Shai Shen-Orr

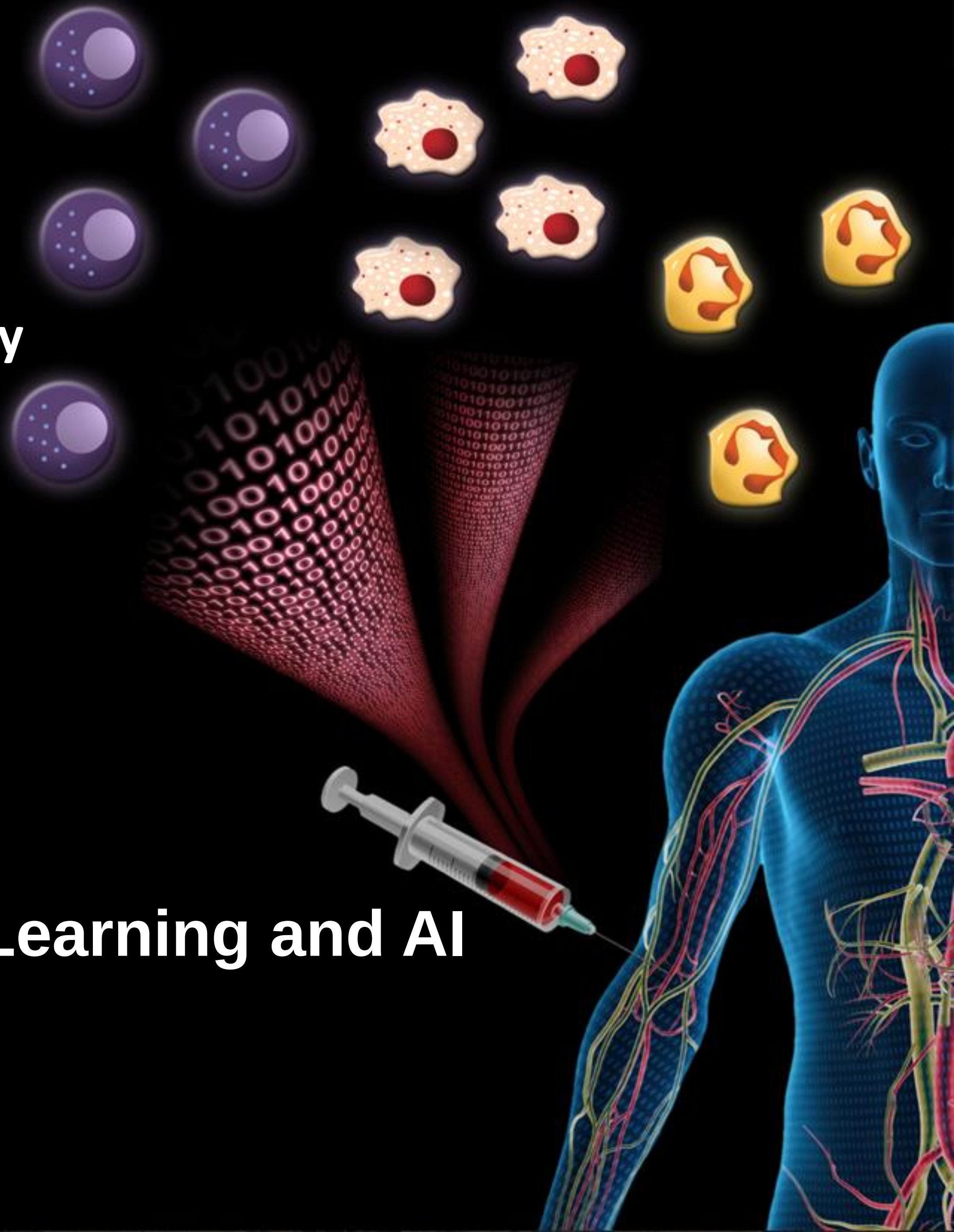
Technion – Israel Institute of Technology

Chief Scientist, CytoReason

Human Immunome Project

Co-CSO

# Unraveling Immunity through Machine Learning and AI



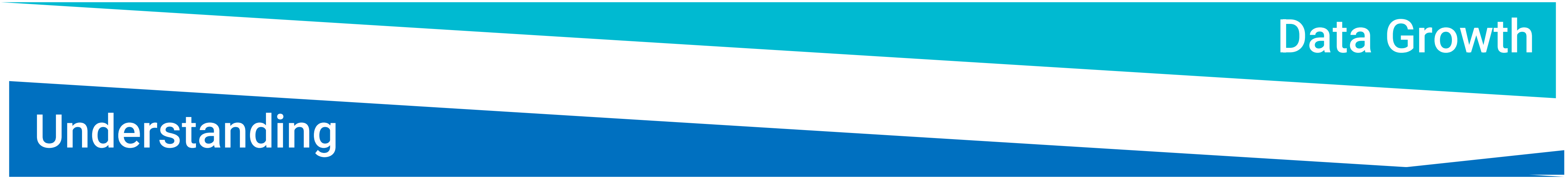
# The complexity I have lived through and where it has led me

E.Coli
**4.6MB**
2000

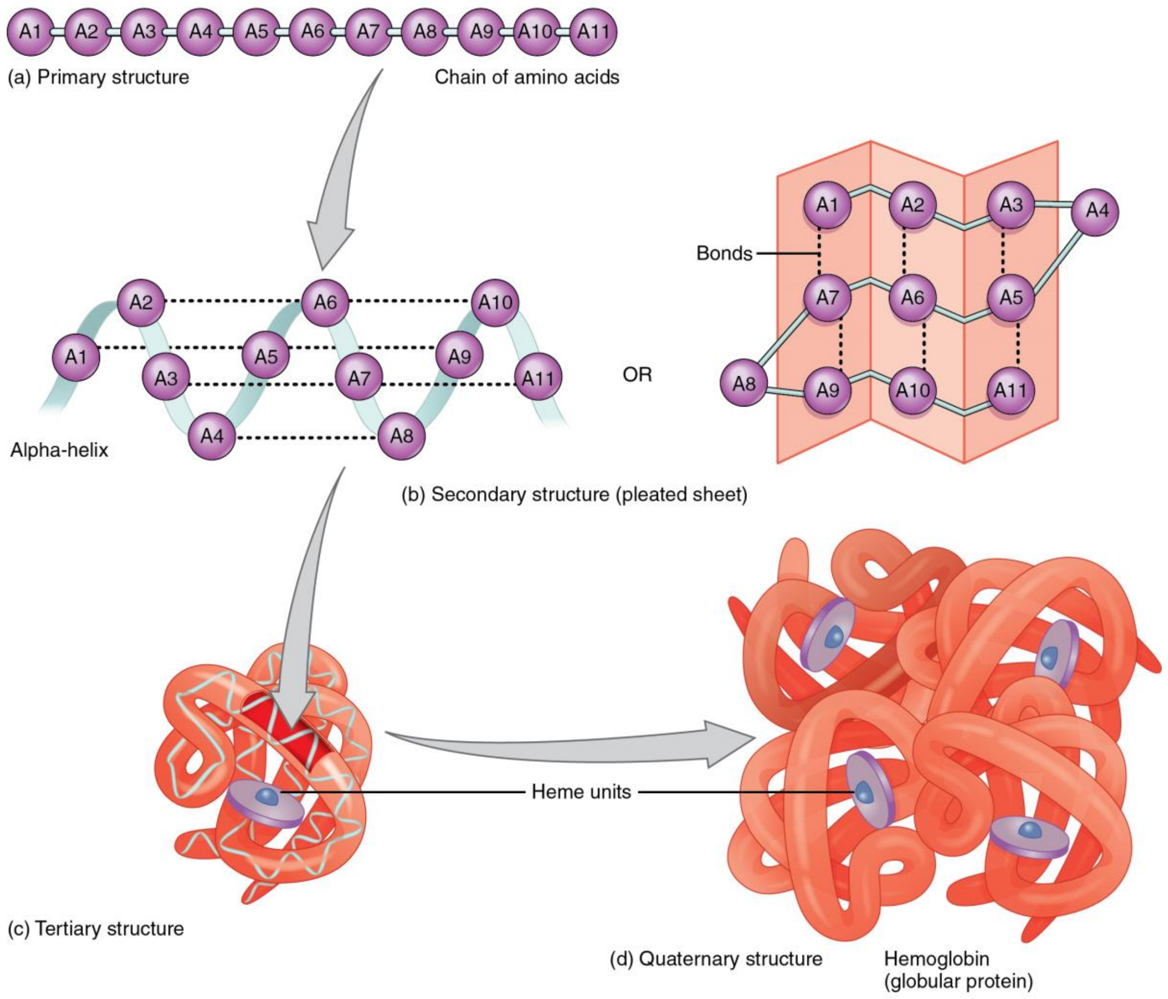
C.elegans
**100MB**
2002

Human
**3000MB**
2007

Many Humans
**~1M * 3000MB**
2023

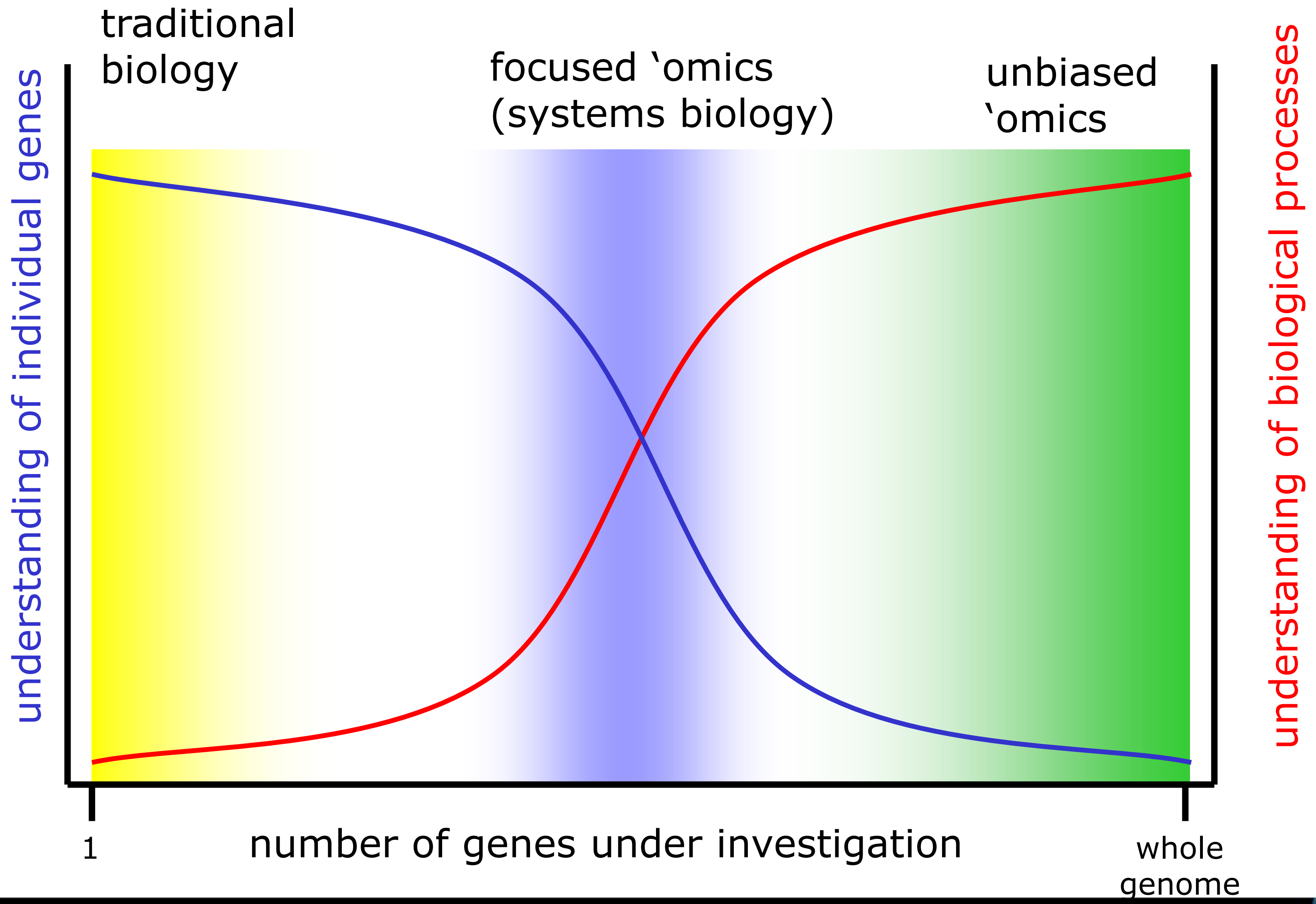


# Understanding structure is the basis of understanding function



Wikimedia Commons https://commons.wikimedia.org/

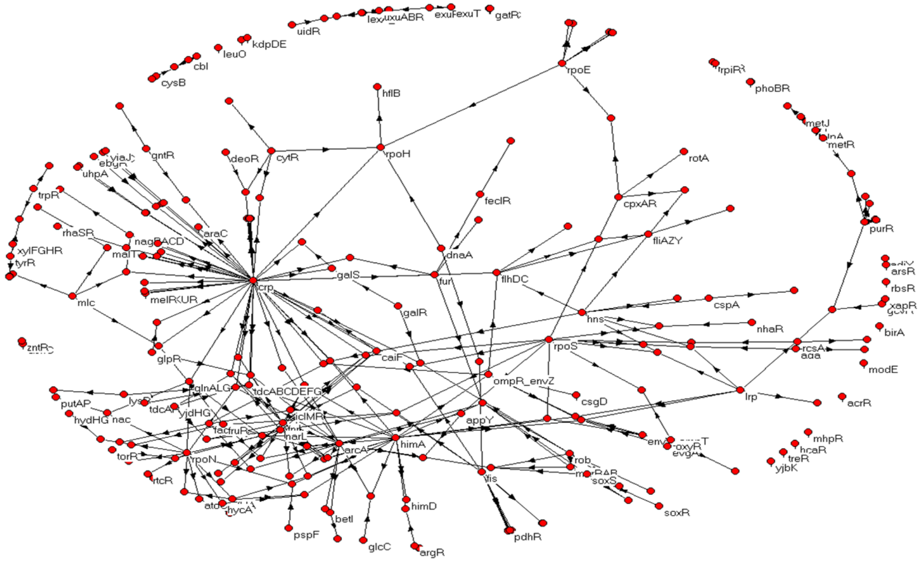
# Why focus?

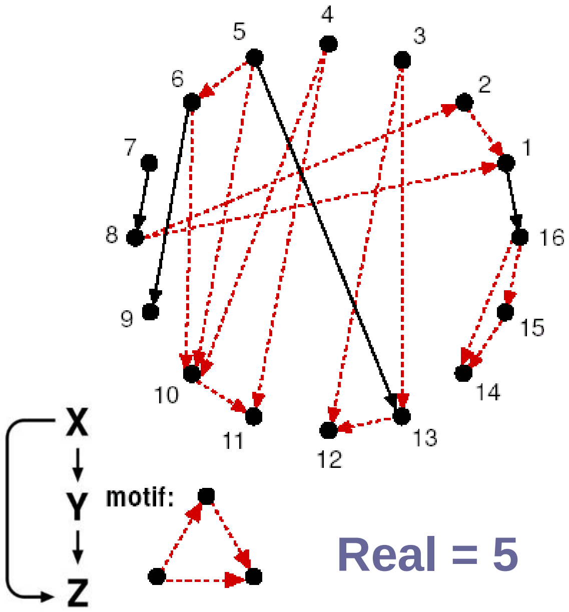


Slide courtesy of Gavin Macbeath

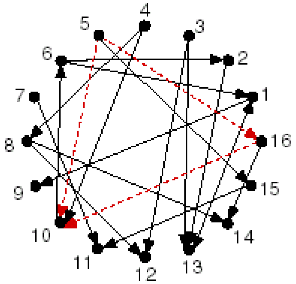
# Where higher order language in biology began for me

## Network motifs as language building blocks

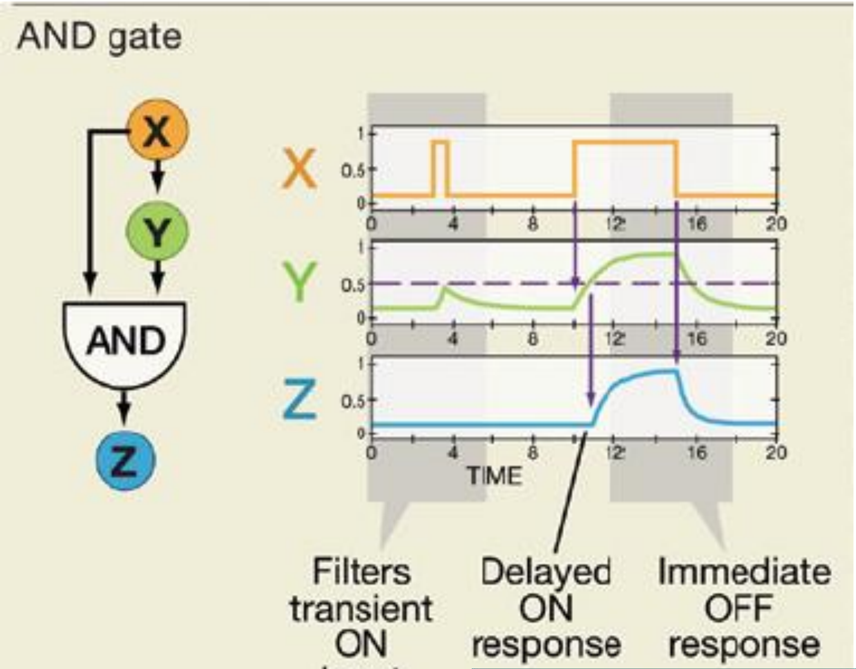




Real = 5



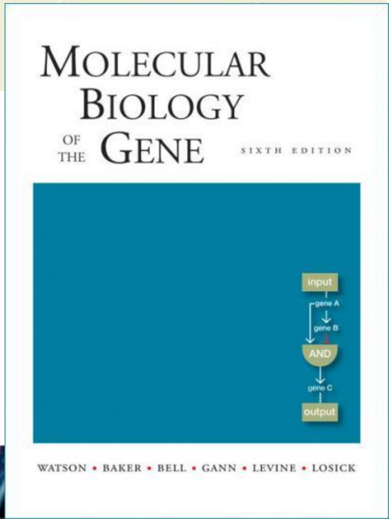
Rand=0.5±0.6



Filters transient ON input signals

Delayed ON response

Immediate OFF response



MOLECULAR BIOLOGY OF THE GENE

SIXTH EDITION

WATSON • BAKER • BELL • GANN • LEVINE • LOSICK

The Feed forward loop is a Network Motif

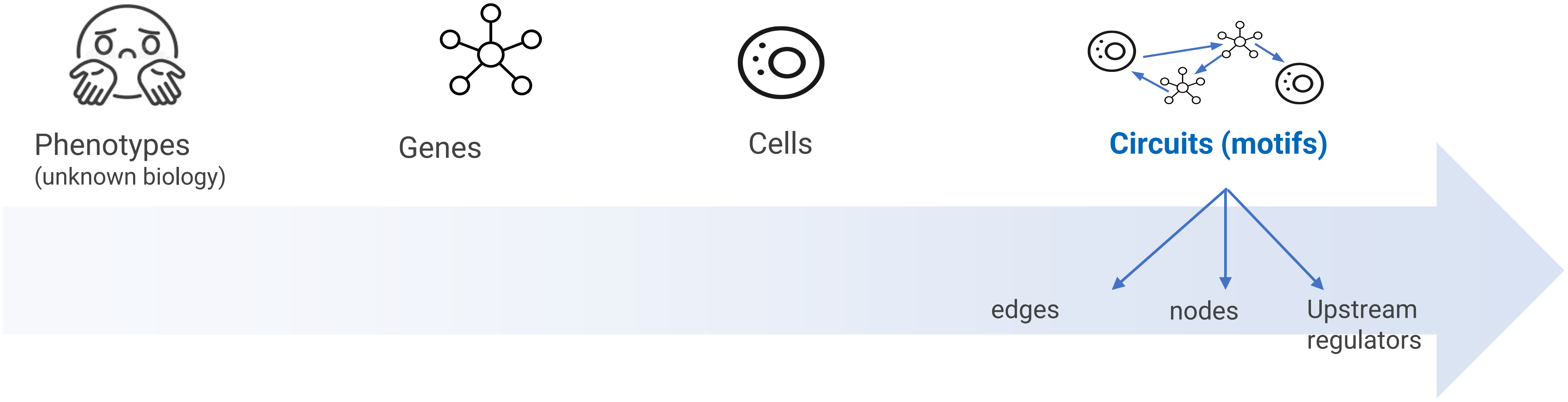
Shen-Orr et al. Nature Genetics 2002

# The immune system is dynamic, complex and highly variable

# Building levels of abstraction in language enables reasoning and manipulation

## Discovering higher-order relations enables thinking of novel treatment paradigms

**Disease target choice**



Phenotypes
(unknown biology)

Genes

Cells

**Circuits (motifs)**

edges

nodes

Upstream regulators

# Low dimensional measurements struggle with capturing complex biology

No longer a relevant analogy

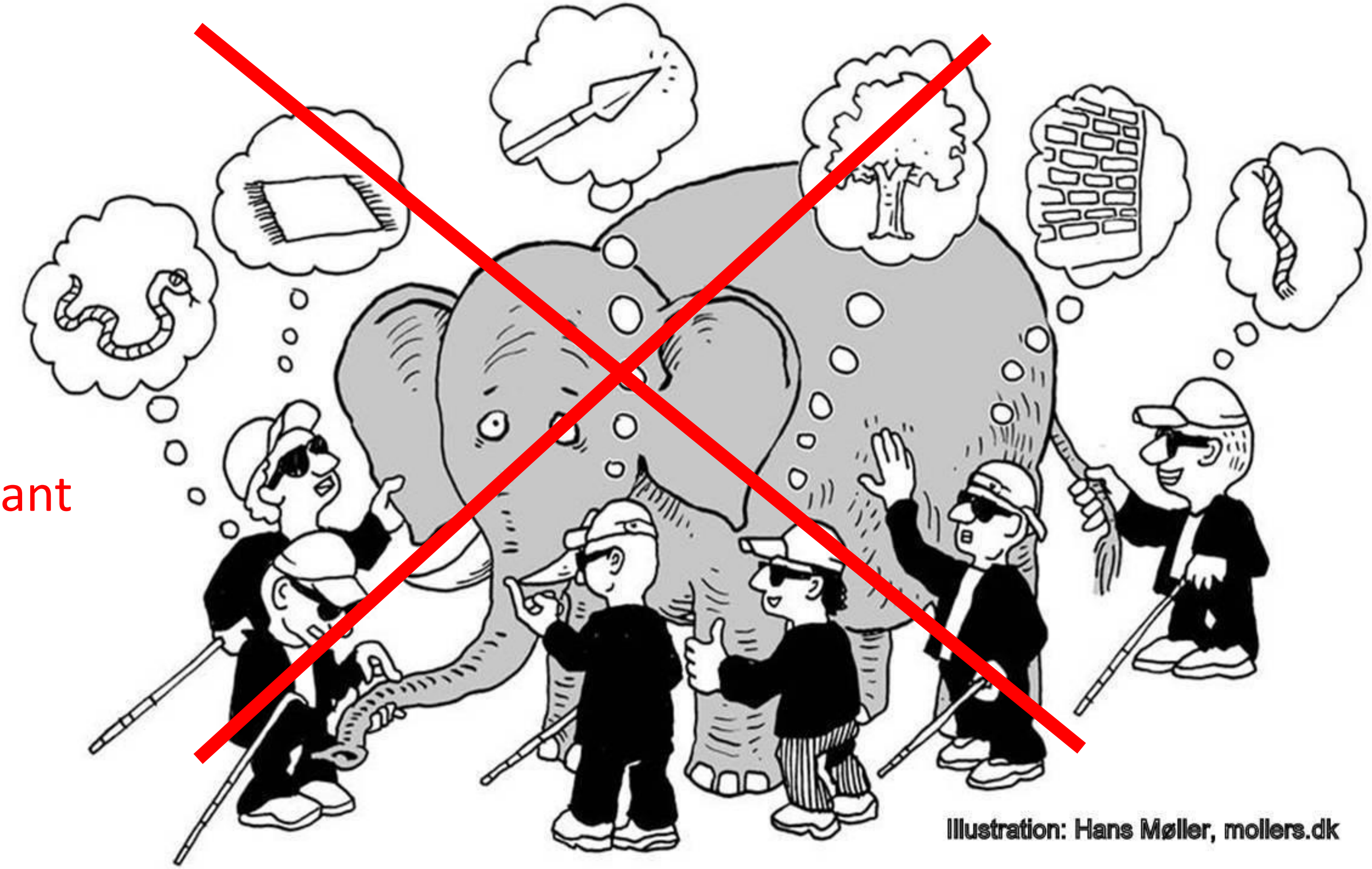
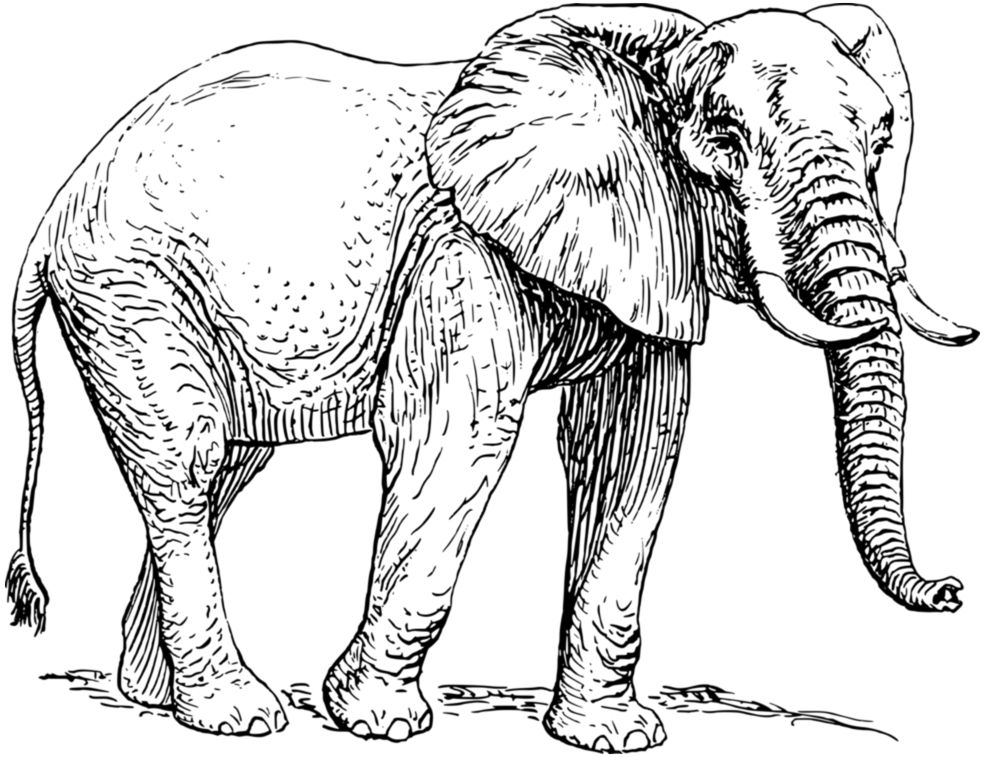


Illustration: Hans Møller, mollers.dk

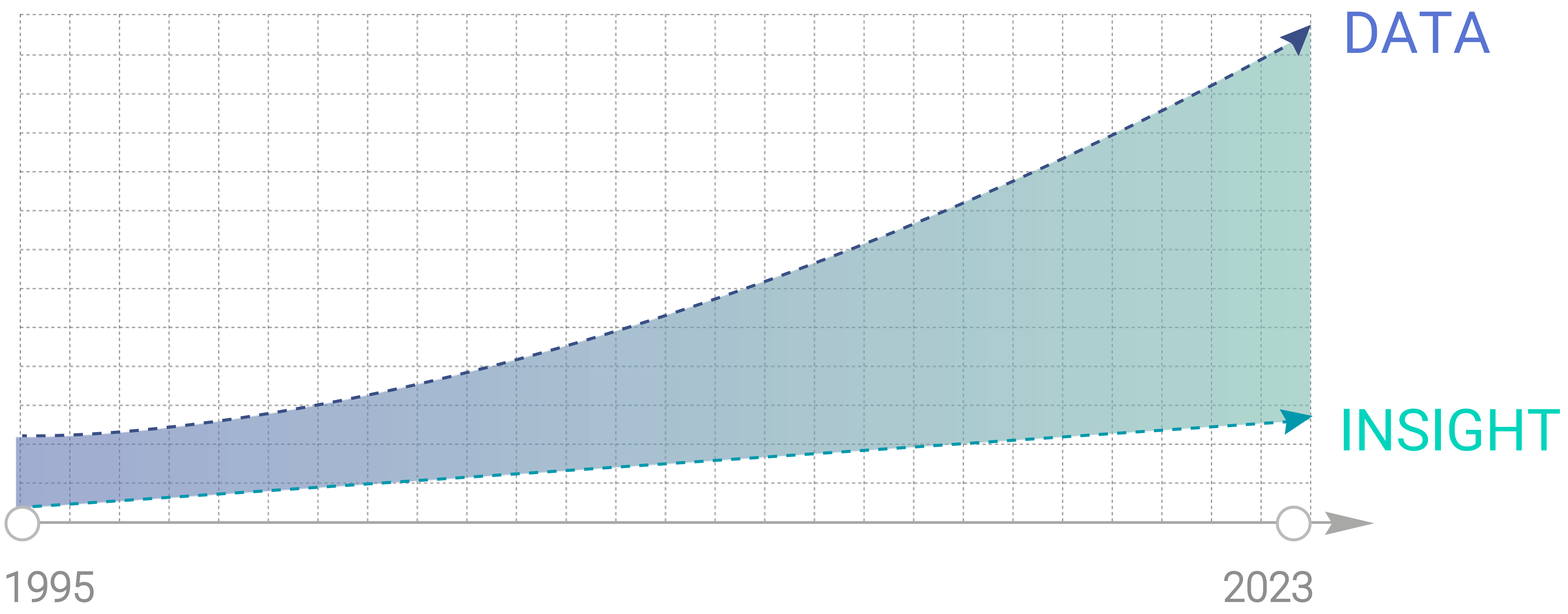
# High dimensional measurements enable studying relations and capture the sum that’s greater than the parts



←

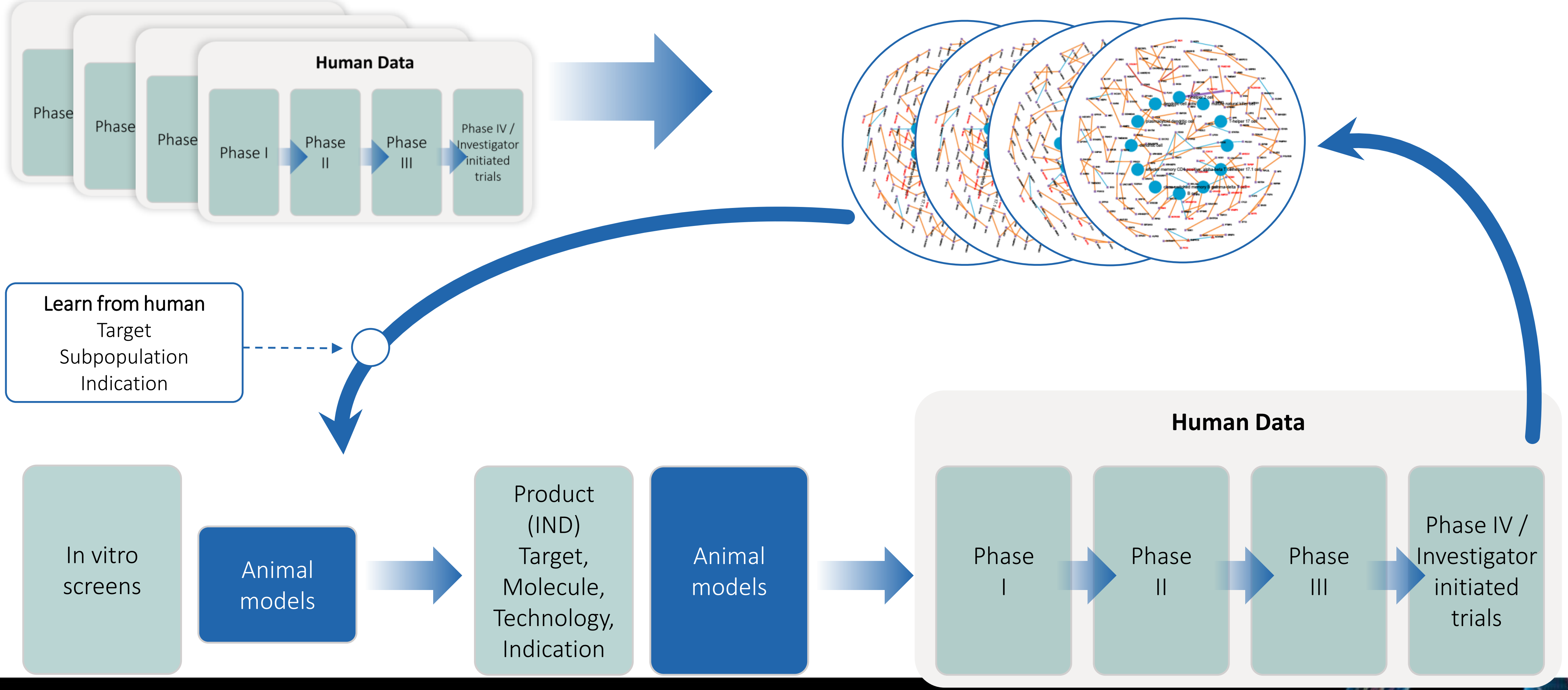
“New School” Scientist Measures All Angles Simultaneously

# The Data-Insight Gap (the scientific problem)



# Bringing human data-driven insight to every decision

Human data collected  from trials across the industry



Human Data

Phase

Phase

Phase

Phase I

Phase II

Phase III

Phase IV / Investigator initiated trials

Learn from human

Target

Subpopulation

Indication

In vitro screens

Animal models

Product (IND) Target, Molecule, Technology, Indication

Animal models

Human Data

Phase I

Phase II

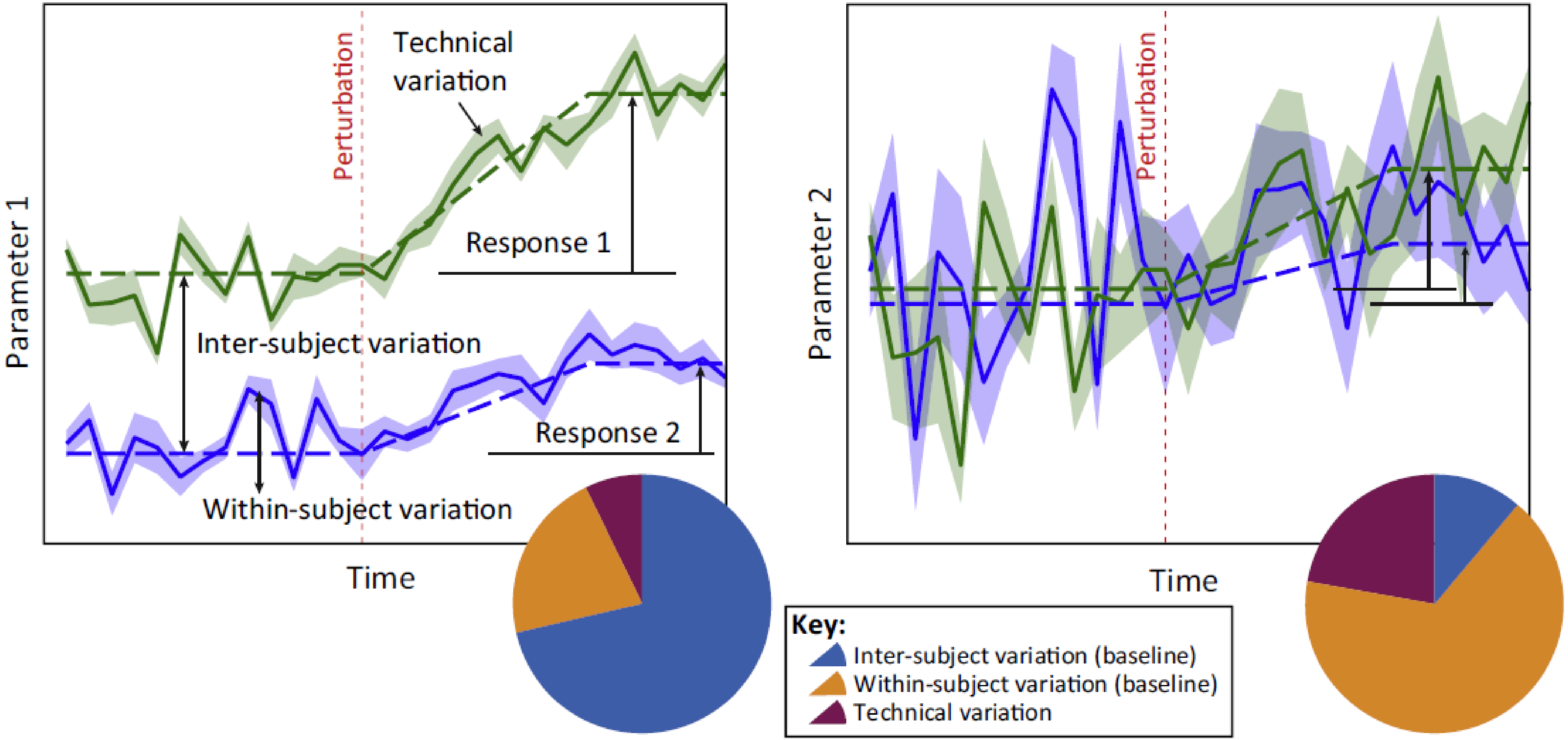
Phase III

Phase IV / Investigator initiated trials

Immune health is an emergent phenomena directly related to physiology

A shift back from reductionism: "The progressive triumphs of physiology over molecular biology" -Sir James Black

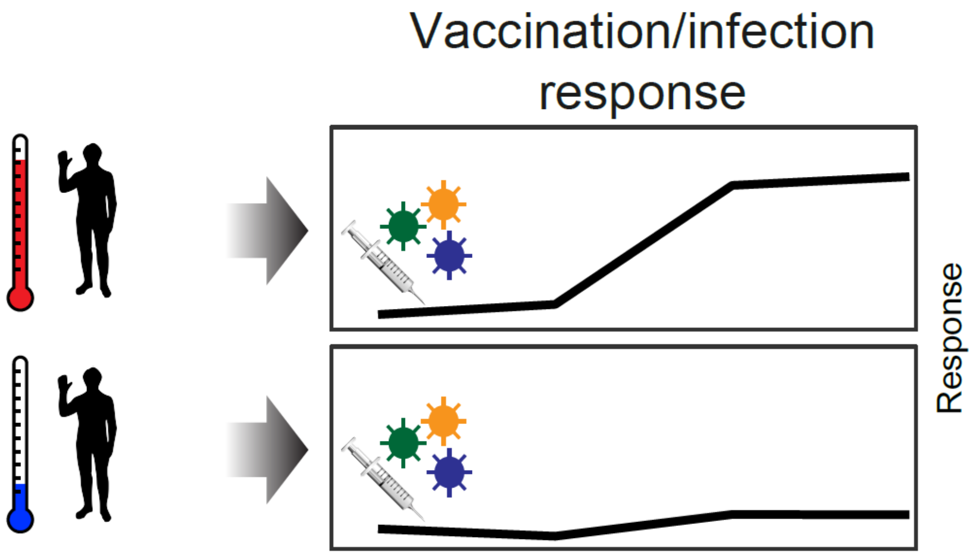
# Understand the origin of variability and design experiments accordingly



Tsang, Trends in Immunology 2015

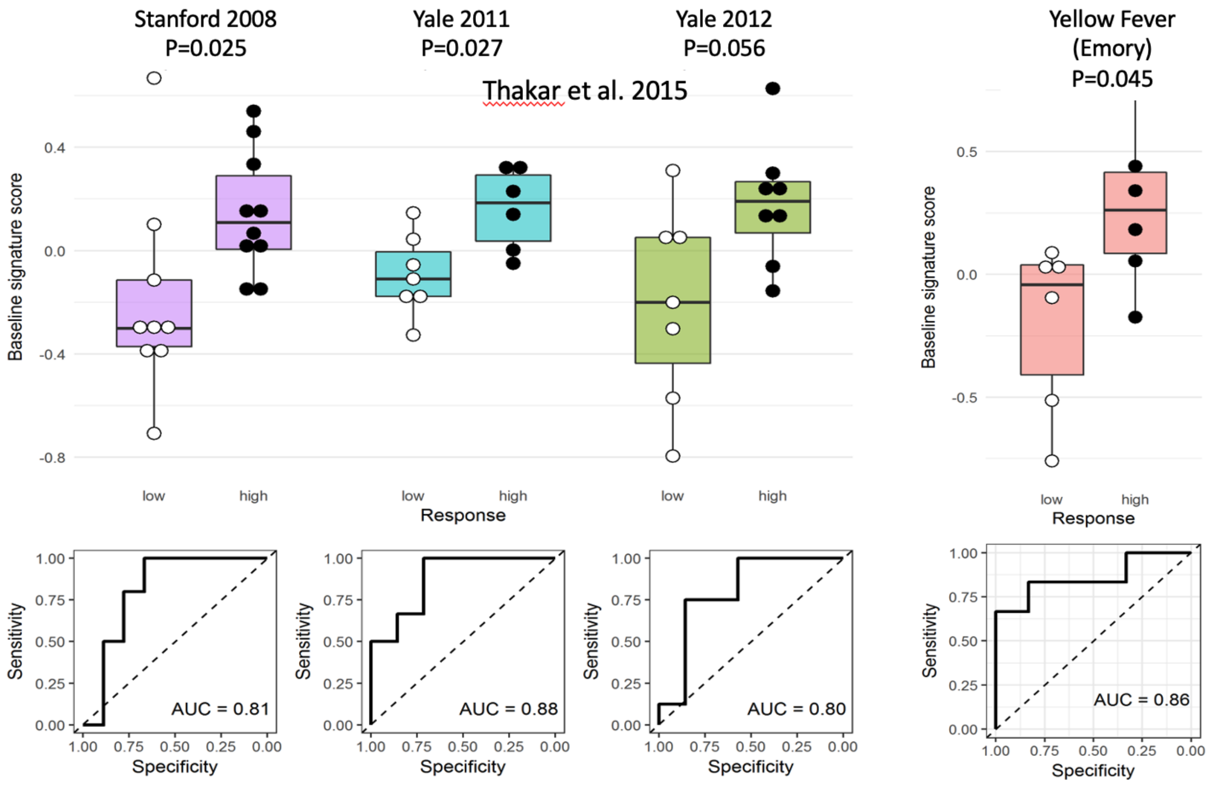
# Personal immune state (setpoint) can predict outcome

### Baseline predictors of influenza vaccine responses



Learned through multi-timepoint analysis
**independent of age, gender, pre-existing immunity**

### Validation in other vaccination cohorts

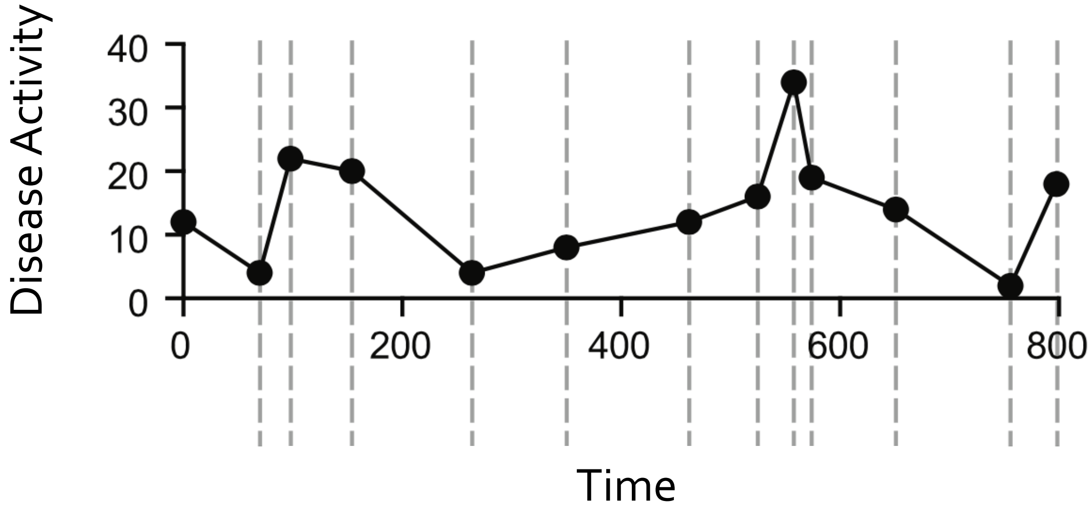


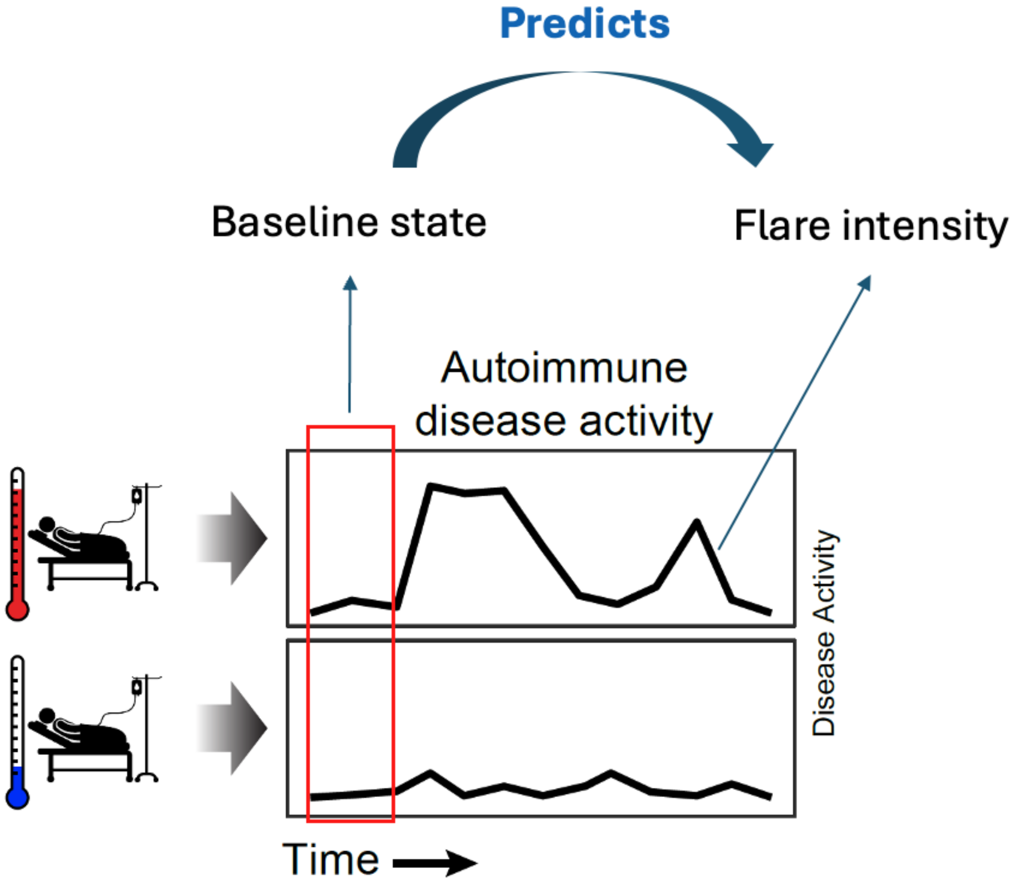
Multiple B & T cell subsets → CD20+CD38++ → 10 gene signature in PBMC

Tsang et al. *Cell* 157(2) 499-513, 2014

Kotliarov, Sparks……Tsang. *Nature Medicine* 2020

# Also predict lupus disease flare in a subset of patients

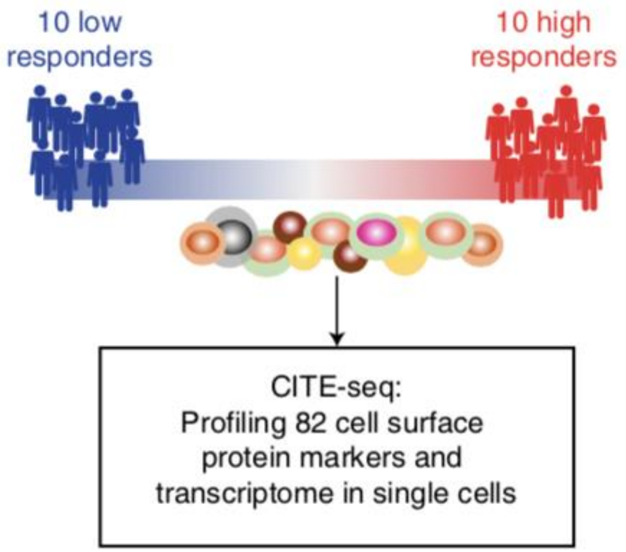




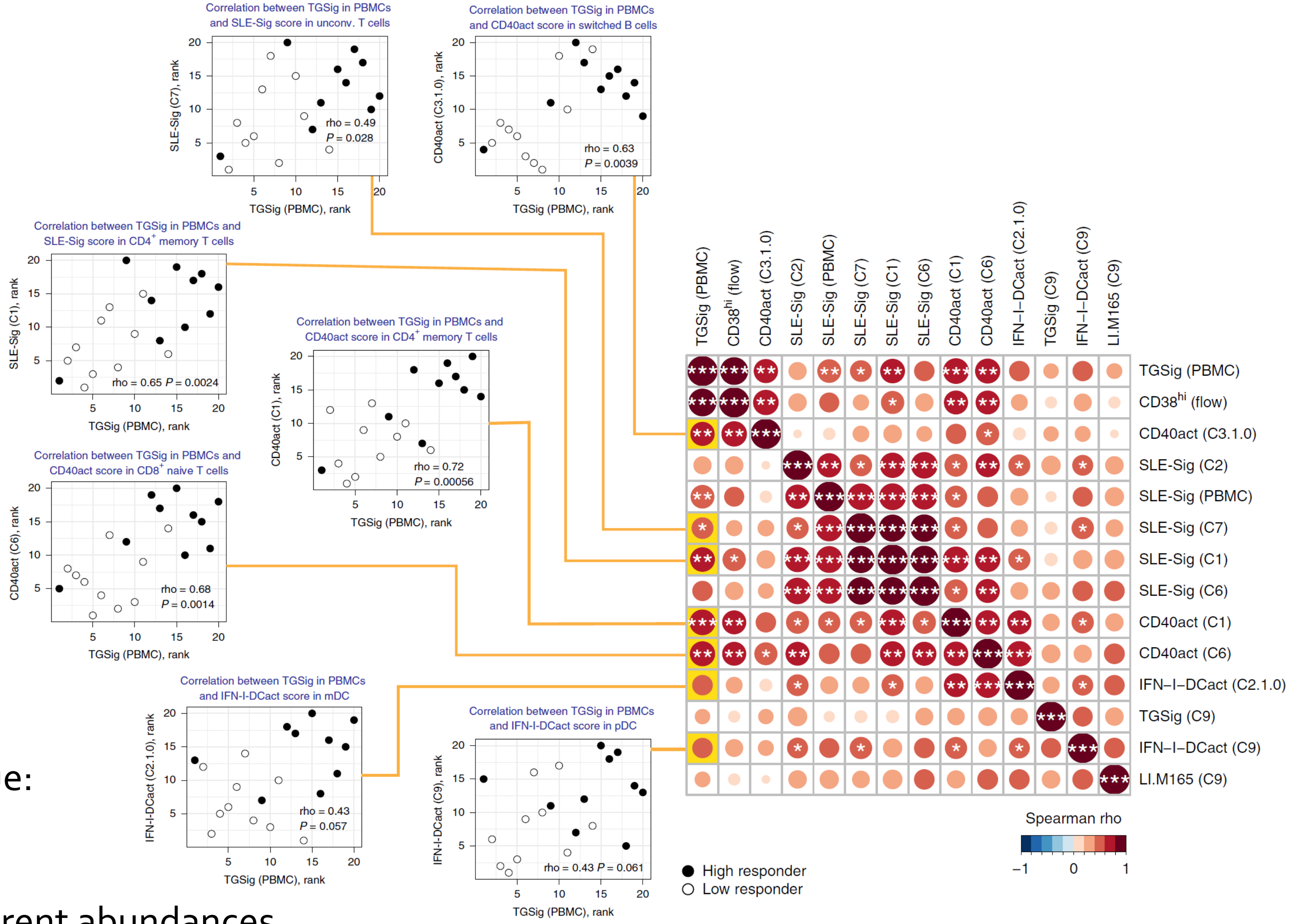
Banchereau et al. *Cell* 2016

Kotliarov, Sparks……Tsang. *Nature Medicine* 2020

# Predictive proxies are highly compressible but may mislead interpretation



CITE-seq:
Profiling 82 cell surface protein markers and transcriptome in single cells



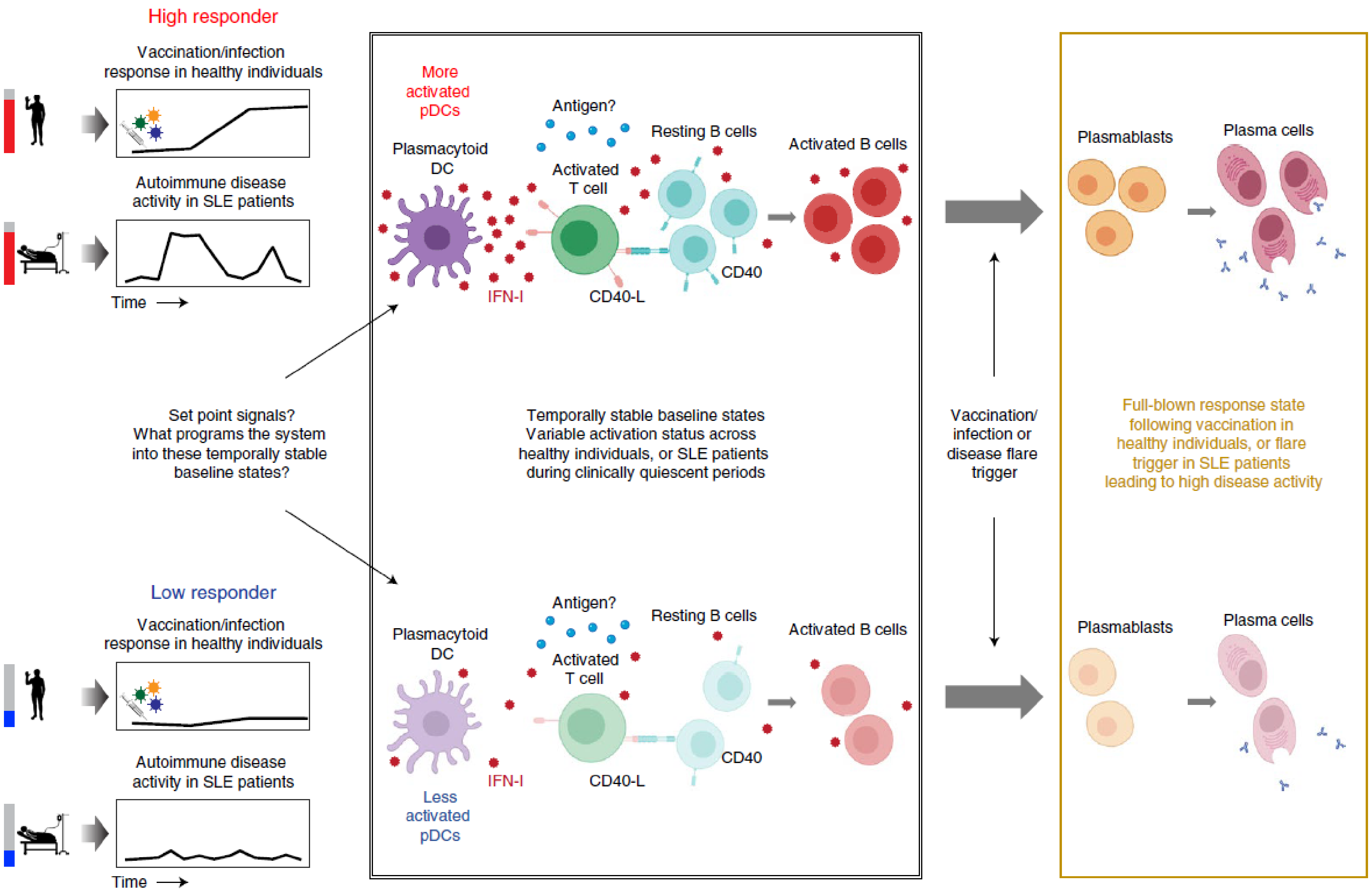
A blood transcriptomic signature origin may be due:

1. cell frequency shifts
2. cell state differences

And may happen in one or more cell-types of different abundances

Kotliarov, Sparks……Tsang. *Nature Medicine* 2020

# A cellular circuit whose “setpoint” determines future response



High responder
Vaccination/infection response in healthy individuals
Autoimmune disease activity in SLE patients
Time
Set point signals? What programs the system into these temporally stable baseline states?
Low responder
Vaccination/infection response in healthy individuals
Autoimmune disease activity in SLE patients
Time
More activated pDCs
Plasmacytoid DC
Antigen?
Activated T cell
Resting B cells
Activated B cells
IFN-I
CD40-L
CD40
Temporally stable baseline states
Variable activation status across healthy individuals, or SLE patients during clinically quiescent periods
Plasmacytoid DC
Antigen?
Activated T cell
Resting B cells
Activated B cells
IFN-I
CD40-L
CD40
Less activated pDCs
Vaccination/ infection or disease flare trigger
Plasmablasts
Plasma cells
Full-blown response state following vaccination in healthy individuals, or flare trigger in SLE patients leading to high disease activity
Plasmablasts
Plasma cells

Kotliarov, Sparks……Tsang. *Nature Medicine* 2020

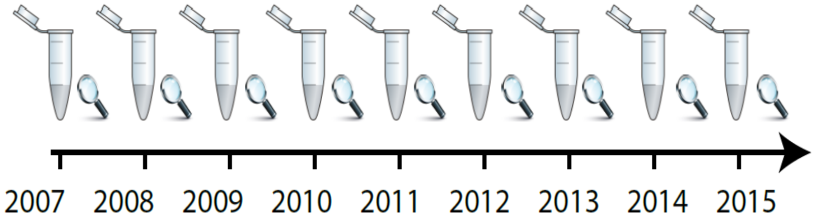
# Immune system dynamics dictate a continuum of setpoints / state shifts

# A longitudinal analysis of immune aging

## Immune-features change over time at rates that differ between individuals

Yearly blood

samples collection



Subjects

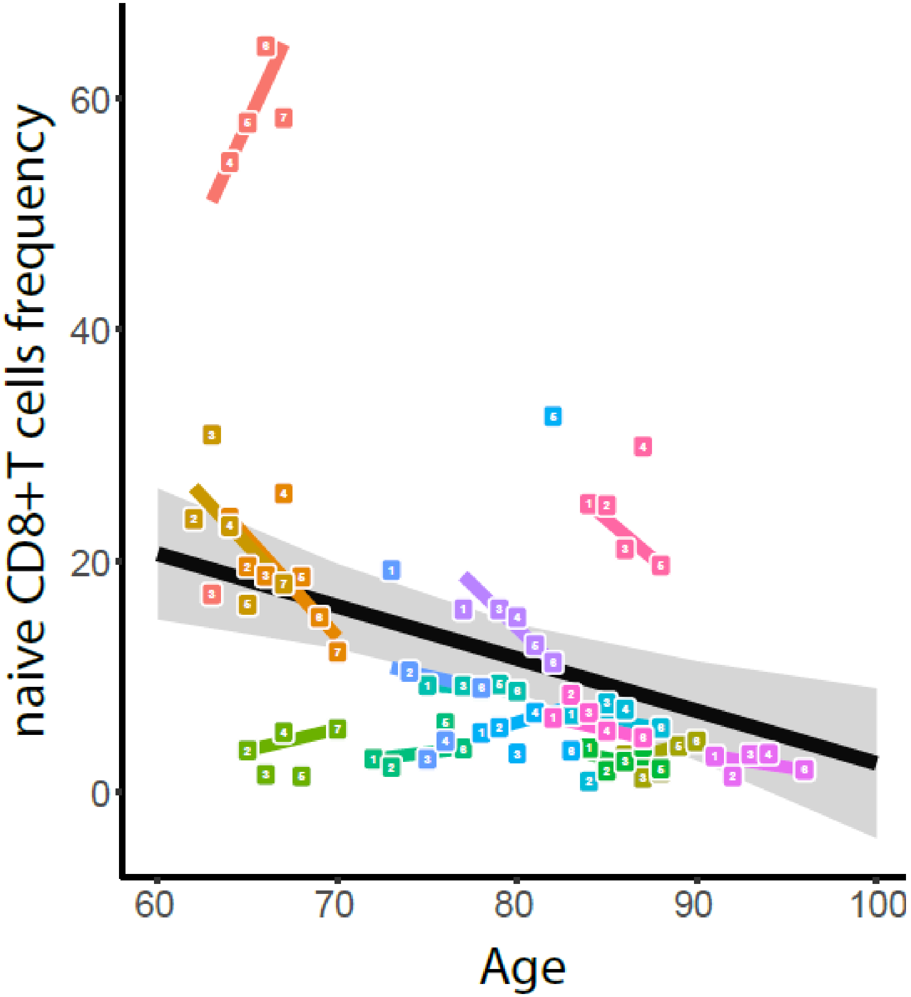
72 older adults (60-96)

63 young adults (20-30)

Healthy on enrollment

Samples (individuals X years)

| Longitudinally changing cell types | 2012 | 2013 | 2014 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| 🔴 | 20% | 24% | 26% | 30% | 32% | 34% |
| 🔵 | 74% | 73% | 78% | 64% | 61% | 57% |
| 🟠 | 13% | 25% | 10% | 5% | 12% | 7% |
| 🟢 | 2% | 1% | 3% | 5% | 8% | 6% |

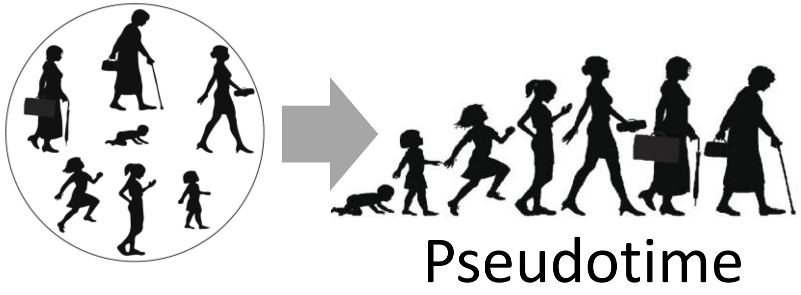


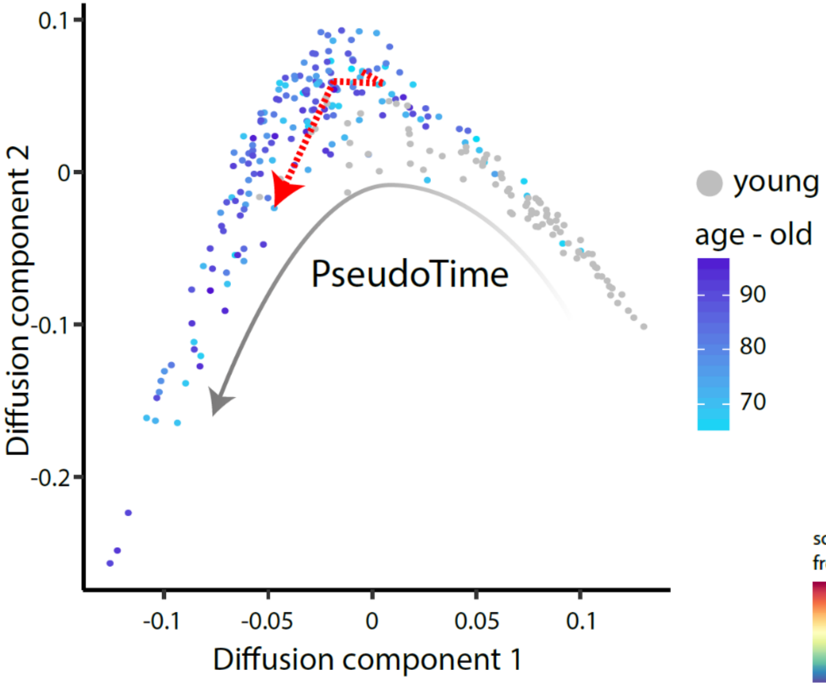
Alpert et al. Nature Med, 2019

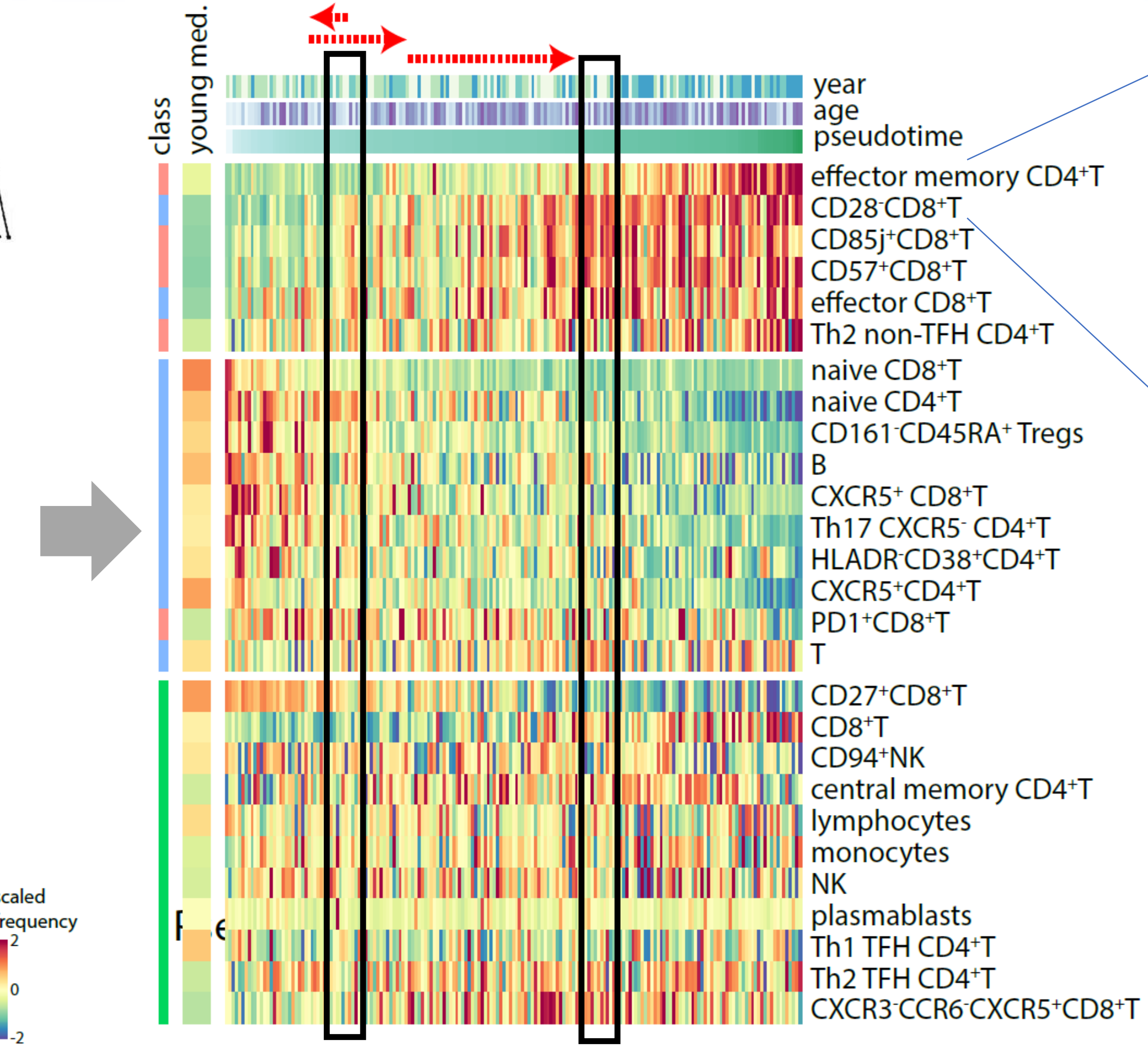
Alpert et al. Nature Med. 2019

# Individuals lie along a trajectory of immune state changes

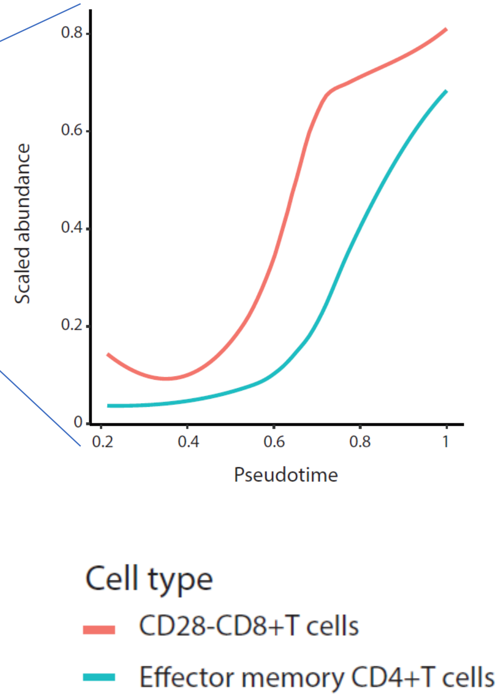
A high-resolution snapshot of the population can be stitched together to approximate long term processes







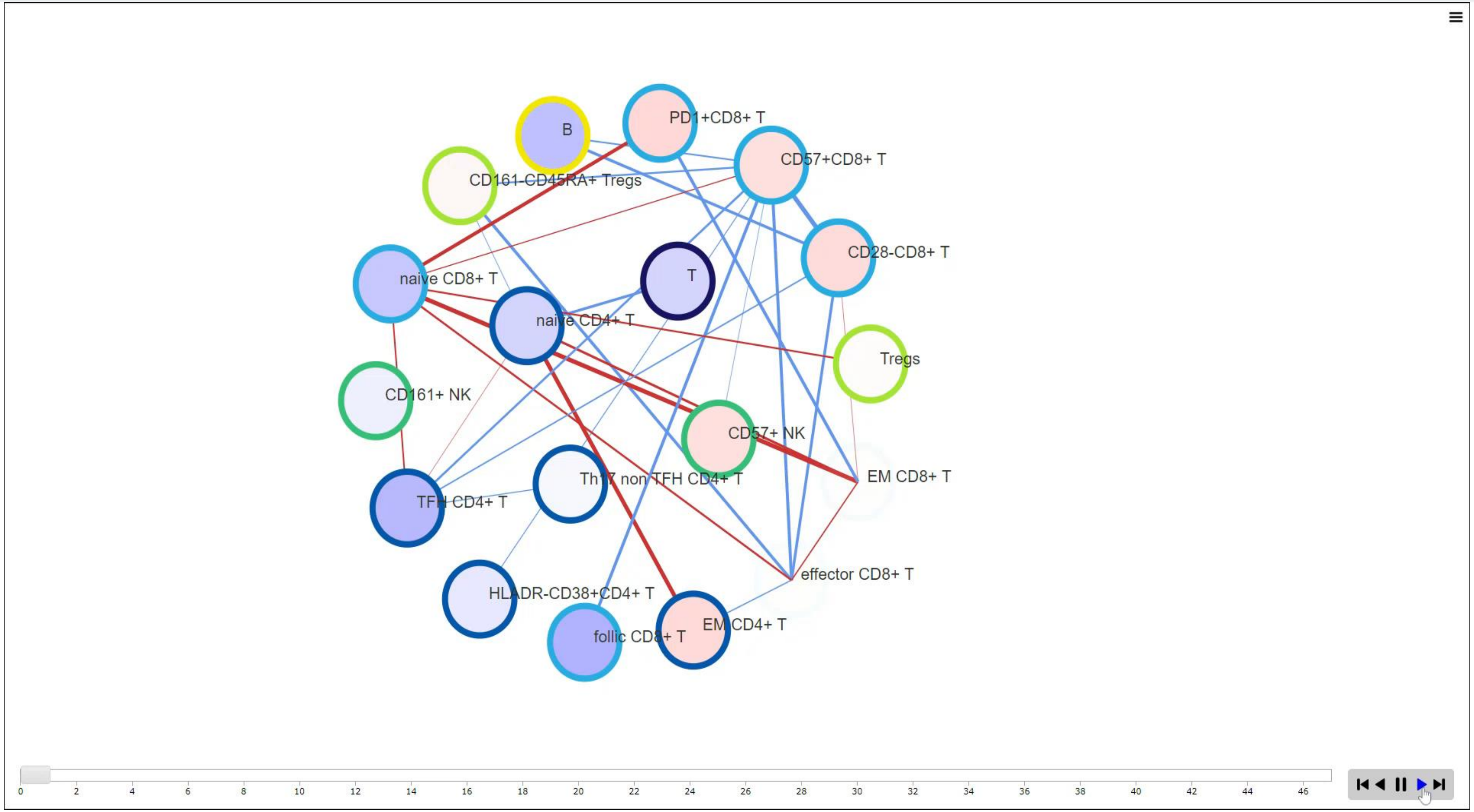
Samples orderd by pseudotime



Alpert et al. Nature Med, 2019

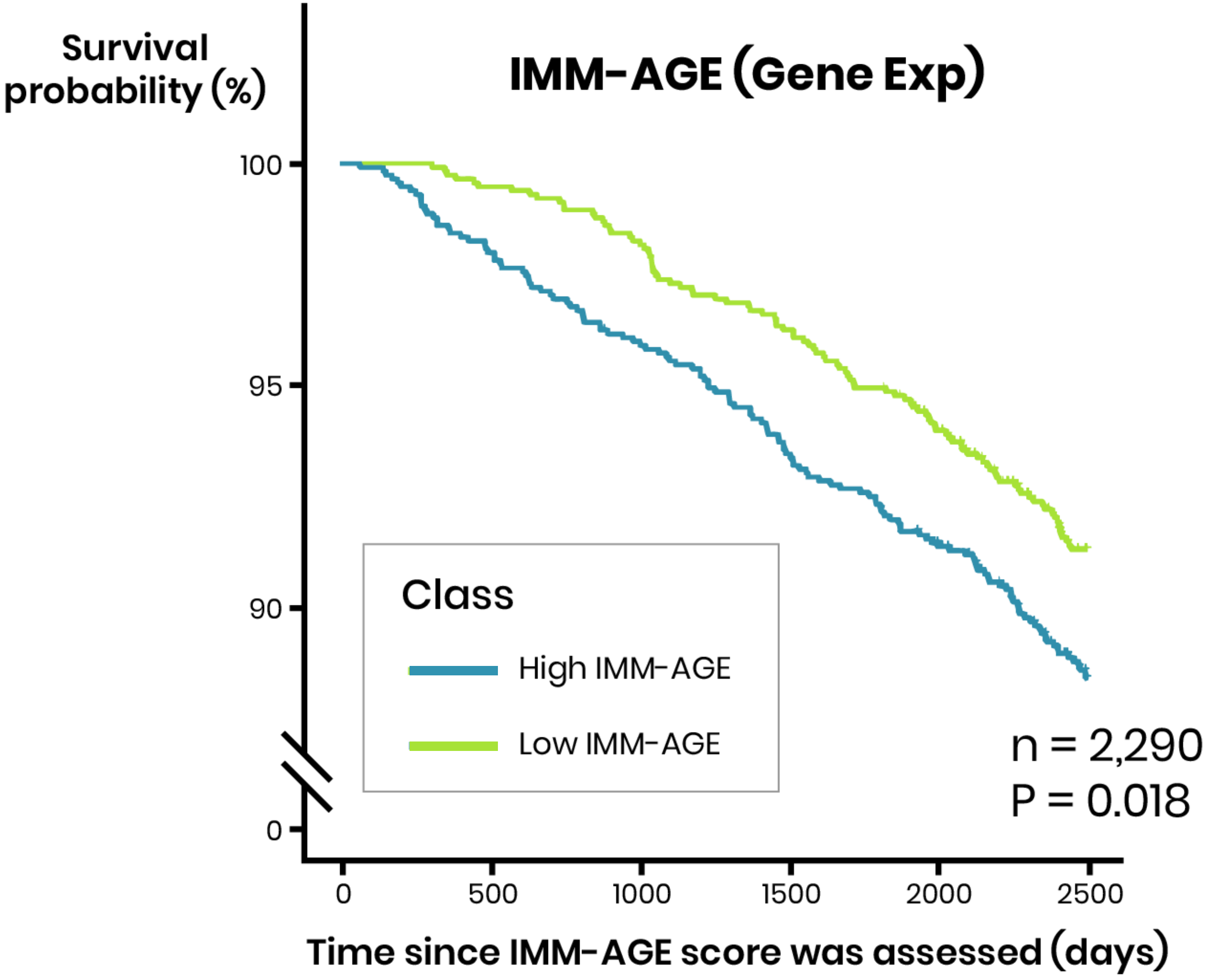
# Under the hood of immune-aging

## Coordinated dynamics among cell-types



Kveler, Peretz, Unpublished

# IMM-AGE predicts mortality beyond standard risk factors



IMM-AGE mass cytometry → Approximated by whole blood gene expression

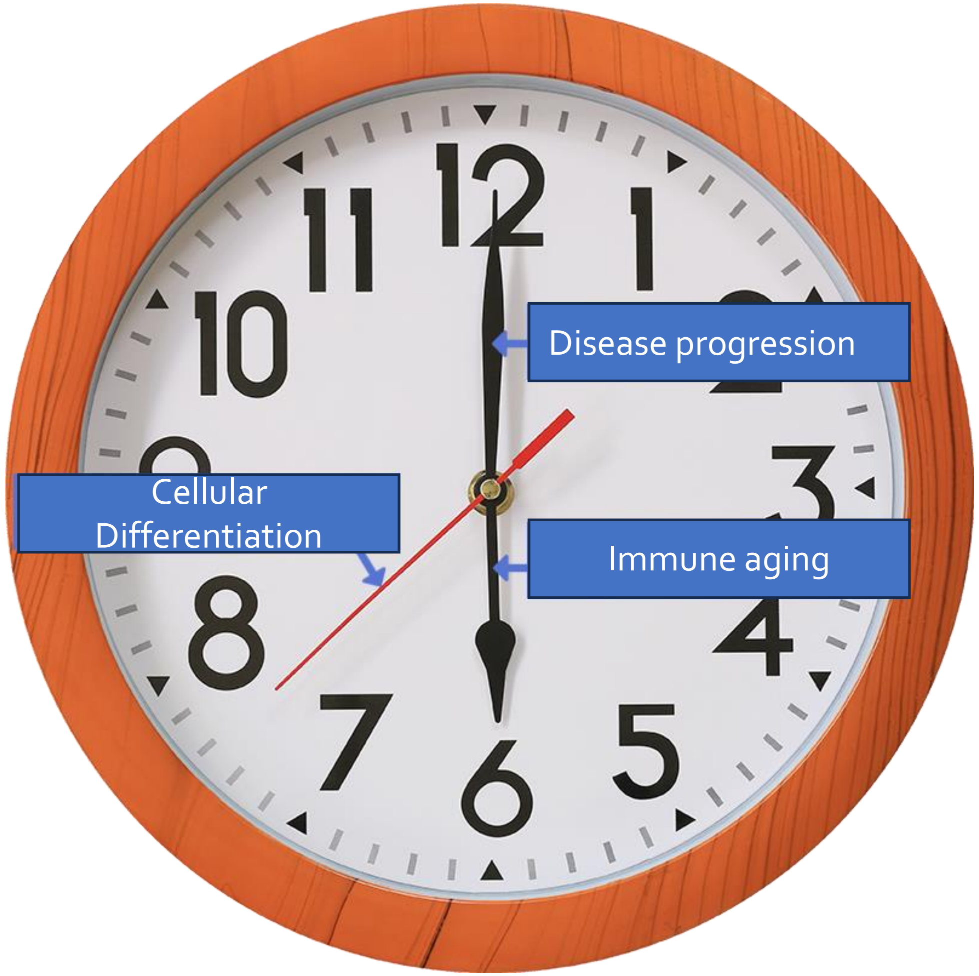
$P<10^{-4}$ for association with survival, by multi-variate Cox adjusted for cardio risk factors and events. Cox regression Hazard ratio = 1.05 per 5-year increment

Model with IMM-AGE versus Methylation Biological Clock:
$P = 8.3 \cdot 10^{-5}$, 0.051 for immune-age and methylation age, respectively

Alpert et al. Nature Med, 2019

# Nothing in immunology makes sense except in the light of time

-paraphrased on T. Dobzhansky



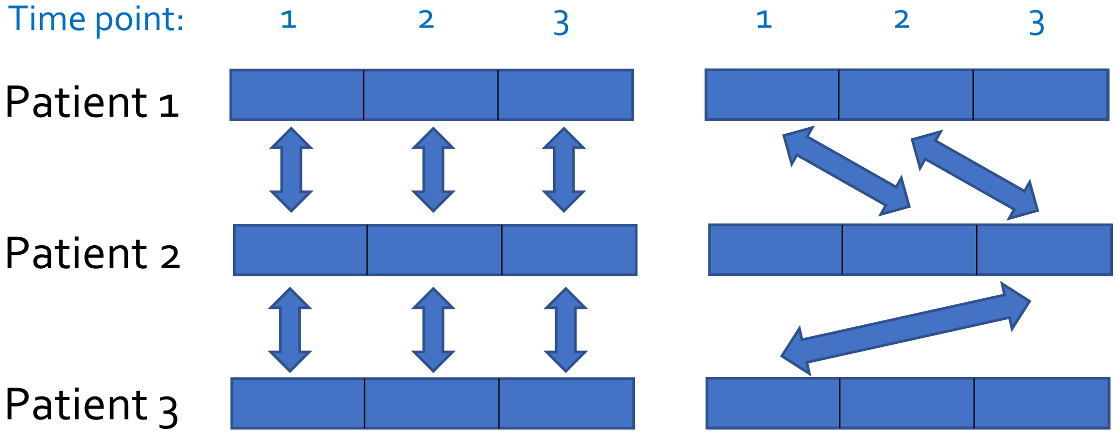
And yet, we struggle with putting time into the equation

- Long experiments
- Typical time scales must be known *apriori*
- Miss intermediate, short-lived states.
- Biological material usually gets destroyed

# Individuals rates vary, alignment may be the way forward

How we usually analyze temporal data...

True biology



Time point: 1 2 3 1 2 3

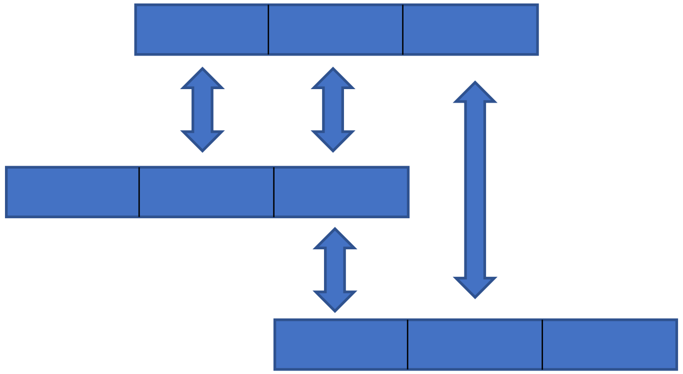
Patient 1

Patient 2

Patient 3

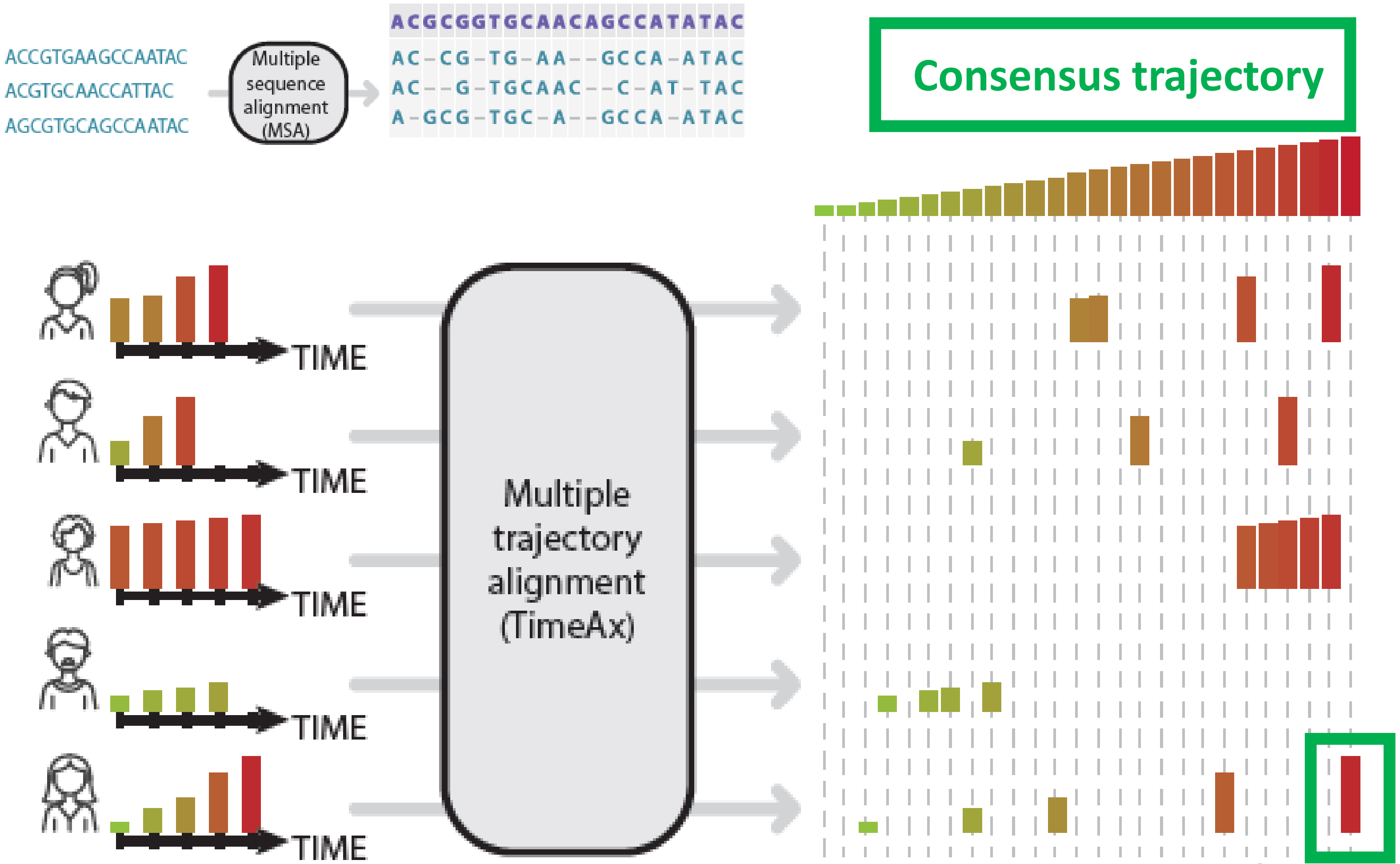
**But disease progression is not the same across patients**

Multiple trajectory alignment

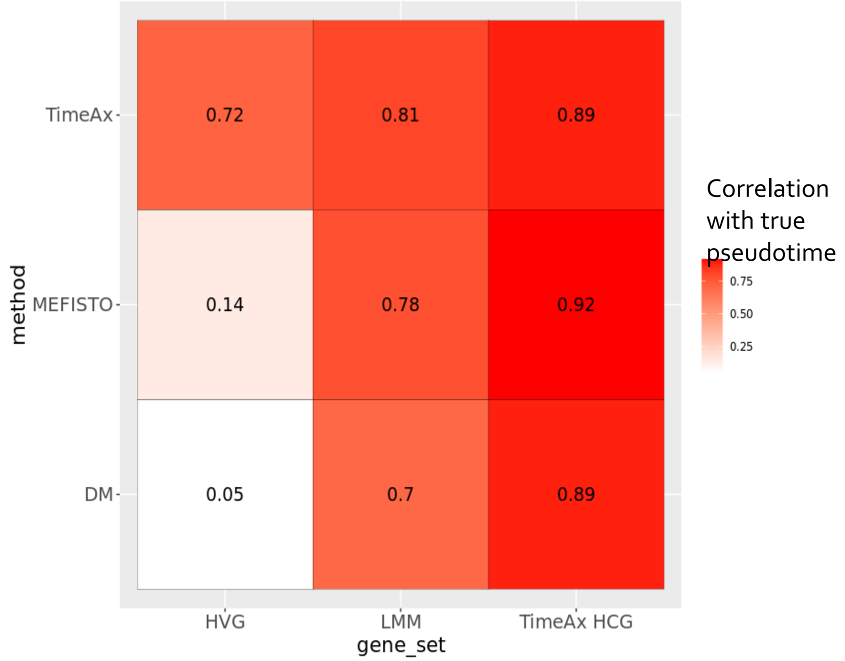


**Consensus trajectory**

# TimeAx does multiple trajectory alignment for high-resolution dynamics



**Pseudotime position**

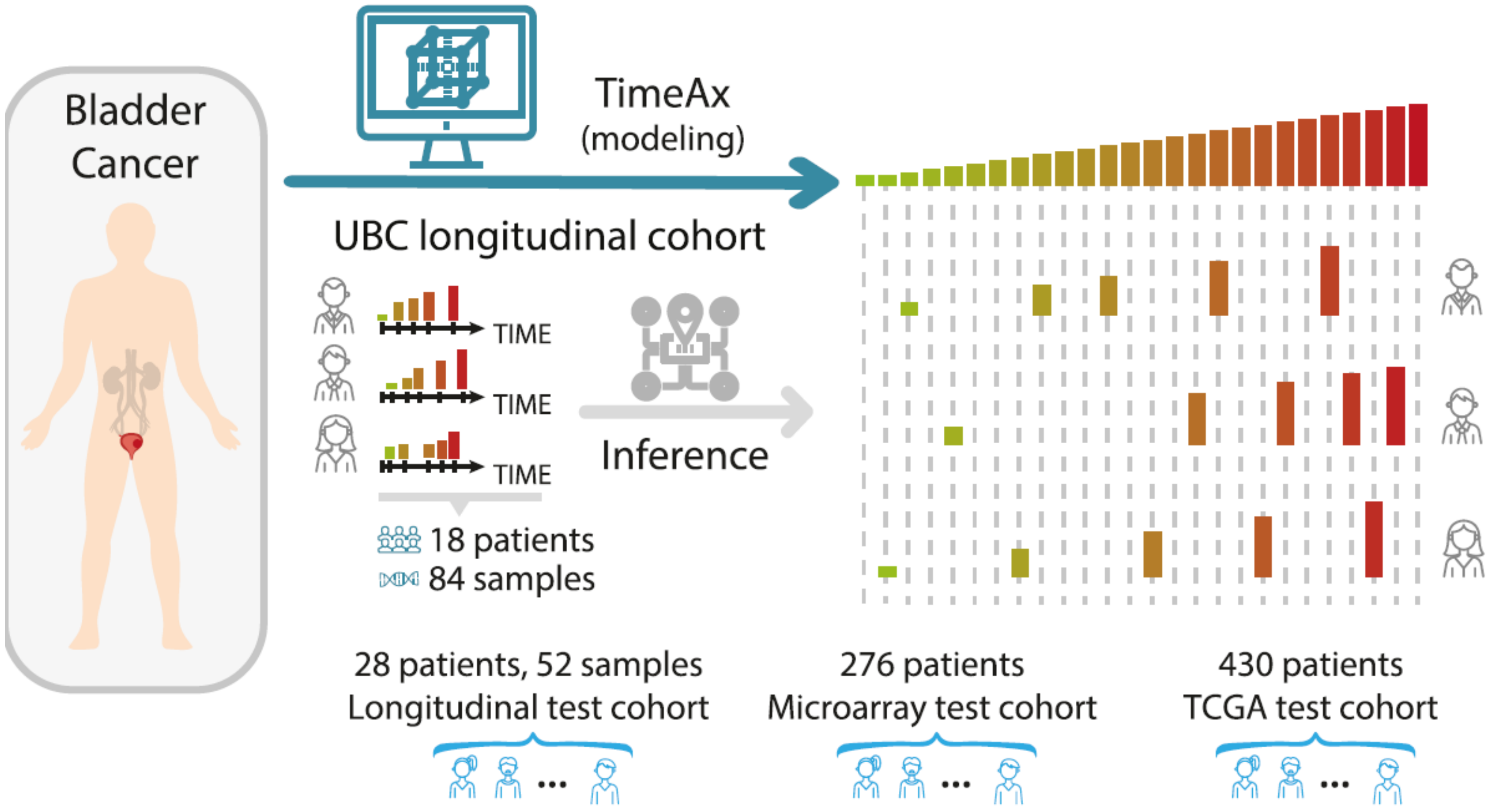


HVG - highly variable genes
LMM - linear mixed model
HCG - highly consistent genes

*Frishberg, Milman et al. Nature Comm. 2023*

Collaboration with F. Theis, J. Schultze

# Disease pseudotime captures progression dynamics better than chronological time



Bladder Cancer

TimeAx (modeling)

UBC longitudinal cohort

TIME

TIME

TIME

18 patients

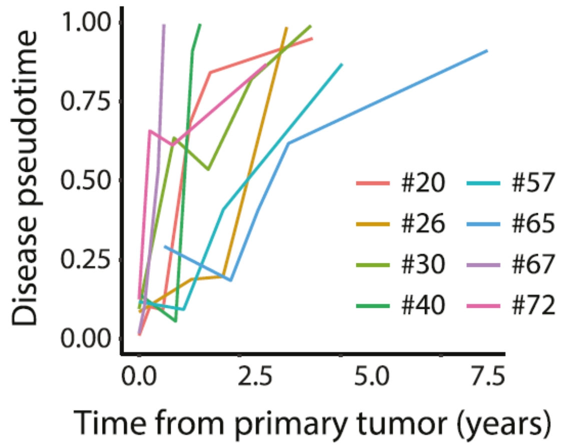
84 samples

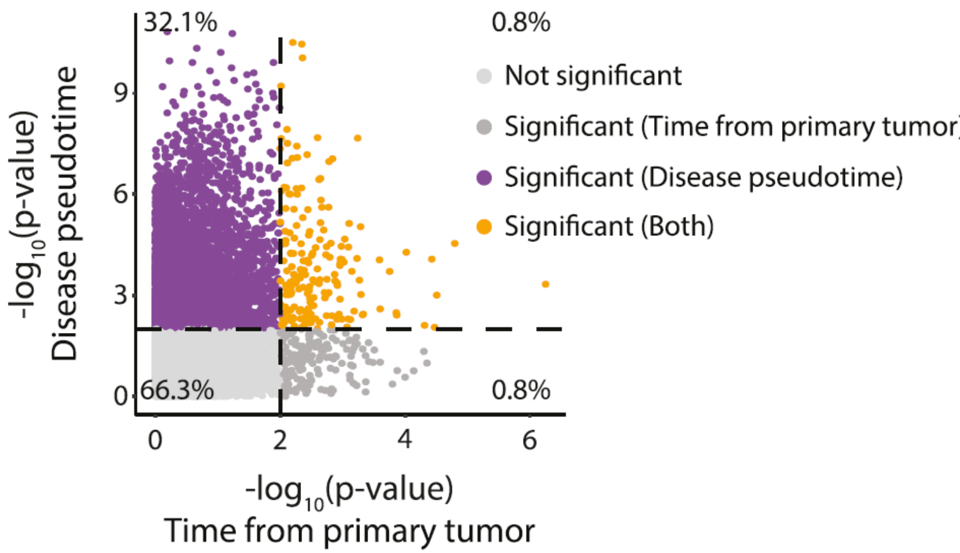
Inference

28 patients, 52 samples
Longitudinal test cohort

276 patients
Microarray test cohort

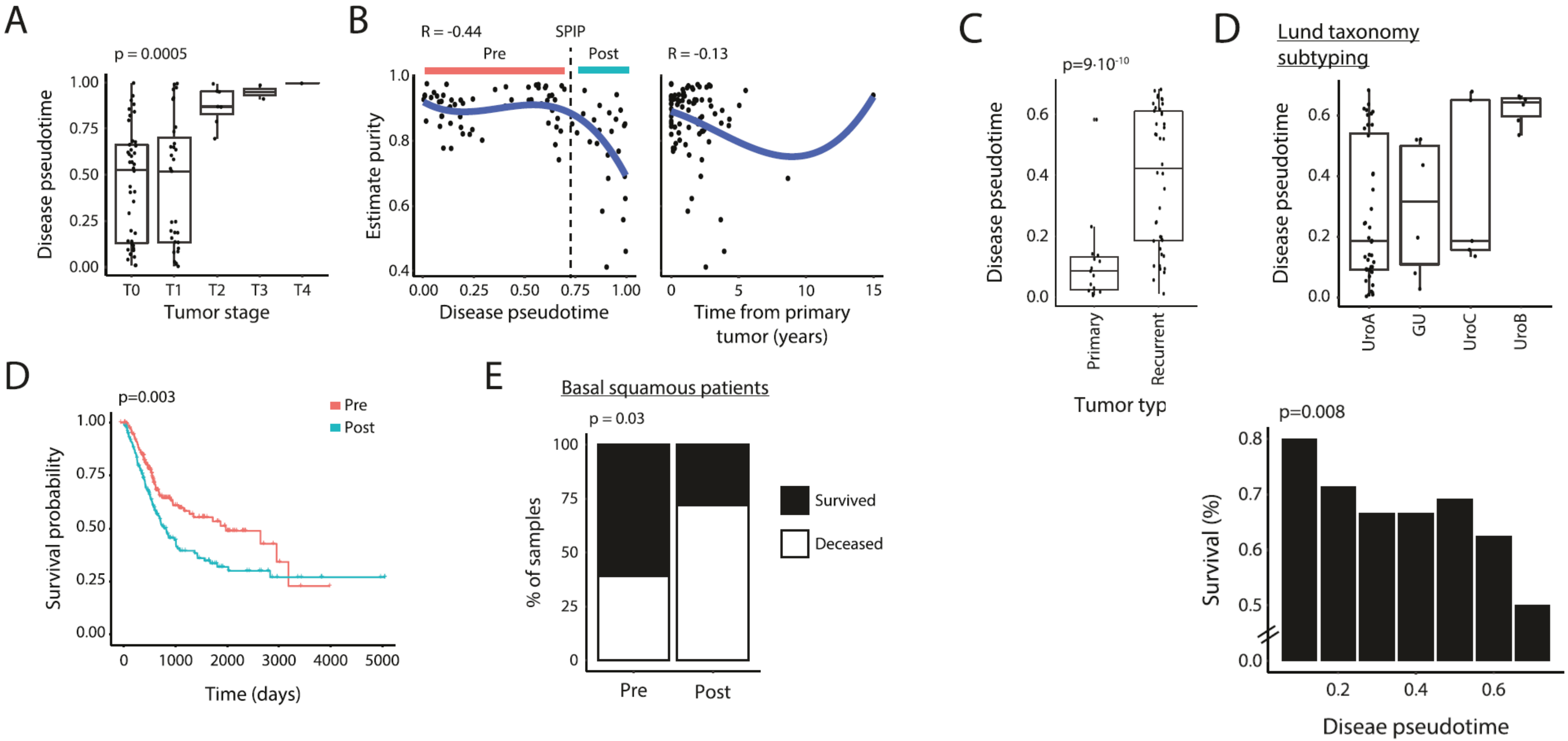
430 patients
TCGA test cohort





*Frishberg, Milman et al. Nature Comm. 2023*

# Disease pseudotime captures variation undetectable by current clinical stratification frameworks

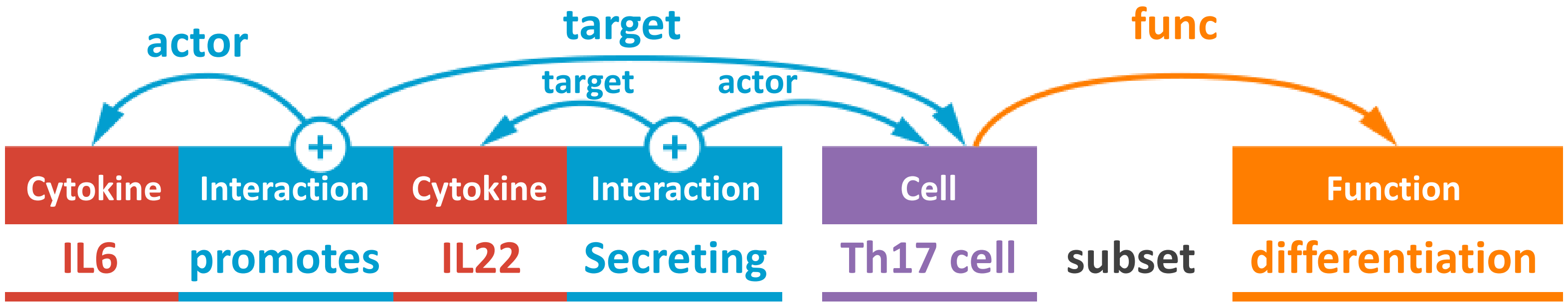


*Frishberg, Milman et al. Nature Comm. 2023*

# Cells are the atomic unit of immune network motifs

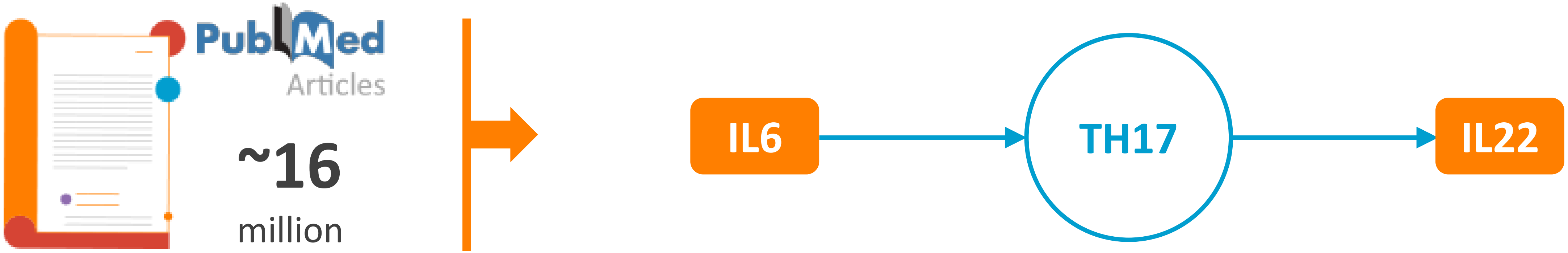
# From literature to machine-readable inter-cellular knowledgebase

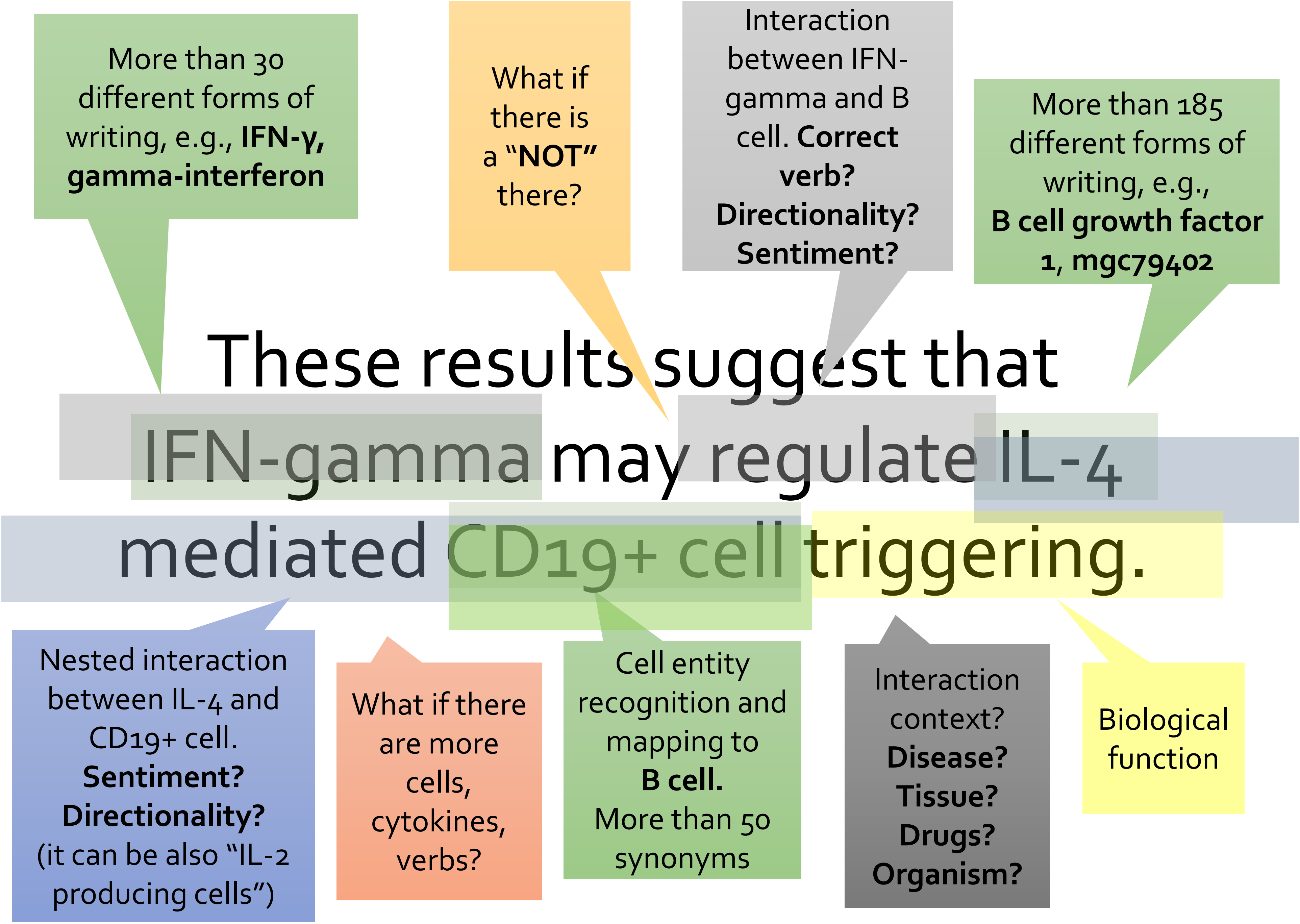
**Immune specific text mining engine optimized for cytokines, immune cells and immune response mapping**



These results suggest that IL6 promotes IL22 Secreting Th17 cell subset differentiation

**Disease Specific Literature Network**





More than 30 different forms of writing, e.g., **IFN-γ, gamma-interferon**

What if there is a “**NOT**” there?

Interaction between IFN-gamma and B cell. **Correct verb? Directionality? Sentiment?**

More than 185 different forms of writing, e.g., **B cell growth factor 1, mgc79402**

These results suggest that IFN-gamma may regulate IL-4 mediated CD19+ cell triggering.

Nested interaction between IL-4 and CD19+ cell. **Sentiment? Directionality?** (it can be also “IL-2 producing cells”)

What if there are more cells, cytokines, verbs?

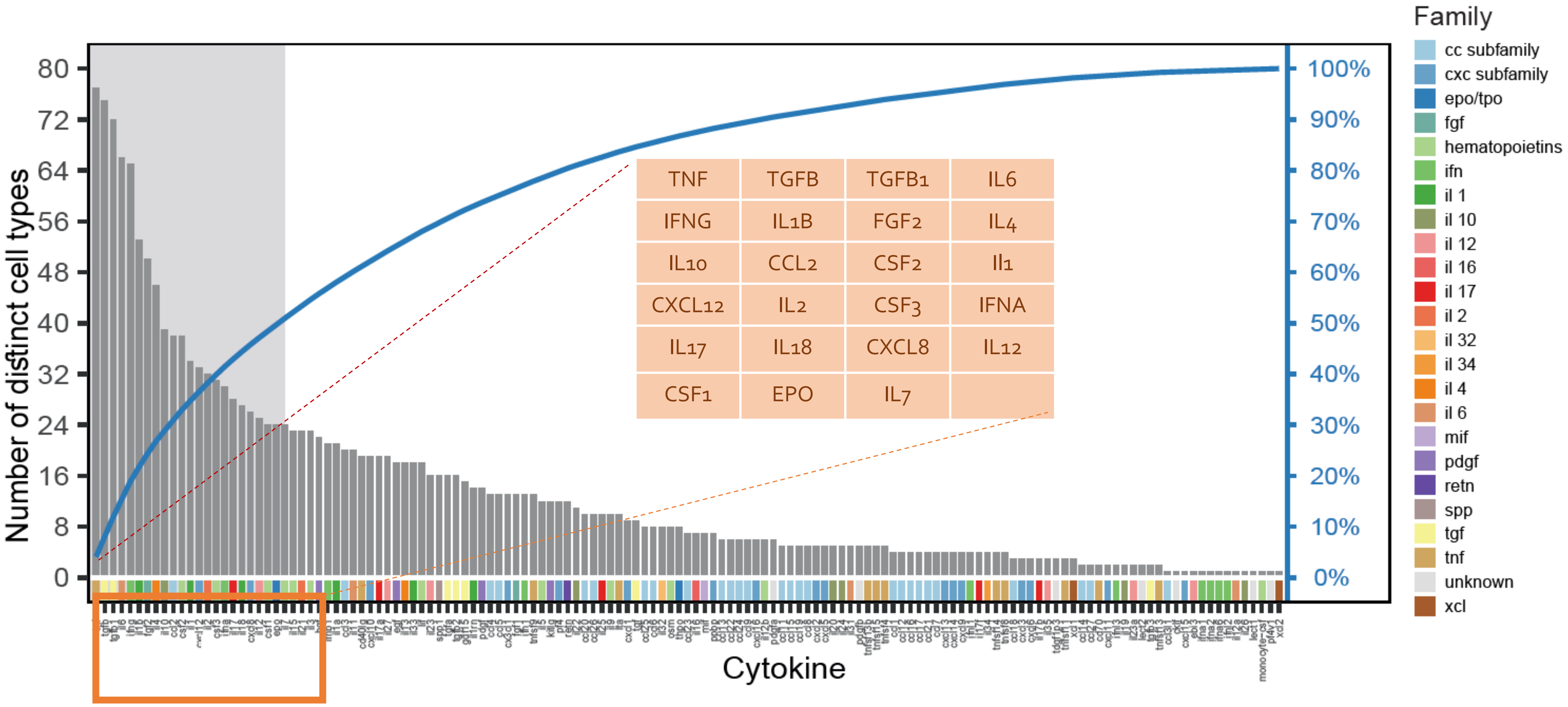
Cell entity recognition and mapping to **B cell.** More than 50 synonyms

Interaction context? **Disease? Tissue? Drugs? Organism?**

Biological function

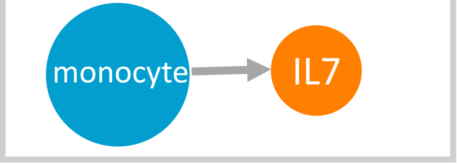
# 23 cytokines account for 50% of human knowledge

## Data model enables prediction of novel insights

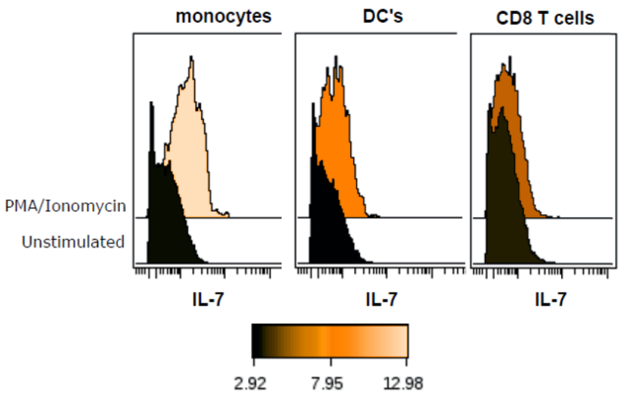


| | | | |
|---|---|---|---|
| TNF | TGFB | TGFB1 | IL6 |
| IFNG | IL1B | FGF2 | IL4 |
| IL10 | CCL2 | CSF2 | Il1 |
| CXCL12 | IL2 | CSF3 | IFNA |
| IL17 | IL18 | CXCL8 | IL12 |
| CSF1 | EPO | IL7 | |

### Prediction

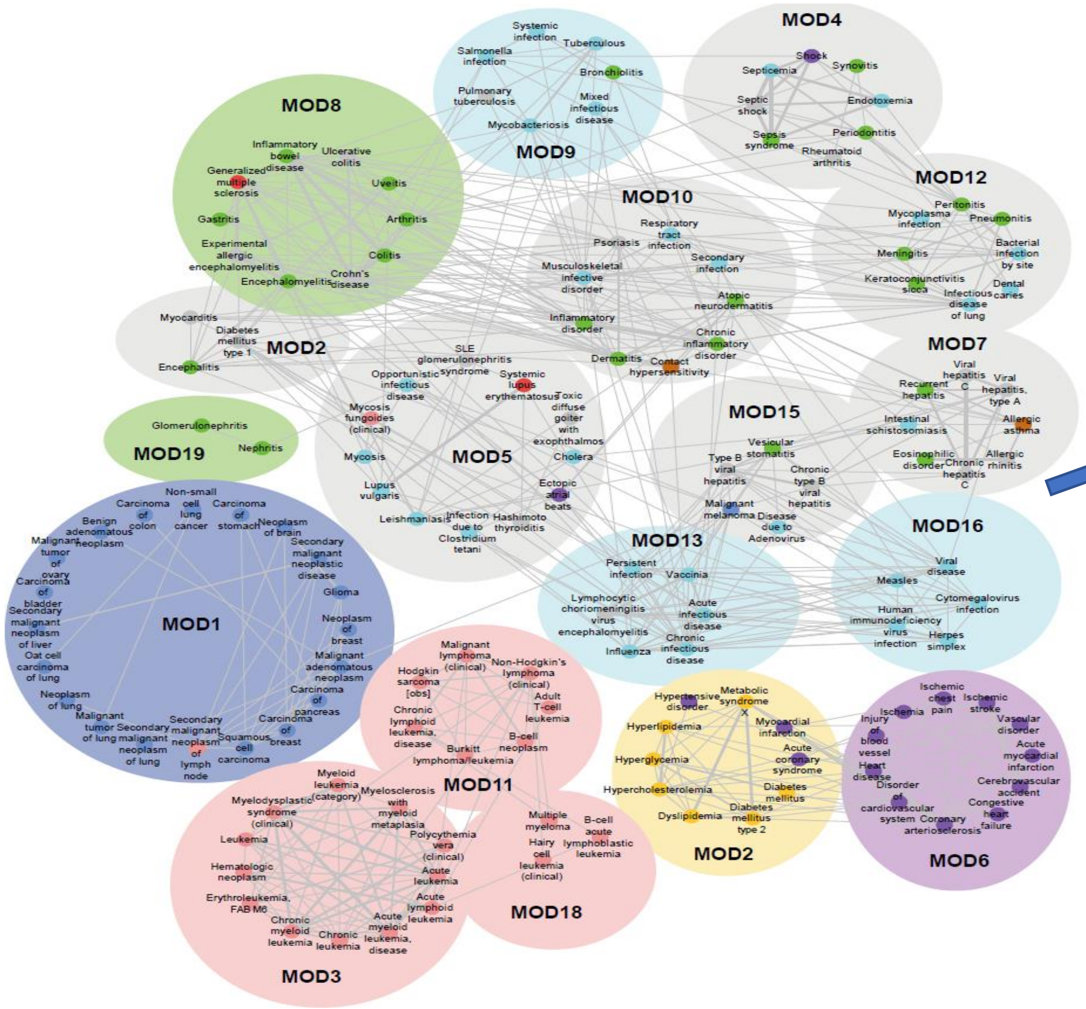


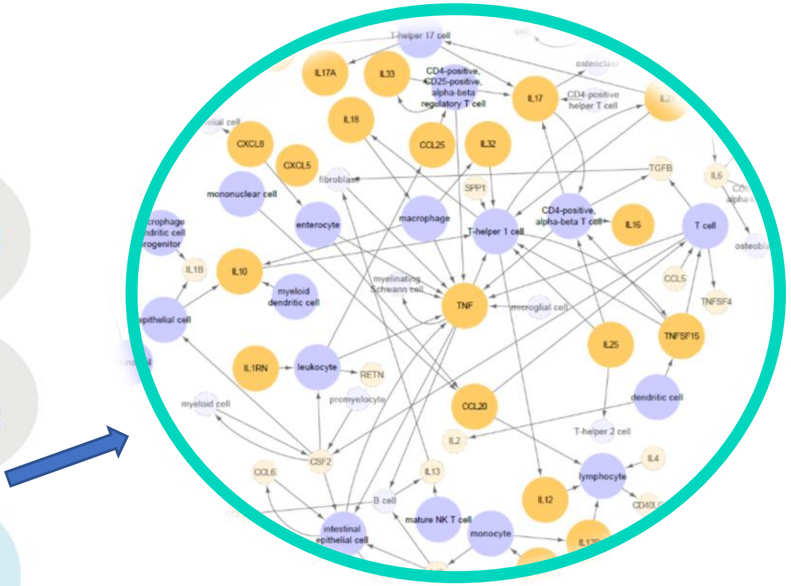
### Experimental Validation

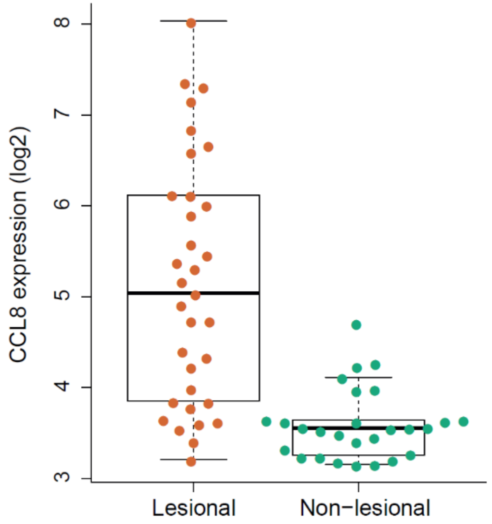


Kveler et al. Nature Biotech, 2018

# An immune-based classification of disease identifies novel targets







**Node Color**

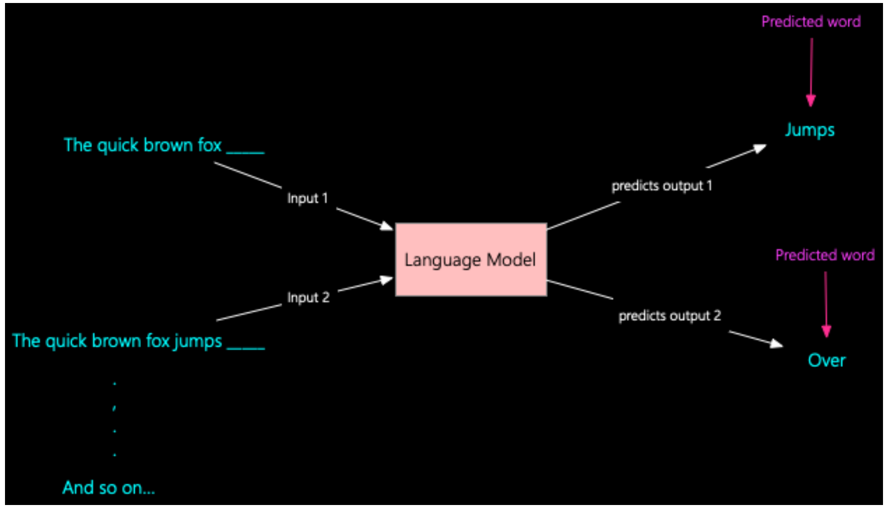
- ● Hypersensitivity condition
- ● Inflammatory disorder
- ● Neoplasm of hematopoietic cell type
- ● Non-hemato neoplasm
- ● Autoimmune disease
- ● Contagious disease
- ● Disorder of cardiovascular system
- ● Generalized metabolic disorder
- ● Multiple

Kveler et al., Nature Biotech, 2018

# Large language models, what's the change ?

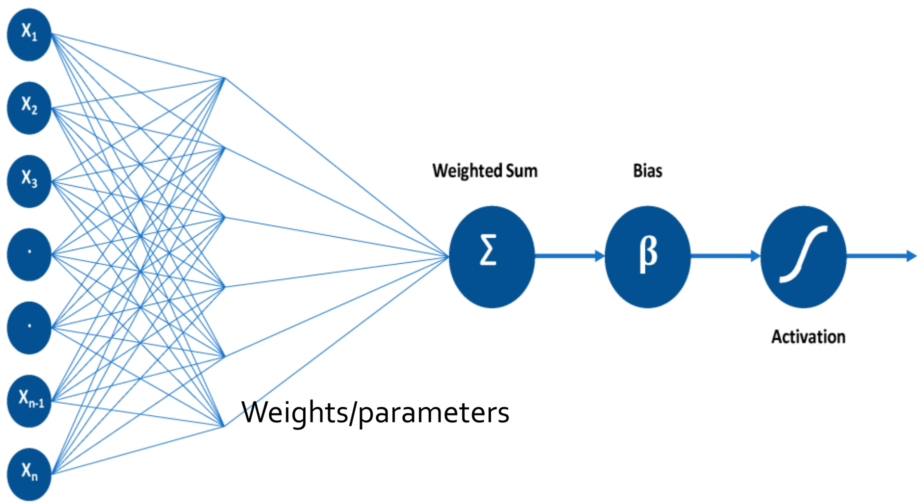
- Trained on enormous amounts of data

- Maturation of a deep learning architecture that suits language problems
  - Leverages everything it saw (training)
  - Probabilistic
  - Compresses the dimension of the data
  - Knows how to take context into account

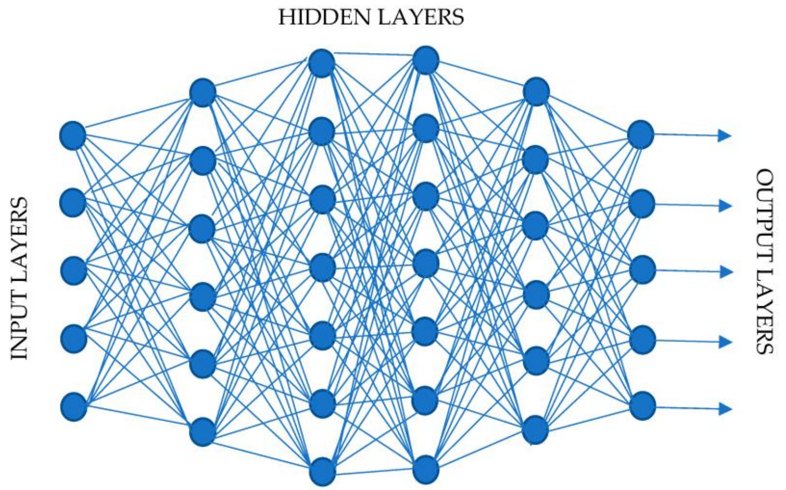


# The development of compressive probabilistic neural nets

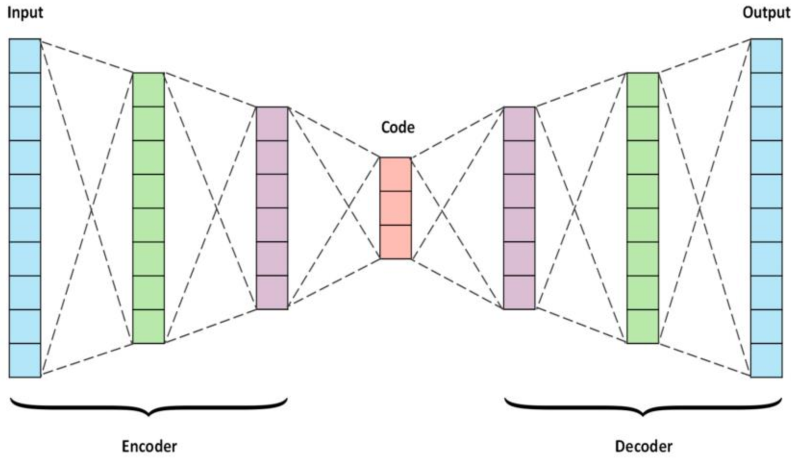
### Artificial neuron



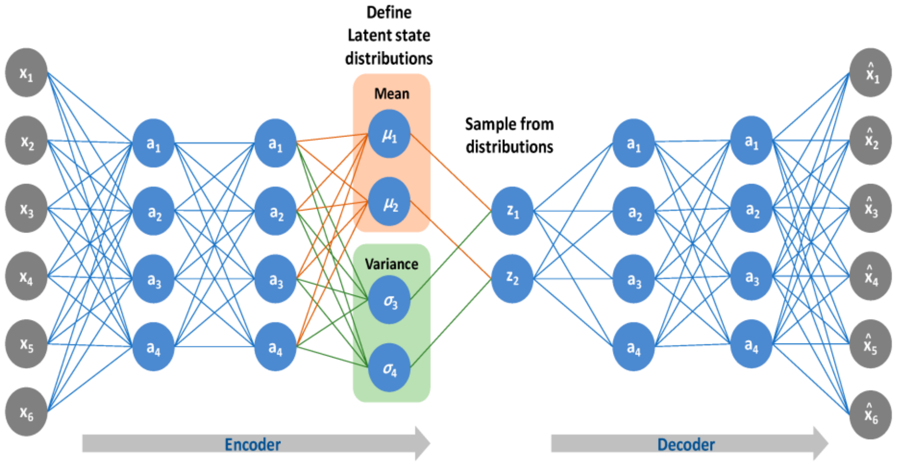
### Deep neural networks



### Compression via Encoder/Decoder architecture



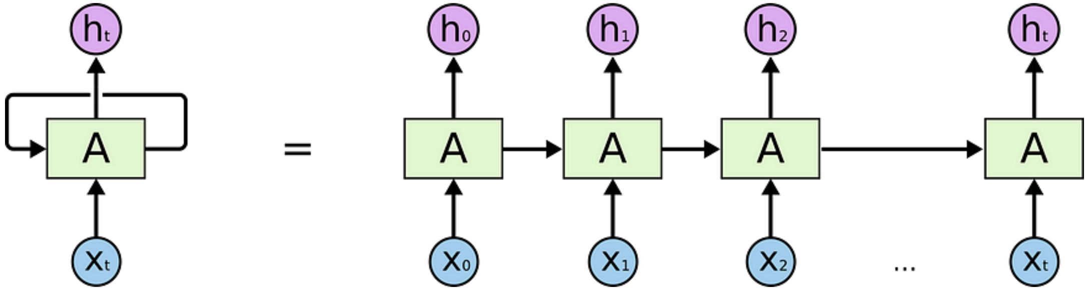
### Variational auto-encoders



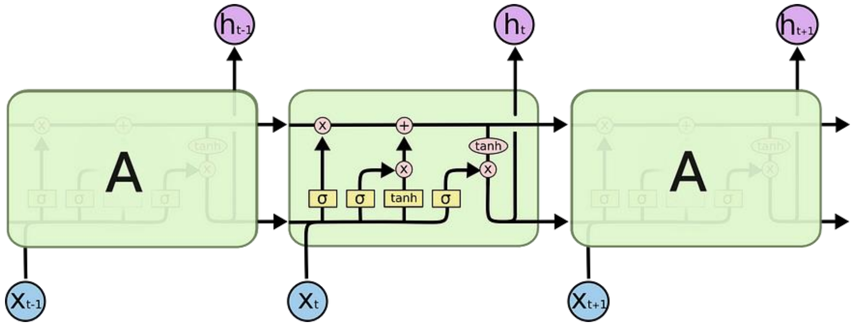
# The right architecture - A stroll down memory lane

SLOW

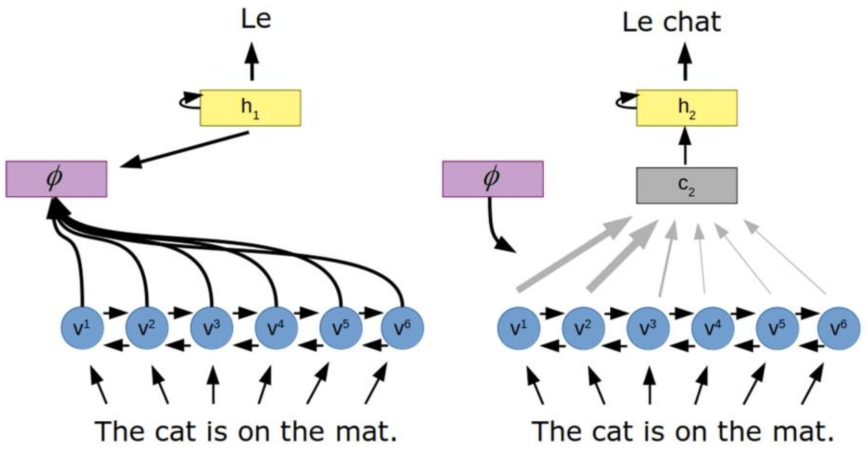
**Recurrent neural net model**: Struggles to maintain memory



**Long-Short term memory** (LSTM): Manipulate importance



**Self-Attention mechanisms via masking** (one at a time)



What you need for attention:

1. Tokenize, Embed – break down input and transform
2. Compute 3 vectors: (Q)uery, (K)ey ,(V)alue
3. Compute (embedded) QXK similarity between token pairs
4. Attention is given to tokens that have similarity score

**Multi-attention mechanisms**

- Repeat with different embeddings

# How transformers work

## Attention is all you need – all tokens get assessed simultaneously for attention

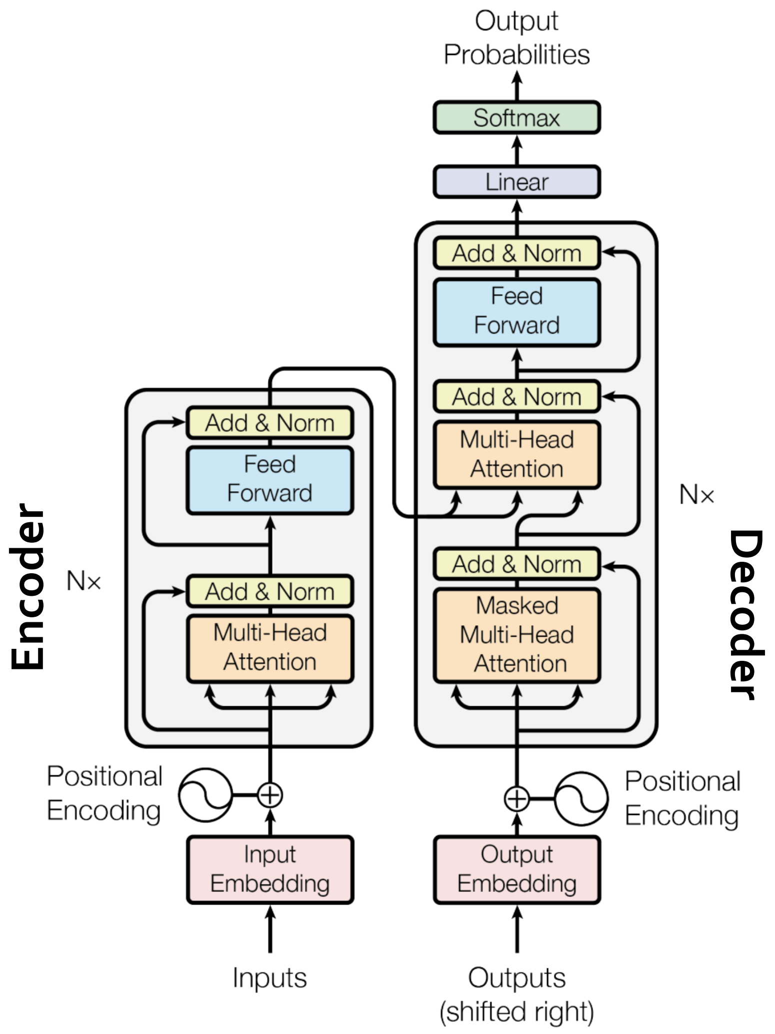
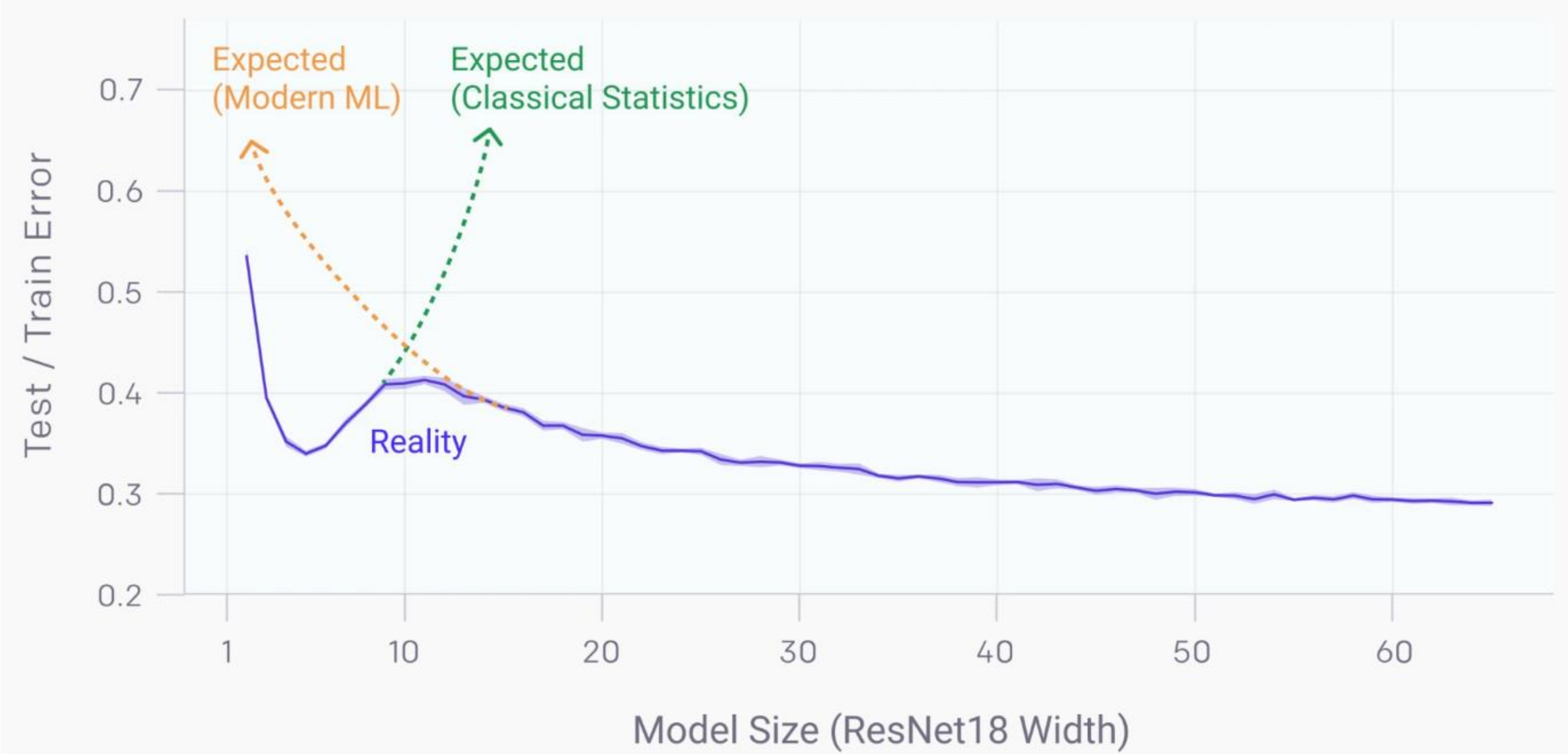


Figure 1: The Transformer - model architecture.

1. **Input & Input Embeddings –** Input text transformed to a numerical format
2. **Positional Encoding –** The order of words numerically transformed
3. **Encoder** – Breaks embeddings to atomic units and transform to an abstraction. Store as hidden state. Repeat many times (multi-attention)
4. **Outputs & output embeddings (shifted right, 1 token)** – Same as input, but masking next token, computes loss function of decoder and update parameters (both in training and inference stages)
5. **Decoder –** Estimates next token (output) based on input.
6. **Linear layer and softmax** – transform output back to high dimension for every token and assign probabilities for most likely output

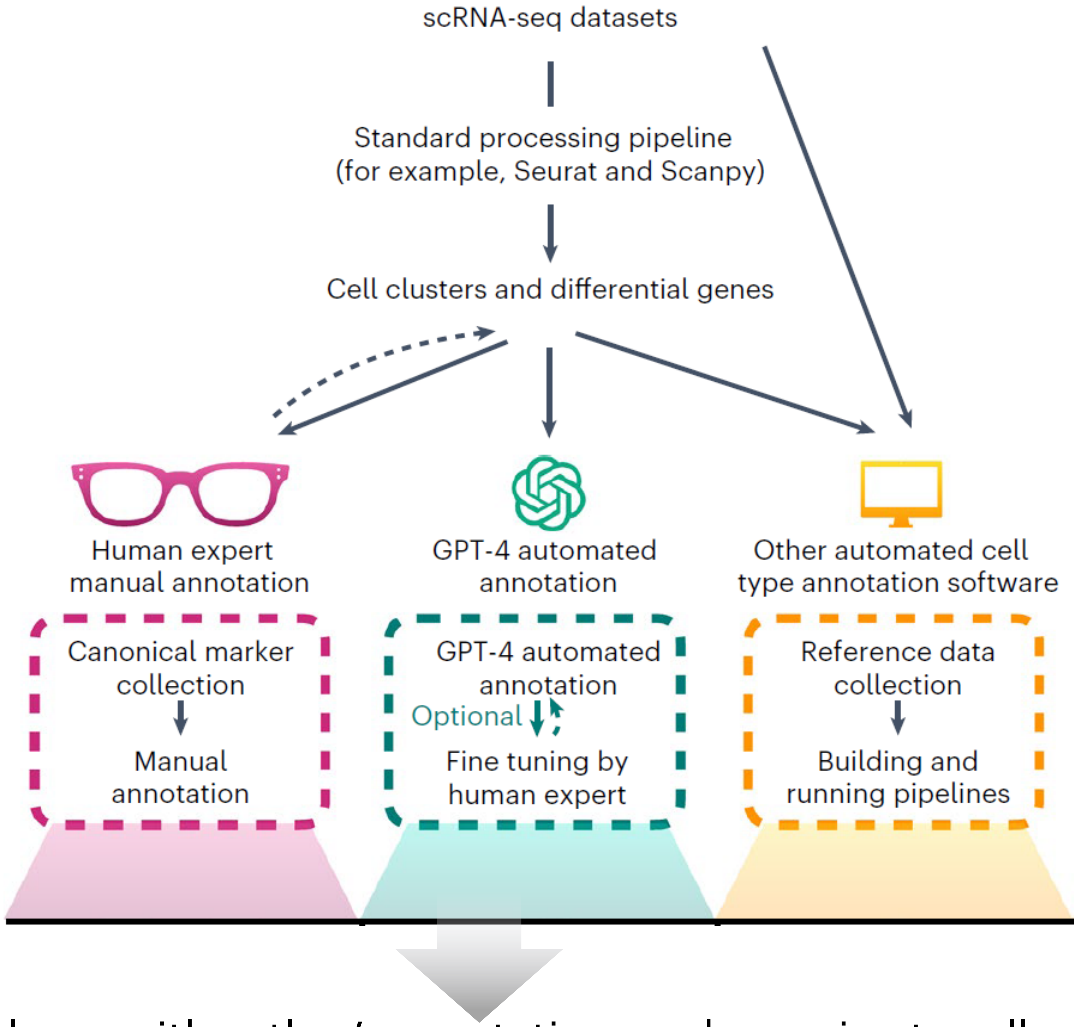
Vaswani et al., Advances in Neural Information Processing Systems, 2017

# These models have many params and yet appear to escape from overfitting



# Getting to cell knowledge

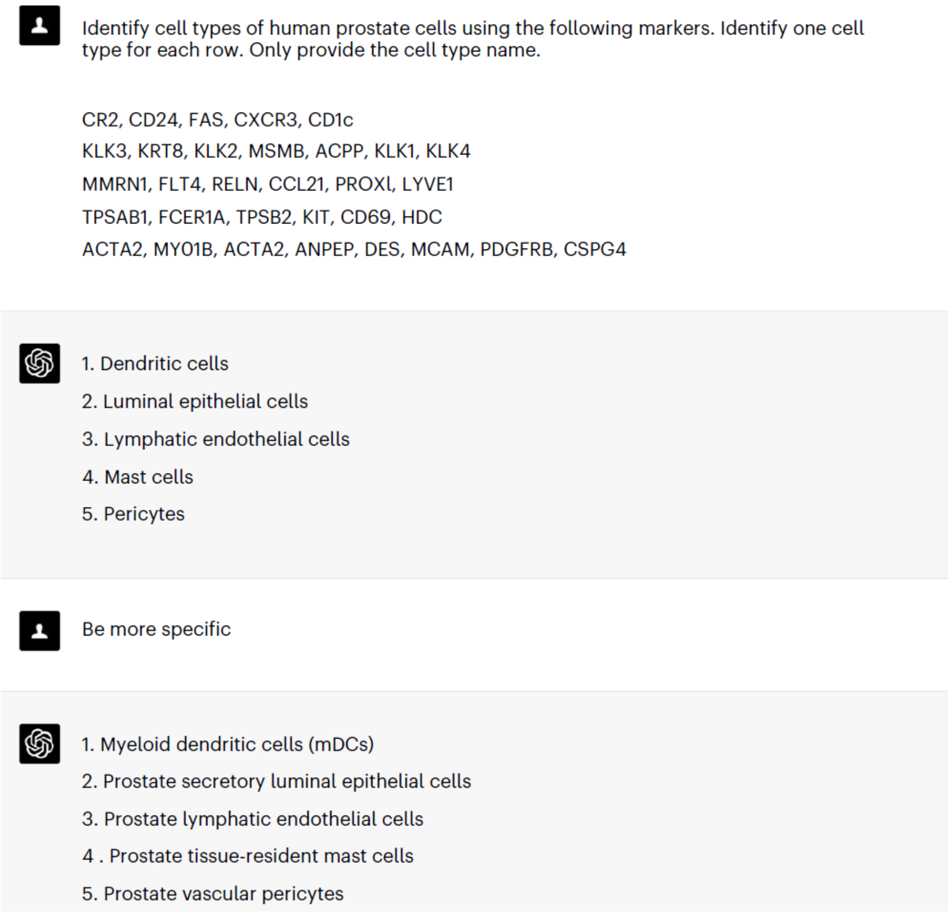
## A comparative study



Concordance with author's annotations and mapping to cell ontology

Hou, Ji, Nature Methods 2024

**b**



Identify cell types of human prostate cells using the following markers. Identify one cell type for each row. Only provide the cell type name.

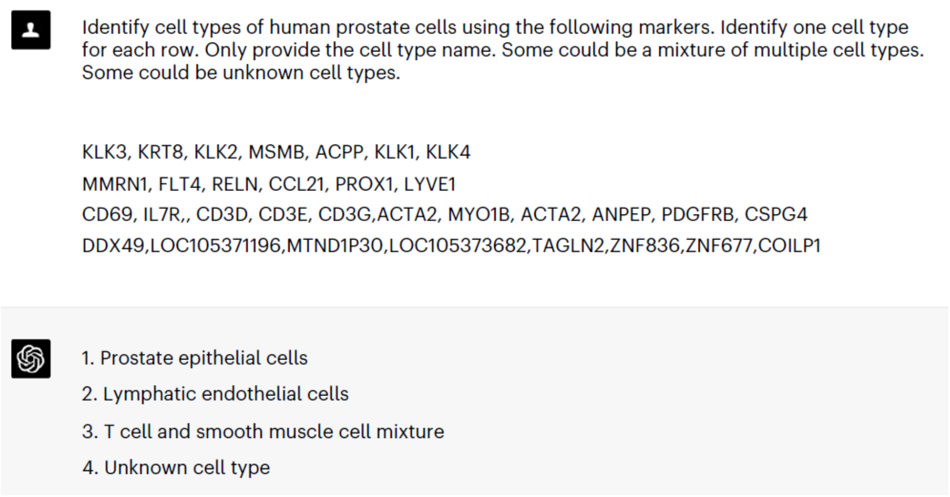
CR2, CD24, FAS, CXCR3, CD1c
KLK3, KRT8, KLK2, MSMB, ACPP, KLK1, KLK4
MMRN1, FLT4, RELN, CCL21, PROXl, LYVE1
TPSAB1, FCER1A, TPSB2, KIT, CD69, HDC
ACTA2, MYO1B, ACTA2, ANPEP, DES, MCAM, PDGFRB, CSPG4

1. Dendritic cells
2. Luminal epithelial cells
3. Lymphatic endothelial cells
4. Mast cells
5. Pericytes

Be more specific

1. Myeloid dendritic cells (mDCs)
2. Prostate secretory luminal epithelial cells
3. Prostate lymphatic endothelial cells
4 . Prostate tissue-resident mast cells
5. Prostate vascular pericytes

**c**



Identify cell types of human prostate cells using the following markers. Identify one cell type for each row. Only provide the cell type name. Some could be a mixture of multiple cell types. Some could be unknown cell types.

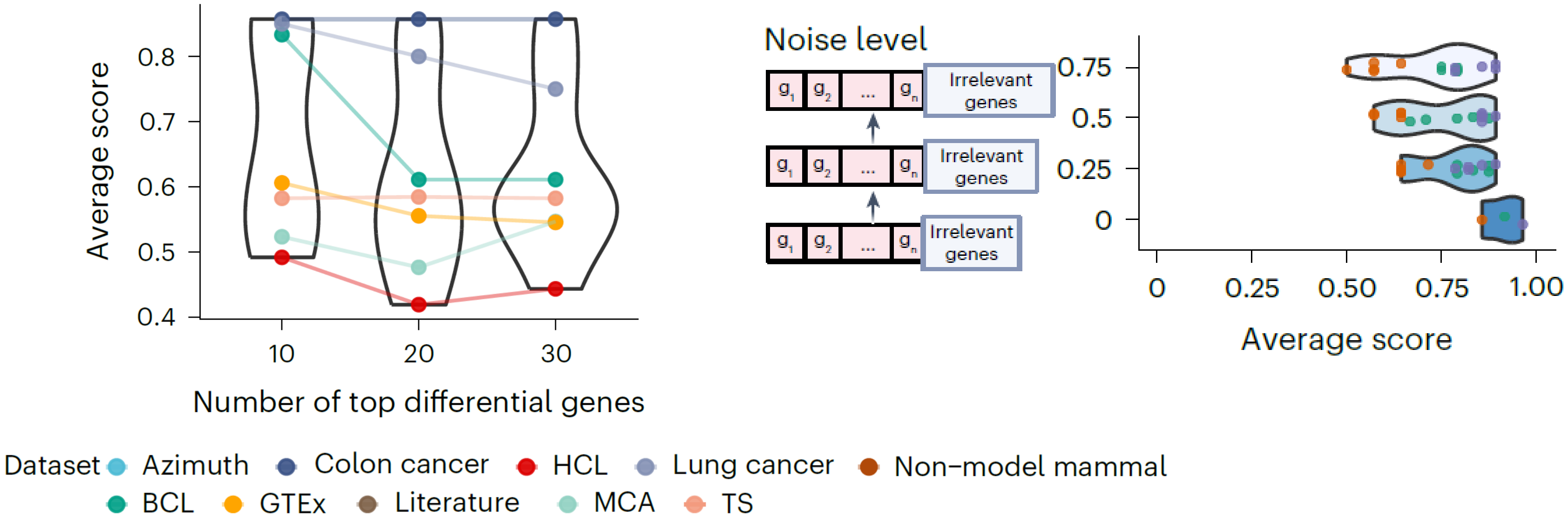
KLK3, KRT8, KLK2, MSMB, ACPP, KLK1, KLK4
MMRN1, FLT4, RELN, CCL21, PROX1, LYVE1
CD69, IL7R,, CD3D, CD3E, CD3G,ACTA2, MYO1B, ACTA2, ANPEP, PDGFRB, CSPG4
DDX49,LOC105371196,MTND1P30,LOC105373682,TAGLN2,ZNF836,ZNF677,COILP1

1. Prostate epithelial cells
2. Lymphatic endothelial cells
3. T cell and smooth muscle cell mixture
4. Unknown cell type

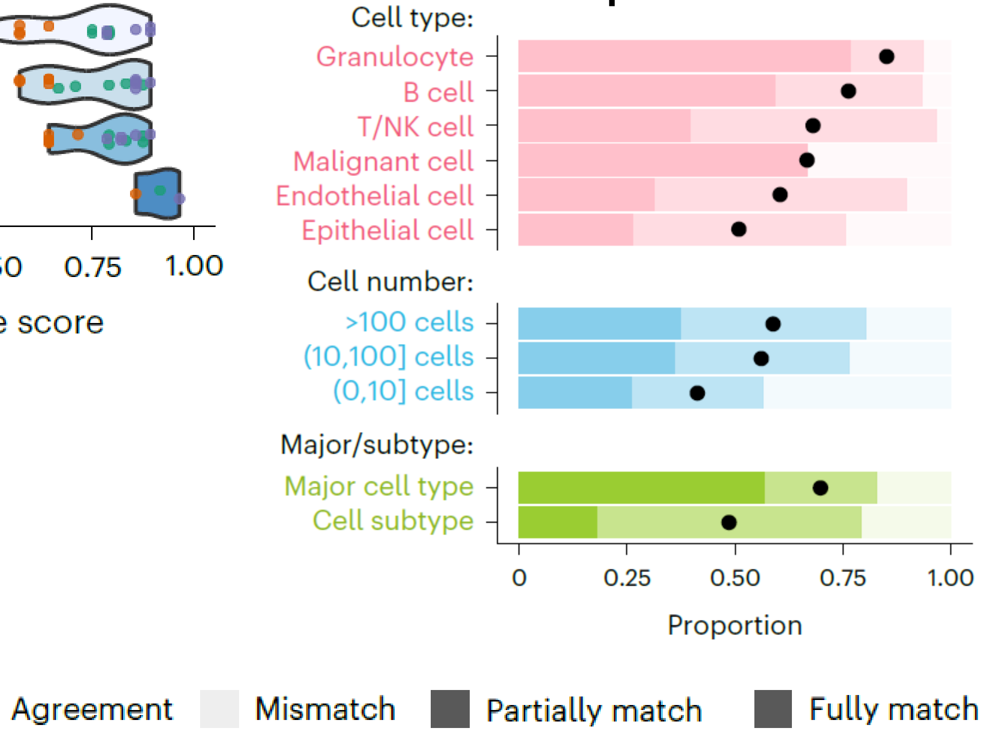
# Cell annotation info exits in the knowledge of top differential genes and is robust to noise

## Few genes needed

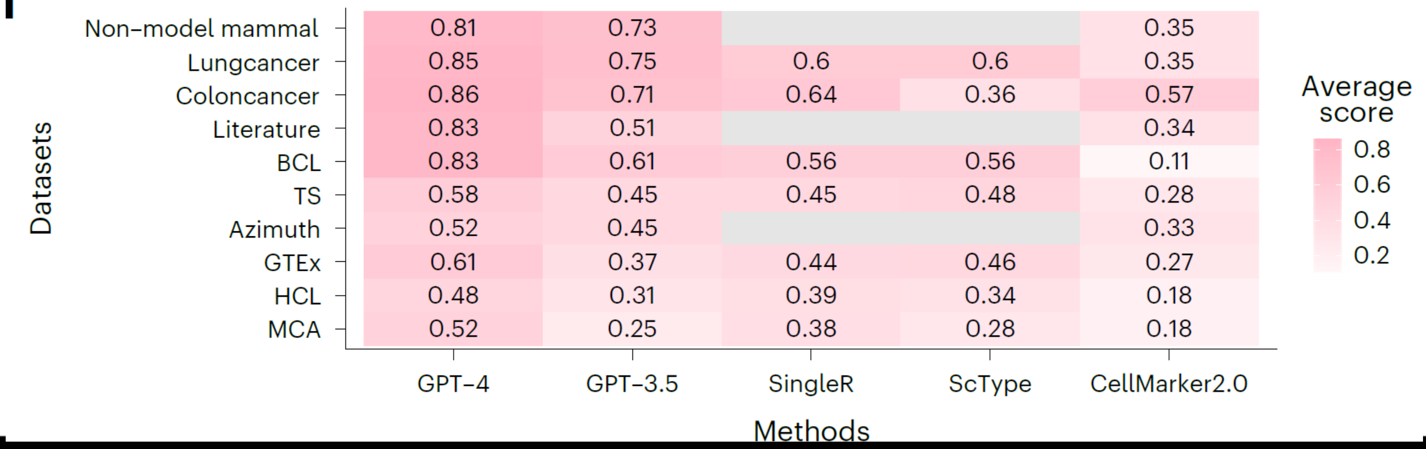
## Annotation robust to noise



## Performance resolution & tissue dependent



## Datasets



| Datasets | GPT-4 | GPT-3.5 | SingleR | ScType | CellMarker2.0 |
|---|---|---|---|---|---|
| Non-model mammal | 0.81 | 0.73 | | | 0.35 |
| Lungcancer | 0.85 | 0.75 | 0.6 | 0.6 | 0.35 |
| Coloncancer | 0.86 | 0.71 | 0.64 | 0.36 | 0.57 |
| Literature | 0.83 | 0.51 | | | 0.34 |
| BCL | 0.83 | 0.61 | 0.56 | 0.56 | 0.11 |
| TS | 0.58 | 0.45 | 0.45 | 0.48 | 0.28 |
| Azimuth | 0.52 | 0.45 | | | 0.33 |
| GTEx | 0.61 | 0.37 | 0.44 | 0.46 | 0.27 |
| HCL | 0.48 | 0.31 | 0.39 | 0.34 | 0.18 |
| MCA | 0.52 | 0.25 | 0.38 | 0.28 | 0.18 |

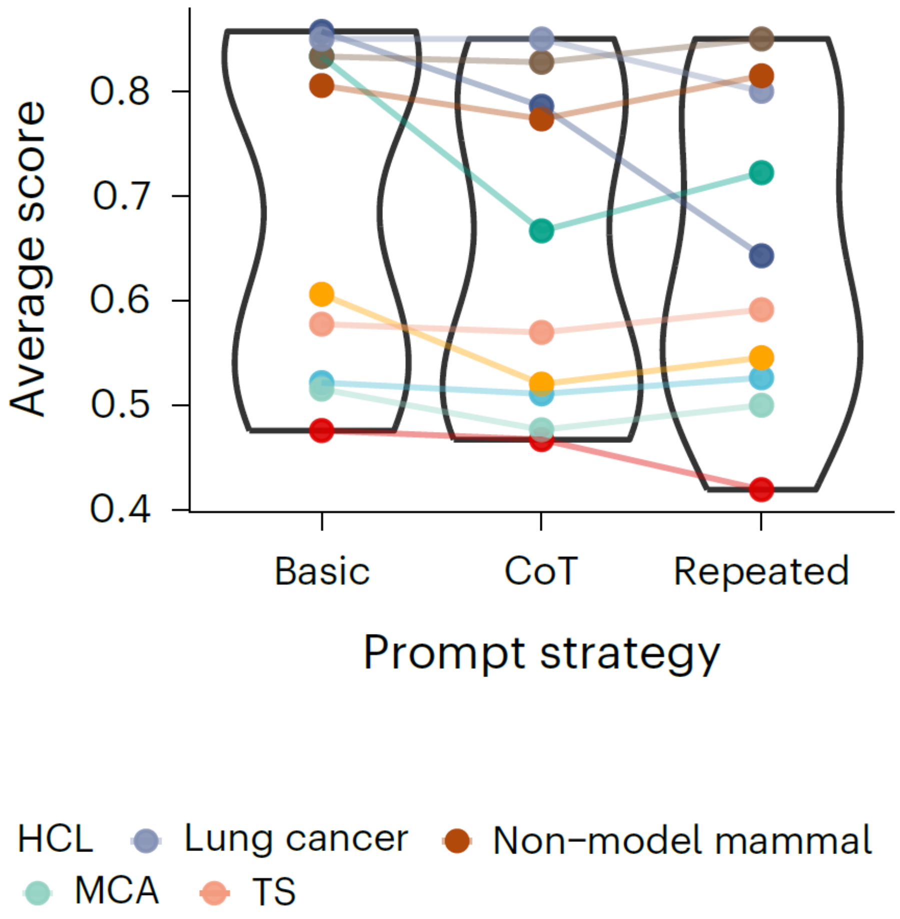
Hou, Ji, Nature Methods 2024

# Simple prompts suffice for most annotations

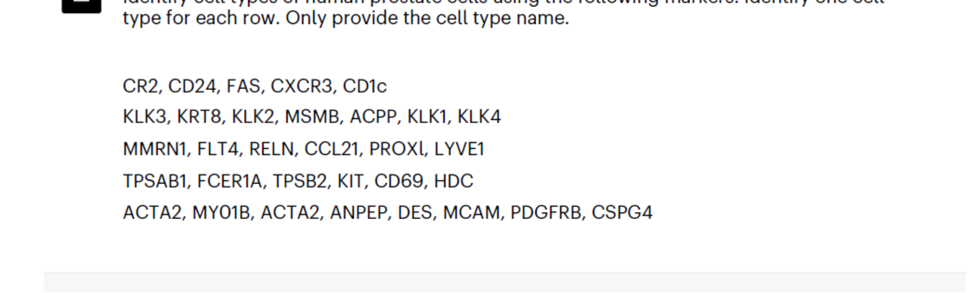
**Basic prompts**: ‘Identify cell types of TissueName cells using the following markers separately for each row. Only provide the cell type name. Do not show numbers before the name. Some can be a mixture of multiple cell types.\n GeneList’.

**Chain of thought** prompts start with: “‘Because *CD3* gene is a marker gene of T cells, if *CD3* gene is included in the marker gene list of an unknown cell type, the cell type is likely to be T cells, a subtype of T cells, or a mixed cell type containing T cells’.”

**Repeated:** Perform basic 5 times and take top hit

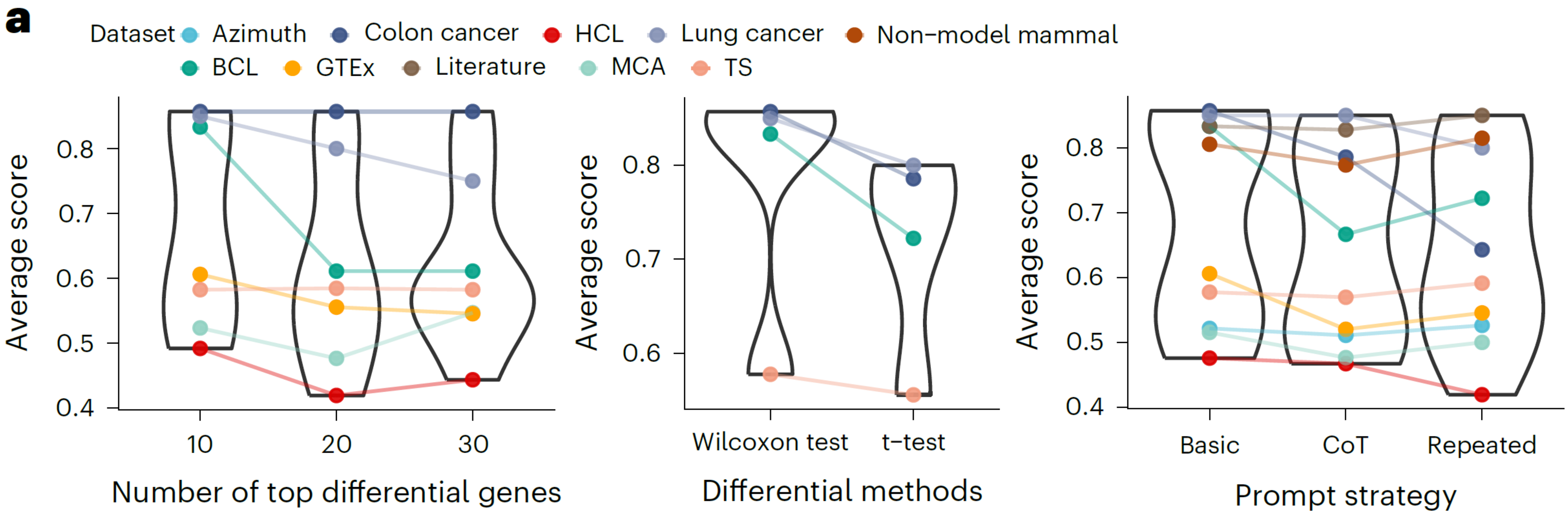


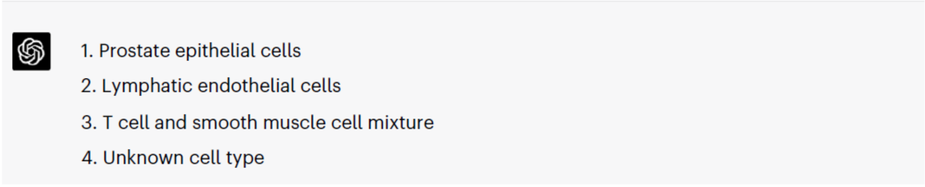
# Getting to cell knowledge



Identify cell types of human prostate cells using the following markers. Identify one cell type for each row. Only provide the cell type name.

CR2, CD24, FAS, CXCR3, CD1c
KLK3, KRT8, KLK2, MSMB, ACPP, KLK1, KLK4
MMRN1, FLT4, RELN, CCL21, PROXl, LYVE1
TPSAB1, FCER1A, TPSB2, KIT, CD69, HDC
ACTA2, MY01B, ACTA2, ANPEP, DES, MCAM, PDGFRB, CSPG4





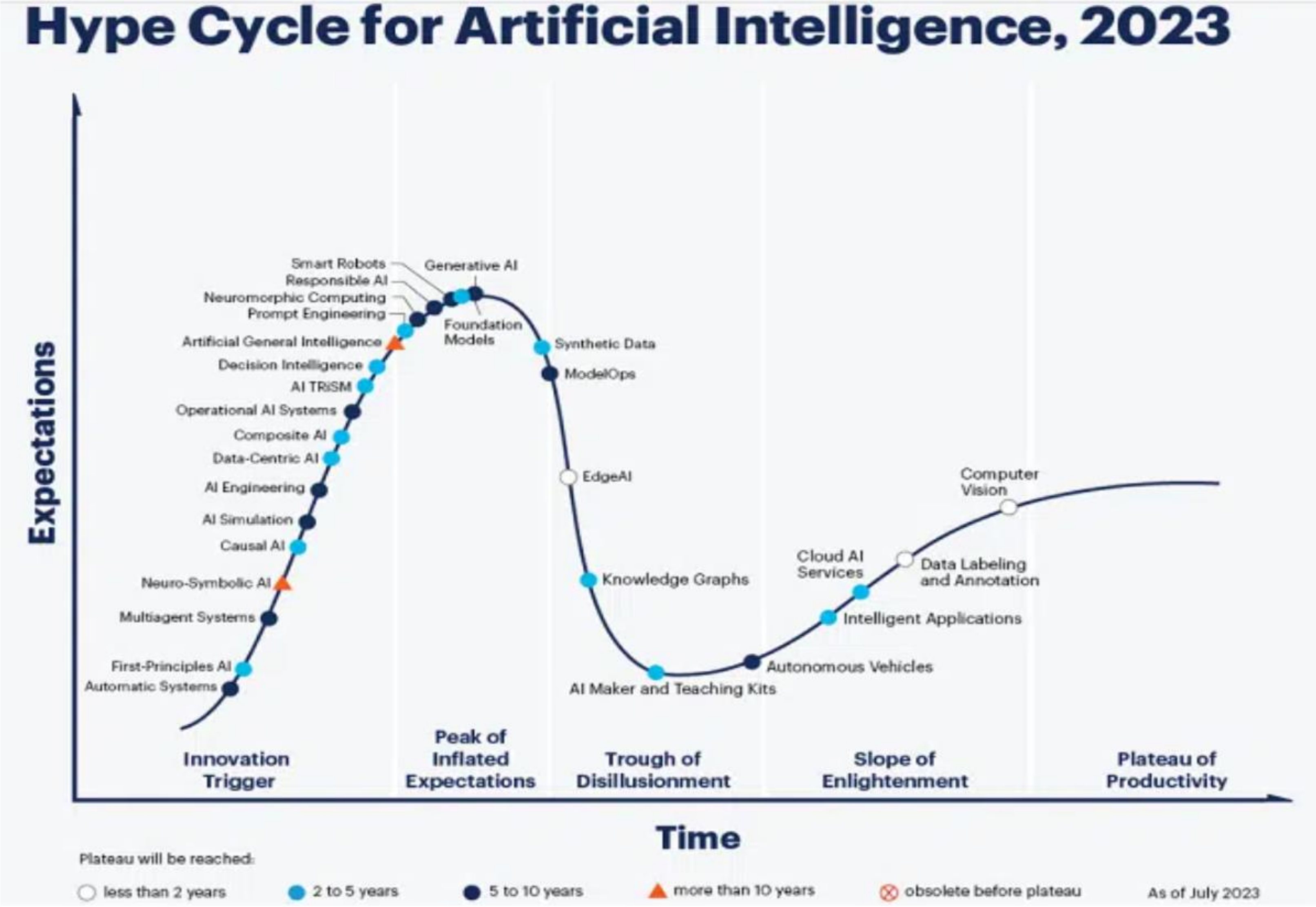
1. Prostate epithelial cells
2. Lymphatic endothelial cells
3. T cell and smooth muscle cell mixture
4. Unknown cell type

Hou, Ji, Nature Methods 2024

# The basis of success of LLM

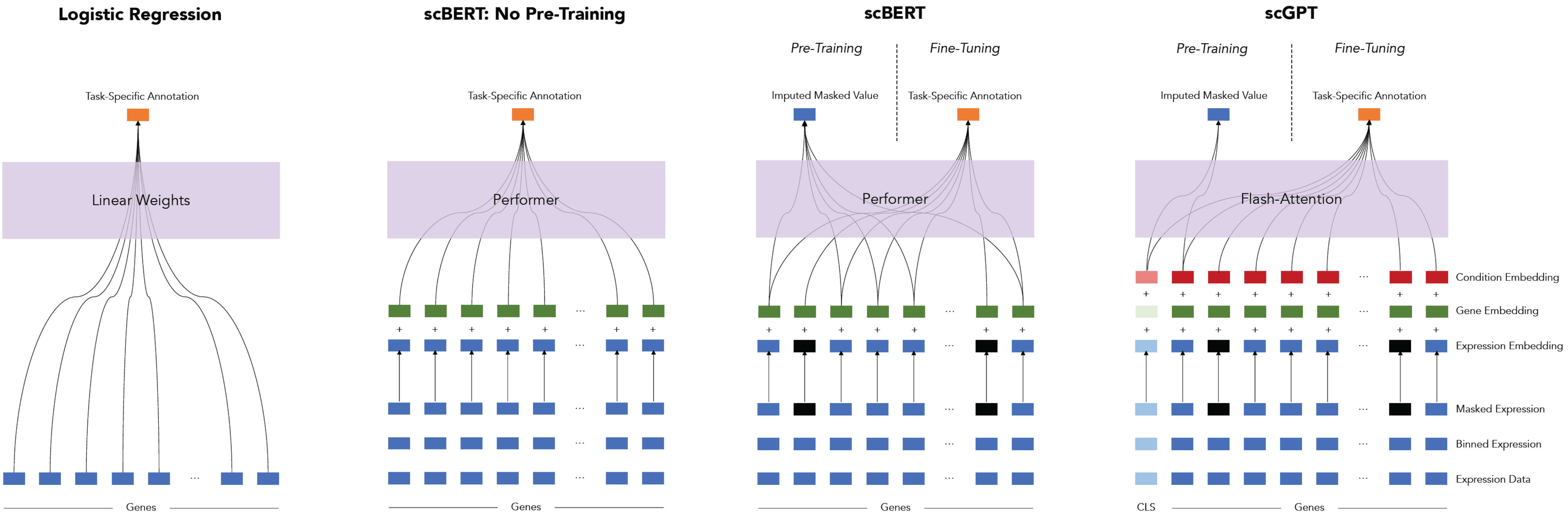
- **<u>There is a language</u>**
- The model is truly **Foundational →** enormous amount of data went it
- **The use cases appear in the data well**
- **Does the same for biological data ?**

# Beware the hype



Gartner, https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle

# Put the hype to the test for a cell-annotation task

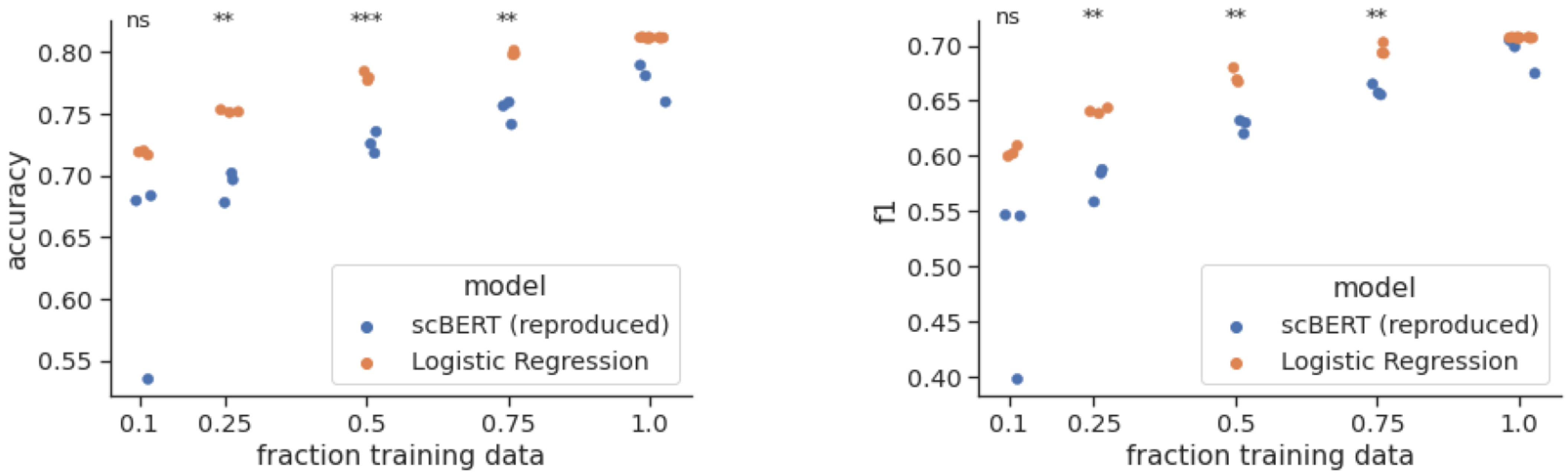


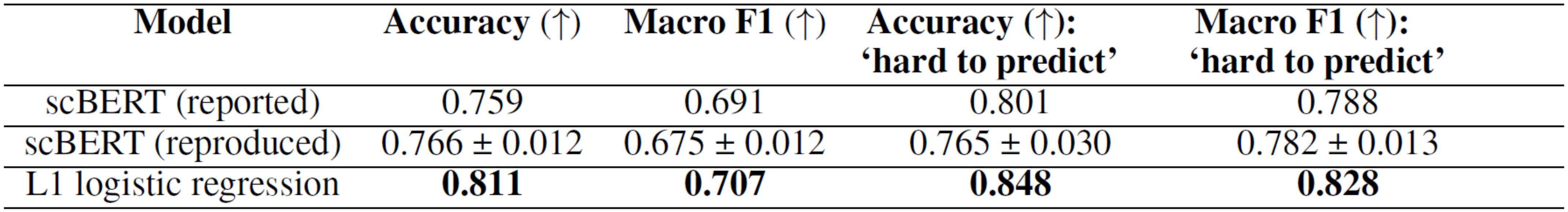
scBERT -Yang et al. Nature Machine Intelligence ,Sept. 2022
scGPT – Cui et al. Nature Methods, Feb. 2024

Boiarsky et al bioRxiv , Oct. 2023

# Logistic regression outperforms foundation models fine-tuned cell annotations

Dataset dependent effects observed



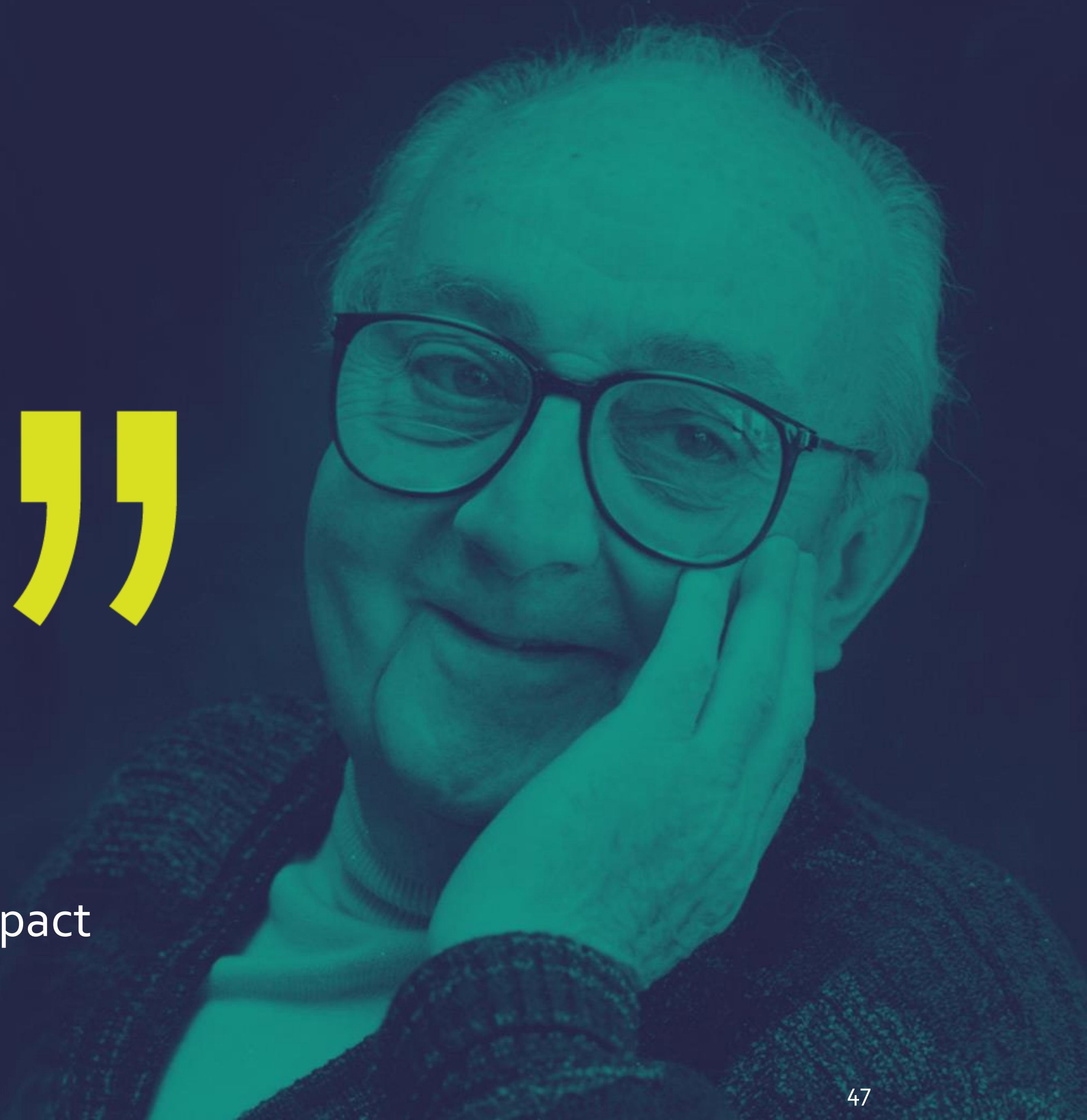


| Model | Accuracy (↑) | Macro F1 (↑) | Accuracy (↑):<br>‘hard to predict’ | Macro F1 (↑):<br>‘hard to predict’ |
|---|---|---|---|---|
| scBERT (reported) | 0.759 | 0.691 | 0.801 | 0.788 |
| scBERT (reproduced) | 0.766 ± 0.012 | 0.675 ± 0.012 | 0.765 ± 0.030 | 0.782 ± 0.013 |
| L1 logistic regression | **0.811** | **0.707** | **0.848** | **0.828** |

Boiarsky et al bioRxiv , Oct. 2023

“All models are wrong, but some are useful”

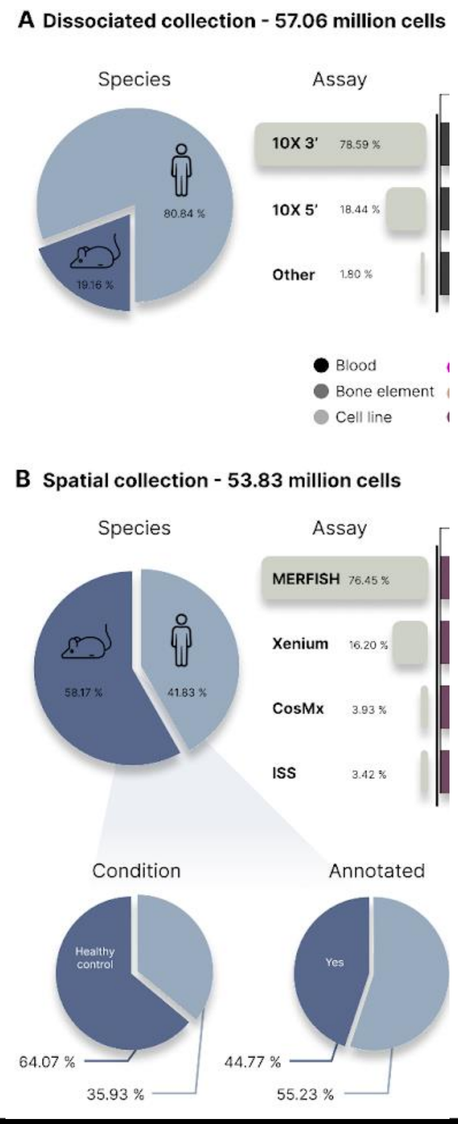
*George E.P. Box*

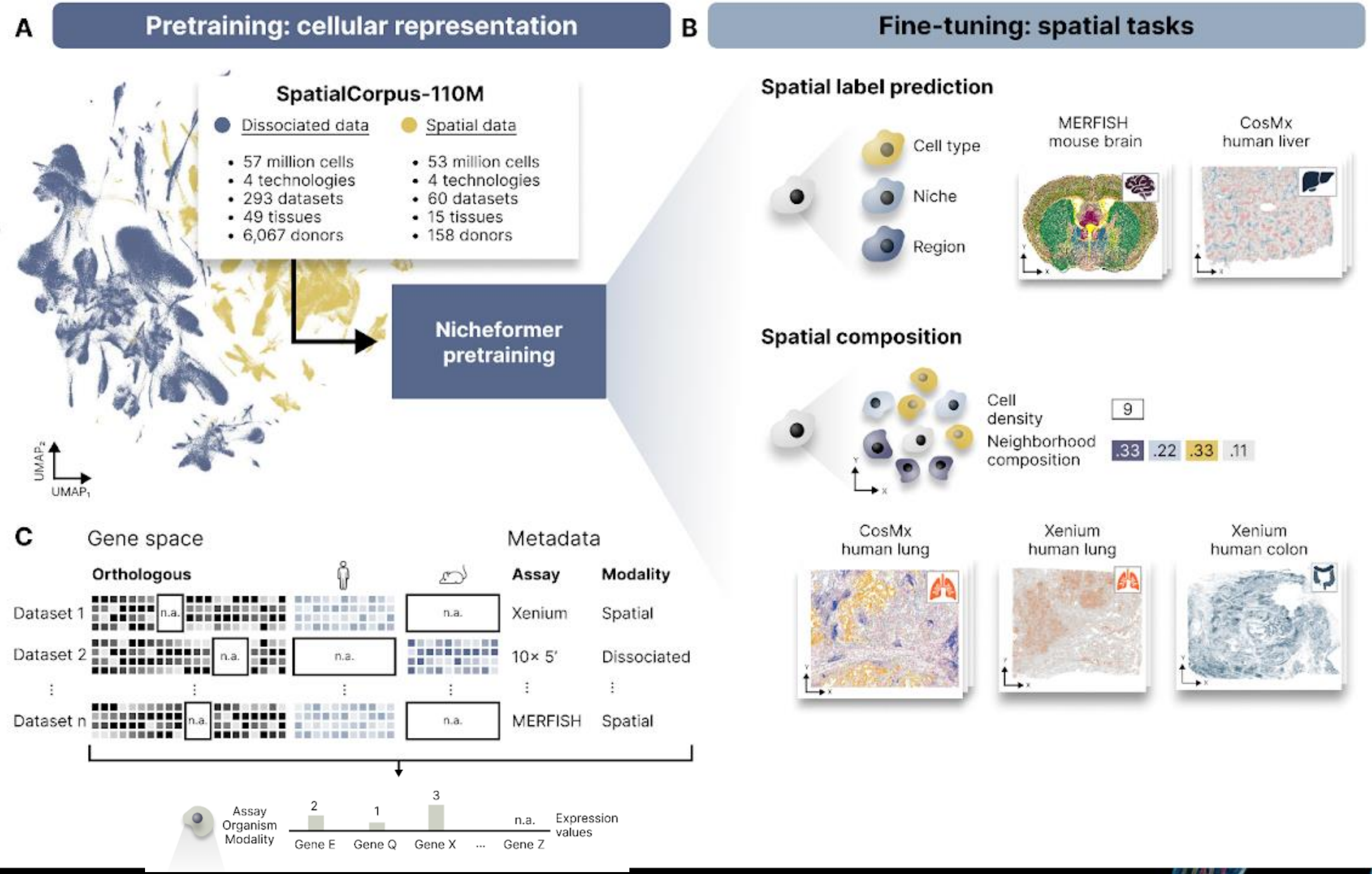


Prioritize by models that bring utility & impact

# Nicheformer – a foundation model for spatial calling tasks

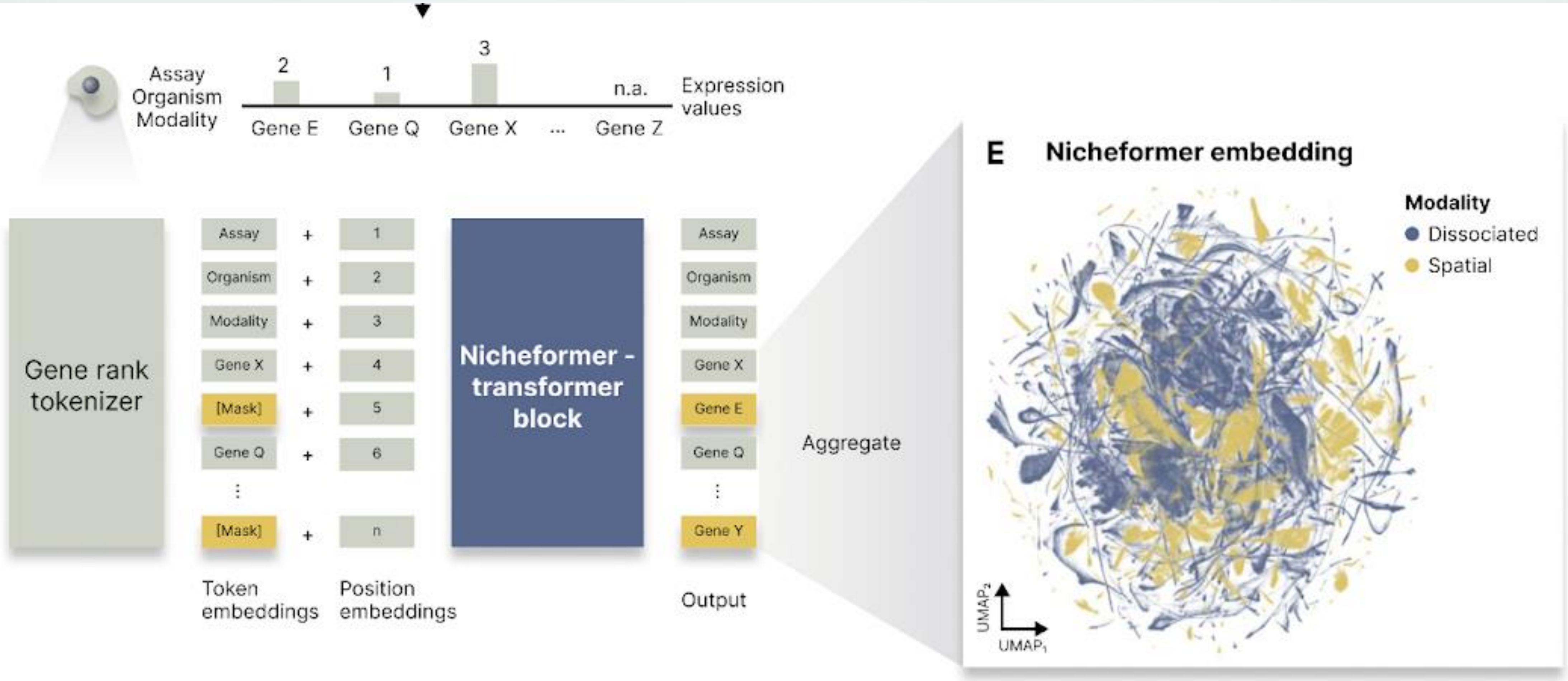
## ~110M cells disassociated and spatial, across species platform





Schaar et al. bioRxiv , April 2024 48

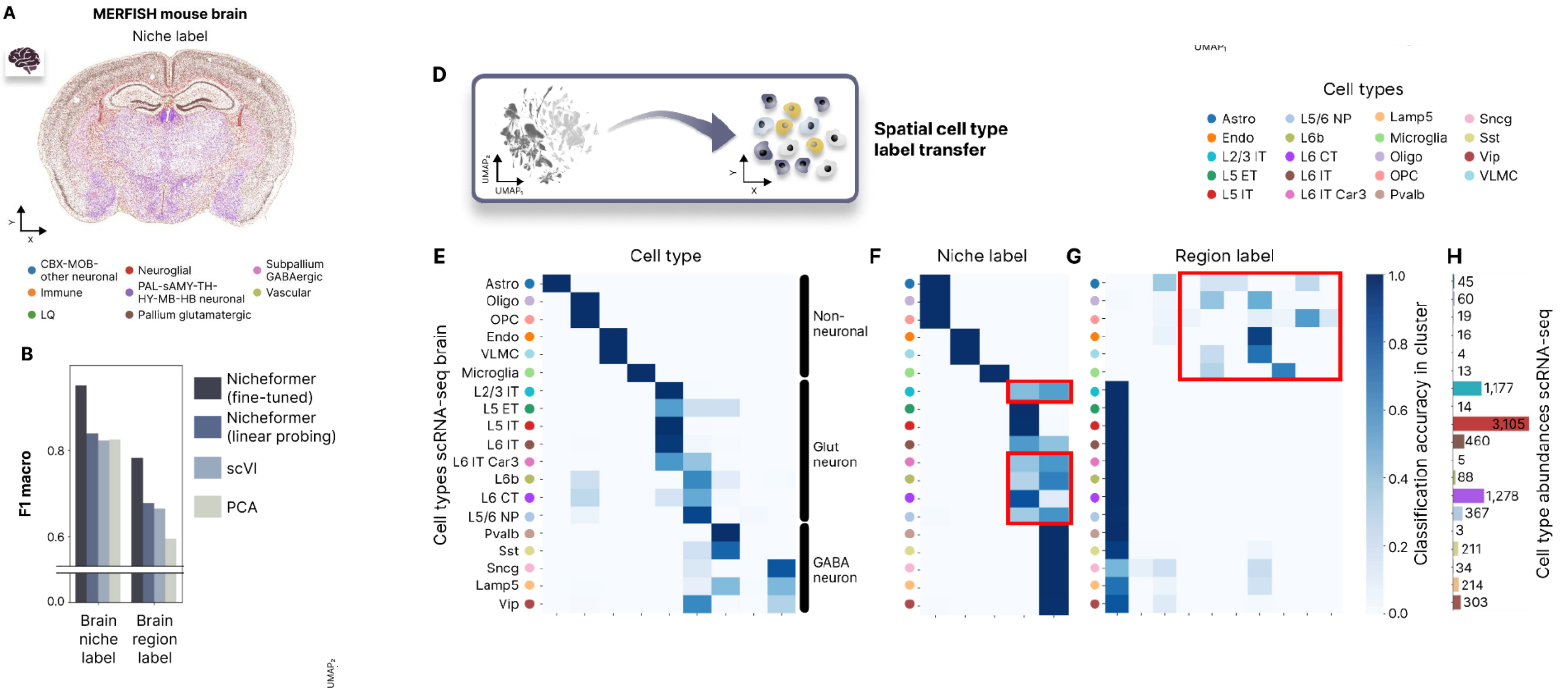
# Nicheformer – architecture



Context length of 1,500 gene tokens as transformer input.
Transformer block consisting of 12 transformer encoder units with 16 attention heads per layer → 512-dimensional embedding

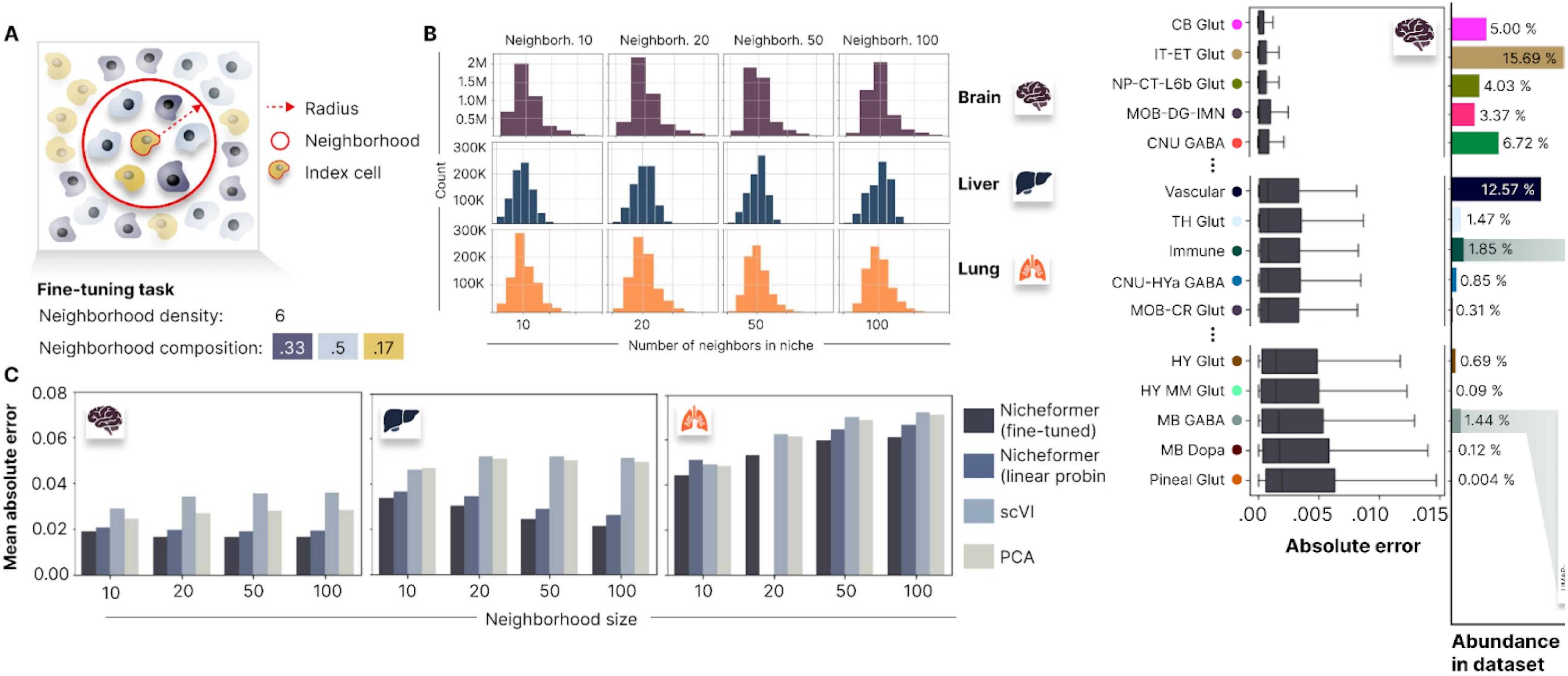
Schaar et al. bioRxiv , April 2024

# There is added value in spatial modeling and an ability to assign spatial info to disassociated cells



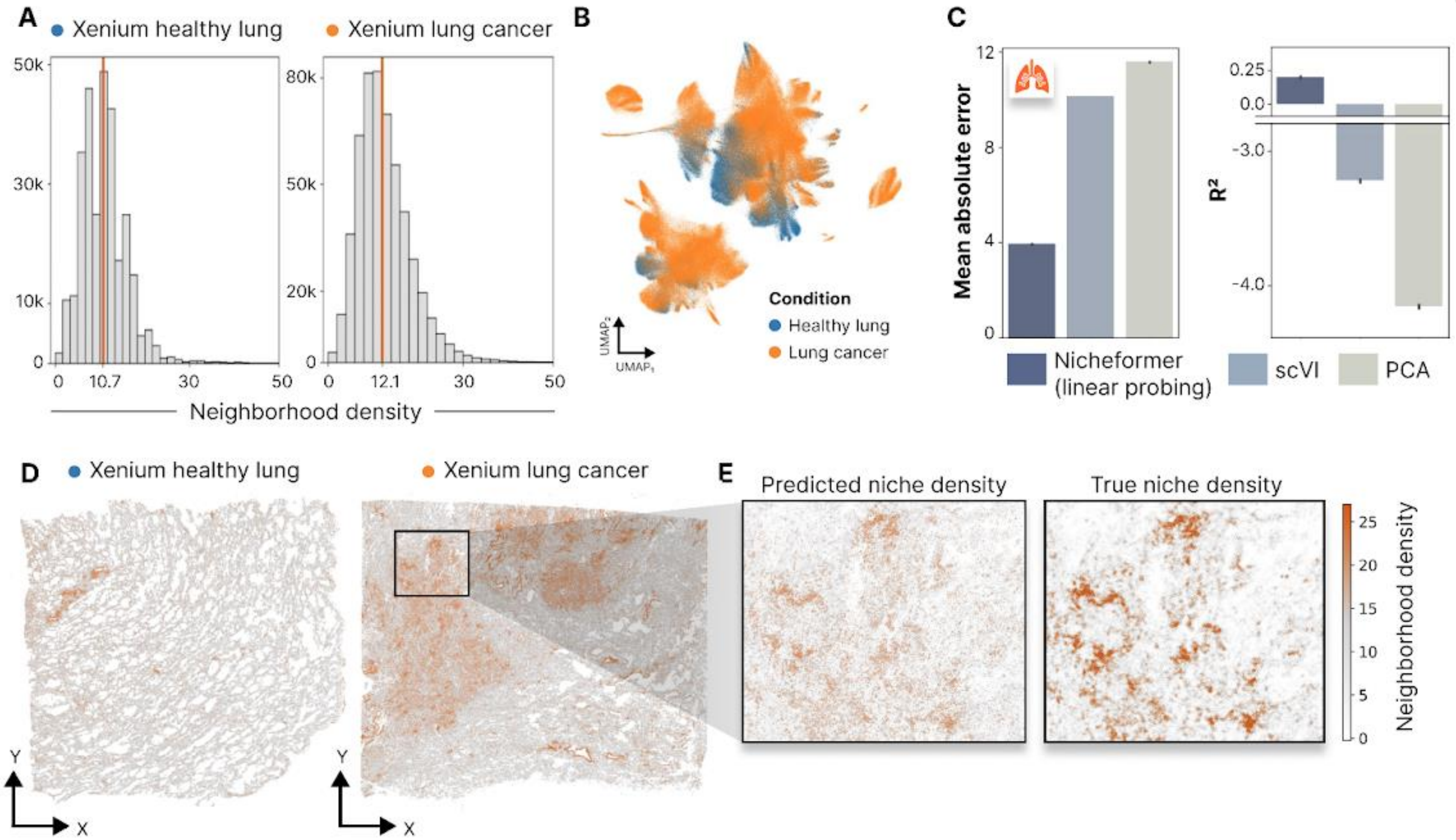
Schaar et al. bioRxiv , April 2024

# Model utility in predicting cell neighborhood composition



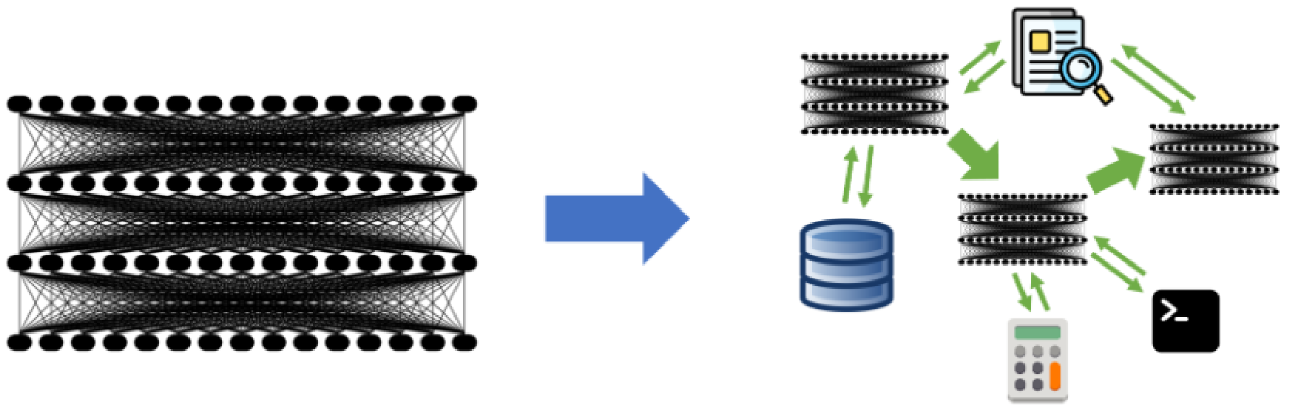
Schaar et al. bioRxiv , April 2024

# Model utility in predicting cell density



Schaar et al. bioRxiv , April 2024

# The future: Compound AI Systems

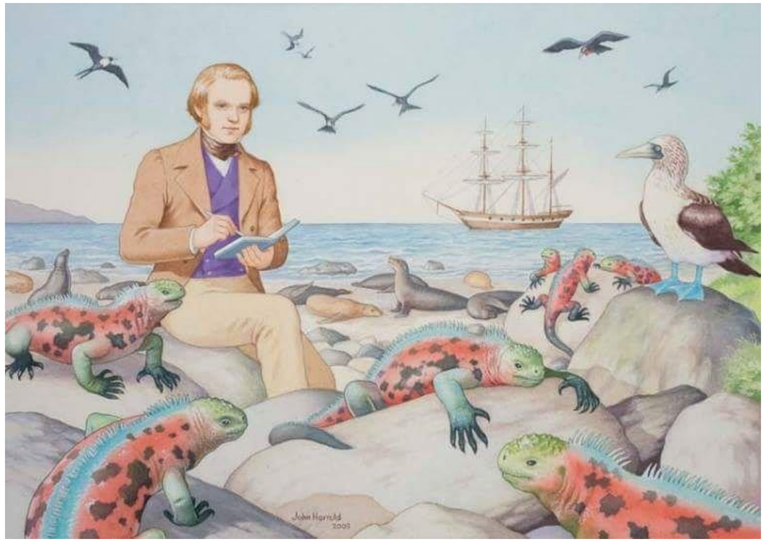


*Increasingly many new AI results are from compound systems.*

- Tasks are easier to improve at a system design
- More control and trust
- Systems can be dynamic
- Flexible performance goals

Zaharia et al. The shift from models to compound systems, April 2024

# Biologist of the future



~19th century

~20th century

~21st century

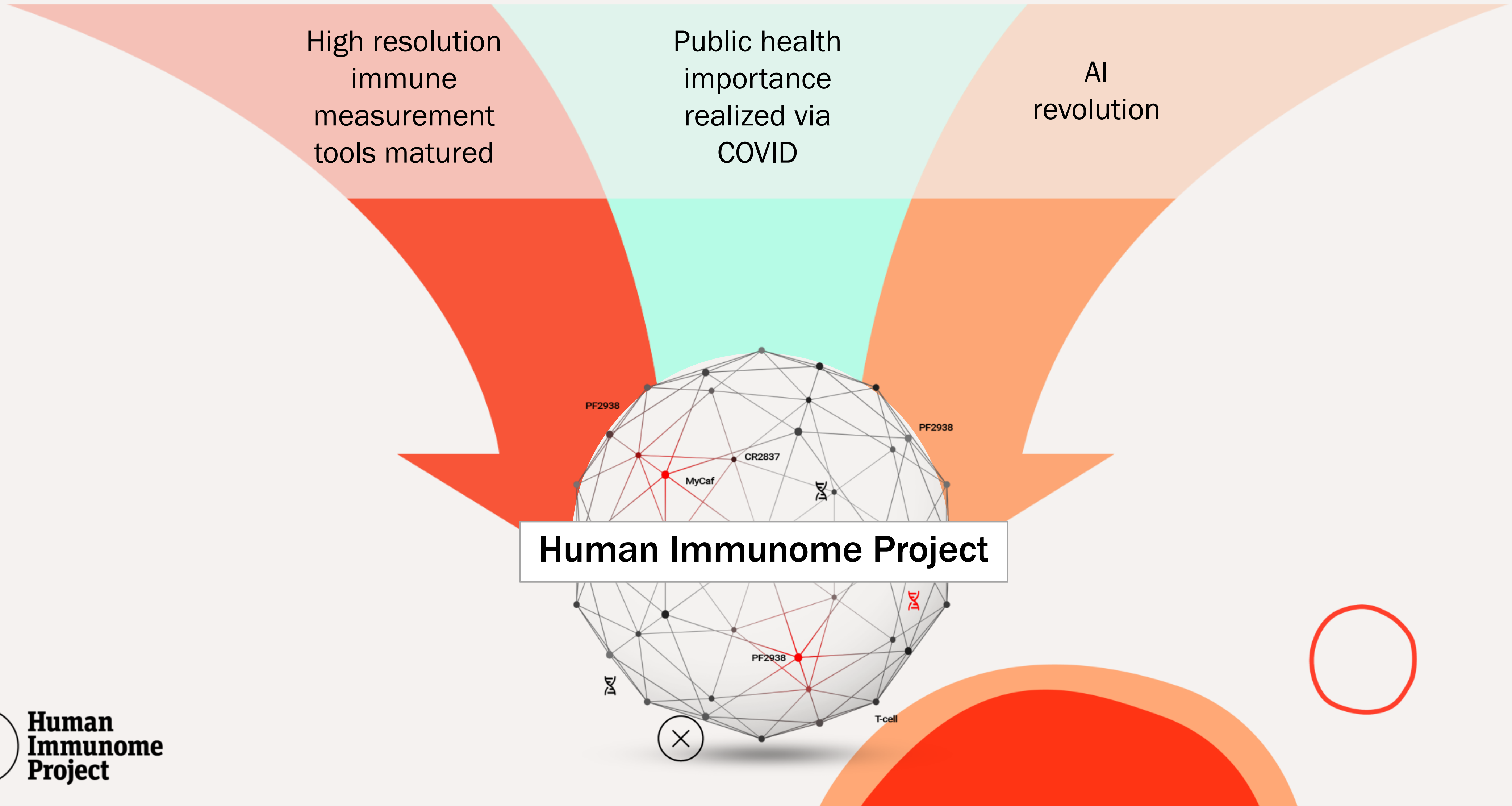
# The problem

Our understanding of immune variation across people and over time is **rudimentary**;

limited data on how baseline immune status is **linked to functional outcomes**;

**difficult to predict** health trajectory, treatment response, and other outcomes at the individual level

# Now is the time to map global diversity of immune health



High resolution immune measurement tools matured

Public health importance realized via COVID

AI revolution

Human Immunome Project

Human Immunome Project

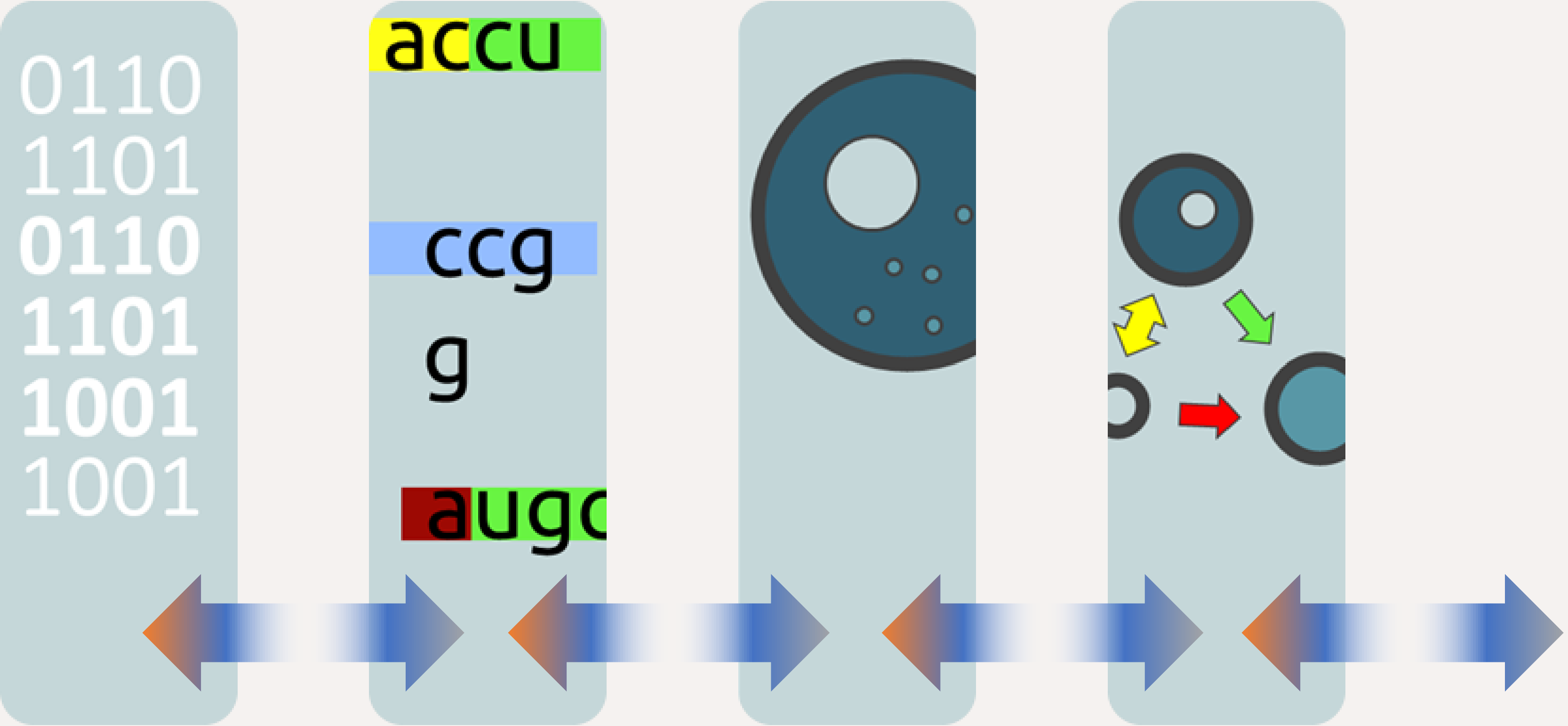
# STRATEGIC PLAN – VISION STATEMENT

**A PREDICTIVE UNDERSTANDING OF** immunological baseline and functional responses encompassing all **POPULATIONS IS NEEDED** to enable research, drug discovery and economy of global health care

**The utilitarian version: Tailored, global reference ranges at high resolution**

Human Immunome Project

# The data we collect will allow predicting immune health



Raw Data

Within Data Type

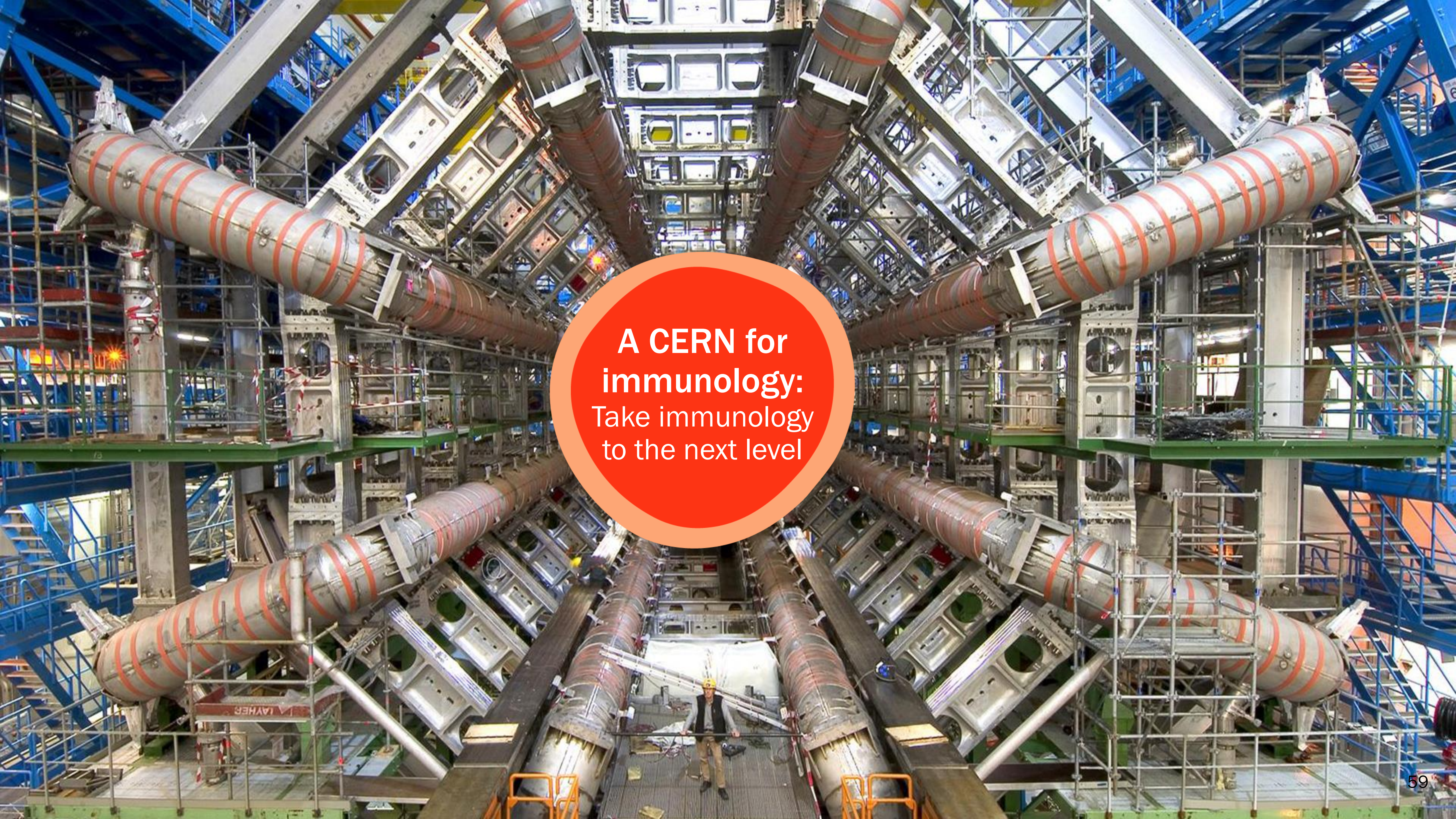
Within cell analysis

Integrated System Analysis

## Benefits

- Predict vaccine response
- Baseline immune state as a co-variate / predictor
- Estimate subpopulation structure for response
- Identify immune correlates of clinical phenotypes
- Identify drivers of immune variation

Human Immunome Project



## A CERN for immunology:

Take immunology to the next level

# Learn from humans to cure humans

CytoReason

TECHNION
Israel Institute of Technology

Shai Shen-Orr

shenorr@technion.ac.il

www.shenorrlab.technion.ac.il