Accelerating biomedicine through genomics and AI

Mehrtash Babadi
Institute Scientist
Director of Computational Methods
Data Sciences Platform
Broad Institute of MIT and Harvard

Fact I

Biomedicine is very rapidly becoming an informational field

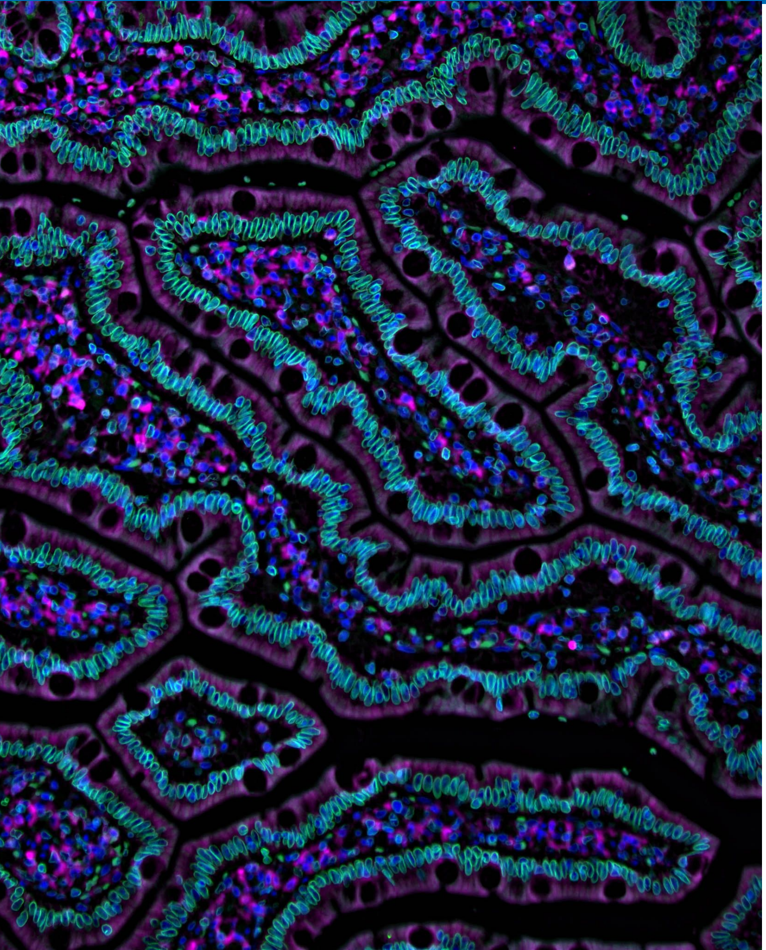
Fact II

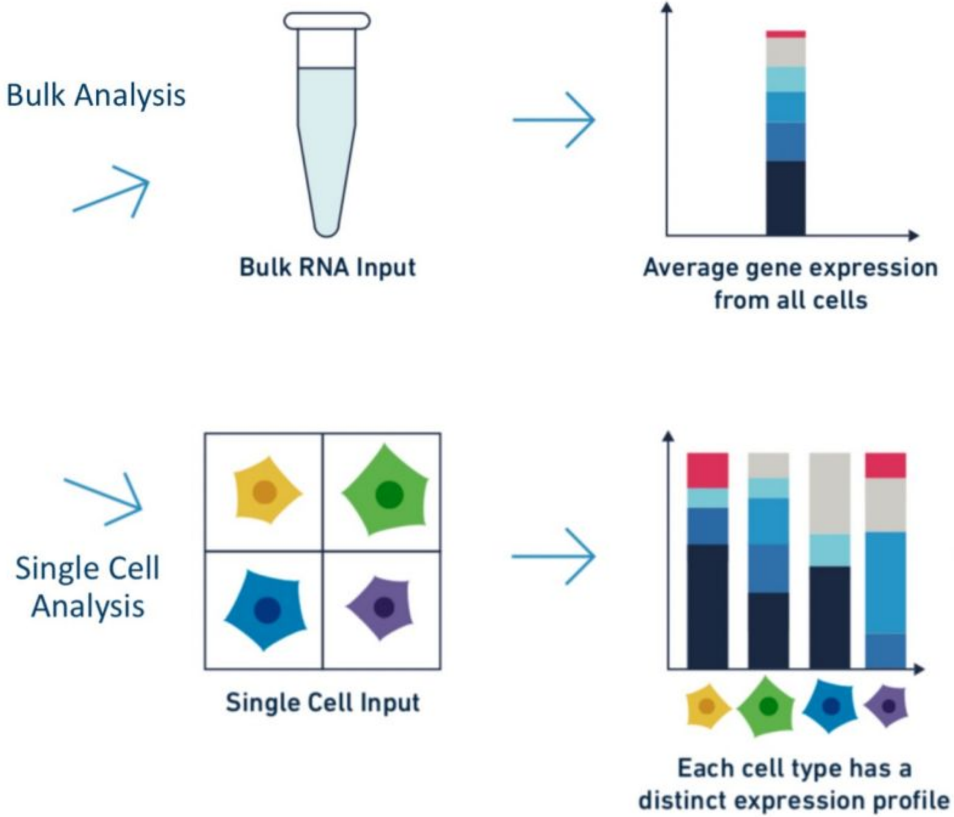
Artificial intelligence accelerates all informational fields alike

Fact I

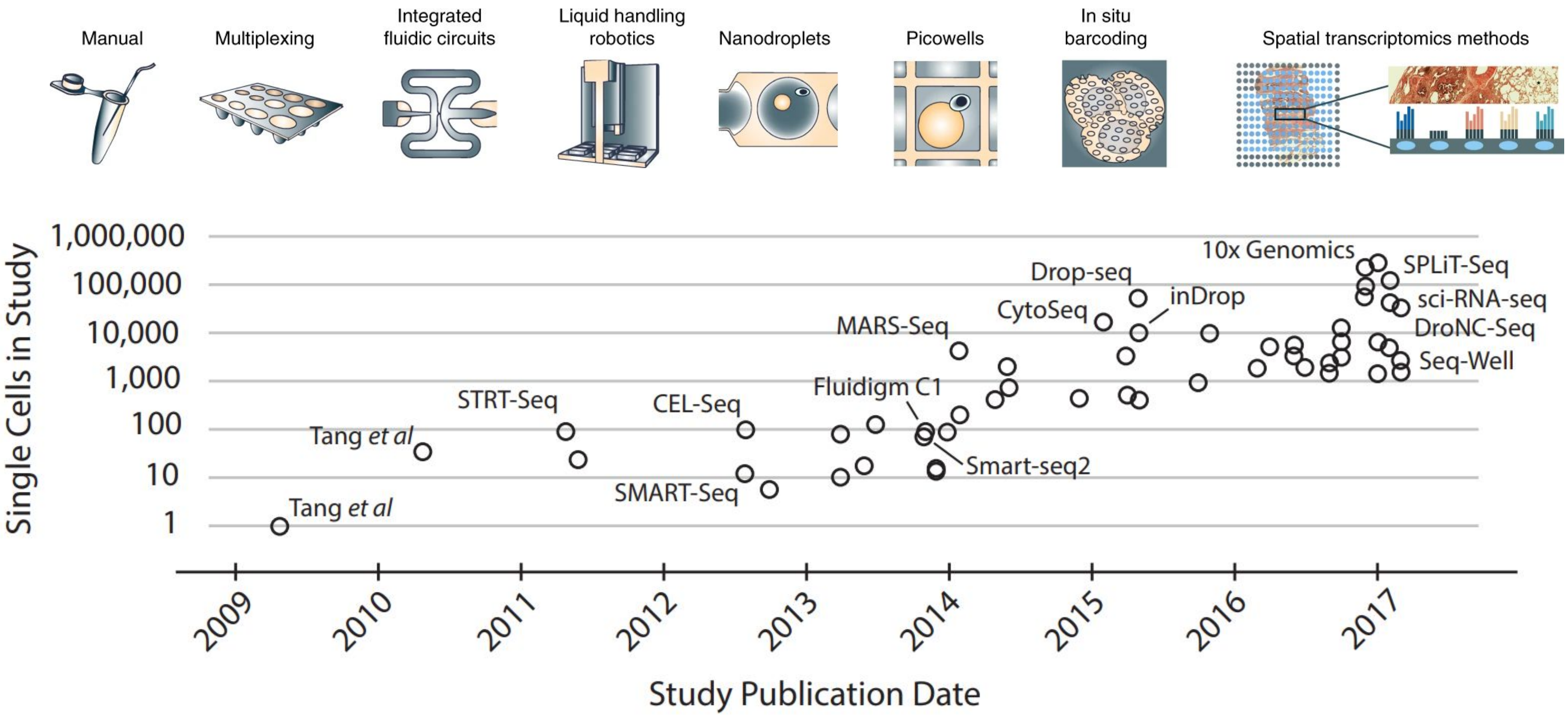
Biomedicine is very rapidly becoming an informational field

The revolution of single-cell biology

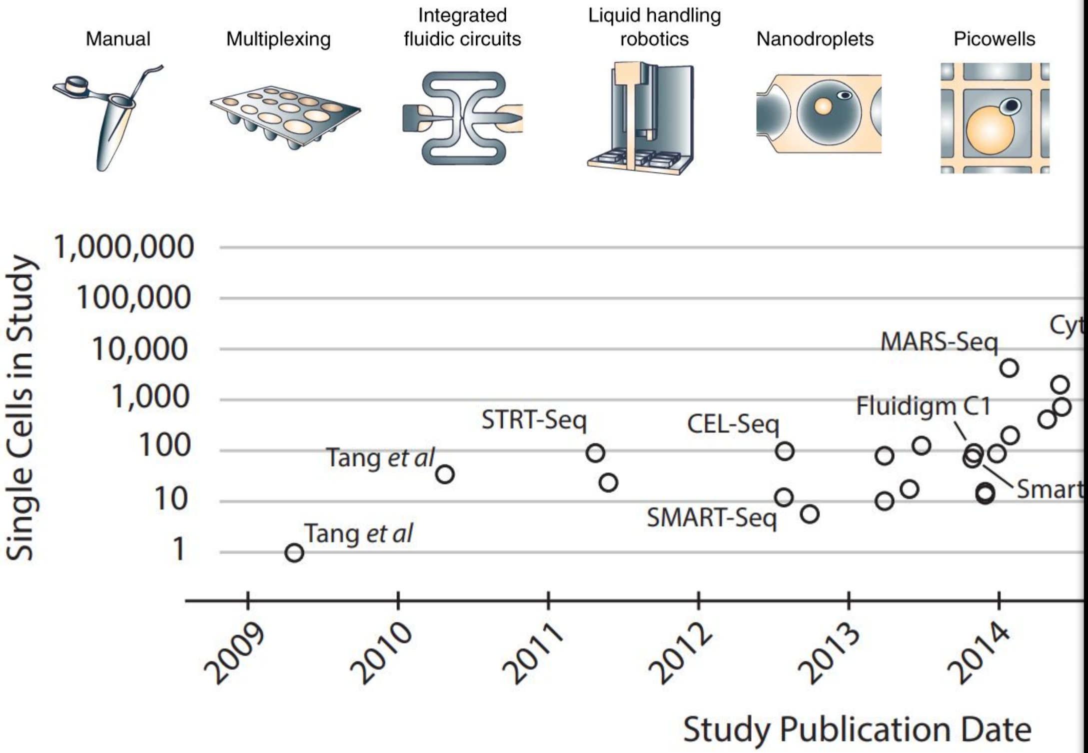


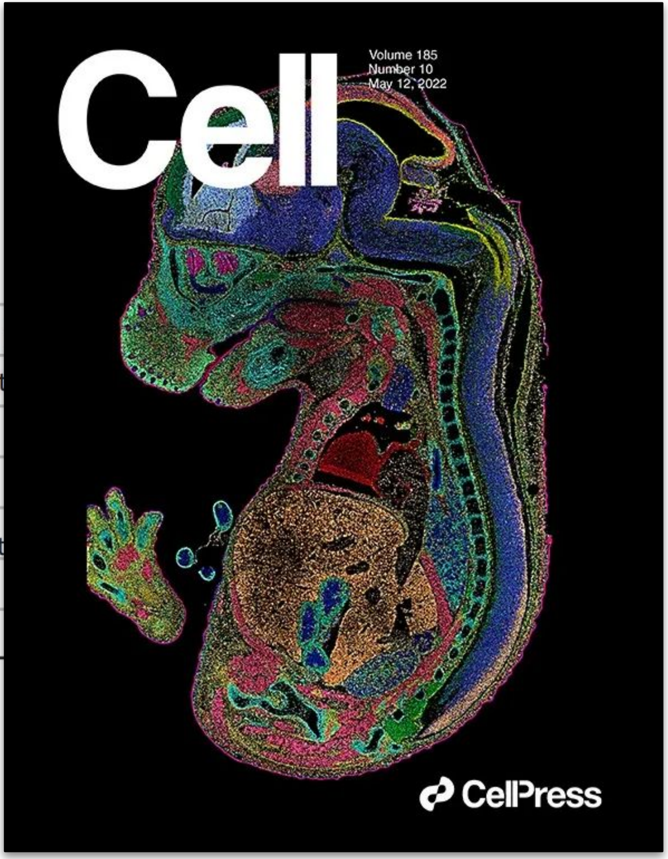


The throughput of single-cell sequencing increased by > **1,000,000 times** in a decade (2010-2020)

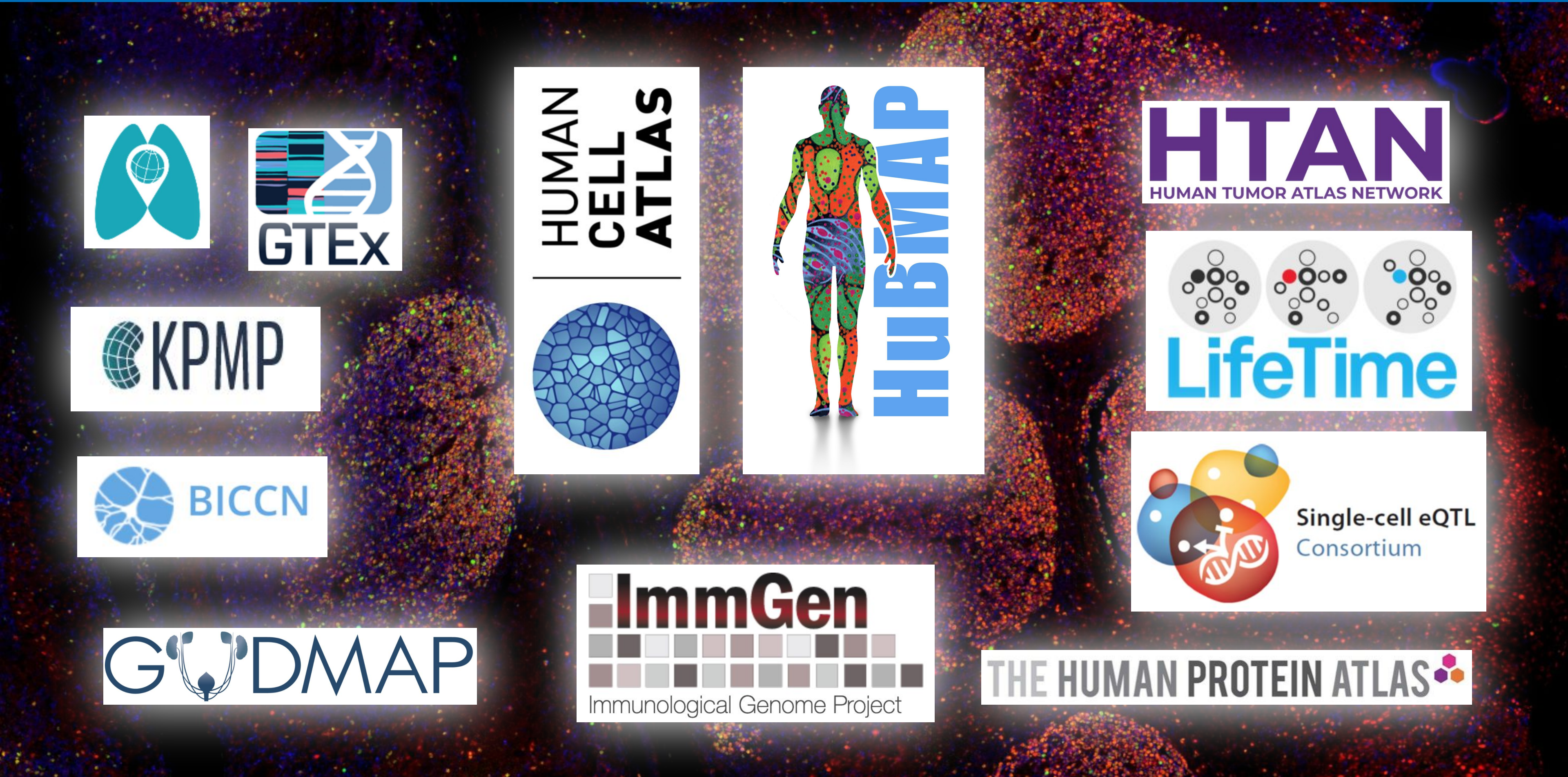


The throughput of single-cell sequencing increased by > **1,000,000 times** in a decade (2010-2020)





The rapid rise of single cell atlases



GTEx

KPMP

BICCN

GUDMAP

HUMAN CELL ATLAS

HuBMAP

ImmGen
Immunological Genome Project

HTAN
HUMAN TUMOR ATLAS NETWORK

LifeTime

Single-cell eQTL Consortium

THE HUMAN PROTEIN ATLAS

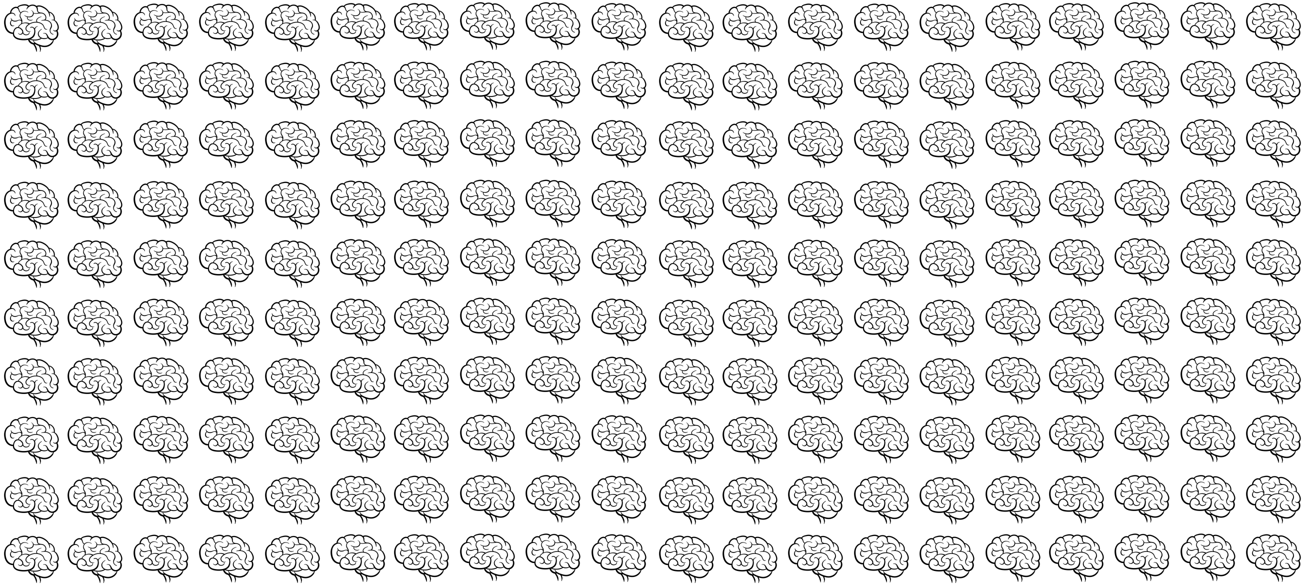
Exhibit — Broad Center for Human Brain Variation

BICAN Center for **Human Brain Variation**

200 human brains, 50 brain regions

20k-50k cells per region spatial gene expression and chromatin accessibility

200M-500M multimodal cell state measurements



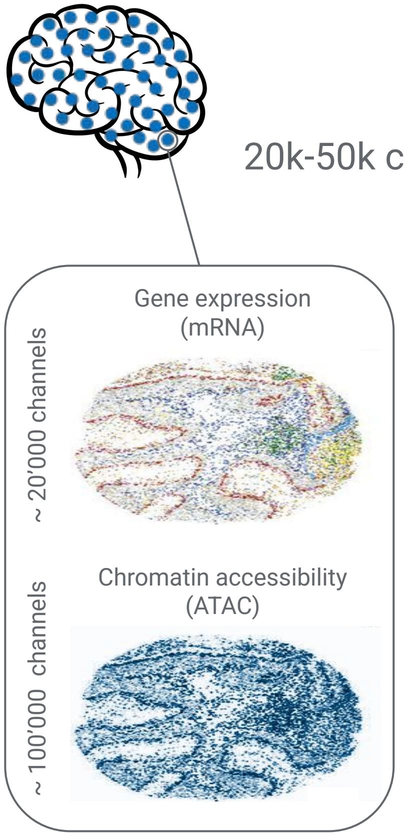
Steven McCarroll

co-PI

Evan Macosko

co-PI

Exhibit — Broad Center for Human Brain Variation



BICAN Center for **Human Brain Variation**
200 human brains, 50 brain regions
20k-50k cells per region spatial gene expression and chromatin accessibility
200M-500M multimodal cell state measurements

Key scientific questions:

- What is the nature of human brain variation across individuals at the resolution of individual cells and molecular processes?
- How genetics influences such variability?
- What are the molecular mechanisms underlying psychiatric, neurodegenerative, and aging-related diseases?

Steven McCarroll
co-PI

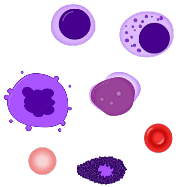
Evan Macosko
co-PI

NIH National Institute of Mental Health

Exhibit — Deorphanizing rare diseases through CRISPR perturbation experiments

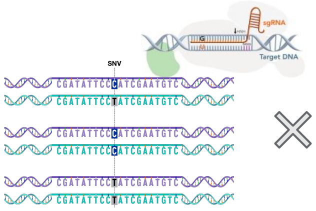
Project RADICAL (“**RA**re **DI**sease **C**ures **A**cce**L**erator”)

Phase 1 — Profiling and Identification of Nodal Pathways



10 hiPSC cell models

×



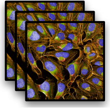
24,000 genomic variants

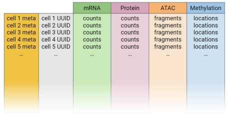
×

hypoxia
oxidative stress
ER stress
growth factors
…

1 ~ 5 conditions

Morphology



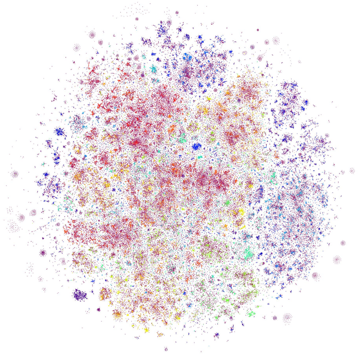


		mRNA	Protein	ATAC	Methylation
cell 1 meta	cell 1 UUID	counts	counts	fragments	locations
cell 2 meta	cell 2 UUID	counts	counts	fragments	locations
cell 3 meta	cell 3 UUID	counts	counts	fragments	locations
cell 4 meta	cell 4 UUID	counts	counts	fragments	locations
cell 5 meta	cell 5 UUID	counts	counts	fragments	locations
...	...	...	...	...	...

scRNA-seq

2 readouts

Representation Learning,
Predictive Modeling



10 billion single cells

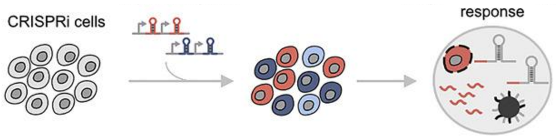
10’s or 100’s rare genetic diseases sharing the same impacted function (“deorphanization”)

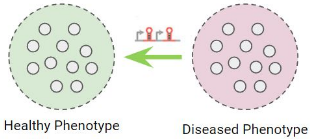
Anna Greka
co-PI

Stacey Gabriel
co-PI

Phase 2 — Identification of Therapeutic Targets via Perturbation Screens

200 genome-wide perturbation screens for a representative cell line x variant per nodal pathway





Fact II

Artificial intelligence accelerates all informational fields alike

AI is changing the way we conduct research

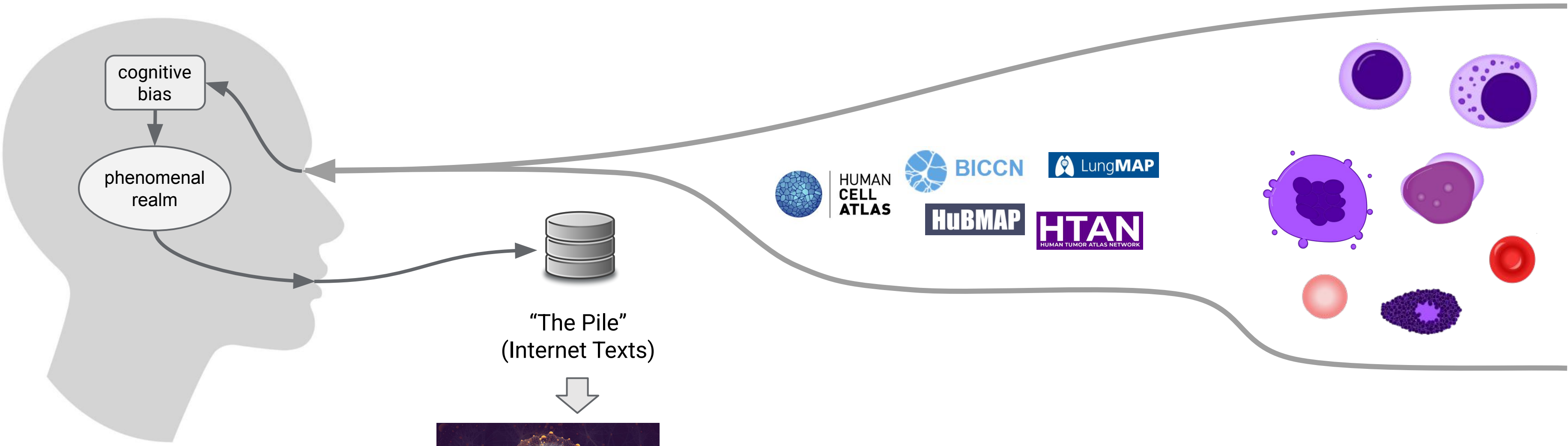
- As more data is generated, conducting biomedical research becomes **increasingly more complex**

- Our traditional analysis methods do not scale to the pace of data generation
    - No single measurement can fully capture the complex and high-dimensional reality of cellular states
    - No single dataset is sufficient to satisfactorily evaluate a complex hypothesis
    - No single human expert can fully intuit and internalize the patterns lying within high-throughput biological measurements

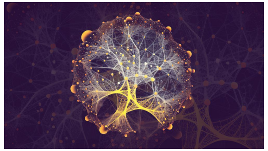
- It is only a matter of time before we generate enough data across all **spatiotemporal scales**, **molecular modalities**, and **contexts** to **"contain the answer"** to our most complex biological questions

- This trajectory parallels all informational fields
    - ***Internet*** *→ Large volume of texts, images, movies, etc.*
      → Development of increasingly capable ***AI models/agents of natural language***
    - ***Genomics*** *→ Large volume of genomes, transcriptomes, tissue slides, etc.*
      → Development of increasingly capable ***AI models/agents of biomedicine***

Beyond natural language: *learning the intrinsic language of cells*



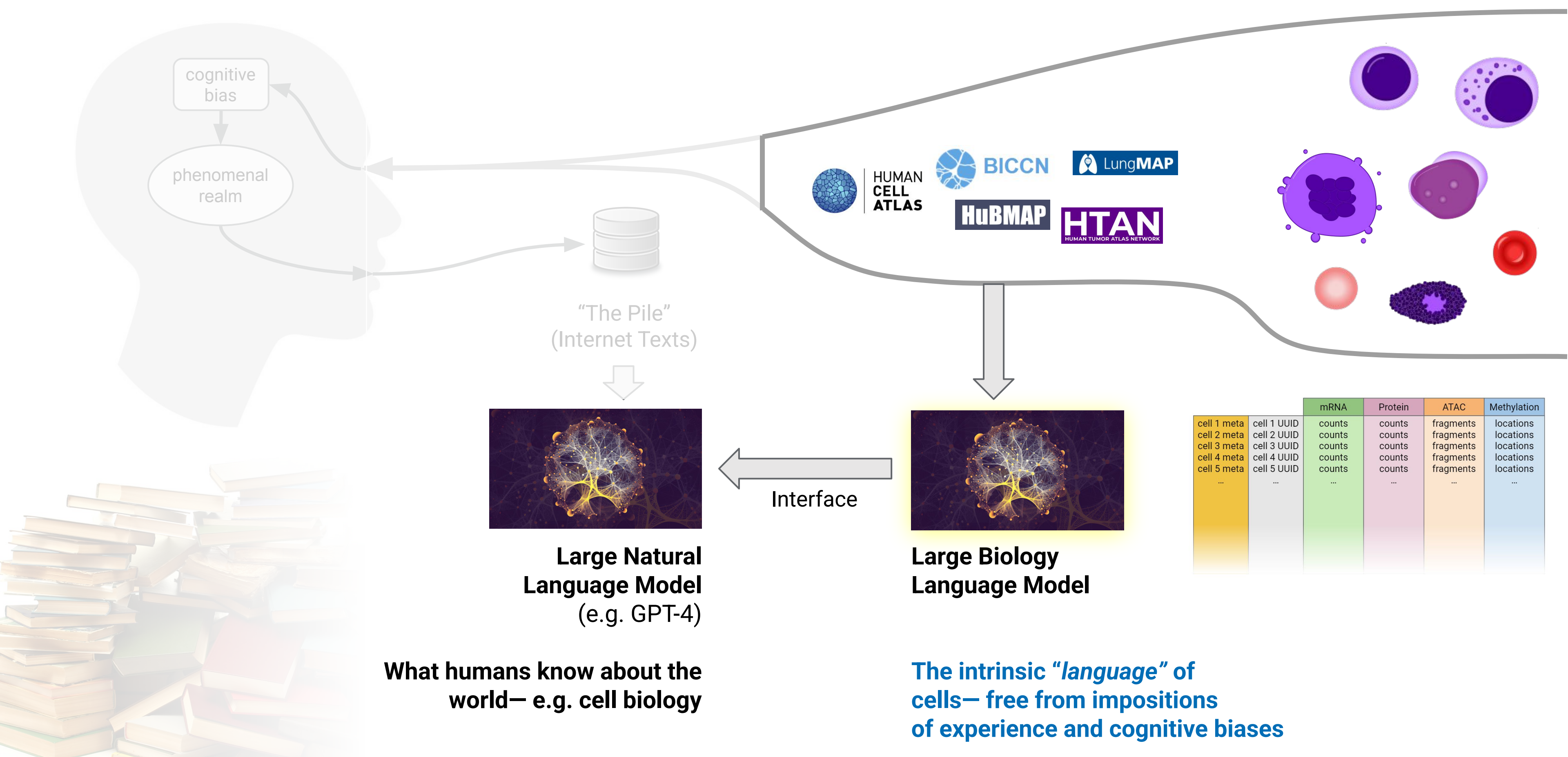
“The Pile”
(Internet Texts)



Large Natural Language Model
(e.g. GPT-4)

What humans know about the world— e.g. cell biology

Beyond natural language: *learning the intrinsic language of cells*



		mRNA	Protein	ATAC	Methylation
cell 1 meta	cell 1 UUID	counts	counts	fragments	locations
cell 2 meta	cell 2 UUID	counts	counts	fragments	locations
cell 3 meta	cell 3 UUID	counts	counts	fragments	locations
cell 4 meta	cell 4 UUID	counts	counts	fragments	locations
cell 5 meta	cell 5 UUID	counts	counts	fragments	locations
...	...	...	...	...	...

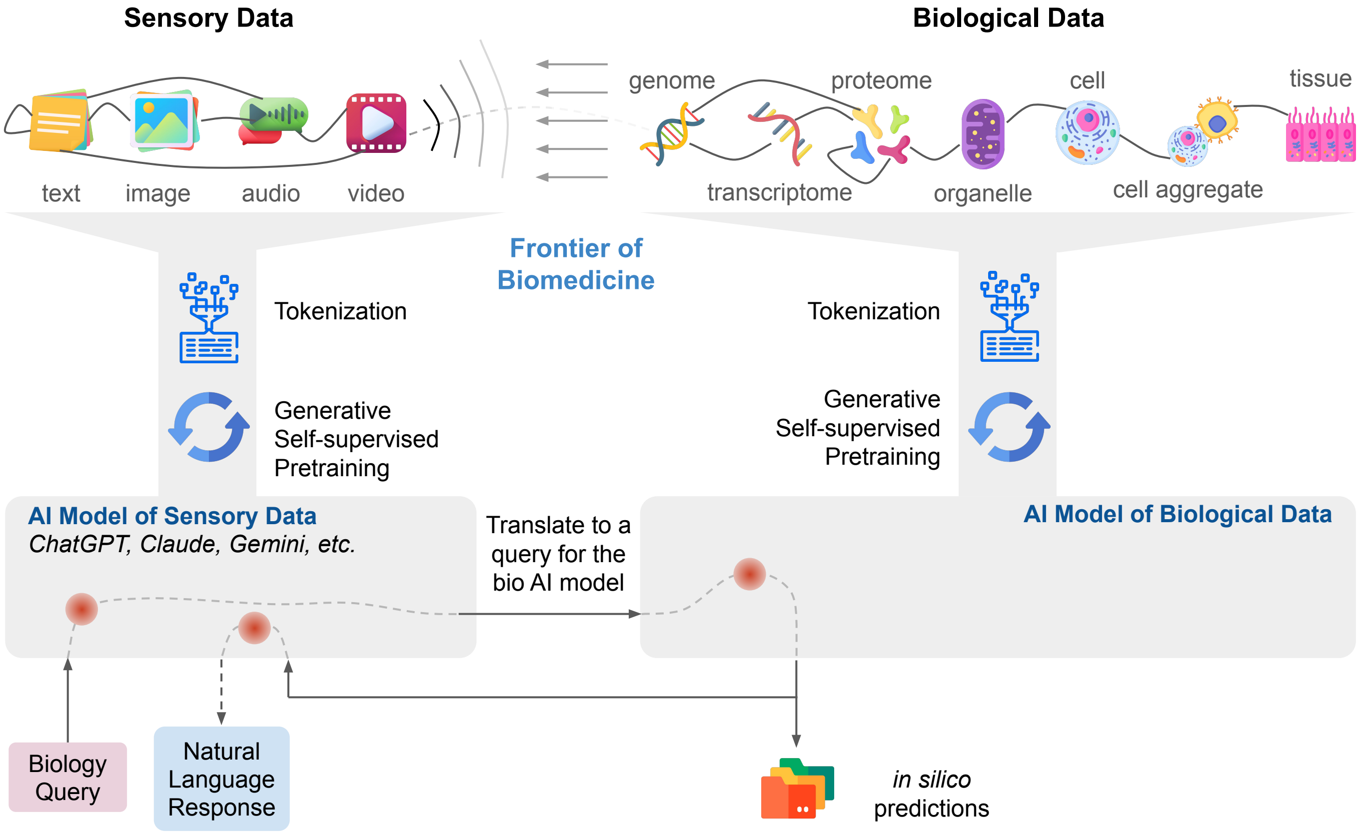
Large Natural Language Model
(e.g. GPT-4)

Large Biology Language Model

What humans know about the world— e.g. cell biology

The intrinsic "*language*" of cells— free from impositions of experience and cognitive biases

We envision a future where biomedical research is largely driven by AI research assistants and scientists



Sensory Data

text image audio video

Biological Data

genome proteome cell tissue

transcriptome organelle cell aggregate

Frontier of Biomedicine

Tokenization

Generative Self-supervised Pretraining

Tokenization

Generative Self-supervised Pretraining

AI Model of Sensory Data
ChatGPT, Claude, Gemini, etc.

Translate to a query for the bio AI model

AI Model of Biological Data

Biology Query

Natural Language Response

in silico predictions

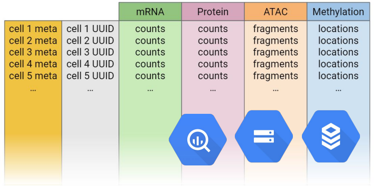
Broad is building an AI platform for cell biology research

Cellarium AI Platform

A cloud-native data and machine learning operations infrastructure for single-cell biology data inspired by modern AI practices

- **storage** and **querying** of billion-scale single-cell omics data
- rapid model **development**, model **training**, model **serving**, and building/serving **applications**

Cellarium
Single-Cell Cloud DataStore



		mRNA	Protein	ATAC	Methylation
cell 1 meta	cell 1 UUID	counts	counts	fragments	locations
cell 2 meta	cell 2 UUID	counts	counts	fragments	locations
cell 3 meta	cell 3 UUID	counts	counts	fragments	locations
cell 4 meta	cell 4 UUID	counts	counts	fragments	locations
cell 5 meta	cell 5 UUID	counts	counts	fragments	locations
...	...	...	...	...	...

Cellarium
Machine Learning Library

distributed anndata
distributed model training
model zoo
(more)

PyTorch Lightning

Cellarium
Autoscaling Microservices

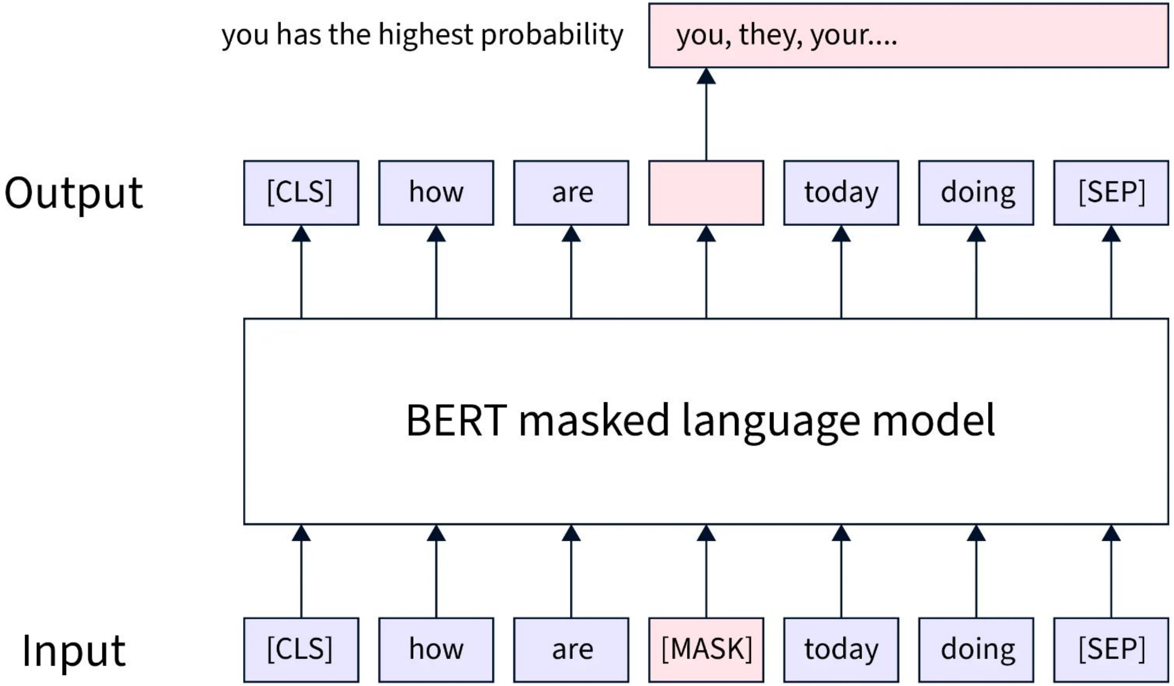
authentication
monitoring
model serving
vector search
(more)

neptune.ai

Cellarium
Applications

CellariumCAS
CellariumGPT
(more)

Large language models (of natural language)



you has the highest probability

you, they, your....

Output

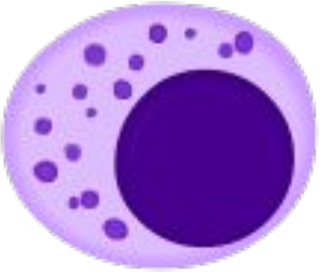
[CLS] how are today doing [SEP]

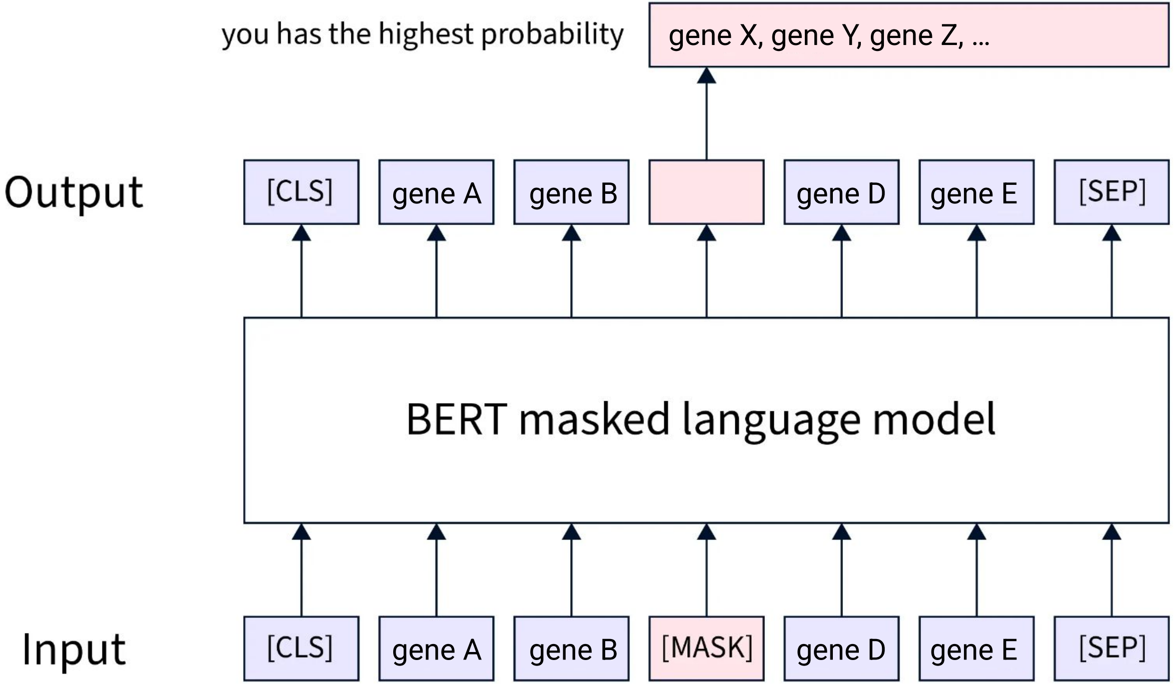
BERT masked language model

Input

[CLS] how are [MASK] today doing [SEP]

Large language models (of cell biology)





you has the highest probability

gene X, gene Y, gene Z, ...

Output

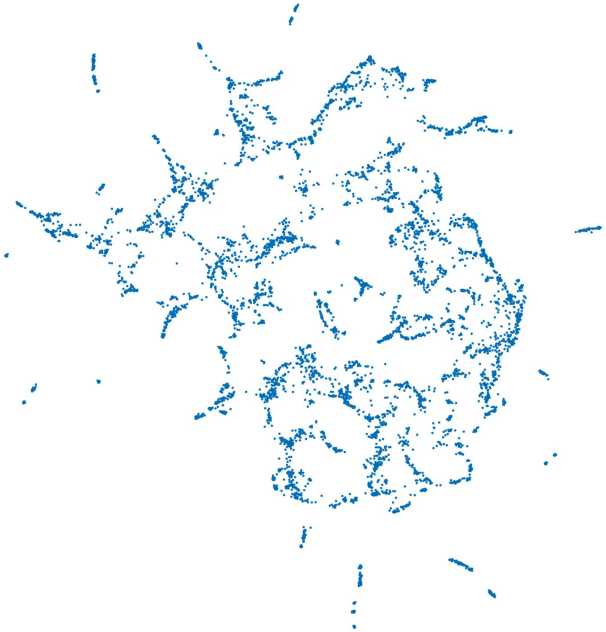
[CLS] gene A gene B gene D gene E [SEP]

BERT masked language model

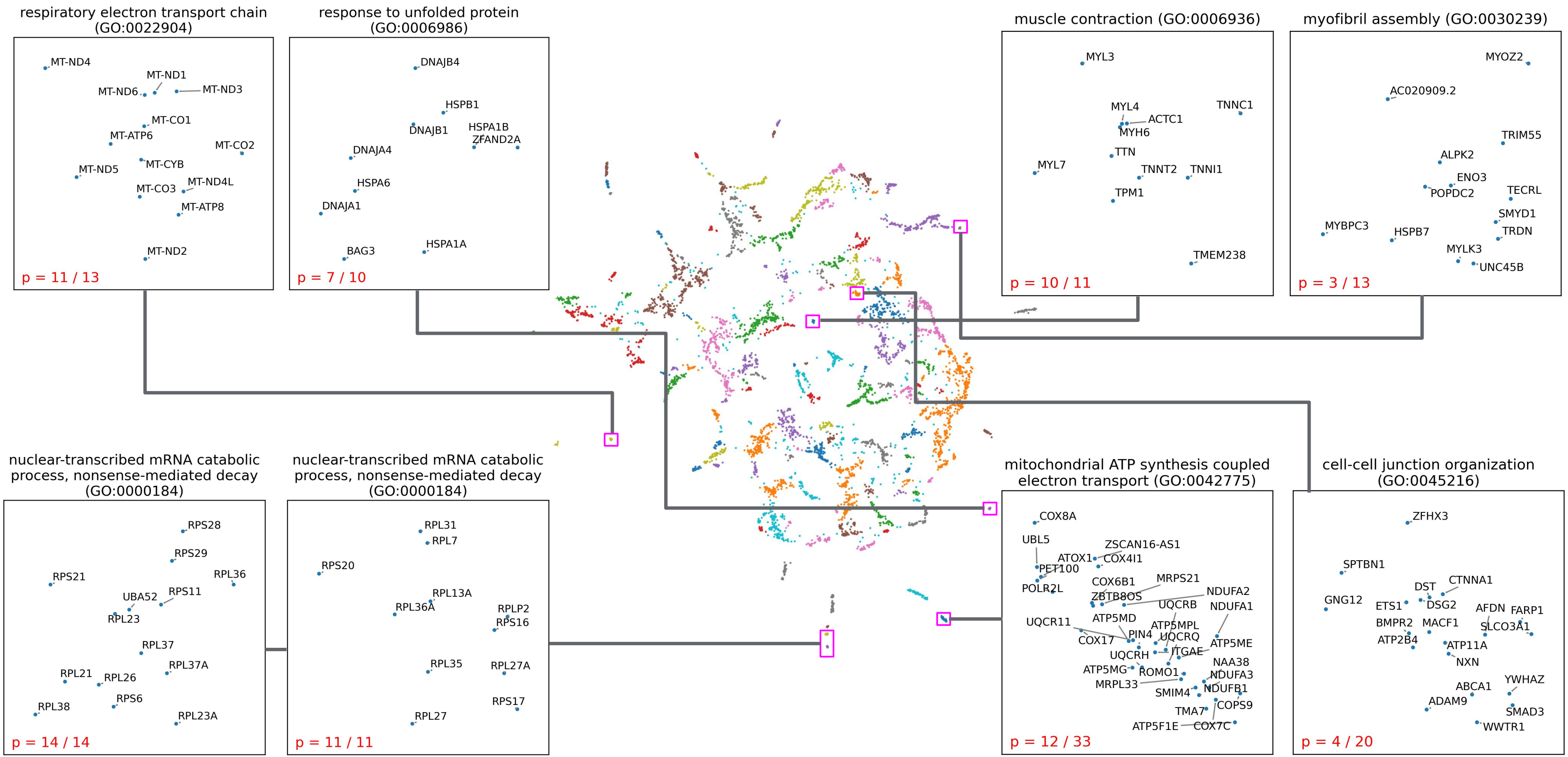
Input

[CLS] gene A gene B [MASK] gene D gene E [SEP]

CellariumGPT reveals cell-type specific gene function



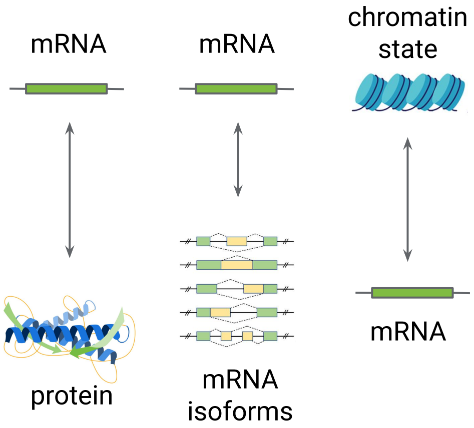
CellariumGPT reveals cell-type specific gene function



We envision a future where biomedical research is largely driven by AI research assistants and scientists

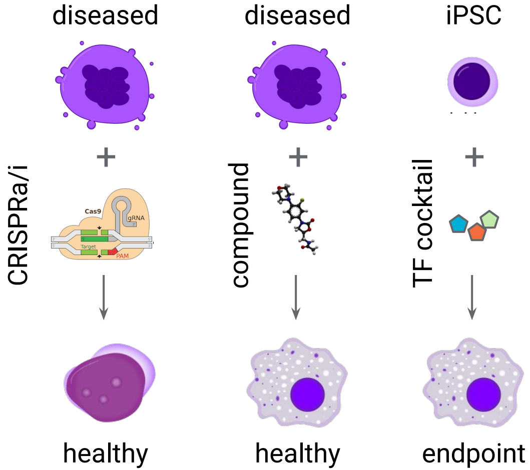
Data Completion

mRNA → protein
chromatin accessibility → mRNa
mRNA → mRNA isoforms



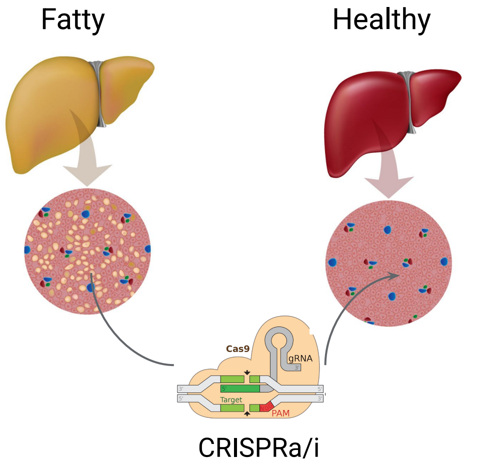
Modifying Cellular States

diseased state + compound/CRISPR → healthy state
pluripotent stem cell + TF cocktail → desired endpoint



Tissue Remodelling

diseased tissue phenotype + compound/CRISPR
→ healthy tissue phenotype



AI models of the **intrinsic language of cells and tissues** will allow us to perform many more experiments and test many more hypotheses *in silico*

We envision a future where biomedical research is largely driven by AI research assistants and scientists

Genomics and AI are accelerating every stage of the drug development process

Discovery & Development

- Precise **identification** of therapeutic targets
- Precise **design** of therapeutic agents
    - small molecules
    - antibodies
    - peptides

Preclinical Research

- Comprehensive ***in silico* evaluation** of therapeutic agents
    - off-target effects
    - pharmacodynamics
    - toxicity
    - clinical benefit

Clinical Research

- Significantly improved design of clinical trials through **personal genomics**
    - patient stratification
    - heterogeneity of clinical benefit