Discovery of cellular reprogramming methodology through single-cell foundation models

Nick Galioto

Department of Computational Medicine and Bioinformatics, University of Michigan

Frontiers in Scientific Machine Learning

November 1, 2024

Collaborators

Faculty and staff:

Indika Rajapakse

Alex Gorodetsky

Lindsey Muir

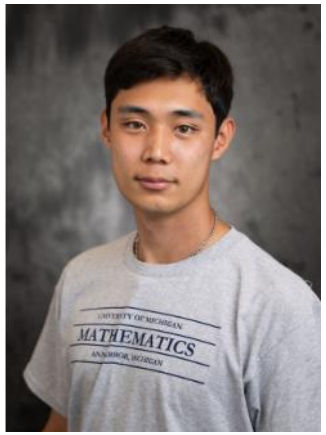
Walter Meixner

PhD students:

Cooper Stansbury

Joshua Pickard

Jillian Cwycyshyn

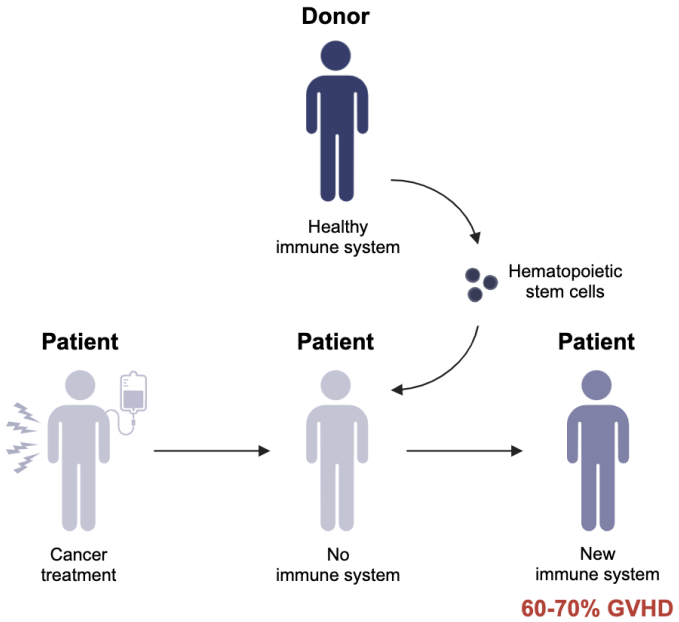


Marc Choi

The ultimate goal: my cells, my cure!

Donnall Thomas

Invented Bone-marrow Transplant
Fred Hutchinson Cancer Center
1990 Nobel Prize in Medicine

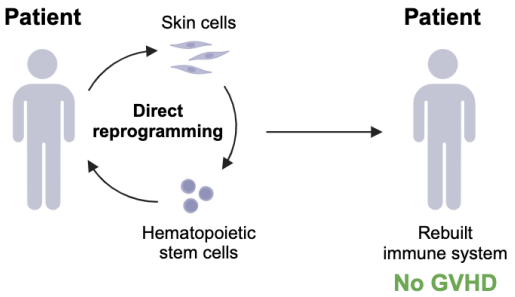


PROBLEM

GVHD is when the patient cells attack the donor cells

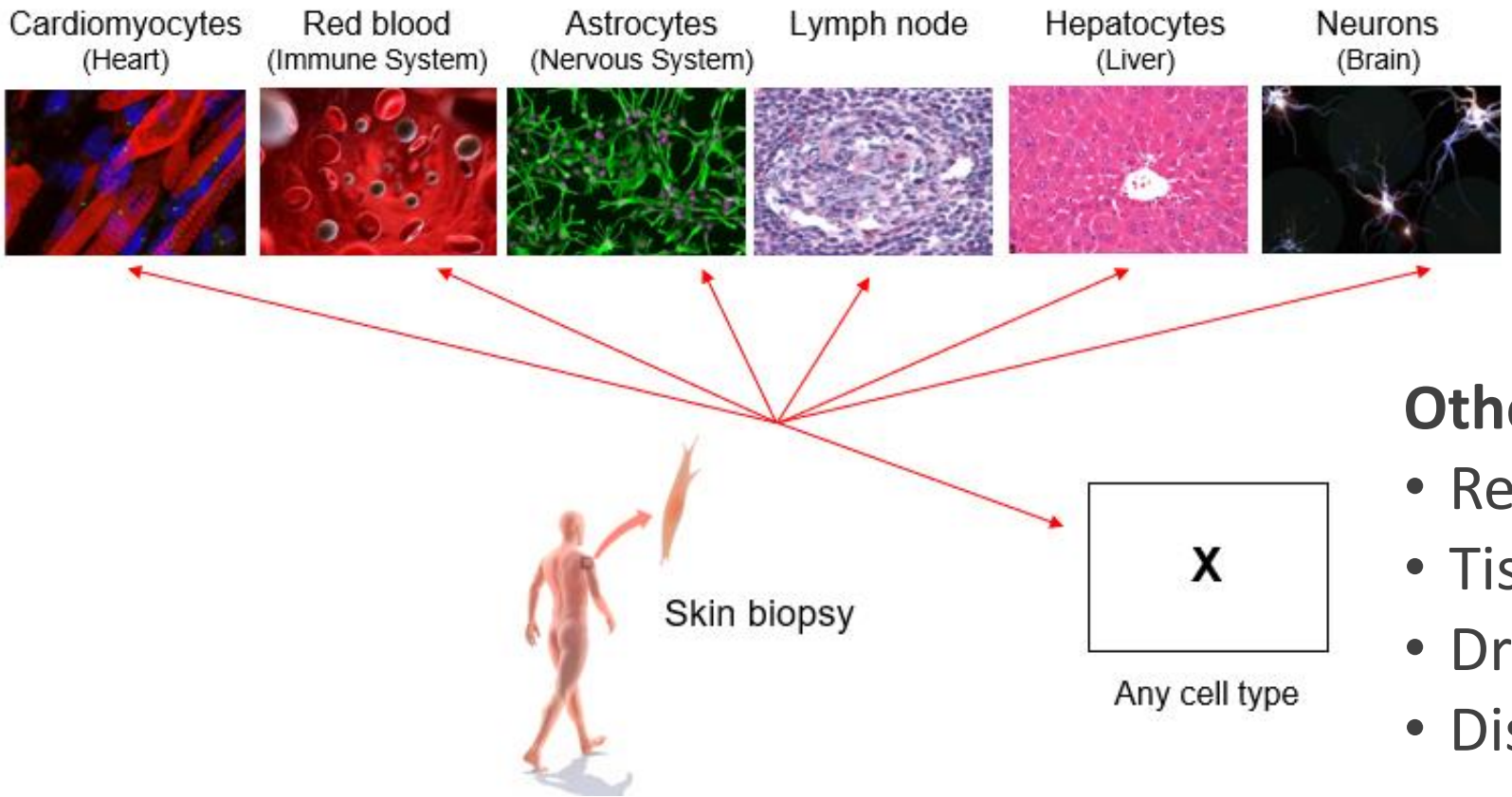
SOLUTION

Autologous cell reprogramming



Bone-marrow Transplant is the **Treatment for the Treatment**

Cell reprogramming



Other applications:

- Replenish immune system
- Tissue regeneration
- Drug discovery
- Disease modeling

Challenges in discovering reprogramming methodologies

Time

- Experiments take several weeks

Money

- 1 experiment $\geq \$15{,}000$

Idea: Can we do *in silico* experimentation?

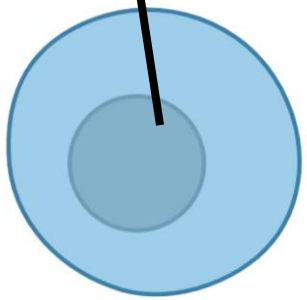
Outline

- Flow of genetic information
- Cell reprogramming
- Digital biology
- Geneformer
- Cell cycle dynamics
- Conclusions

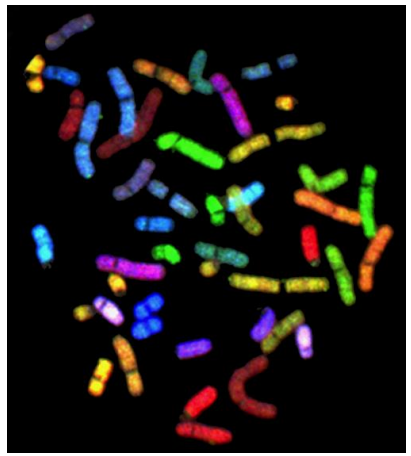
Flow of genetic information

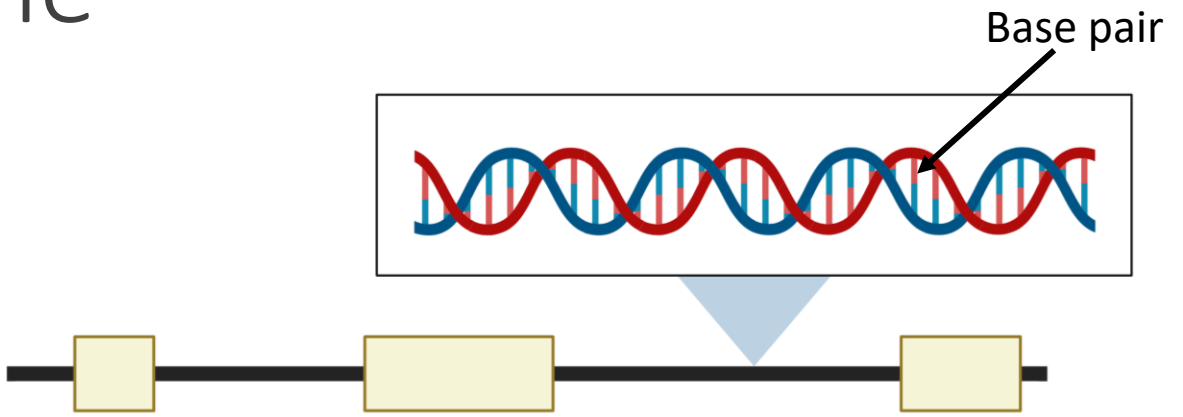
The human genome

Nucleus: contains the DNA

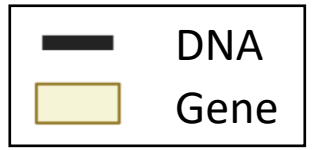


Chromosomes:





Genes are specific segments of the DNA

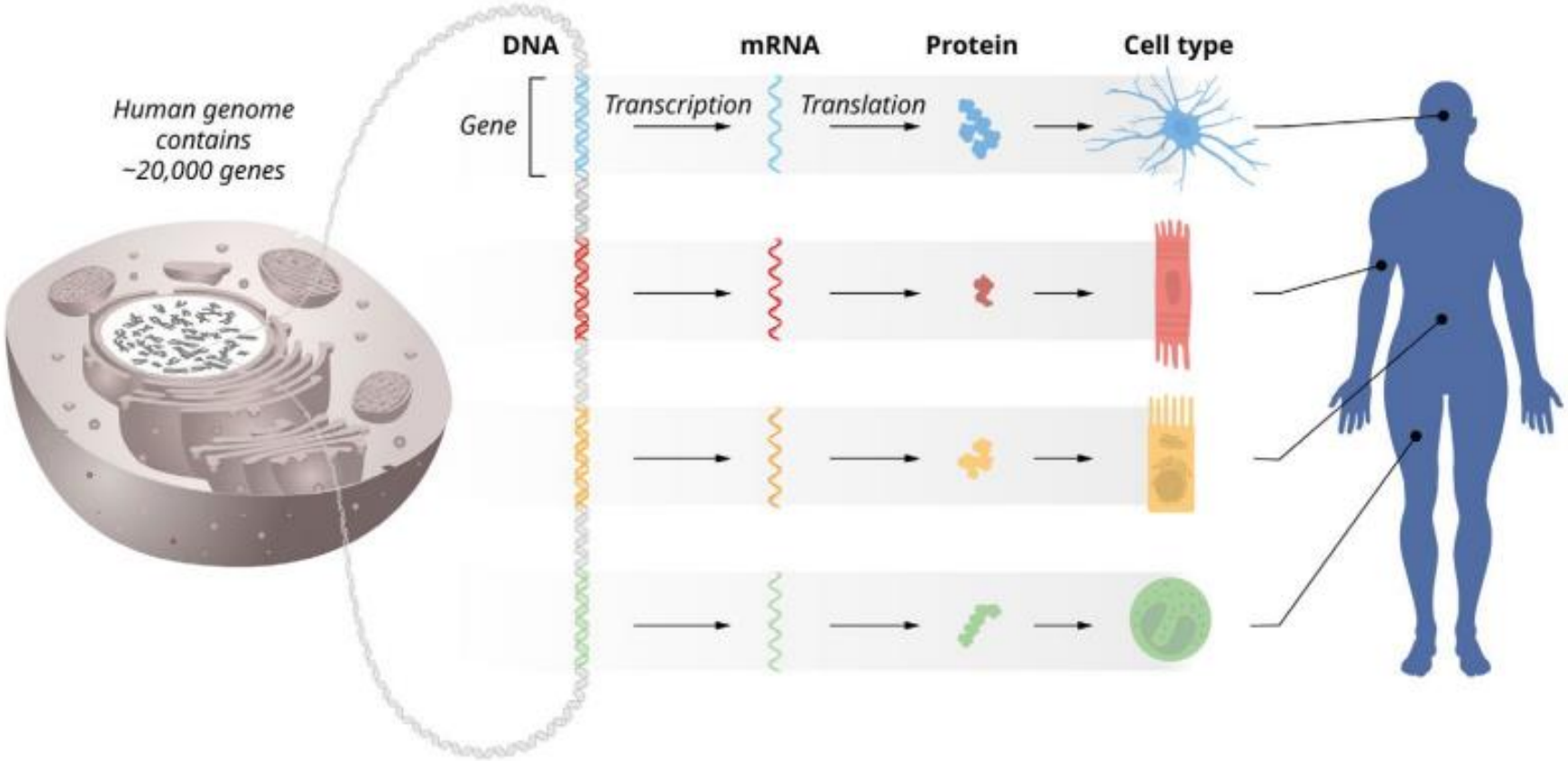


Full length of human genome (30B base pairs)

Total length of genes (50M base pairs)

DNA makes RNA, RNA makes protein

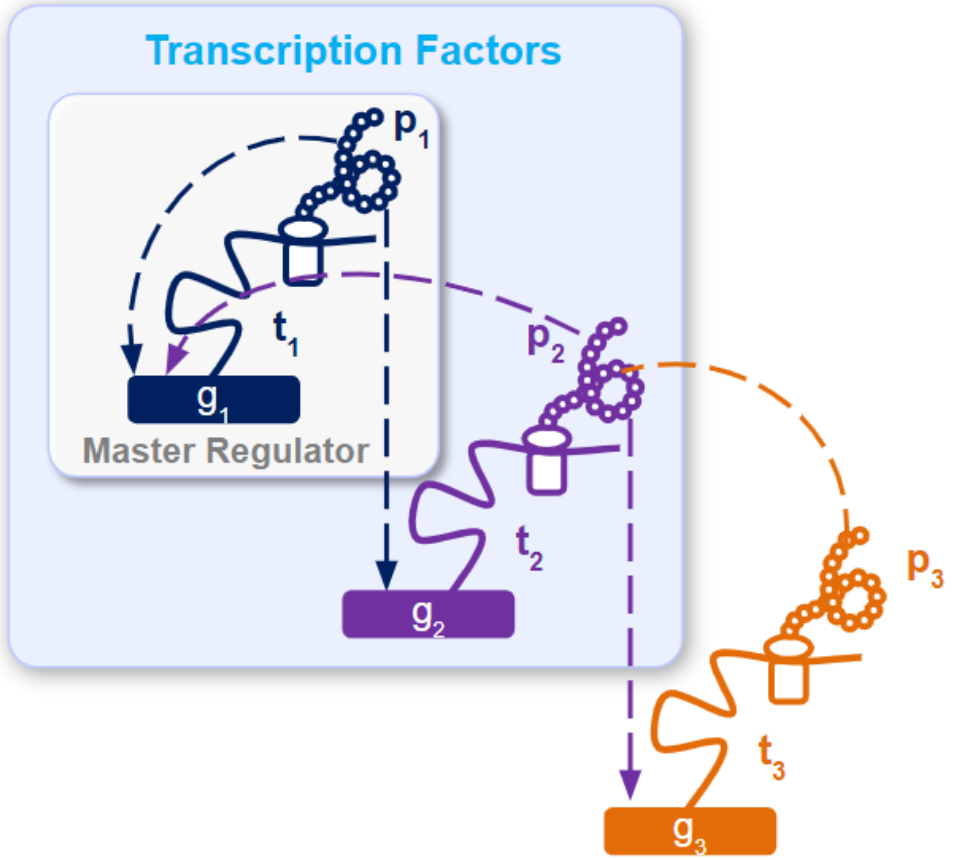
Proteins are responsible for nearly all cellular functions



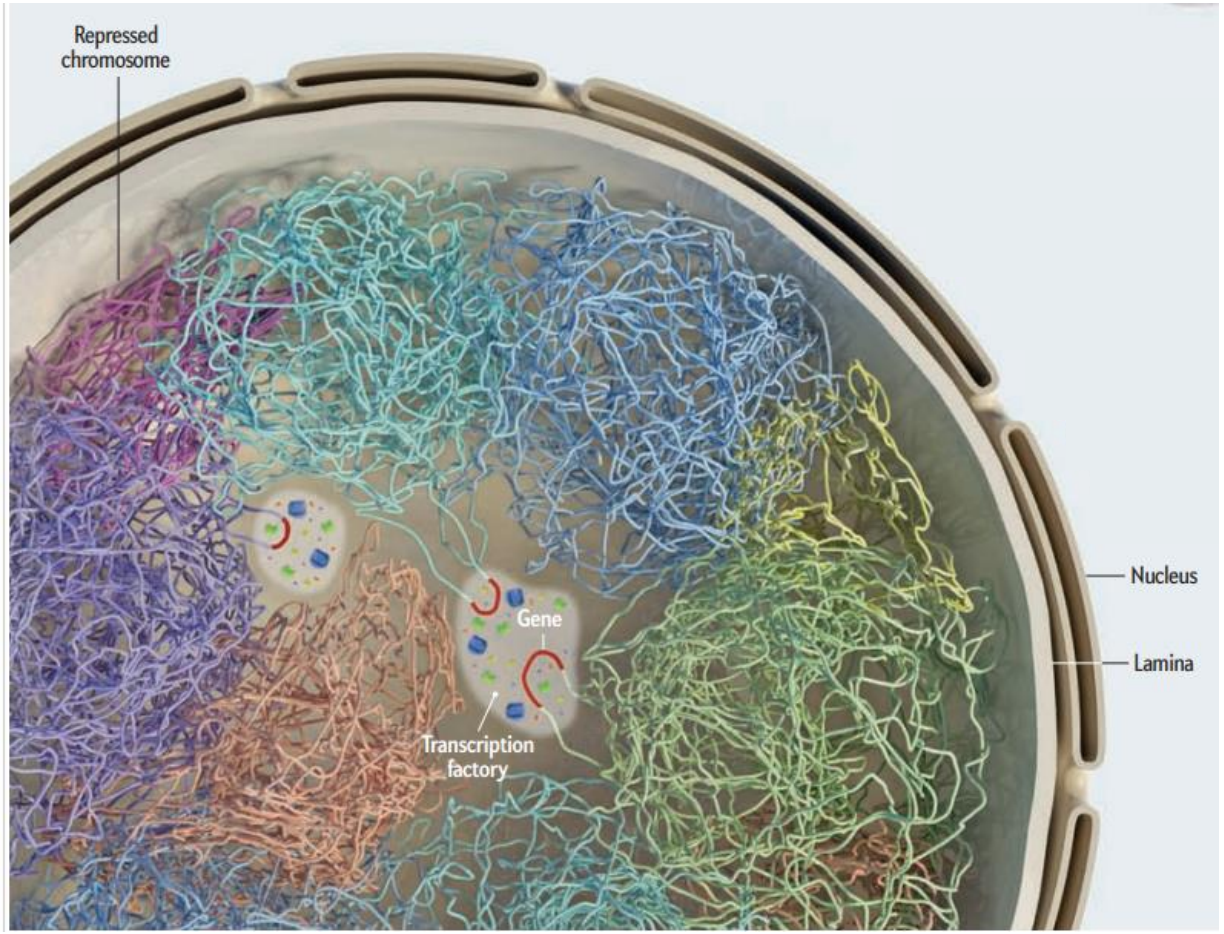
ngalioto@umich.edu

Transcription factors

- Transcription factors (TFs) are proteins that can turn RNA transcription on and off at certain genes
- ‘Master regulators’ can influence their own synthesis

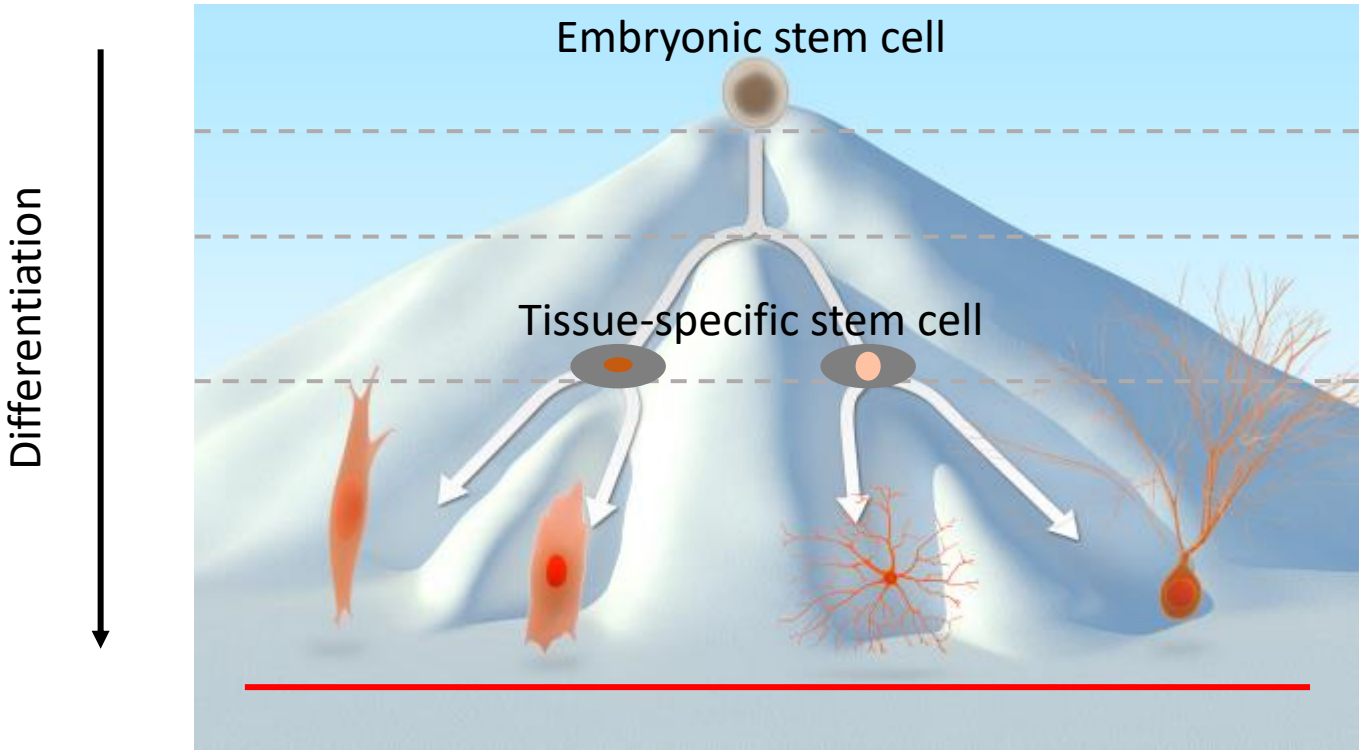


Chromatin organization: how DNA is packed within the nucleus



Cell reprogramming

Typical cell development



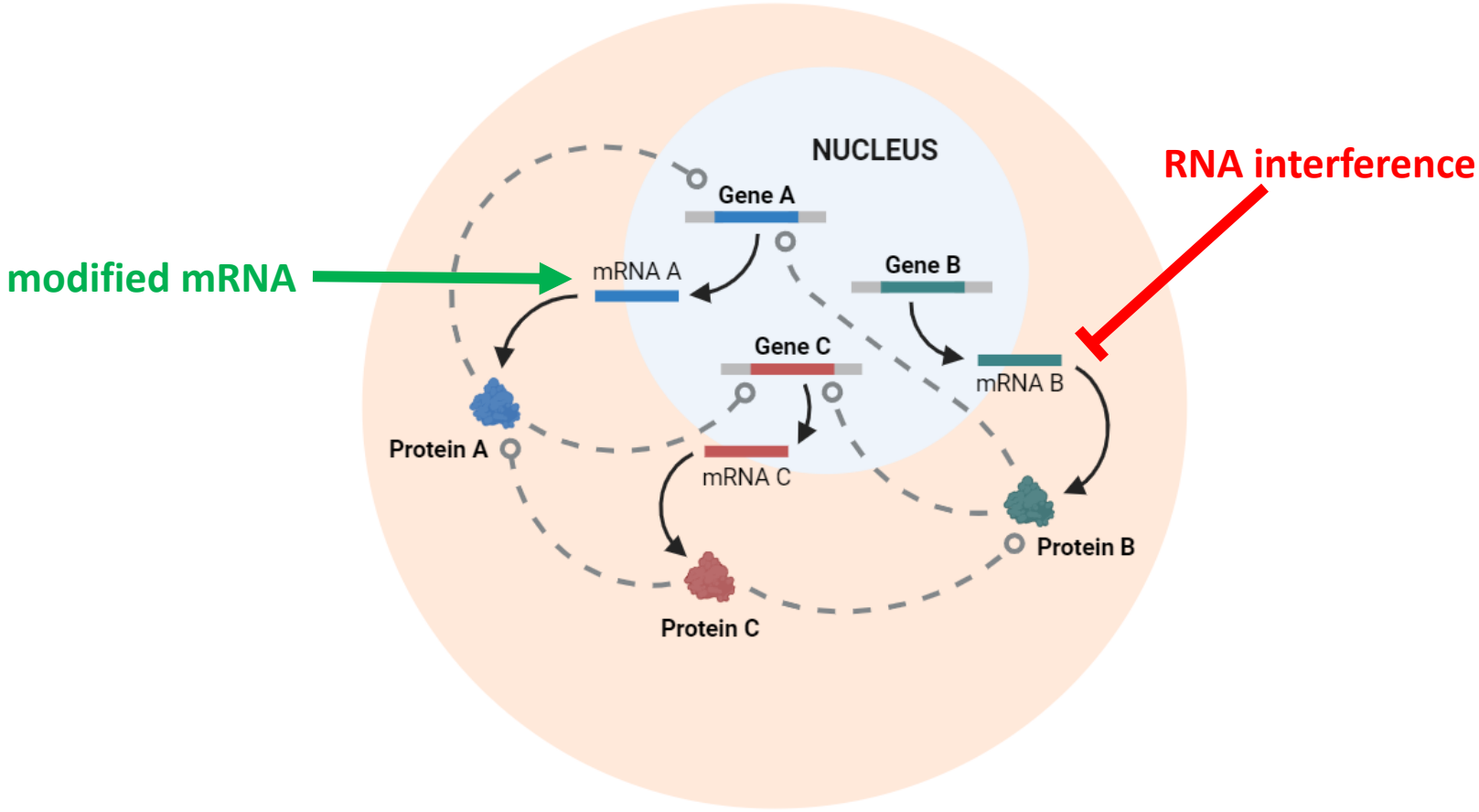
All cell types in the human body (estimated to be about 300)

Cell reprogramming

Shinya Yamanaka: iPSC reprogramming (INDIRECT)
2006: Nobel Prize: 2012

Harold Weintraub: DIRECT Reprogramming
1989: (1945-1995)

Controlling the flow of information

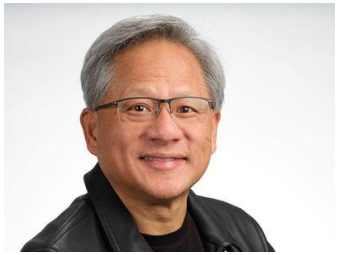


Number of Genes in the Human Genome = 20,000
Number of **Transcription Factors** = 1,800
Number of Master Regulators (subset of Transcription Factors) = 800

Digital biology

Digital biology

"Where do I think the next amazing revolution is going to come? And this is going to be flat out one of the biggest ones ever. There's no question that digital biology is going to be it."

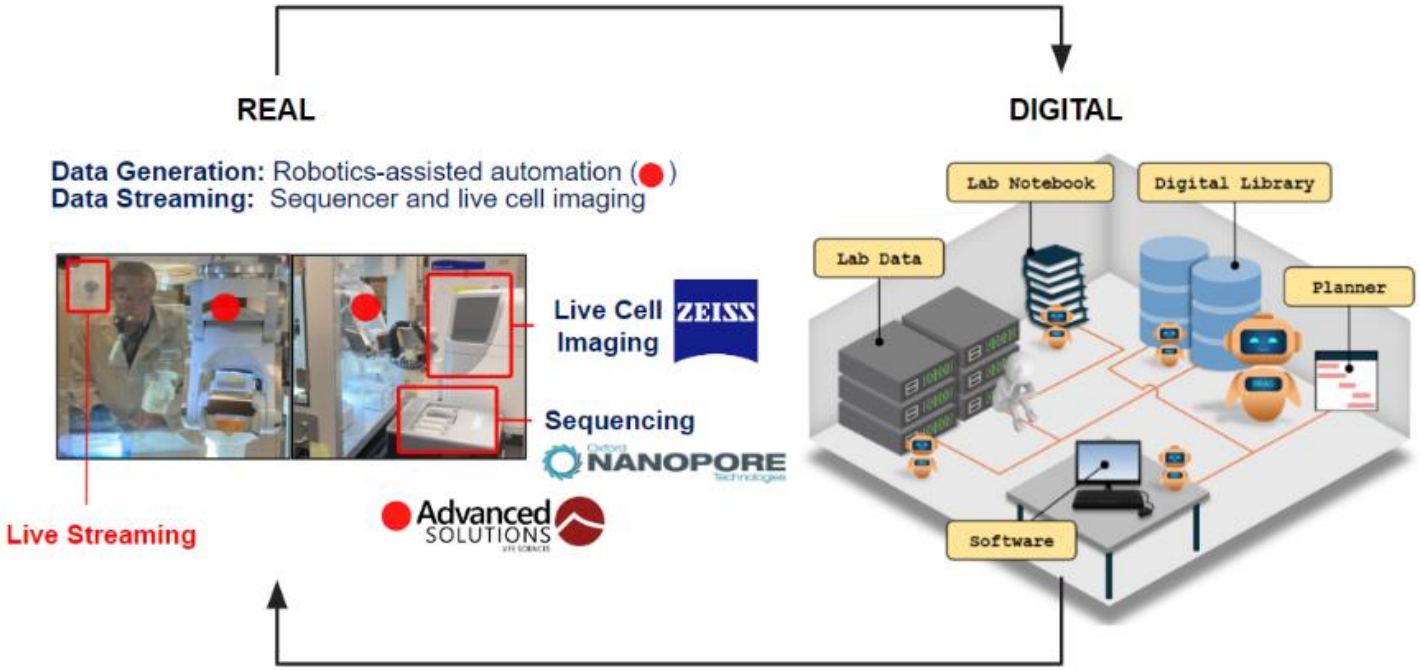


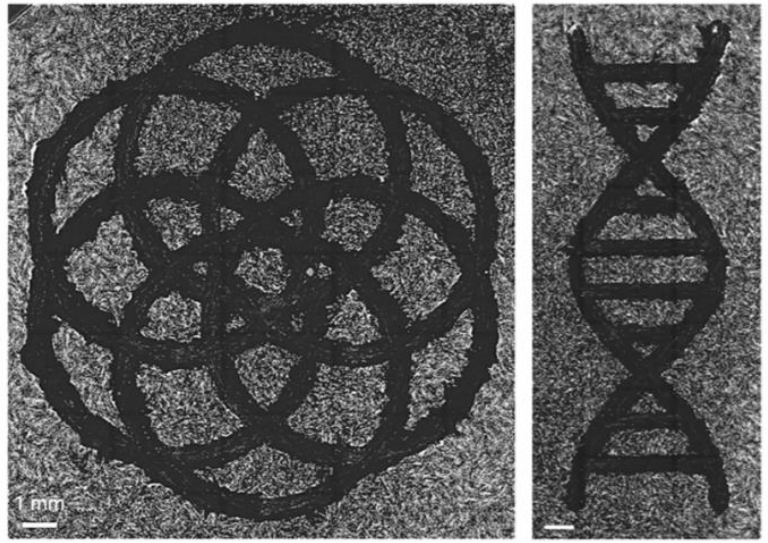
-Jensen Huang, founder & CEO of NVIDIA

- Exponential increase in data availability
  - 1990: 13 years and $3B to sequence human genome
  - Today: 1 day and $600
- Accelerated ability to read, write, and edit DNA
- Technology enabling robotic/automated labs
- Advancements in deep learning and AI
  - AlphaFold can now design novel proteins

BioAssemblyBot

Digital laboratory



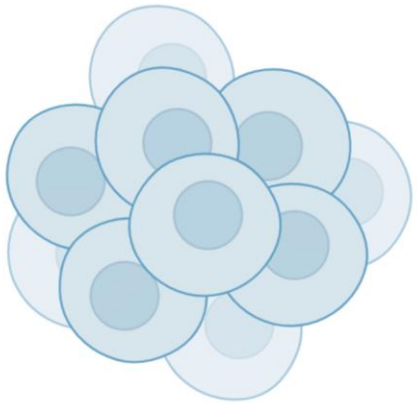


A digital assistant for biology:
Pickard, Joshua, et al. “Language model powered digital biology." arXiv preprint arXiv:2409.02864 (2024).

Automated wound generation:
Cwycyshyn, Jillian, et al. "A programmable platform for probing cell migration and proliferation." APL Bioengineering 8.4 (2024).

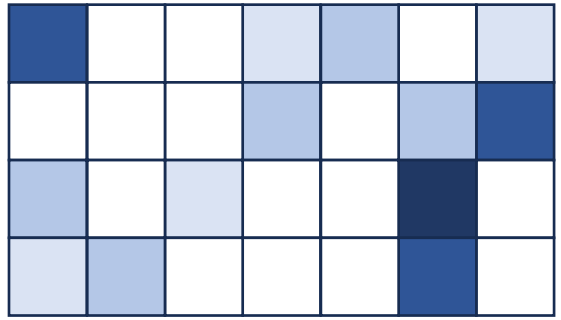
Single-cell RNA sequencing

Counts the number of RNA molecules produced by each gene



Collection of cells

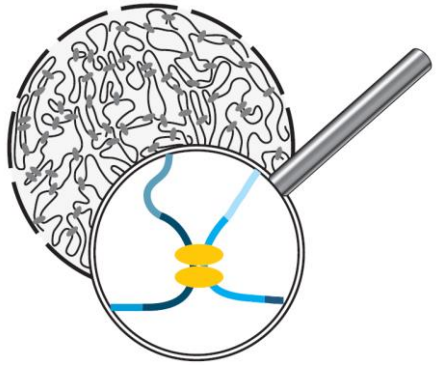
Oxford Nanopore
PromethION 2 Solo™

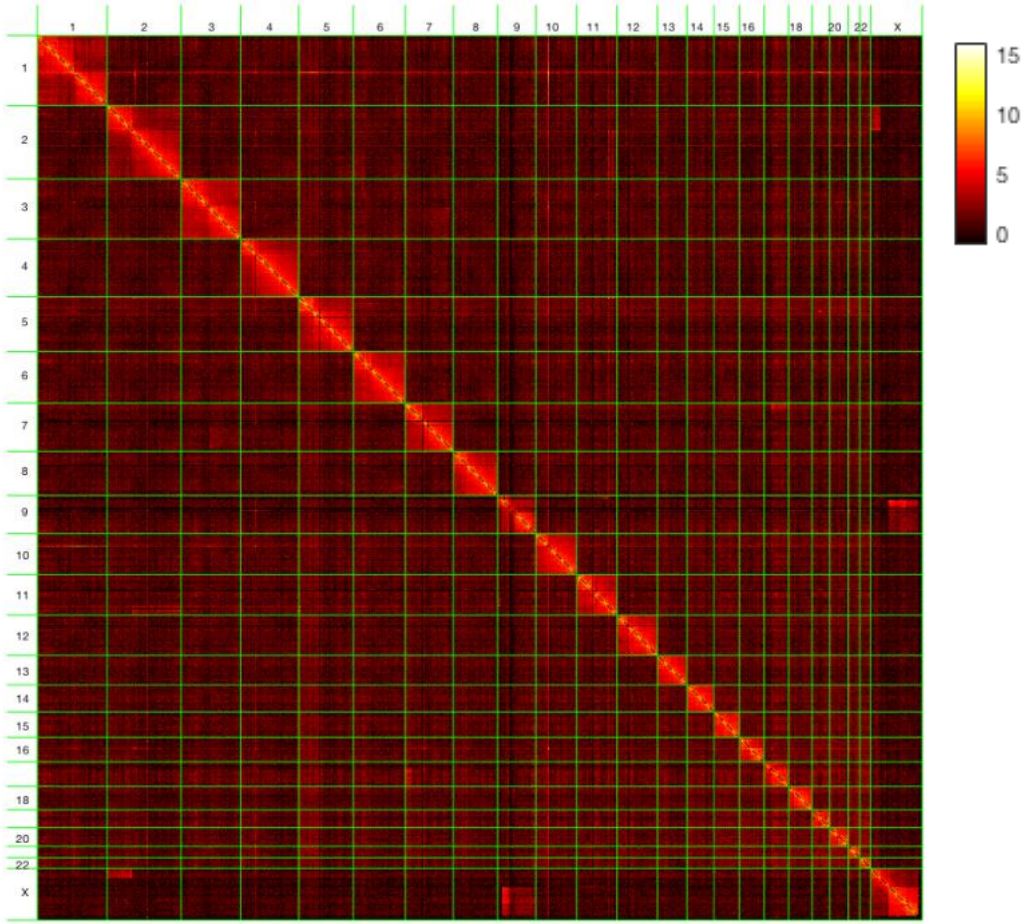


Sparse counts matrix
Dim: #cells × #genes

$\approx 15{,}000$ genes

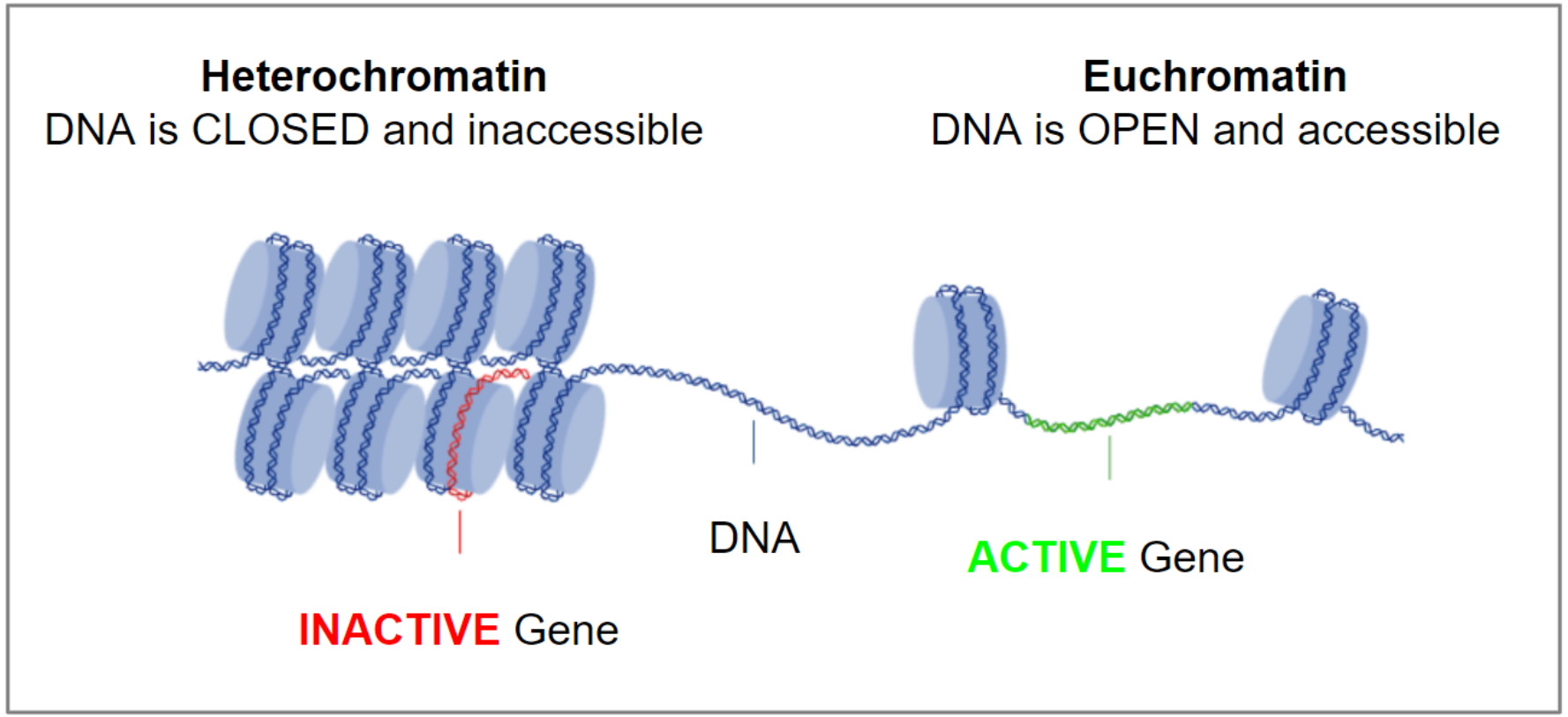
Chromatin conformation data: Hi-C





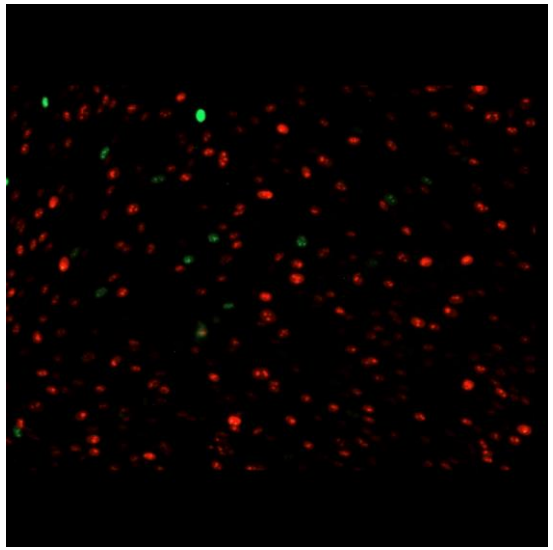
1MB = 3000 X 3000, 50 BP = 62M X 62M

Chromatin accessibility data: ATAC-seq

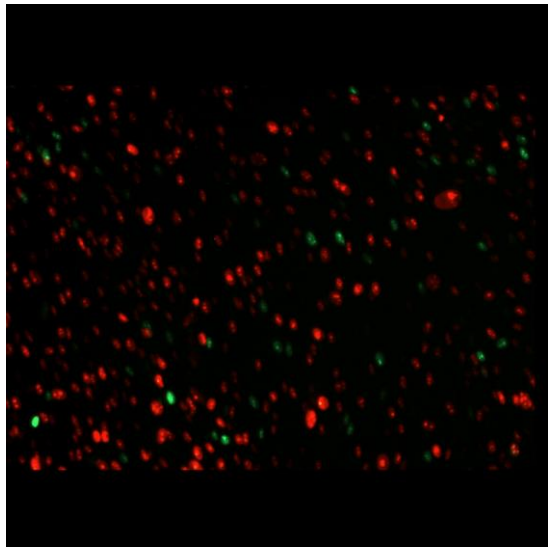


Live cell imaging

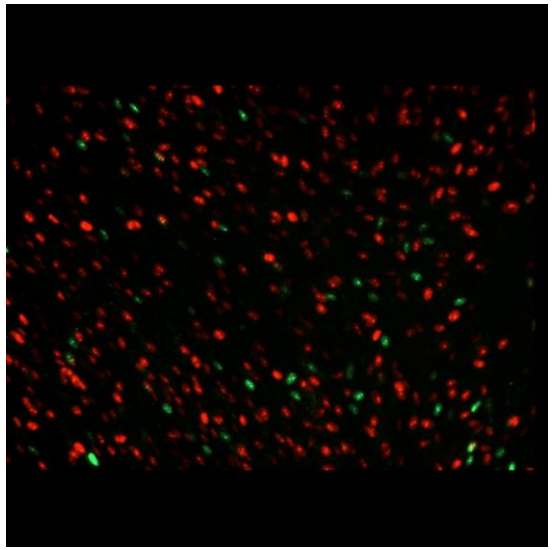
Cells only

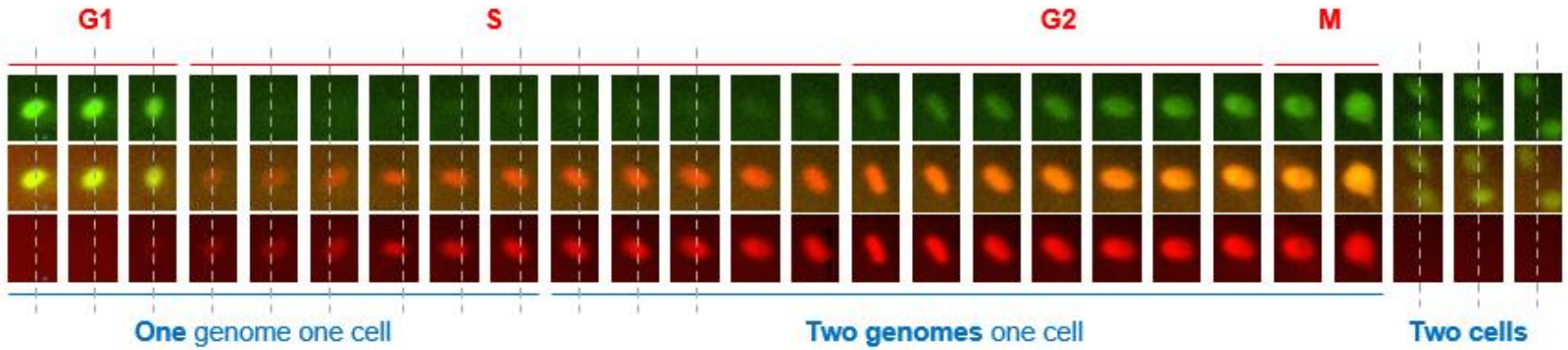


+MYOD1



-PRRX1

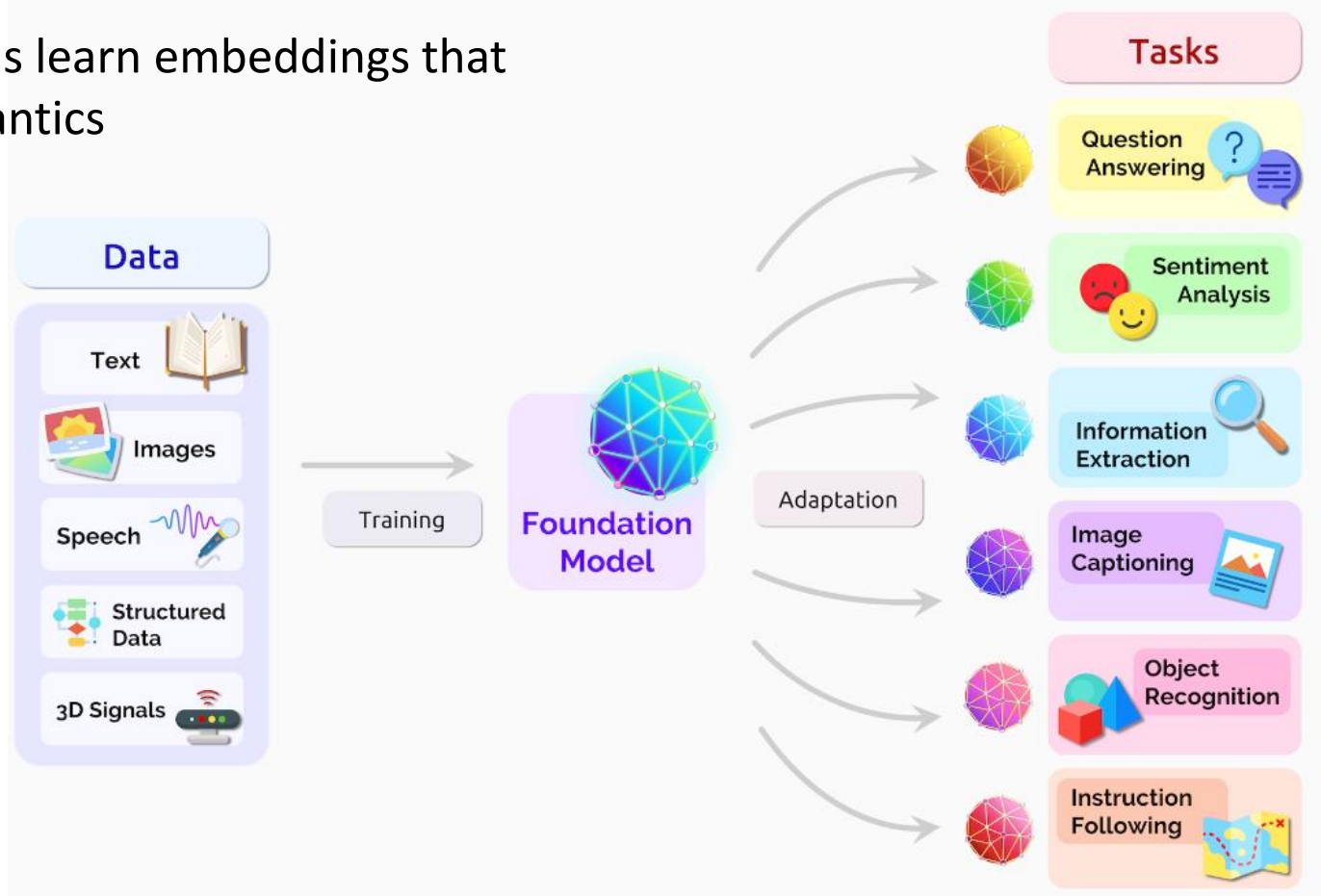




Geneformer

Foundation models

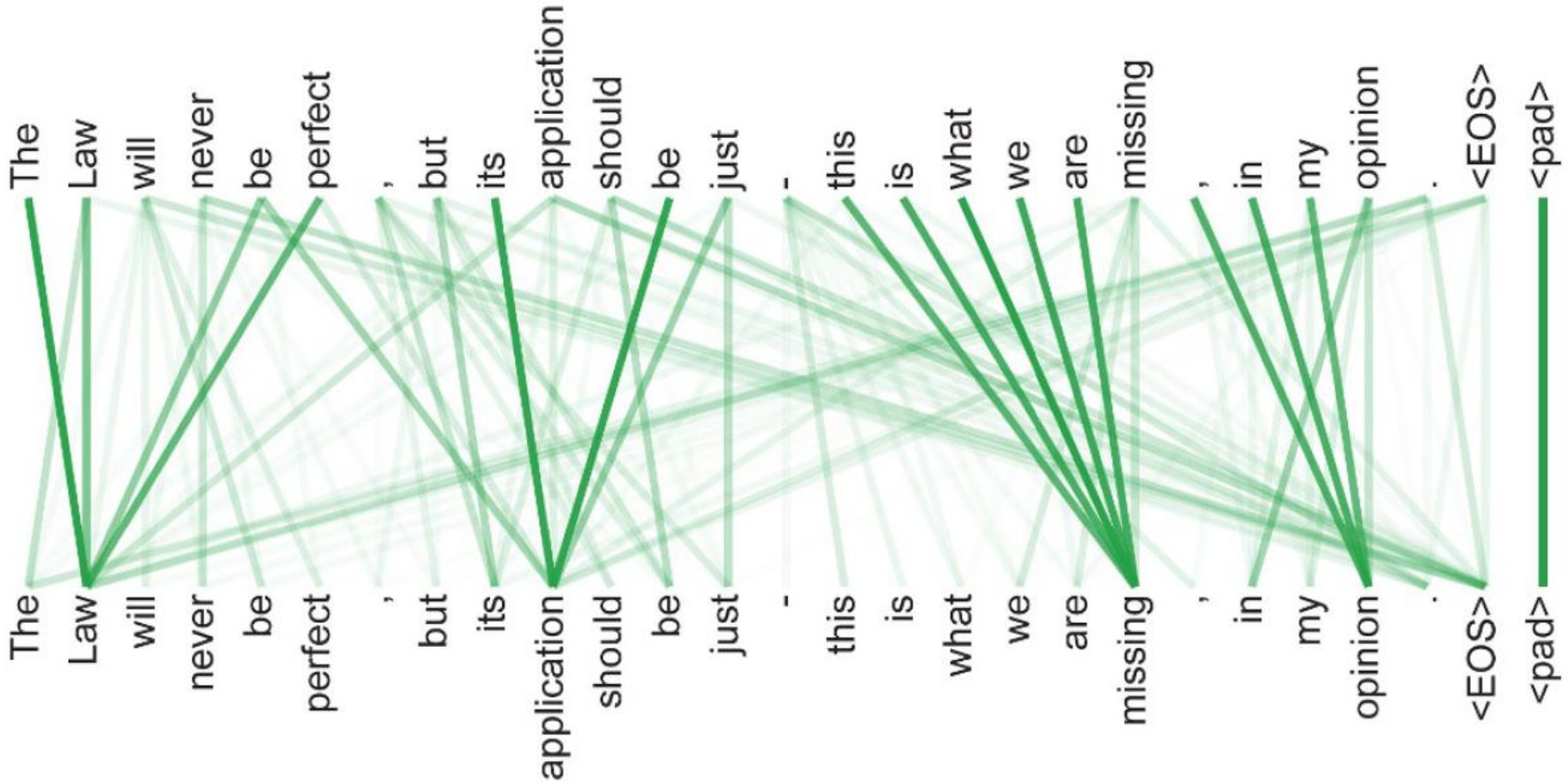
Foundation models learn embeddings that capture data semantics



Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2021).

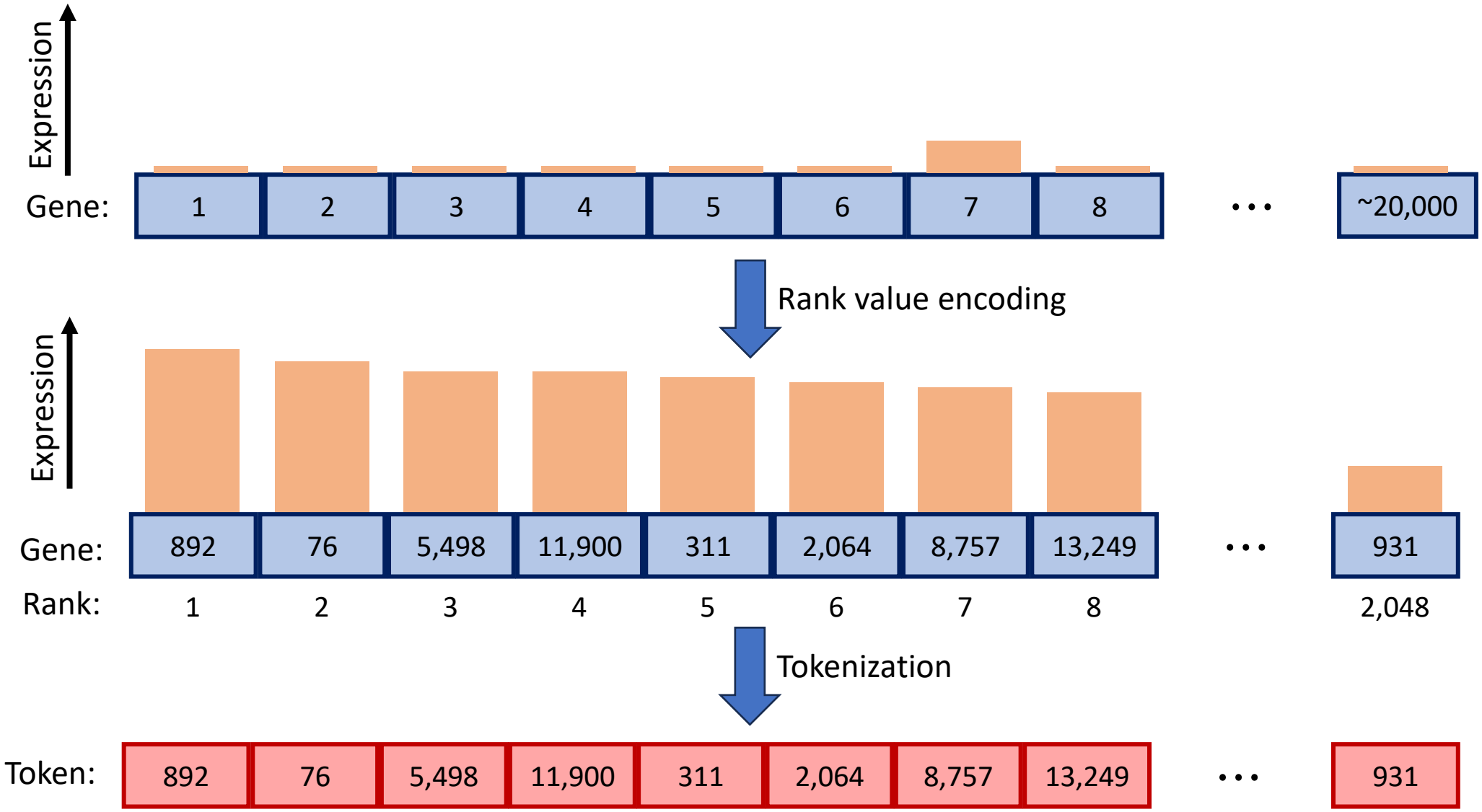
Attention mechanism

The embedding of each token depends on other tokens in the sequence

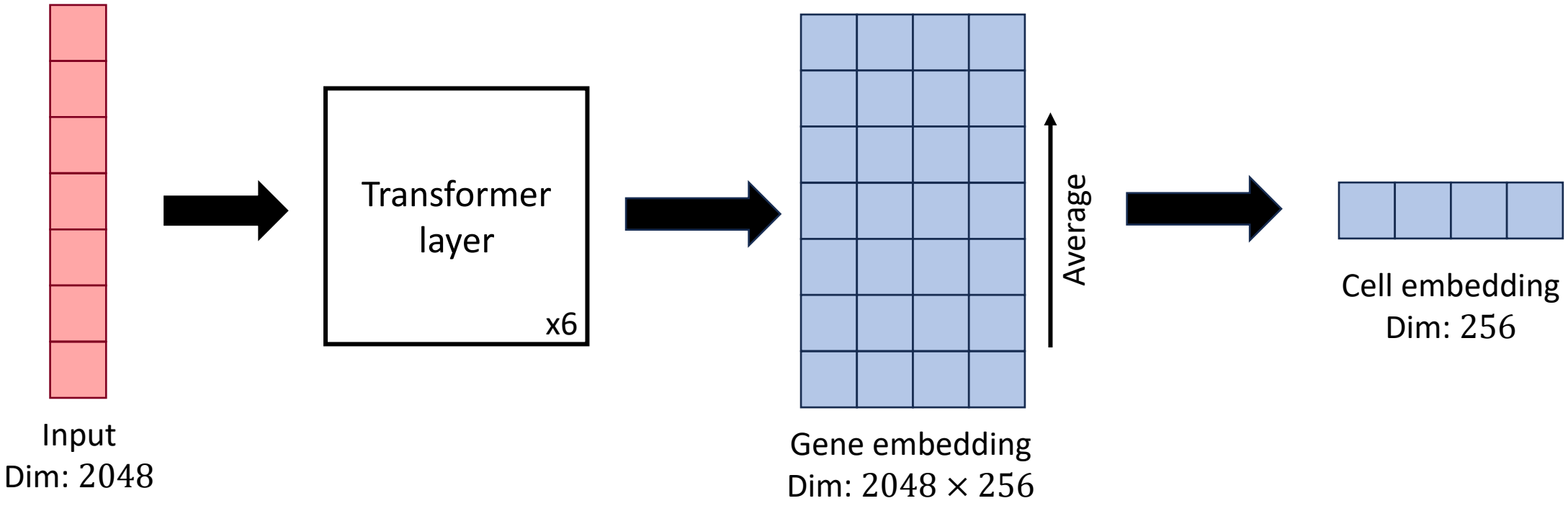


Bishop, Christopher M., and Hugh Bishop. *Deep learning: Foundations and concepts*. Springer Nature, 2023.

Rank value encoding allows for a meaningful positional encoding



Geneformer architecture

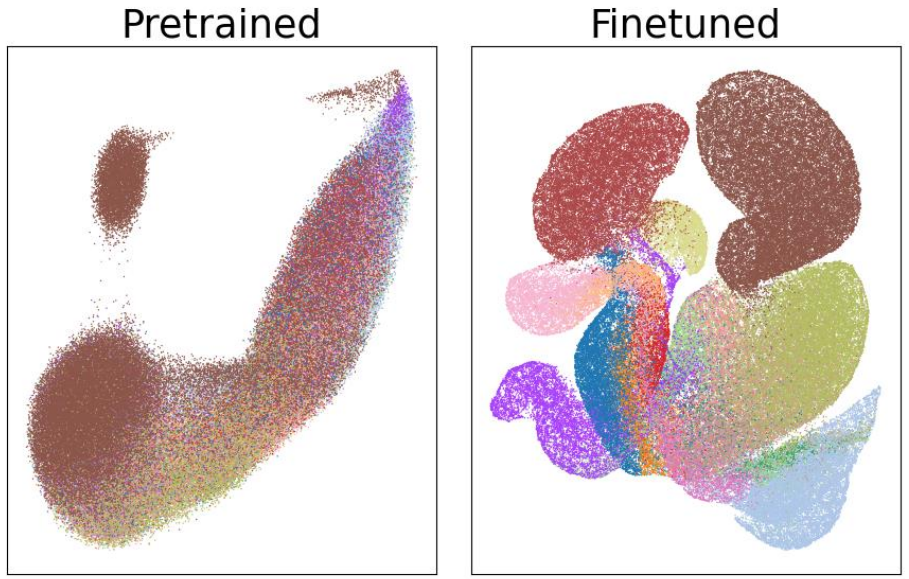


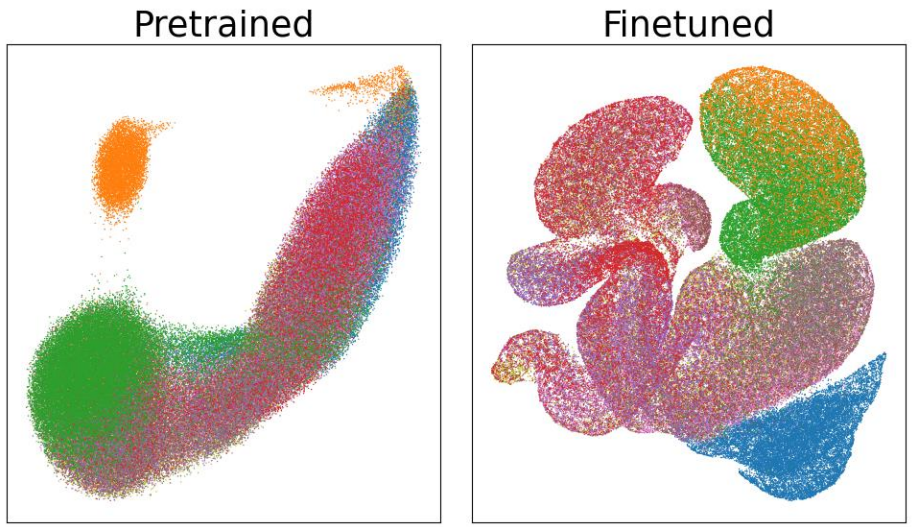
Theodoris, Christina V., et al. "Transfer learning enables predictions in network biology." *Nature* 618.7965 (2023): 616-624.

Fine-tuning

We use cell type classification as the fine-tuning task

Fine-tuning results in clustering by cell type instead of by dataset





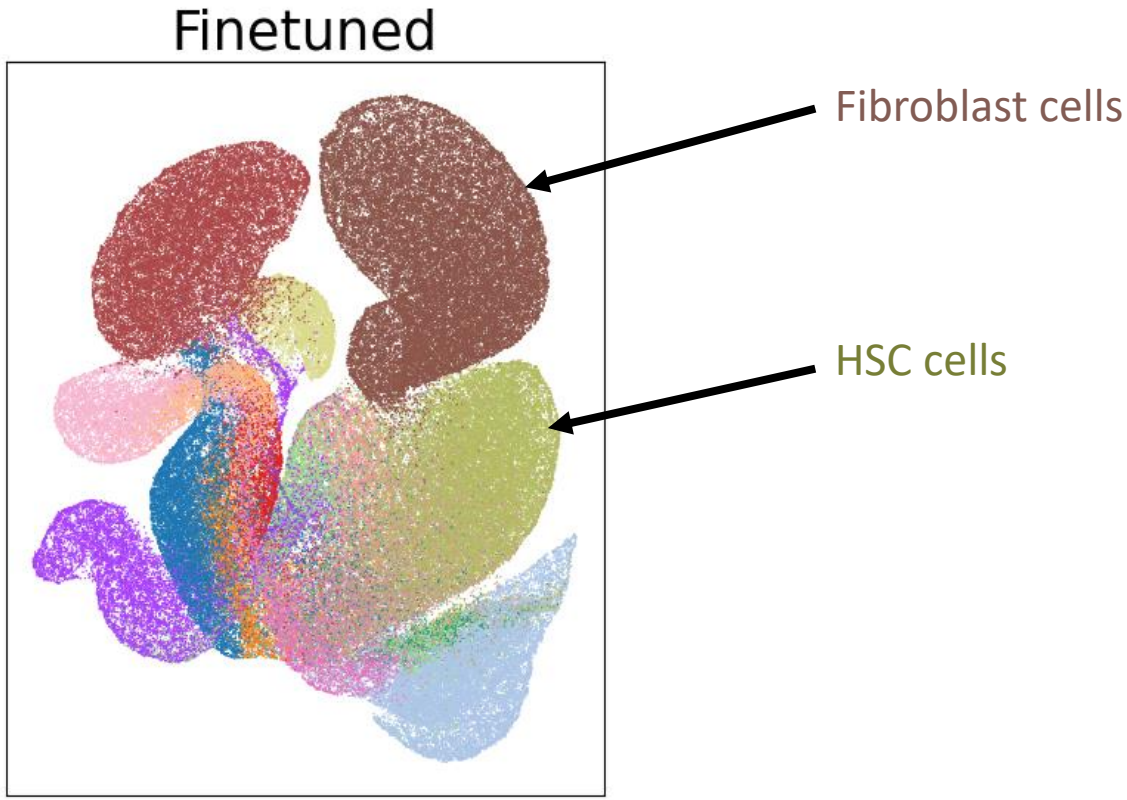
Cell type

- B_cell
- CLP
- CMP
- Dendritic_cell
- EryP
- Fib
- GMP
- HSC
- LMPP
- LinNeg
- MDP
- MEP
- MKP
- MLP
- MPP
- Mono
- NK
- PreBNK
- T_cell

Dataset

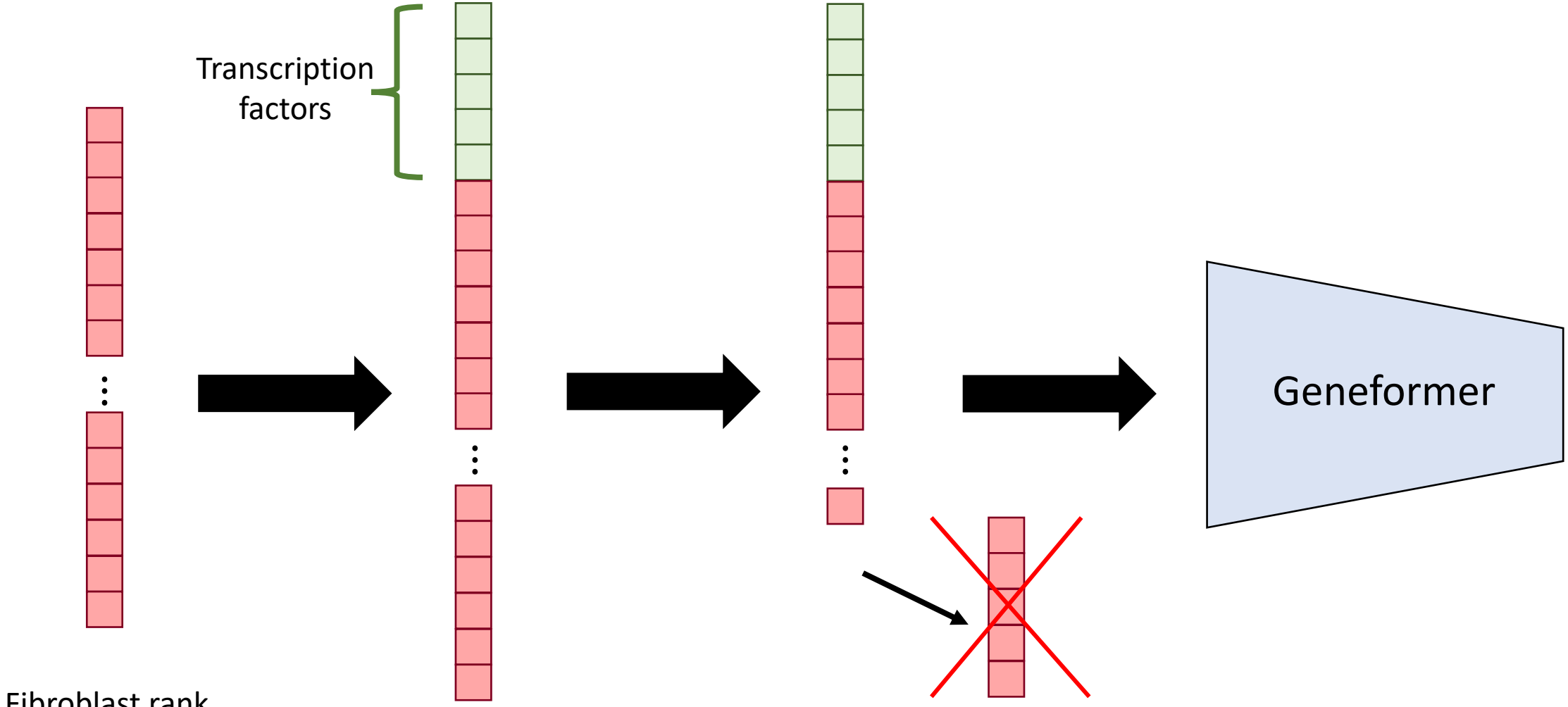
- pellin
- sc_fib
- tabula_sapiens
- weng_old1_BMMC_HSPC
- weng_old2_BMMC_HSPC
- weng_young1_all_t1
- weng_young1_all_t2
- weng_young2_HSC
- weng_young2_all

Fine-tuning



Perturbation experiment

Simulating cell reprogramming



Caveats

- All transcription factors are added to the top
- The order of transcription factors is not considered
- Only models the first step of cell state transition

Relevant measure: the directionality of the shift

Experimental setup

Candidate transcription factors
GATA2
GFI1B
FOS
STAT5A
REL
FOSB
IKZF1
RUNX3
MEF2C
ETV6

10 choose 5
=252 total recipes!

The distance metric:

$$\text{cosine distance} = 1 - \cos\theta$$

$$= 1 - \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

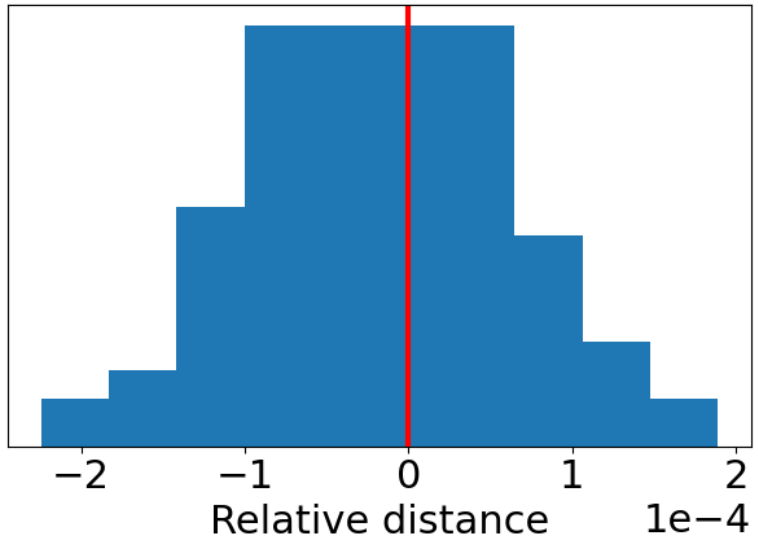
$$\text{Relative distance} = \frac{d^p - d^u}{d^u}$$

d^p: cosine distance between *perturbed* cells centroid and HSC centroid

d^u: cosine distance between *unperturbed* cells centroid and HSC centroid

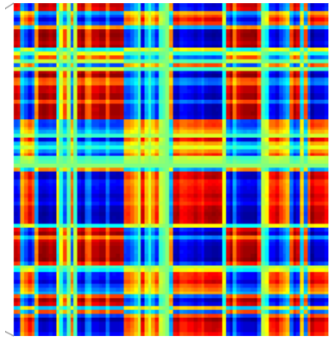
Results

Perturbation	Relative distance ($\times 10^{-4}$)
STAT5A, REL, IKZF1, MEF2C, ETV6	-2.24
STAT5A, FOSB, IKZF1, MEF2C, ETV6	-1.89
FOS, STAT5A, IKZF1, MEF2C, ETV6	-1.87
GFI1B, STAT5A, IKZF1, MEF2C, ETV6	-1.86
FOS, STAT5A, REL, MEF2C, ETV6	-1.83
⋮	⋮
GATA2, GFI1B, FOS, IKZF1, RUNX3	1.49
GATA2, GFI1B, REL, FOSB, RUNX3	1.52
GATA2, FOS, REL, FOSB, RUNX3	1.52
GATA2, GFI1B, FOS, REL, RUNX3	1.53
GATA2, GFI1B, FOS, FOSB, RUNX3	1.89

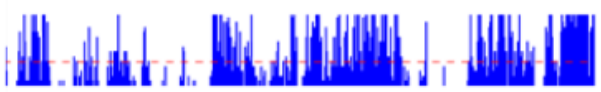


Improvements to Geneformer

- Increase the scale
  - GPT-3 has upwards of 175B parameters and was trained on 45 TB of data
  - Geneformer has 10M parameters and was trained on 30M cells
- Incorporate different types of data
  - Chromatin conformation (Hi-C)
  - Chromatin accessibility (ATAC-seq)



Hi-C



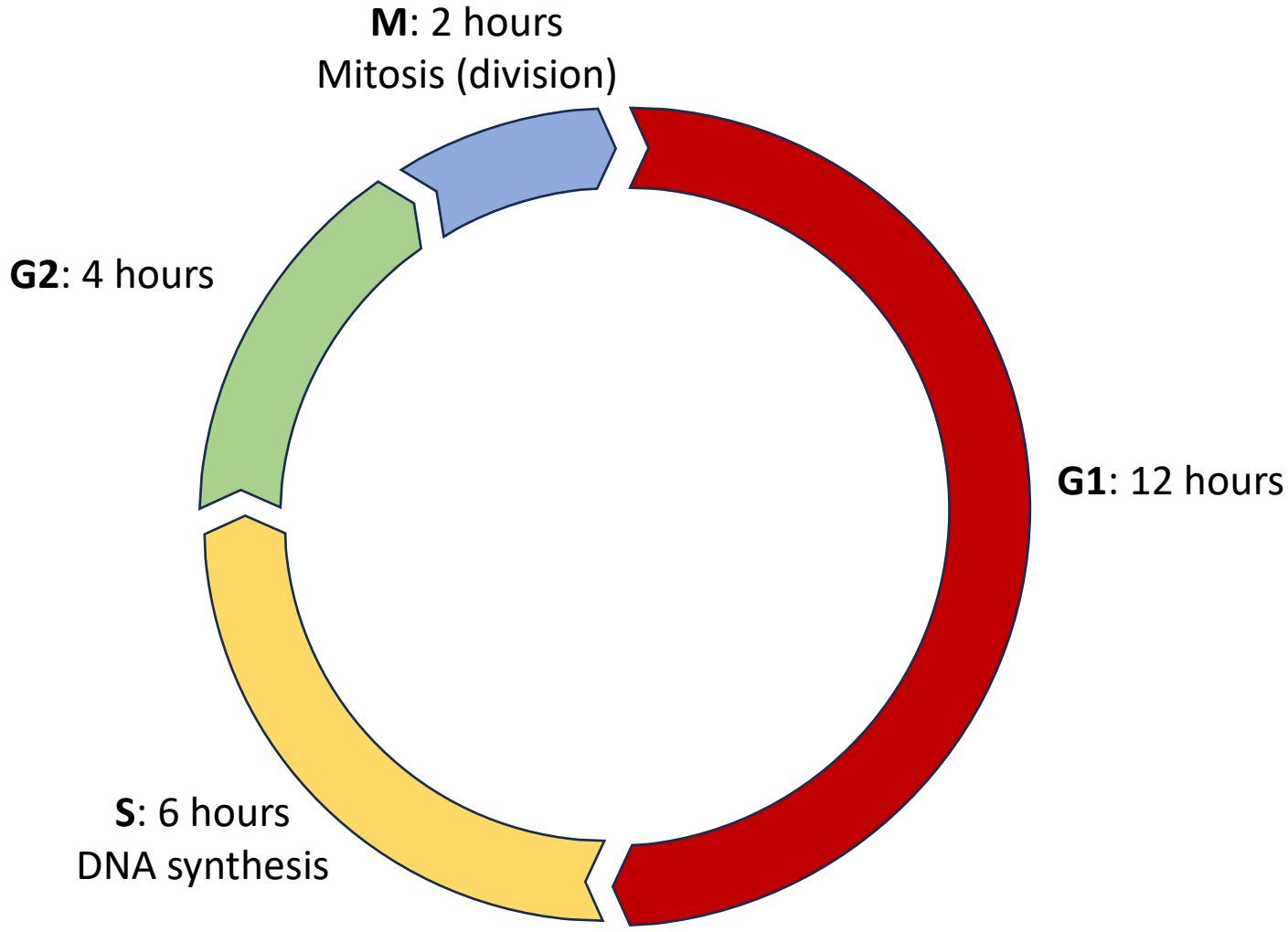
ATAC-seq

Challenges

- The model should work with both unimodal and multi-modal data
- Training the network is both labor and compute intensive

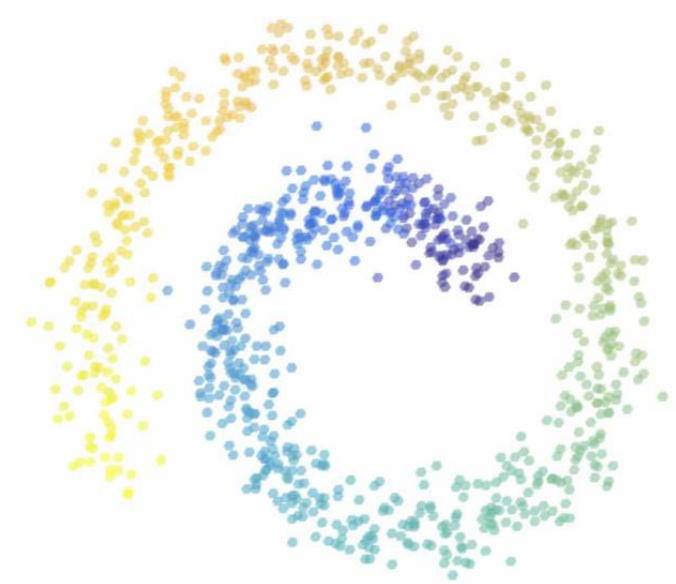
Cell cycle dynamics

Cell cycle

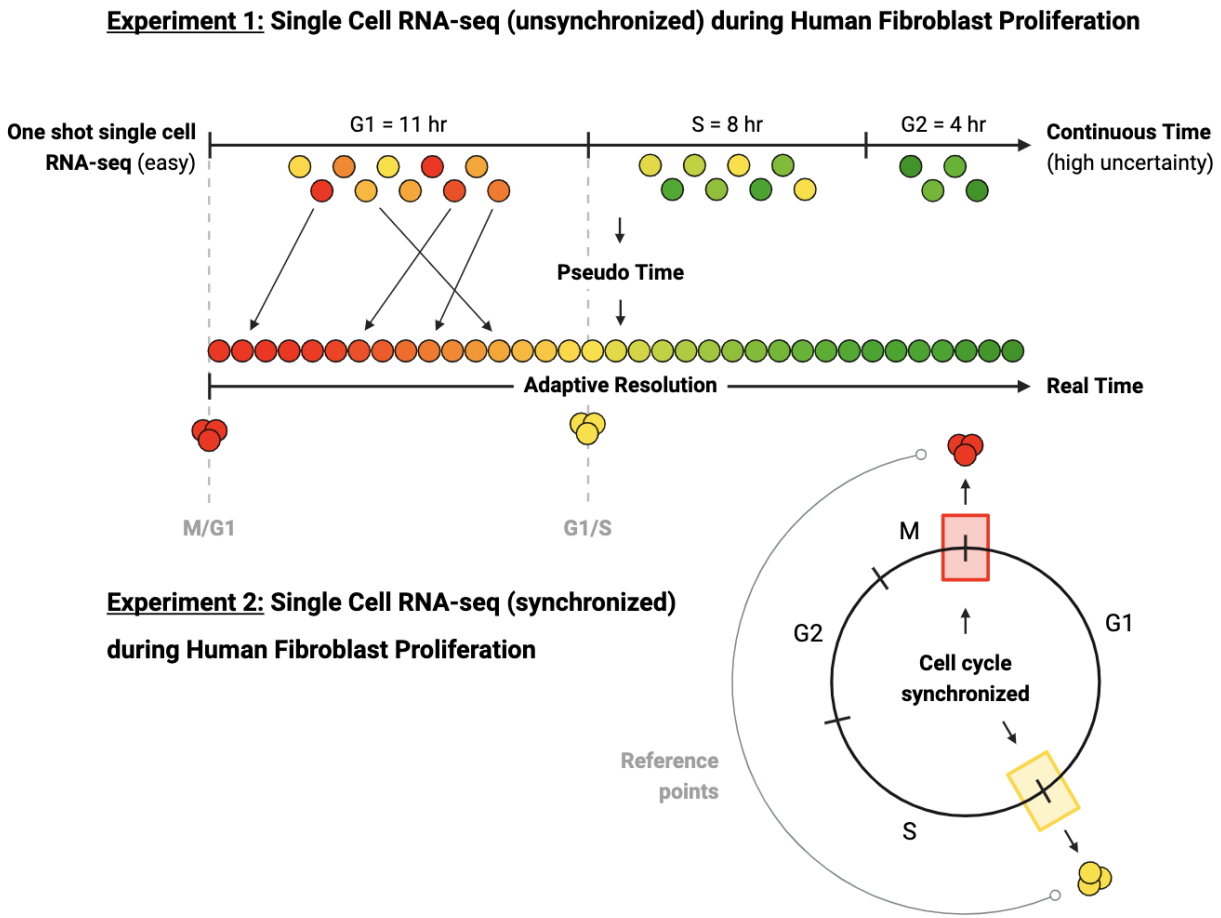

M: 2 hours
Mitosis (division)
G2: 4 hours
G1: 12 hours
S: 6 hours
DNA synthesis

Ordering the cells

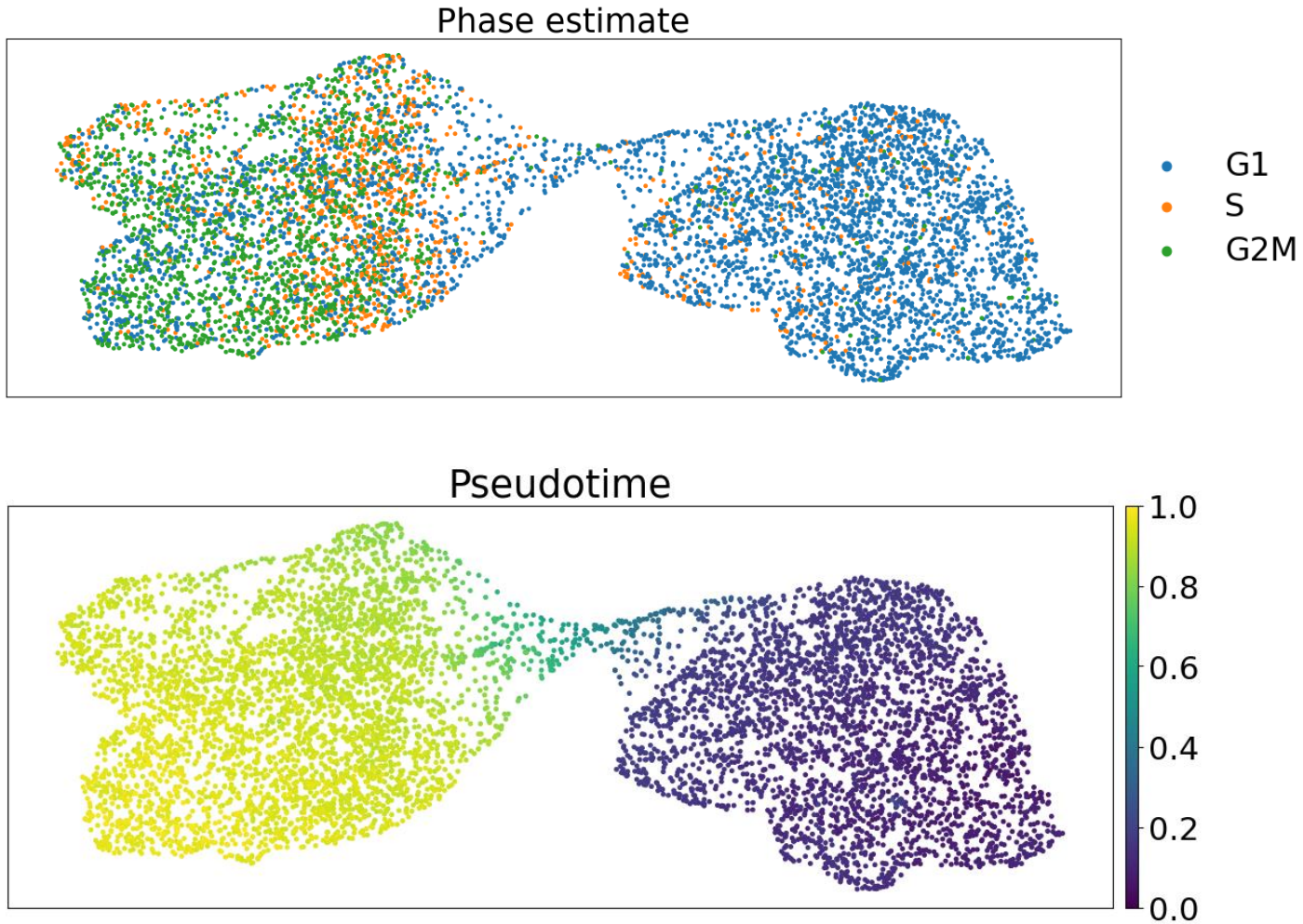
1. Impute missing values
2. Compute the *pseudotime* of each cell relative to a chosen root cell
   - Construct a transition kernel across the cells
   - The distance between cells x and y is the accumulated probability of starting from x and arriving at y over all random walk lengths



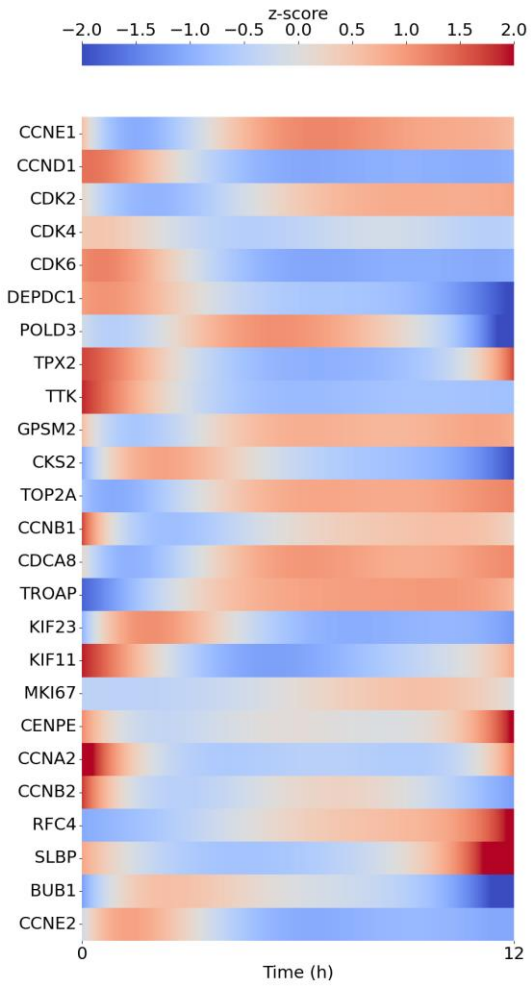
Collecting cell cycle data

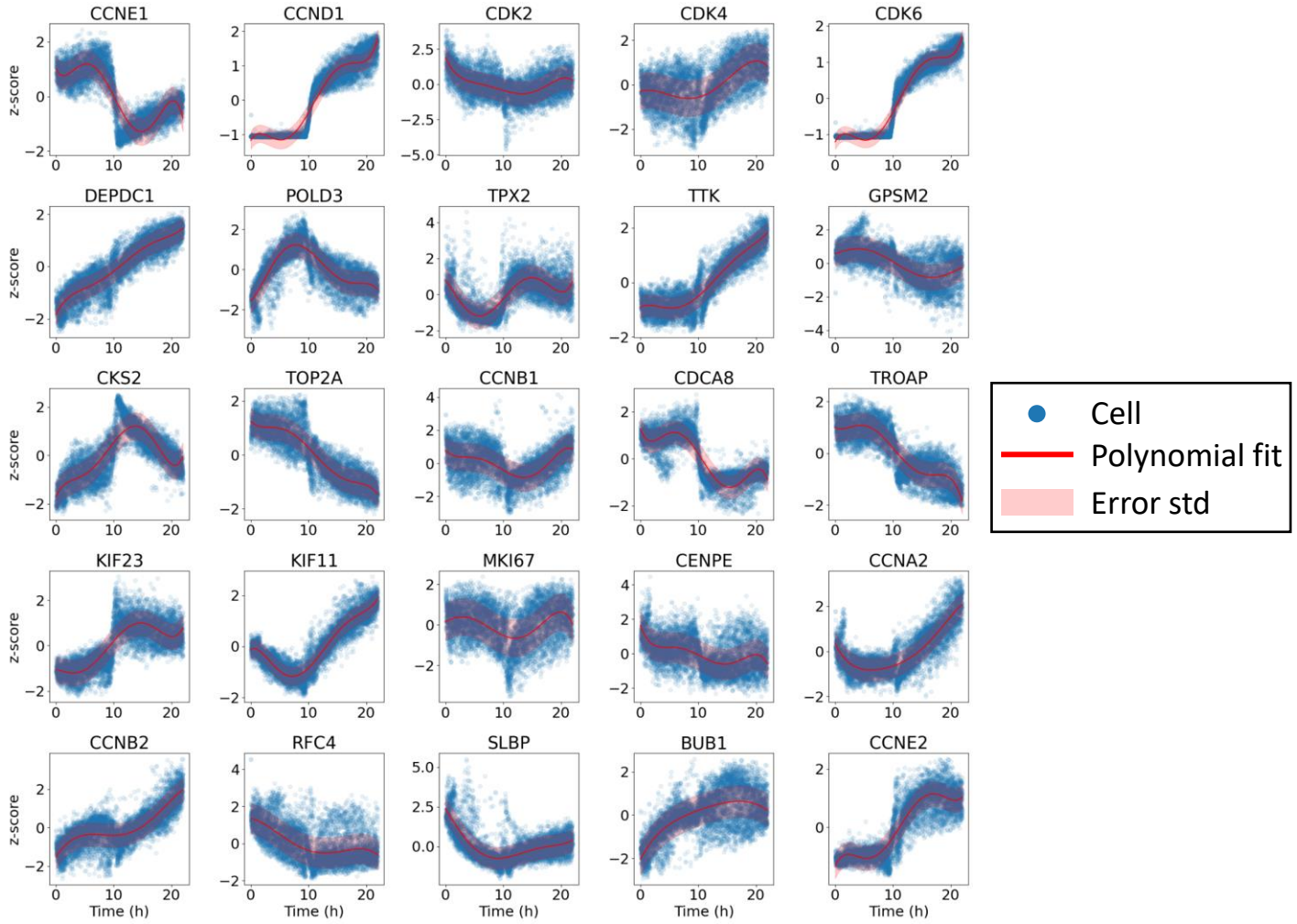


Pseudotime ordering shows agreement with phase estimates



Gene trajectories





Next steps

Learn the dynamics

$$x_{k+1} = Ax_k + Bu_k$$

Remaining design decisions:

- The coordinate frame and dimensionality
- The length of timesteps
- Sparsity pattern of A
  - Which genes interact with each other

Conclusions

Conclusions

- Digital biology has the potential to revolutionize medicine
- There is still a lot of work to be done to fully reap the benefits of advancements in other fields such as machine learning and robotics
- Breakthroughs in AI can be brought to biology to accelerate discovery

Funding

- DARPA
  - TwinCell Blueprint: Foundation for AI-Assisted Cell Reprogramming
- AFOSR
  - Data-guided Learning and Control of Higher Order Structures

Thank you!

References:

- Theodoris, Christina V., et al. "Transfer learning enables predictions in network biology." Nature 618.7965 (2023): 616-624.
- Pickard, Joshua, et al. “Language model powered digital biology." arXiv preprint arXiv:2409.02864 (2024).
- Cwycyshyn, Jillian, et al. "A programmable platform for probing cell migration and proliferation." APL Bioengineering 8.4 (2024).

Lab website: https://rajapakse.lab.medicine.umich.edu/