Large-scale feature learning of chromosome conformation for genome modeling

Nick Galioto

Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, University of Michigan

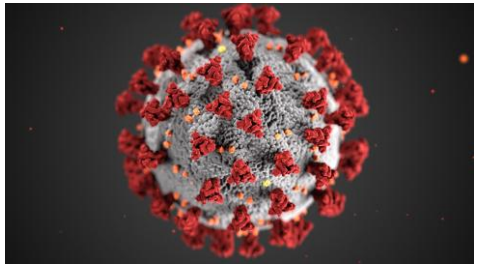
Argonne National Laboratory Seminar Presentation

November 17, 2025

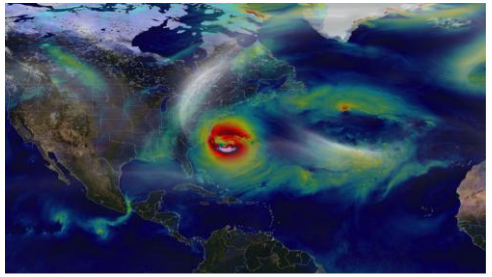
My background

The need for data-driven estimation of dynamics is widespread

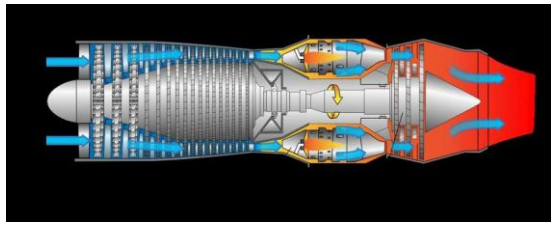
- Epidemiology (SIR model)
- Predicting medical events
  - Blood glucose levels in diabetic patients
- Weather and climate
- Industrial and manufacturing processes
- Financial markets
- Traffic patterns
- Energy grid demands
- Model predictive control
- Surrogate modeling
  - Molecular dynamics
  - Fluid dynamics



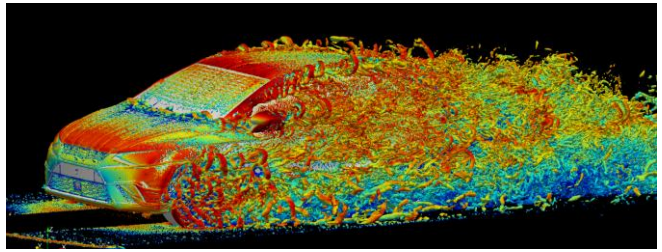
U.S. Center for Disease and Control, 2020.



Earth Science at Ames. National Aeronautics and Space Administration, 2017.



Naji, Anees. Polytechnic University of Bucharest, 2017.



Large-Scale Computational Fluid Dynamics. Barcelona Supercomputing Center, 2023.

Vigilancia Quadcopter. TechnoSys, Embedded Systems (P) Ltd., 2023.

Adhikari, Ganesh. *Investopaper*. Investopaper, 2023.

Sullivan, Justin. *WBUR*. WBUR, 2023.

Hidden Markov model

Joint parameter-state estimation with stochastic dynamics

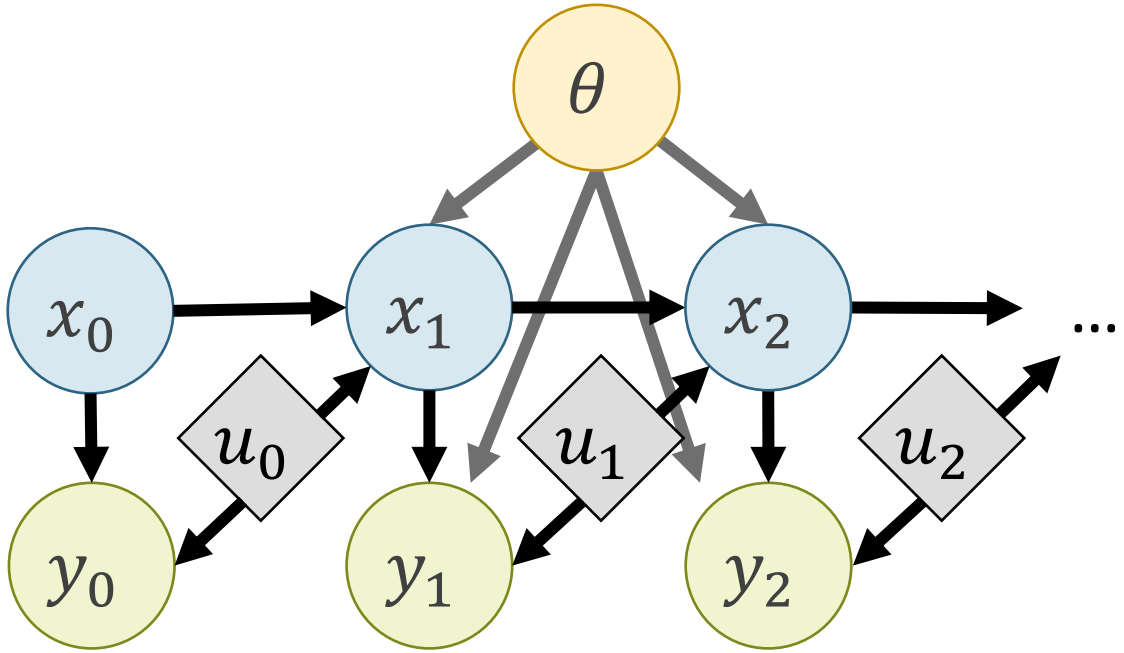
$$X_k \in \mathbb{R}^{d_x}, \qquad Y_k \in \mathbb{R}^{d_y}, \qquad \theta = (\theta_\Psi, \theta_h, \theta_\Sigma, \theta_\Gamma) \in \mathbb{R}^{d_\theta}$$

$$X_k = \Psi(X_{k-1}, u_{k-1}, \theta_\Psi) + \xi_k; \quad \xi_k \sim \mathcal{N}\left(0, \Sigma(\theta_\Sigma)\right)$$

$$Y_k = h(X_k, \theta_h) + \eta_k; \qquad \eta_k \sim \mathcal{N}\left(0, \Gamma(\theta_\Gamma)\right)$$

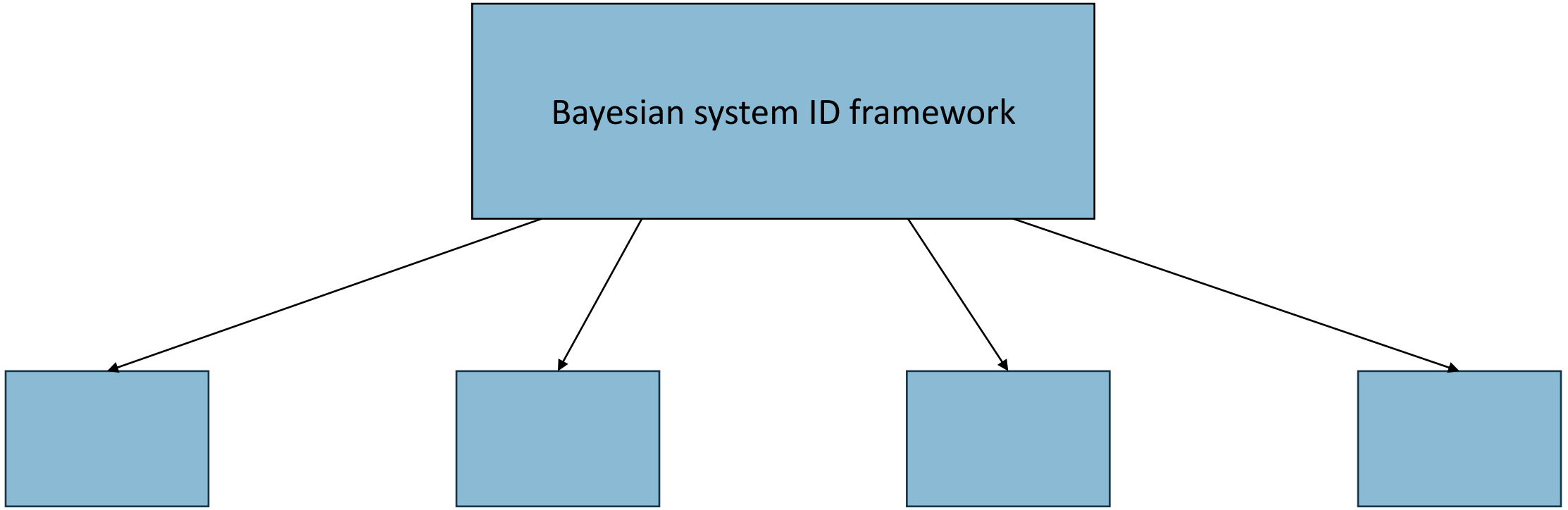
The process noise term ξ_k accounts for model error

- Parameter error
- Integration error
- Insufficient model expressiveness



1. Parameter Uncertainty
2. Model Uncertainty
3. Measurement Uncertainty

Bayesian framework gives conditions for optimality of many popular approaches



Tu, 2013.

Brunton et al., 2016.

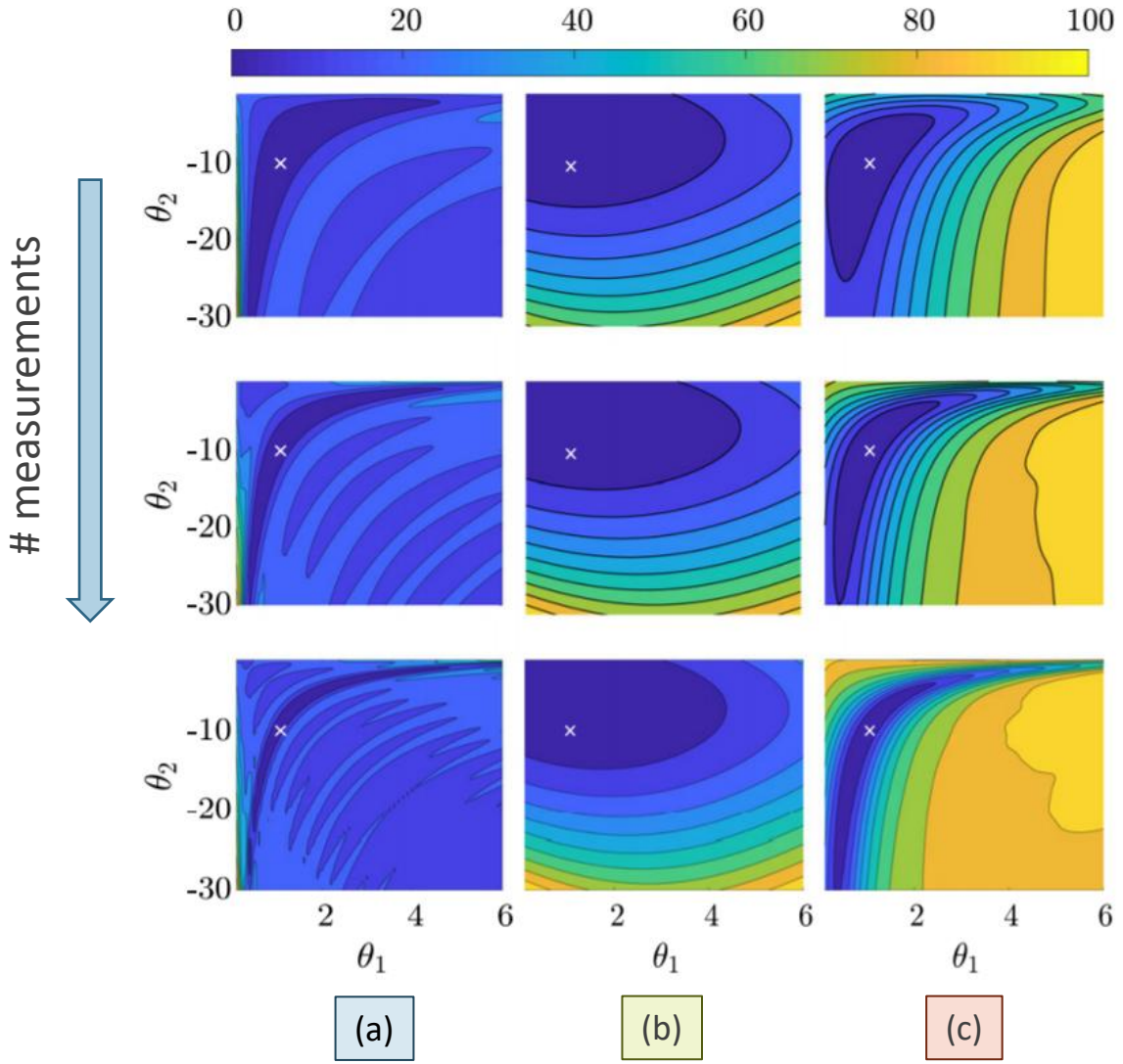
Oymak and Ozay, 2019.

Ribeiro et al., 2020.

Bayesian framework yields advantageous changes to the objective function surface

- (a) Long time horizon simulation
- (b) Short time horizon simulation
- (c) PhD work
  - Optimal combination of (a) and (b)

	(a)	(b)	(c)
Assesses long-term behavior	✓	✗	✓
Smooths local minima	✗	✓	✓
Increased confidence with data	✓	✗	✓



Publications

Journals

- Mustaev, Artem, **Nicholas Galioto**, et al. "A switching Kalman filter approach to online mitigation and correction of sensor corruption for inertial navigation." *arXiv preprint arXiv:2412.06601* (2024). (under review at *ION Navigation*)
- **Galioto, Nicholas**, et al. "Bayesian identification of nonseparable Hamiltonians with multiplicative noise using deep learning and reduced-order modeling." *Computer Methods in Applied Mechanics and Engineering* 430 (2024): 117194.
- **Galioto, Nicholas**, and Alex Arkady Gorodetsky. "Likelihood-based generalization of Markov parameter estimation and multiple shooting objectives in system identification." *Physica D: Nonlinear Phenomena* 462 (2024): 134146.
- **Galioto, Nicholas**, and Alex Arkady Gorodetsky. “Bayesian system ID: Optimal management of parameter, model, and measurement uncertainty.” *Nonlinear Dynamics*, vol. 102, no. 1, 2020, pp. 241-267.

Conferences

- Sharma, Harsh*, **Nicholas Galioto***, Alex Arkady Gorodetsky, and Boris Kramer. “Bayesian Identification of Nonseparable Hamiltonian Systems Using Stochastic Dynamic Models.” 2022 61st IEEE Conference on Decision and Control (CDC). IEEE, 2022.
- **Galioto, Nicholas**, and Alex Arkady Gorodetsky. “A new objective for identification of partially observed linear time-invariant dynamical systems from input-output data.” Learning for Dynamics and Control. PMLR, 2021.
- **Galioto, Nicholas**, and Alex Arkady Gorodetsky. “Bayesian identification of Hamiltonian dynamics from symplectic data.” 2020 59th IEEE Conference on Decision and Control (CDC). IEEE, 2020.

* Denotes equal contribution

Outline

- Cell reprogramming
- Introduction to Hi-C data
- ARCH3D: Architecture and pre-training
- Results
- Conclusions and future work

Introduction

Cell reprogramming

High-throughput chromosome conformation capture (Hi-C)
ARCH3D: Architecture and pre-training
Results
Conclusions and future work

Collaborators

Faculty and staff:

Indika Rajapakse
Mathematics, Bioinformatics

Alex Gorodetsky
Aerospace Engineering

Lindsey Muir
Bioinformatics

Walter Meixner
Experimentalist

Alumni:

Cooper Stansbury
iReprogram

Joshua Pickard
Broad Institute

PhD students:

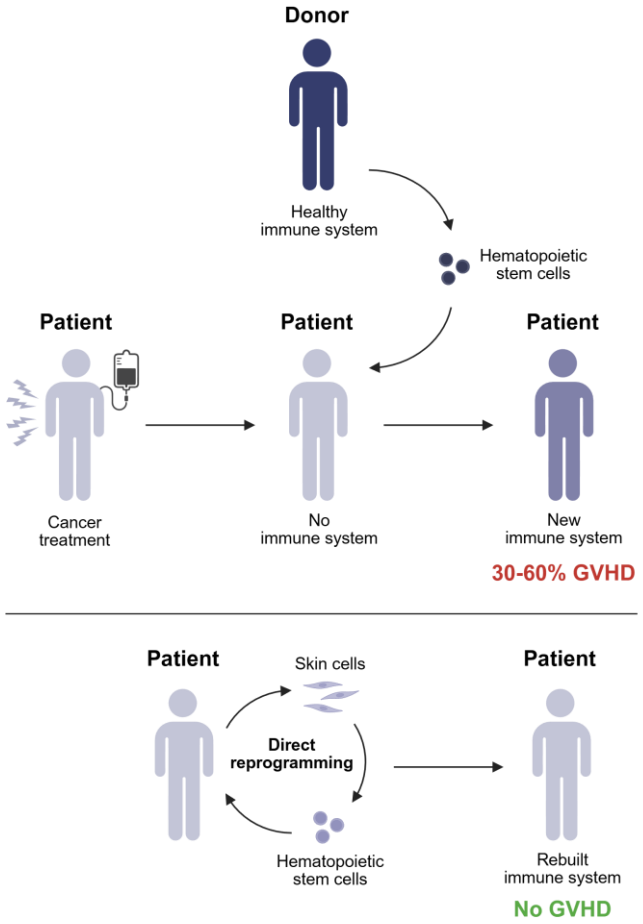
Jillian Cwycyshyn
Bioinformatics

Ram Prakash
Bioinformatics

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

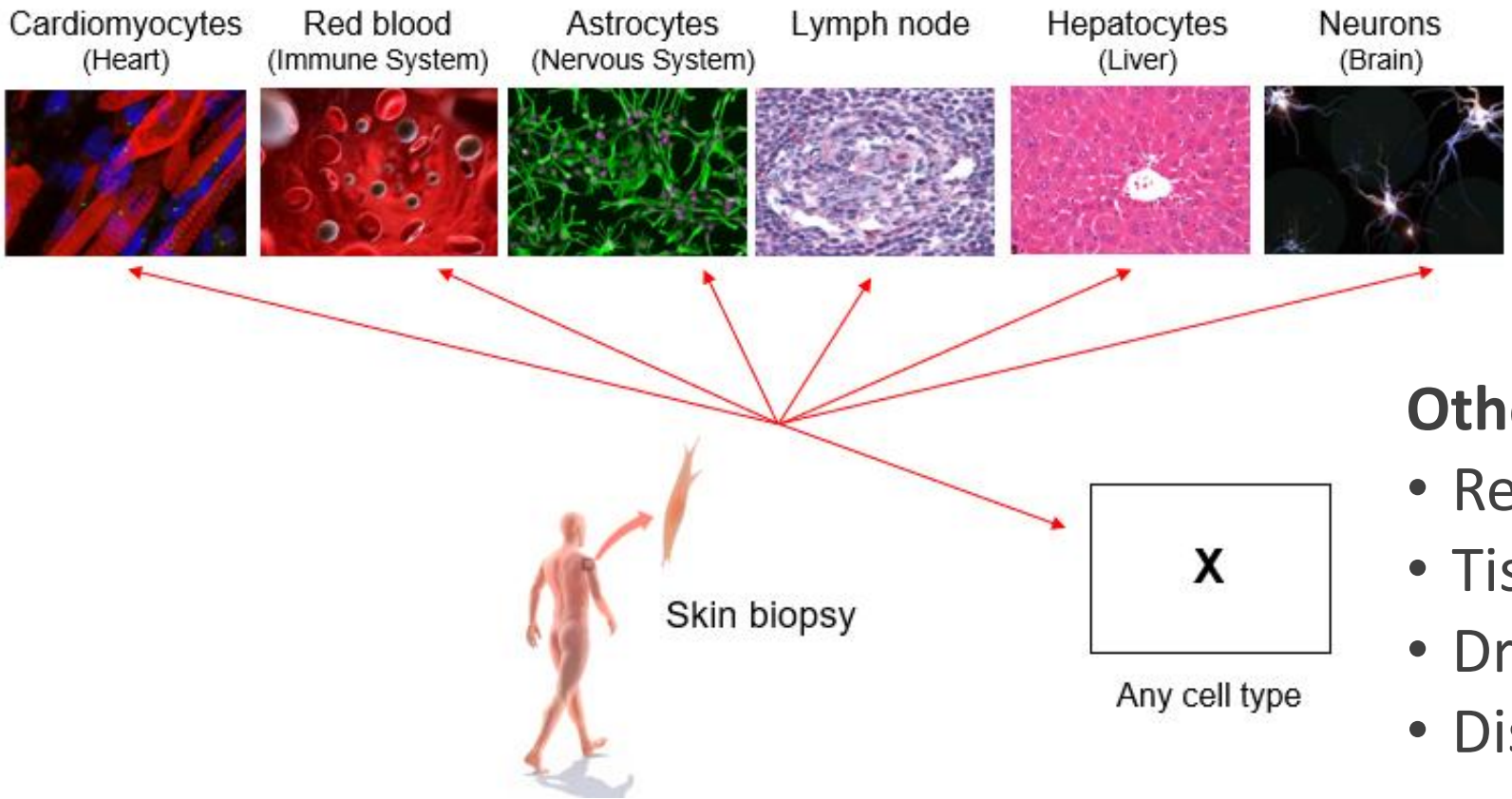
SOLUTION

Autologous cell reprogramming

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

Holtan, Shernan G., et al. "Disease progression, treatments, hospitalization, and clinical outcomes in acute GVHD: a multicenter chart review." *Bone marrow transplantation* 57.10 (2022): 1581-1585.

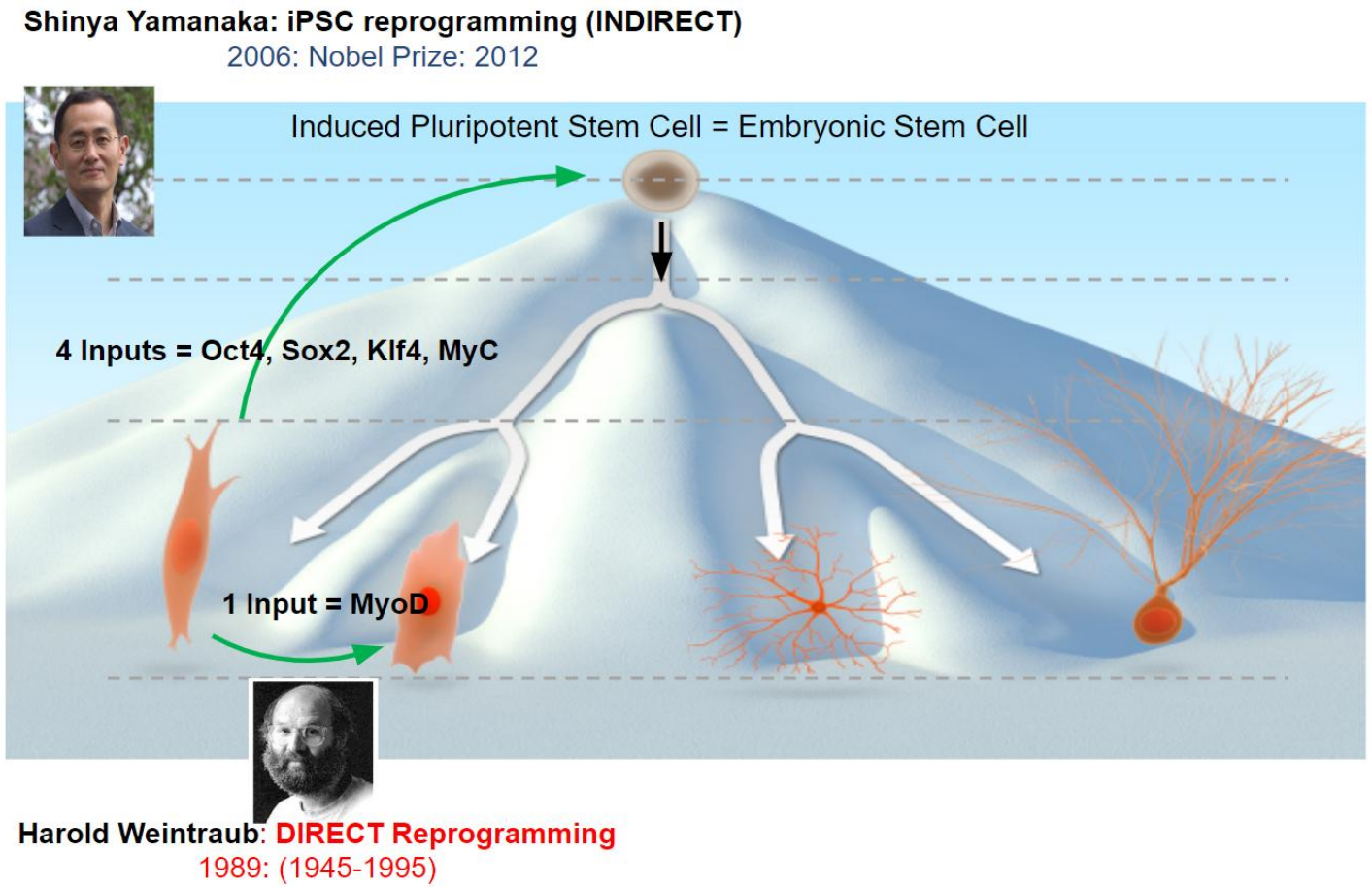
Cell reprogramming



Other applications:

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

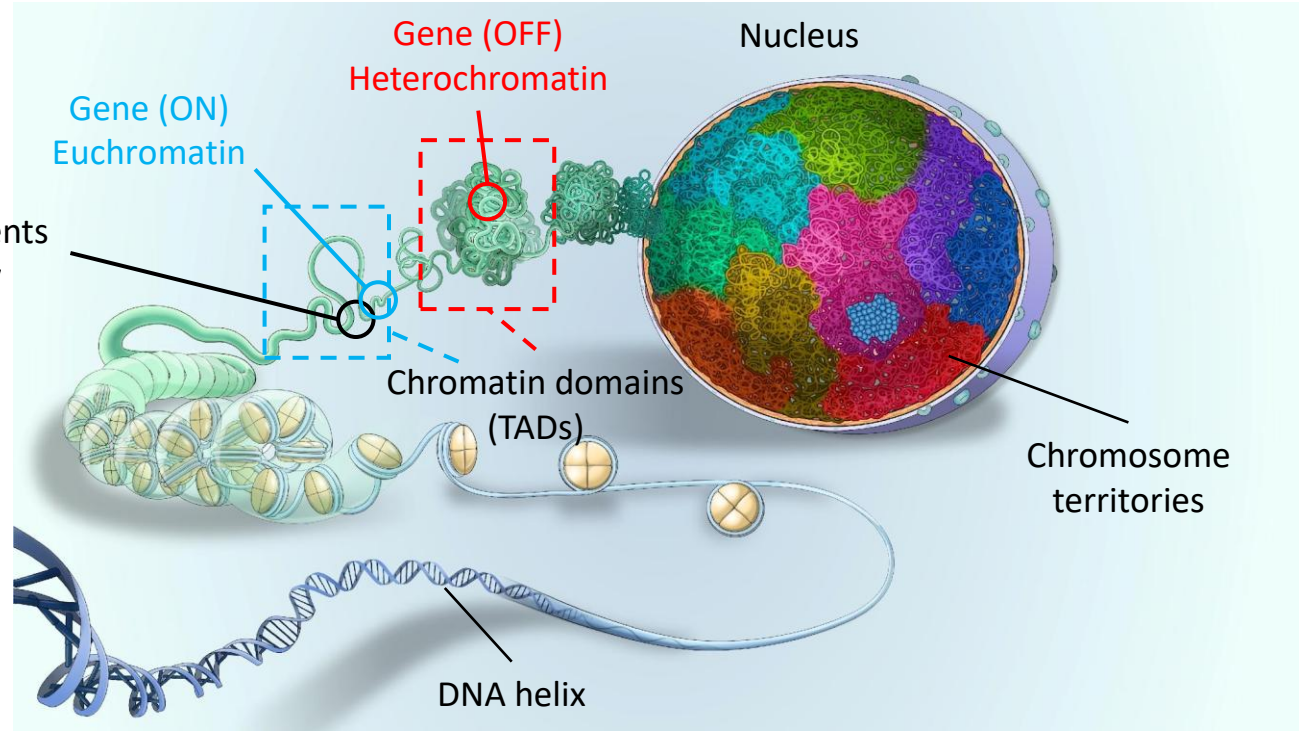
Cell reprogramming can be achieved through introduction of expertly-chosen transcription factors (TFs)



Challenges:

- Experiments are costly
- The set of possible TFs is vast
- Reprogramming efficiency remains low (1-3%)

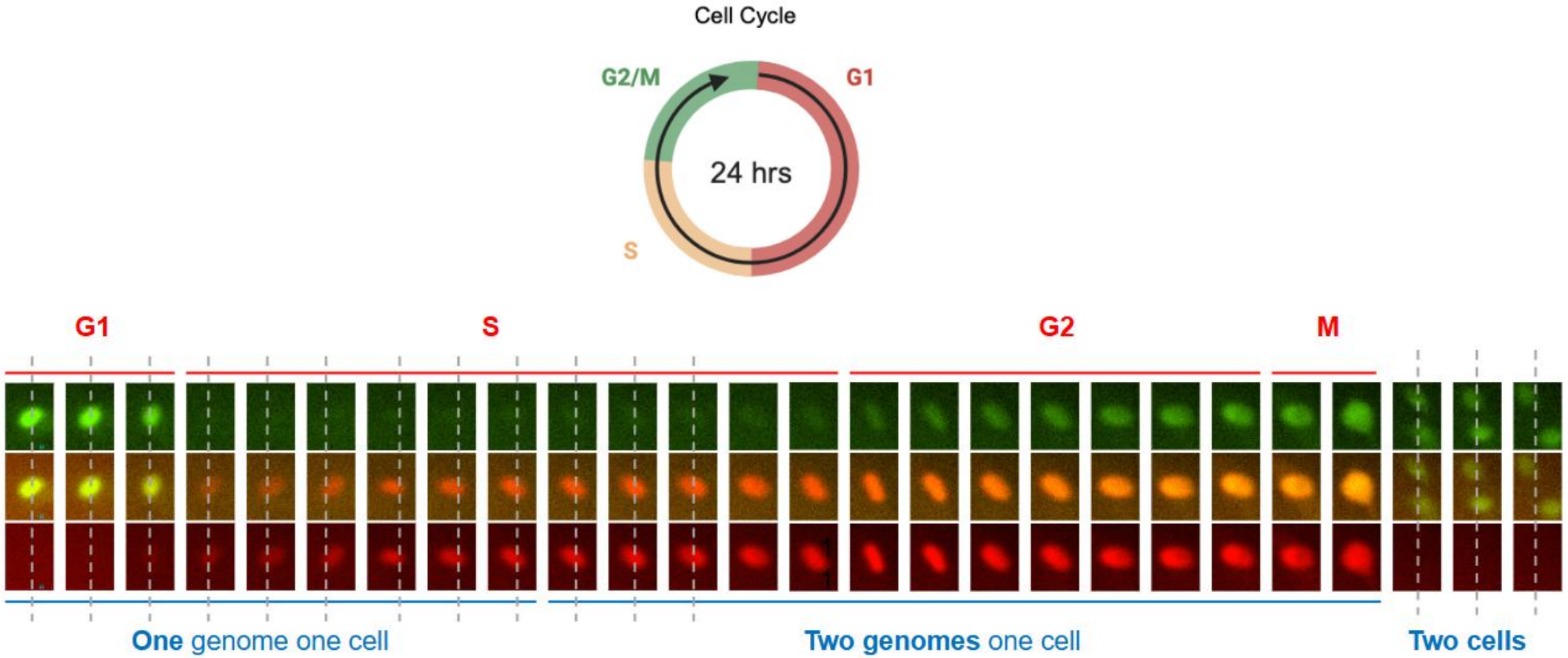
Genome structure regulates cell identity



Adapted from: Misteli, Tom. "The self-organizing genome: principles of genome architecture and function." *Cell* 183.1 (2020): 28-45.

- Chromosomes occupy distinct regions of the nucleus known as “chromosome territories”
- Active genes are located in areas of loosely-packed chromatin (euchromatin)
- Topologically associating domains (TADs) insulate sections of the genome from each other
- Enhancers are brought into proximity of promoters through chromatin looping

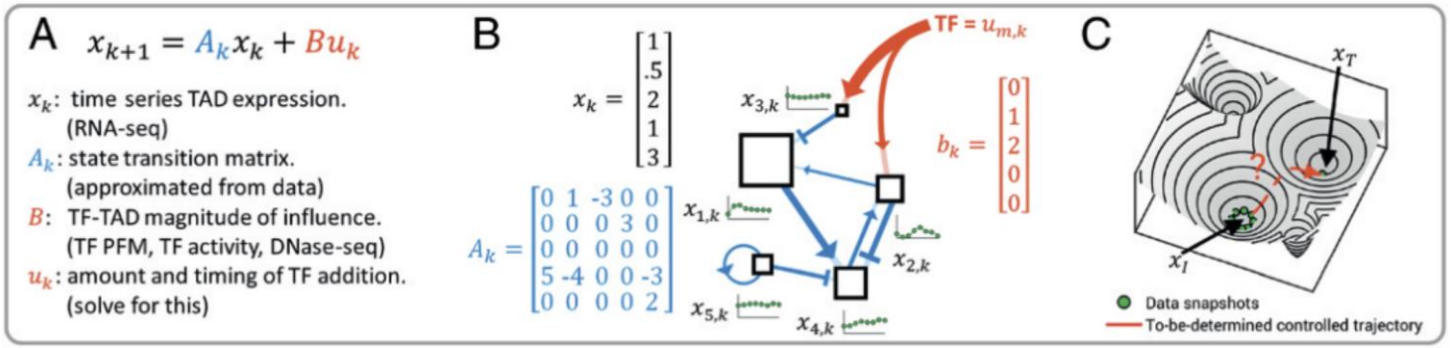
The genome is a dynamical system



Existing method: Data-guided control (DGC)

Formulates cell reprogramming as a control problem

- State is represented using RNA-seq data, grouped into TADs
- Selection of TFs is modeled as a control policy



Optimal TF policy

$$u_t^* = \underset{u_t}{\operatorname{argmin}} \|x_T - z_6(u_t)\|$$

Target cell state

Estimated cell state

- **Limitation:** Cannot account for changes in TAD structure

Ronquist, Scott, et al. "Algorithm for cellular reprogramming." *Proceedings of the National Academy of Sciences* 114.45 (2017): 11832-11837.

Foundation models show promise in producing multi-purpose representations of biological data

DNA Sequence
- GenSLM
- AlphaGenome
- Evo2

Transcriptomic
- Geneformer
- scGPT
- scBERT

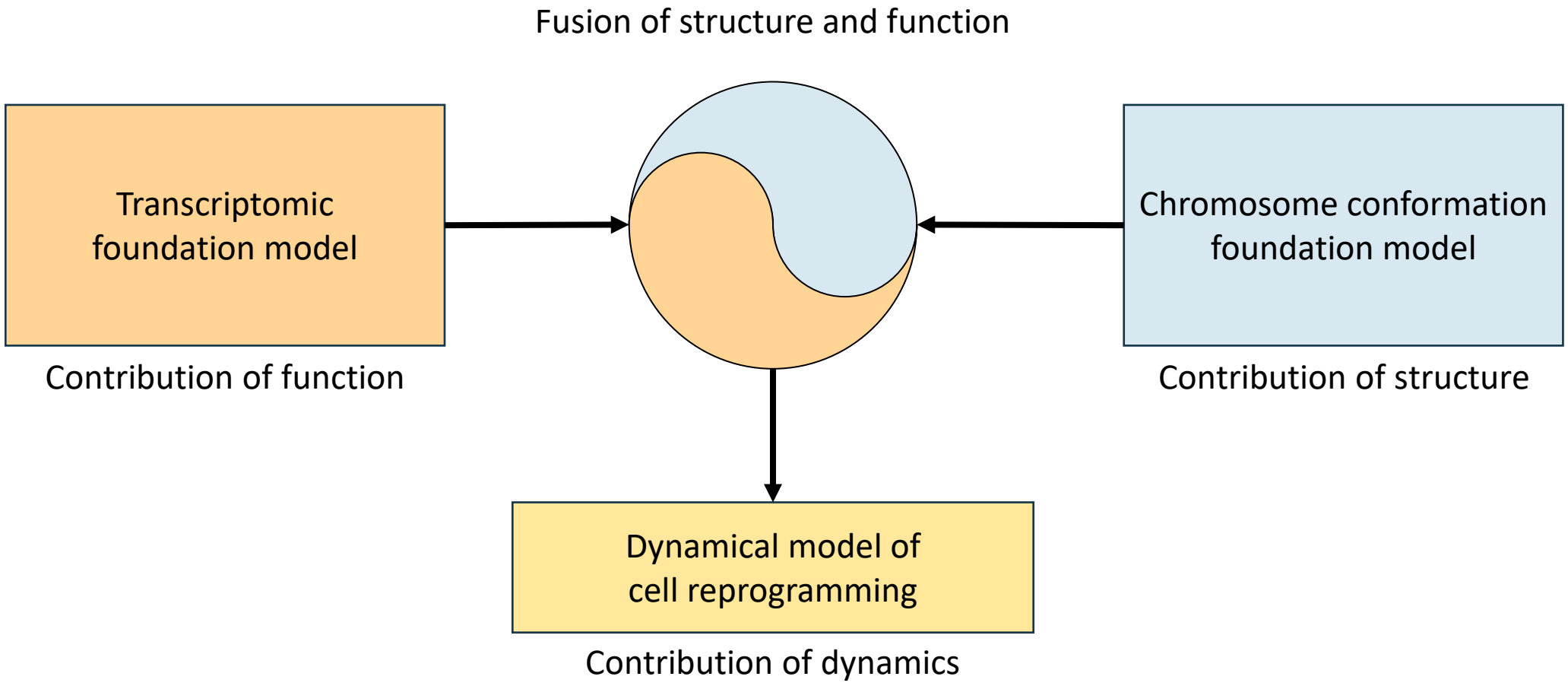
Protein sequence
- AlphaFold
- ESM-2, 3

ATAC-seq + DNA
- EPCOT
- GET

Spatial transcriptomics
- scGPT-spatial
- Nicheformer

Genome structure remains underexplored!

AI-powered state representation

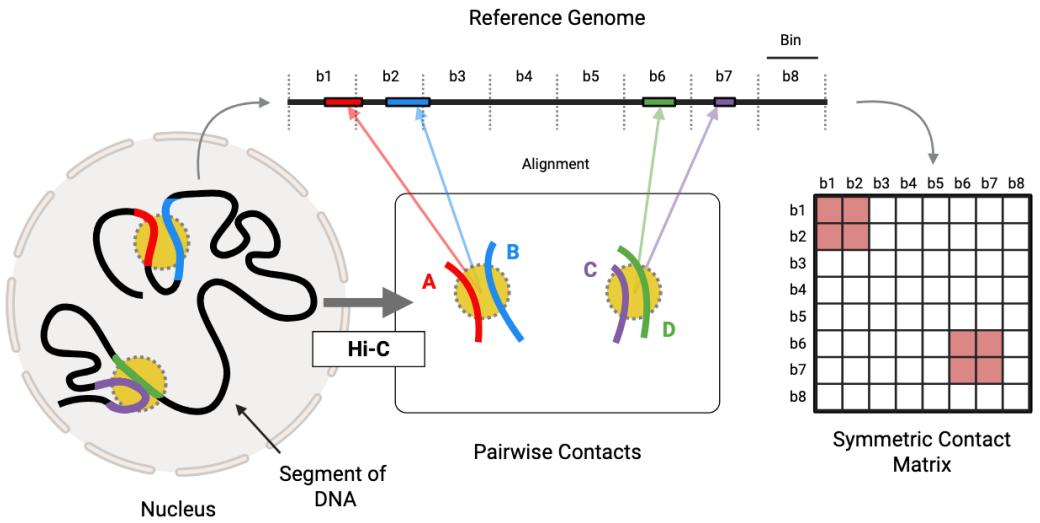


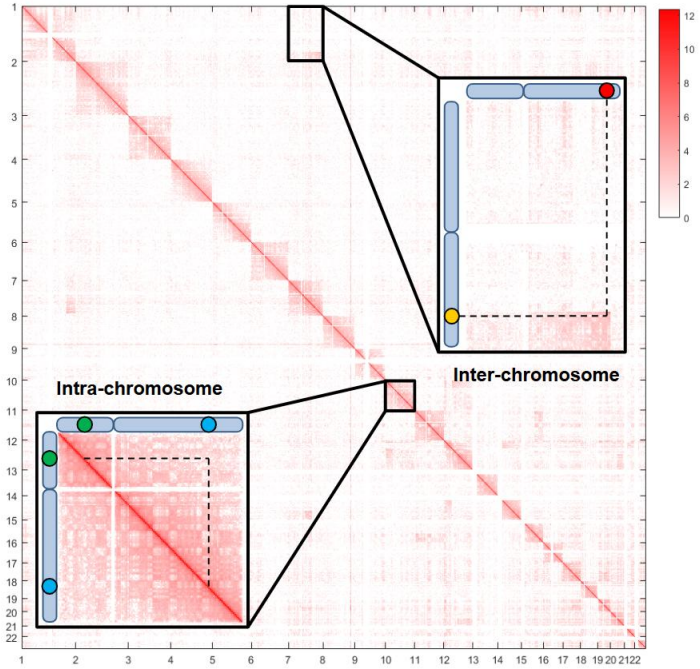
Introduction
Cell reprogramming

High-throughput chromosome conformation capture (Hi-C)

ARCH3D: Architecture and pre-training
Results
Conclusions and future work

Hi-C records the number of times two loci come into contact





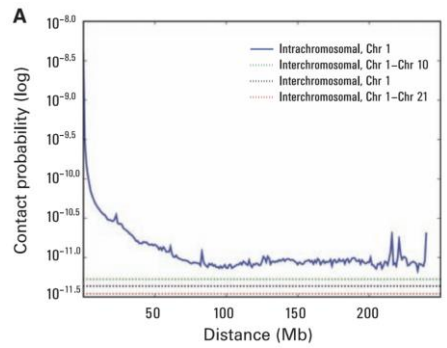
- Each entry in the contact matrix is known as a *pixel*
- Each pixel can be interpreted as a contact frequency

Block diagonal structure reflects chromosome territories

Lieberman-Aiden, Erez, et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *science* 326.5950 (2009): 289-293.

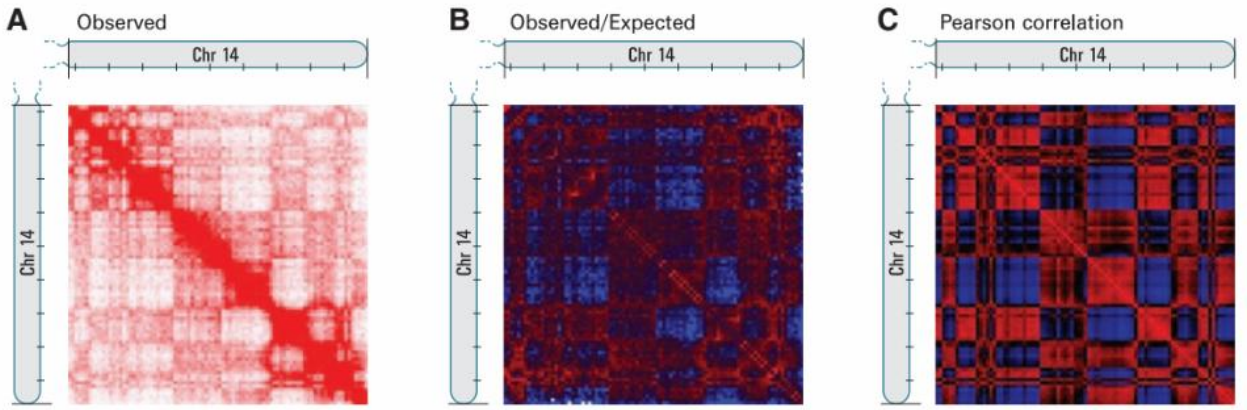
Original Hi-C paper

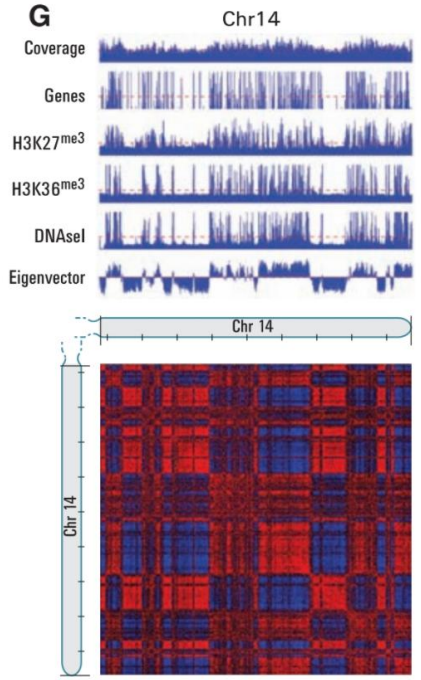
Plaid pattern reflects A/B compartmentalization of genome structure



Contact probability follows
a power-law scaling

Dividing every diagonal by its average
(observed/expected) mitigates the diagonal dominance

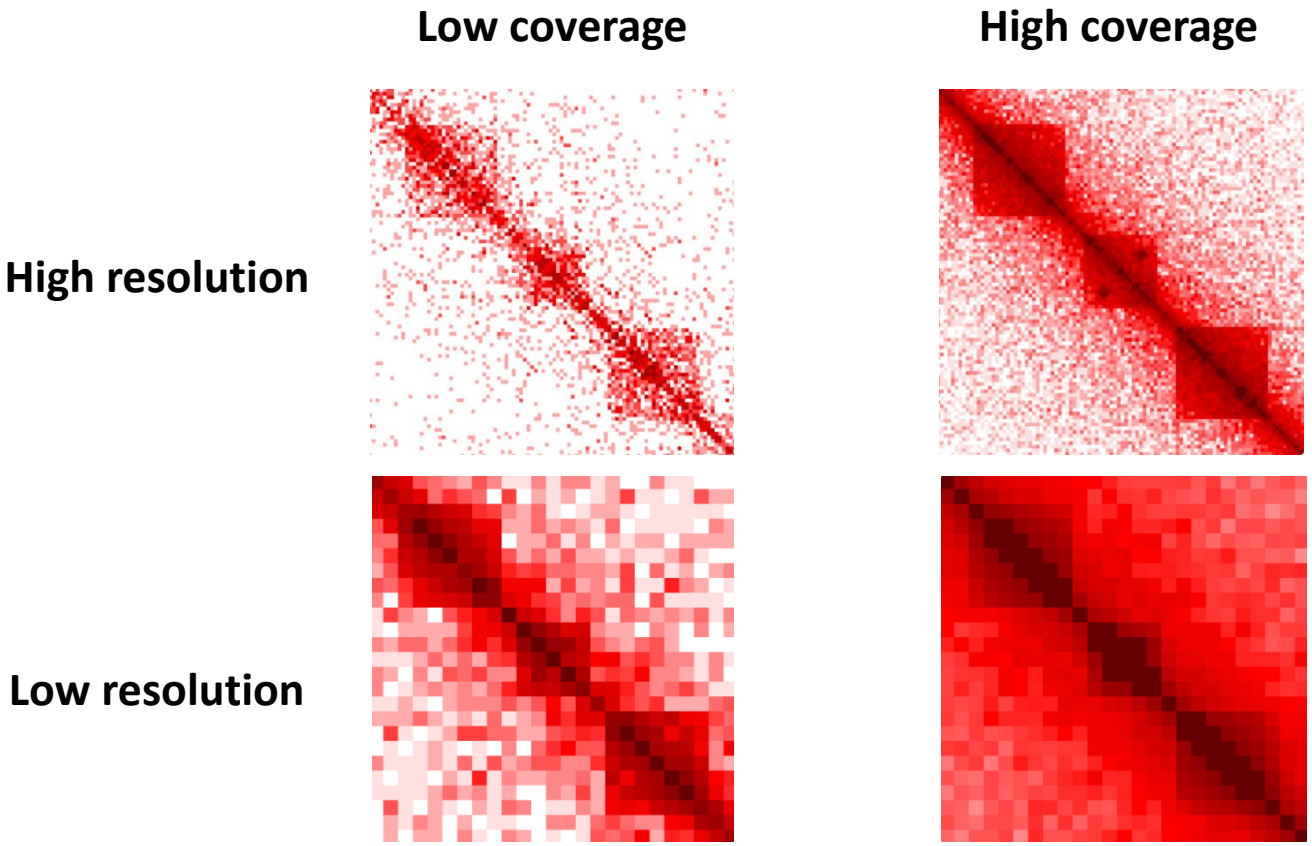




The first eigenvalue correlates
with chromatin accessibility

Lieberman-Aiden, Erez, et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* 326.5950 (2009): 289-293.

Hi-C: Resolution and coverage



- Low coverage cannot capture fine-scale structures (e.g., loops)
- However, low coverage can represent low-resolution Hi-C with similar accuracy as the high-coverage experiment

Introduction
Cell reprogramming
High-throughput chromosome conformation capture (Hi-C)

ARCH3D: Architecture and pre-training

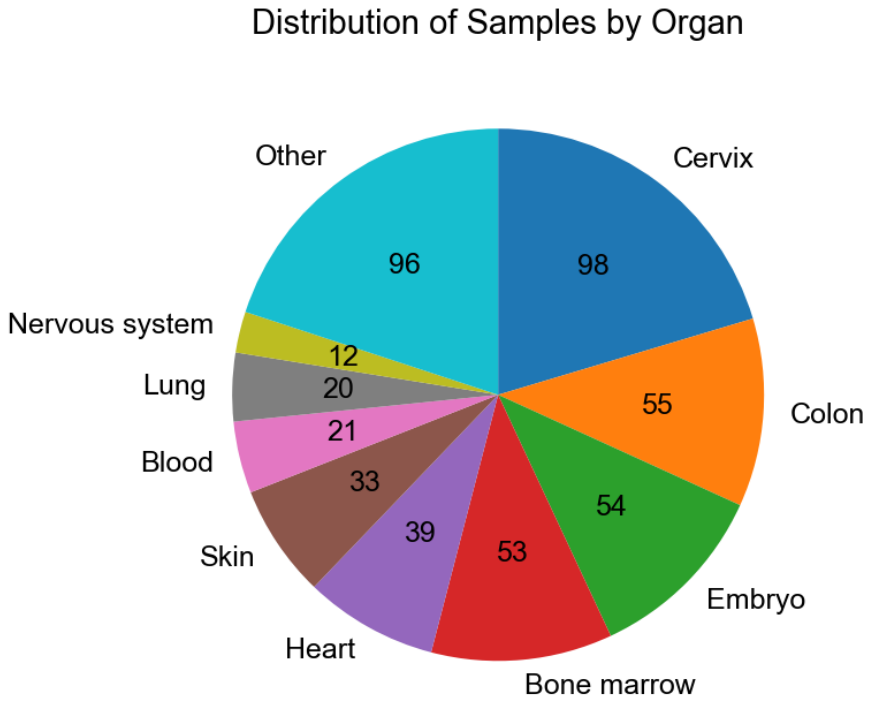
Results
Conclusions and future work

Pre-training corpus

Consortia:
- 4DNucleome
- ENCODE

Experiments:
- In-situ Hi-C
- Dilution Hi-C
- DNase Hi-C

Preprocessing:
- KR normalization
- Observed/expected



Other organs

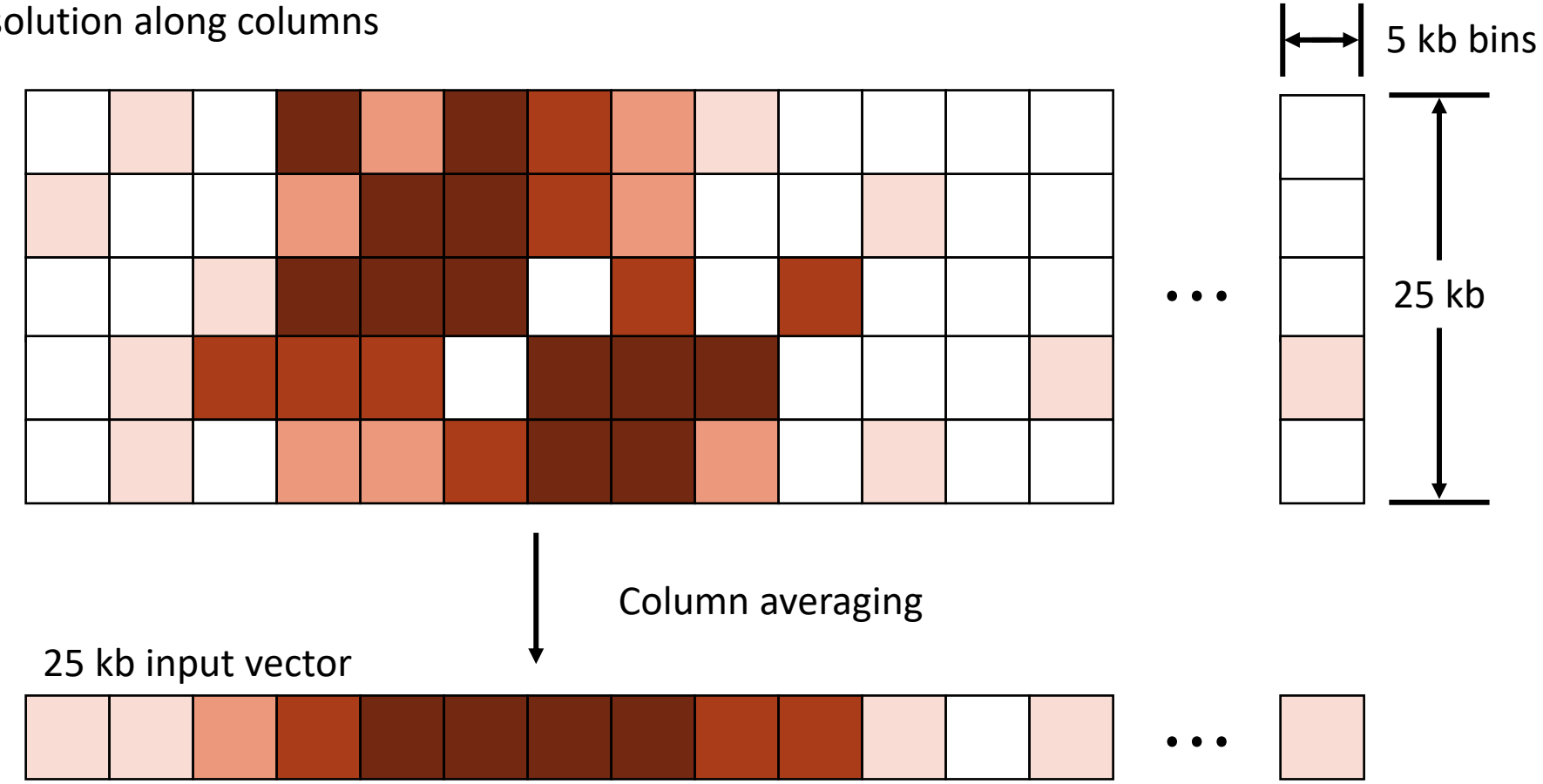
Muscle: 9	Thymus: 4
Breast: 9	Ovary: 4
Brain: 9	Spinal cord: 2
Uterus: 7	Esophagus: 2
Pancreas: 7	Spleen: 2
Kidney: 7	Bone: 1
Blood vessel: 7	Eye: 1
Nose: 6	Stomach: 1
Adrenal gland: 5	Testis: 1
Liver: 5	Thyroid: 1
Prostate: 5	Vagina: 1

481 total experiments (> 10M contacts)

Tokenization scheme

- Represents genomic loci, not patches
- Permits loci of any length (multiple of 5kb)
- Retains high resolution along columns

Locus lengths:
- 5 kb
- 10 kb
- 25 kb
- 50 kb
- 100 kb
- 250 kb
- 500 kb
- 1 Mb

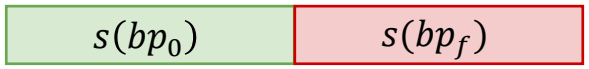


Biology-informed encodings provide the model with positional information

Genomic locus:

chr: bp_0 chr: bp_f

Base pair encodings



s is a vector-valued sinusoidal encoding with ith entry defined as:

if i is even:

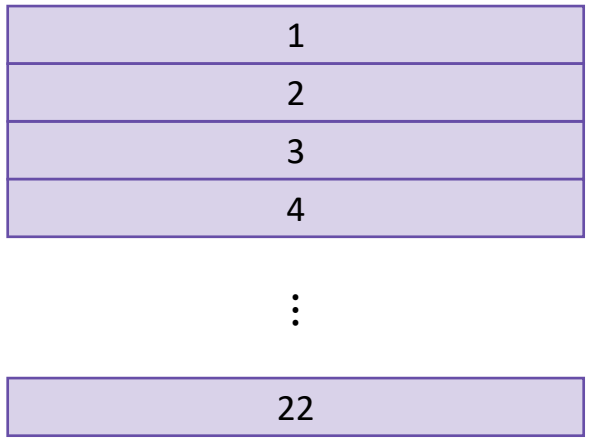
$$s_i(bp) = \sin\left(\frac{bp}{1000^{6i/d_{model}}}\right),$$

if i is odd:

$$s_i(bp) = \cos\left(\frac{bp}{1000^{6i/d_{model}}}\right)$$

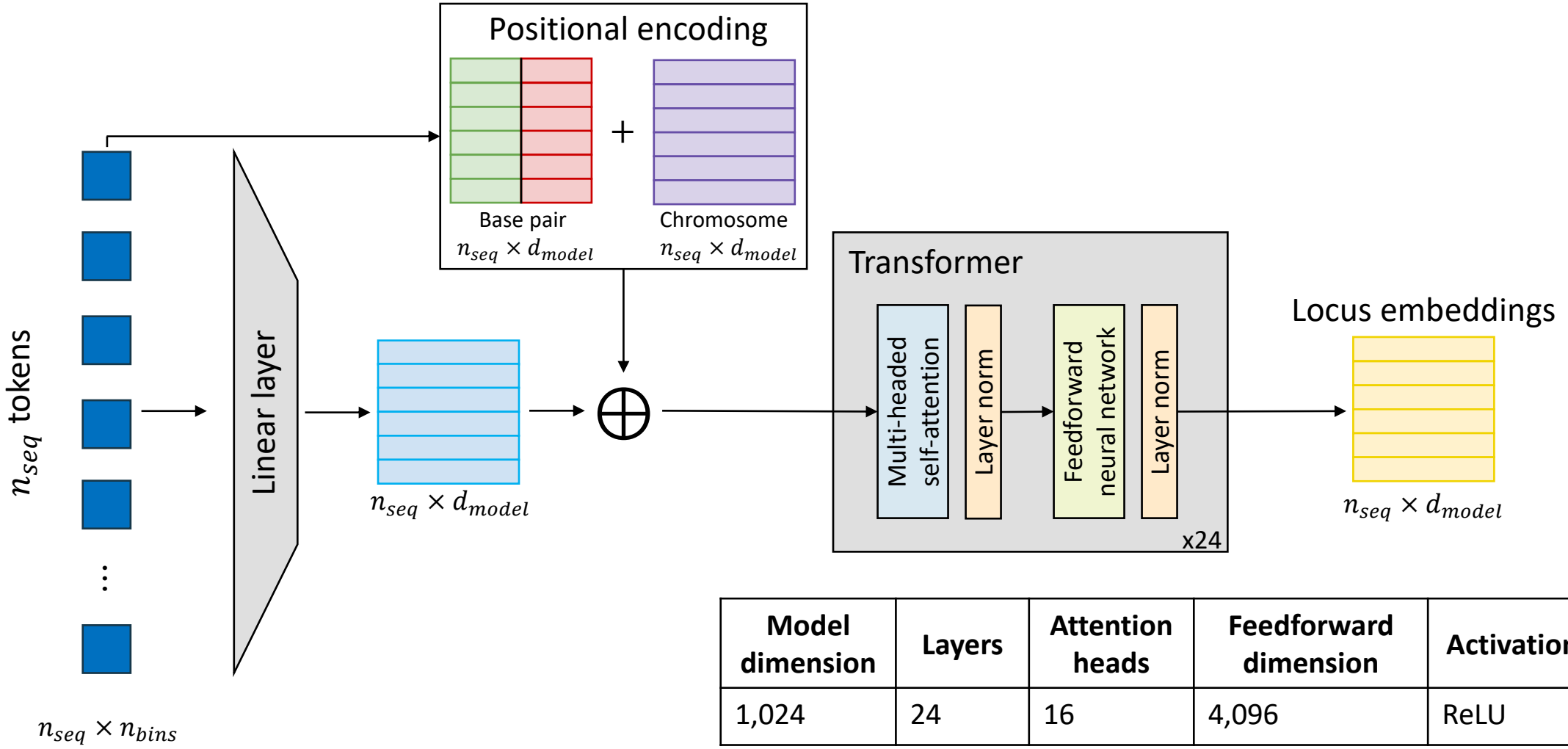
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Chromosomal encodings



Final positional encoding:

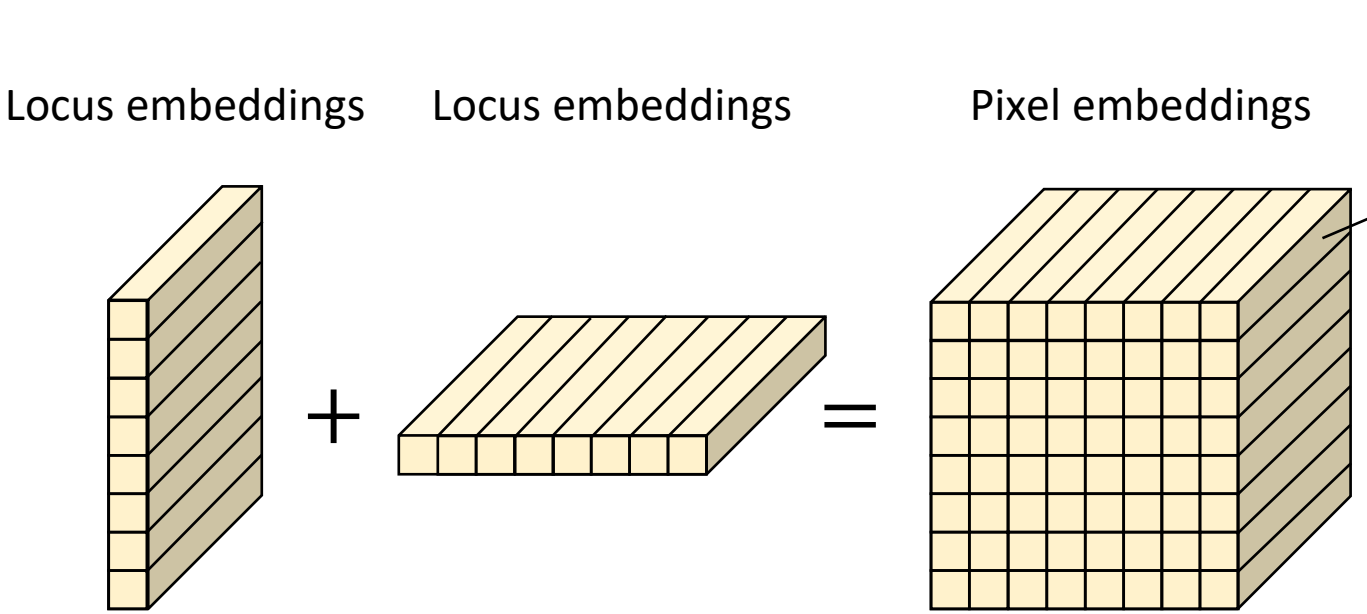
ARCH3D architecture

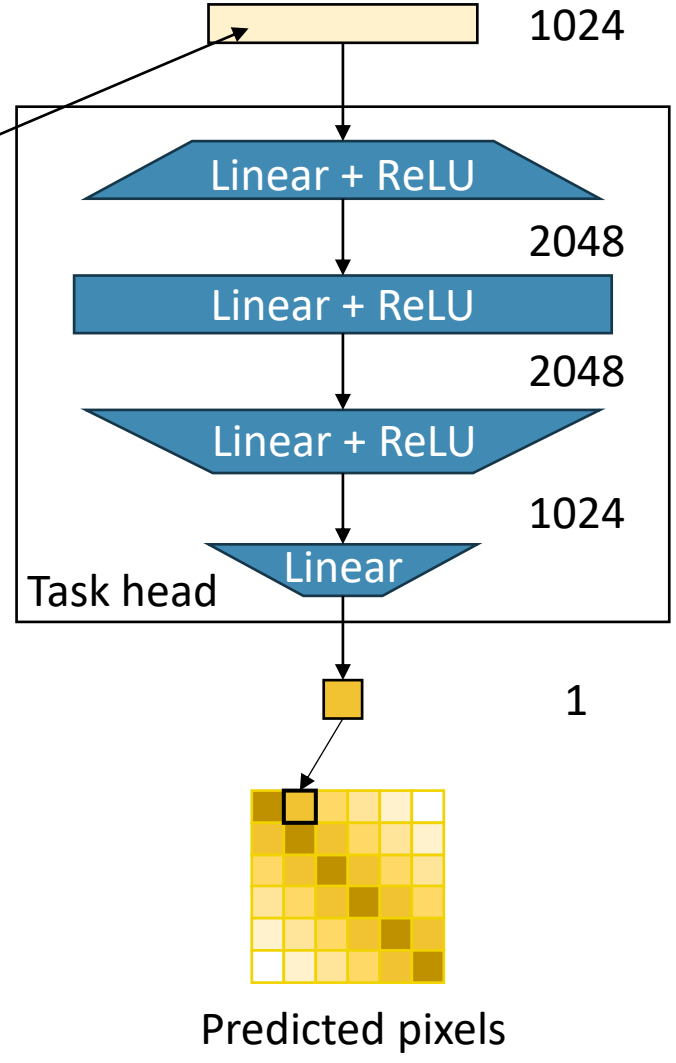


Model dimension	Layers	Attention heads	Feedforward dimension	Activation
1,024	24	16	4,096	ReLU

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 conference of the NAACL: human language technologies, volume 1 (long and short papers). 2019.

Task head architecture



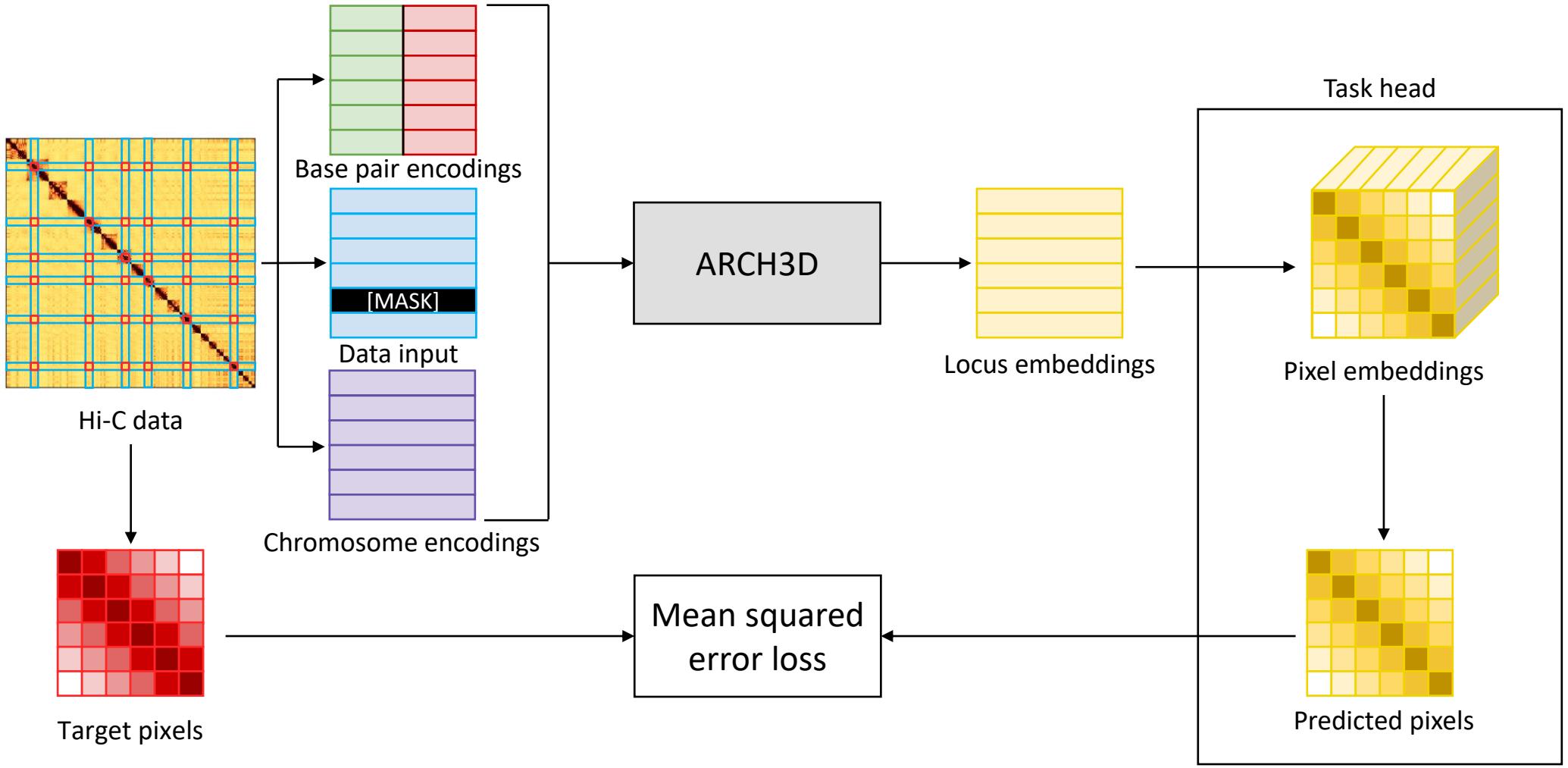


Locus embeddings are transformed to pixel embeddings through pairwise addition

$$p_{ij} = \ell_i + \ell_j, \qquad i, j = 1, \ldots, n_{seq}$$

p_{ij} is the ijth pixel embedding and ℓ_i the ith locus embedding

Pre-training task: Masked locus modeling



Training approach

University of Michigan Lighthouse HPC Cluster

17 nodes, each with:

- 8 NVIDIA H100 GPUs (80 GB VRAM)
- 1 TB RAM
- 96 cores

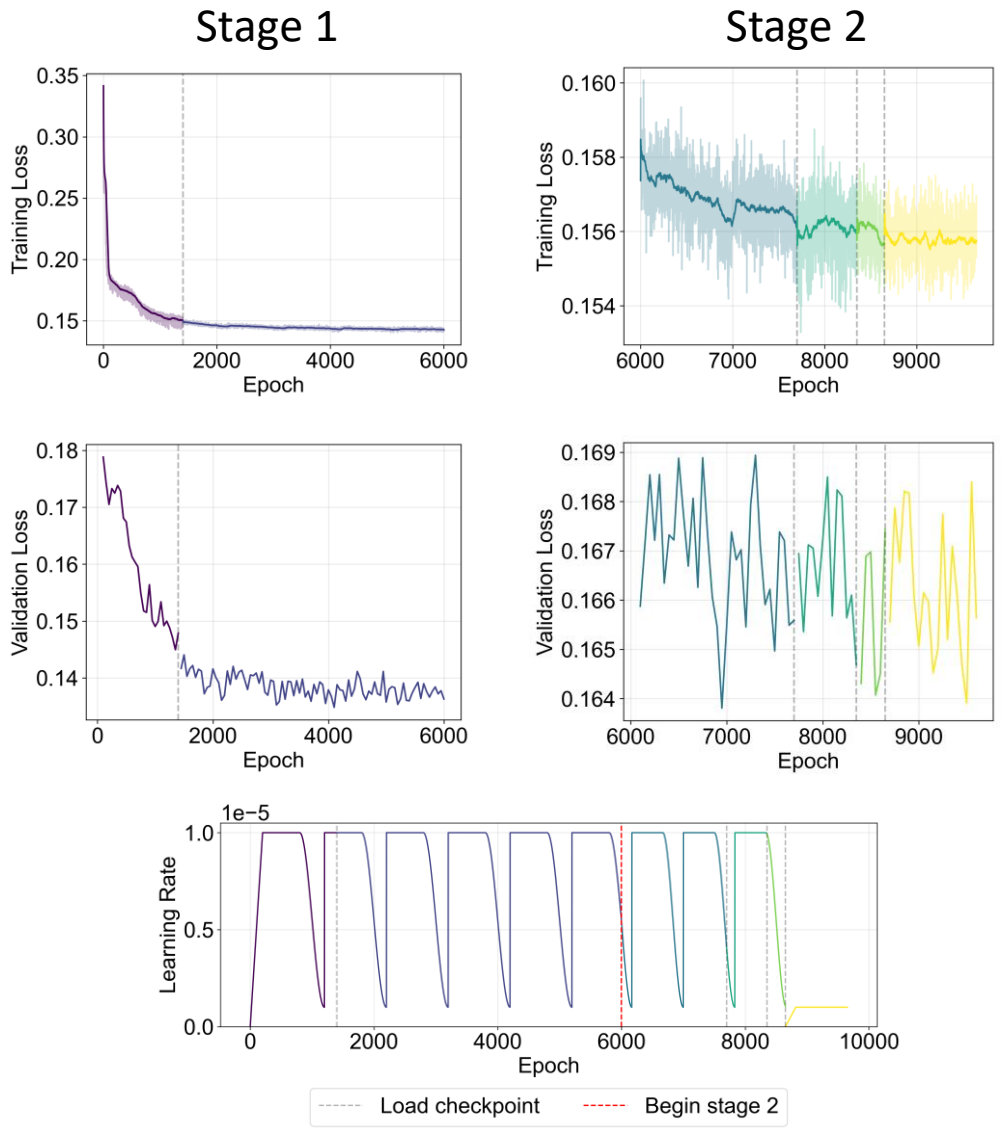
Optimizer: Adam

Learning rate schedule:

1. Linear warmup to 1e-5 over 500 steps
2. Constant 1e-5 for 3,000 steps
3. Cosine anneal to 1e-6 over 2,000 steps
4. Repeat 2—3

Stage	GPUs	RAM (TB)	Epochs	Time (h)	GPU hours
1	8	1.0	6,000	504	4,032
2	16 / 32	3.2 / 4.0	5,700	384	9,600

Stage 1: 194 Hi-C experiments; **Stage 2:** 481 Hi-C experiments



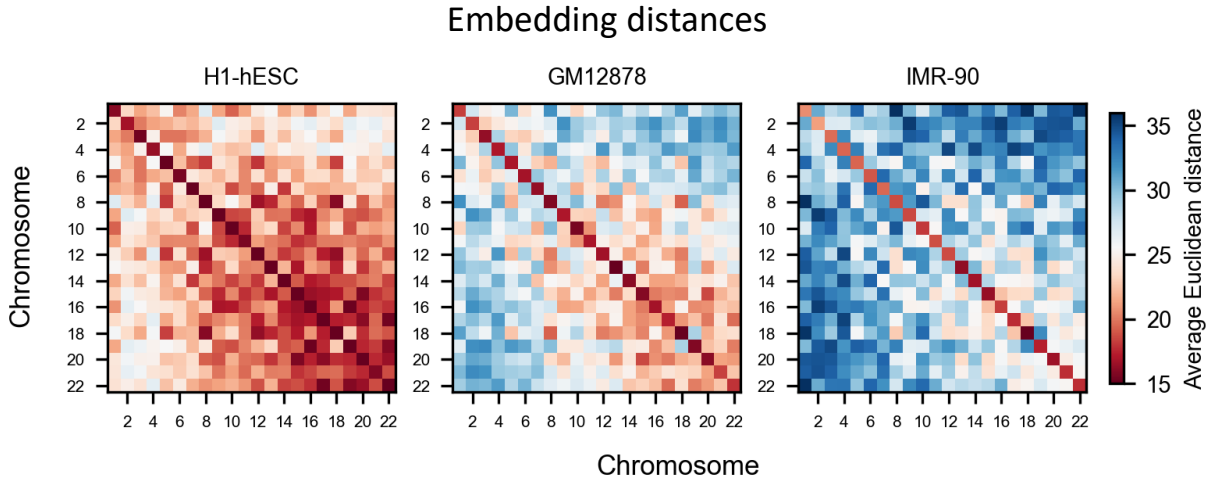
In final run, learning rate held at 10% of max.

Introduction
Cell reprogramming
High-throughput chromosome conformation capture (Hi-C)
ARCH3D: Architecture and pre-training

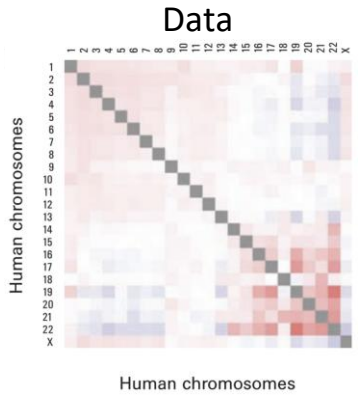
Results

Conclusions and future work

Positioning of embeddings reflects genomic structure



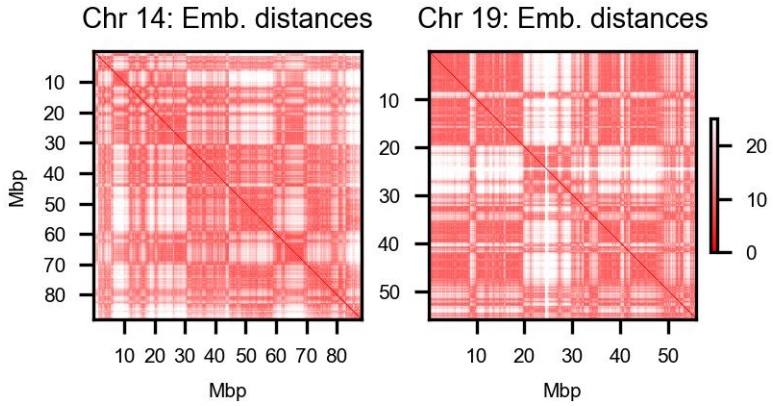
Average distance between embeddings within and across chromosomes

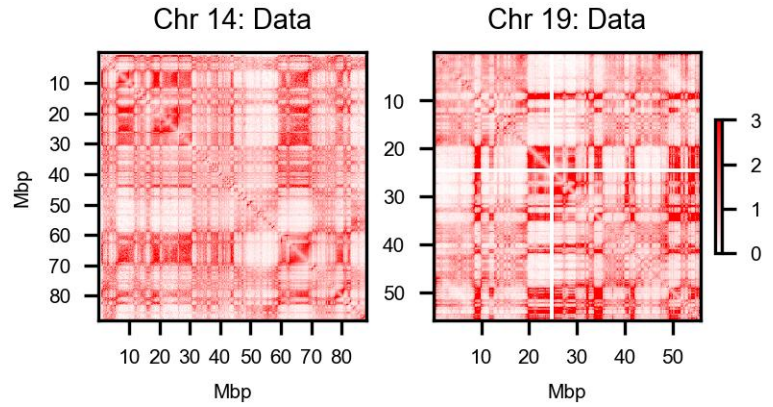


Average inter-chromosomal contact probabilities

Main observations:

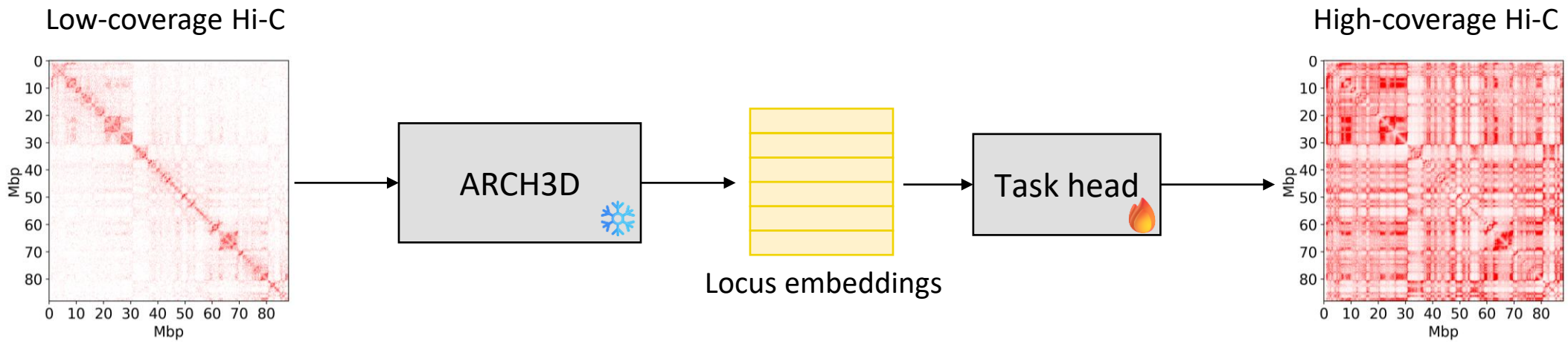
1. Diagonal blocks show shortest average distance, similar to chromosome territories
2. More differentiated cell lines show greater distance between inter-chromosomal embeddings
3. Embeddings from smaller chromosomes are positioned closer together than embeddings from larger chromosomes, mirroring experimental data





Lieberman-Aiden, Erez, et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *science* 326.5950 (2009): 289-293.

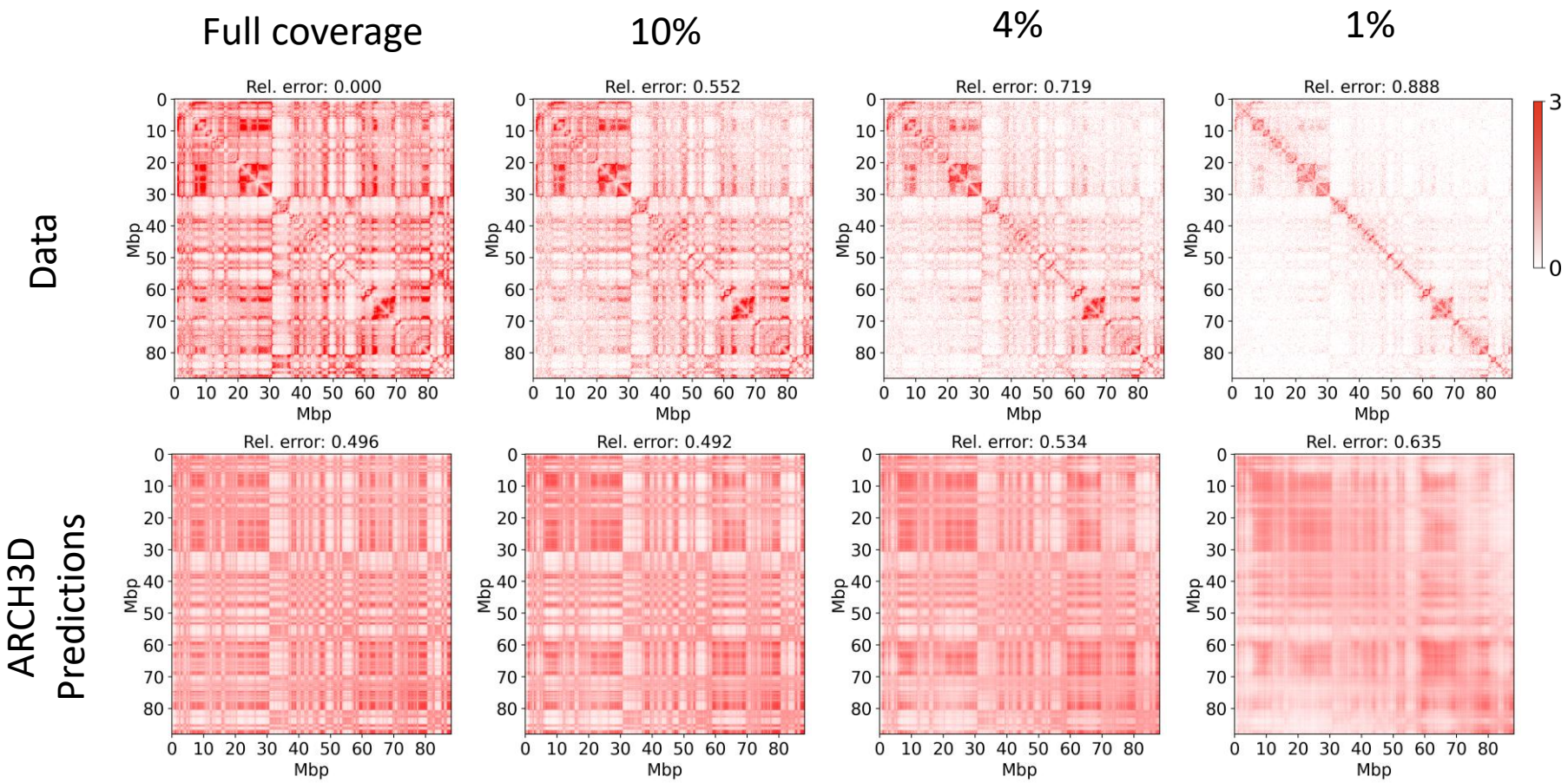
Resolution enhancement training scheme



Training:

- **Cell:** GM12878
- **Coverage:** 1%, 10%, 100%
- **Locus lengths:** 10kb, 25kb, 50kb, 100kb, 500kb, 1Mb

ARCH3D predictions degrade gracefully with decreasing coverage

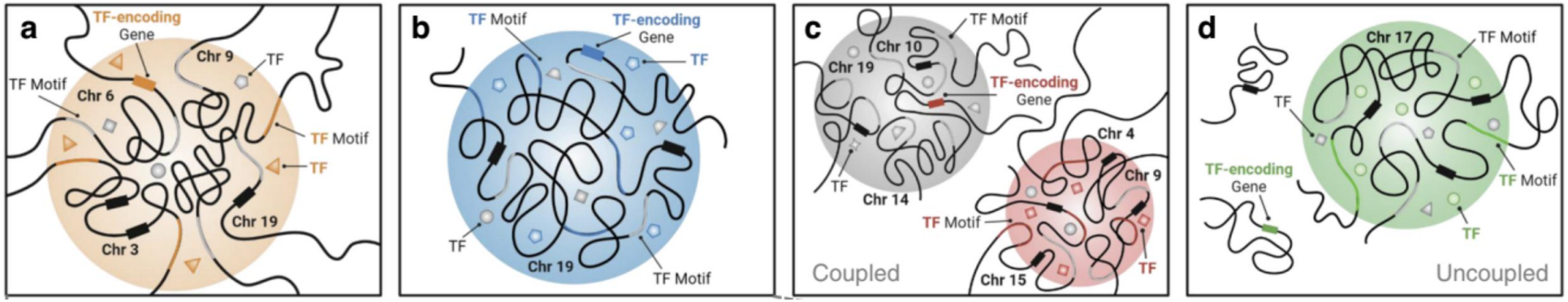


- **Cell:** IMR-90
- **Chromosome:** 14
- **Resolution:** 100 kb
- **HC contacts:** 1B

$$\text{Rel. error} = \frac{\|LC - HC\|_F}{\|HC\|_F},$$ where LC is data/predictions from low-coverage and HC is high-coverage data

Evidence suggests genes cluster into transcription factories

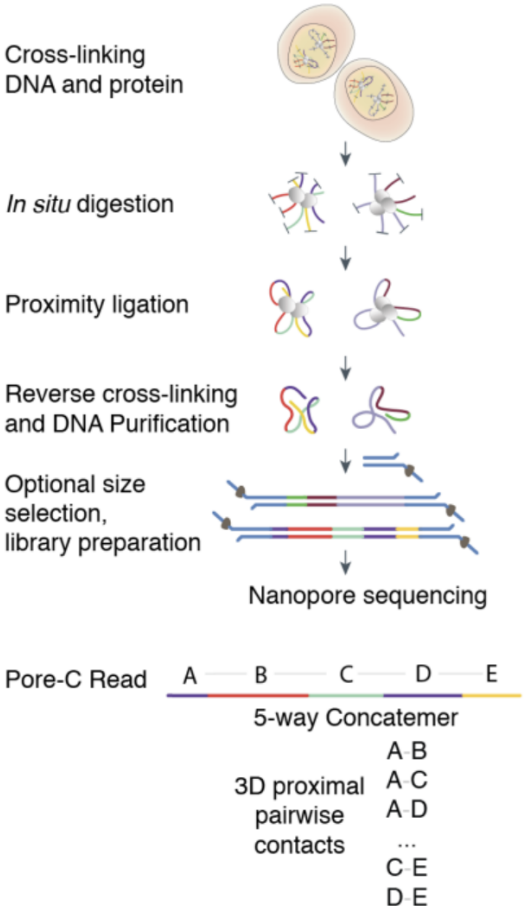
- Gene transcription is localized to a small number of sites known as “transcription factories”
- Genes within a transcription factory are co-regulated
- Pore-C records multi-way interactions using long-read sequencing



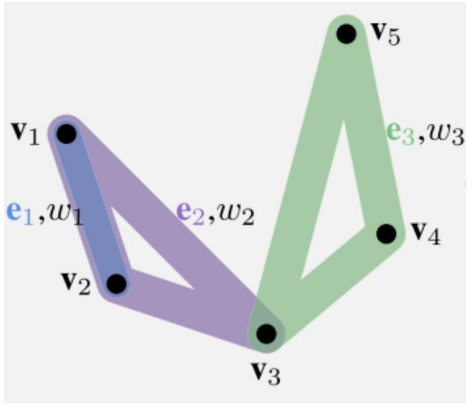
Dotson, Gabrielle A., et al. "Deciphering multi-way interactions in the human genome." *Nature Communications* 13.1 (2022): 5498.

Pore-C creates a hypergraph

Experimental procedure

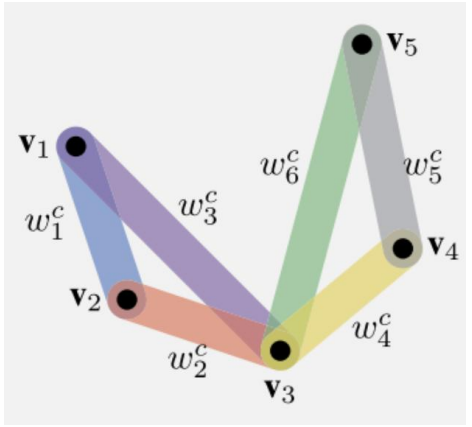


Multi-way interactions
(Pore-C)



Clique expansion

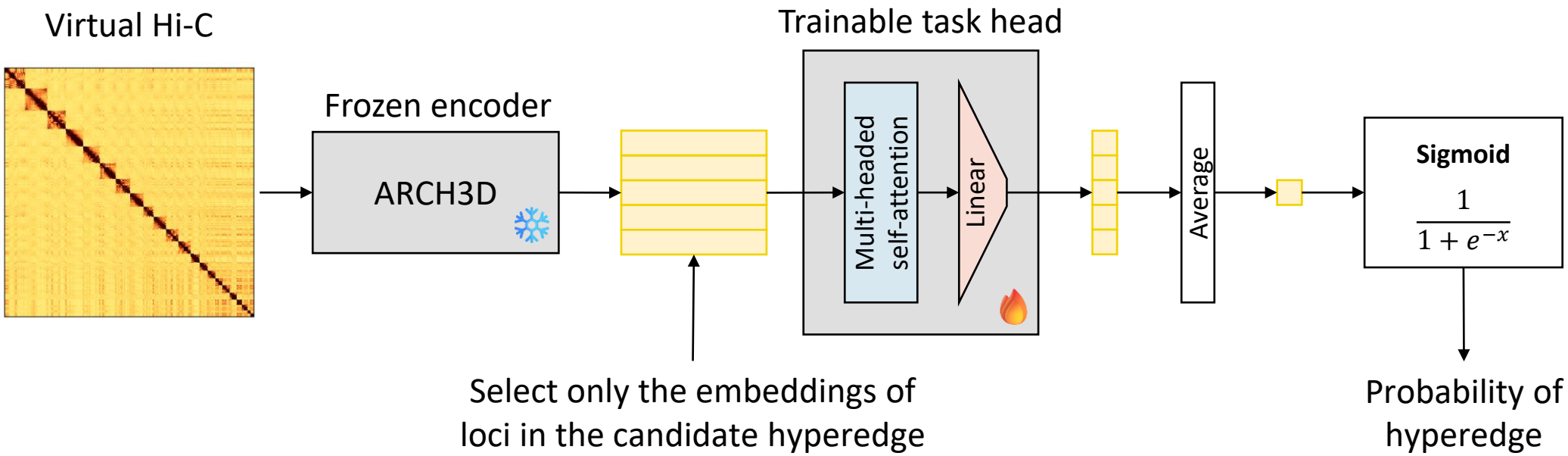
Pairwise interactions
(Virtual Hi-C)



Clique-expansion gives an approximation of Hi-C referred to as “virtual Hi-C”

Surana, Amit, Can Chen, and Indika Rajapakse. "Hypergraph similarity measures." *IEEE Transactions on Network Science and Engineering* 10.2 (2022): 658-674.

Hyperedge prediction training scheme



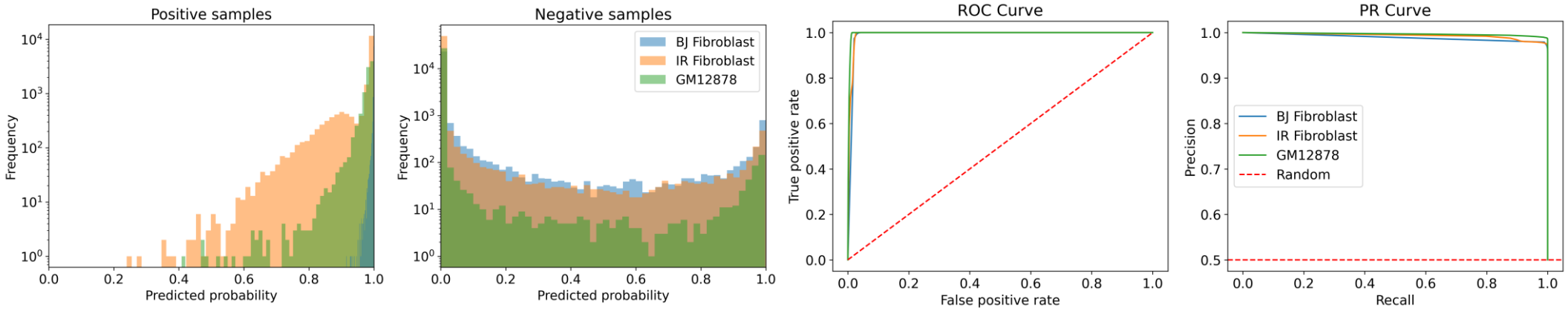
Three negative samples are generated for every positive sample

- Replace a random selection of nodes in the positive sample with randomly-chosen nodes from the same chromosome

Zhang, Ruochi, and Jian Ma. "MATCHA: probing multi-way chromatin interaction with hypergraph representation learning." *Cell systems* 10.5 (2020): 397-407.

Prediction of multi-way interactions generalizes to unseen cell lines using virtual Hi-C

Log scale!



Training set:

- GM12878

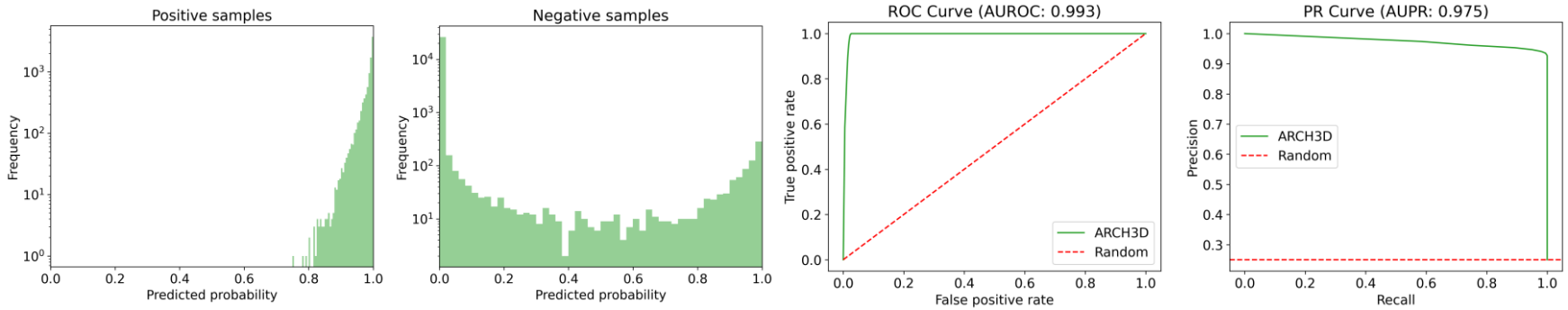
Testing set:

- BJ fibroblasts
- IR fibroblasts

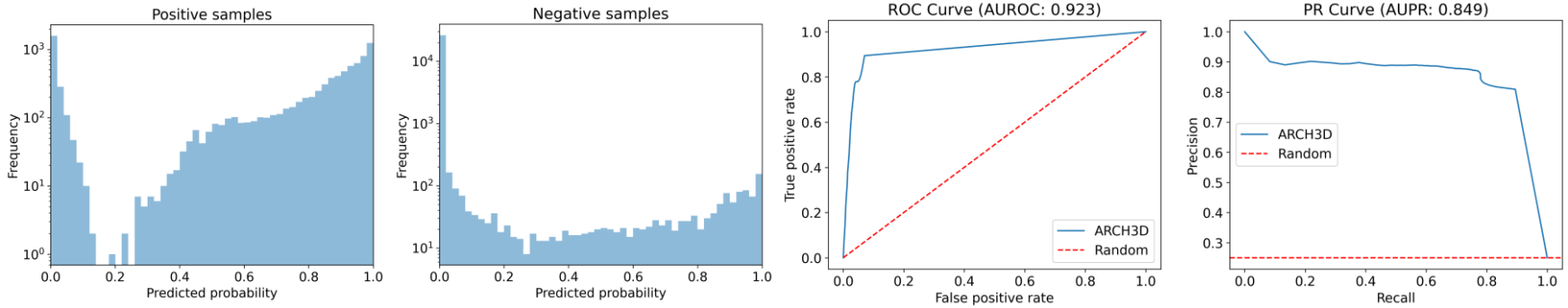
	AUROC	AUPR
GM12878	0.997	0.997
BJ fibroblasts	0.989	0.989
IR fibroblasts	0.993	0.994

ARCH3D predicts Pore-C directly from Hi-C

GM12878
216M Hi-C contacts

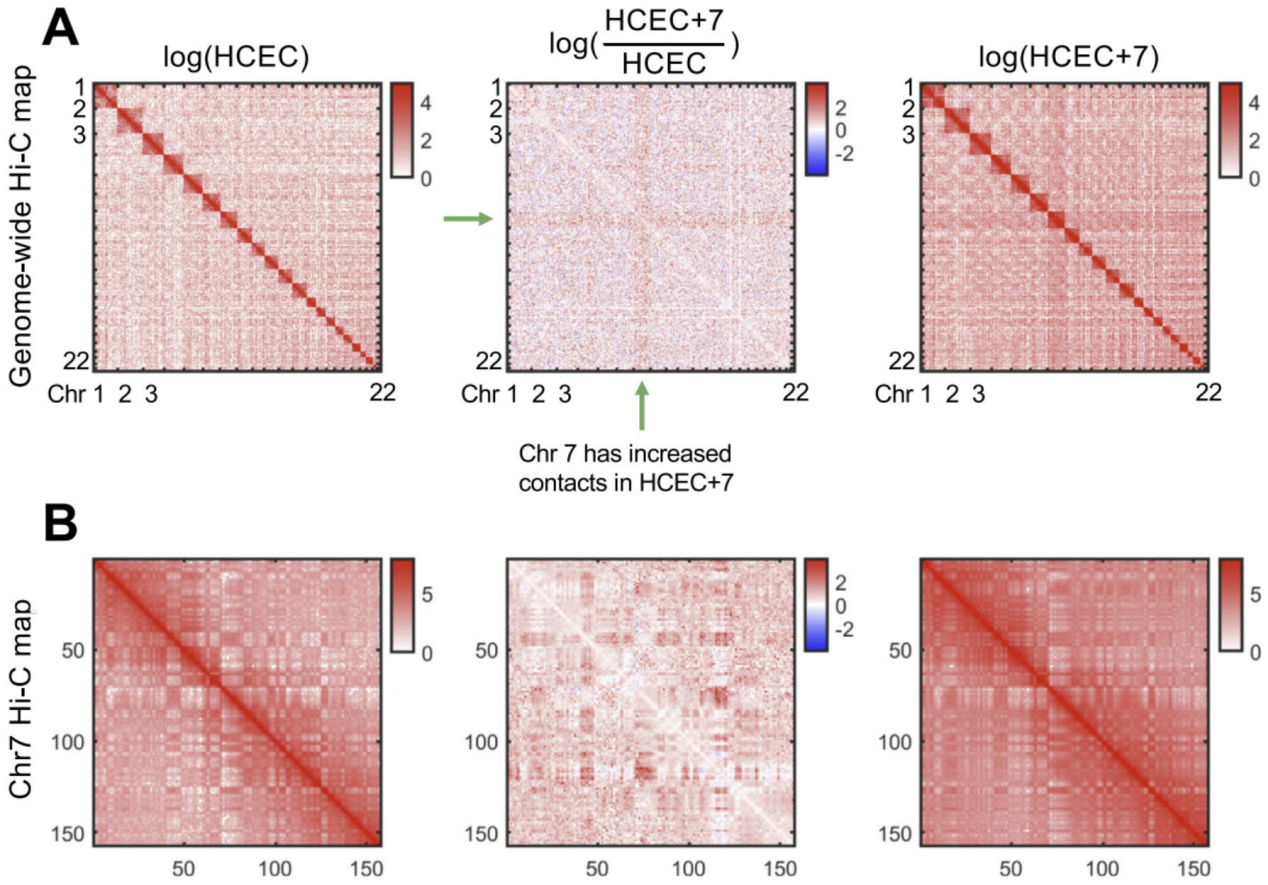


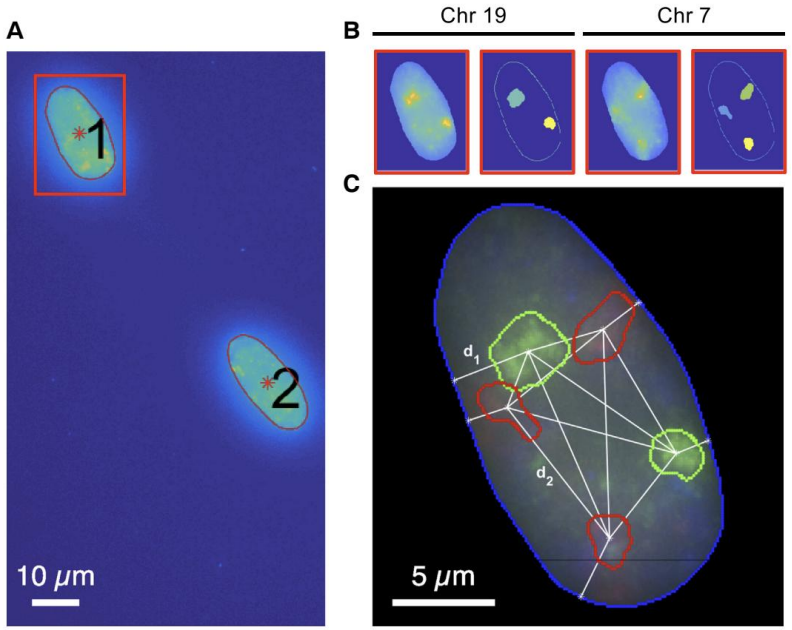
BJ Fibroblast
106M Hi-C contacts



BJ fibroblast Hi-C was not contained in pre-training corpus—totally new to ARCH3D!

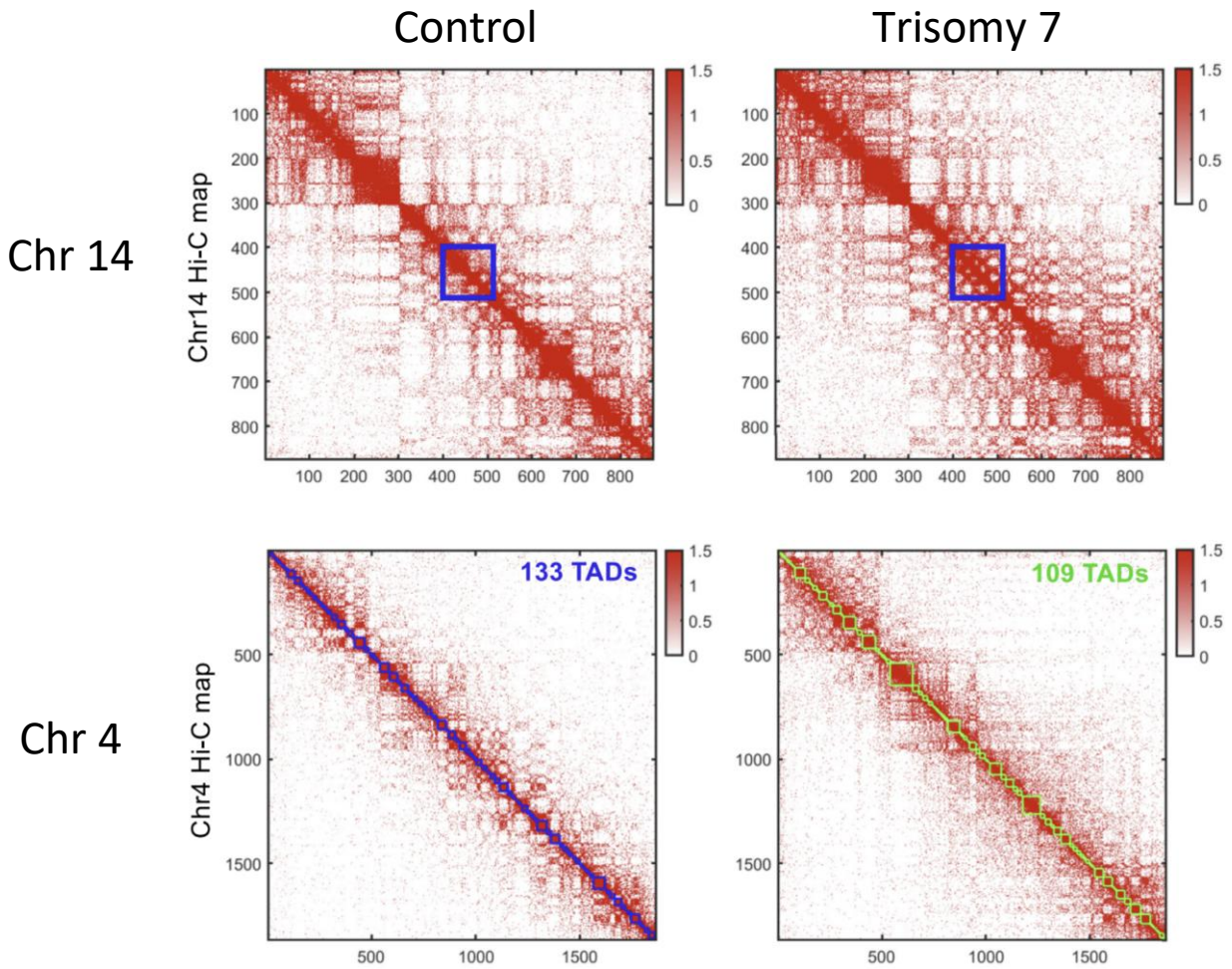
Perturbations in genome architecture: extra chr7





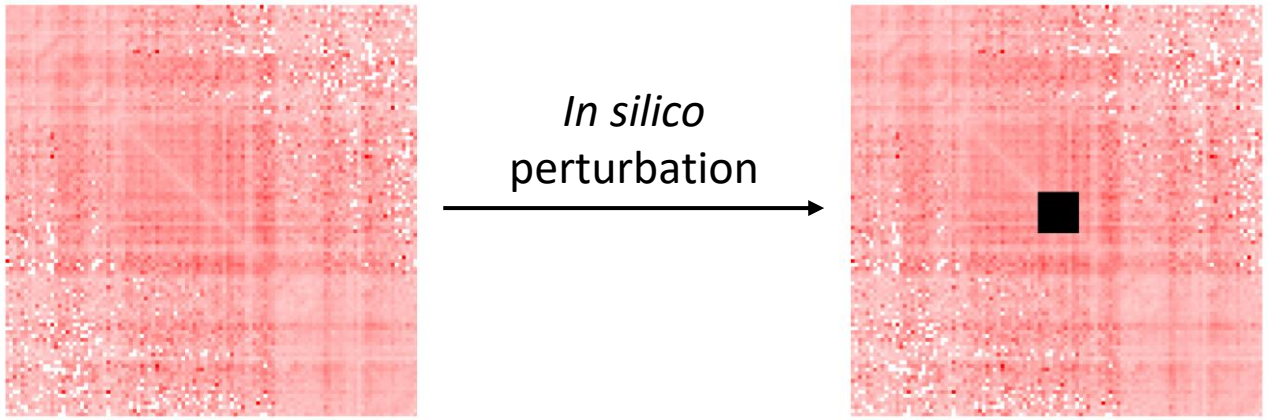
Braun, Rüdiger, et al. "Single chromosome aneuploidy induces genome-wide perturbation of nuclear organization and gene expression." *Neoplasia* 21.4 (2019): 401-412.

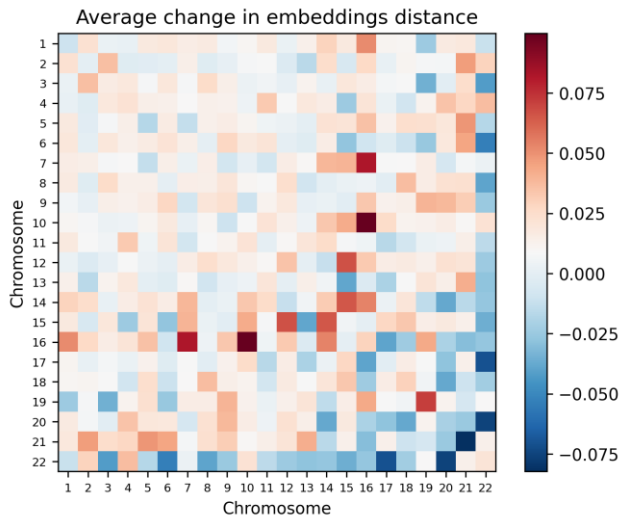
Introduction of a third chromosome 7 yields genome-wide disruptions in structure

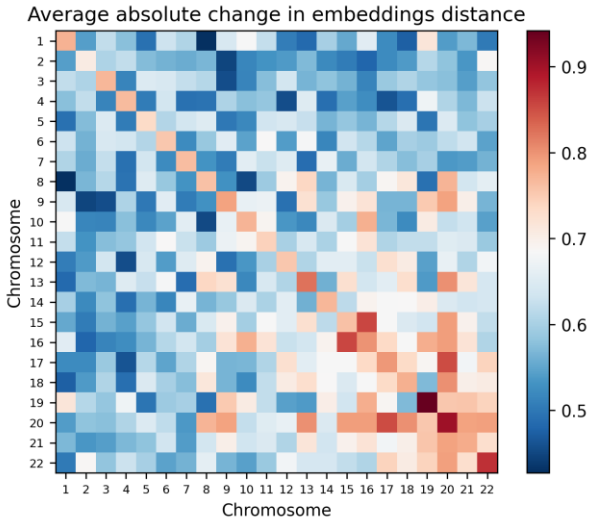


- Changes in compartmentalization
- Fewer TADs

Perturbation results







Smaller chromosomes show greater sensitivity to perturbations

Introduction
Cell reprogramming
High-throughput chromosome conformation capture (Hi-C)
ARCH3D: Architecture and pre-training
Results

Conclusions and future work

Conclusions

- The organization of ARCH3D's embedding space mirrors that of the nucleus
- ARCH3D enhances the coverage of low-coverage experiments
- ARCH3D identifies multi-way interactions from Hi-C data

Future work

- Integrate ARCH3D embeddings with transcriptomic embeddings
- Extend to single-cell Hi-C data

Funding

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