Mayo Clinic
Digital Pathology

Bridget Toomey, Josh Meehl, Jake Matras

June 2026

ABOUT US

LEADERSHIP + CORE TECHNICAL TEAM

- Leadership with decades of experience at Mayo
- Rapidly growing technical team

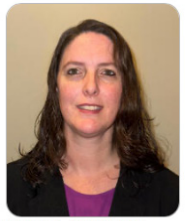
Reporting Unit Owner
Rogers, James A. III
CEO-MAYO CLINIC DIGITAL PATHOLOGY

Panos Korfiatis, Ph.D.
CONS-SENIOR AI SCIENTIST

Matt F. Redlon
CHAIR-AI PROGRAM

Andrew P. Norgan, M.D., Ph.D.
CONS-PATHOLOGY MD

Bridgette A. Wagner
DIRECTOR-BUSINESS DEVELOPMENT OPERATIONS

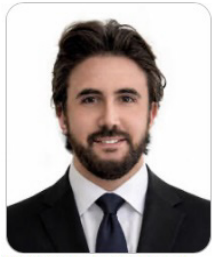


Crystal D. Dison, M.B.A.
BUSINESS ADMIN PARTNER-DIGITAL PATHOLOGY

Nicole K. Hanf, M.A.
BUSINESS ADMIN PARTNER-DIGITAL PATHOLOGY

Caitlin K. Riley
SENIOR COMMUNICATIONS SPECIALIST

Bridget K. Toomey, M.S.
PRINCIPAL AI/ML ENGINEER

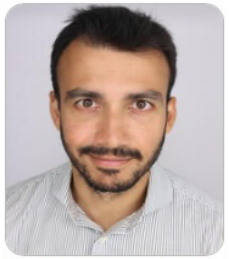


Derek P. Alexander, Ph.D.
PRINCIPAL AI/ML ENGINEER

Patrick H. Duffy
PRINCIPAL DATA ENGINEER

Jake R. Matras
SENIOR AI/ML ENGINEER

Josh D. Meehl, M.S.
SENIOR DATA SCIENCE ANALYST



Vivekananda Sarangi, Ph.D., M.S.
SENIOR DATA SCIENCE ANALYST

Abrar Zahin, Ph.D.
RTP-RESEARCH FELLOW-ENG

Carl P. Molnar, M.S.
SOFTWARE ENGINEER

REDEFINING DIAGNOSIS

Advancing digital pathology through AI innovation

OUR AIM

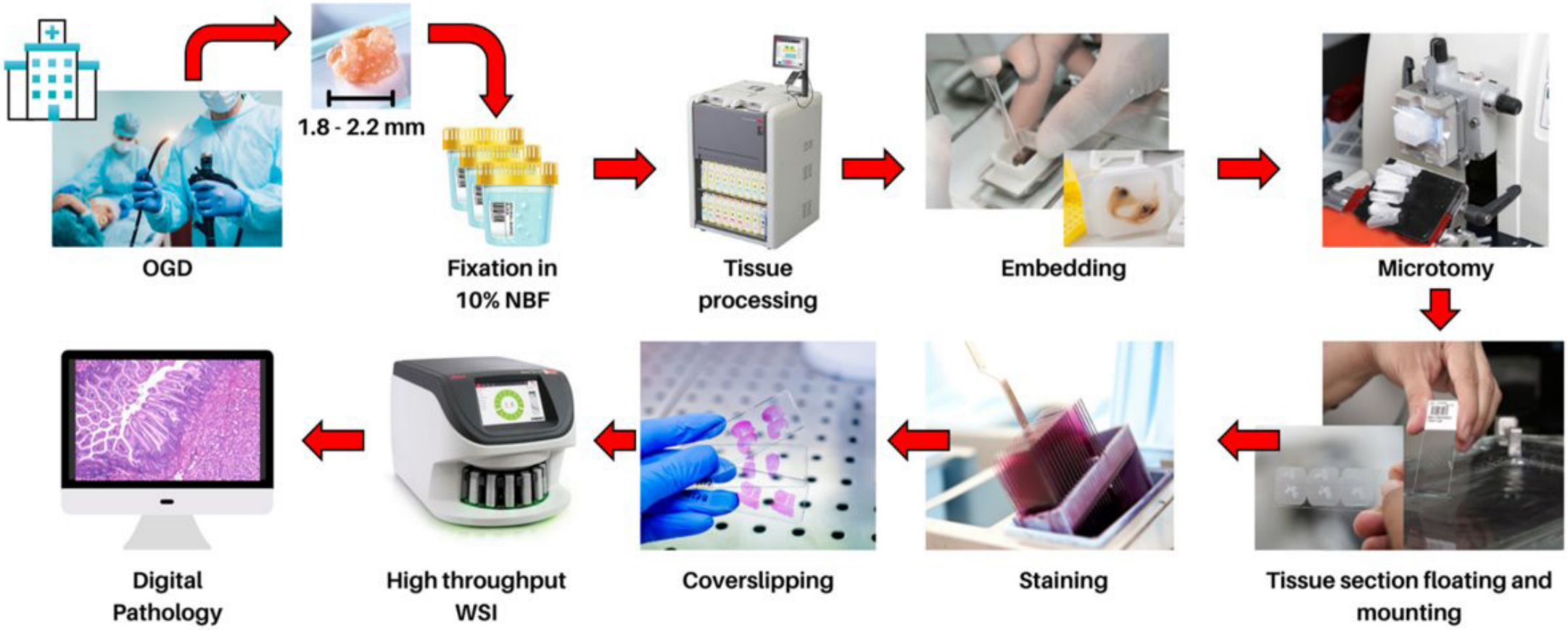
Transform pathology from a manual, subjective process into a data-driven, precise science

By digitizing slides and integrating AI, we can deliver:

- Faster diagnoses
- Greater accuracy
- Lower costs
- Personalized care

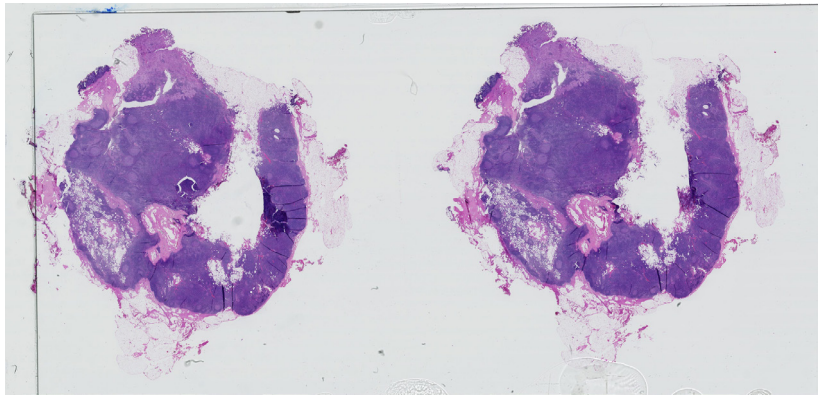
BACKGROUND

WHAT IS DIGITAL PATHOLOGY?



BACKGROUND

WHAT IS A WSI?



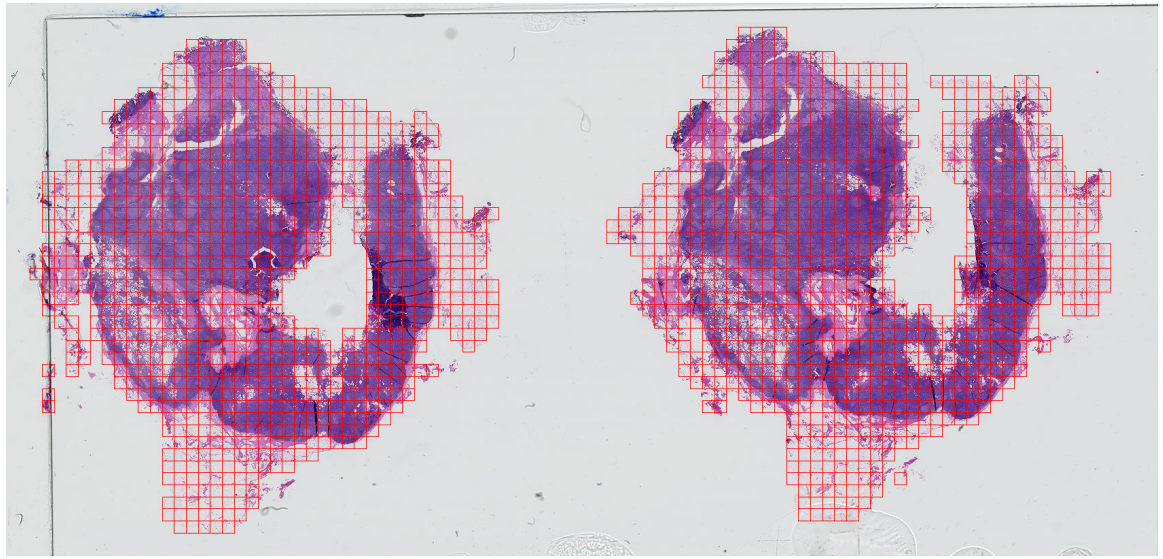
H&E stained lymph node section from CAMELYON16 dataset

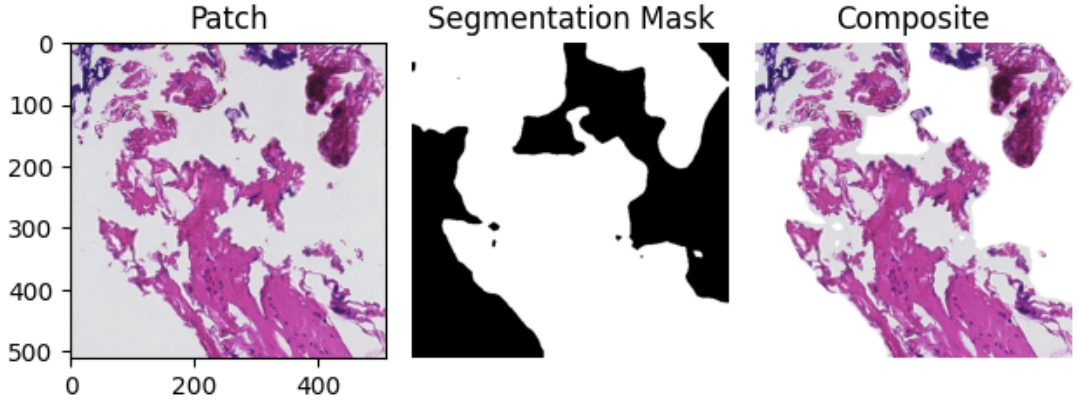
- Whole Slide Images (WSIs) are:
  - very high resolution (e.g. 0.25 µm per pixel)
  - 10s of Gigapixels
  - ~1-4 GB per WSI (compressed)
  - just a bunch of JPEGs in a tiled arrangement
- Fundamental challenges include:
  - **Diversity** of tissue prep, staining, scanning (generalizable models are elusive)
  - **Lack of labels** at pixel level (weak correlation to report text)
  - **Scale of storage** & **computation** (20M WSIs ~ 500 PBs of uncompressed pixel data)
  - **Few public datasets** due to privacy — precludes accessible and reproducible science

BACKGROUND

HOW ARE WSI TYPICALLY PROCESSED?

- Step 1: Tissue Detection
  - simple: thresholding of RGB/HSV
  - more complex: tissue segmentation model
  - reduces computational burden & useless information

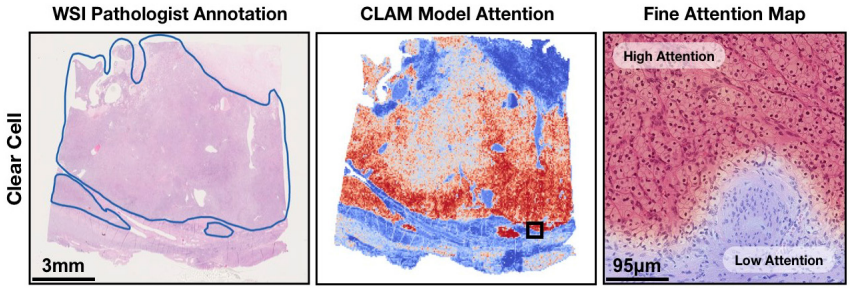


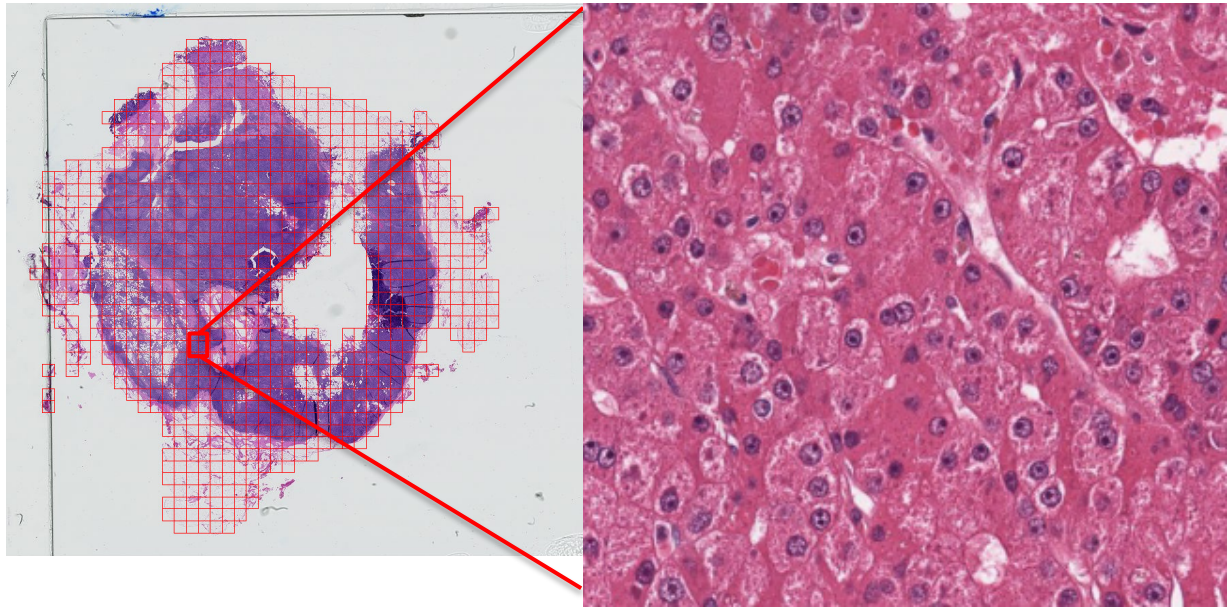


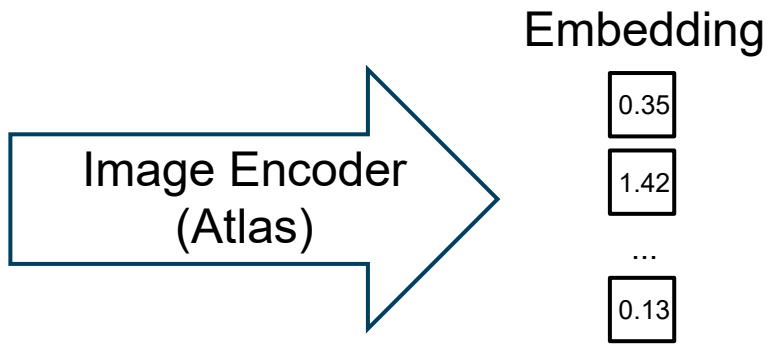
BACKGROUND

HOW ARE WSI TYPICALLY PROCESSED?

- Step 2: Feature extraction
  - split WSI into tiles eg 224x224 pixels @ .25 µm/px
  - embed each tile with an image encoder
  - train classifiers and other models on downstream task



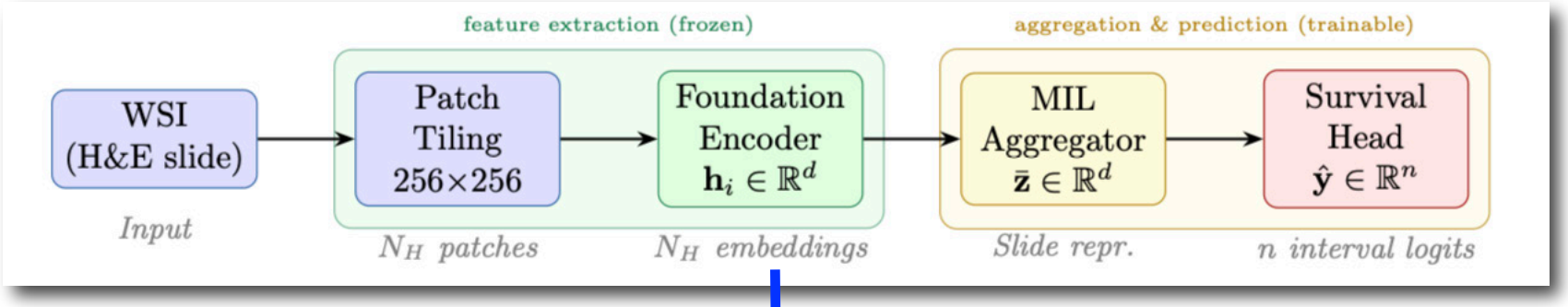




EXAMPLE: SURVIVAL PREDICTION

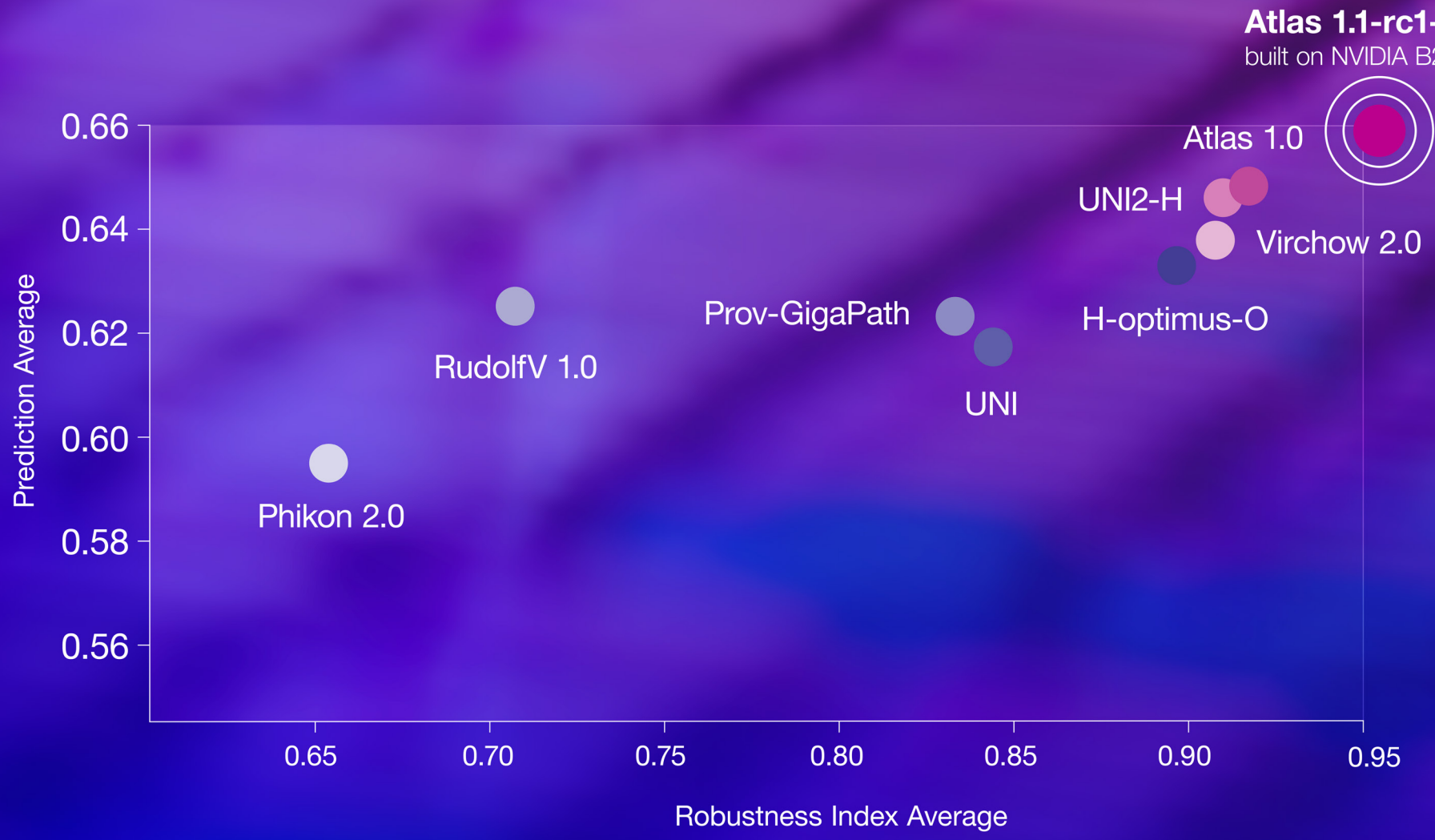
Survival prediction can be framed as a classification task

Whole slide images → Classifier → Output: The probability that patient dies or survives



ATLAS v1 (trained on 1.2M WSIs) and ATLAS v2 (trained on 5.5 WSIs) used as frozen encoders

Atlas Foundation Model



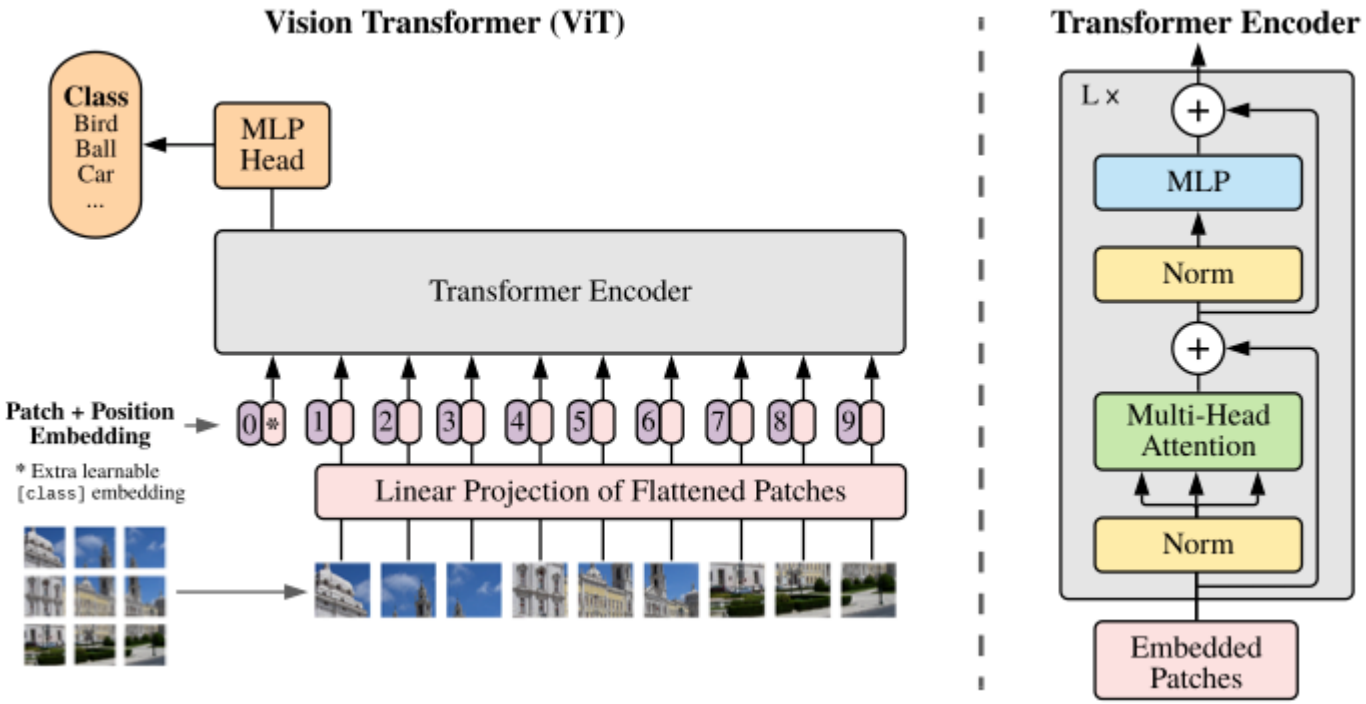
↑3.7%

Our latest model improves by 3.7% on robustness benchmarks over Atlas 1.0 and thereby extends the lead on available competitor models.

ATLAS FOUNDATION MODEL

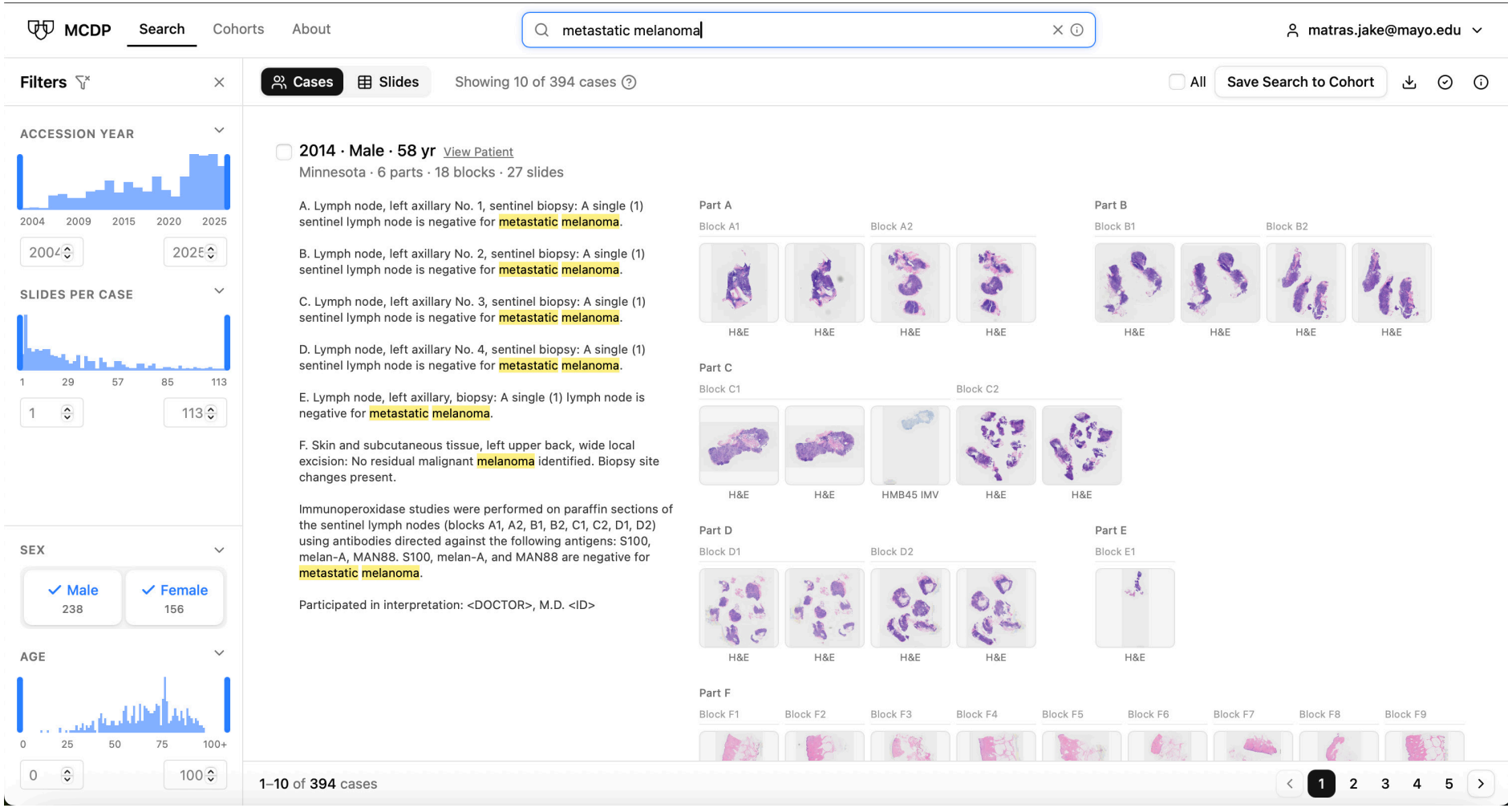
ARCHITECTURE

- Vision Transformer (ViT)
  - 640M parameters
  - ~220 GFlops
  - 32 layers
  - 16 heads
  - embedding dim 1280
  - 8 register tokens
  - patch size 14px
  - flash attention



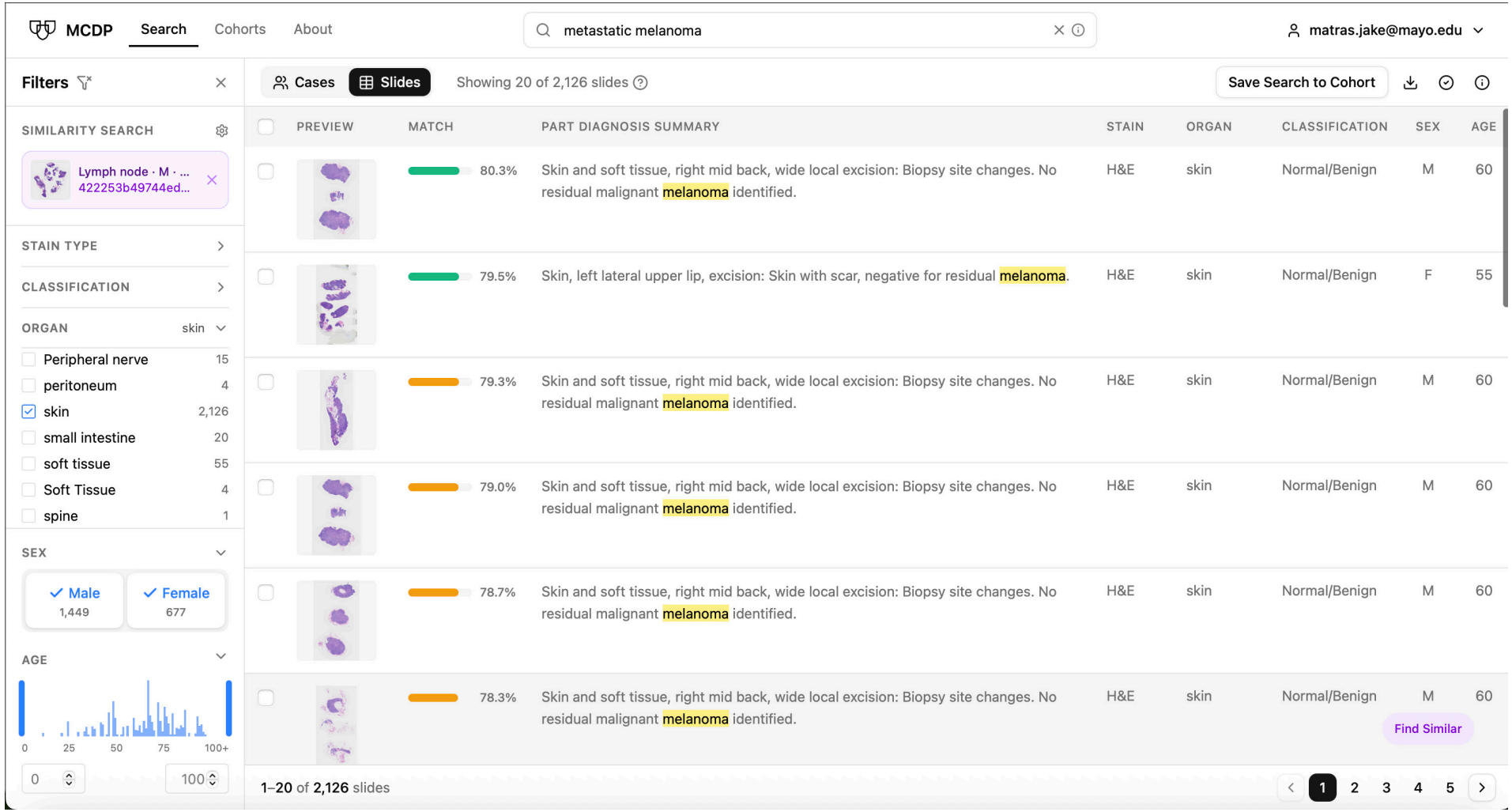
EXAMPLE: VECTOR SEARCH

UNLOCKING THE VALUE OF DIGITAL PATHOLOGY ARCHIVES



EXAMPLE: VECTOR SEARCH

UNLOCKING THE VALUE OF DIGITAL PATHOLOGY ARCHIVES



The Future Is Multimodal

Biology

Body

Organ

Tissue

Cellular

Subcellular

Data

Models

Digital twin

Predictions

Diagnosis

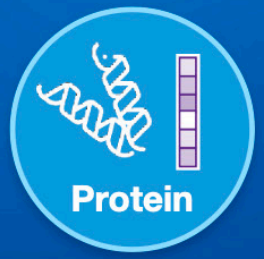
Prognosis

Therapy optimization

Simulate and predict biology
before touching a lab bench

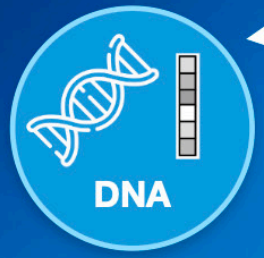
Molecular Embeddings

Analyzes protein sequences for structure, interactions, and functional disruptions.



Predicts changes in transcript abundance and regulatory dynamics.

Predicts patient-specific genetic variants and disease associations.



Cellular Representations

Predicts cell-specific transcriptome-proteome states and lineage trajectories.

Cell State

Predicts subcellular organization and morphological features in individual cells.

Morphology

Multi-cellular Representations

Aggregates single-cell embeddings to represent overall tissue composition and interactions.

Cell Aggregation

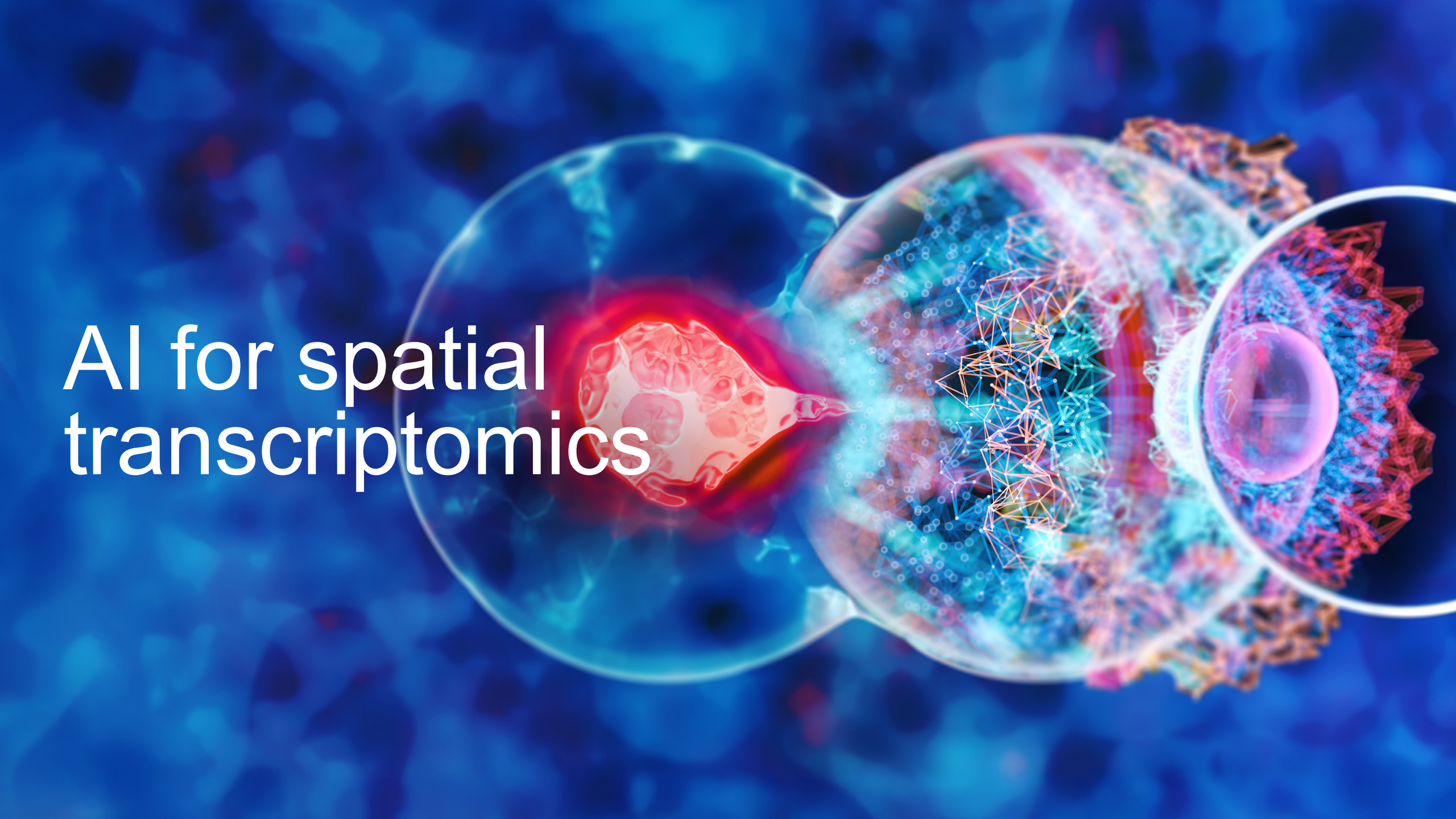
Analyzes tissue architecture and cell organization from 2D/3D microscopy images.

Tissue Imaging

Links gene expression profiles to precise locations in 2D/3D tissues.

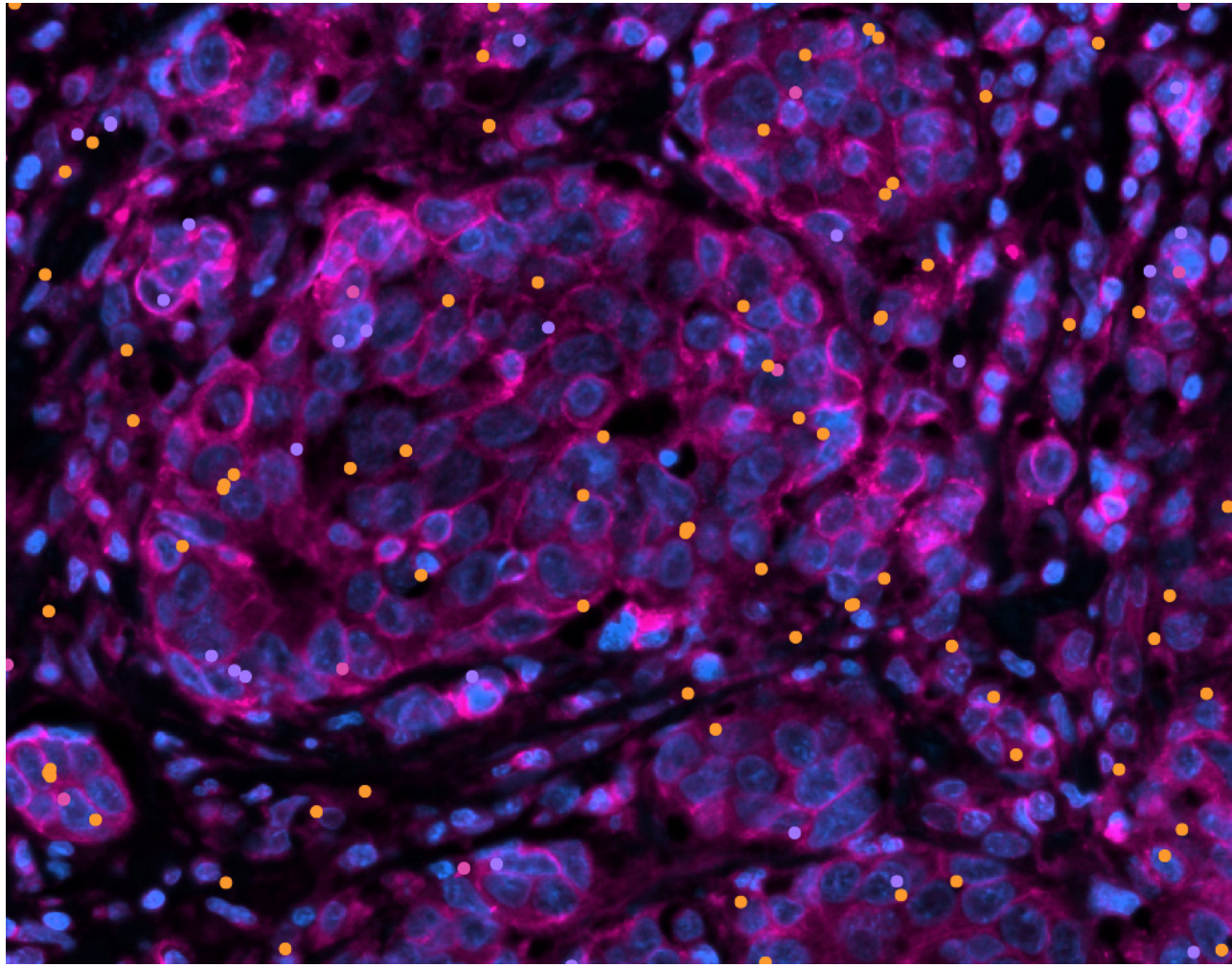
Spatial Transcriptomics

AI for spatial transcriptomics



WHAT IS SPATIAL TRANSCRIPTOMICS?

- Transcriptomics: study of RNA, or the transcribed part of the genome
- Spatial transcriptomics (ST): focus on spatial arrangement of transcripts within a sample
- Too expensive for clinical use
- Can vary in spatial resolution & breadth of transcripts detected
  - Visium: full transcriptome, less spatial resolution
  - Xenium: 5k panel, fine-grained resolution



Example Xenium output from a breast cancer sample. The purple dots are BRCA1, the pink ones BRCA2, and the orange ones PTEN. The blue stains are the nuclei, and the pink are the cell boundaries.

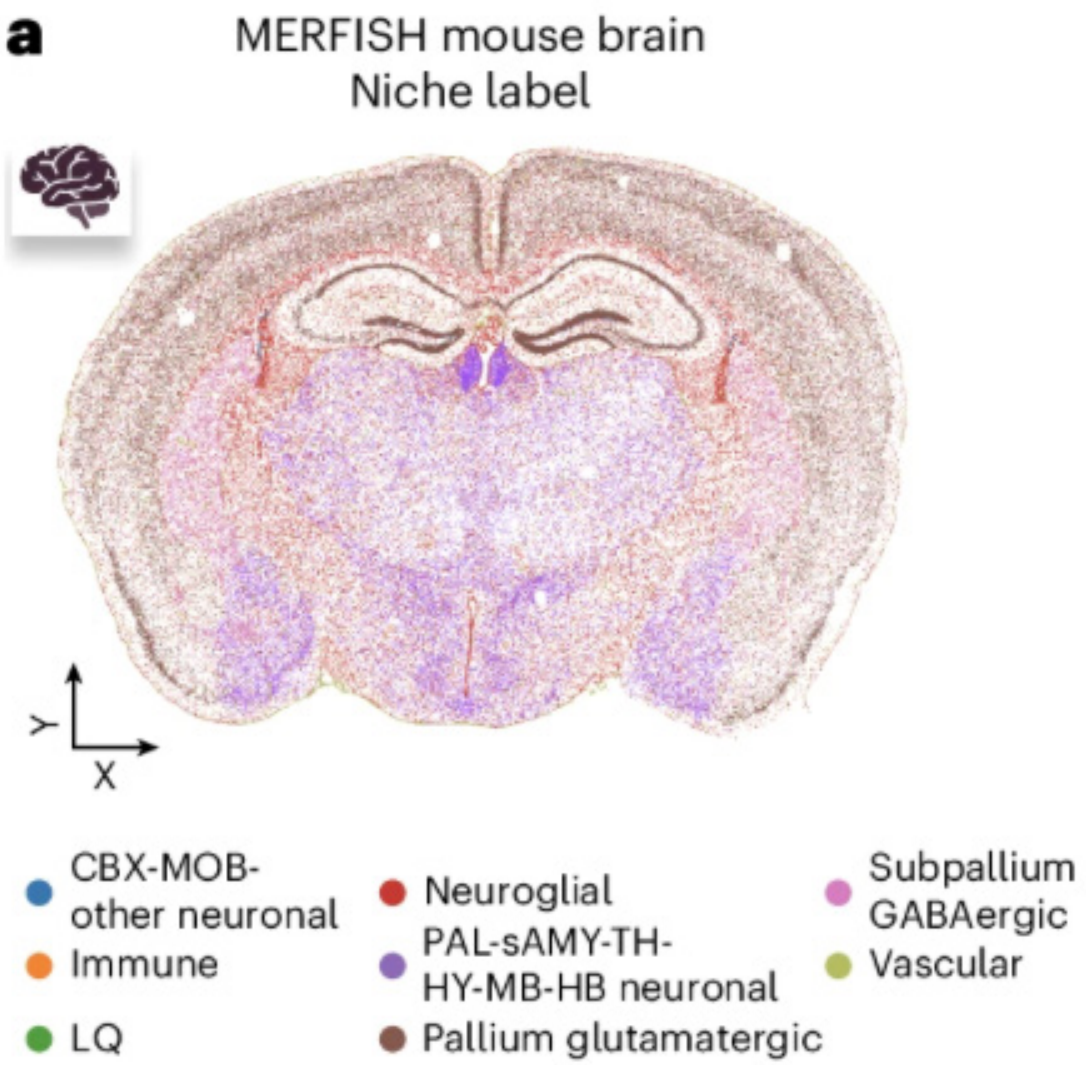


Figure 4a from the Nicheformer paper

EXISTING ST MODELS: NICHEFORMER

- Tejada-Lapuerta et al, 2025
- Trained base model on both spatial transcriptomics and scRNA-seq data
- Treats the cell as the fundamental unit
- Fine-tuned on spatial transcriptomics data to do downstream spatial tasks
- Only performs comparably to PCA on some downstream tasks

EXISTING ST MODELS: SCGPT-SPATIAL

- Wang et al, 2025
- Performs continual pre-training on scGPT (trained on scRNA-seq)
- Uses a mixture of experts decoder to route samples

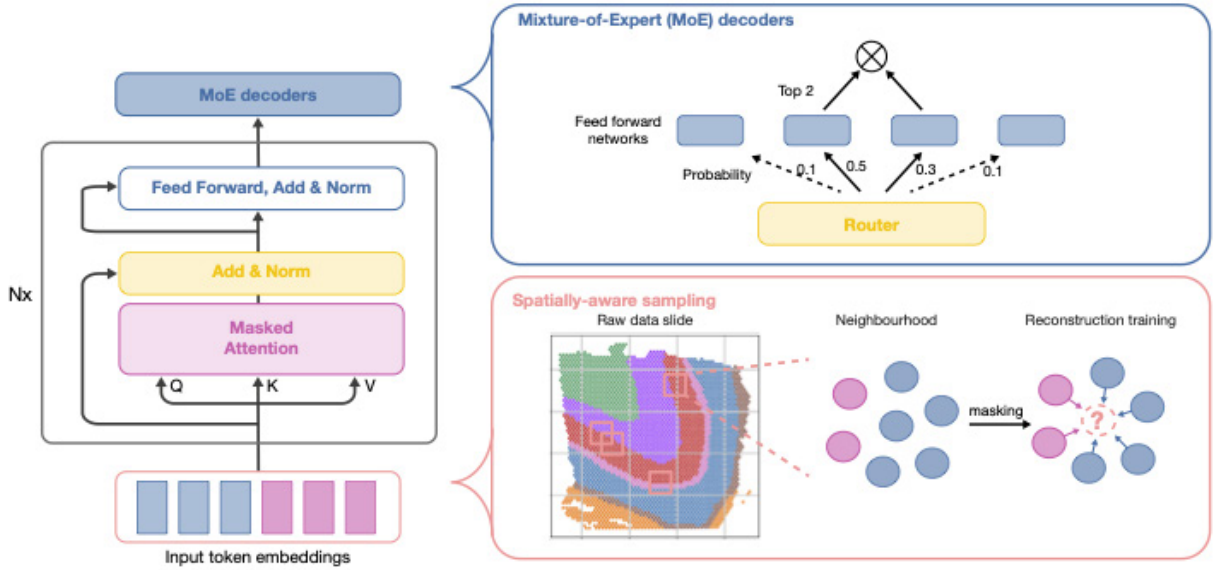
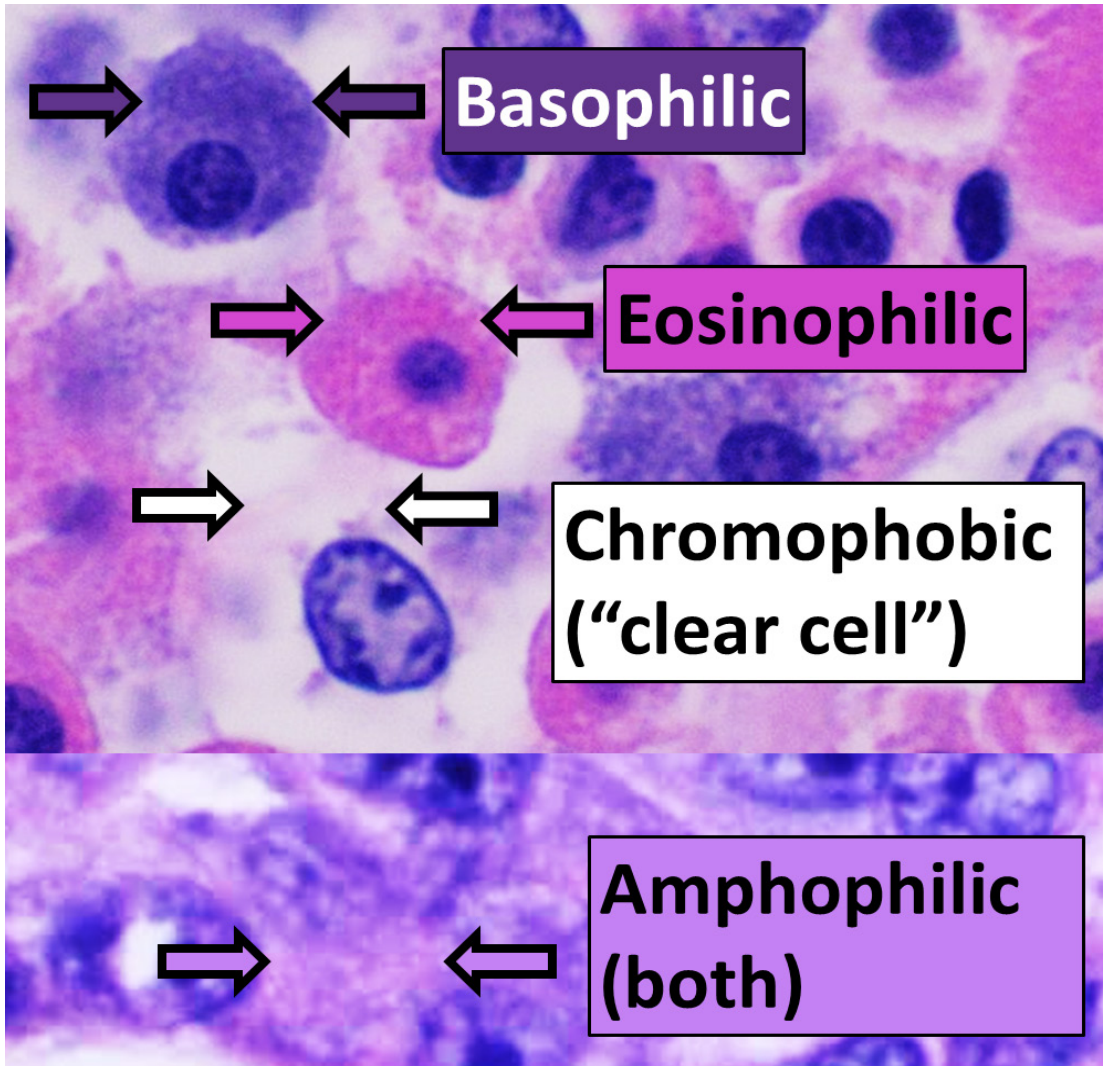


Figure 1B from the scGPT-spatial paper. The spatially-aware sampling is an interesting technique, too.

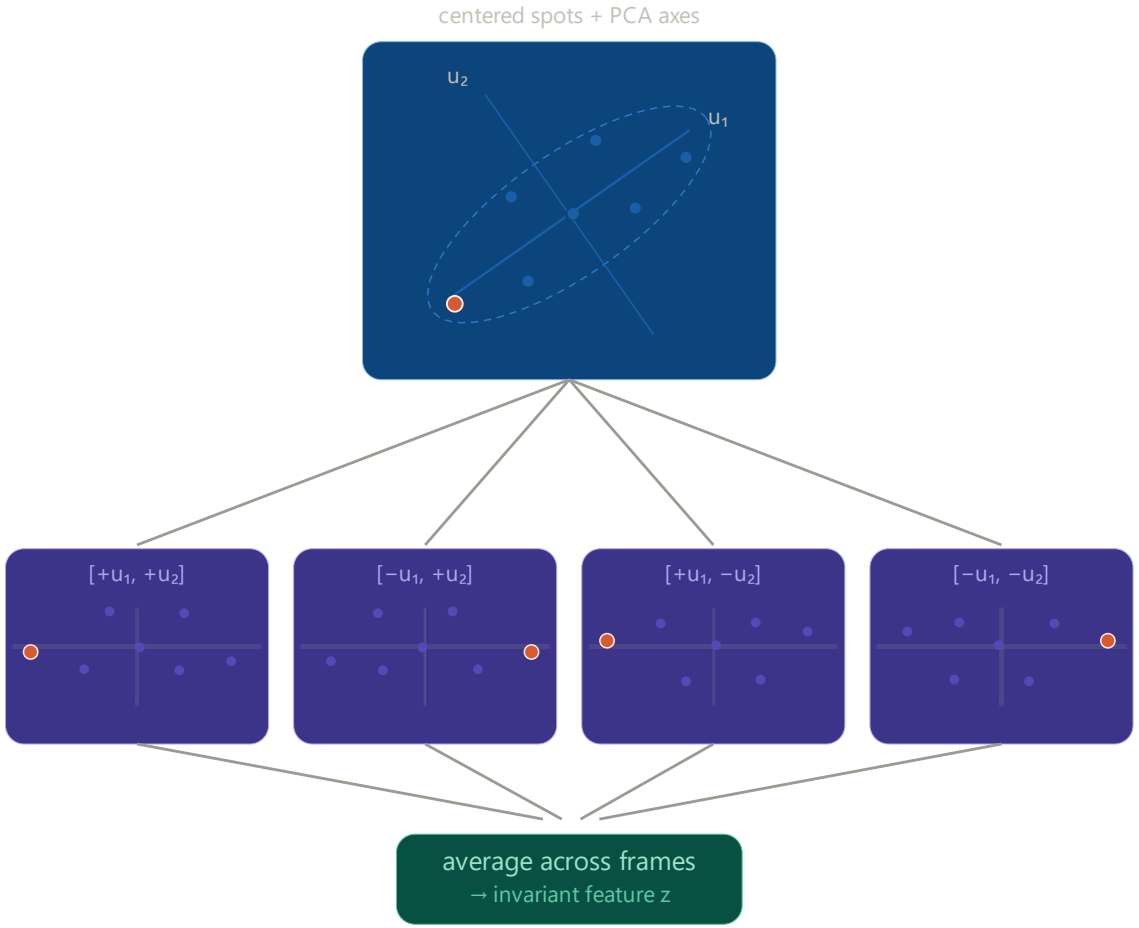


PAIRING H&E AND ST

- Hemotoxylin and eosin (H&E): primary stain used in pathology
- Hemotoxylin: stains nuclei purple/blue
- Eosin: stains cytoplasm and extracellular matrix pink
- Expense of ST has encouraged search for alternatives
- One option: train in an "ST-aware" way, infer on H&E only

Source: Dr. Mikael Haggstrom, M.D.

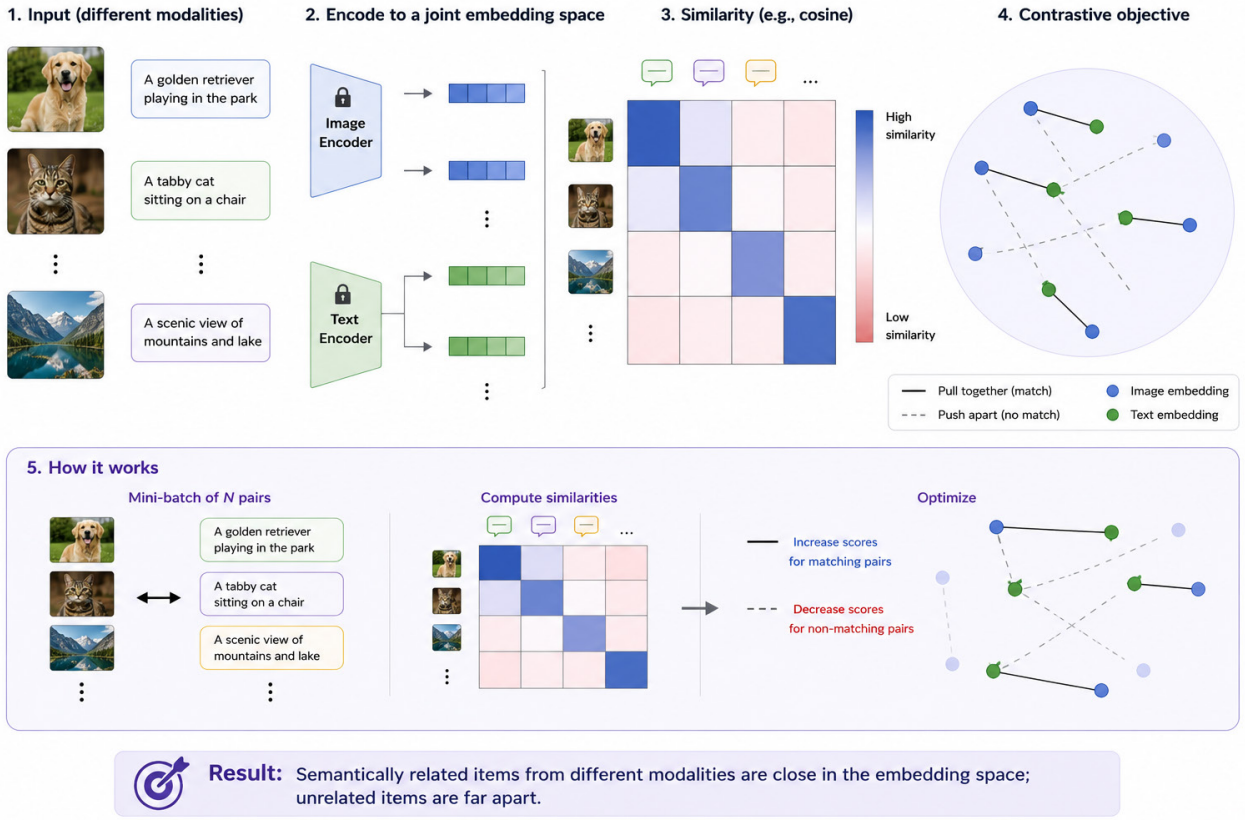
License



ML ARCHITECTURAL PRE-REQS 1

- Naïve attention: no indication of tokens' relative positions
- Common method in LLMs: positional encoding. Can be added to token embeddings or to logits.
- Frame averaging: perform PCA on relative distances between pairs of points. Then project the distances onto PCs, and finish with a linear projection to a scalar.
- Ultimate result: a scalar that preserves anisotropy but is also invariant to rotations, translations, and reflections

ML ARCHITECTURAL PRE-REQS 2



- Parameter-efficient fine-tuning (PEFT): fine tune a model without updating all parameters
- Low-rank adaptation (LoRA): version of PEFT where a low-rank matrix adapters are learned
- Contrastive loss:
  - Useful for simultaneous learning across multiple modalities
  - Push pairs of the same sample across modalities together, pairs of different samples apart

PAIRED H&E AND ST: STPATH

- Training corpus:
  - Open source data from Visium, Visium HD, Xenium, and Spatial Transcriptomics
  - 16+ organs
- Key design choices:
  - Gene masking schedule -> can infer ST from H&E only
  - Loss focus on top 200 highly variable genes
  - Frame averaging

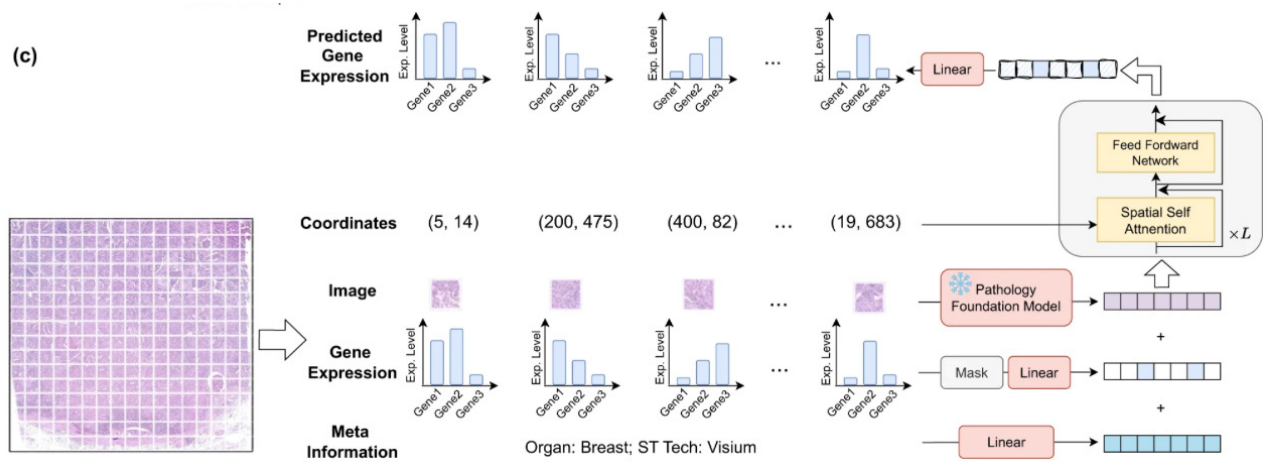


Figure 1c from the STPath paper

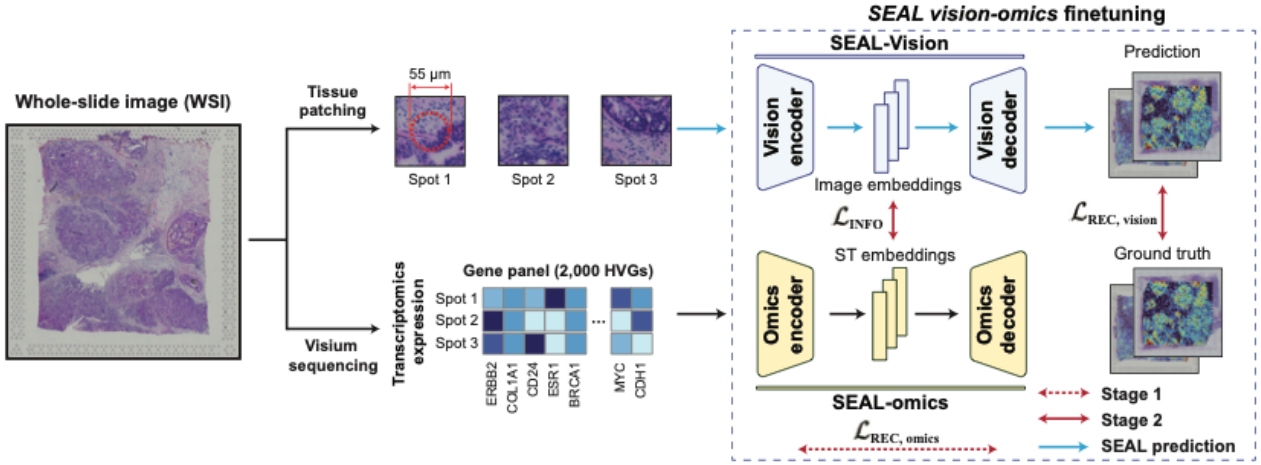


Figure 1C from the SEAL paper

PAIRED H&E AND ST: SEAL

- Train an ST encoder from scratch using a variational autoencoder
  - Use a special loss to mitigate sparsity and heterogeneity
- Then, train a LoRA on top of an existing image encoder using contrastive loss. Also include a loss to reconstruct ST from H&E.
- Only use the LoRA at inference time
- Compared to a full fine-tuning, the LoRA is better on slide-level tasks but slightly worse on patch-level tasks

WHAT CAN WE USE THESE MODELS FOR?

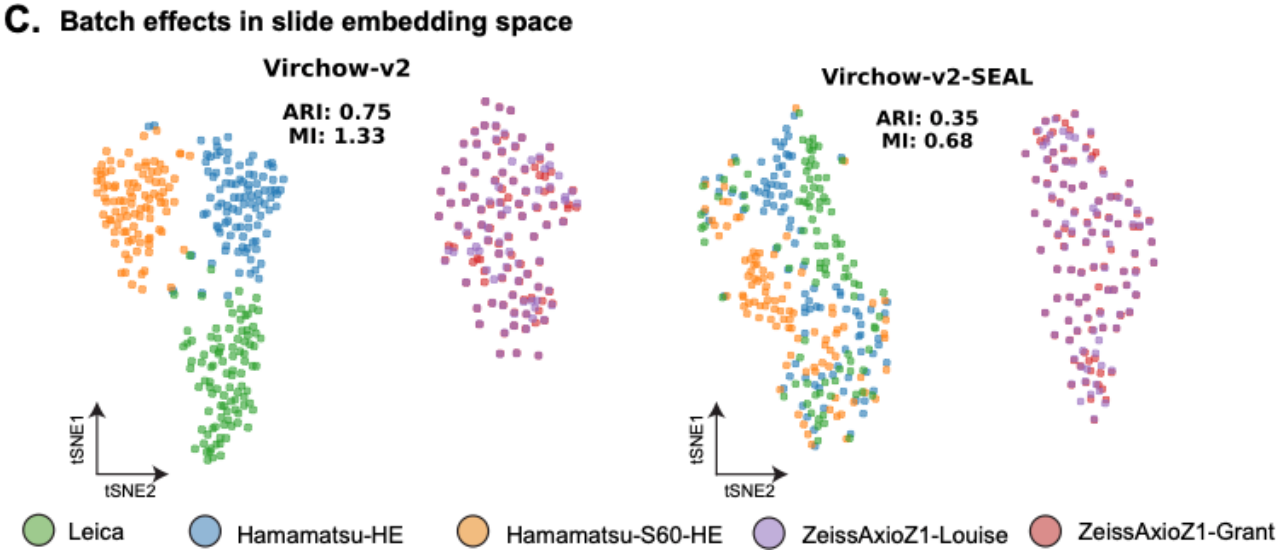


Figure 2c from the SEAL paper

- As Jake showed, we can use MIL with patch-level embeddings to make slide/patient-level predictions
- Hope in this field: guiding H&E embeddings with ST data will yield better predictions
- Results from STPath and SEAL papers indicate improvements on some tasks, but many overlapping confidence intervals
- Interesting SEAL result: improves scanner confounding

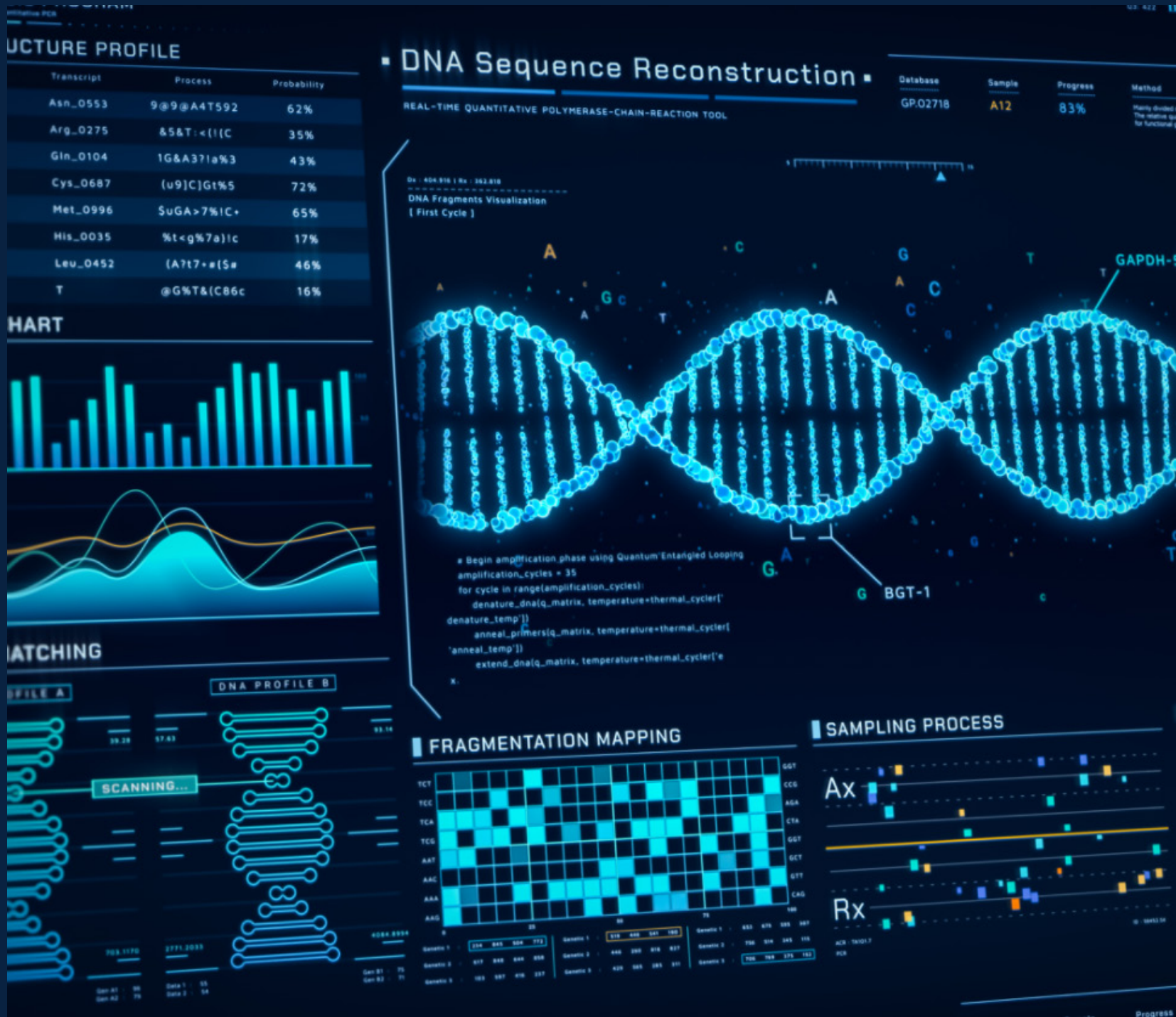
MAYO CLINIC

Genomic Foundation Models

What they predict

How they learn from unlabeled sequence

And where they stop working



WHY THIS MATTERS

We can sequence faster than we can interpret.

Sequencing a genome is routine. Saying what a variant actually does is not.

- Most variants carry no functional label, and a wet-lab assay can take months to years.
- Foundation models can score a variant straight from sequence, with no per-variant label.

~2 million

variants of uncertain significance in ClinVar.

~89 million

missense variants scored zero-shot by AlphaMissense.

0 labels

pathogenicity labels used during pretraining.

WHAT THESE MODELS DO

Three things that were not possible before.

1

Variant effect prediction

AlphaMissense

Scores every possible missense variant by how much it violates learned protein grammar. No pathogenicity labels in training.

2

Noncoding interpretation

SpliceAI · Enformer

Predicts splicing changes near a site, including cryptic sites, and tissue-specific expression effects of regulatory variants.

3

Polygenic risk scoring

FM embeddings

Richer sequence representations can add to classical PRS for complex-disease risk stratification.

Notice where all three live. They sit at the **molecular layer**. That is not a coincidence, and we return to it.

APPLICATION ONE, UP CLOSE

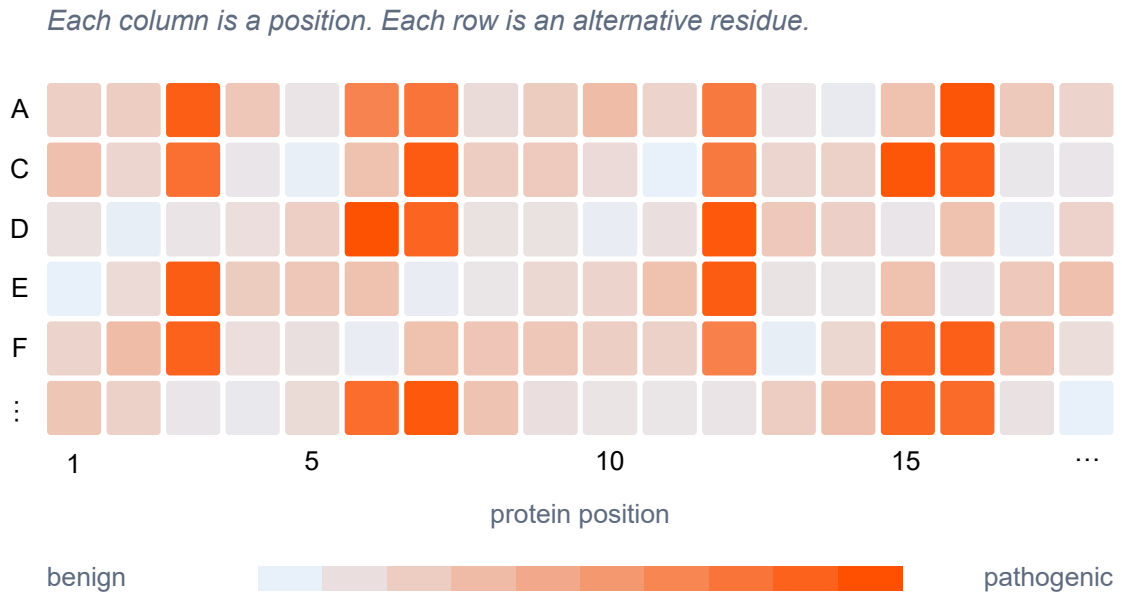
Scoring every possible missense variant.

AlphaMissense assigns a pathogenicity score to every missense substitution in the proteome, about 89 million of them, with no pathogenicity labels used in training.

How. A substitution that breaks patterns conserved across species scores toward pathogenic. Conservation is the supervision.

Output. A calibrated score per variant, usable as ACMG PP3 or BP4 computational evidence.

Payoff. A triage path for the roughly 2 million variants of uncertain significance in ClinVar, without waiting for assays.



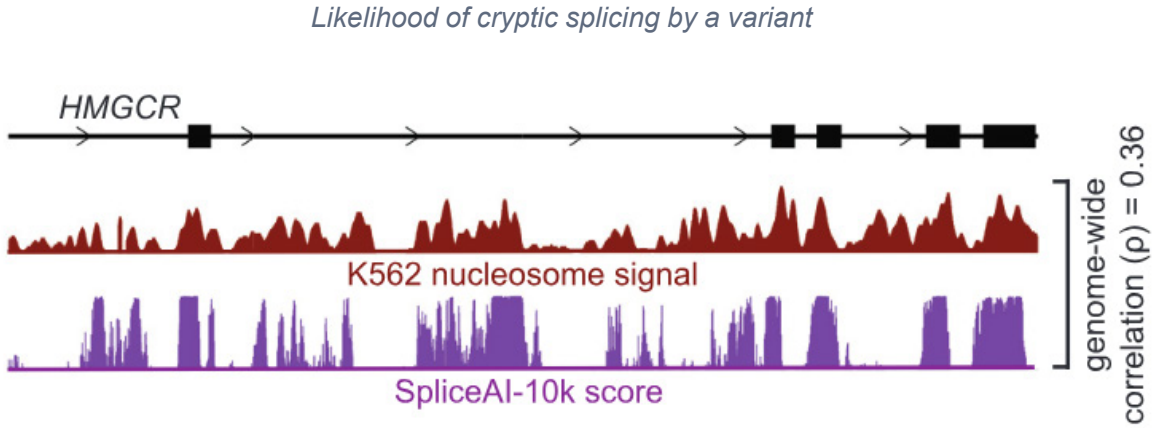
APPLICATION TWO, UP CLOSE

Reading variants outside the coding regions.

Most of the genome, and most GWAS signal, sits outside protein-coding regions.

SpliceAI. Flags variants that create or destroy splice sites, including cryptic sites that rule-based tools miss. In clinical pathology use at Mayo and elsewhere.

Enformer. Predicts tissue-specific expression changes, which helps explain noncoding GWAS hits that older methods left as associations with no mechanism.



THE SHARED IDEA

One principle behind all of them.

Self-supervised pretraining. Every model learns sequence representations from massive unlabeled data first, then adapts to a task.

Pretrain

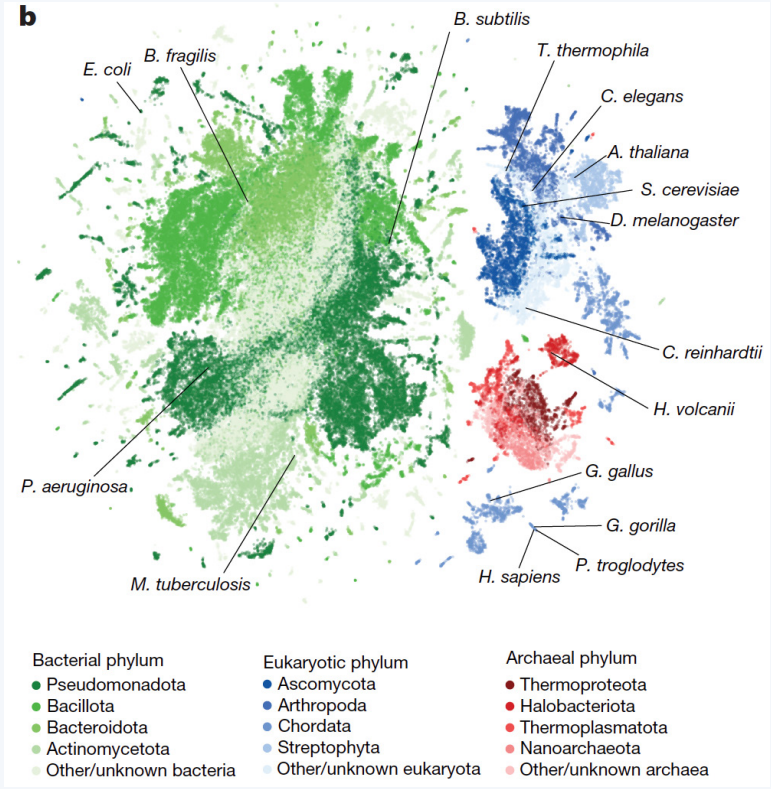
Predict masked tokens across billions of nucleotides from thousands of species.

Adapt

Apply it directly, or attach a small task head and train on a few thousand labels.

THE KEY INSIGHT

Evolution already labeled the data. A base conserved across mammals is doing something. A model trained to fill in masked sequence has to learn those patterns to succeed.



One model's learned sequence space separates bacterial, eukaryotic, and archaeal genomes. It spans the tree of life from unlabeled sequence alone.

Evo 2, Fig. 1B · Brixi et al. 2026

TWO WAYS TO USE A PRETRAINED MODEL

Zero-shot or fine-tuned.

Zero-shot

Apply the pretrained model as it is. Score a new variant by how far it departs from learned patterns. No task labels needed.

Fine-tuning

Add a small task head and train on a few thousand labeled examples. The pretrained representations do the heavy lifting.

Think of a biologist with broad undergraduate training (pretraining)

They can reason about unfamiliar problem with the fundamentals alone (zero-shot)

Makes doctoral specialization faster and more robust (fine-tuning).

HOW IT WORKS, NO EQUATIONS

Masked language modeling.

Hide about 15 percent of the bases in a sequence. Ask the model to predict what was there.
To do this well, it has to learn which bases belong in which context.

MASKED

A	T	G	C	A	?	G	C	T	A

Predicted distribution for the masked base

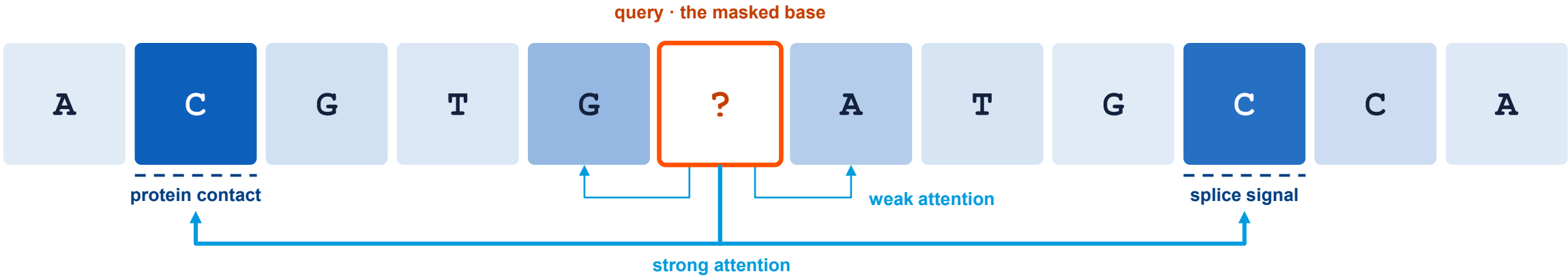
ATTENTION, IN ONE LINE

To fill the gap, the model weighs every other position and leans on the few that are informative, like a splice signal and the exon it governs.

HOW IT WORKS, NO EQUATIONS

Attention decides what to look at.

To fill the masked base, the model weighs every other position and leans on the few that carry information. It learns those links from data, not from rules.

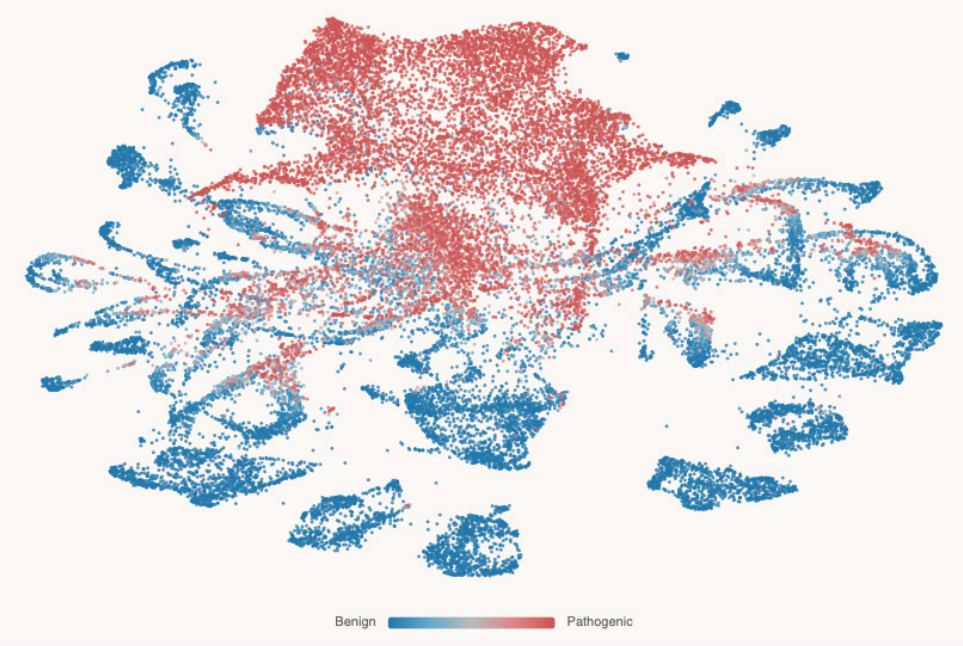


WHY THIS MATTERS FOR INTERPRETATION

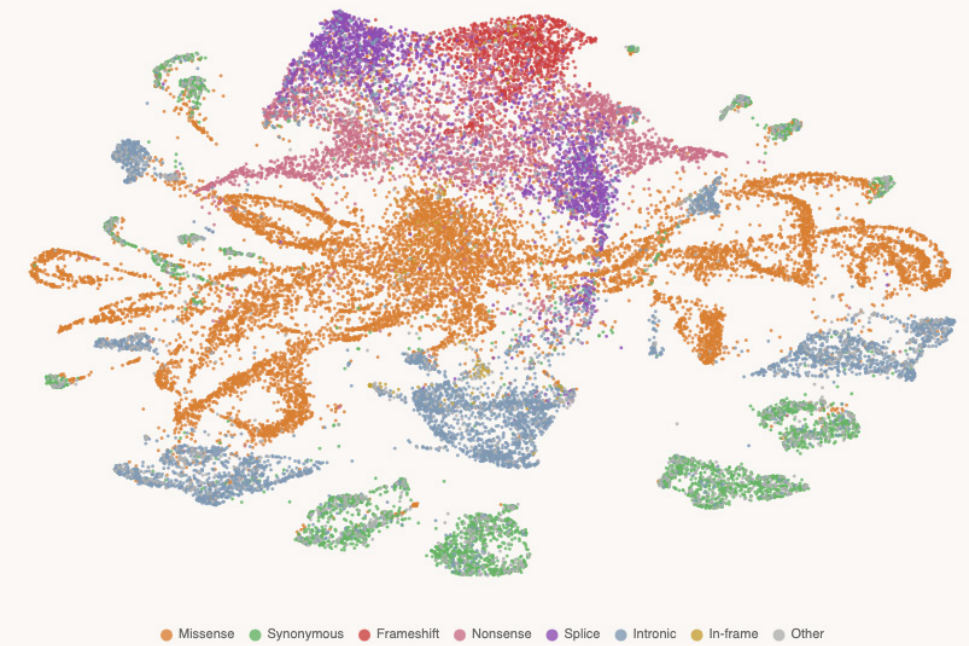
These weights are also what interpretability tools read to partially explain a score. Partial, and not yet clinical-grade on their own.

WHY A THIN PREDICTOR WORKS

The embedding already separates variants.



Colored by predicted pathogenicity. Benign and pathogenic variants settle into different regions of the space.



The same space, by consequence class. Structure the model was never given as a label.

WHY THIS MATTERS

EVEE builds a light, interpretable predictor on top of these Evo 2 embeddings. The representation has already done most of the work, which is why a thin head reaches high accuracy and generalizes zero-shot.

HOW IT WORKS, NO EQUATIONS

Context length is a biological limit.

A model that reads only 500 bases cannot learn that an enhancer 50 kb away controls a gene. Wider reading windows make longer-range biology visible.

DNABERT era. Sees local sequence grammar and splice motifs.

Mid-range models. Sees the gene body plus nearby regulatory elements.

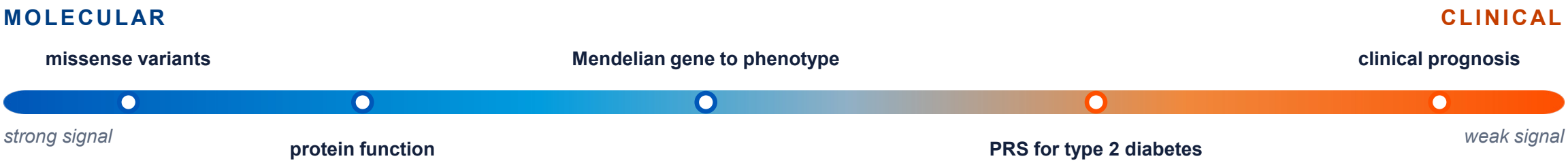
HyenaDNA, Evo2. Sees distal enhancers and long-range regulation.

A wider window means more of the regulatory genome the model can actually see.

STRENGTHS AND LIMITATIONS

The molecular-to-clinical spectrum.

GFMs are strong where sequence determines the outcome. They weaken as clinical complexity stacks above the molecular layer.



Excels at variant to protein function, splicing, expression, and one-variant Mendelian disease.

Struggles with type 2 diabetes, IBD, and CAD, where risk comes from many variants plus environment plus time.

WHY THE LIMITATION IS FUNDAMENTAL

The model has never seen a patient's life.

WHAT PRETRAINING GIVES

Evolutionary conservation and sequence grammar. This aligns tightly with molecular consequence, because what evolution rejects is what breaks protein function.

WHAT CLINICAL RISK REQUIRES

Genetic architecture, plus environment, plus developmental timing. A GFM sees the genome. It does not see the diet, the microbiome, or the trigger of onset.

PRS, FRAMED HONESTLY

FM-enhanced PRS shows modest gains over classical PRS for common disease. Promising, not transformative. The classical baseline is already predictive.

NOISY LABELS

The label Crohn's disease covers stricturing, penetrating, and inflammatory subtypes with different architectures. The model learns a clinical consensus, not a biological truth.

USING A SCORE RESPONSIBLY

Three questions before you trust a GFM score.

1

Right model for this variant?

A protein LM scores missense. A regulatory model scores noncoding. Using the wrong one gives a number with no meaning.

2

Calibrated and validated here?

A confidence score is not a probability of pathogenicity unless it is calibrated, and validated on populations relevant to this patient.

3

Evidence, or verdict?

Is the output one piece of evidence among many, or is it being treated as the final answer? It should be the former.

Integration, not replacement. GFM scores enter ACMG-AMP as computational evidence, PP3 or BP4. They do not replace functional assays, segregation, or population frequency.

WHERE THIS STANDS TODAY

Already in use, not speculative.

IN THE CLINIC NOW

SpliceAI is part of molecular pathology workflows at Mayo and elsewhere for splicing interpretation.

IN RESEARCH EVALUATION

AlphaMissense is being assessed as PP3 computational evidence in research settings.

IN ACTIVE STUDY

DNA language models are under clinical research, with calibration and equity still open questions.

Two caveats travel with every model. Scores can be miscalibrated, and models trained on mostly European ancestry can flag variants common in underrepresented populations as unusual.

TRY IT YOURSELF

Explore a real variant with EVEE.

EVEE, the Evo Variant Effect Explorer, puts this lecture together. It builds an interpretable pathogenicity predictor on top of Evo 2 embeddings and covers every ClinVar variant. A Mayo and Goodfire collaboration.

evee.goodfire.ai

4.2M
ClinVar variants with pre-computed predictions

7B
parameters in Evo 2, no human variant data in training

0.997
AUROC on about 833k ClinVar SNVs

YOUR TASK

1. Open evee.goodfire.ai and search a gene with validated benchmarks, like BRCA1, BRCA2, TP53, or LDLR, or one you work on.
2. Pick a variant of uncertain significance. Read its predicted pathogenicity and its disruption profile, the set of annotation categories the model predicts are broken.
3. Read the synthesized explanation. Ask whether the evidence stays molecular and mechanistic, or drifts toward a clinical claim the model cannot support.
4. Run it through the three questions from the responsible-use slide, then place it on the molecular-to-clinical spectrum.

REFLECTION

Where did the explanation help, and where did you still need a human? EVEE turns embeddings into structured disruption predictions, then a reasoning model writes them up in plain language.

Computational predictions are not diagnoses. Same point as the responsible-use slide, now on a tool you can touch.

QUESTIONS & ANSWERS