STAT 992: Foundation Models for Biomedical Data

Ben Lengerich

Lecture 13: Foundation Models for Genomics

March 16, 2026

Last Time: Imaging

- Massive unlabeled image archives
- Expensive annotation
- Representation reuse
- Distribution shift
- Shortcut learning

Key question: What invariances transfer across hospitals and datasets?

Today: Genomics

- What invariances exist in genomic data? Transfer across:
  - species
  - cell types
  - experiments
  - diseases
  - perturbations

If genomic meaning transfers, foundation models work.
If not, specialized models dominate.

Genomics is attractive for foundation models

- Massive biological datasets:
  - genome sequencing
  - RNA-seq
  - ATAC-seq
  - single-cell atlases
  - perturbation screens

But expensive labels…

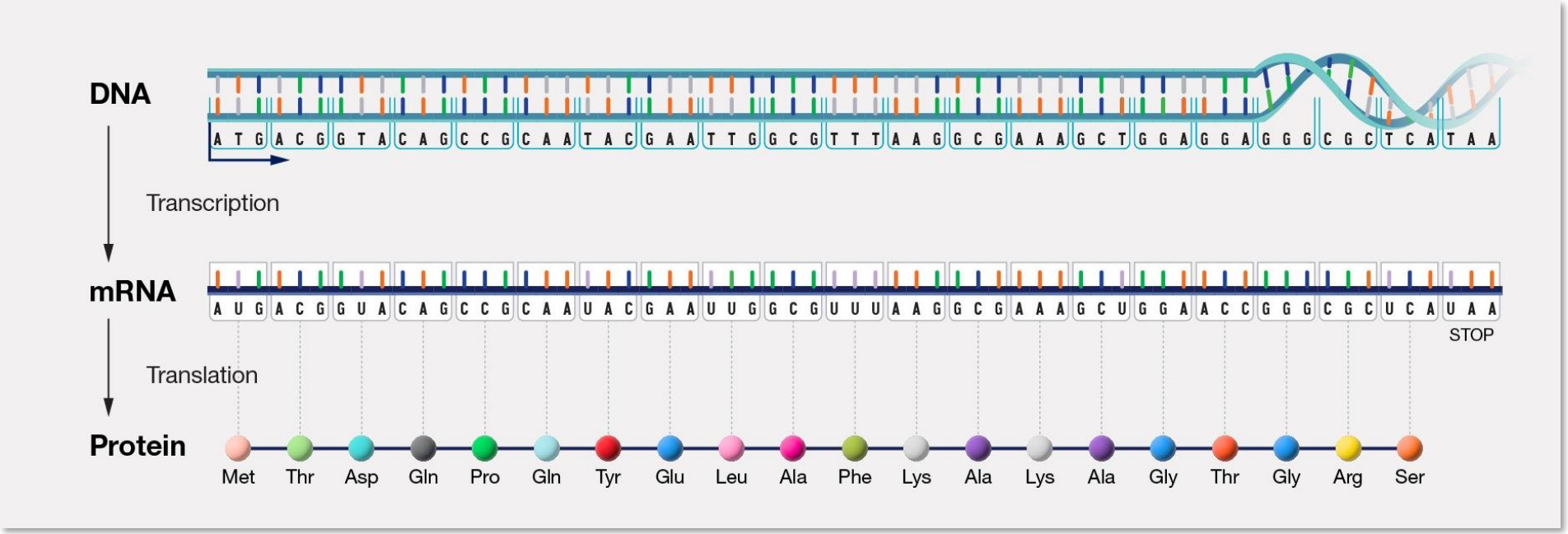
The supervision bottleneck in genomics

Biological measurements are often destructive.

- Many assays destroy the sample:
  - RNA-seq → cell lysis
  - ATAC-seq → chromatin fragmentation
  - ChIP-seq → DNA-protein extraction
  - CRISPR screens → system perturbation
- Consequence: Limited (usually one) measurements per cell
  - Multi-modal datasets are more rare in genomic; can we learn to merge predictions/embeddings from multiple uni-modal models?
- Consequence: Incomplete biological observations
  - We observe different modalities in different cells.
  - Foundation models can help integrate these?

A core difficult in genomics

- Sequence alone does not determine biological behavior



https://www.genome.gov/genetics-glossary/Central-Dogma

Implication for foundation models

- **Uni-modal models are limited, but we rarely observe multi-modal assays**

Genomics Regimes

Two main genomics regimes

- Sequence foundation models
  - Sequence-to-function models
- Cell-state foundation models
- Sometimes: multimodal regulatory models

Sequence foundation models

- Examples:
  - DNABERT-2
  - Nucleotide Transformer
  - Evo
  - HyenaDNA
- Goal: Learn regulatory grammar

- BPE tokenization
- Masked sequence modeling

- What it sounds like, large-scale pretraining, main “scaling study”
- Train on thousands of genomes
- Long-range genomic modeling without attention

Sequence foundation models

- What sequence models learns
  - TF motifs
  - Motif combinations
  - Enhancer grammar
  - Evolutionary signals
- Tokenization choices
  - Nucleotide tokens
  - k-mers
  - Learned tokens
- Common objectives
  - Masked sequence modeling
  - Autoregressive prediction

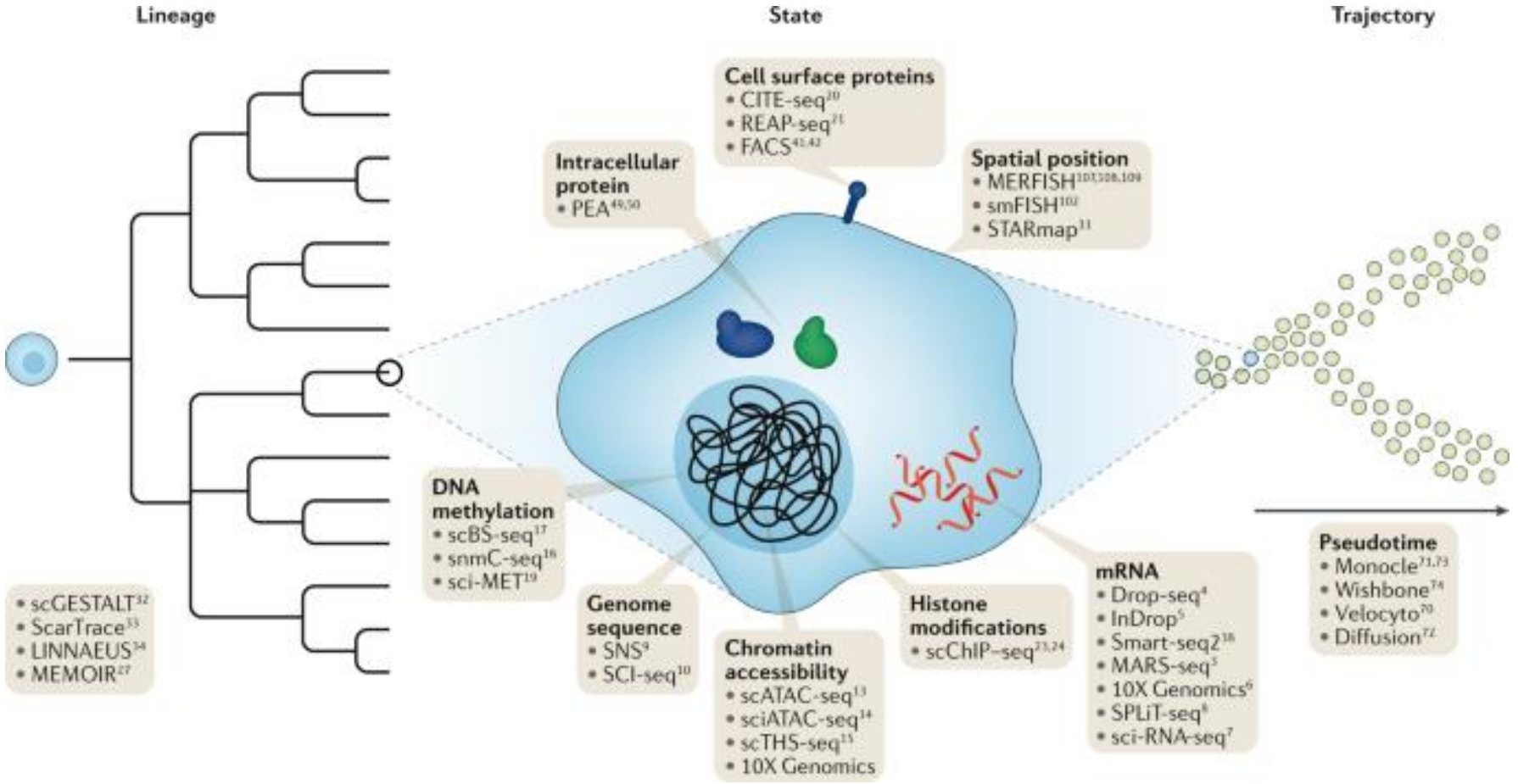
"DNA is a language"(?)

Sequence-to-function models

- Challenge: Long-range dependence
  - Genomic regulation often occurs across 100k+ base pairs.
  - DNALongBench
- How much context length is needed?

Cell-state foundation models

- The rise of single-cell assays



https://www.nature.com/articles/s41576-019-0093-7

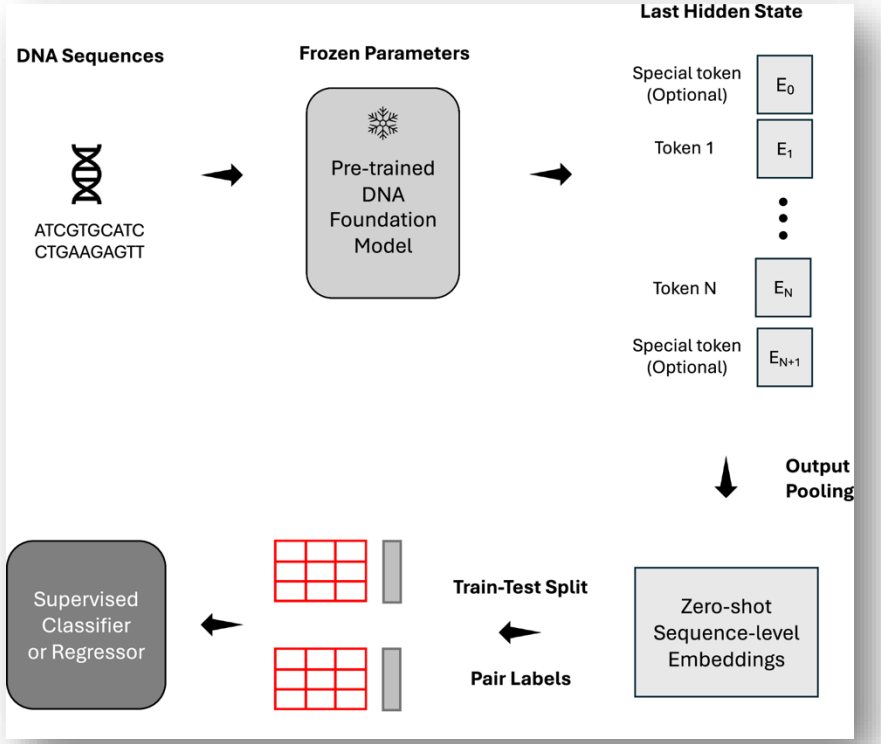
Cell-state foundation models

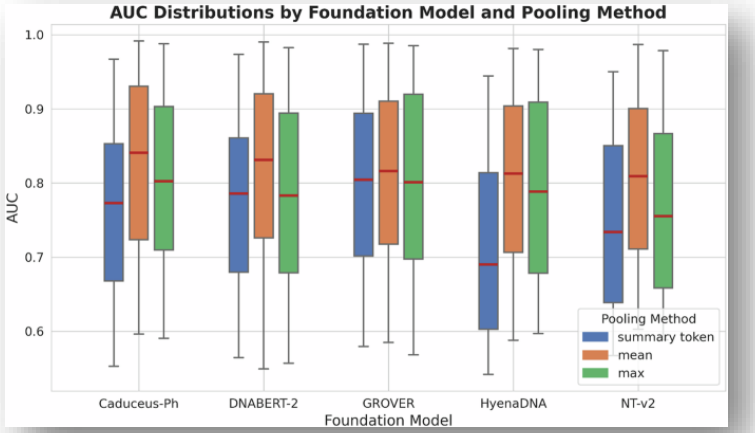
- Examples:
  - Geneformer
  - scGPT
- What these models learn:
  - Cell types
  - Regulatory programs
  - Differentiation paths

- Transformer for single-cell expression
- Trained on tens of millions of cells

Benchmarking DNA foundation models

- [Nature Communications 2025]

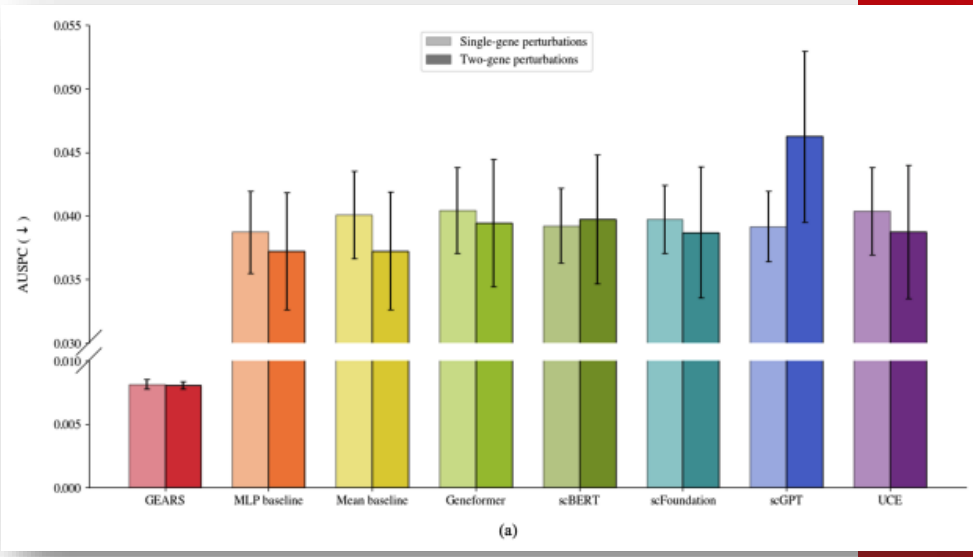




Model	Input Sequence Length	Average Prediction Correlation
DNABERT-2	6000 bp	0.121
NT-v2	6000 bp	0.122
HyenaDNA	6000 bp	0.122
Caduceus-Ph	6000 bp	0.123
GROVER	2048 bp	0.114
HyenaDNA-450K*	196K bp	0.137
Caduceus-Ph Long Sequence Input*	131K bp	0.127
Enformer*	196K bp	0.129

Perturbations as benchmarks

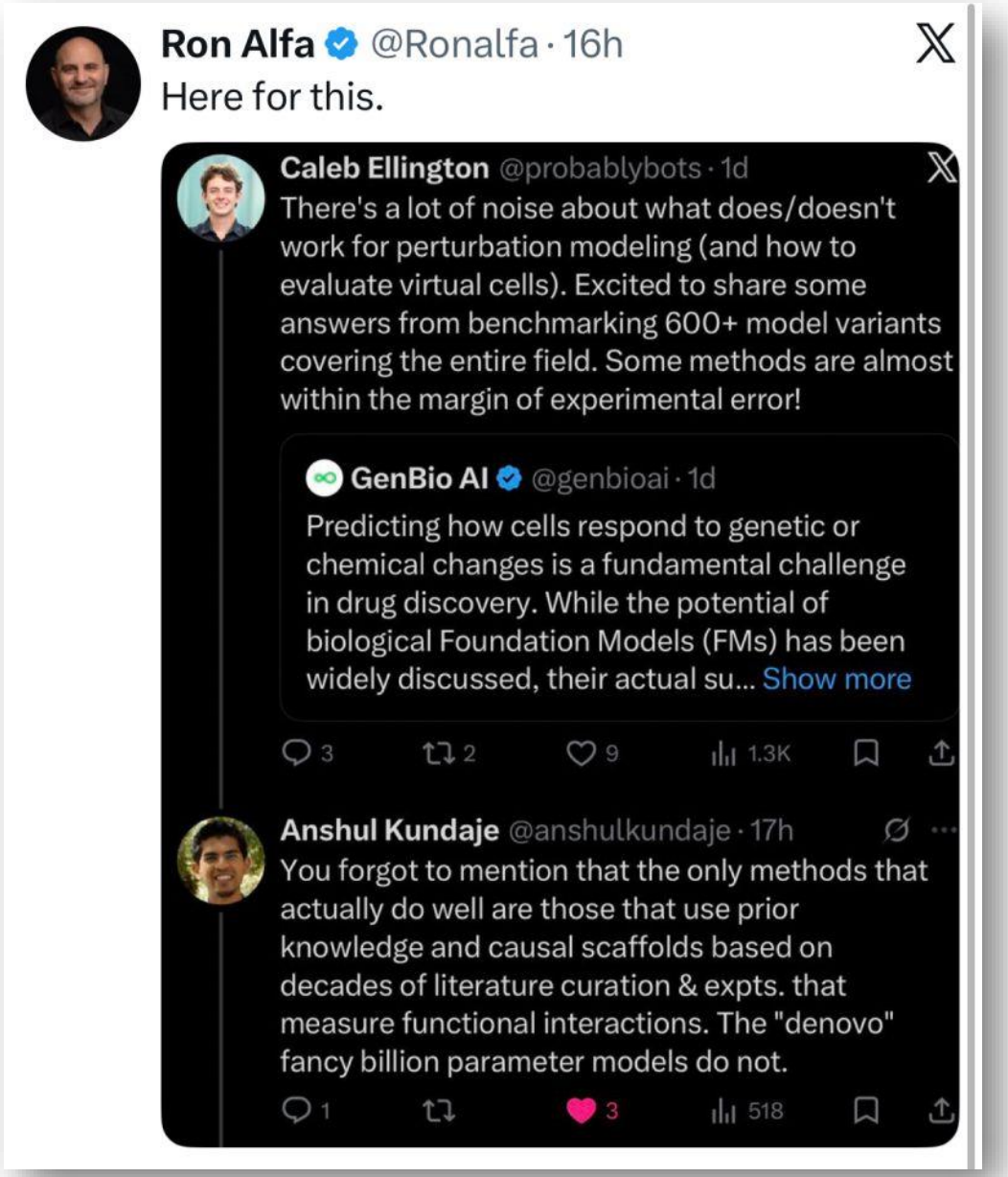
- ICML 2025
- Task: Predict gene expression changes after genetic perturbation
- Key result: Zero-shot embeddings from single-cell foundation models **provide limited improvement over simple baselines**.
- BMC Genomics 2025



In this study, we benchmarked two recently published foundation models, scGPT and scFoundation, against baseline models. Surprisingly, we found that even the simplest baseline model—taking the mean of training examples—outperformed scGPT and scFoundation. Furthermore, basic machine learning models that incorporate biologically meaningful features outperformed scGPT by a large margin. Additionally, we identified that the current Perturb-Seq benchmark datasets exhibit low perturbation-specific variance, making them suboptimal for evaluating such models.

The virtual cell vision

- Idea: Simulate cellular behavior with large models.
- Challenges
  - causal biology
  - perturbation data?
  - context?



The big picture: Imaging vs Genomics

- Imaging invariances
  - physics
  - anatomy
- Genomics invariances
  - evolutionary constraints
  - regulatory grammar

Questions?