-term survivors
Bags of image patches from shorter-term survivors
Weibull module
α
β
Prediction
C
Model prediction
Overall survival prediction
MSI prediction
BRAF mutation prediction
Weight map
1
0.8
0.6
0.4
0.2
0
Multi-Label Classification Model
Results of 7 concepts
• Cancer-associated stroma (STR) : %
• Lymphocytes (LYM) : %
• Mucus (MUC) : %
• Colorectal adenocarcinoma epithelium (TUM) : %
• Tissue debris (DEB) : %
• Smooth muscle (MUS) : %
• Adipose tissue (ADI) : %

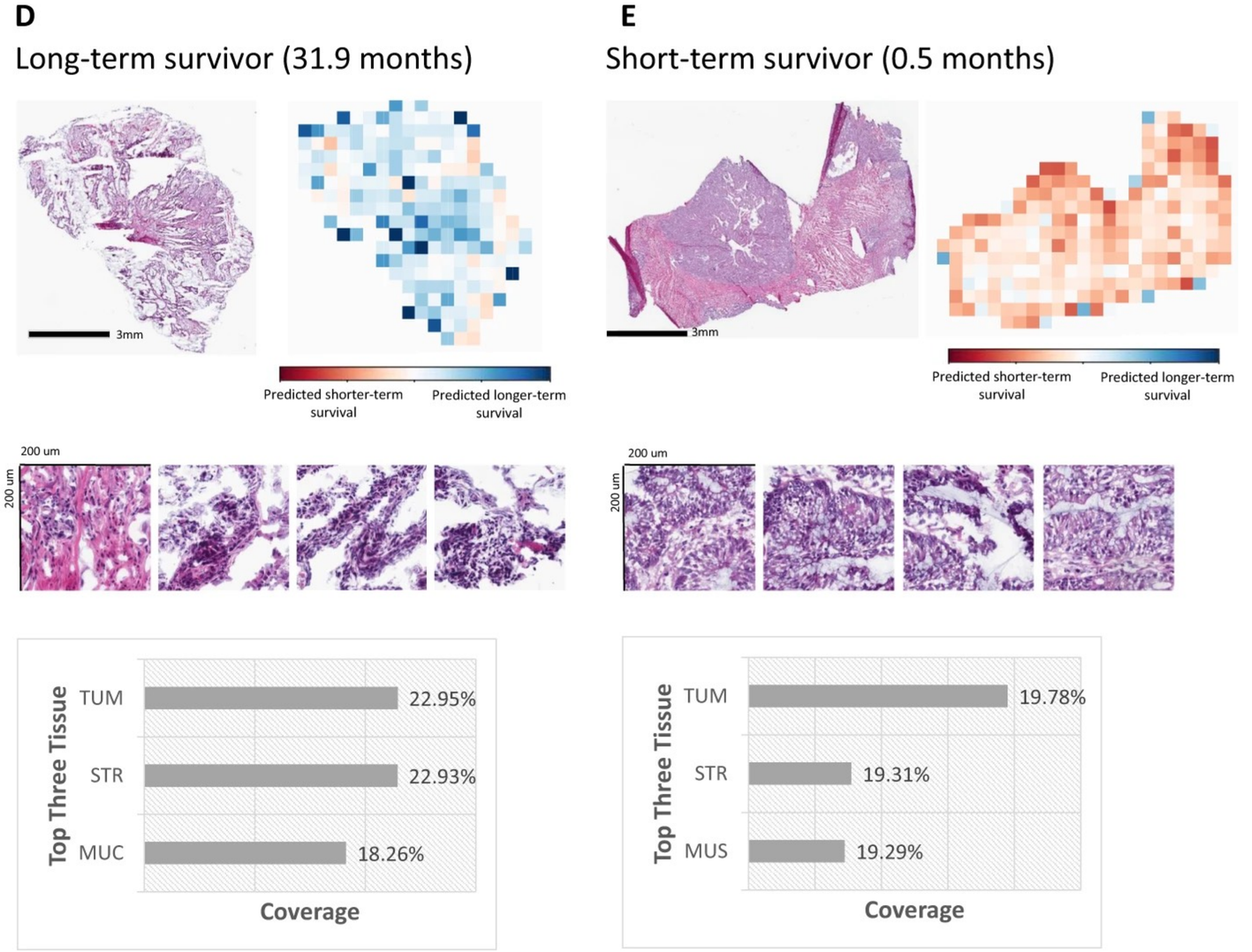
MOMA predicts survival

And replicates…

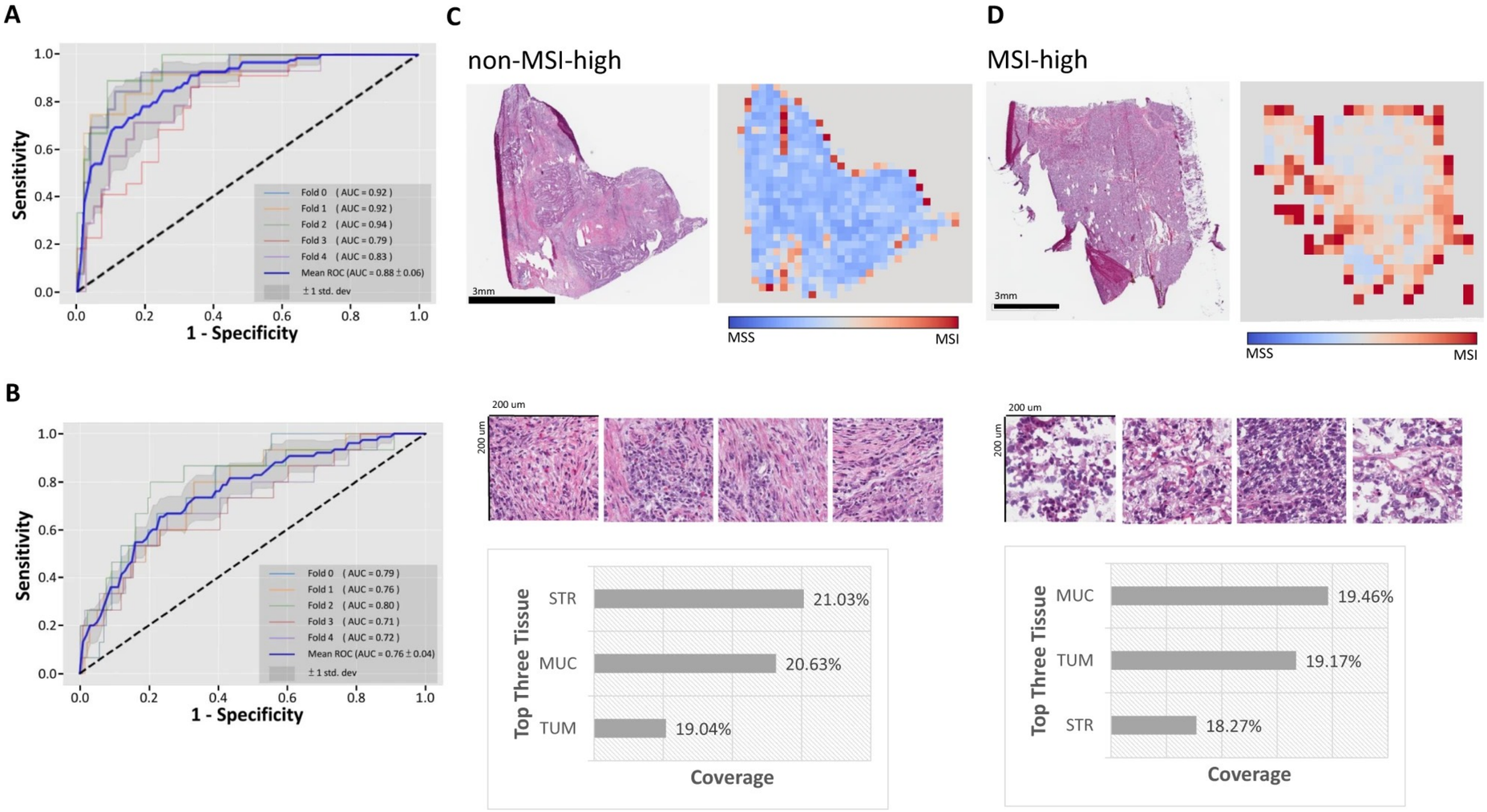
And replicates…

…and makes pretty maps that explain prediction

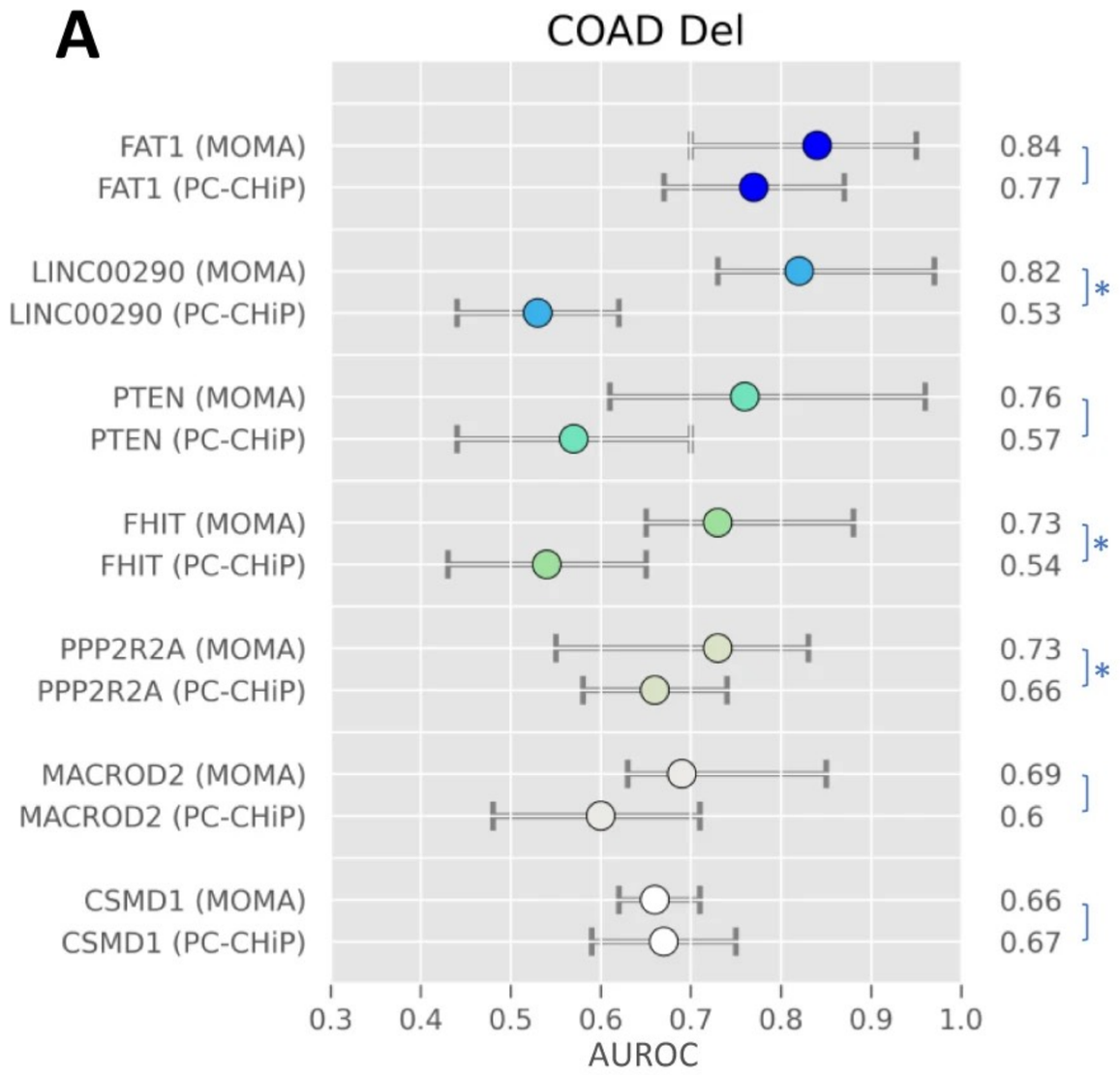




MOMA predicts Micro-satellite Instability Status



Significantly improved copy number variation detection compared to state-of-the-art



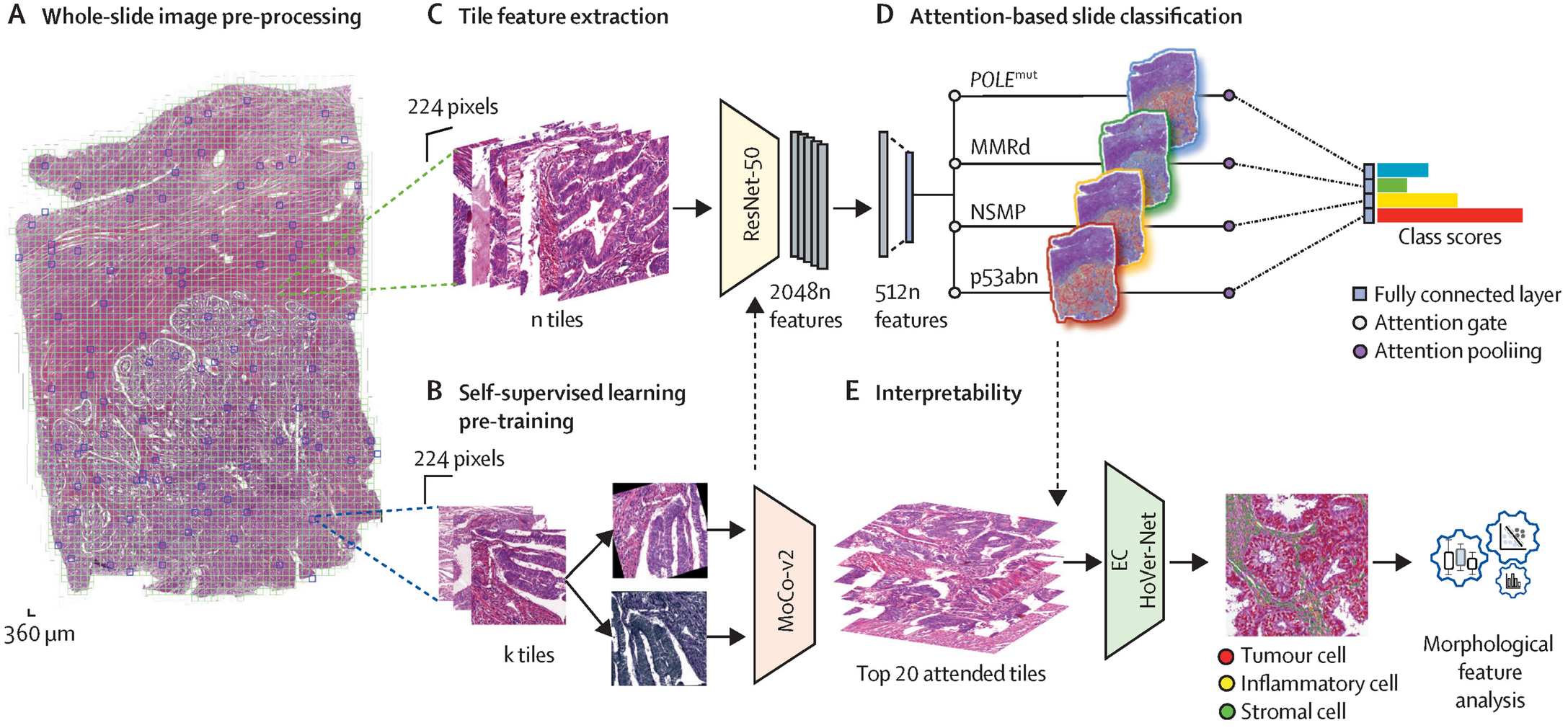
 https://doi.org/10.1016/s2589-7500(22)00210-2

Interpretable deep learning model to predict the molecular classification of endometrial cancer from haematoxylin and eosin-stained whole-slide images: a combined analysis of the PORTEC randomised trials and clinical cohorts (Fremond and Andani et al, *The Lancet Digital Health*)

- Goal: Use H&E stained slides to predict the molecular subtype of endometrial cancer
- Method:
  - Whole-slide-image dataset of 2,028 patients from 3 RCTs
  - MoCo-v2 (tile selection) + ResNet-50 (CNN) + HoVer-Net (nuclear segmentation and classification)
- Result:
  - AUROC of 0.874 in x-fold val; 0.876 in independent test set
  - Identified low tumor-to-stroma ratio as novel subtype characteristic
  - Misclassifications identified patients that show features atypical for their class and survival more consistent with the misclassification
- Conclusion: These models will start to become routine practice as their continued clinically validity is established

Integrates several models

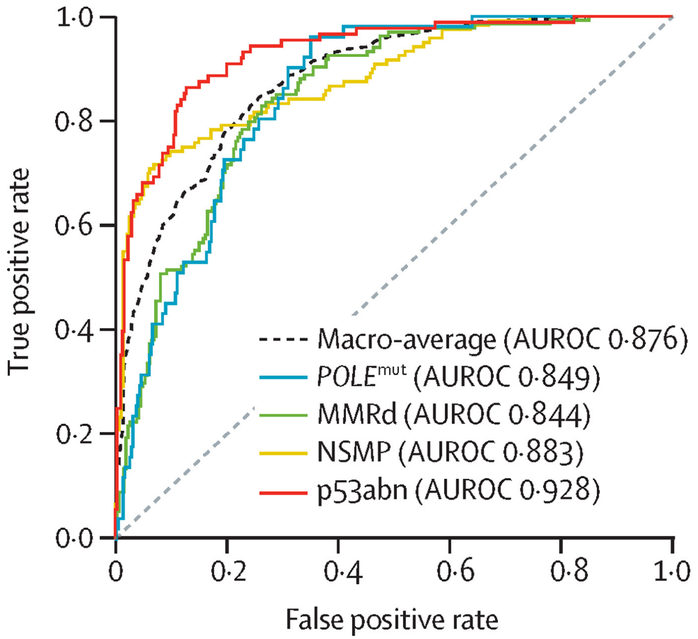
Classification

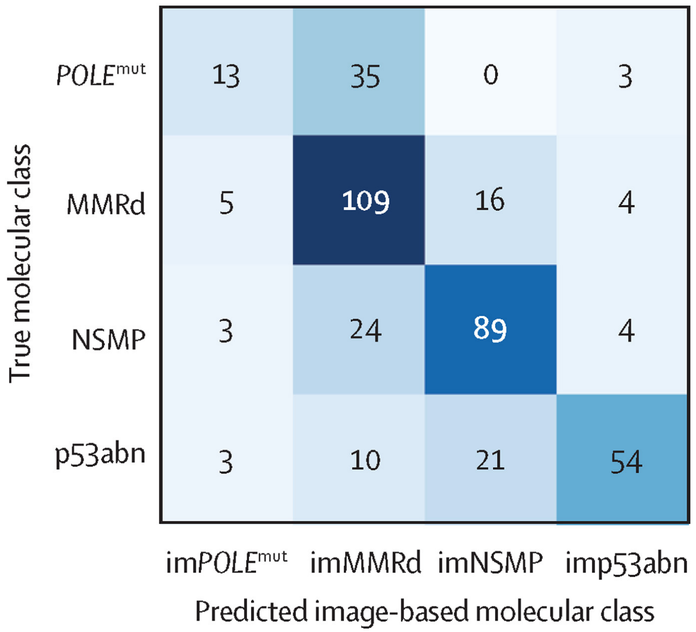


Tile selection

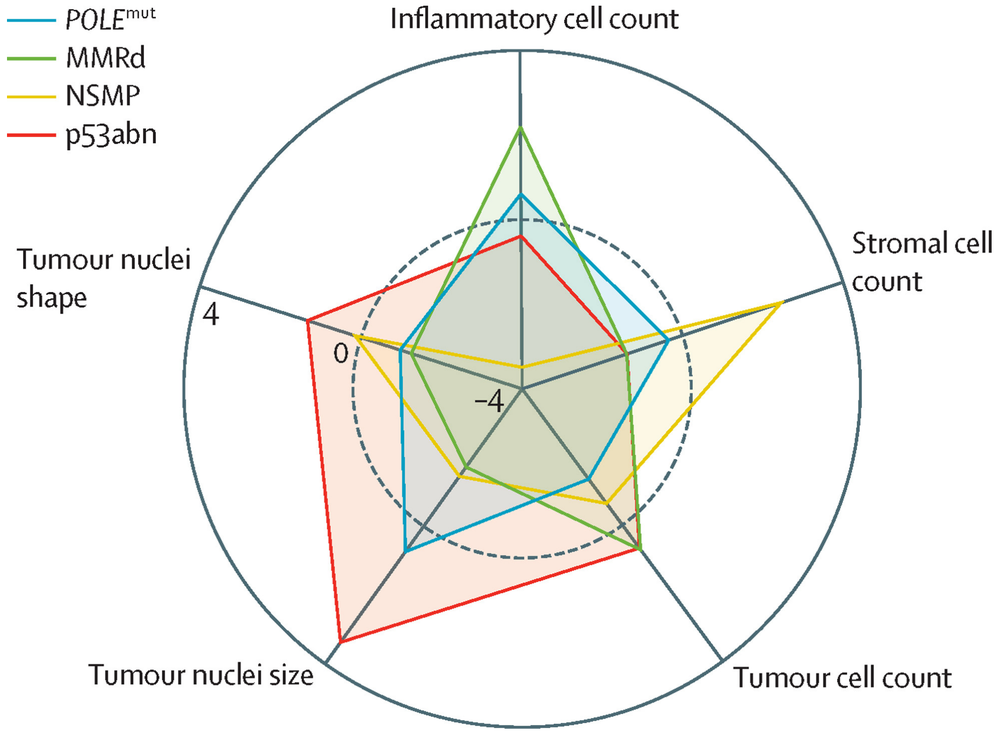
Interpretation

Good performance

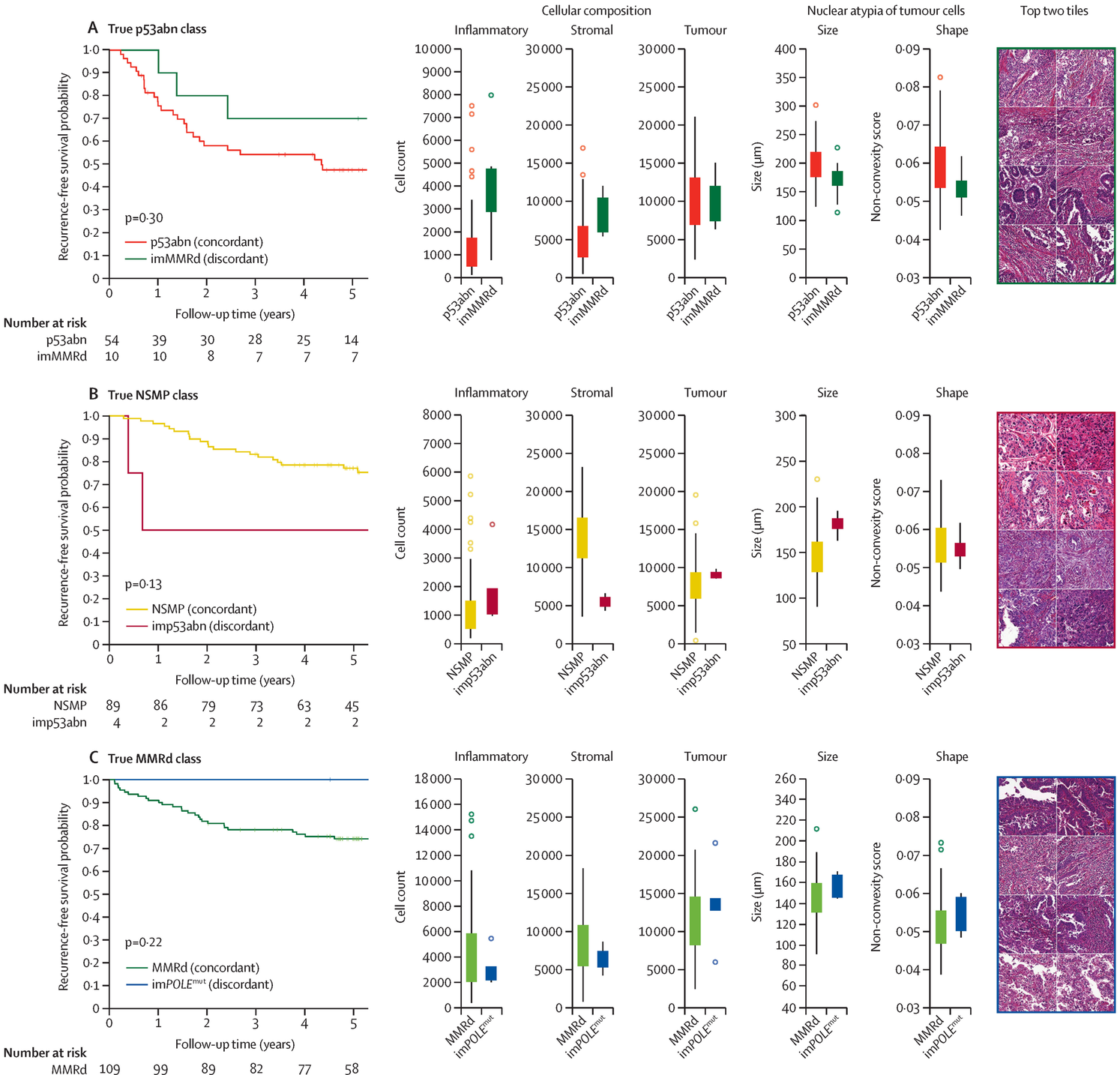




Used a post-hoc analysis using SVM to look investigate model



Examples which were misclassified had discordant outcomes





“You Belong With Me”

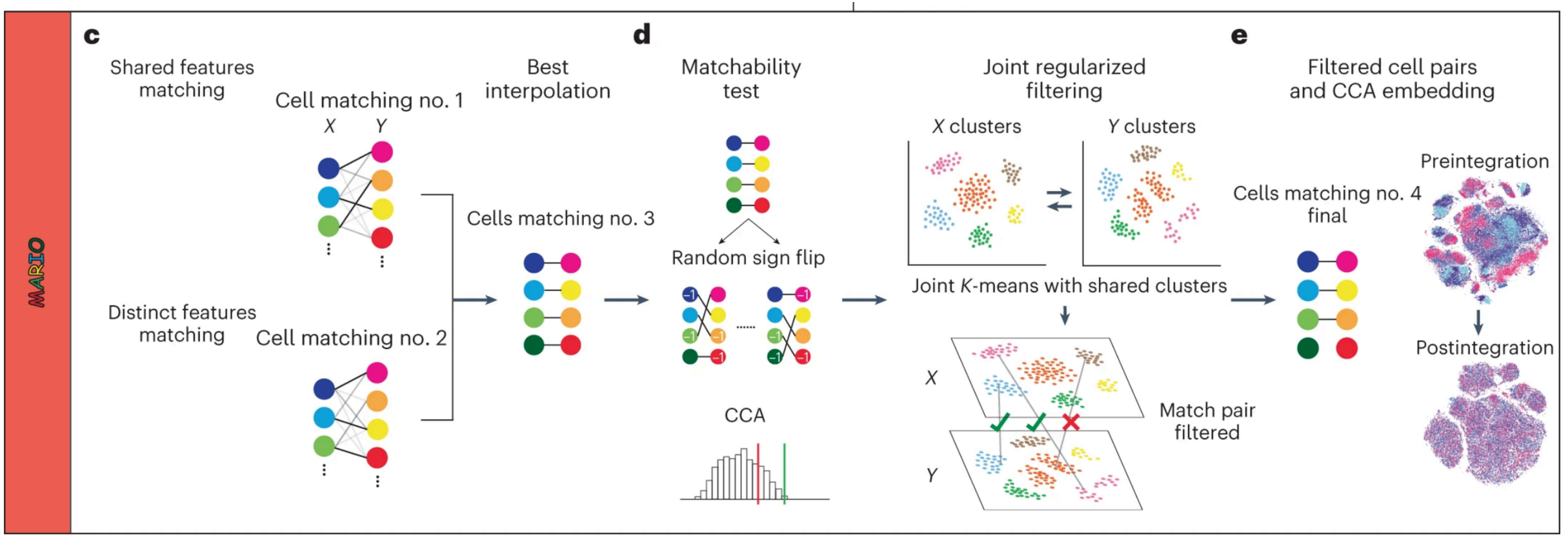
Exploring Interactions and Networks

 https://doi.org/10.1038/s41592-022-01709-7

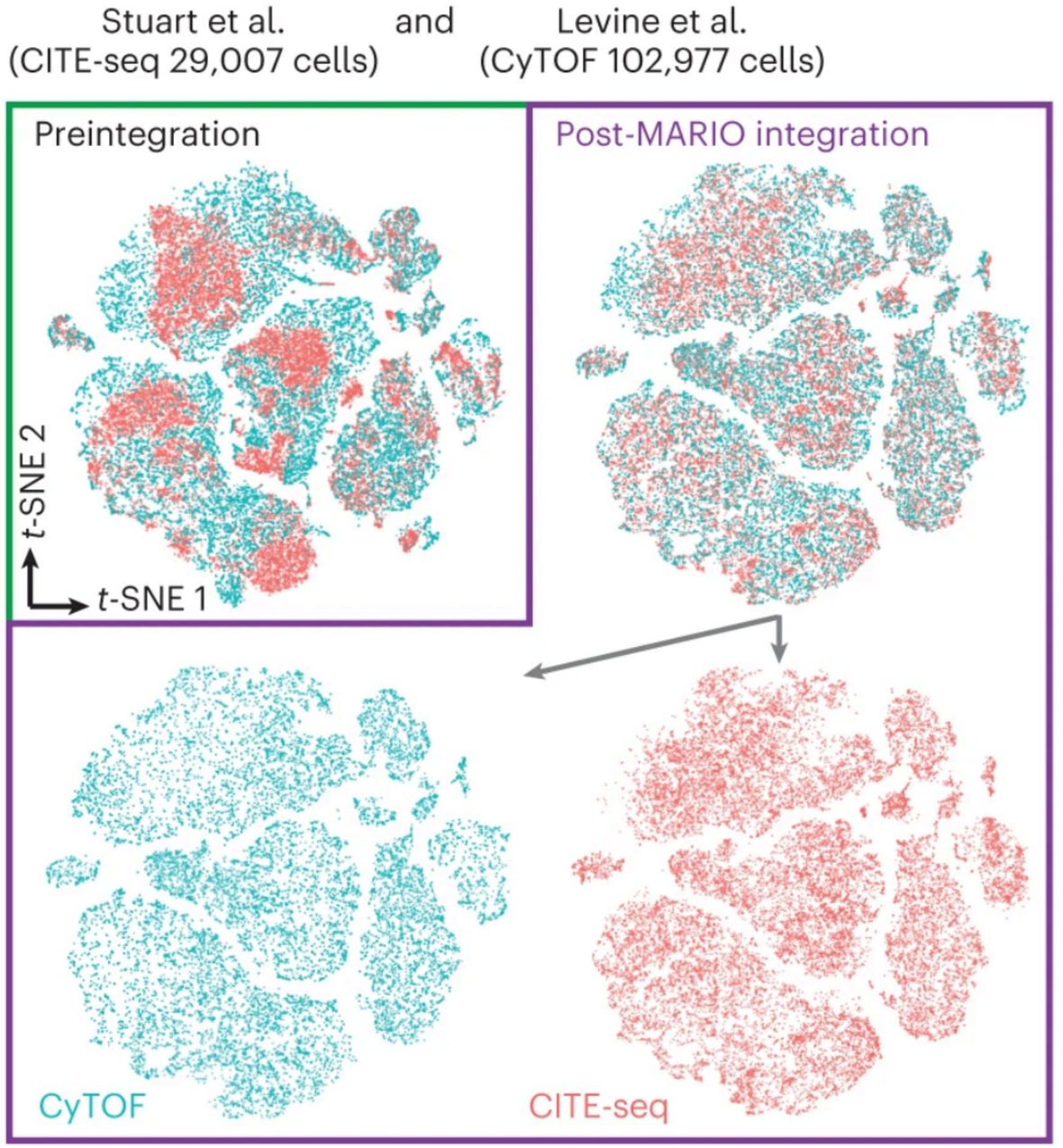
Robust single-cell matching and multimodal analysis using shared and distinct features (Zhu et al, *Nature Methods*)

- Goal: Align multimodal data from multiple single cell experiments
- Method:
  - Flexible model that allows for partial overlap to identify stable matches between cells (MARIO)
  - A little brute force, but also solves a previously intractable problem, so…
- Result: Better matching when compared to benchmarks
- Conclusion: Single cell methods are in the “magic becomes engineering” phase

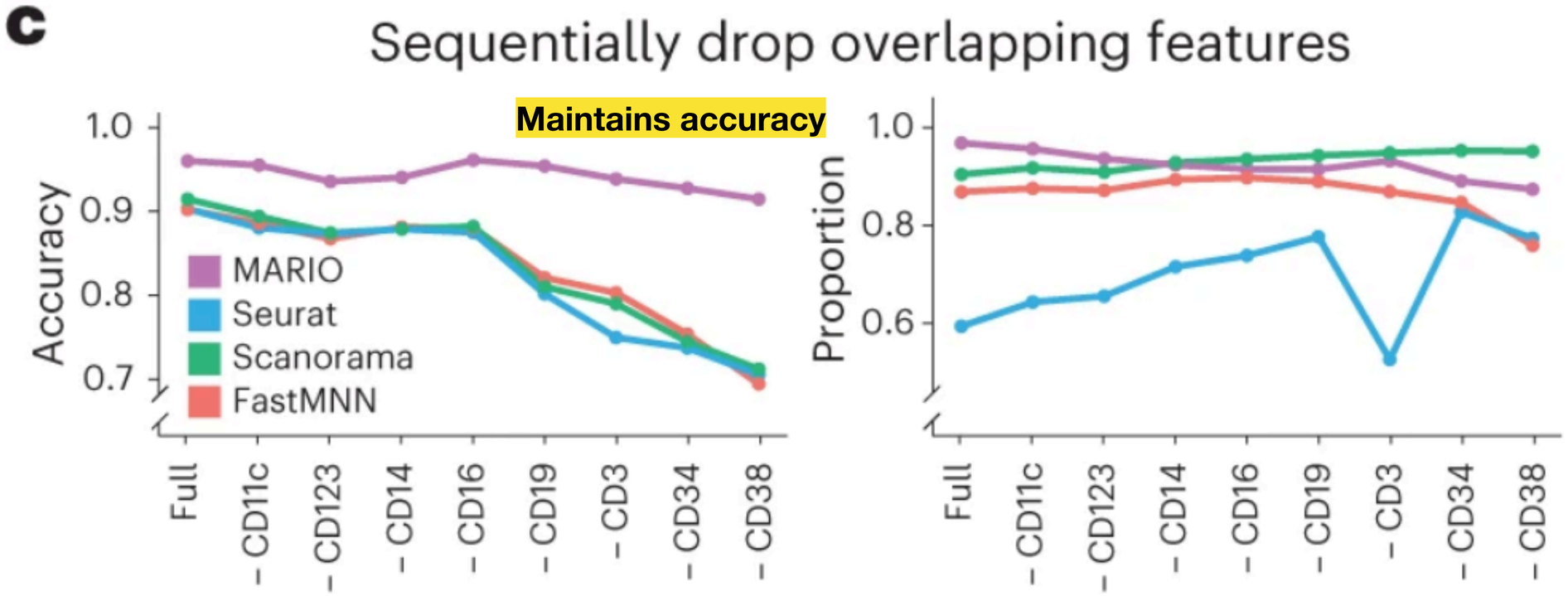



MARIO
c
Shared features matching
Cell matching no. 1
X
Y
Distinct features matching
Cell matching no. 2
Best interpolation
Cells matching no. 3
d
Matchability test
Random sign flip
CCA
Joint regularized filtering
X clusters
Y clusters
Joint K-means with shared clusters
X
Y
Match pair filtered
e
Filtered cell pairs and CCA embedding
Cells matching no. 4 final
Preintegration
Postintegration

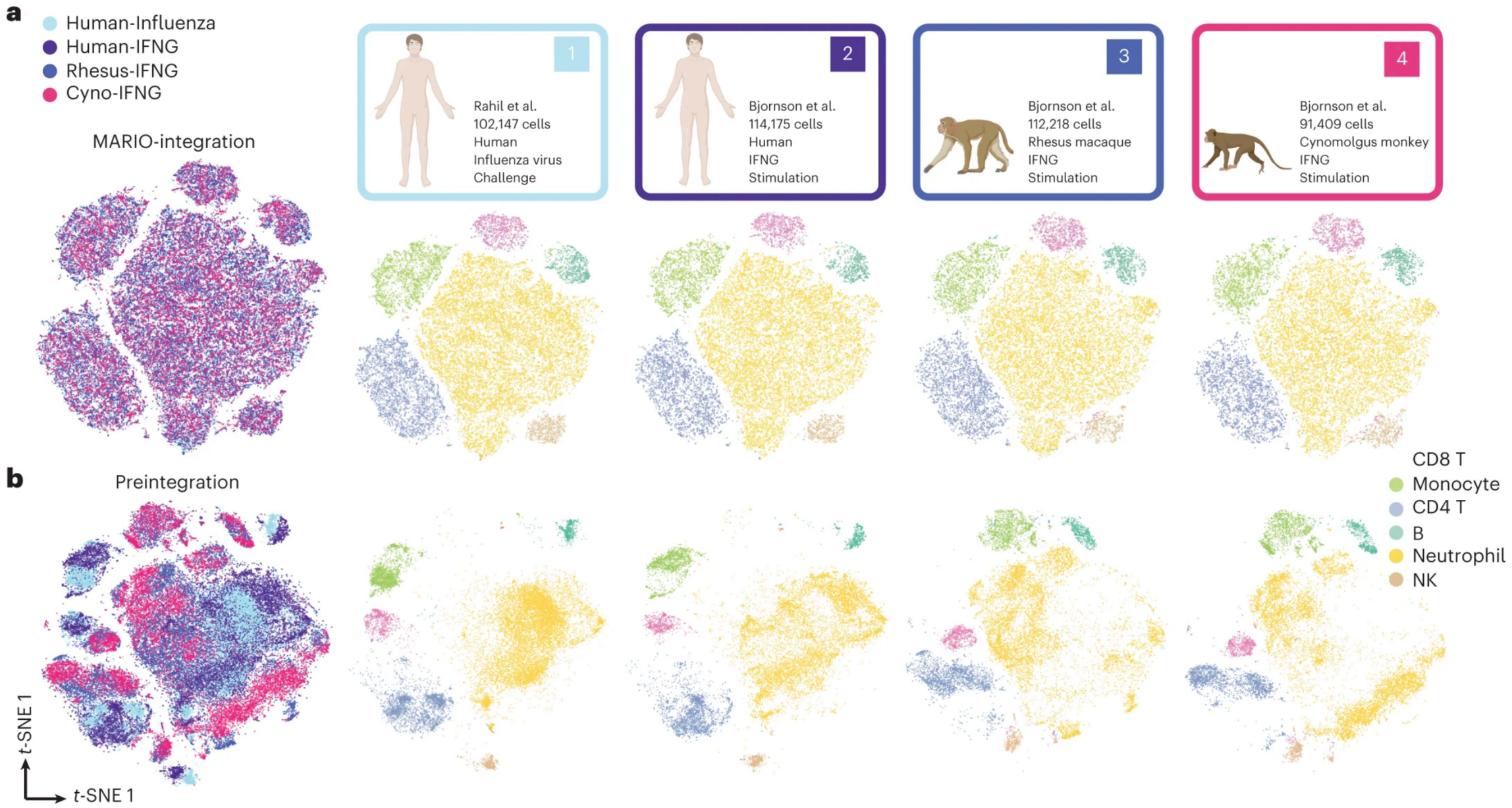
Blotchy and biased :(



Nicely integrated!


c
Sequentially drop overlapping features
Maintains accuracy
Accuracy
1.0
0.9
0.8
0.7
MARIO
Seurat
Scanorama
FastMNN
Full
– CD11c
– CD123
– CD14
– CD16
– CD19
– CD3
– CD34
– CD38
Proportion
1.0
0.8
0.6

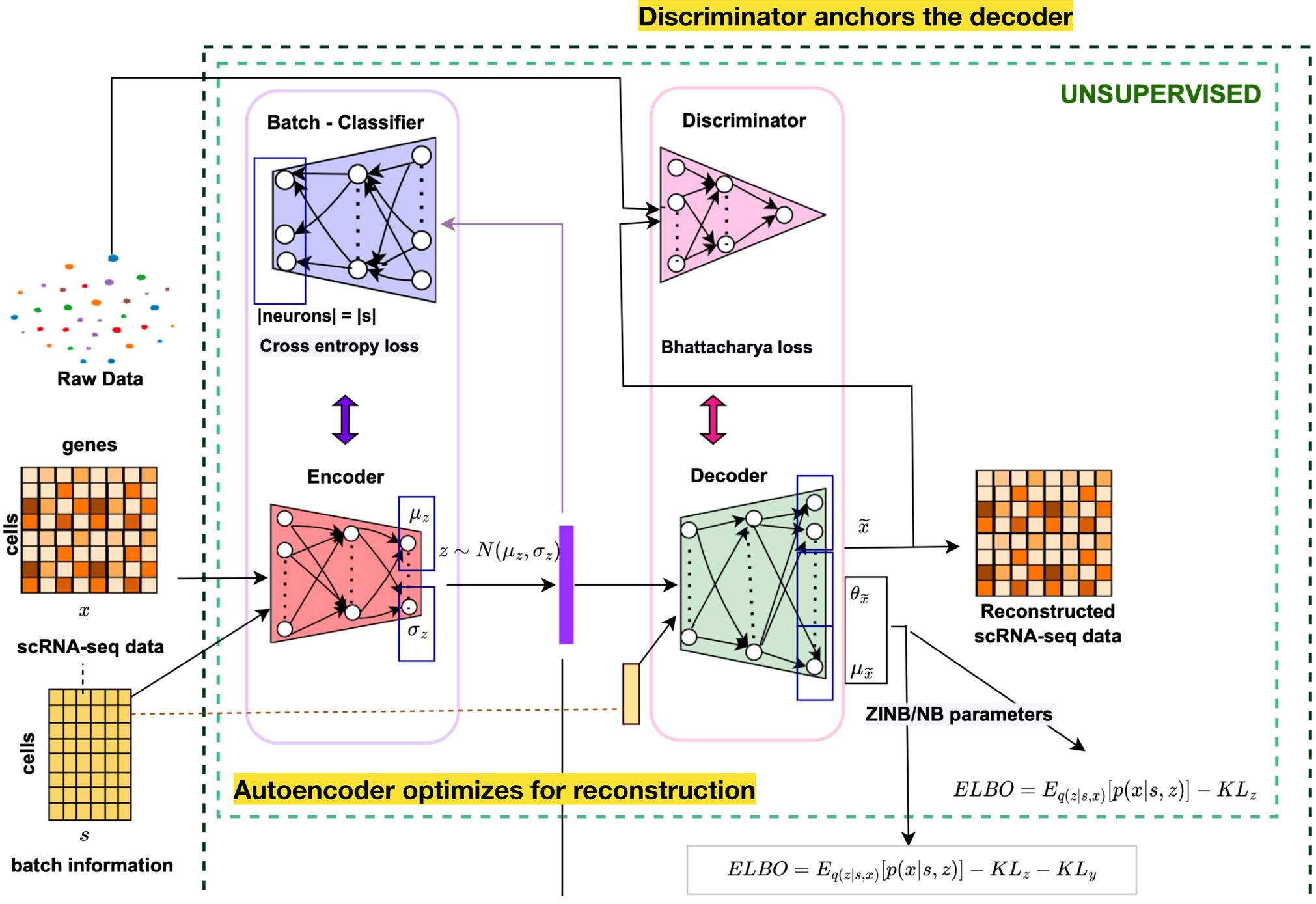
Allows for cross-species integration!



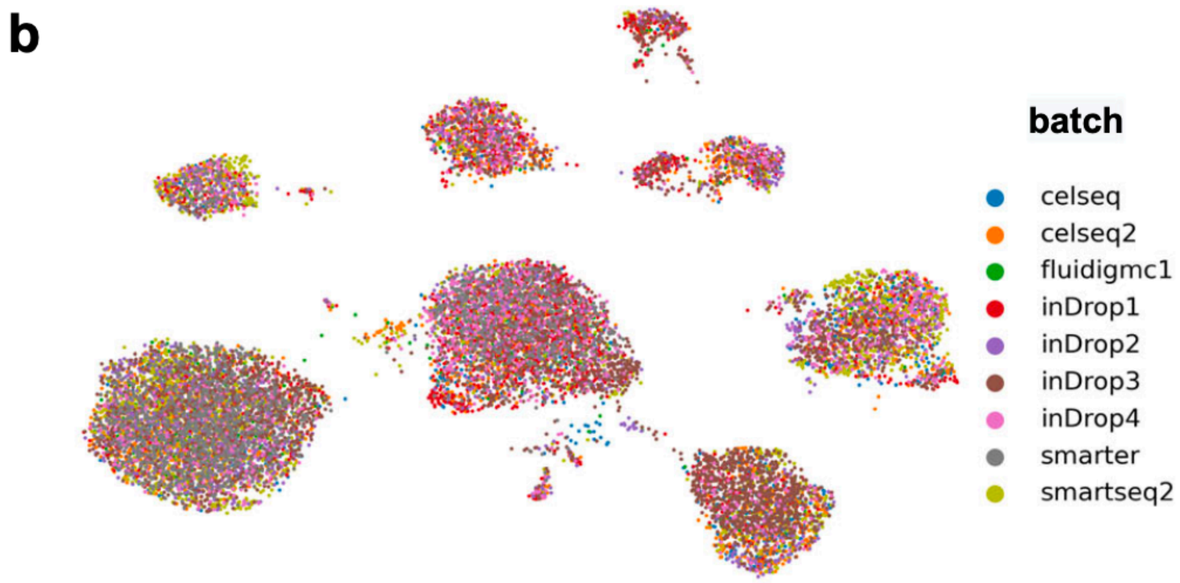
 https://doi.org/10.1038/s41467-023-43590-8

scDREAMER for atlas-level integration of single-cell datasets using deep generative model paired with adversarial classifier (Shree, Pavan, and Zafar, *Nature Communications*)

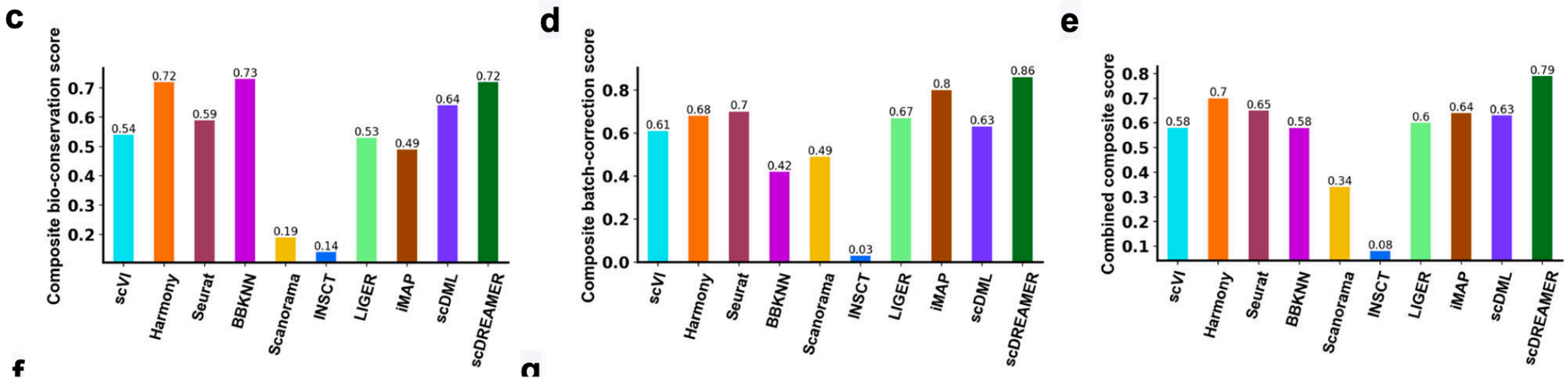
- Goal: Produce a better method to integrate disparate single-cell datasets
- Method:
  - Take a combo unsupervised, semi-supervised, and supervised approach
  - Unsupervised uses adversarial variational autoencoder + batch classifier for learning embeddings
  - Supervised uses ^^ plus cell type classifier
  - Train on 1mm cells and 147 batches
- Result: Overcomes known integration challenges:
  - Nested batch effects, skewed cell types among batches, and conservation of development trajectory
- Conclusion: It’s a good time to be a single cell scientist


Discriminator anchors the decoder
UNSUPERVISED
Batch - Classifier
|neurons| = |s|
Cross entropy loss
Discriminator
Bhattacharya loss
Raw Data
genes
cells
x
scRNA-seq data
cells
s
batch information
Encoder
μ_z
σ_z
$z \sim N(\mu_z, \sigma_z)$
Decoder
$\widetilde{x}$
$\theta_{\widetilde{x}}$
$\mu_{\widetilde{x}}$
Reconstructed scRNA-seq data
ZINB/NB parameters
Autoencoder optimizes for reconstruction
$ELBO = E_{q(z|s,x)}[p(x|s,z)] - KL_z$
$ELBO = E_{q(z|s,x)}[p(x|s,z)] - KL_z - KL_y$

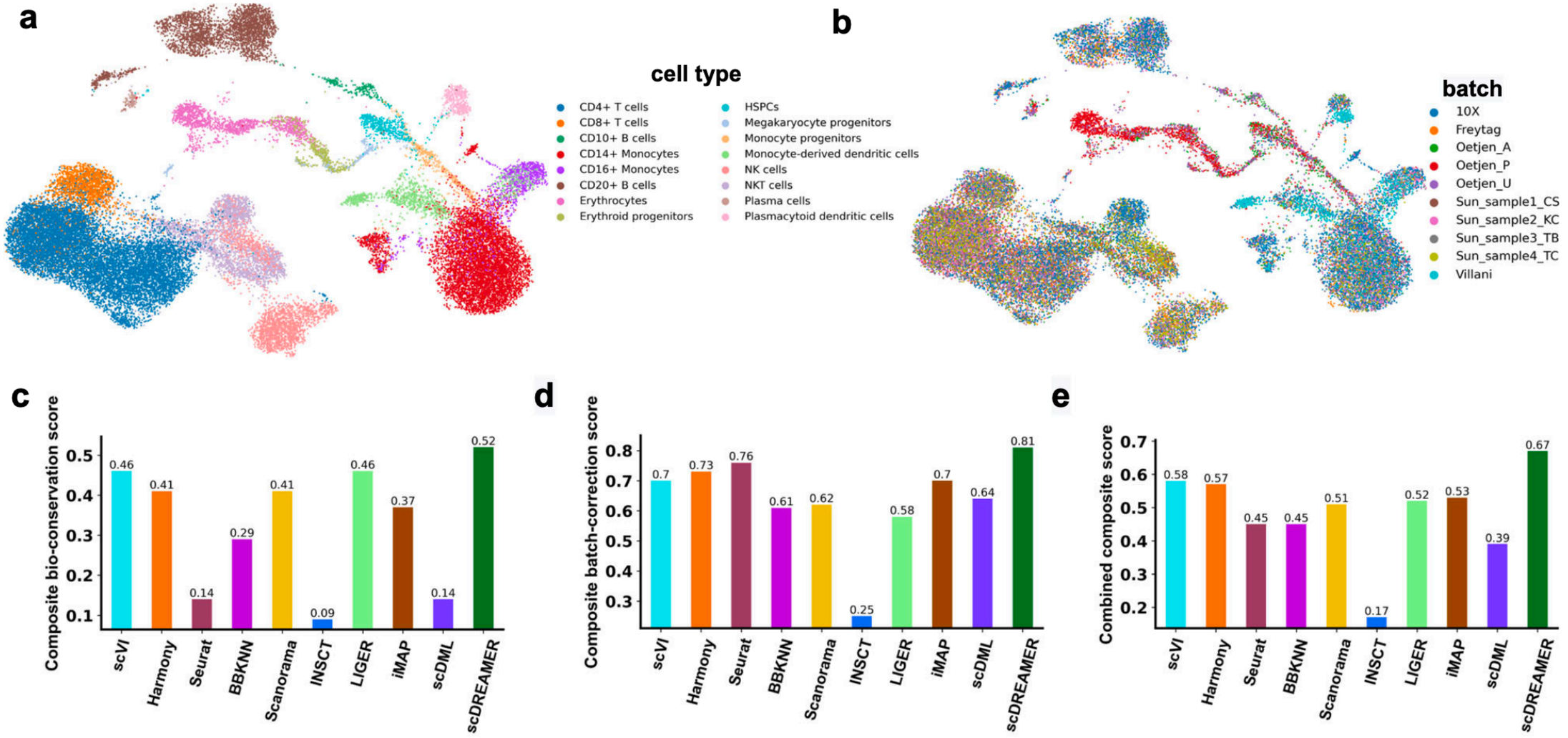
Look Ma, no batches!



Best across a lot of metrics



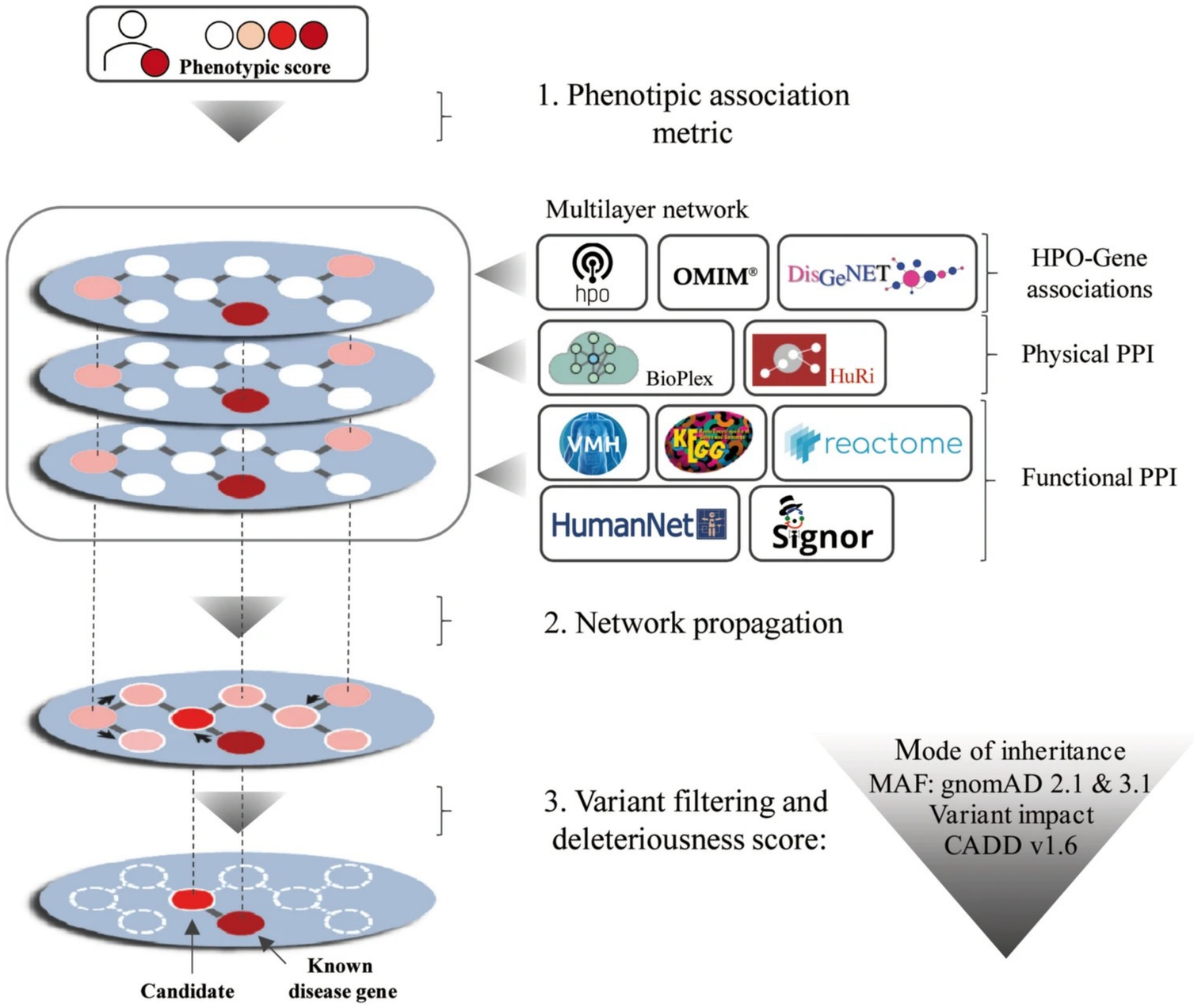
Integrates well even when there's cell type skewing between batches



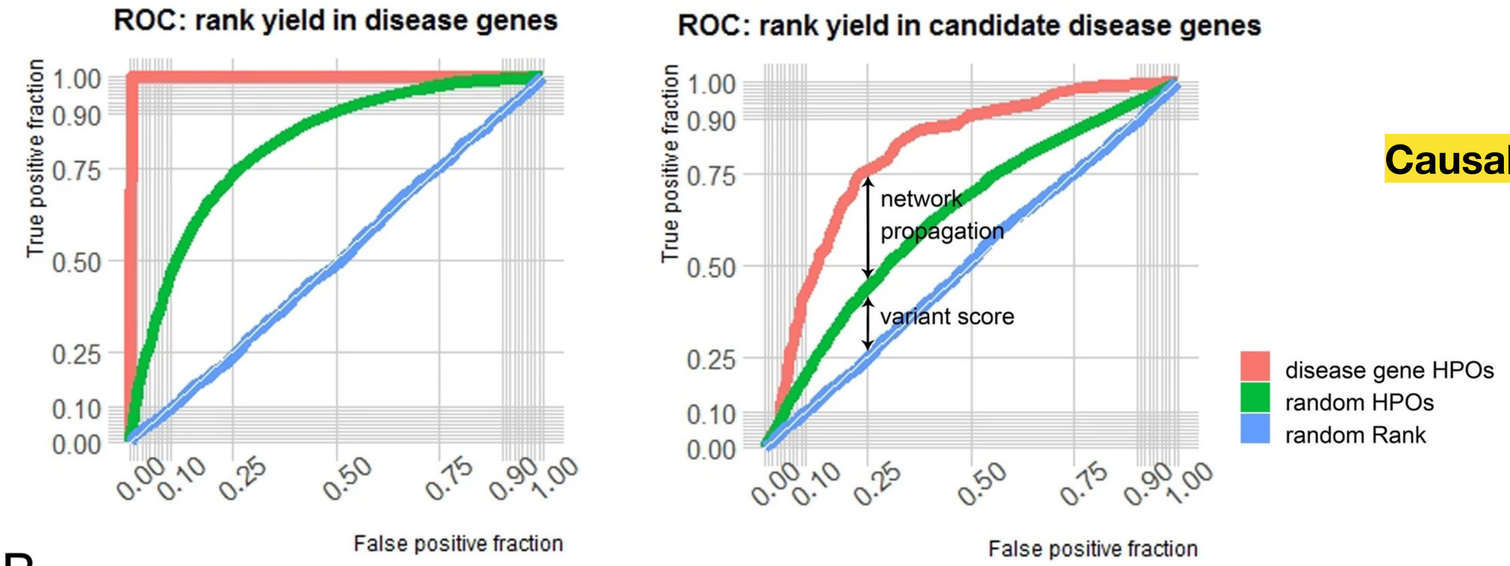
 https://doi.org/10.1186/s13073-023-01214-2

ClinPrior: an algorithm for diagnosis and novel gene discovery by network-based prioritization (Schlüter, Vélez-Santamaría, and Verdura et al, *Genome Medicine*)

- Goal: Improve discovery of causal rare variants
- Method: Use “interaction network-based” parameters as priors
- Result:
  - achieved diagnostic yield of 70%
  - including 10 novel candidates not previously associated with disease
  - 7 of which were functionally validated
- Conclusion: Rigorous ontological development to the rescue

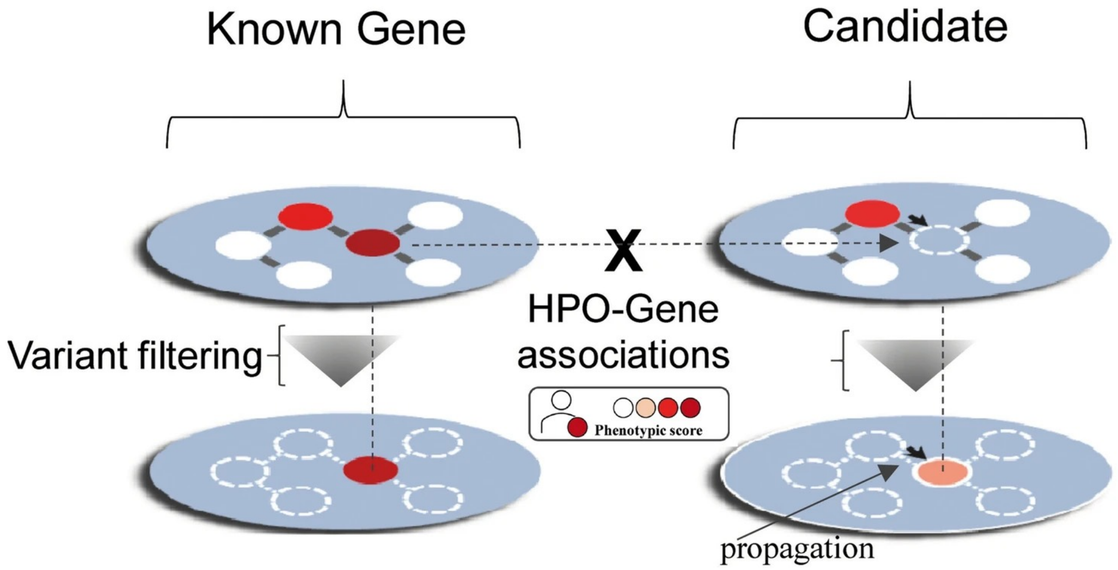

Phenotypic score
1. Phenotipic association metric
Multilayer network
hpo
OMIM®
DisGeNET
HPO-Gene associations
BioPlex
HuRi
Physical PPI
VMH
KEGG
reactome
HumanNet
Signor
Functional PPI
2. Network propagation
3. Variant filtering and deleteriousness score:
Mode of inheritance
MAF: gnomAD 2.1 & 3.1
Variant impact
CADD v1.6
Candidate
Known disease gene

A

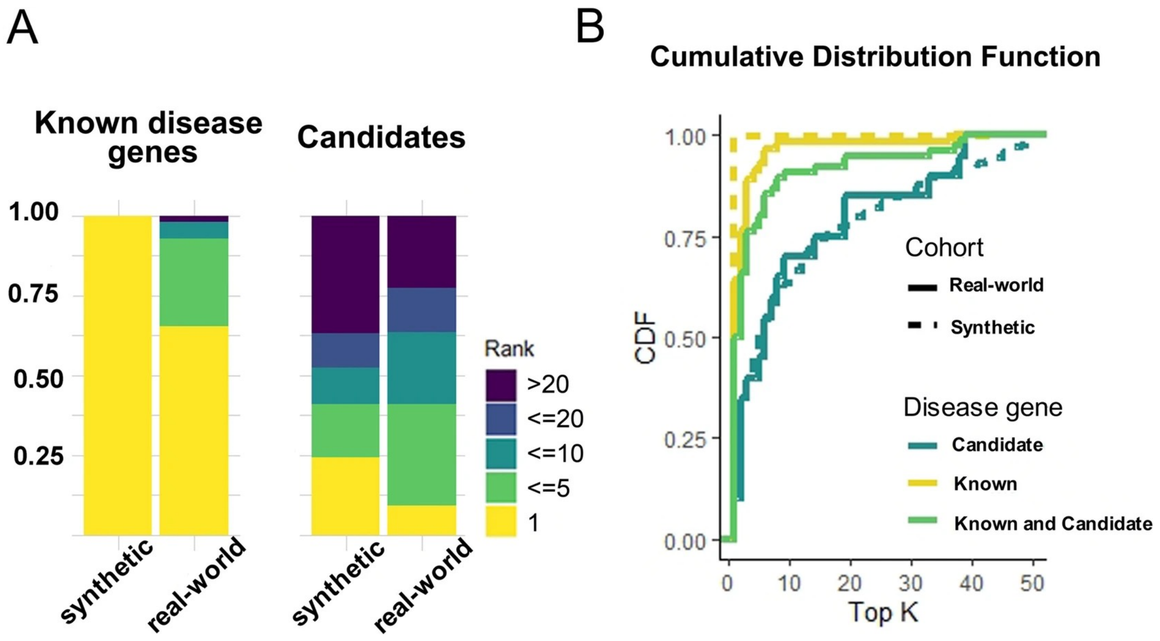


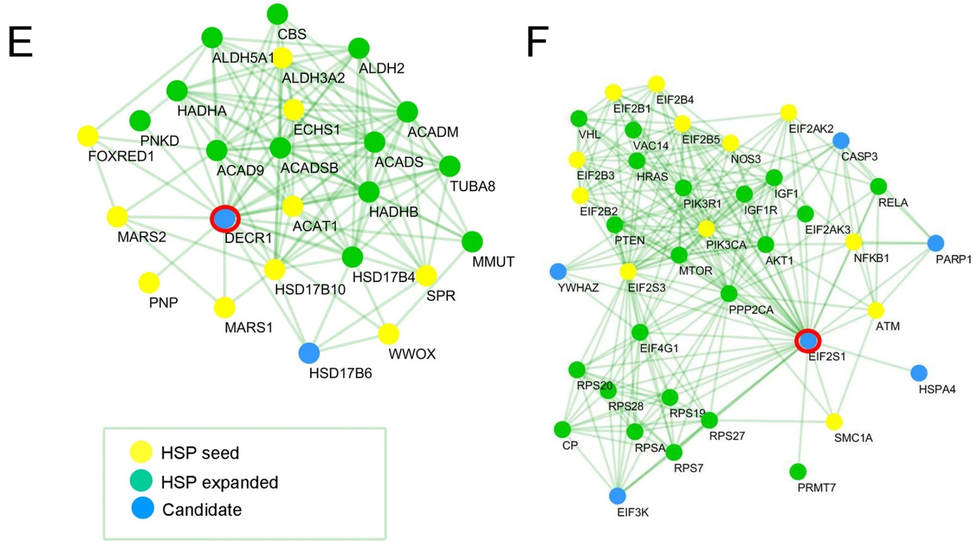
Causal gene is hidden and rediscovered

B



Good performance concordance between synthetic and real-world data





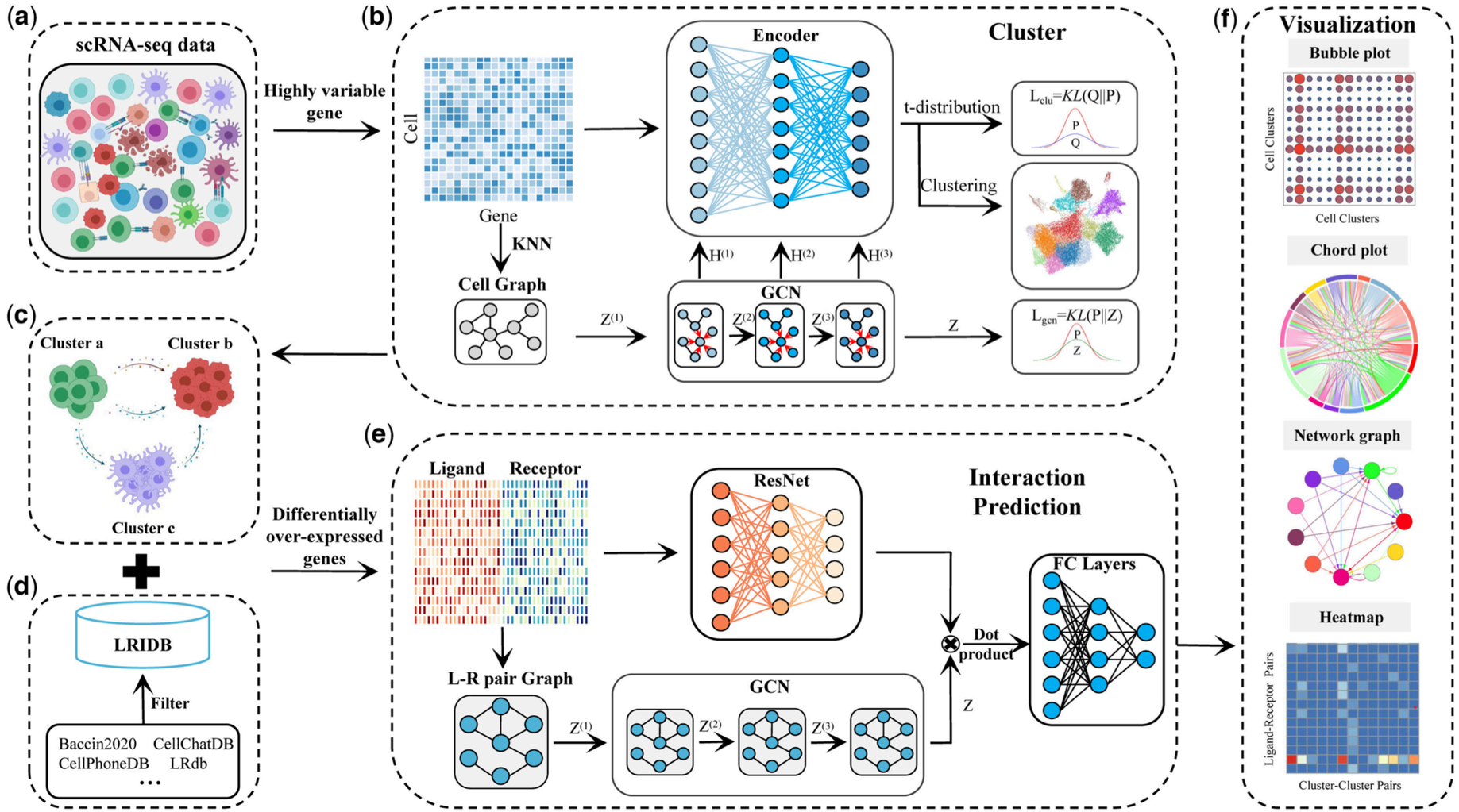
No network study is complete with a hairball!

DeepCCI: a deep learning framework for identifying cell–cell interactions from single-cell RNA sequencing data (Yang et al, *Bioinformatics*)

- Goal: Cell-cell interactions are critical for differentiation, immune response, cellular homeostasis, all kinds of things
  - but scRNA Seq loses this relationship
  - Use deep learning to rebuild cell-cell interactions
- Method:
  - Graph convolutional neural networks (GCN) and ResNet for cell clustering and cell-cell interaction identification
  - Evaluate on several publicly available scRNASeq datasets, validate against a set of known cell-cell interactions
- Result:
  - Best in class performance and lots of pretty plots
- Conclusion: Good to use while you're waiting on your spatial single cell experiments to run

Autoencoder first

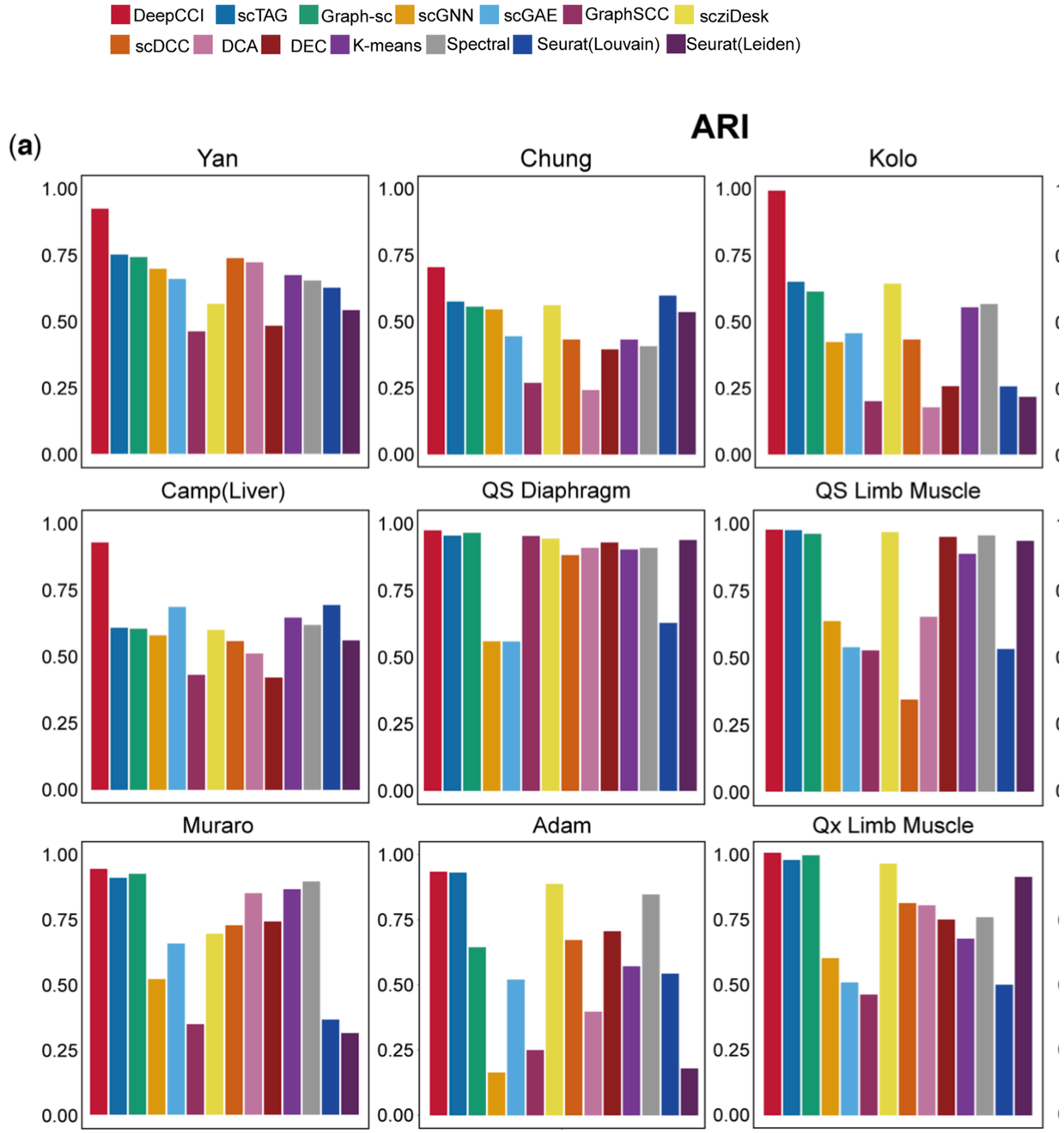
Unsupervised clustering



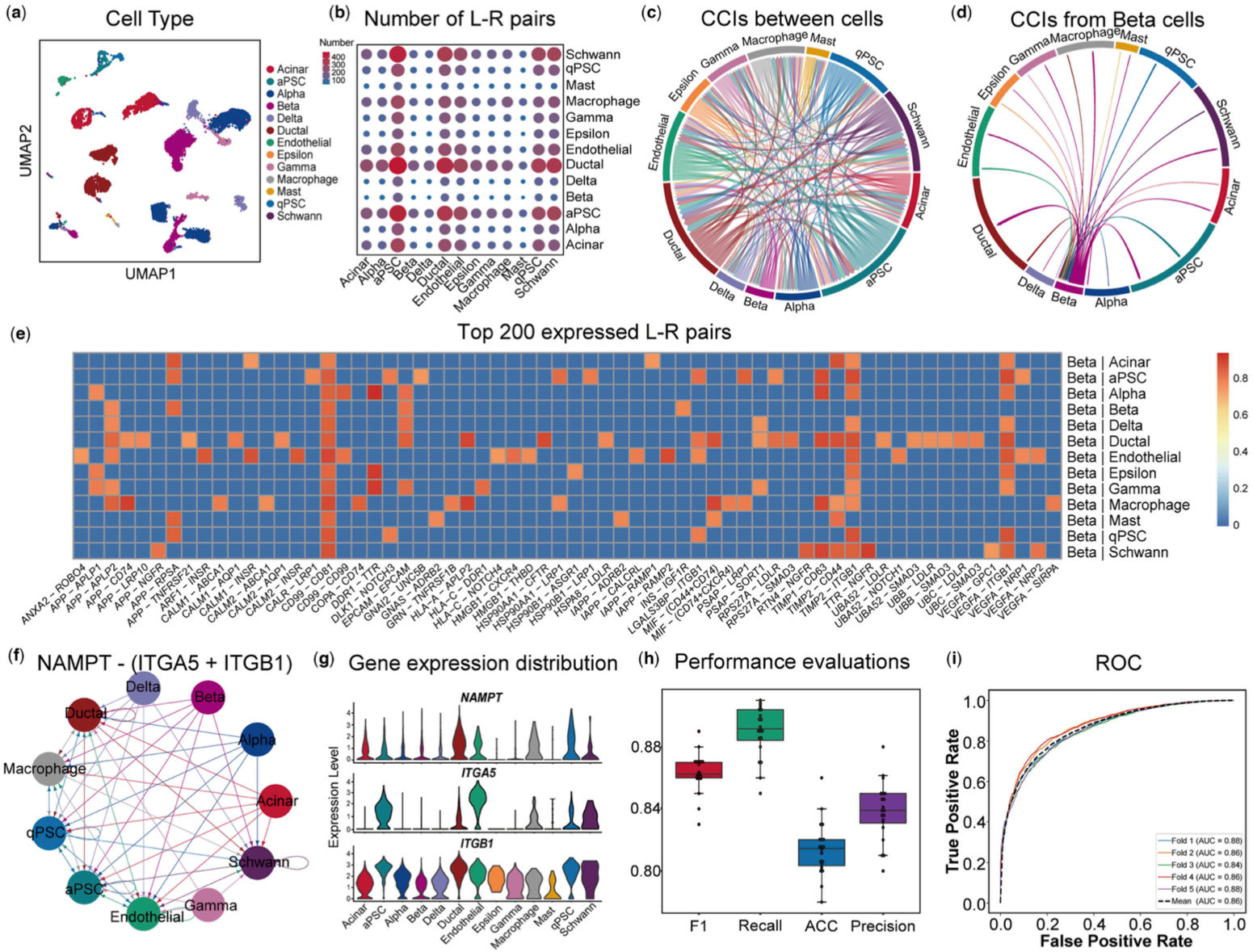
Cell-cell interaction predictions and data viz

Training labels for CCI

Compared to several other methods … looks good



Pretty



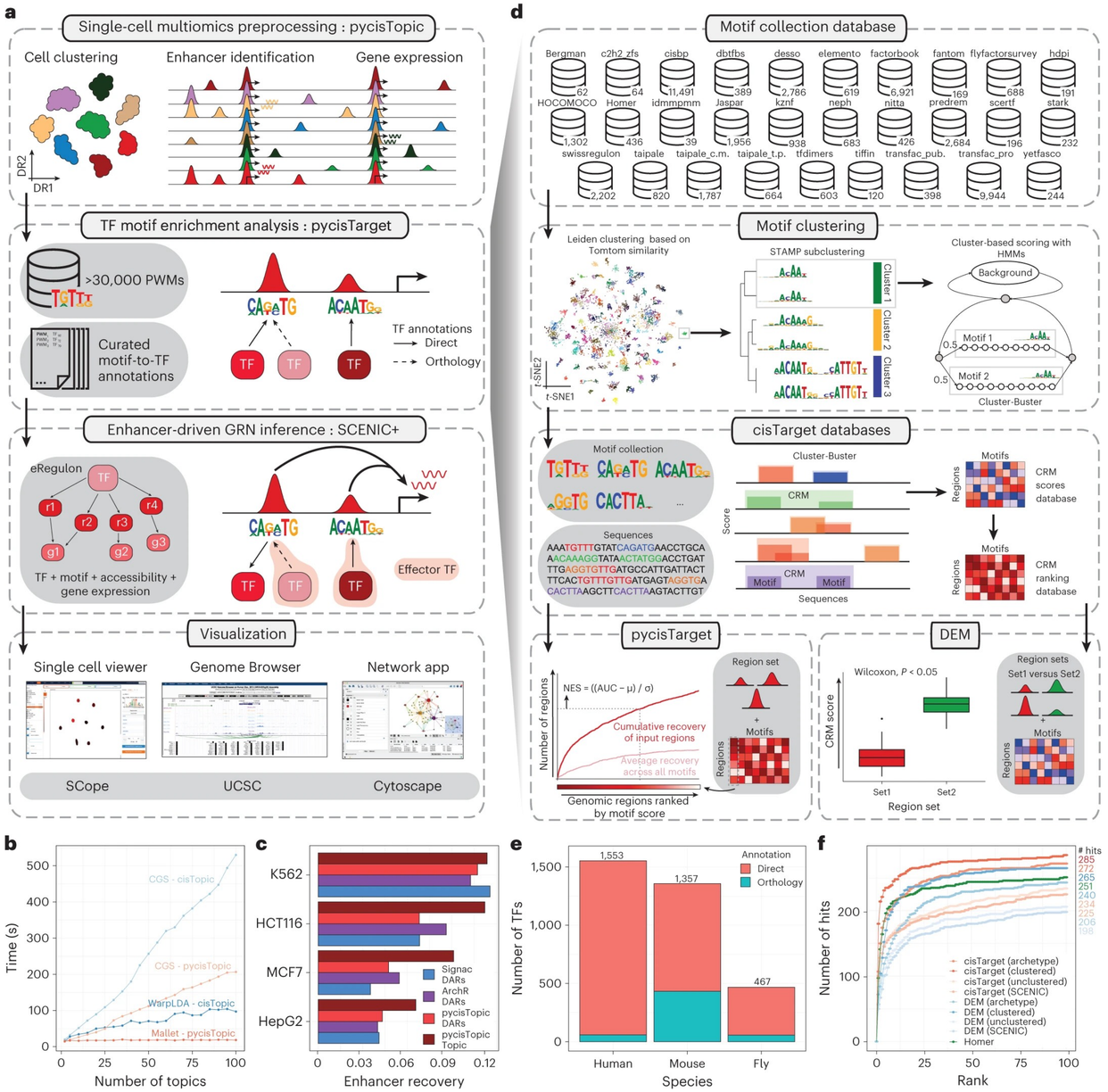
Performance at predicting CCIs

 https://doi.org/10.1038/s41592-023-01938-4

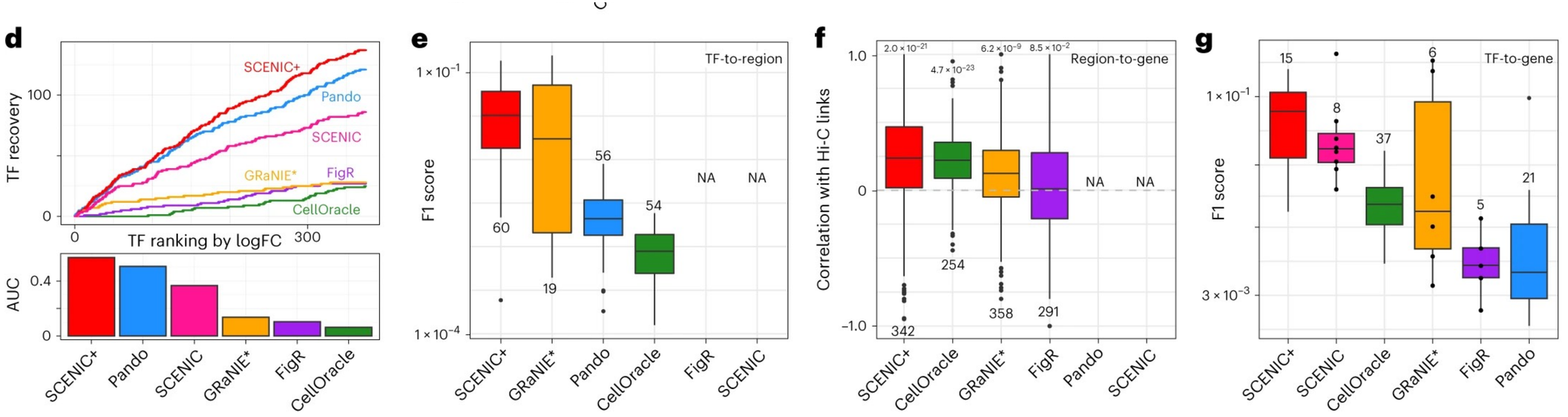
SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks (González-Blas and De Winter et al, *Nature Methods*)

- Goal: Derive single cell gene regulatory networks
- Method:
  - Build an integrative bioinformatics pipeline to identify TFs
  - Using sc-ATACSeq and scRNA-Seq
  - Integrate established bioinformatics knowledge bases
- Result:
  - Curated a TF motif collection of 30,000 motifs
  - State-of-the-art performance on external and diverse benchmarks
- Conclusion: Tour-de-force for single cell analysis

Massive integrative pipeline



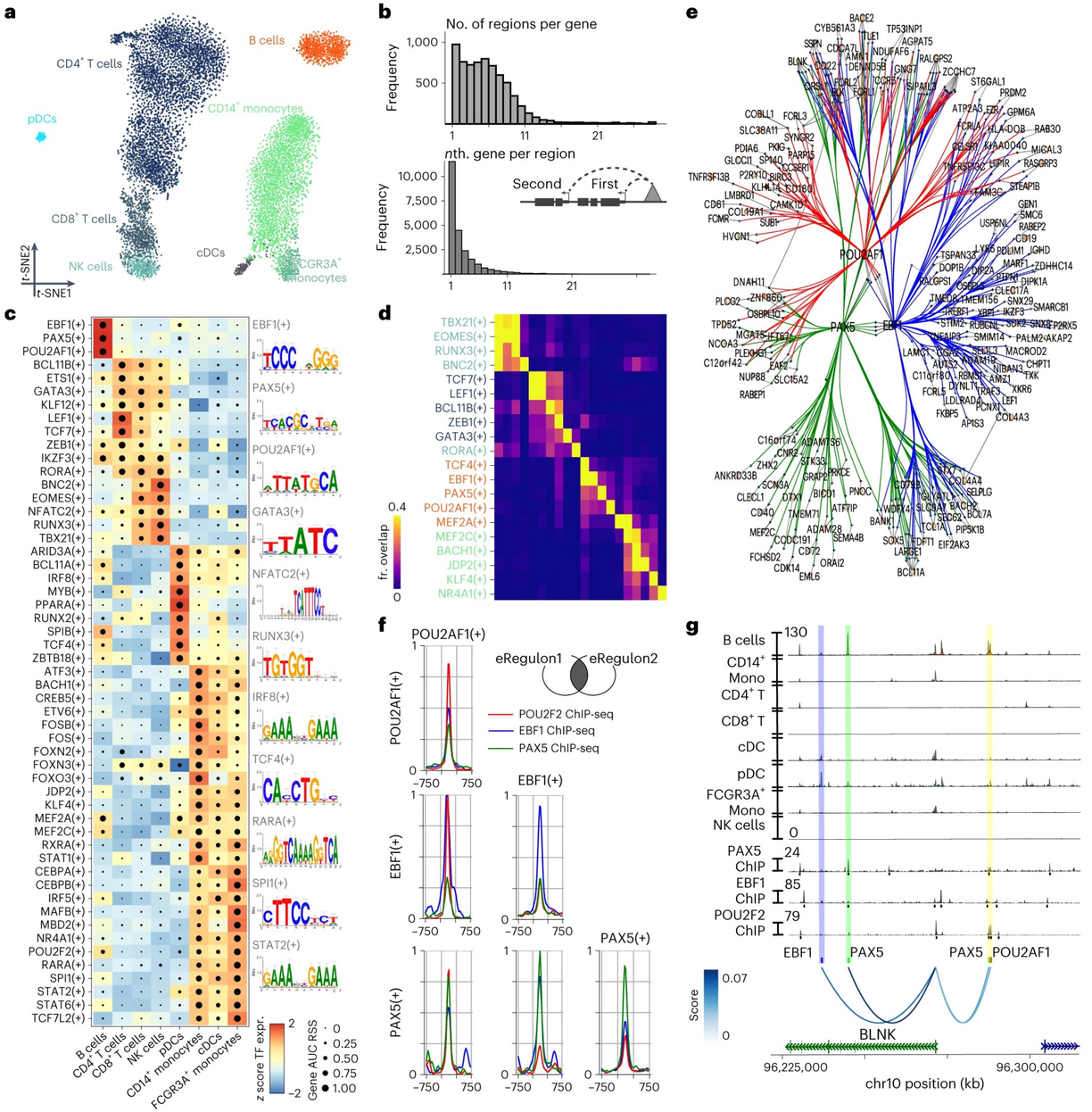
Excellent performance on range of benchmarks



TF-to-region **Region-to-gene** **TF-to-gene**

#IS24 #YIR24 @proftatonetti

Rediscovered well established TF motifs

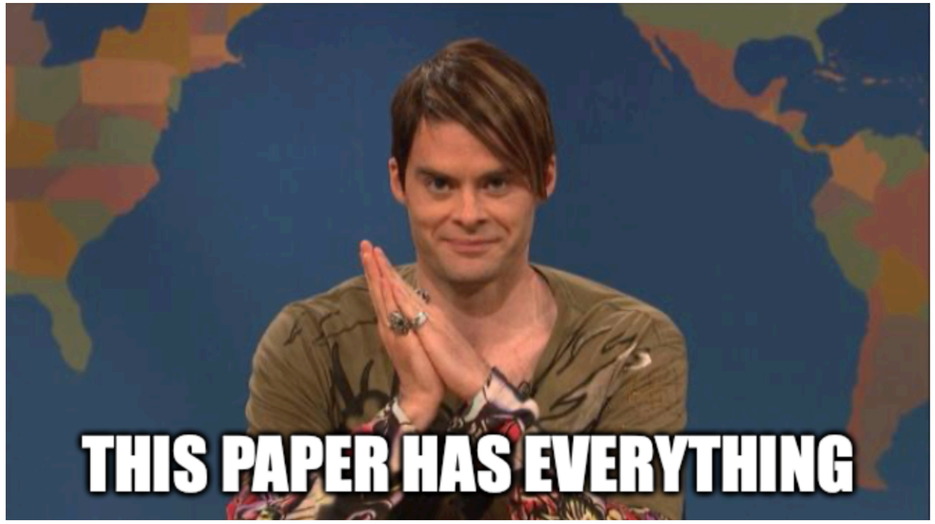


Build new Gene Regulatory Networks

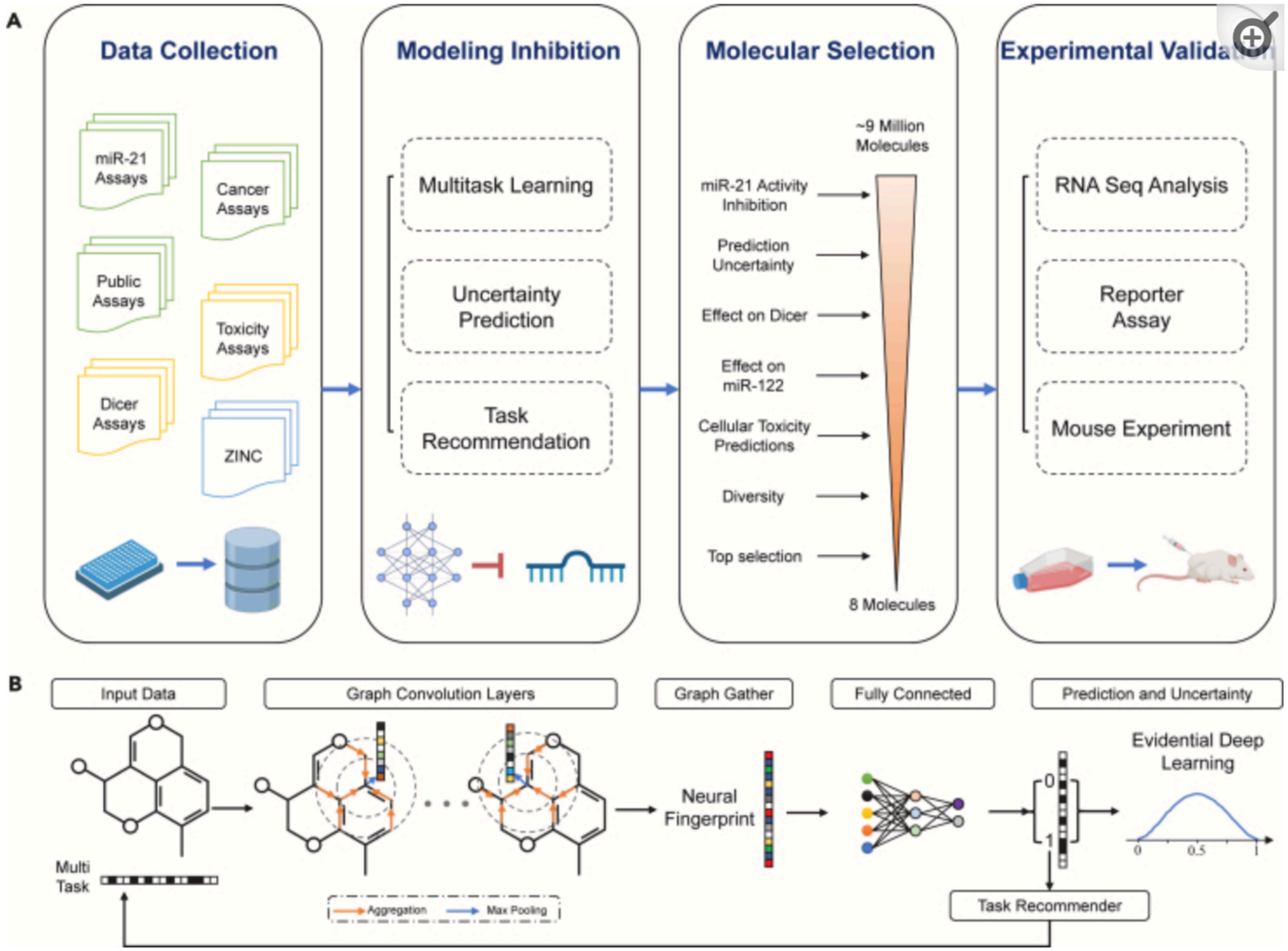
 https://doi.org/10.1016/j.patter.2023.100909

Functional microRNA-targeting drug discovery by graph-based deep learning (Arshadi et al, *Patterns*)

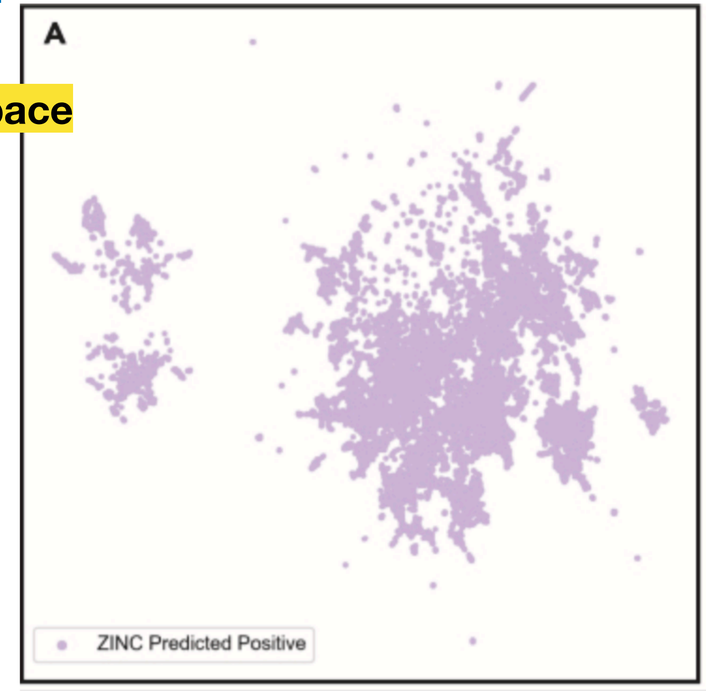
- Goal: Identify small molecules to target oncogenic microRNAs (typically very hard to drug)
- Method:
  - Graph-CNNs + post-hoc toxicity ranking
  - Screen pool of 9 million molecules
  - Identified 8, evaluated in cells
  - tested top candidate in xenograft mice
- Result: significantly reduced lung metastasis
- Conclusion: —->



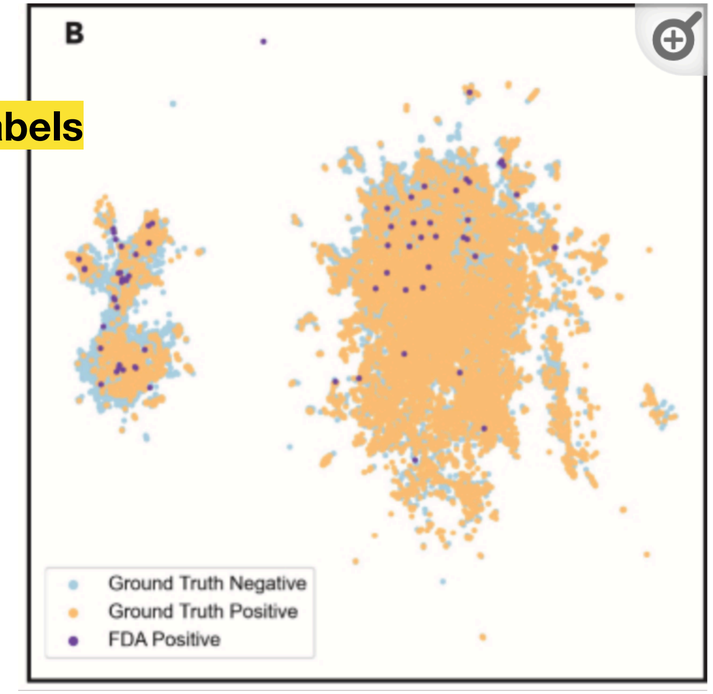
- Learning multiple tasks like this can degrade the performance of a specific task of interest (i.e. miR-21 inhibition)
- Task recommendation will prioritize a target task and identify which tasks support
- This is learned along with everything else



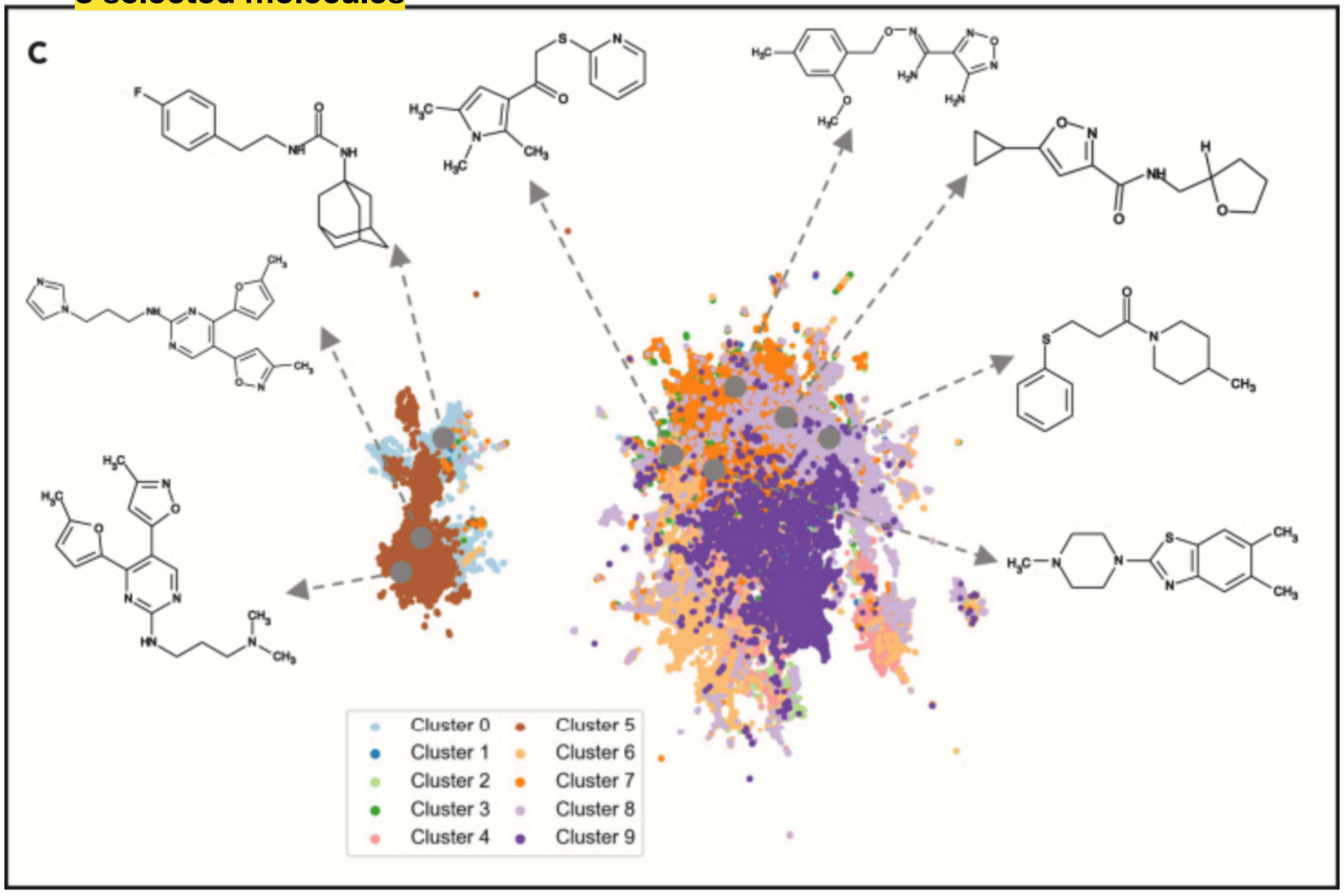
Space



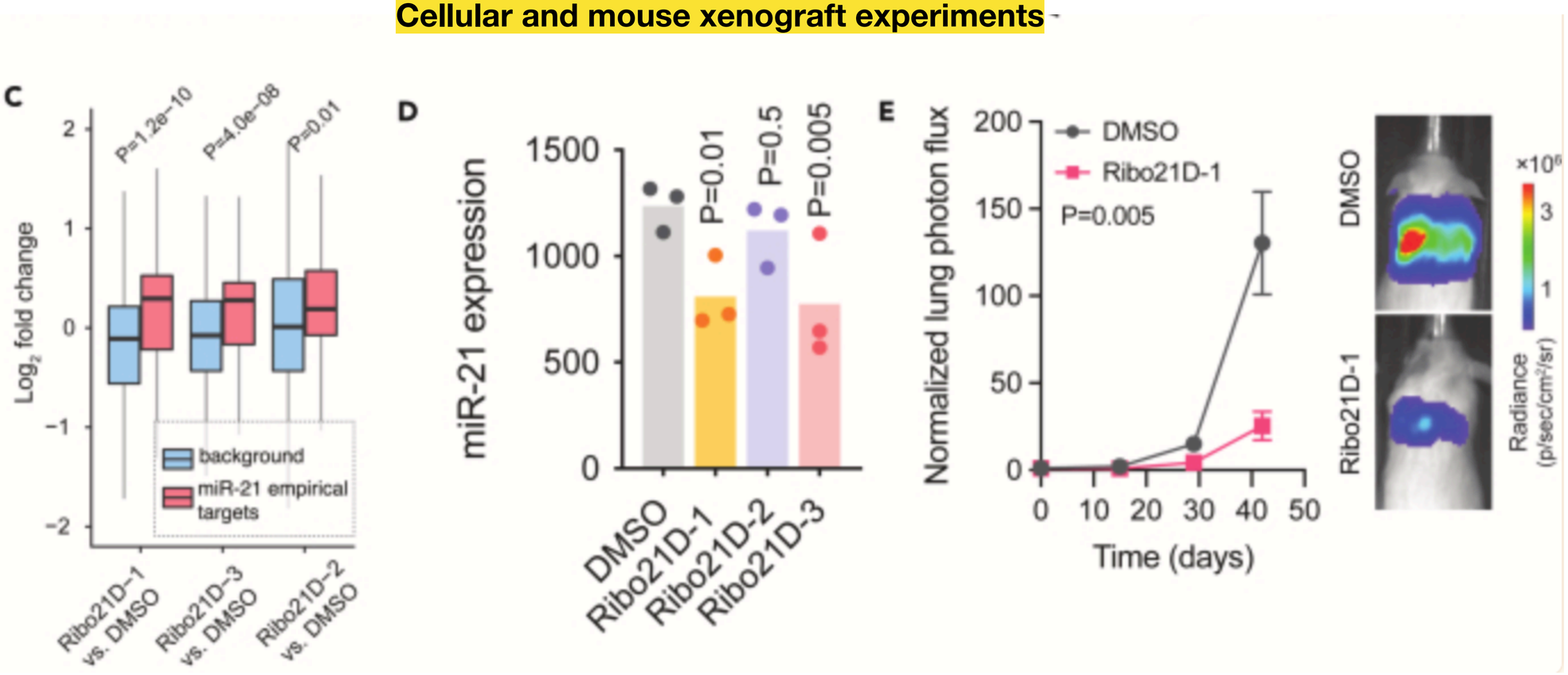
Labels



8 selected molecules



Cellular and mouse xenograft experiments



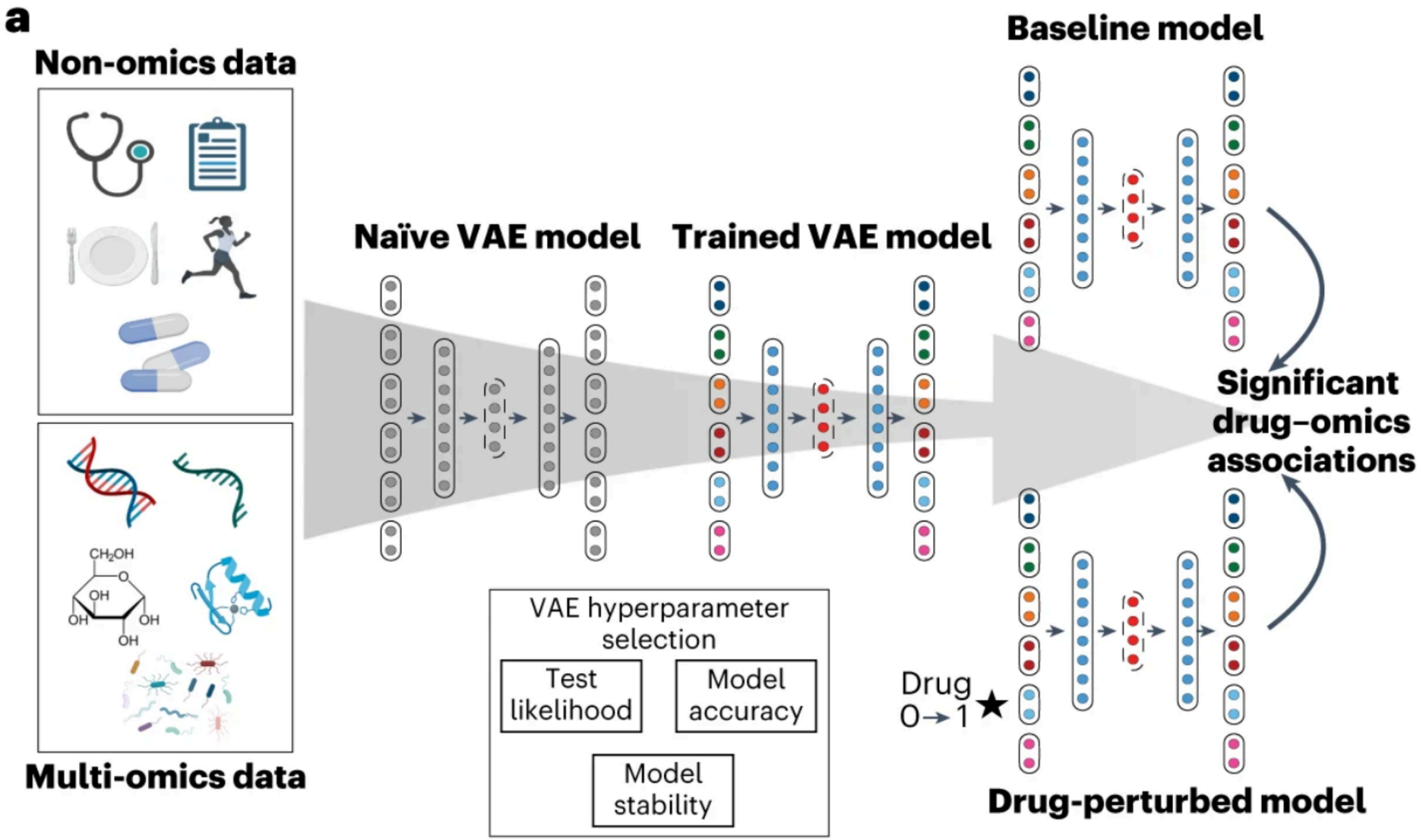


"High Hopes"

Pushing the Boundaries in Drug Discovery

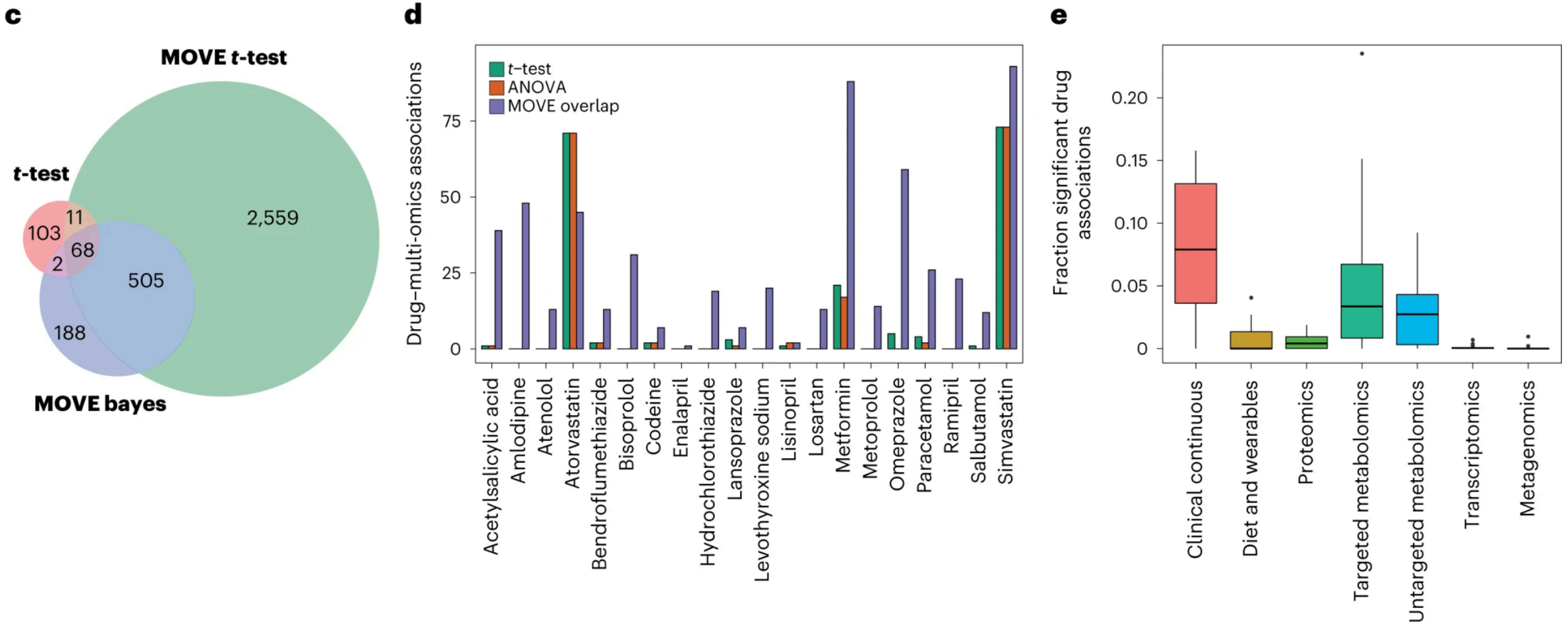
Discovery of drug–omics associations in type 2 diabetes with generative deep-learning models (Allesøe et al, *Nature Biotechnology*)

- Goal: Integrate multiomics data in a seamless and *interrogate-able* way
- Method:
  - variational auto-encoders
  - cohort of 789 Type 2 Diabetes patients with genomics, proteomics, transcirptomics, metabolomics, and microbiome + data from medication, diet questionnaires, and clinical measurements
  - Statistical analysis leveraging generative encoders
- Result: Founds lots of associations, some of which are even real
- Conclusion: Generative models and statistical analysis are an exciting combination!

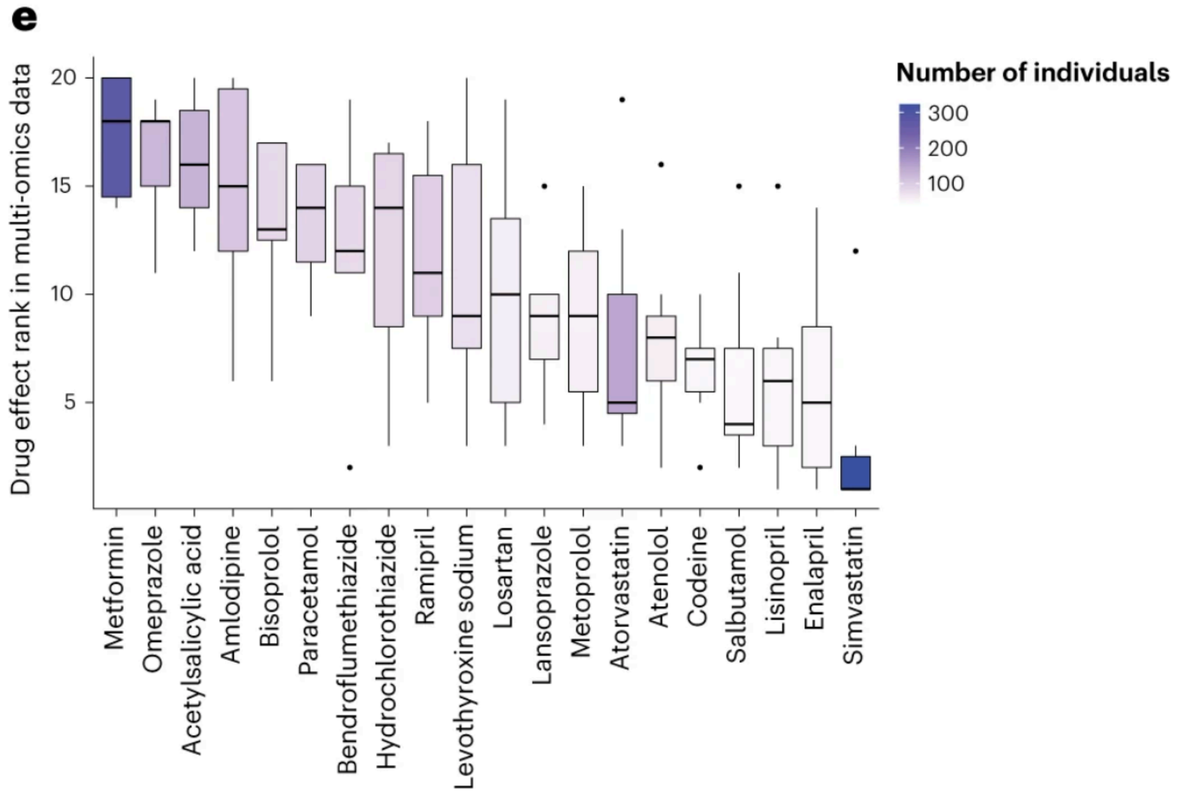

a
Non-omics data
Multi-omics data
Naïve VAE model
Trained VAE model
Baseline model
Significant drug–omics associations
VAE hyperparameter selection
Test likelihood
Model accuracy
Model stability
Drug 0→1
Drug-perturbed model

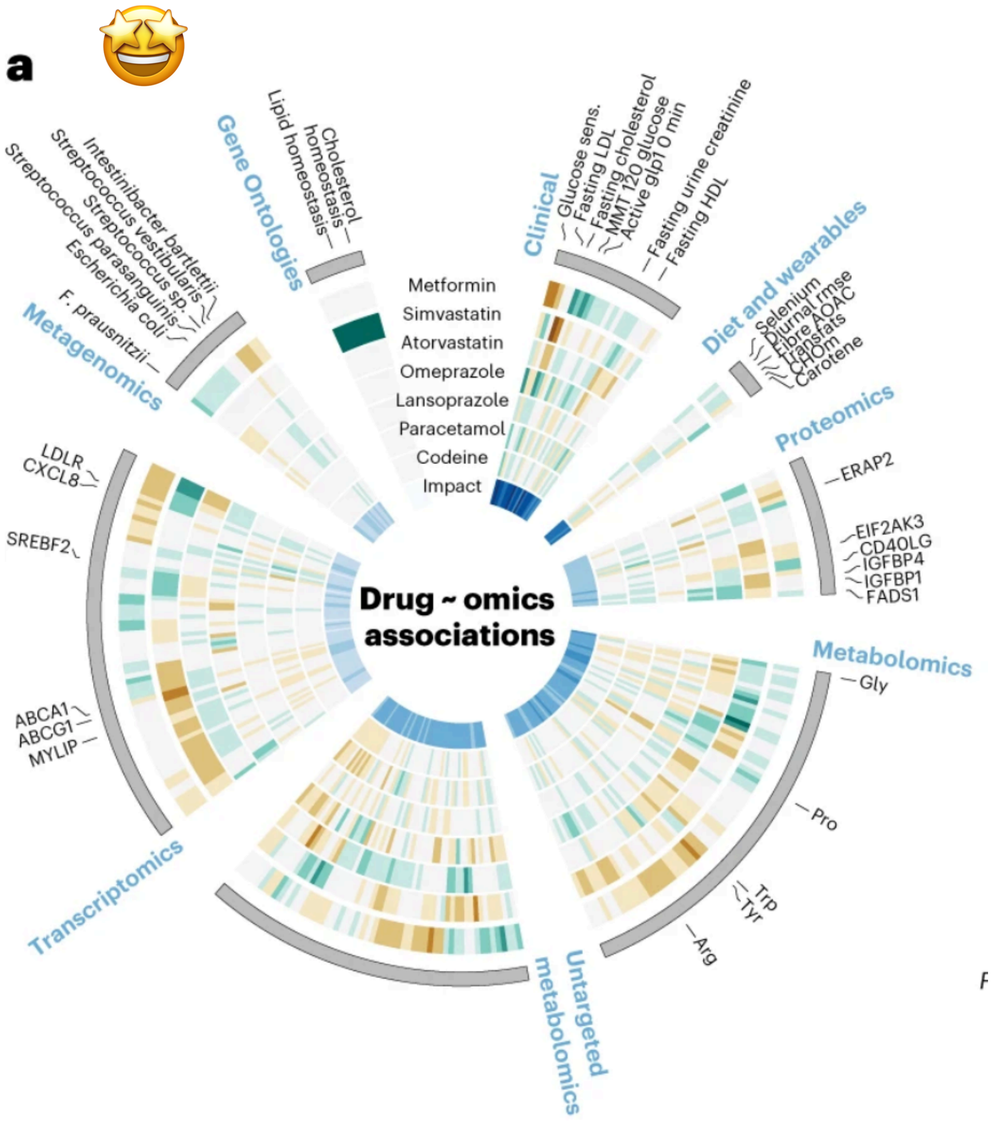
More associations…

…across modalities



Metformin on top! (that's good)



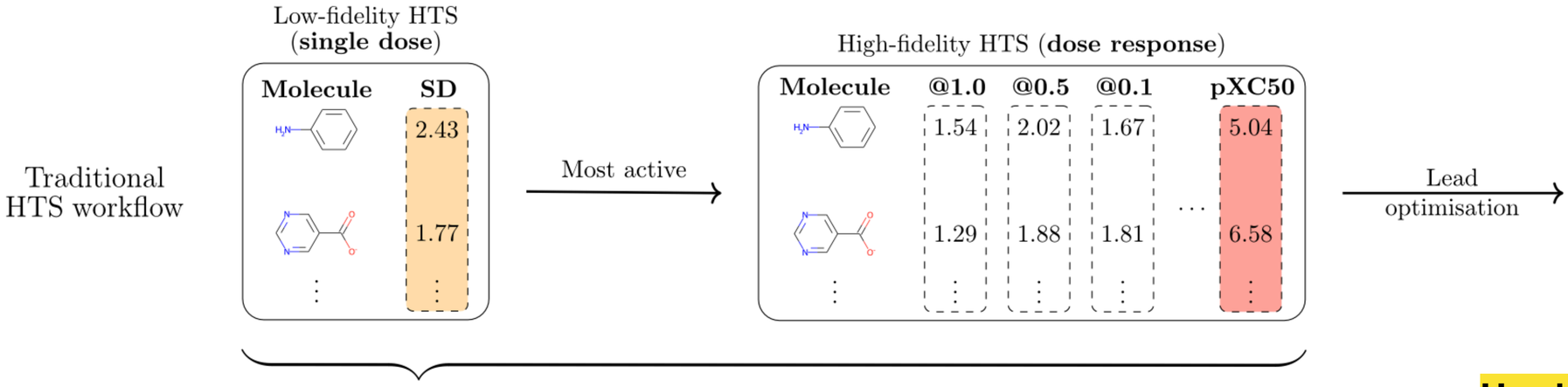


Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting (Buterez et al, *Nature Communications*)

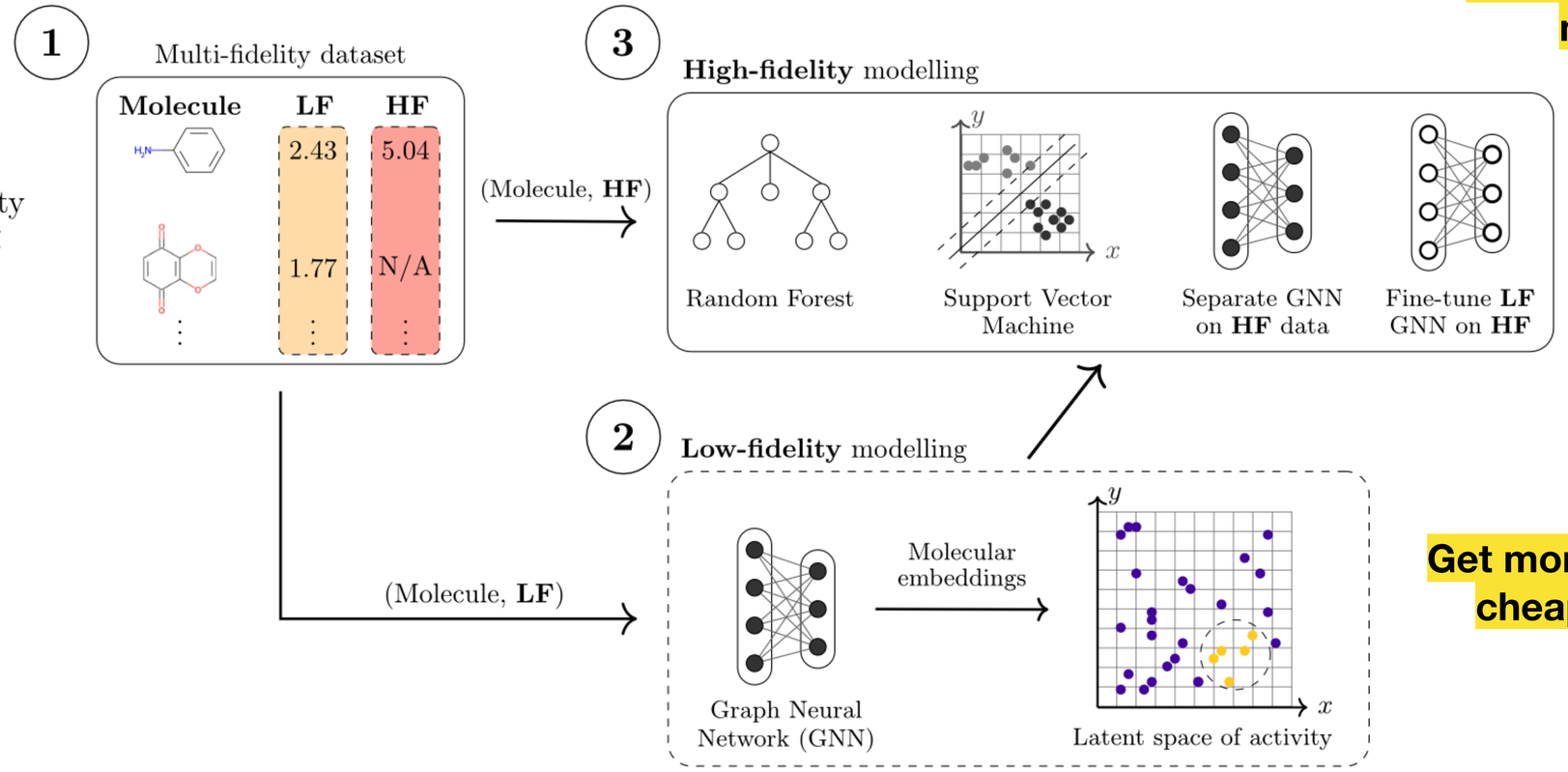
- Goal: A lot of important molecular properties are difficult and/or expensive to assay; use transfer learning from plentiful cheap data to improve their prediction
- Method:
  - Build a generic workflow to go from low-fidelity data to high-fidelity predictions
  - Uses graph neural networks (obvi)
  - Evaluate on several chemical biology and quantum chemistry tasks
- Result:
  - Significantly improved accuracy (measured by mean absolute error (MAE))
- Conclusion: Organizing the data by their ease to collect is a clever way to leverage vast amounts of chemical biology data in PubChem in ChEMBL

#IS24 #YIR24 @proftatonetti

Trad



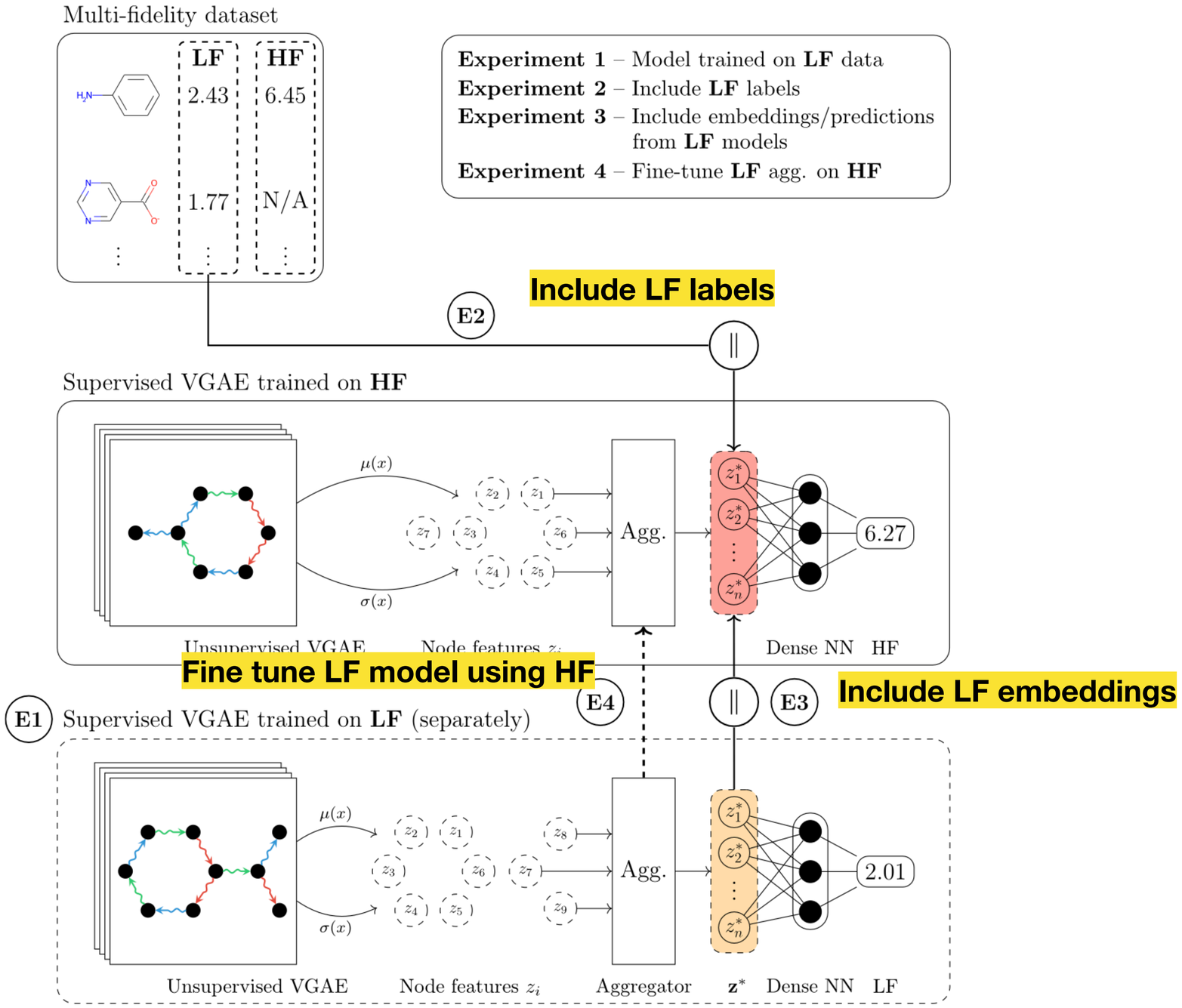
Use in downstream modeling



Transfer Learning Approach

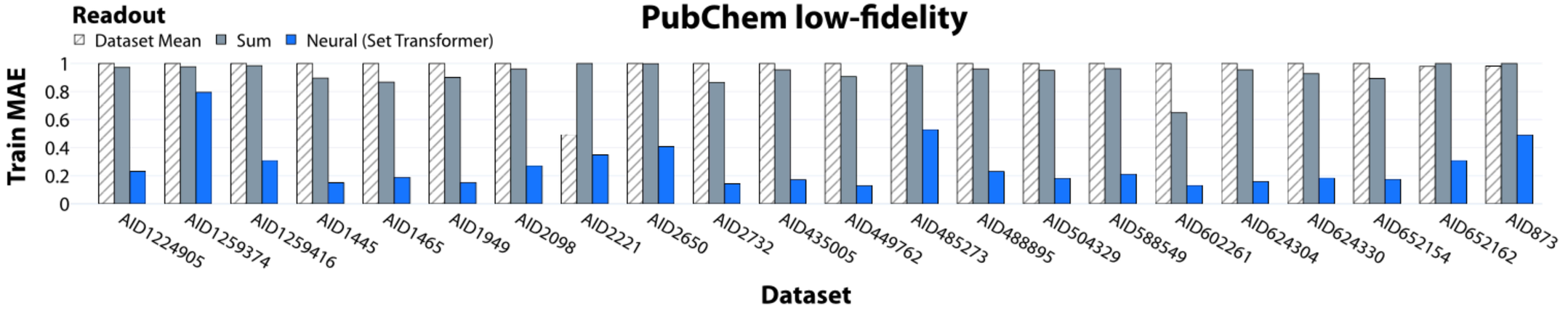
Get more out of your cheap (LF) data

Evaluated each step

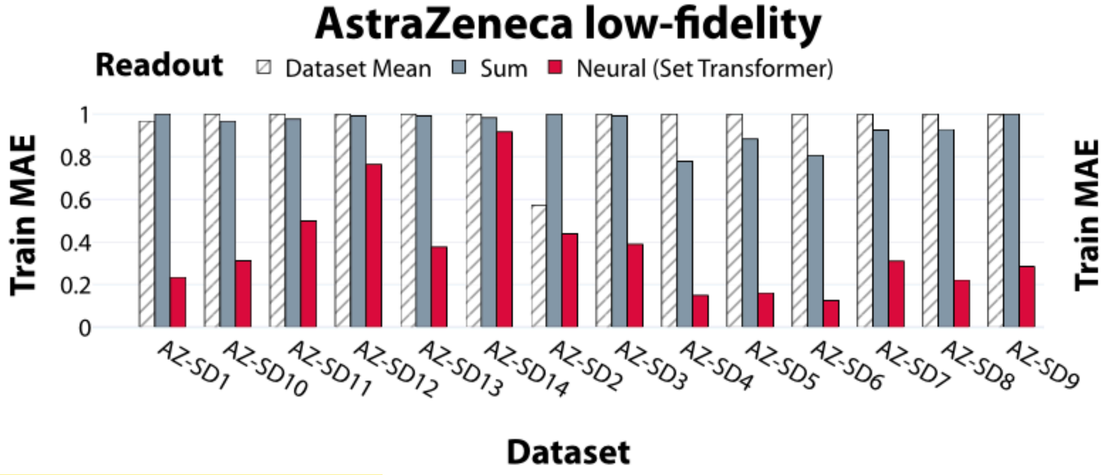


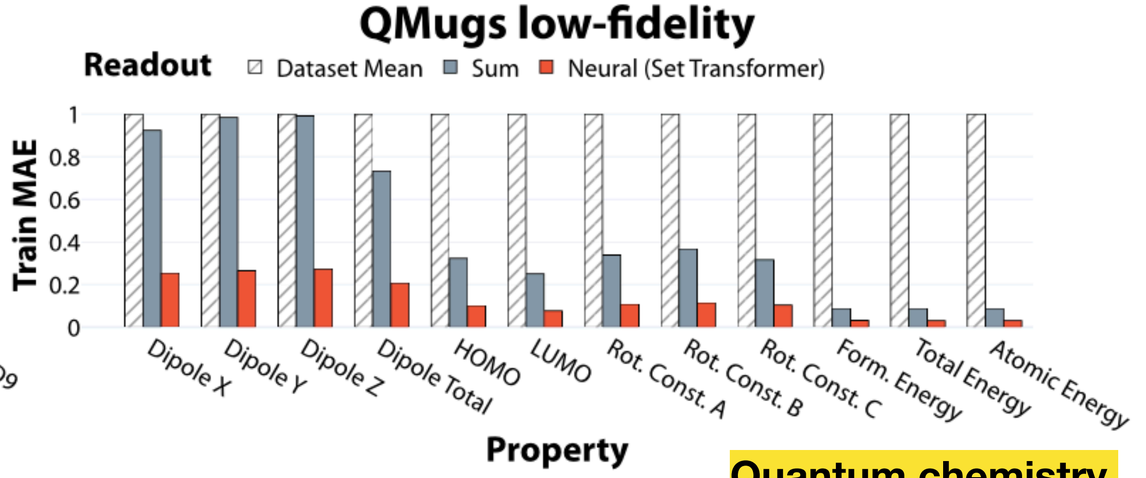
Aggregate of HTS

A



B





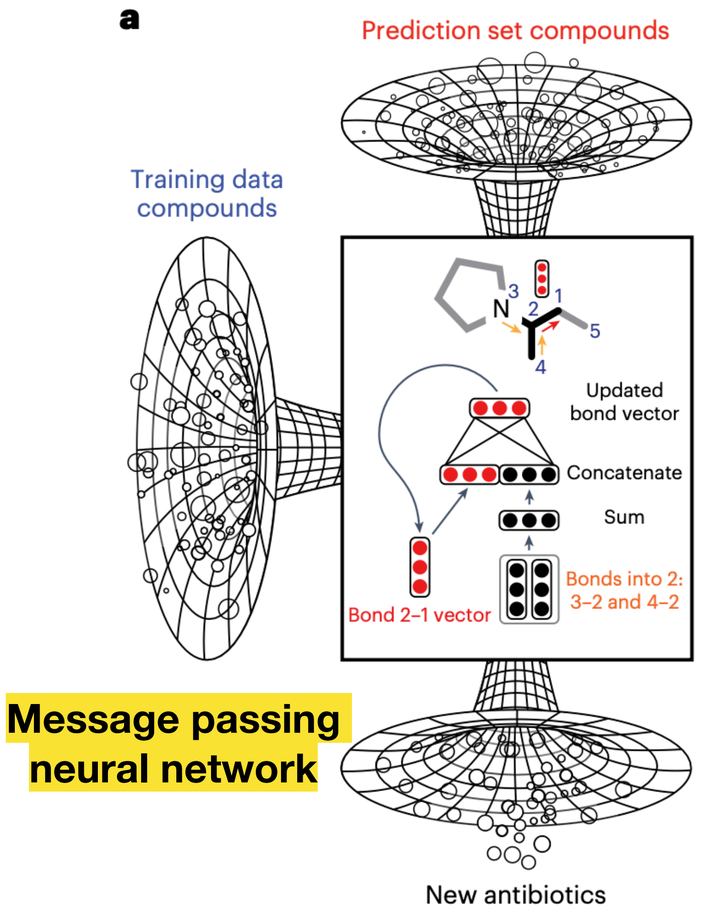
Biochemical screens

Quantum chemistry properties

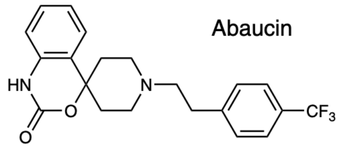
 https://doi.org/10.1038/s41589-023-01349-8

Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii* (Liu et al, *Nature Chemical Biology*)

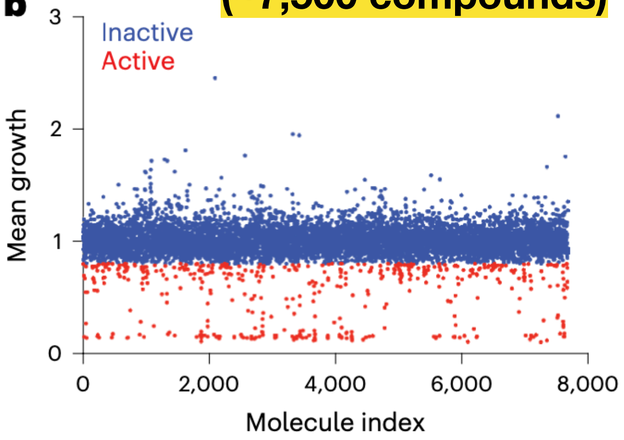
- Goal: Identify novel chemicals with activity against *A. baumannii*
- Method: Use graphical neural networks and data from a screen of 7,500 compounds
- Result:
  - Discovered a new antibiotic, abaucin
  - Identified mechanisms of action (inhibits lipoprotein tracking)
  - Validated that abaucin controls infection in a mouse model
- Conclusion: Relatively small data shows some very strong results. Exciting that not everything has to be trained on 400 gazillion examples.

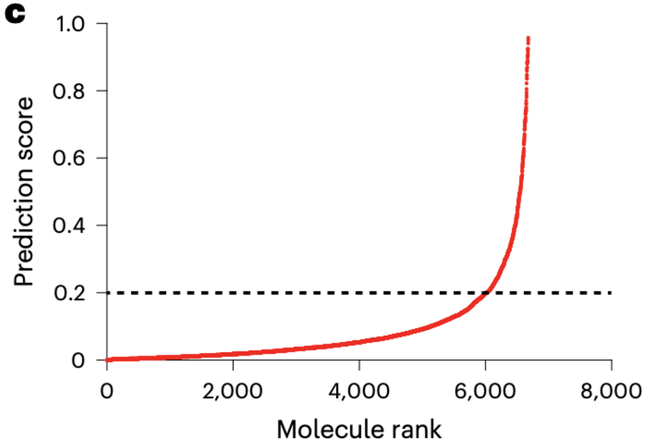


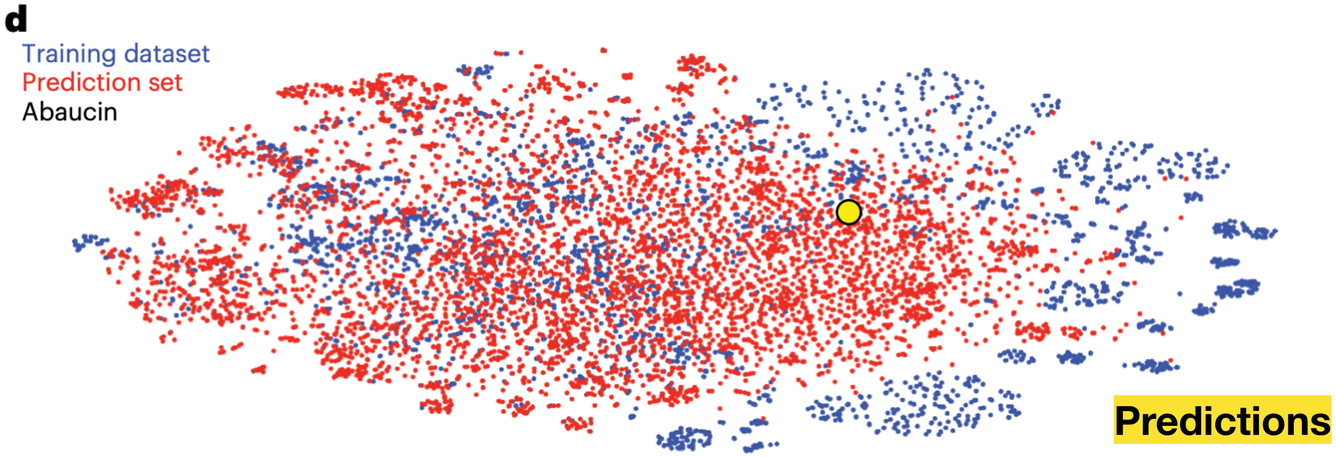
Message passing neural network



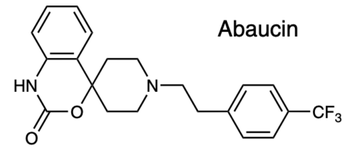
Training data (~7,500 compounds)



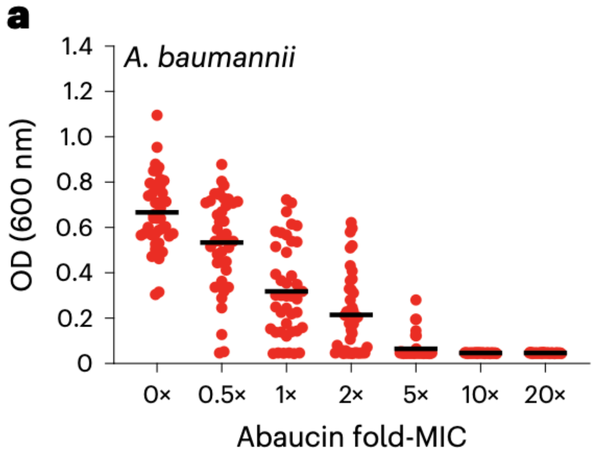


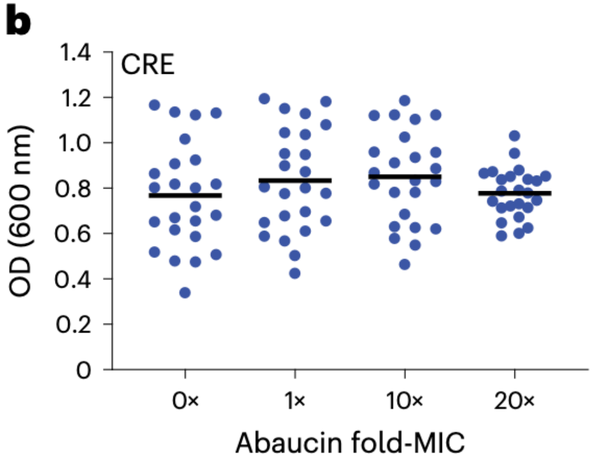


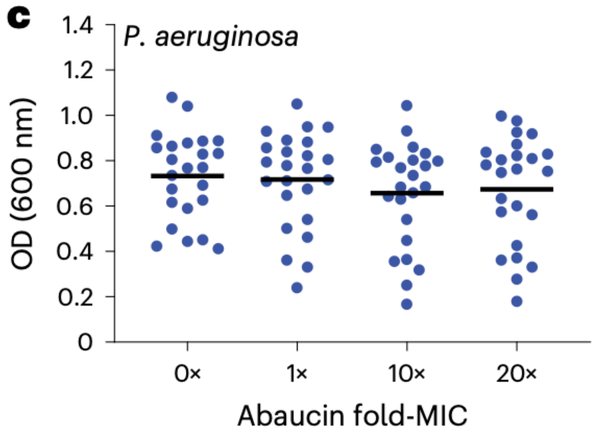
Predictions

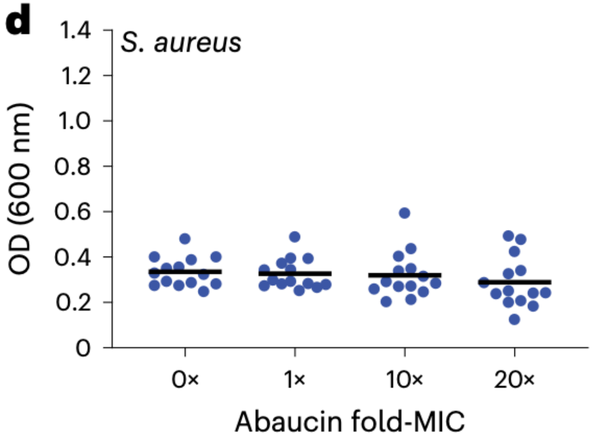


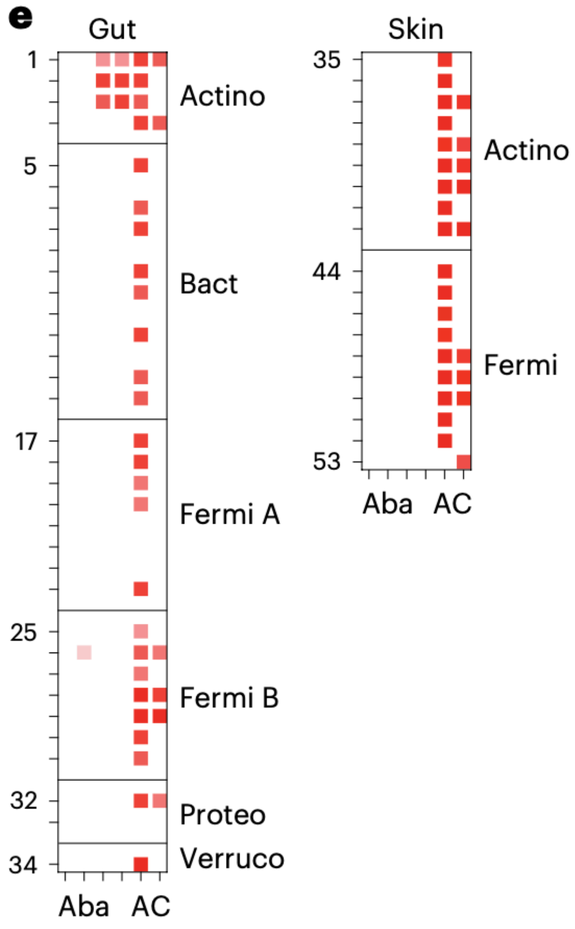
Selective inhibitor of *A. baumannii*



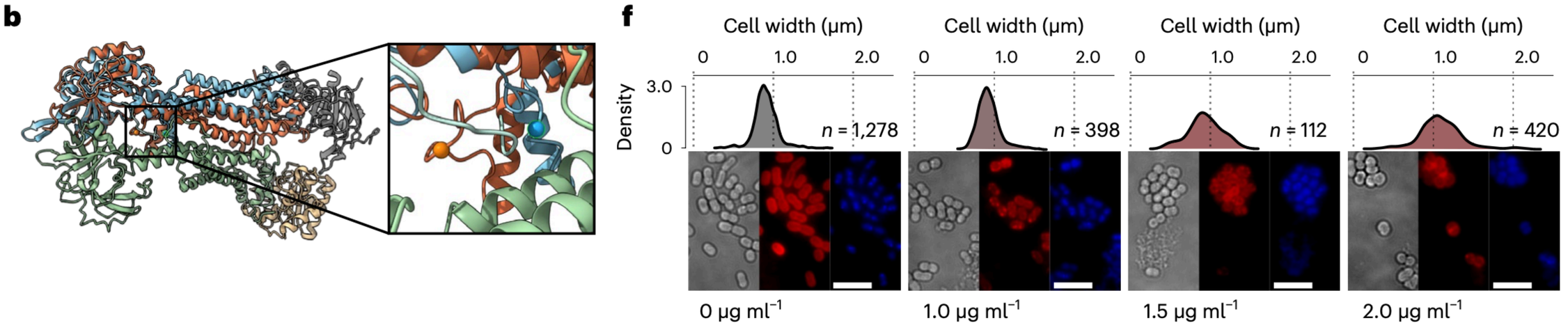






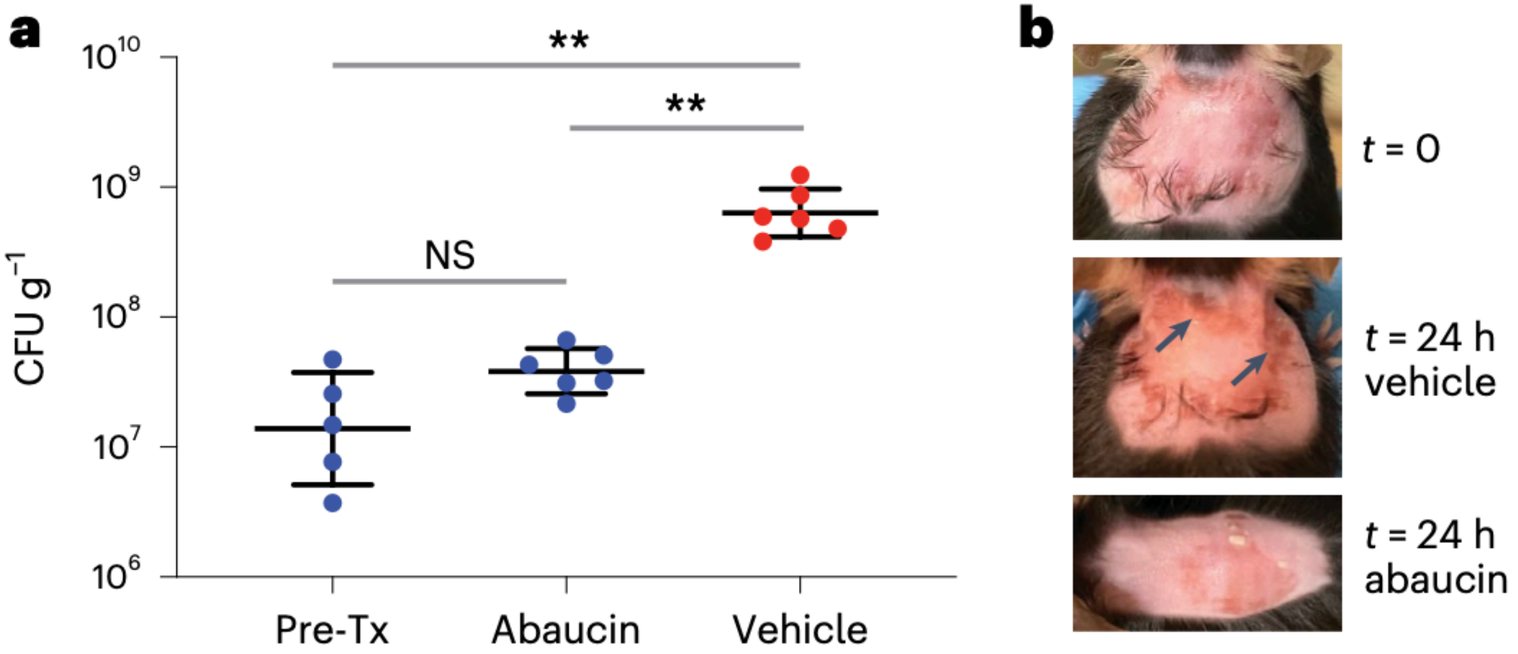


Inhibits lipoprotein trafficking



See it in action!
<— less drug, more drug —>

Cured this mouse's infection



Title (Author et al, *Journal*)

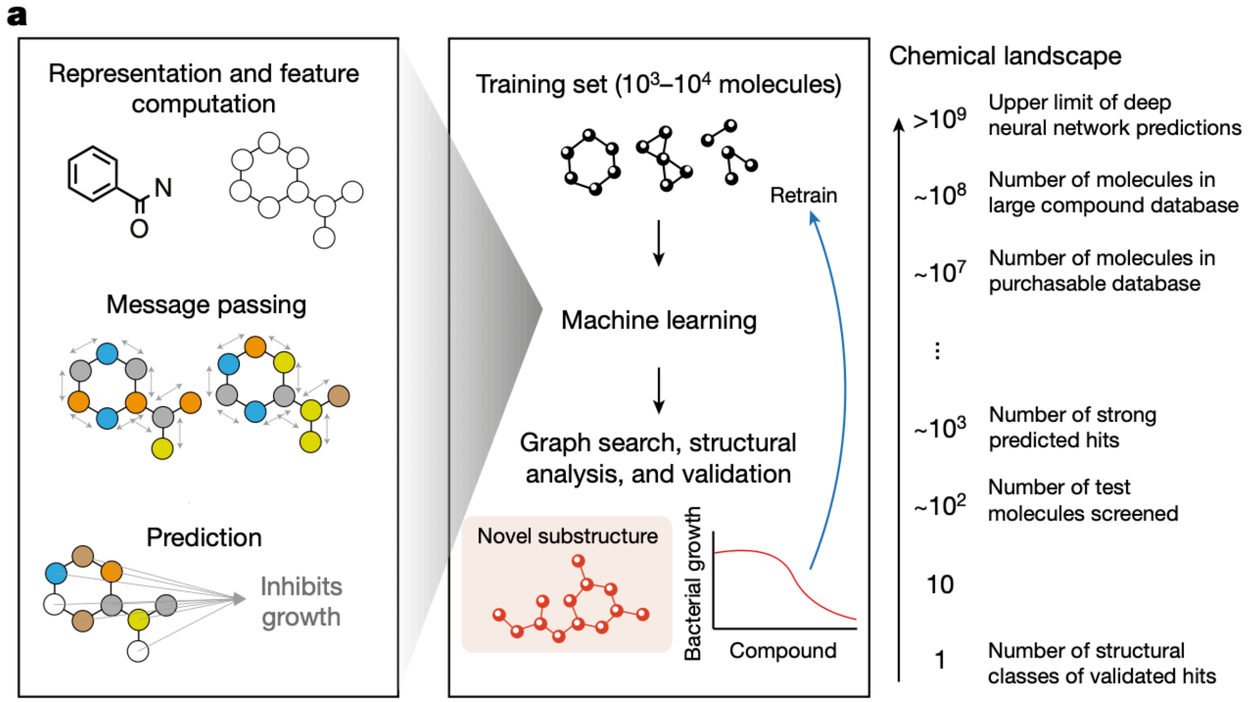
- Goal:
- Method:
- Result:
- Conclusion:

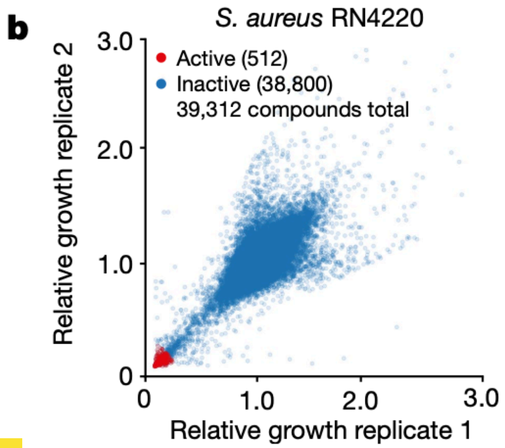
 https://doi.org/10.1038/s41586-023-06887-8

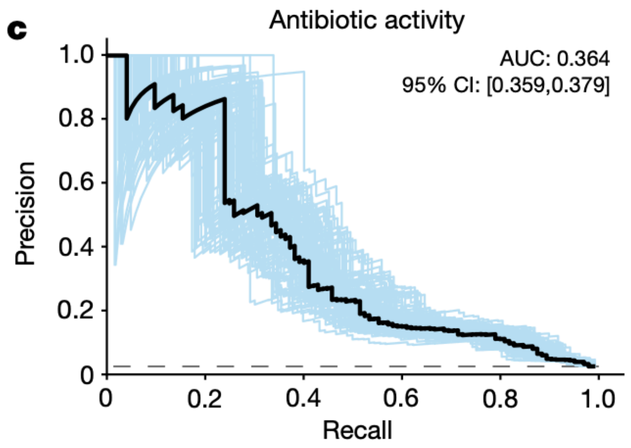
Discovery of a structural class of antibiotics with explainable deep learning (Wong et al, *Nature*)

- Goal: Identify a novel class of antibiotics
- Method:
  - Graph/message-passage neural network that is substructure-based and interpretable
  - 40k assay results applied to 12 million predictions
  - Empirically tested 283 compounds
- Result:
  - Found a structurally novel class of antibiotics with activity against MRSA
- Conclusion: Solution to our antibiotic resistance problem???

GNN



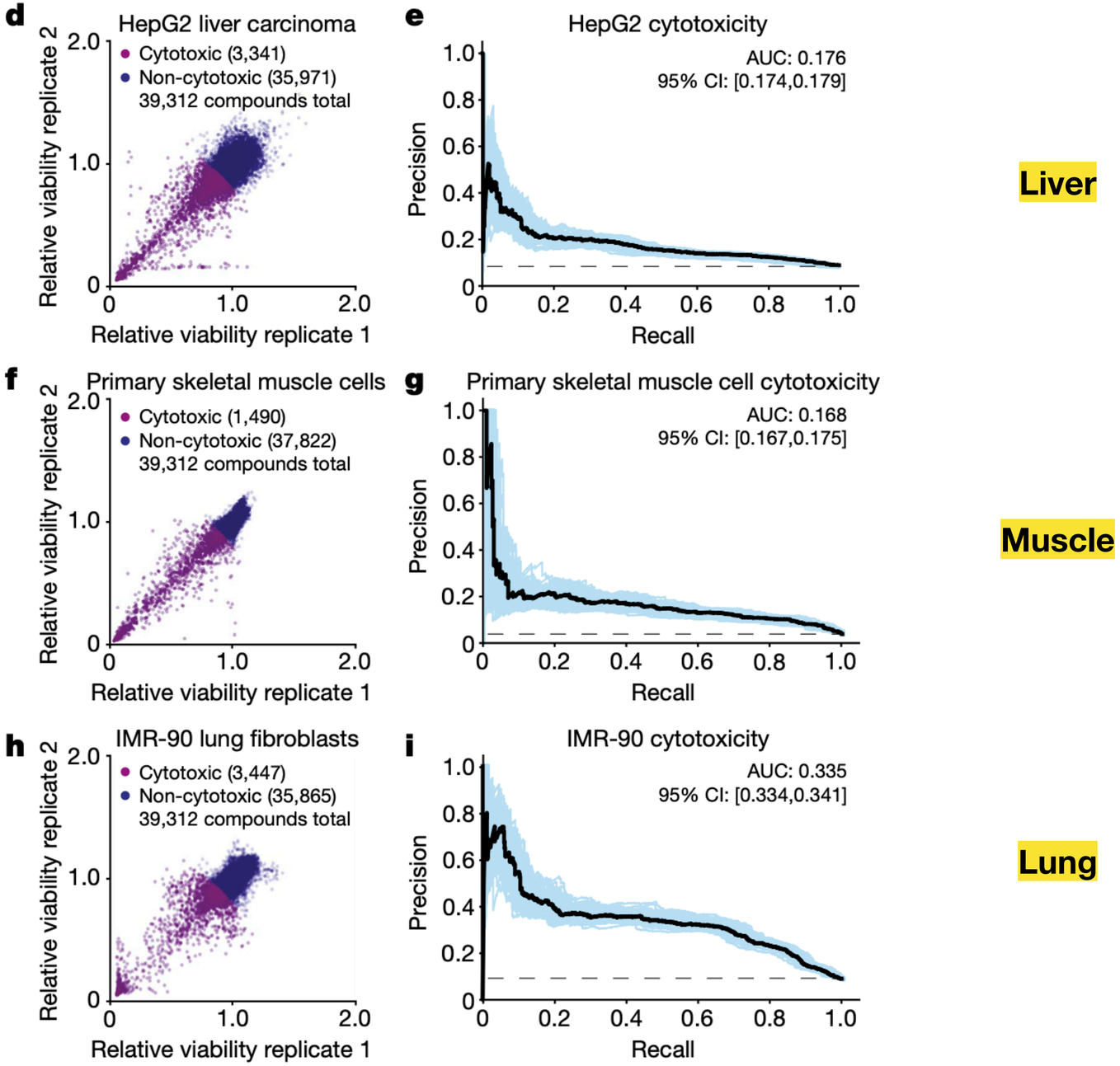




Performances of model ensemble

Training data

Also built models for relevant human cell toxicities



Restrict to compounds with predicted low human cytotoxicity

a

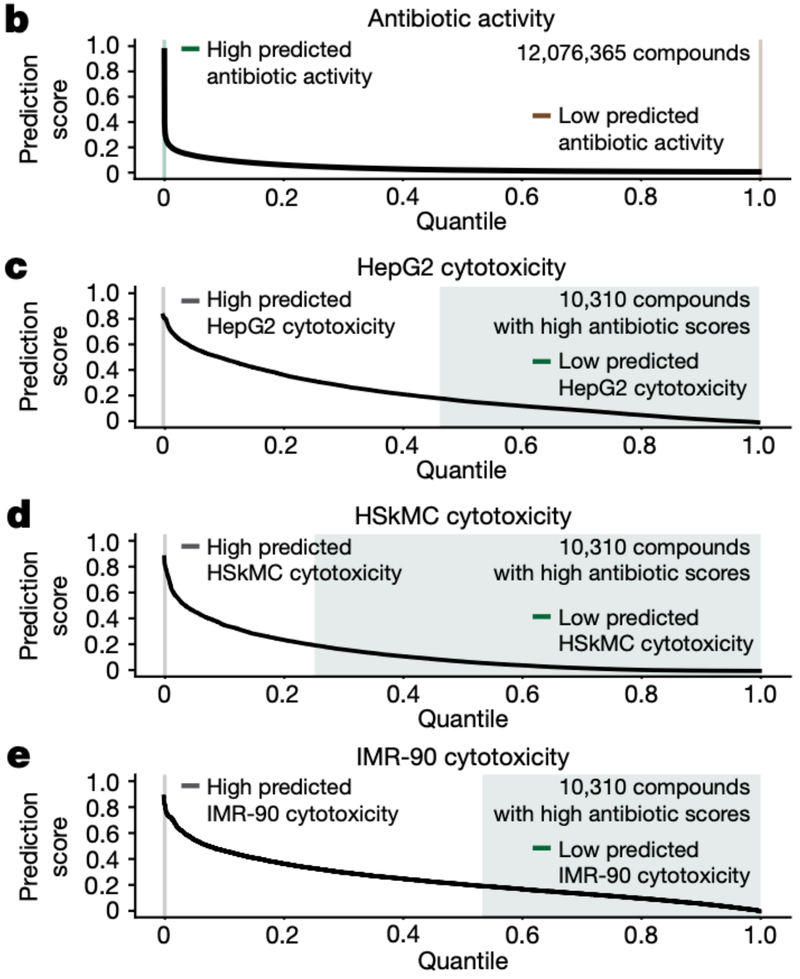
12,076,365 compounds
Mcule purchasable database, 11,277,225 compounds
Broad Institute database, 799,140 compounds

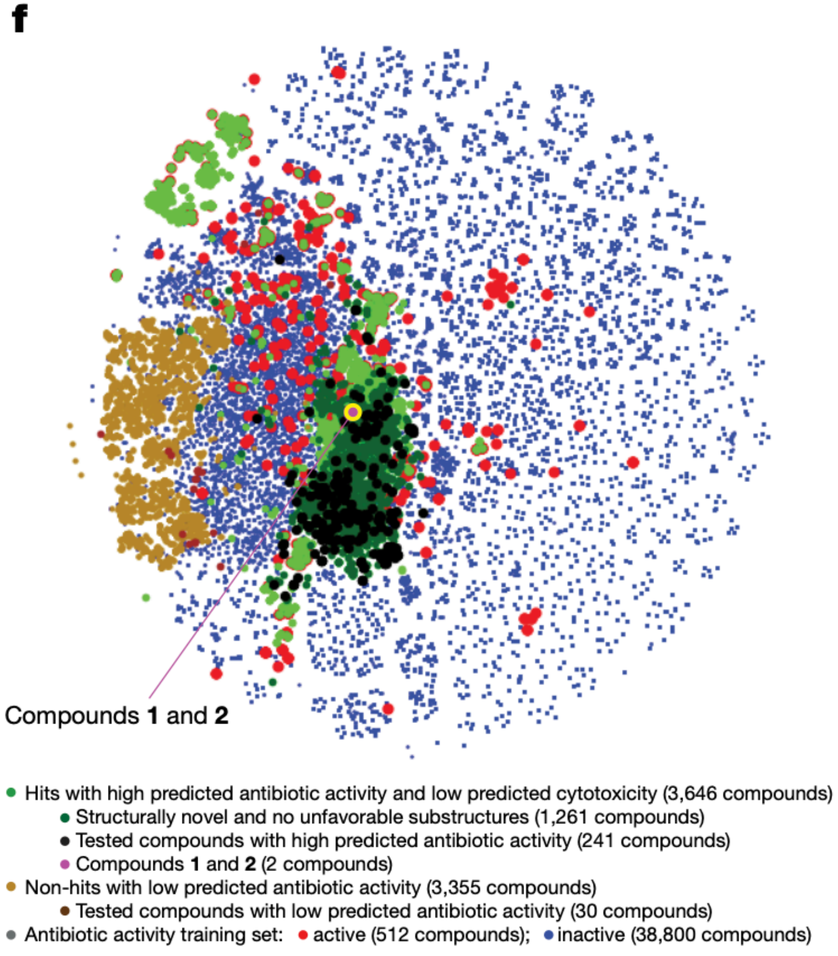
High antibiotic prediction score: 10,310 compounds
Mcule purchasable database, 3,004 compounds
Broad Institute database, 7,306 compounds

Low cytotoxicity prediction score: 3,646 compounds
Mcule purchasable database, 1,210 compounds
Broad Institute database, 2,436 compounds

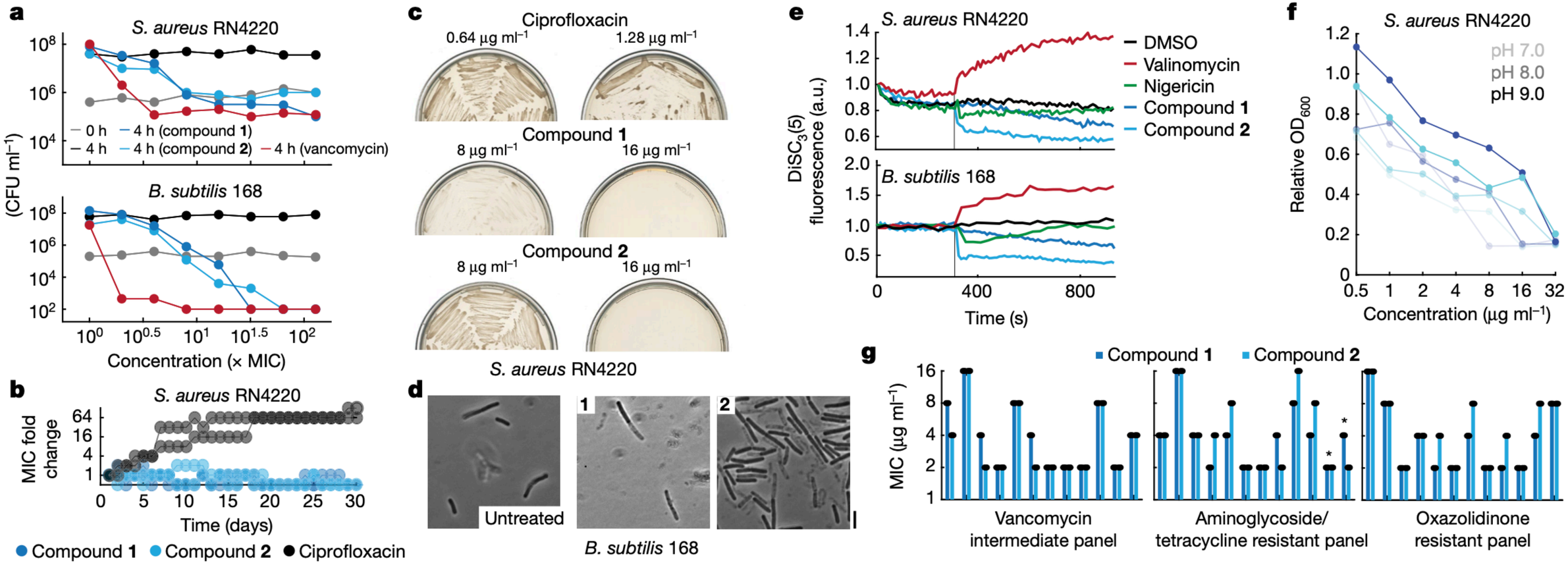
No unfavourable (PAINS/Brenk) substructures: 2,209 compounds
Mcule purchasable database, 721 compounds
Broad Institute database, 1,488 compounds

Structural novelty: 1,261 compounds
Mcule purchasable database, 162 compounds
Broad Institute database, 1,099 compounds



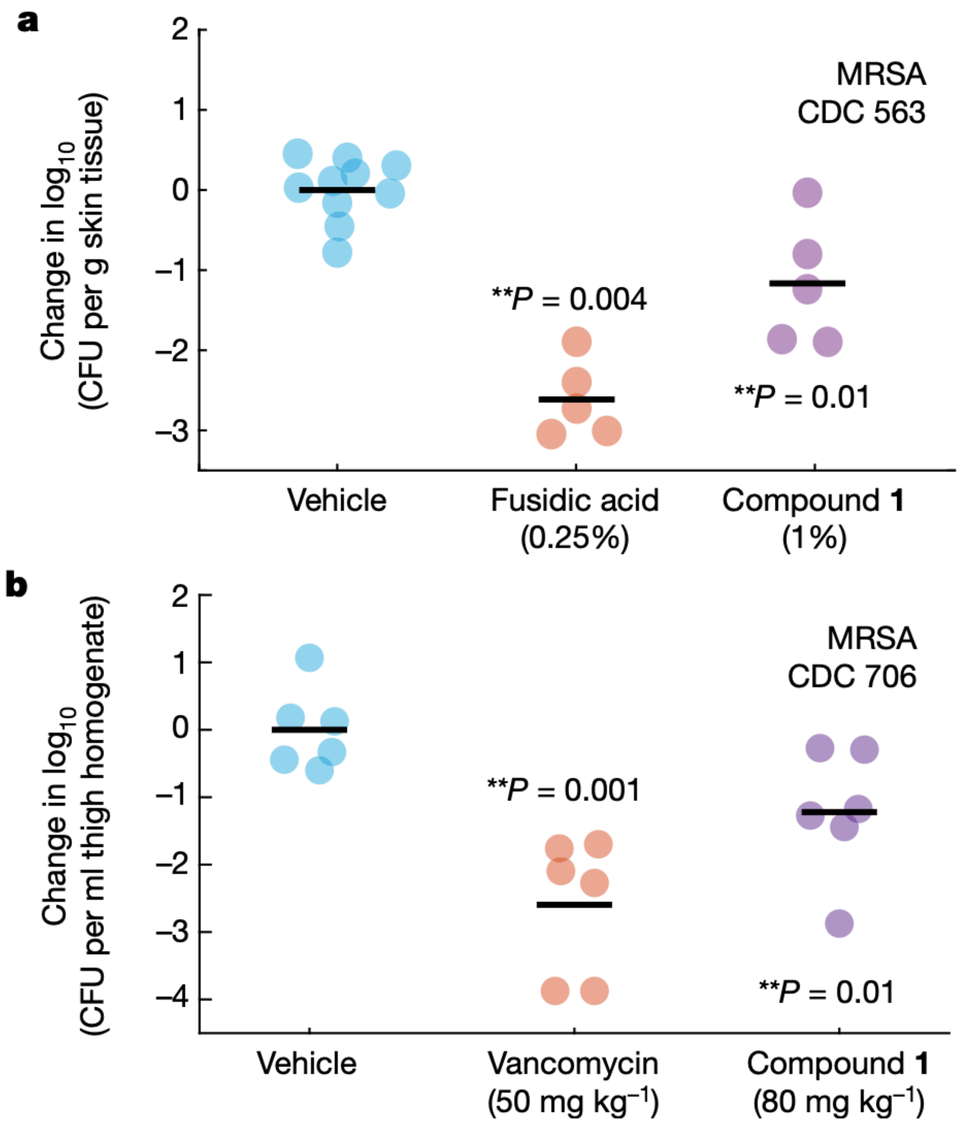


Validation experiments



Working better than Cipro

Effective against two different MRSA strains

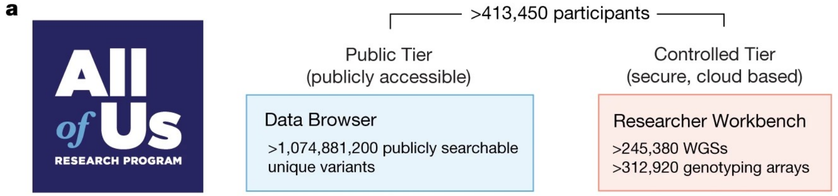


Shout Outs

Genomic data in the All of Us Research Program (The All of Us Research Program Genomics Investigators et al, *Nature*)

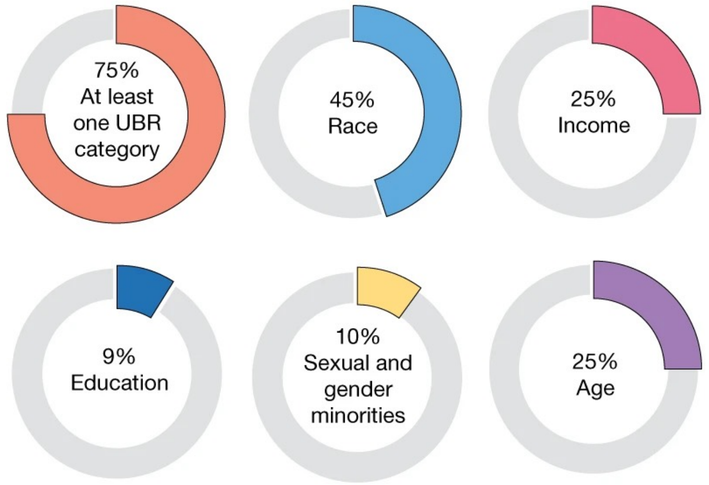
- Goal: Announce the release and initial analysis of the All of Us genomics data
- Method: Clinical-grade genome sequencing of 245k participants
- Result:
  - 77% of participants are from underrepresented communities
  - 46% are from under-represented racial and ethnic minorities
  - 1 billion genetic variants identified, 275mm previously unreported, 3.9mm with coding consequences
  - Found high replication rates across ancestry groups for 3,724 variants associated with 117 diseases
  - Summary level data are publicly available
- Conclusion: Great resource for the world! Maybe we can finally move on from UMAP?

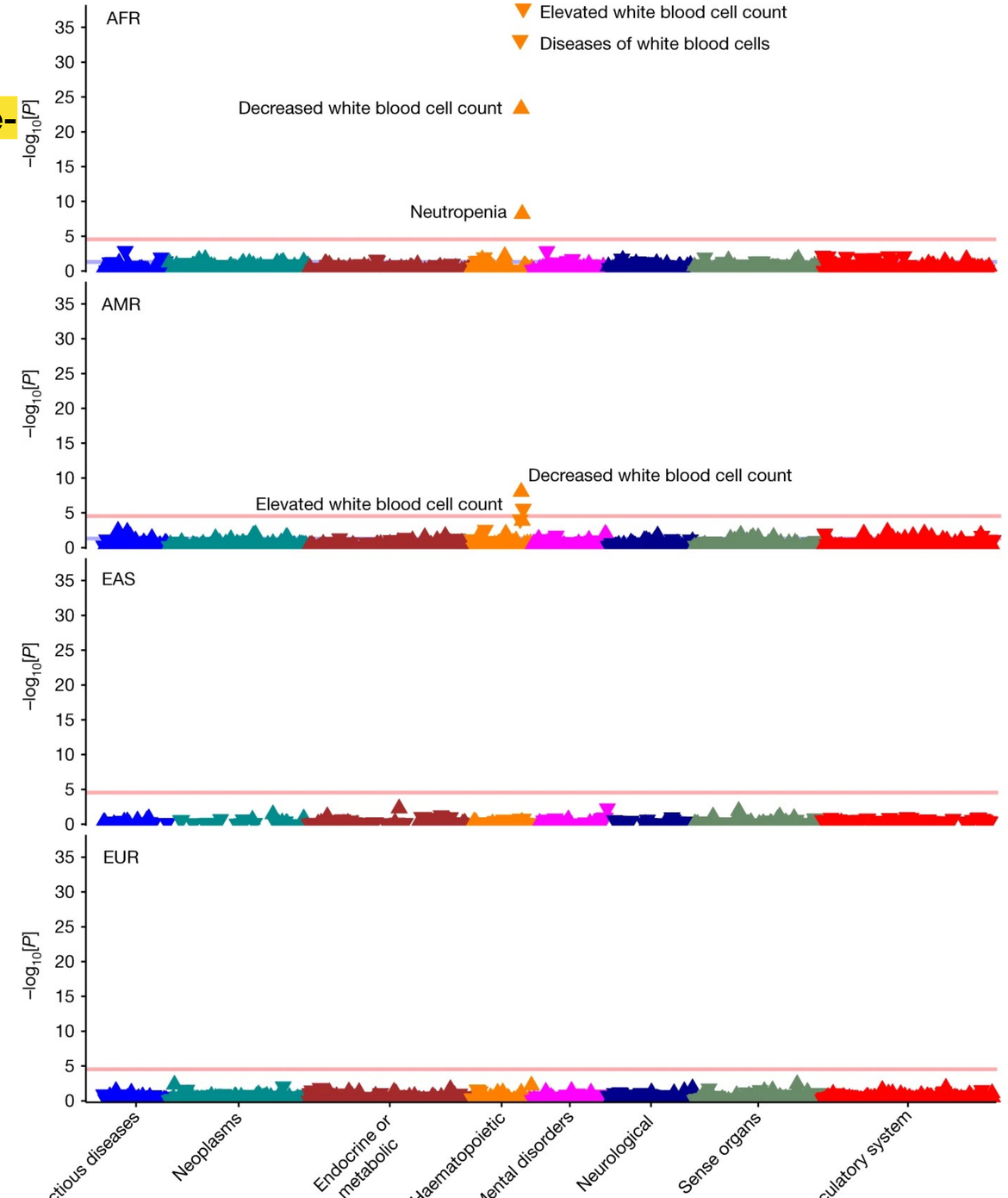
Ancestry-stratified phenome-wide association analysis



b

Participants	>245,380 WGSs	>413,370 Survey responses	>337,540 Physical measurements	>287,000 EHRs
>206,100	●	●	●	●
>39,000	●	●	●	





Can we chat about UMAP?

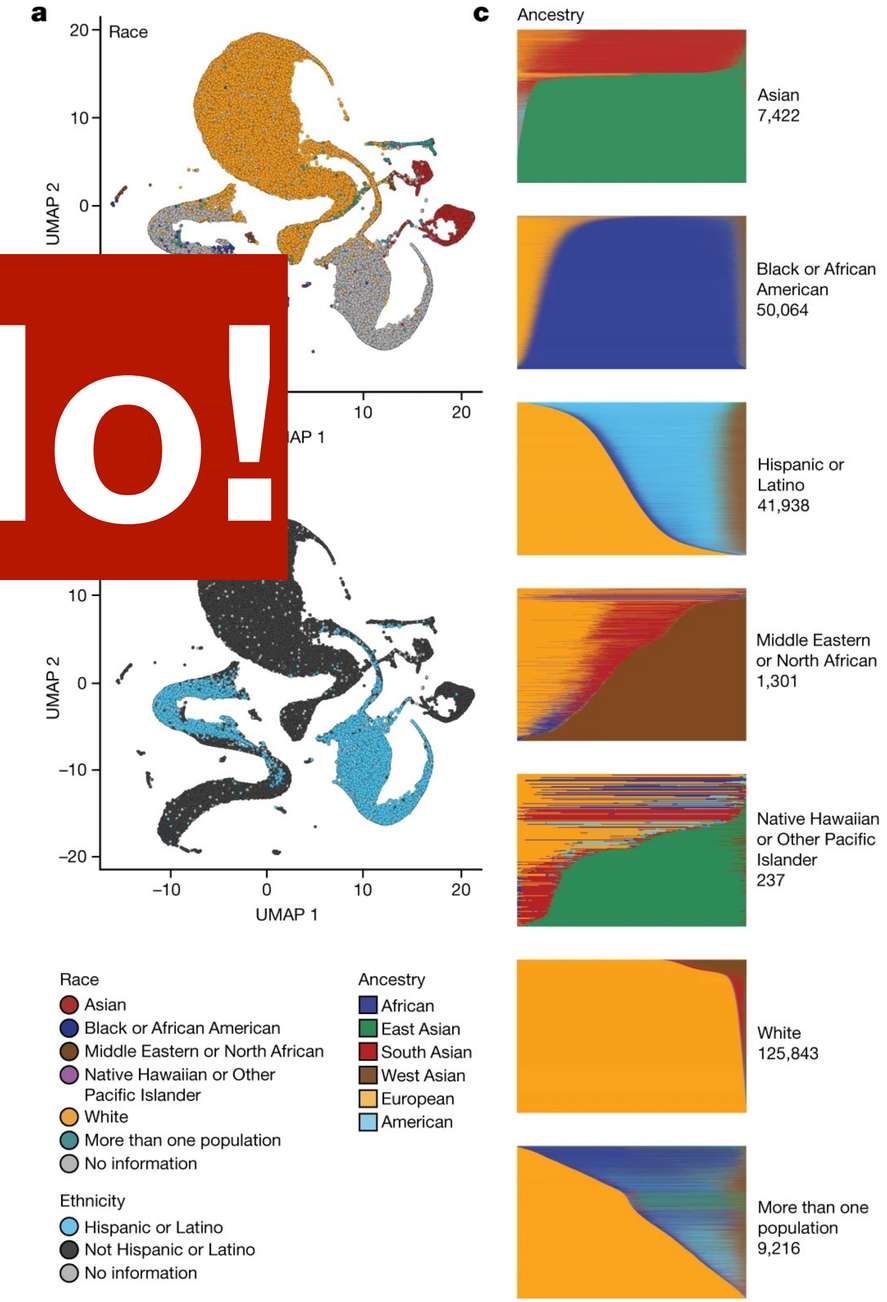
Cons

- Distorts distances
- Sensitive to parameters
- Loss of global structure
- Non-deterministic
- Meaningless axes values
- Scalability
- Overemphasizes clusters/ biases

Pros

- It's pretty

No!



You know what else distorts the truth, reinforces biases, shows you what you want to see?

FOX NEWS channel == UMAP

 https://doi.org/10.2196/45225

Potential Target Discovery and Drug Repurposing for Coronaviruses: Study Involving a Knowledge Graph-Based Approach (Lou et al, *JMIR*)

- Goal: Identify drug repurposing candidates for the next coronavirus
- Method:
  - Integrate domain specific knowledge from the literature with molecular knowledge bases of molecular drug action
  - Use a heuristic filtering method to make massive KG traversals possible
- Result:
  - Identified 33 targets and 18 drug repurposing candidates
  - 41 new drug mechanisms of action detected
- Conclusion: Massive knowledge bases will be useful for pre-training

Figure 4

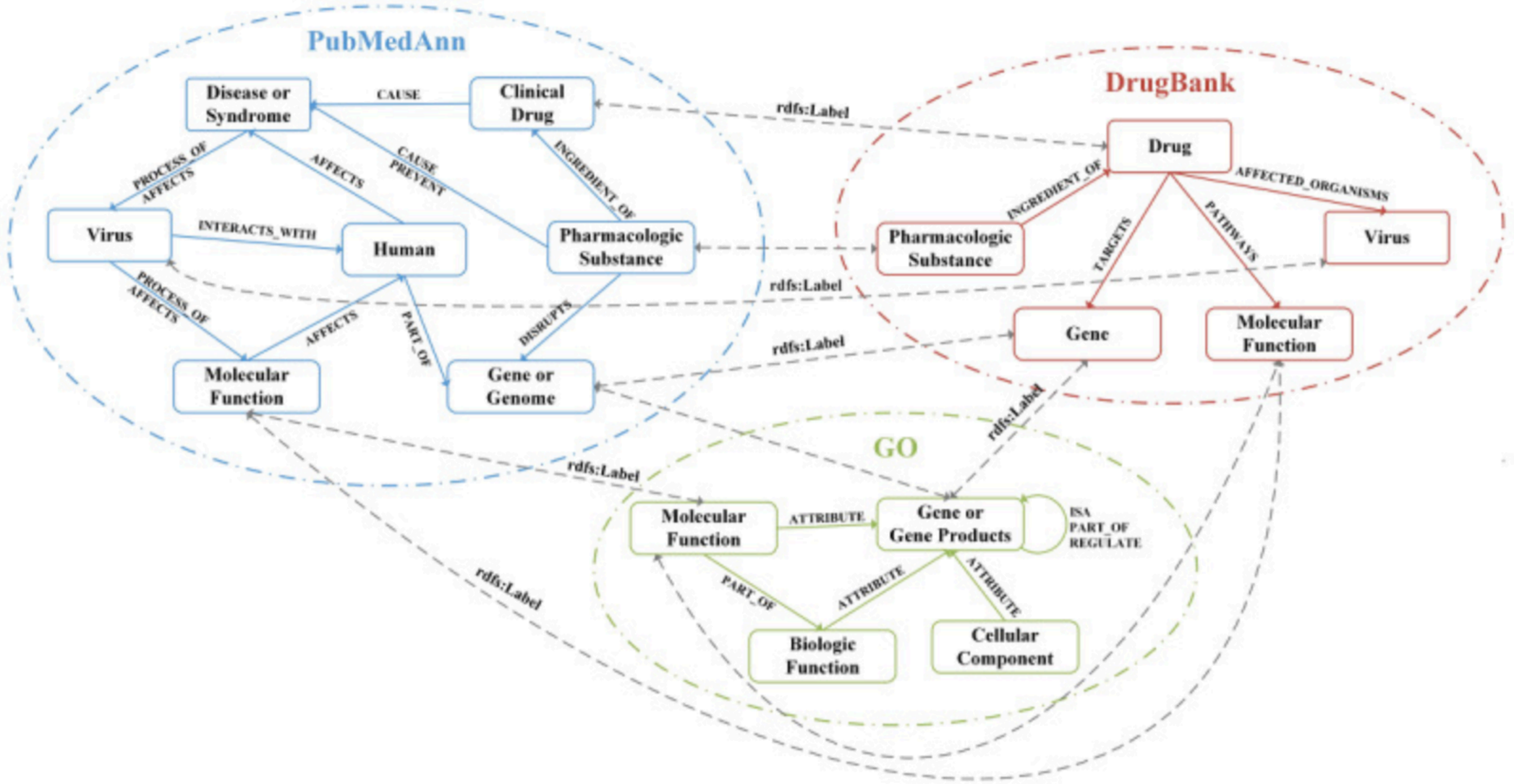
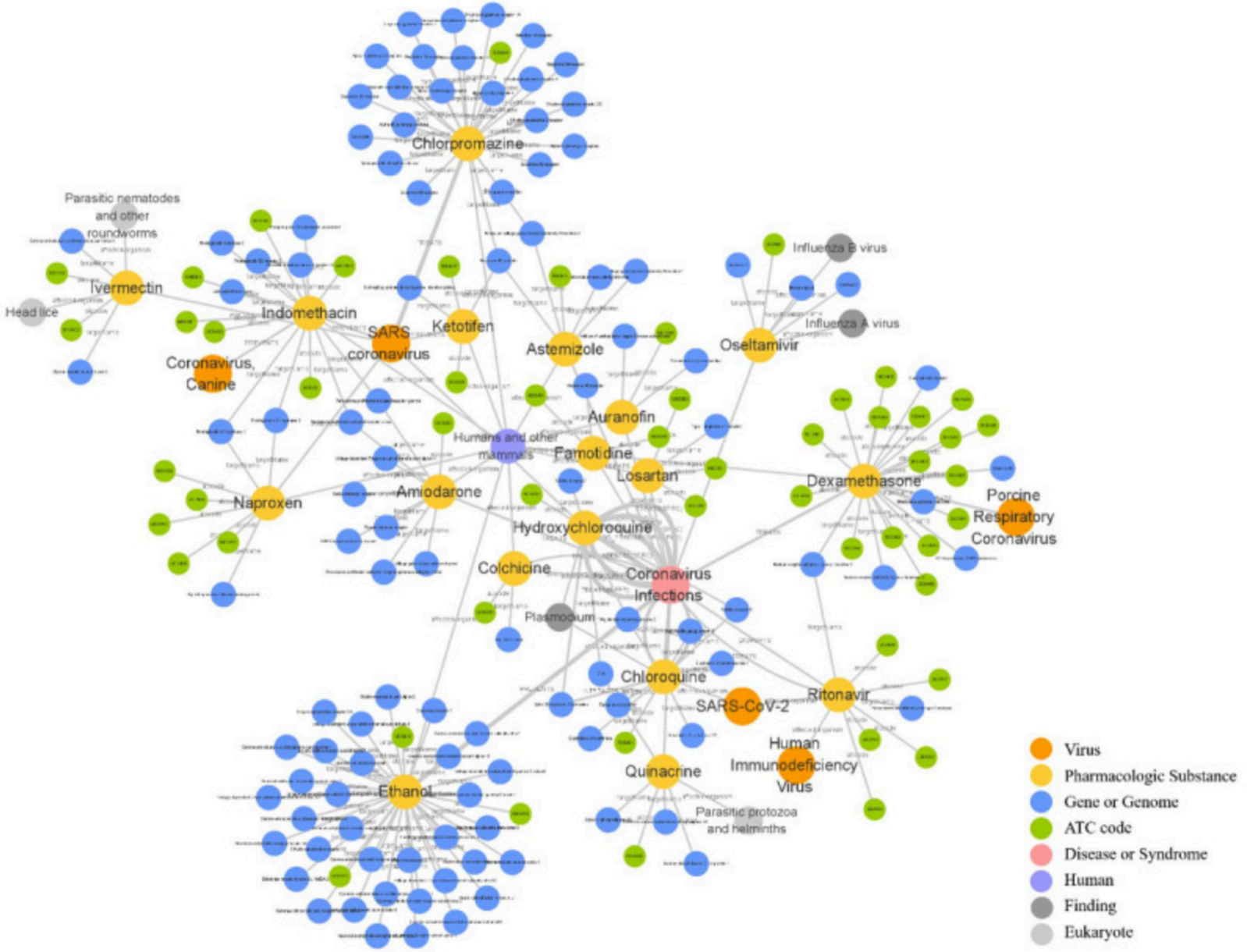

PubMedAnn
Disease or Syndrome
CAUSE
Clinical Drug
PROCESS_OF
AFFECTS
AFFECTS
CAUSE
PREVENT
INGREDIENT_OF
Virus
INTERACTS_WITH
Human
Pharmacologic Substance
PROCESS_OF
AFFECTS
AFFECTS
PART_OF
DISRUPTS
Molecular Function
Gene or Genome
rdfs:Label
rdfs:Label
rdfs:Label
rdfs:Label
rdfs:Label
DrugBank
Drug
INGREDIENT_OF
AFFECTED_ORGANISMS
Pharmacologic Substance
TARGETS
PATHWAYS
Virus
Gene
Molecular Function
GO
Molecular Function
ATTRIBUTE
Gene or Gene Products
ISA
PART_OF
REGULATE
PART_OF
ATTRIBUTE
ATTRIBUTE
Biologic Function
Cellular Component

Figure 8


Chlorpromazine
Parasitic nematodes and other roundworms
Ivermectin
Head lice
Indomethacin
SARS coronavirus
Ketotifen
Coronavirus, Canine
Asemizole
Influenza B virus
Influenza A virus
Oseltamivir
Auranofin
Humans and other mammals
Famotidine
Losartan
Amiodarone
Naproxen
Dexamethasone
Porcine Respiratory Coronavirus
Hydroxychloroquine
Colchicine
Coronavirus Infections
Plasmodium
Chloroquine
SARS-CoV-2
Ritonavir
Human Immunodeficiency Virus
Quinacrine
Parasitic protozoa and helminths
Ethanol
Virus
Pharmacologic Substance
Gene or Genome
ATC code
Disease or Syndrome
Human
Finding
Eukaryote

 https://doi.org/10.1093/bioinformatics/btad606

Aliro: an automated machine learning tool leveraging large language models (Choi et al, *Bioinformatics*)

- Goal: Make advanced machine learning tools available through natural language interface
- Method: Combine Large Language Models and AutoML
- Result: Aliro, open source and pretty tool
- Conclusion: Democratization of AI powered by AI

Aliro Your Friendly AI Assistant | Datasets | Expert | Experiments | Admin

Results: Iris.tsv

Experiment: #64628346786053003 1e6a8dc

Download

Algorithm

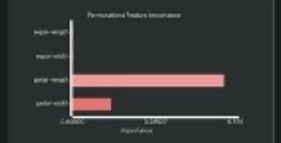
Decision Tree Classifier

Criterion: gini	Max Depth: 3
Min Samples Split: 2	Min Samples Leaf: 1
Min Weight Fraction Leaf: 0	Max Features: sqrt

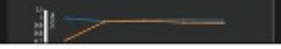
Run Details

Started: 5/15/23 12:08 pm	Finished: 5/15/23 12:08 pm
Duration: 0h 0m 4s	Launched By: testuser

Feature Importance



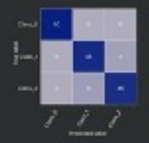
Learning Curve



Class Rate



Confusion Matrix

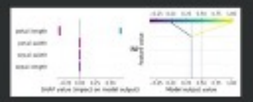


ROC curve only available for binary classification.

SHAP Analysis

Class: 0

Name



TSNE 2D



Balanced Accuracy ✓



AUC

AUC is only available for binary classification.

Precision ✓



Recall ✓



F1 Score ✓

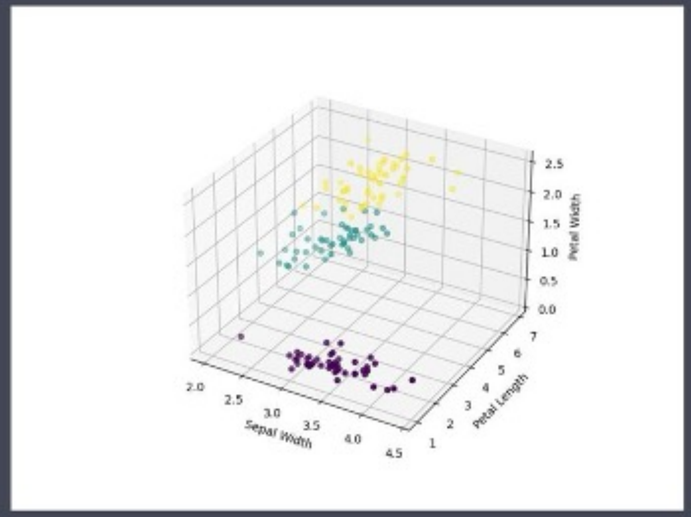


```
ax.set_ylabel('Petal Length')
ax.set_zlabel('Petal Width')
plt.savefig('scatter_plot.jpg')
#
# This code will create a scatter plot with sepal width on the
x-axis, petal length on the y-axis, and petal width on the z-
axis. The data points will be differentiated by color based on
the 'target' column in the dataframe. Finally, the plot will
be saved as a JPEG file named 'scatter_plot.jpg'.
```

Run

AI: Please wait while I am running your code on Aliro...

AI: Please review the content below. If the output contains any files, you can download them by clicking on the respective links.



Type your message here. Submit

 https://doi.org/10.1016/j.gim.2023.100006

Returning integrated genomic risk and clinical recommendations: The eMERGE study (Linder et al, *Genetics in Medicine*)

- Goal: Understand how to assess and return genetic risk to patients from diverse ancestral backgrounds
- Method:
  - Prospective cohort study of 25,000 patients across 10 sites
  - Return cross-ancestry polygenic risk scores, monogenic risks, family history, and clinical risk via a Genome-Informed Risk Assessment (GIRA)
- Result: TBD
- Conclusion: eMERGE continues to be an engine of innovation and I'm excited to see the results of this study

 https://doi.org/10.1093/jamiaopen/ooad084

Medical Informatics Operating Room Vitals and Events Repository (MOVER): a public-access operating room database (Samad et al, *JAMIA Open*)

- Goal: Publicly release data on Operating Room Vitals and Events
- Method: UC Irvine data from 2015-2022
  - ECG, Arterial Line, Pulse Ox Waveforms
  - EMR data including demogs, meds, labs, procedure, and several other modalities
- Result: 58,799 patients and 83,468 surgeries represented
- Conclusion: Amazing new AI resource!

A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study (Swen et al, *The Lancet*)

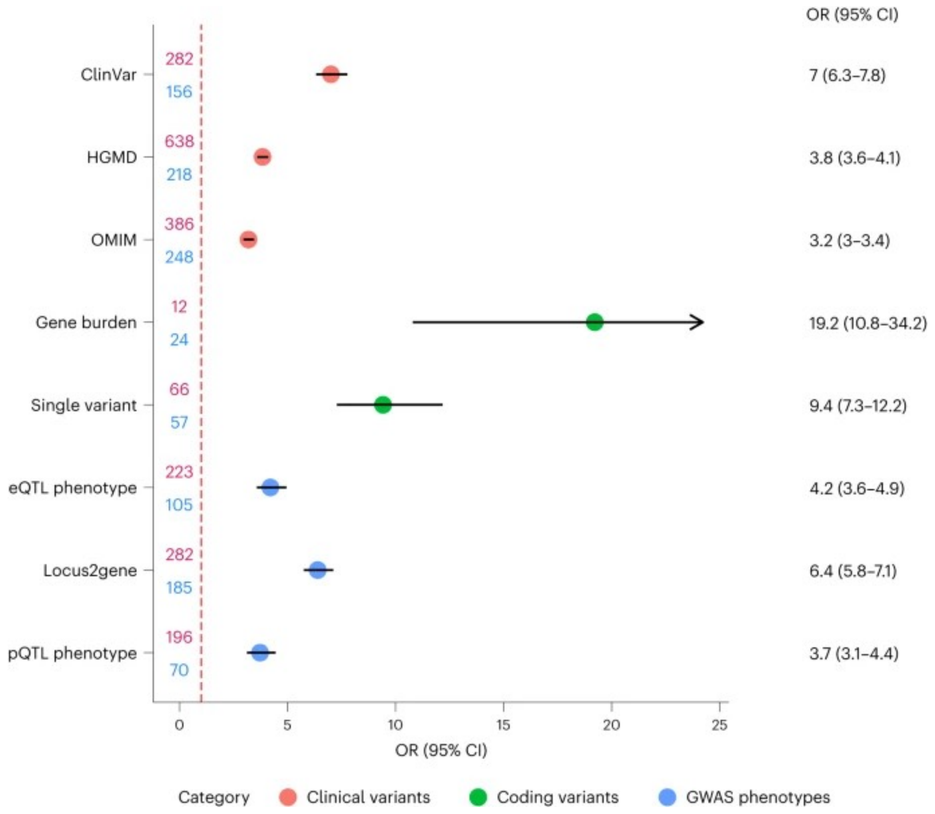
- Goal: Evaluate the benefit of pharmacogenetic testing
- Method:
  - Open-label, multisite, controlled, cluster-randomized crossover study
  - 18 hospitals, 9 health centers, 28 pharmacies
  - 12-gene PGx panel, 50 variants with actionable result
  - Primary outcome: clinically relevant adverse drug reactions
- Result:
  - ~7k patients enrolled, 1.5k with actionable result
  - Significant ~1.4X reduction in ADRs for study group
- Conclusion: Strongest evidence so far that PGx is clinically relevant

 https://doi.org/10.1038/s41588-023-01609-2

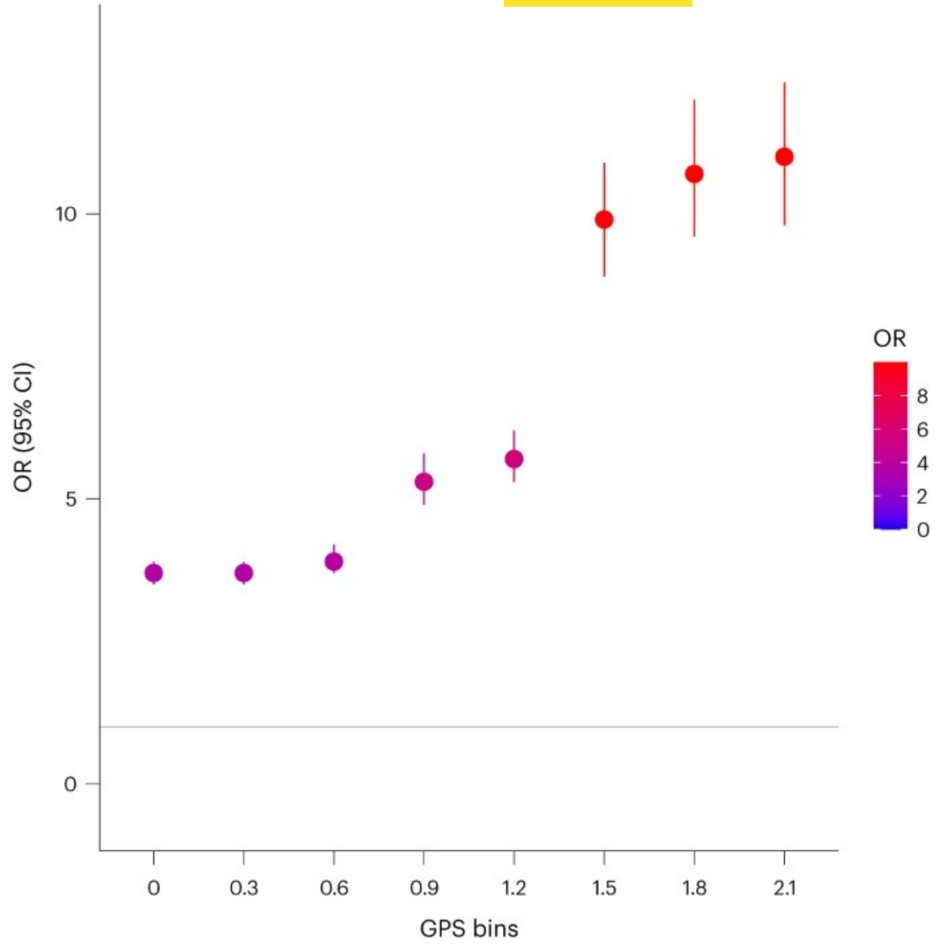
Development of a human genetics-guided priority score for 19,365 genes and 399 drug indications (Duffy et al, *Nature Genetics*)

- Goal: Build a holistic score for a drug target's "genetic evidence"
- Method:
  - Integrate eight features of genetic evidence
  - Introduce the Genetic Priority Score (GPS) using OpenTargets and SIDER
- Result:
  - Top GPS are associated with up to 11X fold enrichment for approved drug targets
- Conclusion: Genes with genetic evidence of an effect make better drug targets

Each of the eight genetic evidences



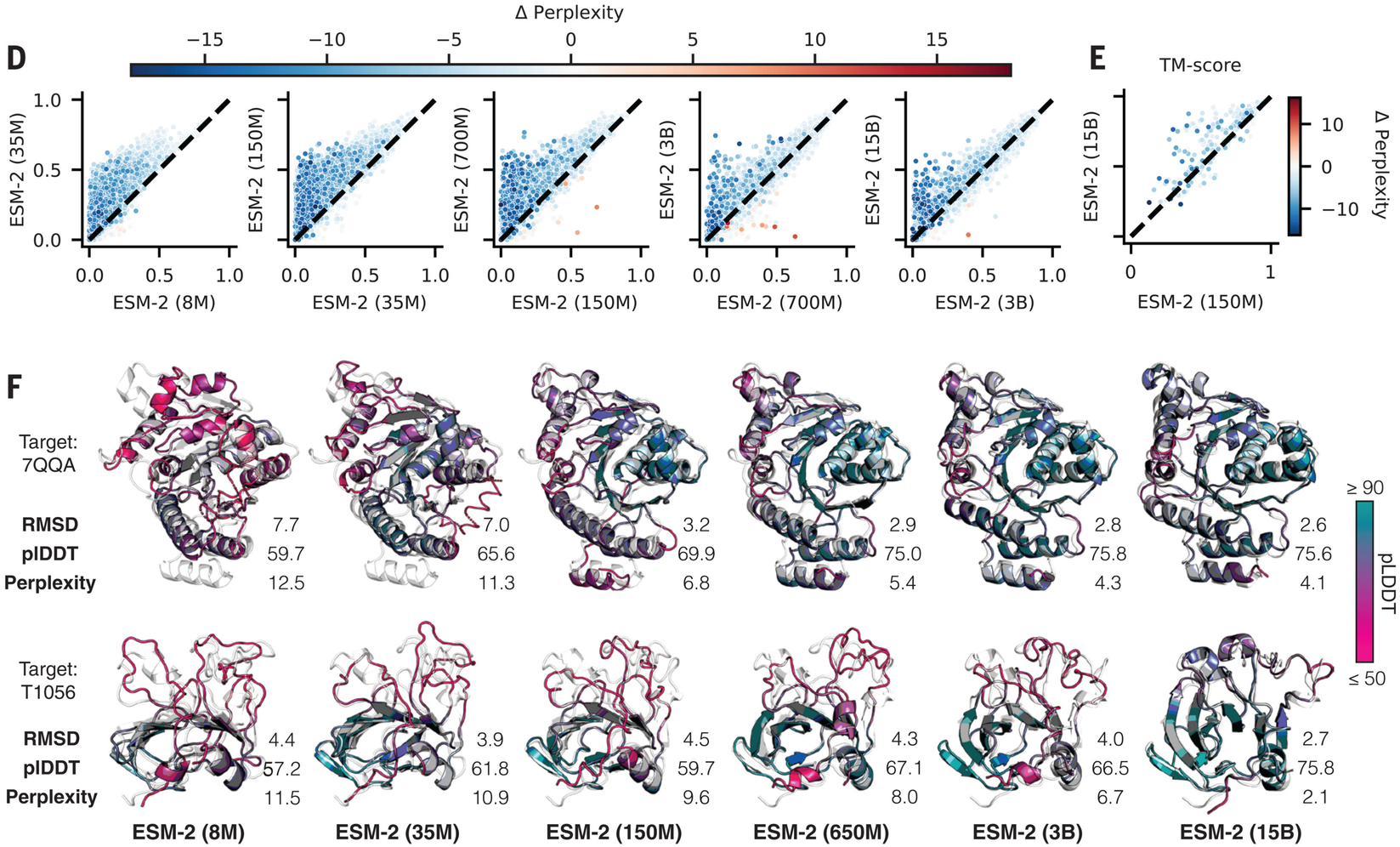
GPS correlates with drug indication



 https://doi.org/10.1126/science.ade2574

Evolutionary-scale prediction of atomic-level protein structure with a language model (Lin, Akin, Rao, Hie et al, *Science*)

- Goal: Build a sequence to structure model
- Method: Combine the ESM protein language model with the AlphaFold-like model to predict structure
- Result:
  - Super fast and super accurate prediction of protein structures
  - Produce >617 million structures from metagenomic sequences (225 million high confidence)
- Conclusion: Powerful new tool for structural biology.

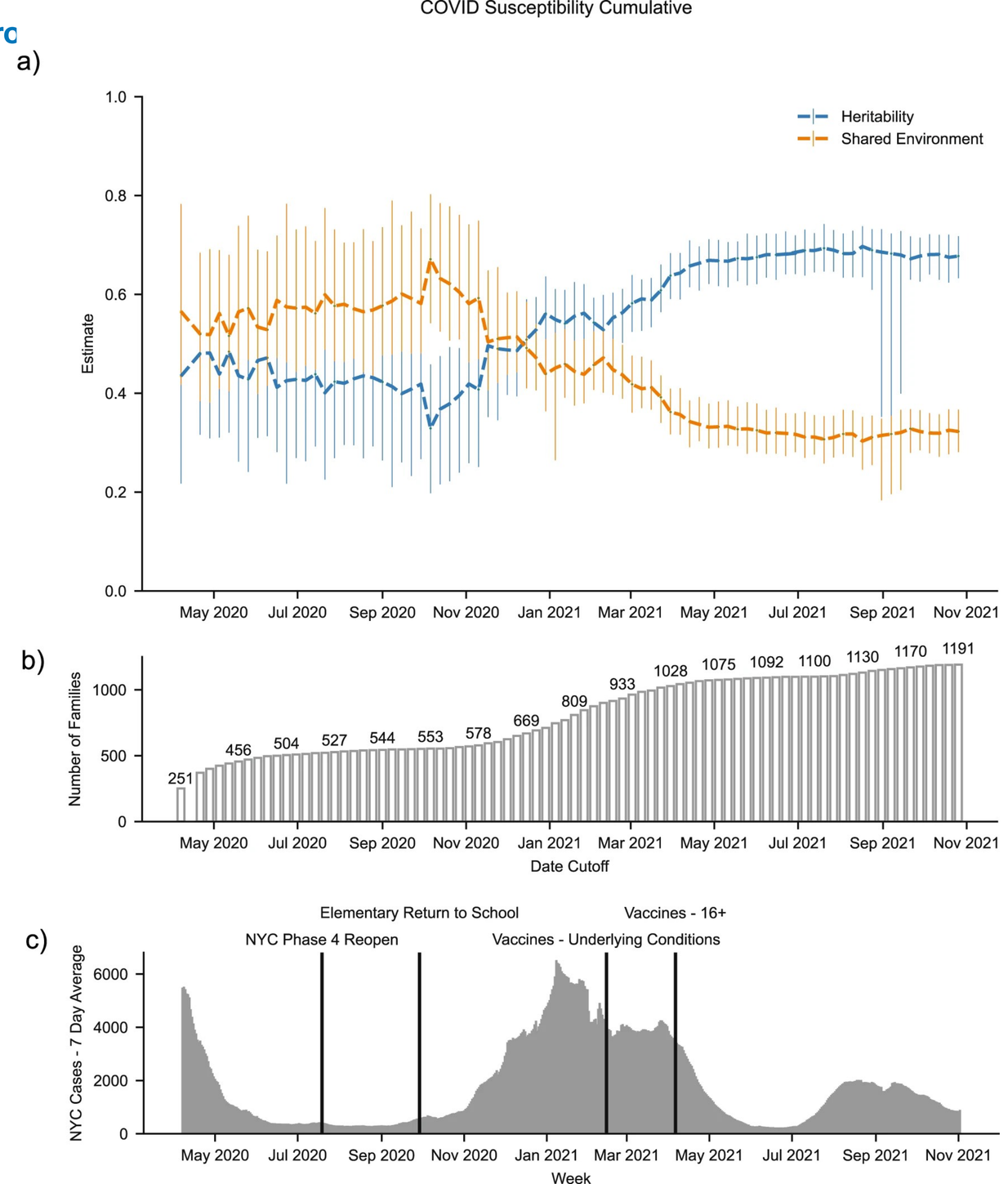


Structure emerges as the number of parameters increases from 8 million to 15 billion

 https://doi.org/10.1038/s41467-023-44250-7

Estimating the heritability of SARS-CoV-2 susceptibility and COVID-19 severity (LaRow Brown et al, *Nature Communications*)

- Goal: Understand the dynamics of covid heritability and environment
- Method:
  - Use EHR-connected pedigree and geographic data for patients in NYC
  - Estimate genetic heritability and environment contribution weekly throughout the pandemic
- Result: Big shifts in heritability from 30% at beginning of pandemic to over 60% later in the pandemic with complementary shift in the environmental contribution
- Conclusion: Public health practices worked to reduce the environmental variation in covid susceptibility and severity


COVID Susceptibility Cumulative
a)
Heritability
Shared Environment
Estimate
1.0
0.8
0.6
0.4
0.2
0.0
May 2020
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
May 2021
Jul 2021
Sep 2021
Nov 2021
b)
Number of Families
1000
500
0
251
456
504
527
544
553
578
669
809
933
1028
1075
1092
1100
1130
1170
1191
Date Cutoff
c)
Elementary Return to School
Vaccines - 16+
NYC Phase 4 Reopen
Vaccines - Underlying Conditions
NYC Cases - 7 Day Average
6000
4000
2000
0
Week

Review of Predictions for 2023

- [x] All of Us will start driving genetic diversity research
- [x] Deep learning will drive advances in molecular dynamics
- [x] More algorithmically-driven and data-driven experimental methods
- [ ] 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
- [x] Data and analysis of single cell proteomics data

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)