Large Language Models in Computational Biology – A Primer (2024 Update)

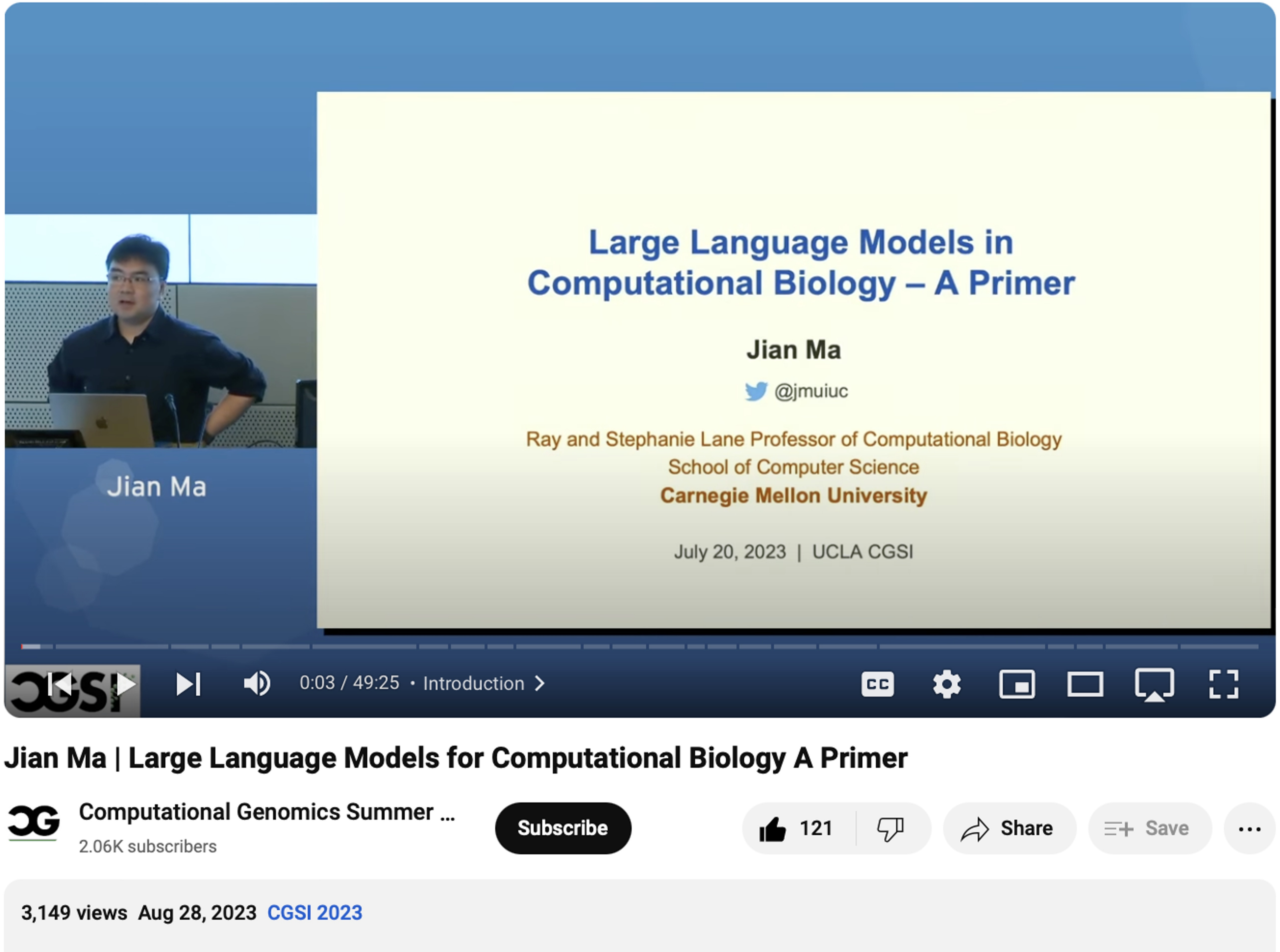
Jian Ma

@jmuiuc

Ray and Stephanie Lane Professor of Computational Biology
Ray and Stephanie Lane Computational Biology Department
School of Computer Science
Carnegie Mellon University

July 15, 2024 | UCLA CGSI

It’s been a year …



Large Language Models in Computational Biology – A Primer

Jian Ma

@jmuiuc

Ray and Stephanie Lane Professor of Computational Biology
School of Computer Science
Carnegie Mellon University

July 20, 2023 | UCLA CGSI

Jian Ma | Large Language Models for Computational Biology A Primer

Computational Genomics Summer ...
2.06K subscribers

3,149 views Aug 28, 2023 CGSI 2023

This presentation was put together with help from –

Ellie Haber

Wenduo Cheng

Shaoheng Liang

Nicholas Ho

Spencer Krieger

Yang Zhang

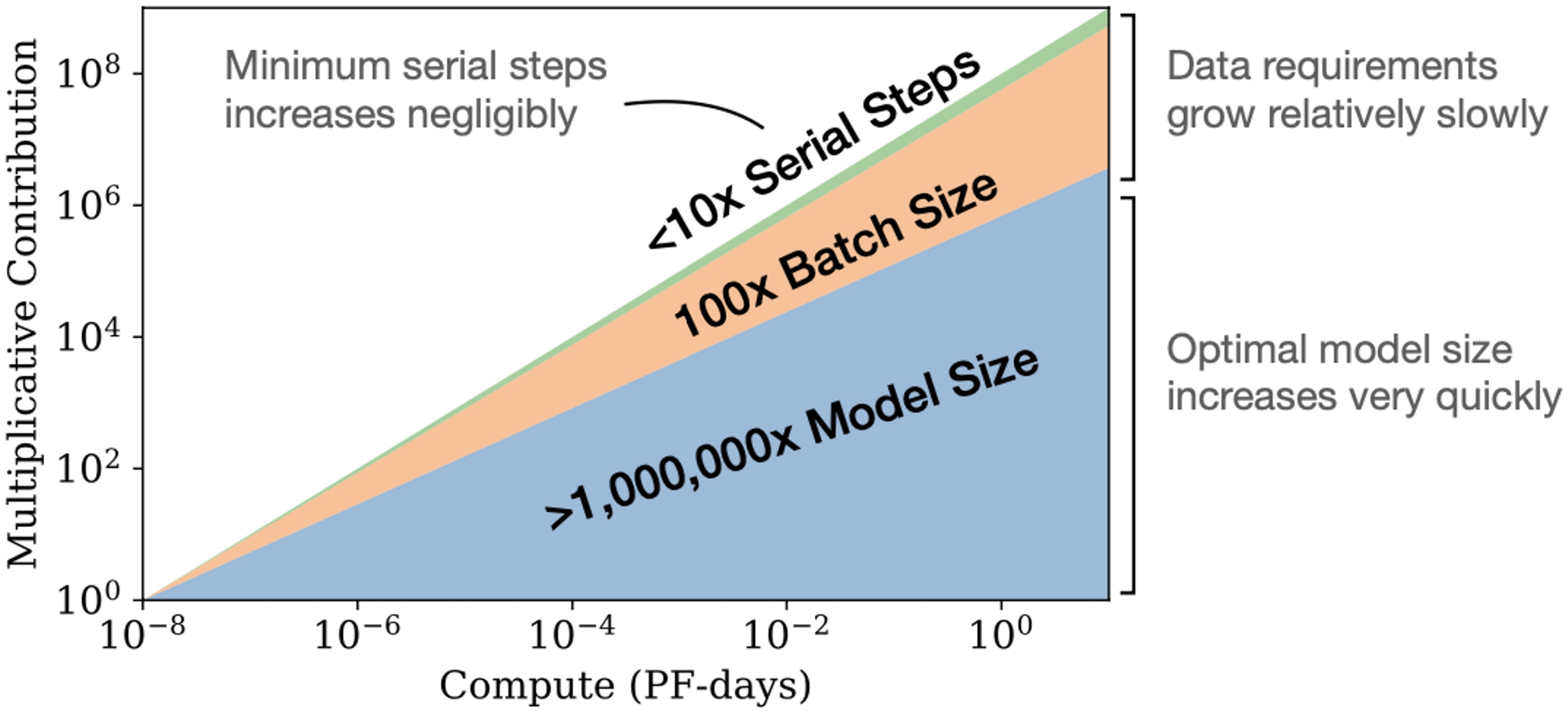
Remy Liu

Junjie Tang

Large Language Models

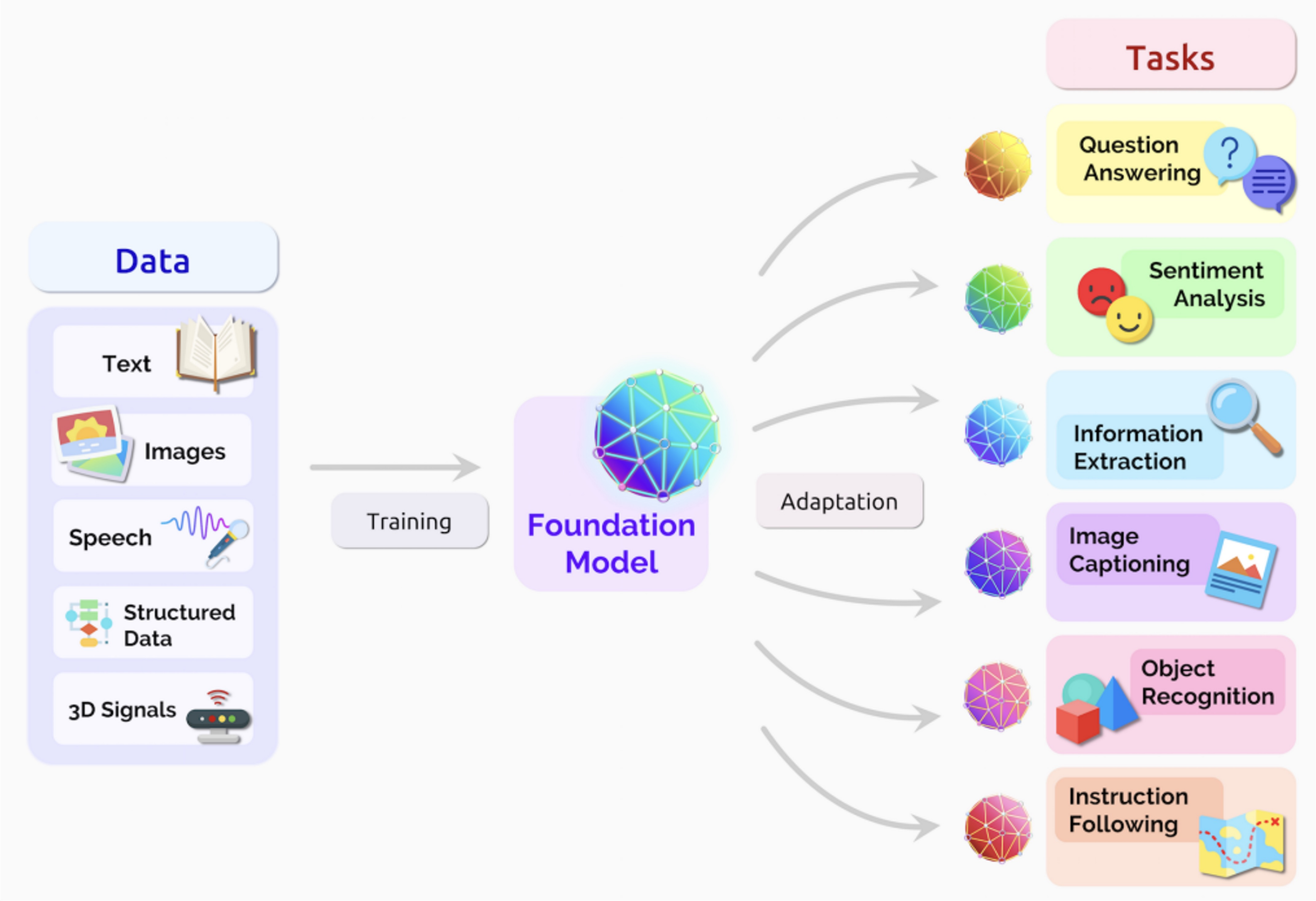
- Large language models
  =
  large-sized pretrained language models

- Scaling laws
  - Kaplan et al. 2020 (OpenAI)
  - Chinchilla scaling – Hoffmann et al. *NeurIPS* 2022



- Differences compared to LMs
  - Large # of model parameters
  - LLMs display some surprising "emergent abilities"
  - LLMs harbor powerful features such as prompting interface (e.g., GPT-4 API)
  - LLMs need tremendous resource to build

What is a Foundation Model?



“On the Opportunities and Risks of Foundation Models”
Bommasani et al. Stanford CRFM 2022

- Foundation models are a replacement for task-specific models

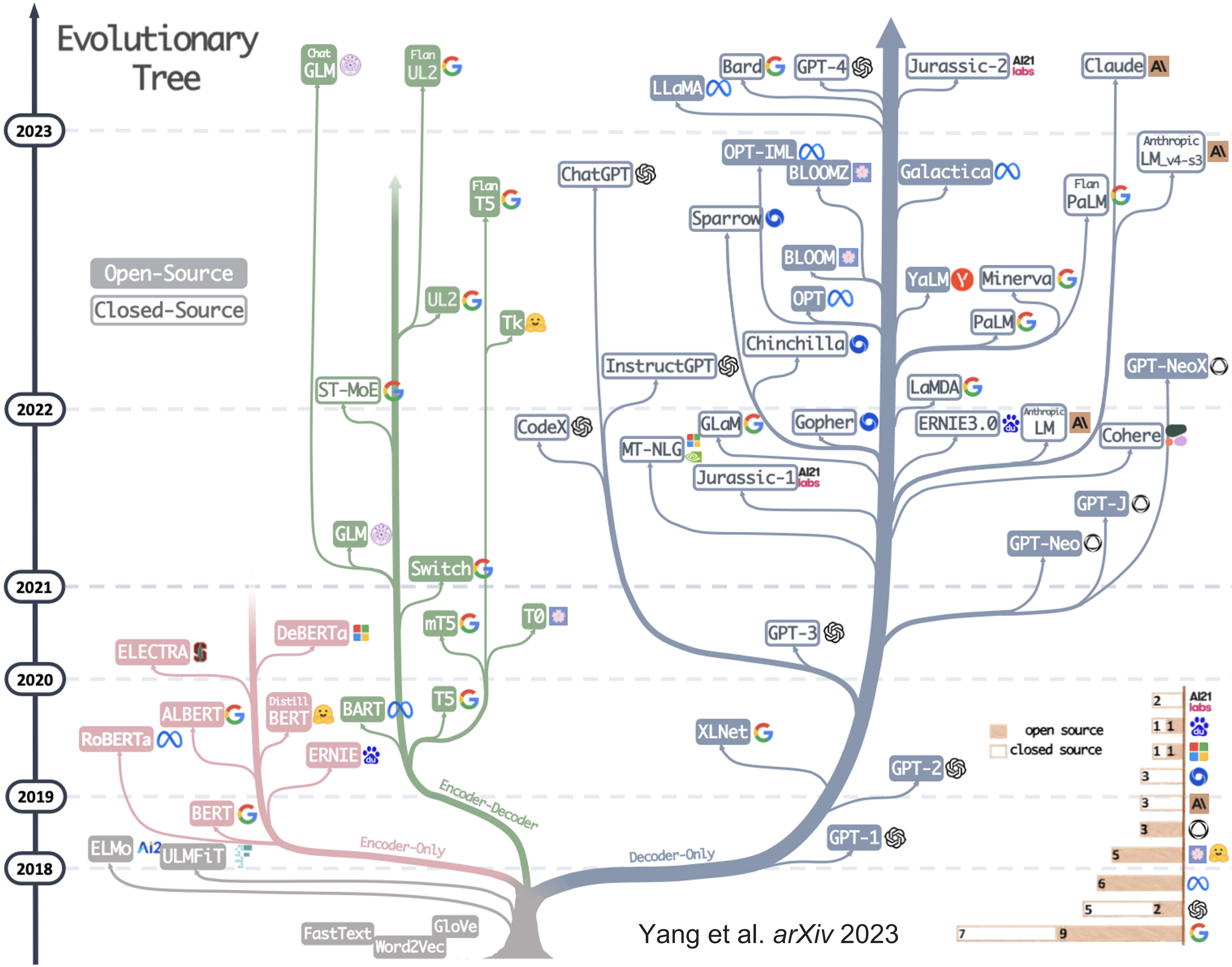
- Large-scale **pretraining** on large unlabeled datasets
- **Finetuning** for diverse downstream tasks

- Self-supervised learning
- Transfer learning

Open questions – from CGSI 2023

- How to better evaluate LLMs? How to make LLMs more accessible?
- How to embed cell/gene to better maintain biological contexts?
- How to incorporate prior knowledge into the neural network?
- How much finetuning is sufficient for a specific task/dataset? Will better designed pre-training tasks help shorten finetuning?
- How to extract the knowledge claimed to be distilled by the model?
- Do we have enough data available to pretrain LLMs or Foundation Models for various modalities in genomics?
- DNA and single-cell LLMs have comparable performance compared to existing approaches – need more challenging problems. What are the important problems for LLMs?
- Specific LLMs from molecular and cell biology literature + genomics data?
- Reliable hallucinations from LLMs => new biological hypothesis?

Genomic DNA Foundation Models

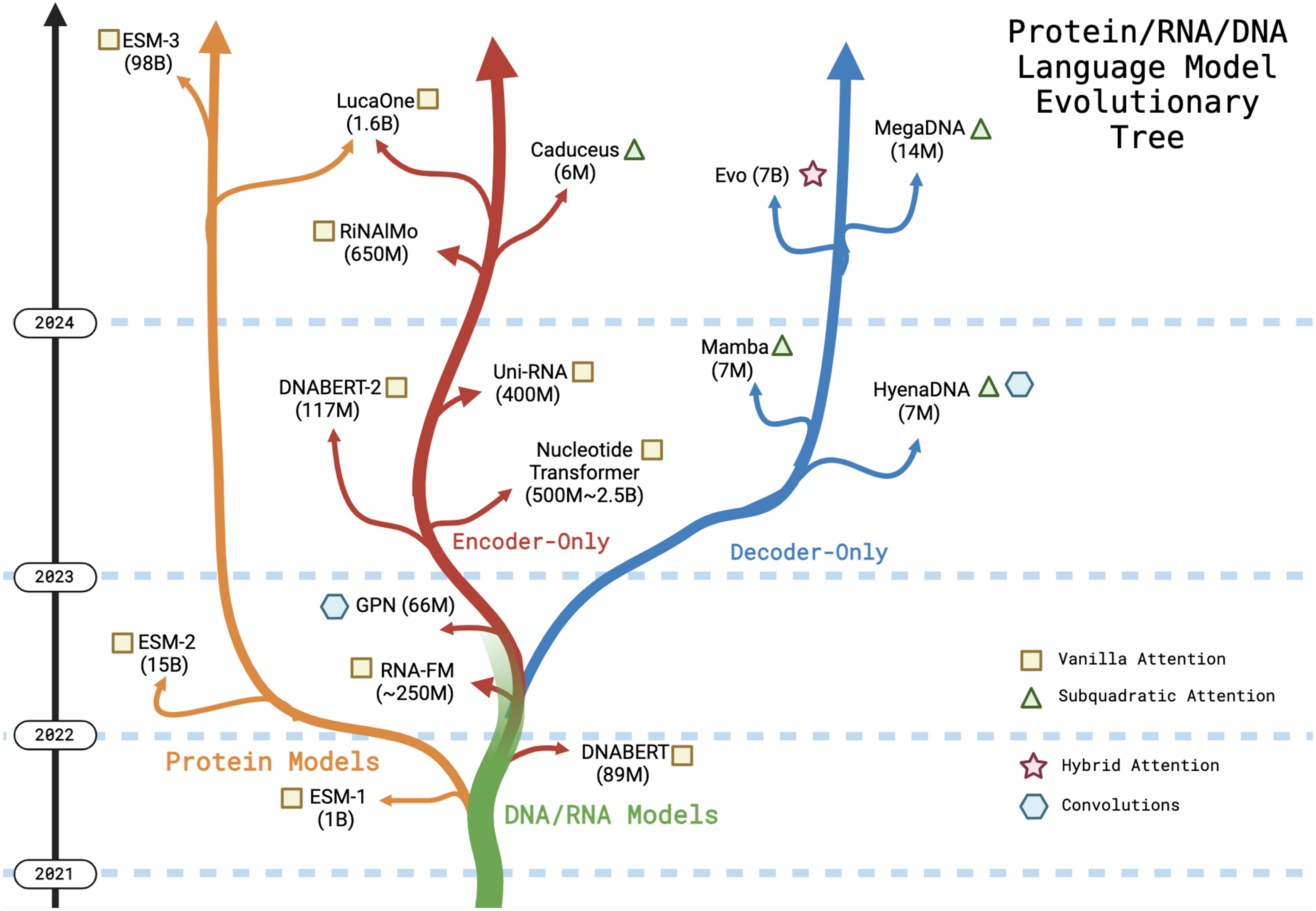


Yang et al. *arXiv* 2023

GPT	0.11B
BERT	0.34B
GPT-2	1.5B
Turing-NLG	17.2B
GPT-3	175B
Switch	1.6T
MT-NLG	530B
JURASSIC-1	178B
GLaM	1.2T
LaMDA	137B
PaLM	540B
OPT	175B
YaLM	100B
BLOOM	176B
Bard	137B
LLaMA	65B
GPT-4	1.7T

Source:
https://github.com/Hannibal046/Awesome-LLM

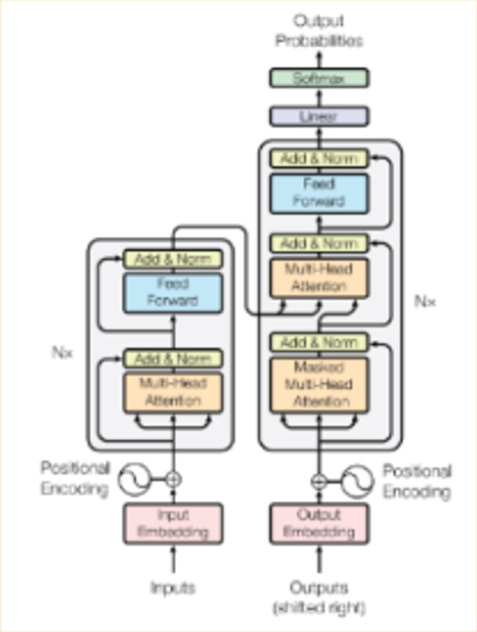
Non-comprehensive evolutionary tree for Protein/RNA/DNA language models



Nicholas Ho

Architecture of LLMs for genomic sequence

Choose Your Fighter (DNA Language Model):

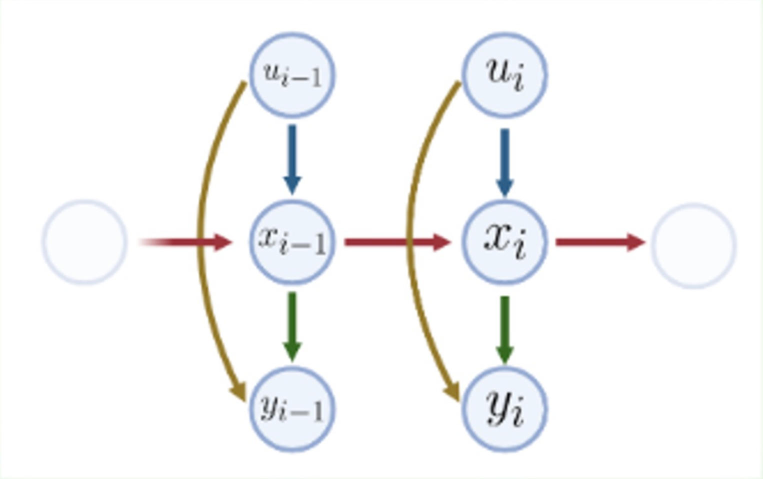


Attention is all you need
(Viswani et al 2015)

Vanilla Attention

- Gave Rise to immense success in vision and NLP

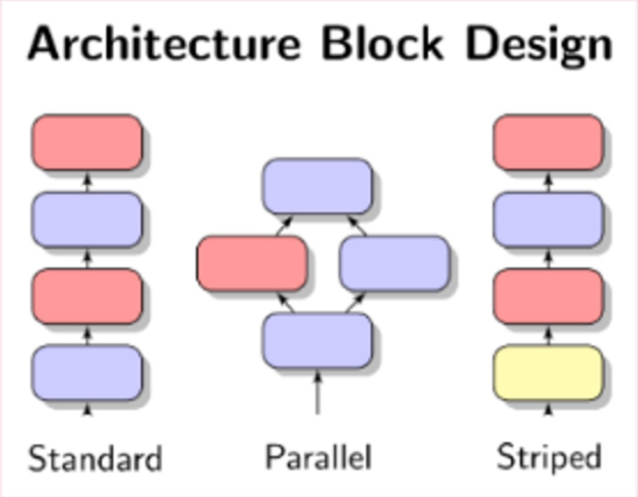
- Pros: Effective, relatively well studied
- Cons: Quadratic Complexity



Subquadratic Attention

- SSMs/Mamba/Hyena/RetNet/RNNs/RWKV/Griffin/BASED

- Pros: Subquadratic complexity
- Cons: Approximates Vanilla Attention, have trade-offs

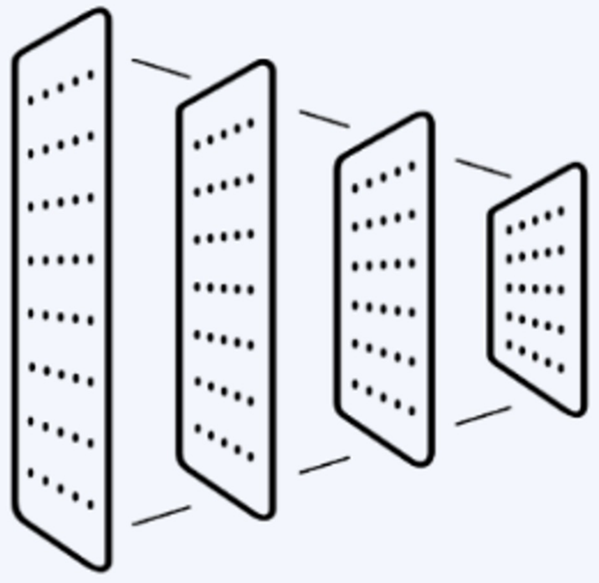


Mechanistic Design of Hybrid Architecture (Poli et al 2024)

Hybrid Attention

- Striped-Hyena
- Striped-Mamba

- Pros: Subquadratic
- Cons: Less well understood, only 2 canonical striped models



Created by Oleksandr Panasovskyi from Noun Project

Convolutions

- Dilated Convolutions
- Hyena Hierarchy (global convs)

- Pros: Widely used, local convs appear effective for DNA
- Cons: may lack global context, less expressive to attention

Some recent LLMs for genomic sequence

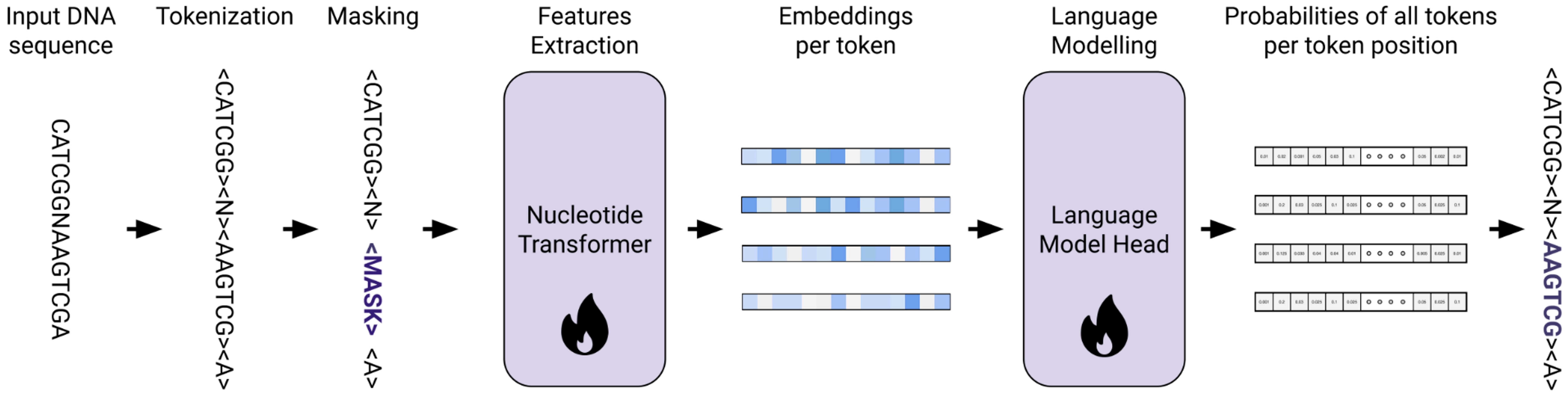
Model	***Paper***	***# Parameters***	***Architecture***	***Training Data***	***Downstream Tasks***
Nucleotide Transformer	Dalla-Torre et al. bioRxiv 2023	500M_human_ref 480M 500M_1000G 480M 2B5_1000G 2537M 2B5_multi_species 2537M	Transformer BERT	human reference, 3202 human genomes, genome from 850 different species	epigenetic marks prediction, promoter and enhancer prediction, splice site prediction
DNABERT-2	Zhou et al. ICLR 2024	117M	Transformer BERT	multi-species genome dataset from 135 species (32.49B)	promoter prediction, TF prediction, splice site prediction, epigenetic marks prediction, variant classification
HyenaDNA	Nguyen et al. NeurIPS 2023	~0.5M to 6.6M	Autoregressive Long convolutions	human reference genome	epigenetic marks prediction, promoter and enhancer prediction, splice site prediction
Caduceus	Schiff et al. ICML 2024	~0.5M to 6.6M	Bidirectional Mamba	human reference genome	epigenetic marks prediction, promoter and enhancer prediction, splice site prediction

Model	*Paper*	*# Parameters*	*Architecture*	*Training Data*	*Downstream Tasks*
Evo	Nguyen et al. bioRxiv 2024	7B Parameters	Striped Hyena	2.7M prokaryotic and phage genomes	Protein, ncRNA, fitness prediction, gene expression prediction, CRISPR and Transposon sequence generation
Genomic Pretrained Network (GPN)	Benegas et al. PNAS 2023	66M Parameters	Dilated Convolutions	TAIR10 reference genome of Arabidopsis thaliana from EnsemblPlants	Variant effect prediction
LucaOne	He et al. bioRxiv 2024	1.8B Parameters	Transformer	DNA, RNA and Protein data across 169,861 species	Protein Interactions with Proteins, ncRNA and DNA, ncRNA interactions with protein, ncRNA and DNA, DNA interactions with protein, ncRNA and DNA

Nucleotide Transformer

- Pre-trained BERT for DNA sequences on humans, 1000 genomes, and multispecies
- Non-overlapping K-mer tokenization
- Context length of 12K bp

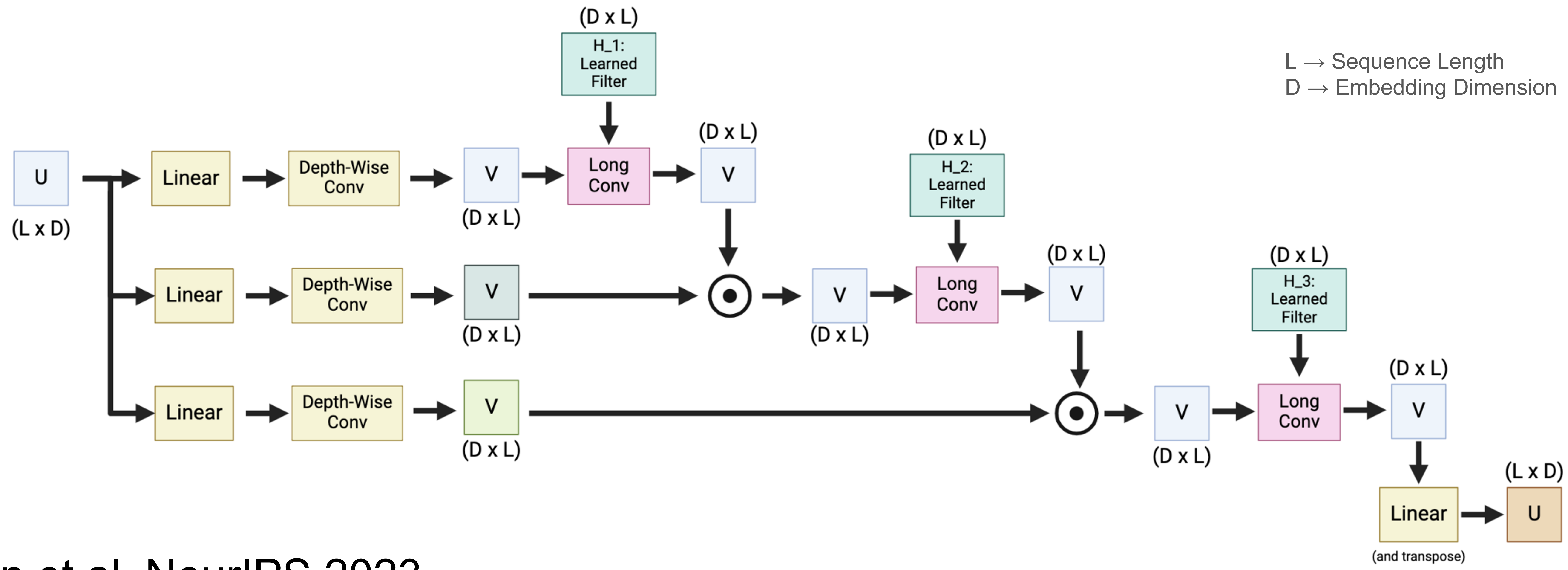
- Downstream prediction tasks:
  - promoter region, TFBS, splice site, functional variants identification



Dalla-Torre et al. bioRxiv 2023

HyenaDNA

- Pre-trained next token prediction for DNA sequences using a convolution-based architecture
- Tokenization: Nucleotide base-pair resolution
- Advantages: Long context modeling (~1M context length)
- Disadvantages: Not quite clear if this convolutional architecture has the capacity to match transformers



Nguyen et al. NeurIPS 2023

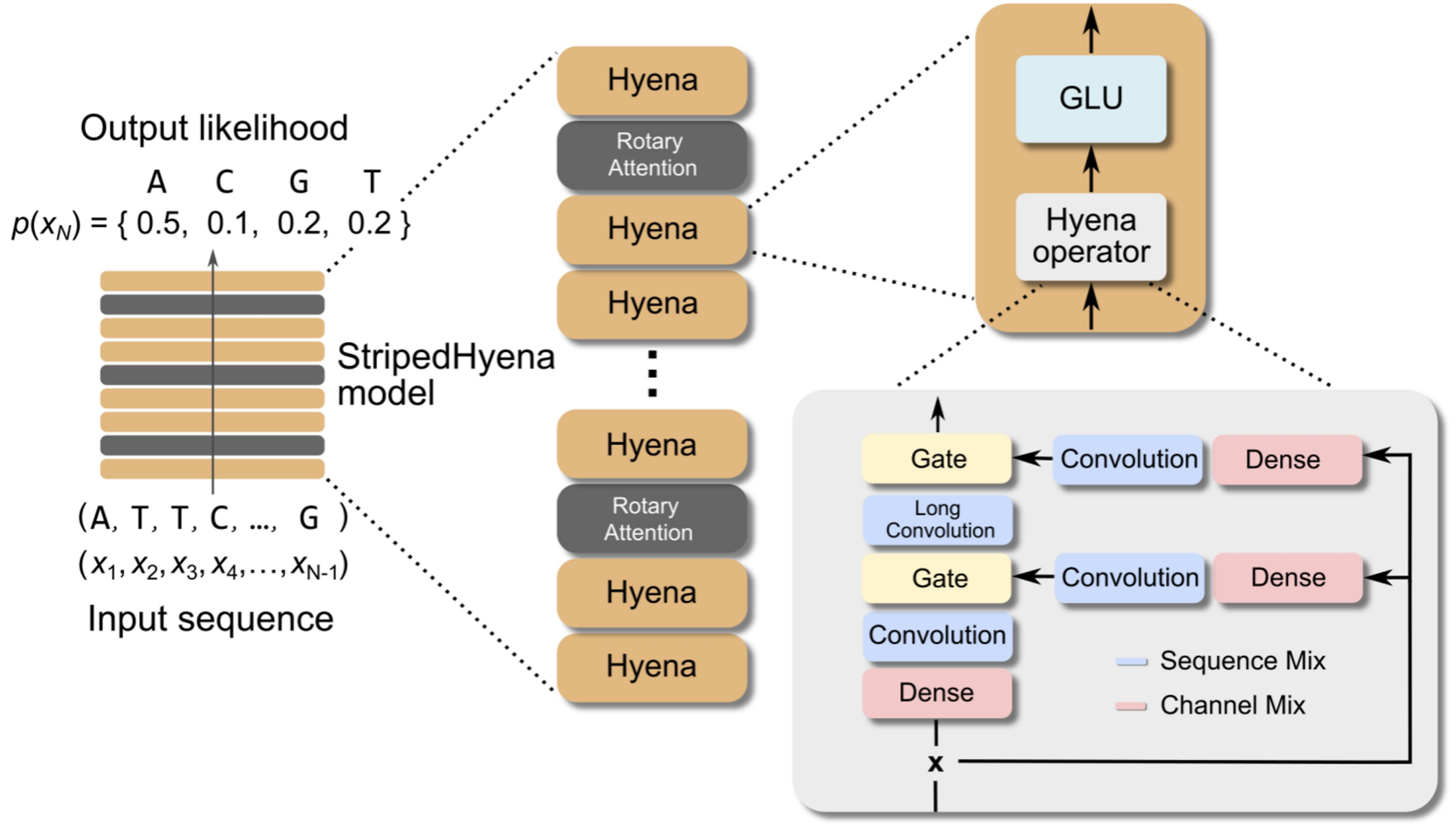
Evo

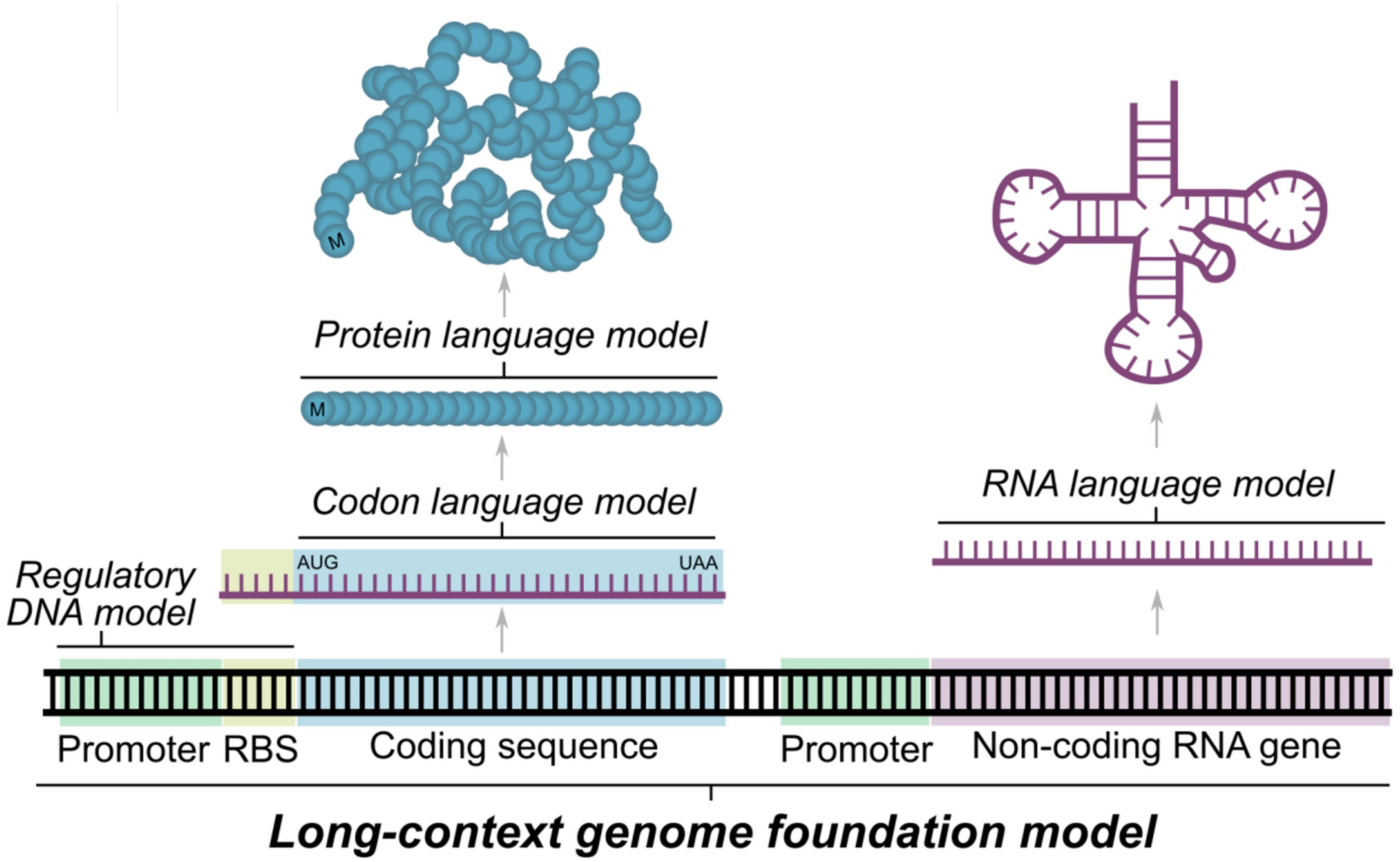
- Autoregressive (next-token prediction) pretrained on prokaryotic and phage genomes

- Striped Hyena architecture: combination of 29 hyena layers and 3 attention layers

- Demonstrates that aspects of protein and ncRNA can be evaluated through a model trained on DNA sequences

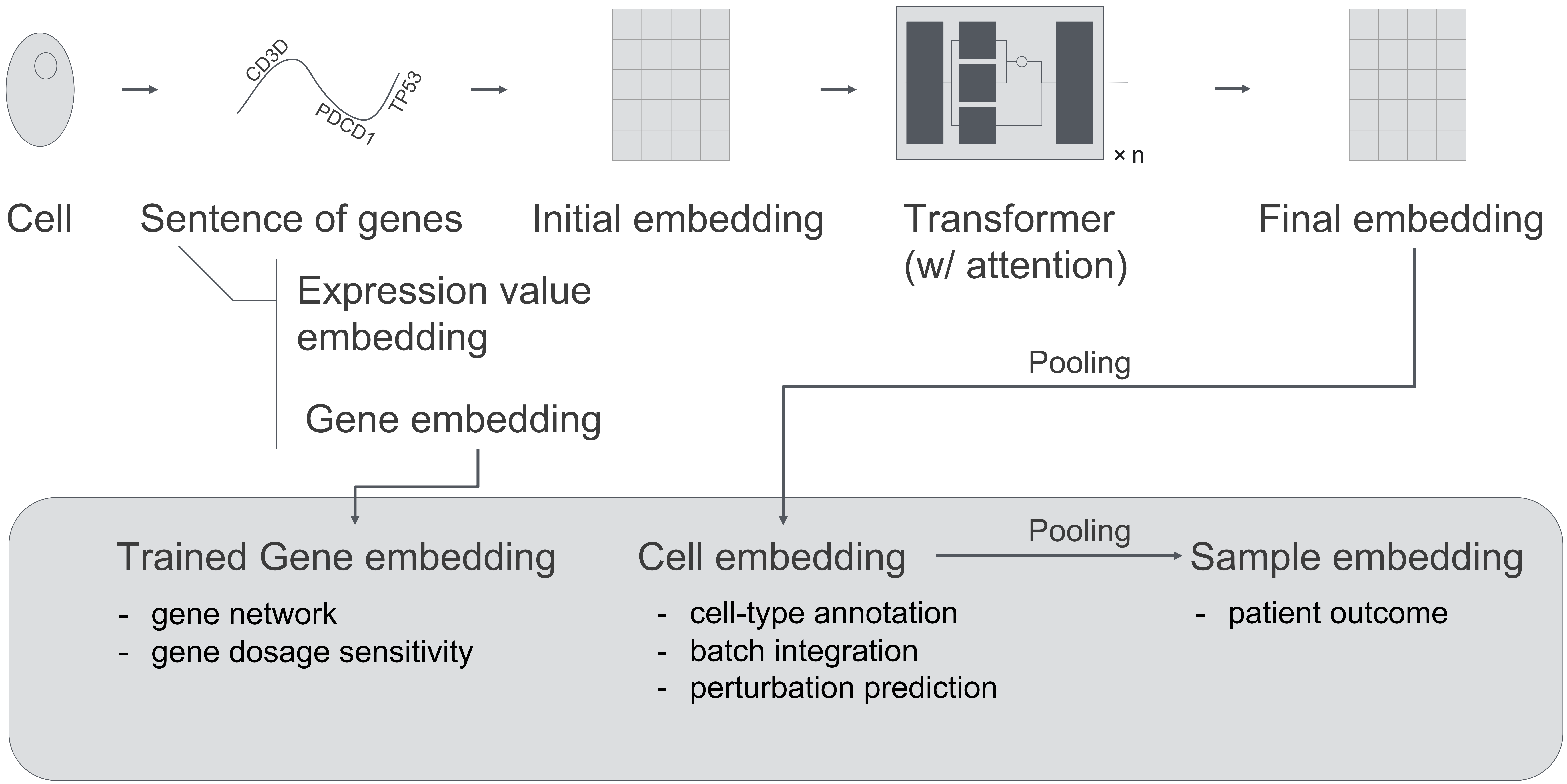
Nguyen et al. bioRxiv 2024





Single-cell Foundation Models (scFMs)

General structure of scFMs



Cell → Sentence of genes → Initial embedding → Transformer (w/ attention) → Final embedding

Expression value embedding

Gene embedding

Pooling

Trained Gene embedding

- gene network
- gene dosage sensitivity

Cell embedding

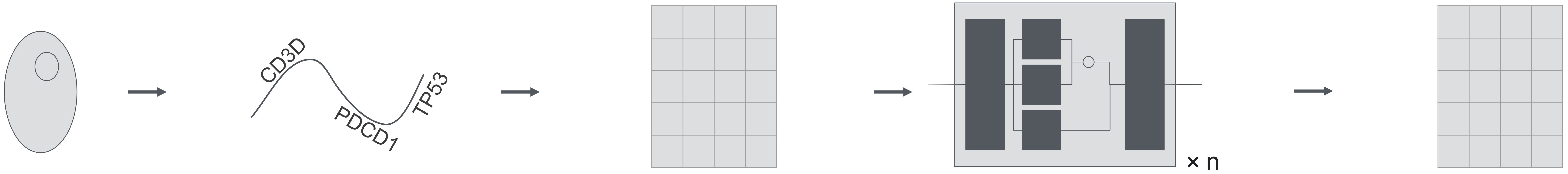
- cell-type annotation
- batch integration
- perturbation prediction

Pooling

Sample embedding

- patient outcome

Tokenization for cells



Cell

Sentence of genes

Initial embedding

Transformer
(w/ attention)

Final embedding

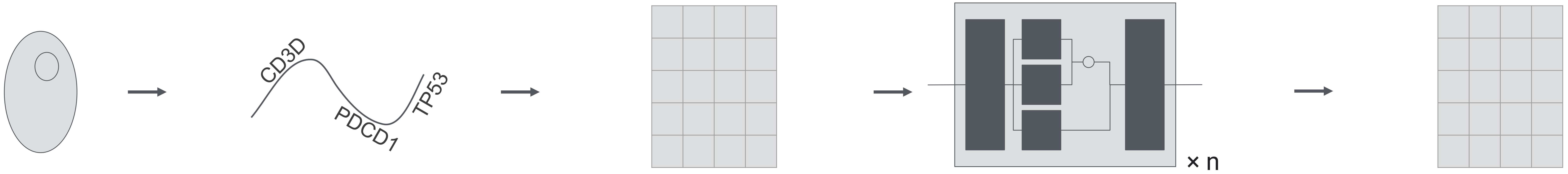
Gene embeddings ordered by their expression value (Geneformer)

Gene embeddings + binned expression value embeddings (scGPT, scBERT)

Gene embeddings + expression value embeddings (scFoundation)

Gene embedding by protein language model (UCE)

Network structures and training strategies



Cell

Sentence of genes

Initial embedding

Transformer (w/ attention)

Final embedding

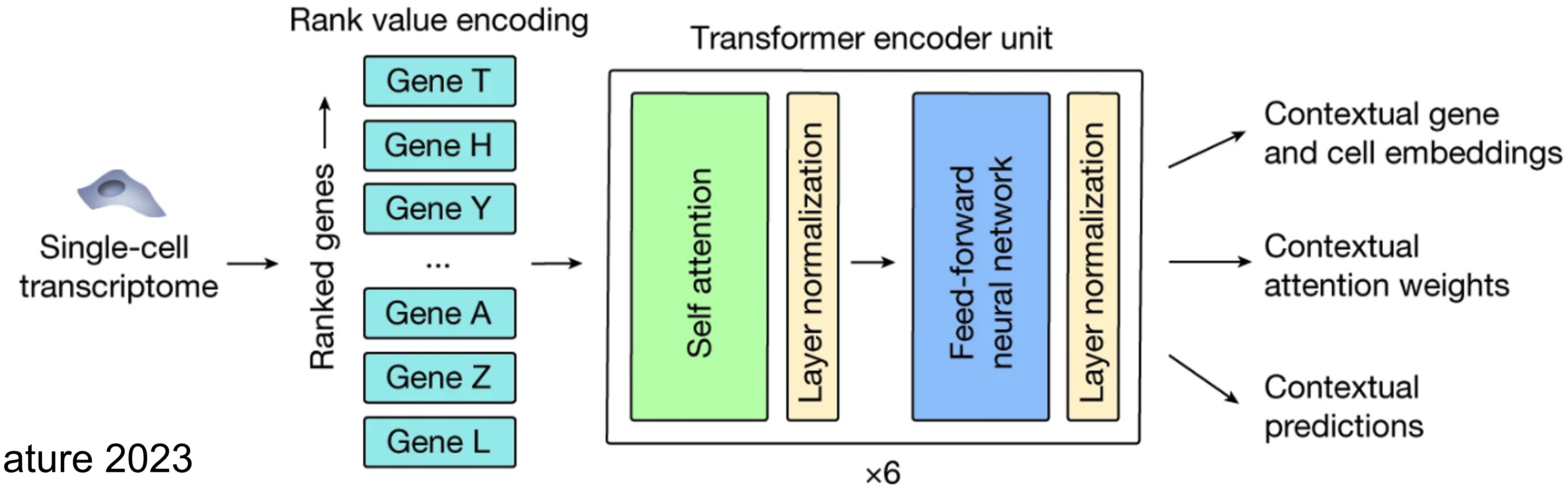
- Autoregressive pretraining (scGPT)
- Masked language modeling (MLM) pretraining (scBERT, Geneformer, scFoundation)

Timeline of scFMs

		# Parameters	*Training data size*	*Highlights*	*Paper*
2022	**scBERT**	5M	1M	> Scalability: Performer	Yang et al., Nat Mach Intell 2022
	Geneformer	40M	30M	> Gene networks inference	Theodoris et al., Nature 2023
	scGPT	51M	33M	> Generative pretraining (cell & gene prompt)	Cui et al., Nat Methods 2024
2023	**scFoundation**	100M	50M	> Scalability: reduced input length > Integration: confounding factors regressed out	Hao et al., Nat Methods 2024
	UCE	650M +15B pLM (fixed)	46M	> Cross-species integration: utilizes pLM (ESM-2) for gene embedding	Rosen et al., bioRxiv 2023
2024	**scMulan**	368M	10M	> Multi-tasking: query by prompts > Richer pretraining: metadata	Bian et al., RECOMB 2024
	NicheFormer	50M	110M	> Integration: dissolved & spatial assays	Schaar et al., bioRxiv 2024

Geneformer

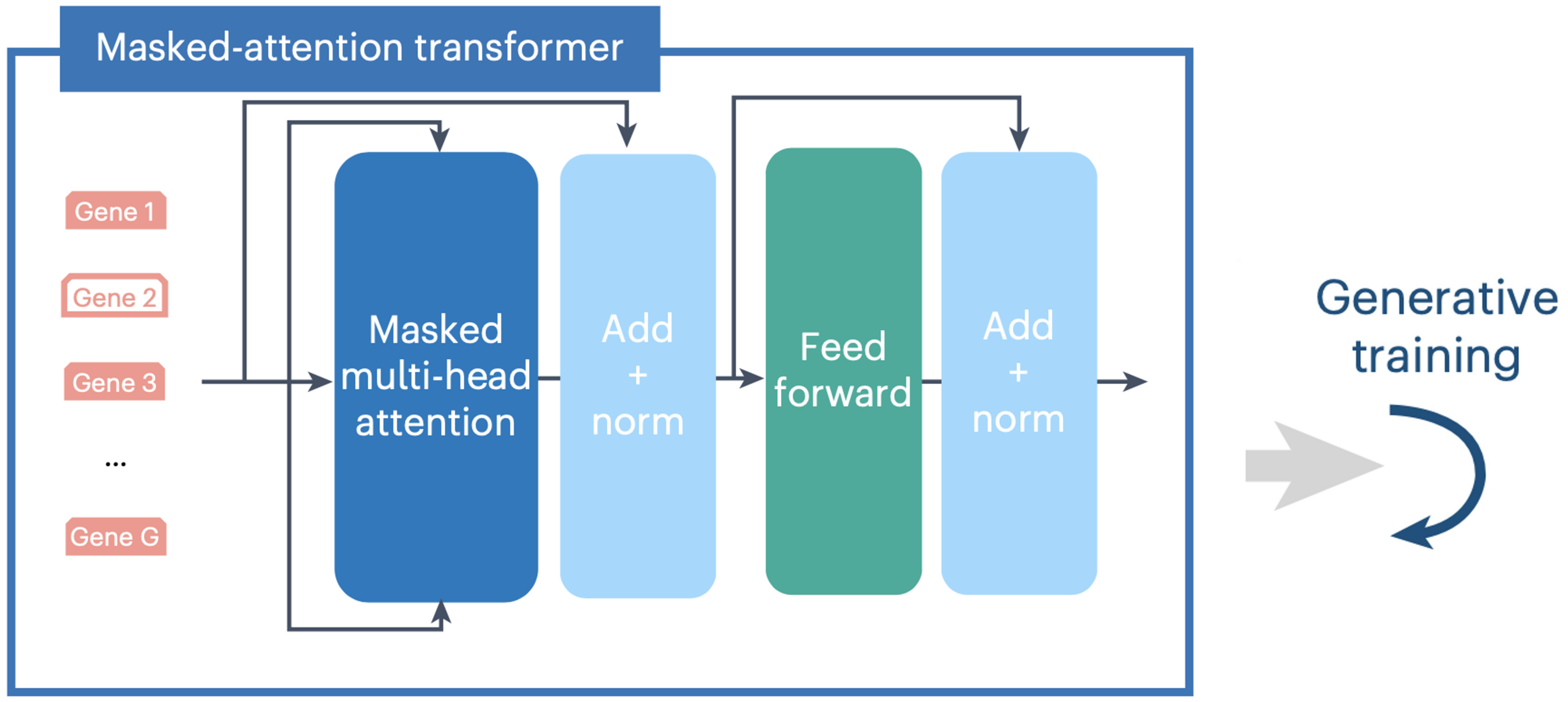
- Pretrained on ~30 million human single-cell transcriptomes
- Utilizes rank value encoding, normalizing gene expression levels across the pretraining corpus
- Transformer with MLM pretraining, incorporating positional encoding to represent a gene’s relative expression level

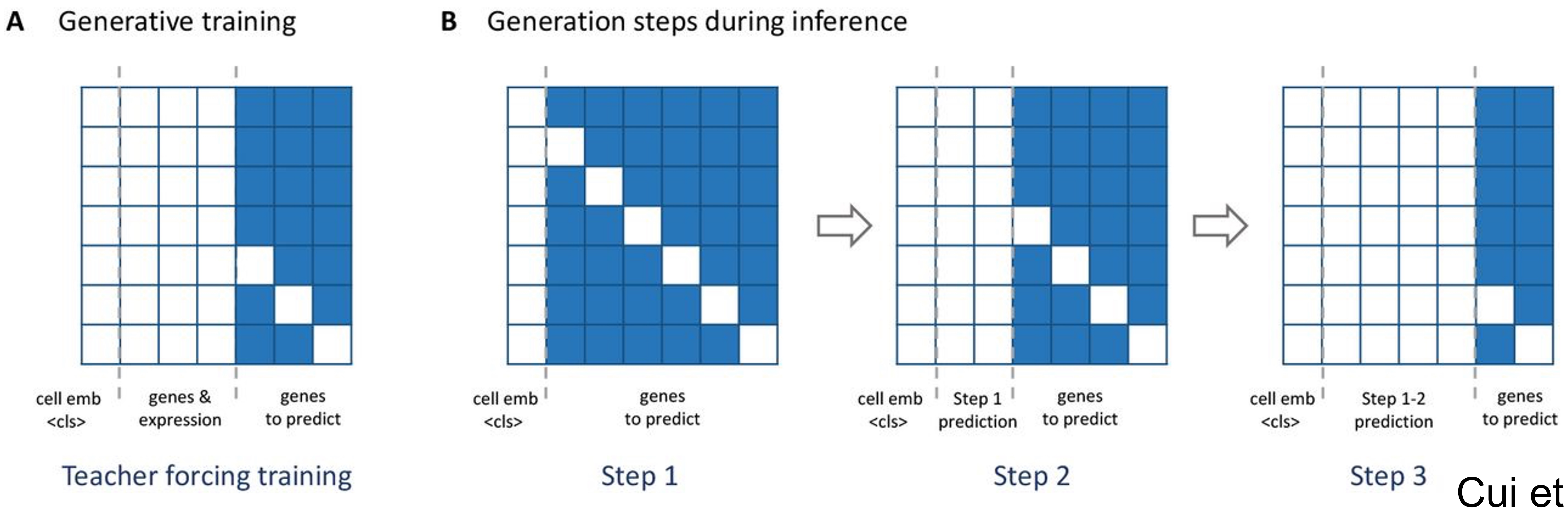


Theodoris et al. Nature 2023

scGPT

- Pretrained on ~33 million single-cell transcriptomes
- Generative pre-training with gene- and cell-prompt
- Utilizes value binning to convert expression counts into relative values



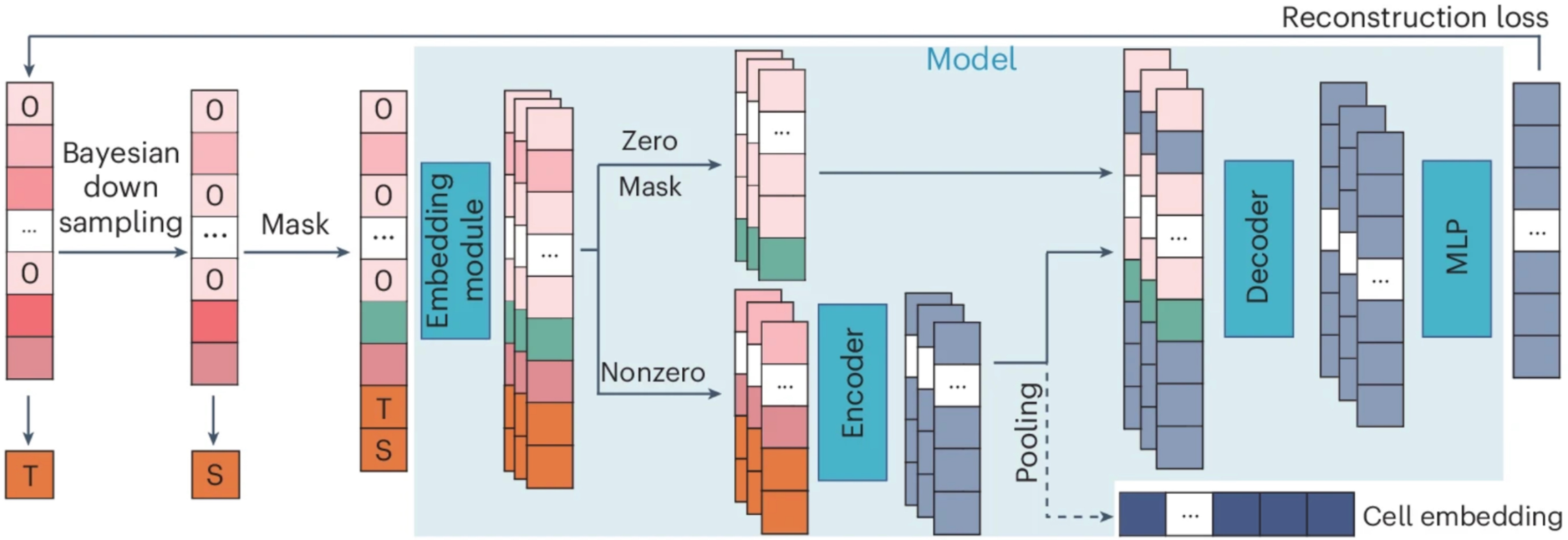

A Generative training

B Generation steps during inference

Cui et al. Nat Methods 2024

scFoundation

- Trained on 50 million cells
- scFoundation designs read-depth-aware (RDA) modeling pretraining task:
  - Downsample gene expression to create cells with varying read counts
  - Reconstructs original expression counts via MLM strategy
- RDA pretraining task enables the learning of relationships between cells with different read depths



Hao et al. Nat Methods 2024

Leveraging prior knowledge for improved gene embeddings

Gene embeddings can be trained de novo, but prior knowledge may help:

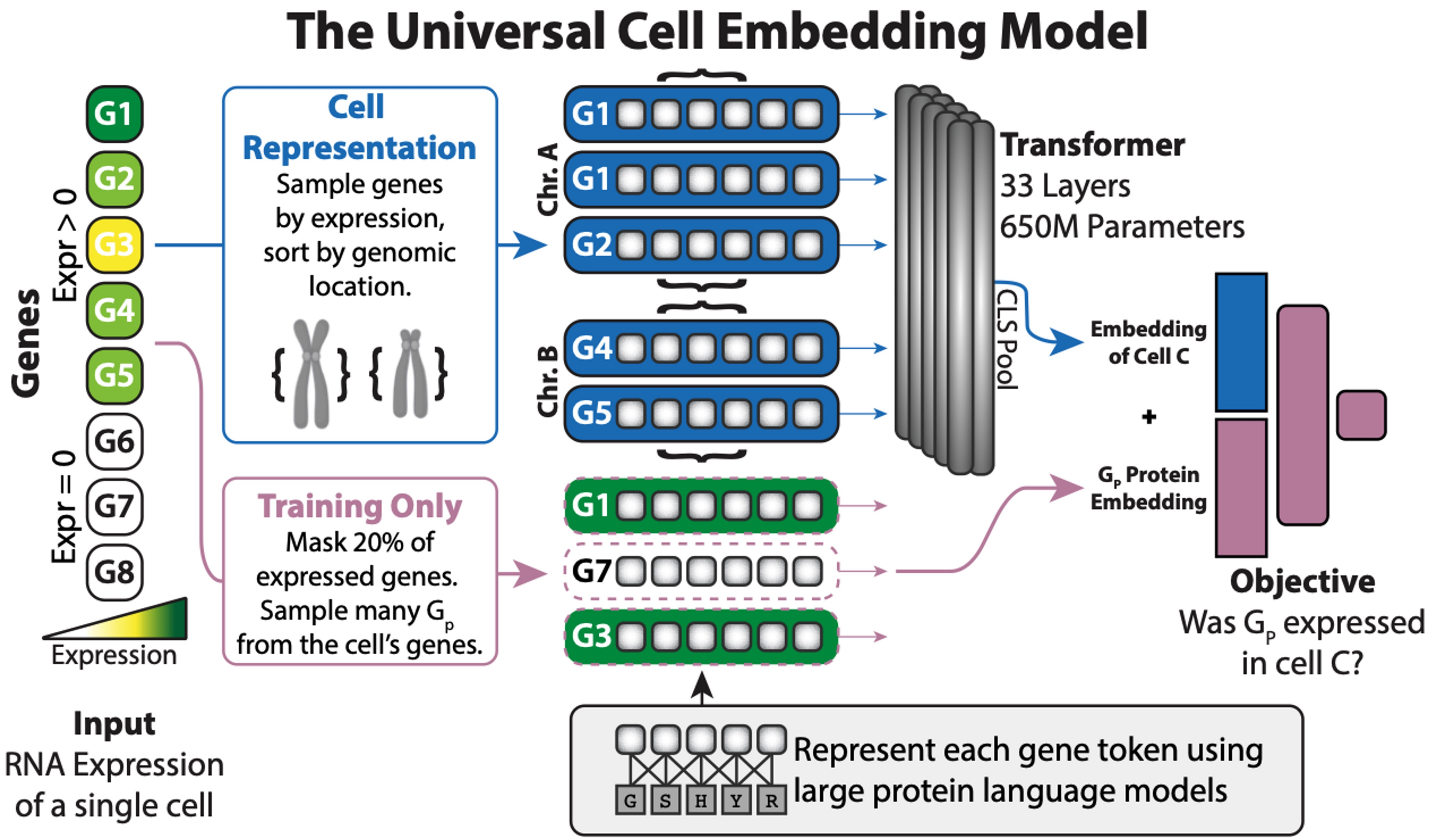
- Gene2vec
  - Distributed representation based on co-expression (used in scBERT)
- GenePT (Chen and Zou)
  - Use GPT-3.5 to generate gene embeddings from gene description.

However, because each gene is treated as a separate entity, knowledge about one gene is not transferable to another. Also, recognizing similarity of genes across species is important for a universal model.

- Universal Cell Embeddings (UCE)
  - Uses protein LLM to embed a sample’s genes with protein products
  - Protein products make genes across species more comparable

UCE framework

UCE samples genes by expression value and orders them by chromosomal loci


The Universal Cell Embedding Model
Genes
Expr > 0
Expr = 0
G1
G2
G3
G4
G5
G6
G7
G8
Expression
Input
RNA Expression of a single cell
Cell Representation
Sample genes by expression, sort by genomic location.
Training Only
Mask 20% of expressed genes. Sample many G_p from the cell's genes.
Chr. A
Chr. B
Transformer
33 Layers
650M Parameters
CLS Pool
Embedding of Cell C
+
G_p Protein Embedding
Objective
Was G_p expressed in cell C?
G S H Y R
Represent each gene token using large protein language models

Binary expression prediction using cell + gene's protein embeddings

Genes embedded by a protein LM

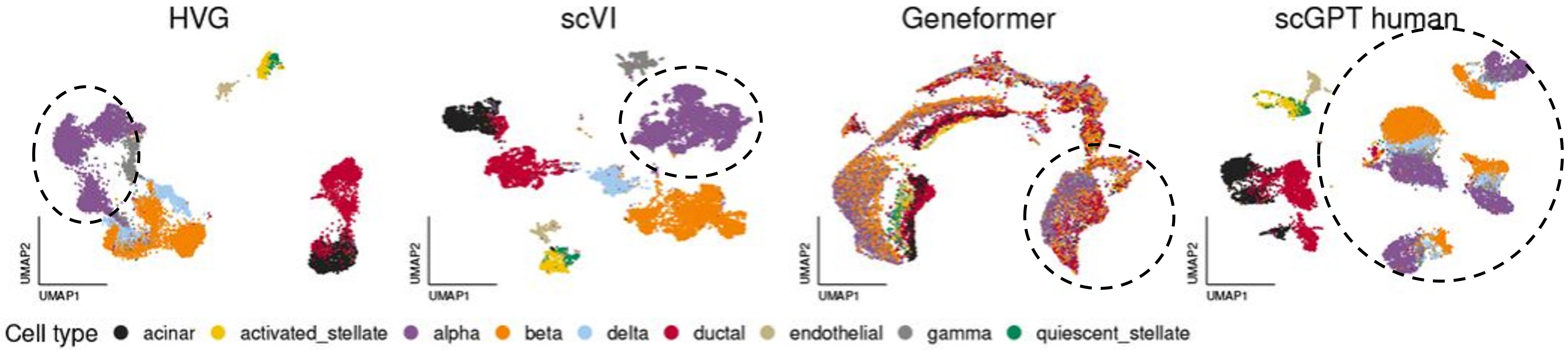
Rosen et al. bioRxiv 2023

Promises and challenges

- Foundation models for genomes and cells seem to have potentials
  - Pretraining on large number of cells will discover intrinsic interaction of genes
  - Pretrained models are easily adaptable to multiple tasks to enable biological findings

- But biology is complicated and its “language” is likely much harder to model than natural languages.
  - Biological data involve many confounding factors
  - Biological questions are often not mathematically well-defined
  - In this data driven era: “what is the best question to ask”

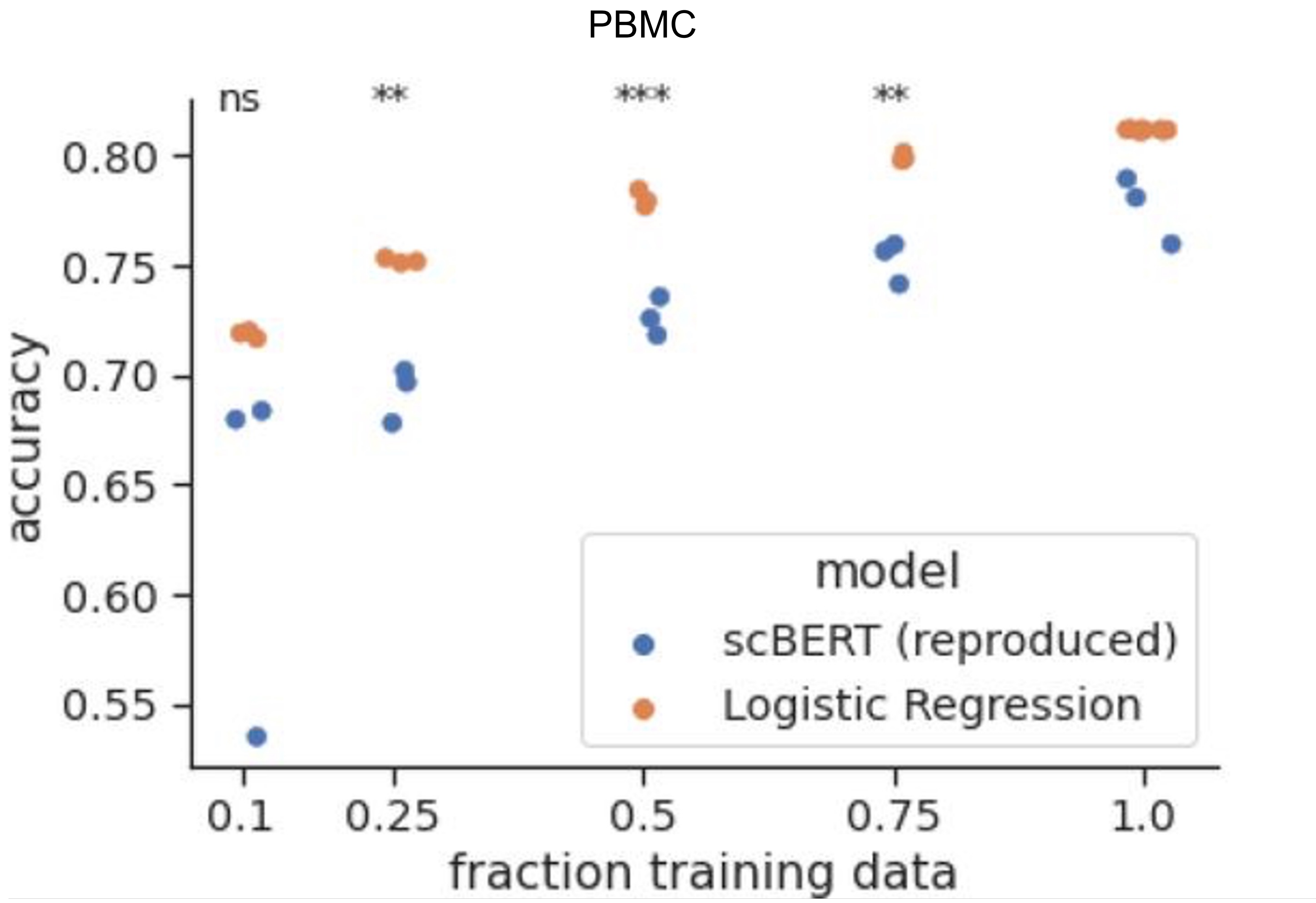
Challenge: how to be more foundational?

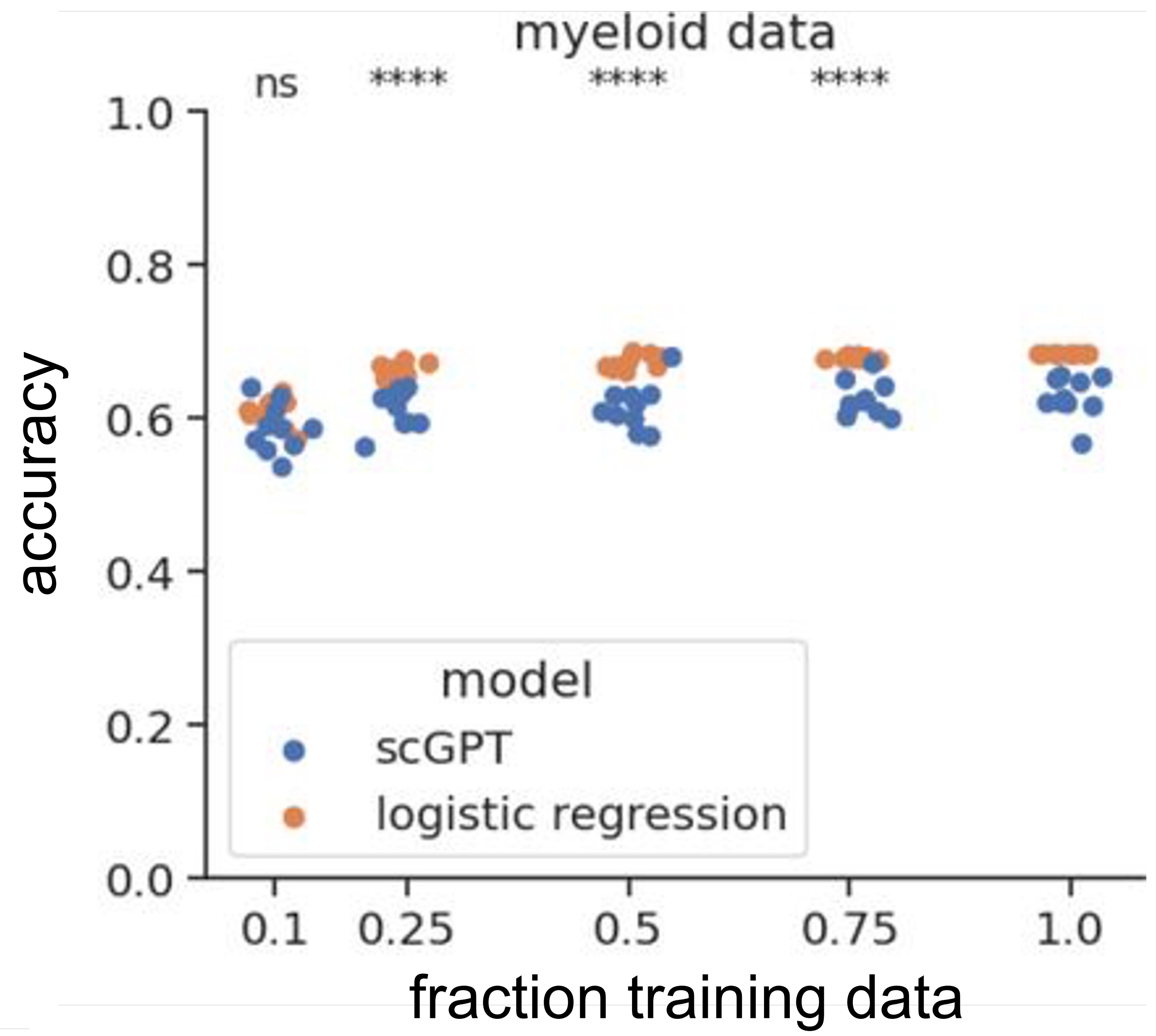


- Limited robustness:
  - Although trained on millions of cells, current foundation models struggle with different assays in zero-shot settings (Kedzierska et al. bioRxiv 2023).
  - Pretraining often fails to separate biology from noise and sometimes has no effect (Boiarsky et al. bioRxiv 2023)

- Do we need more training data or different modeling approaches to make the models truly FMs?

Challenge: Jack of all trades, master of none





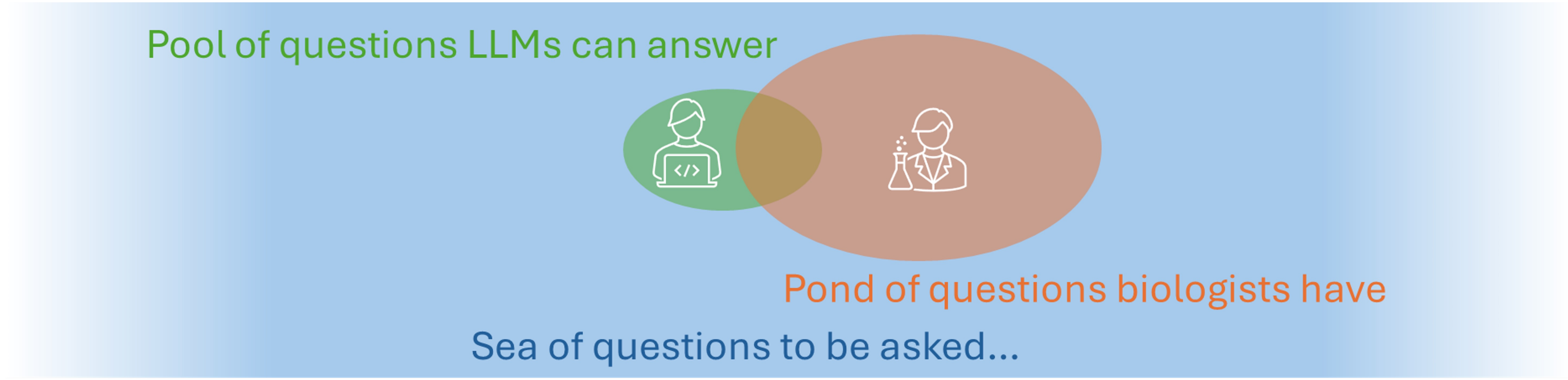
- Training inefficiency (Boiarsky et al. bioRxiv 2023):
  - Expensive pretraining and finetuning do not always outperform simple specialized methods

- Questions:
  - How do we assess the effectiveness of the strategy and output embeddings?
  - How can we better harness the power of the models?

Challenge: Where are the nails?

- FMs promise to be powerful hammers, but surprisingly, many biological questions do not look like nails
- Many methods excel in supervised tasks, but single-cell omics is mainly exploratory, with significance beyond routine supervision

- What **exciting biological discoveries** can foundation models enable?
  - Can a model answer a question it was not trained for?
  - Can a model uncover unique data characteristics without a predefined question?


Pool of questions LLMs can answer
Pond of questions biologists have
Sea of questions to be asked...

Challenge: Mechanistic insights from the FMs?

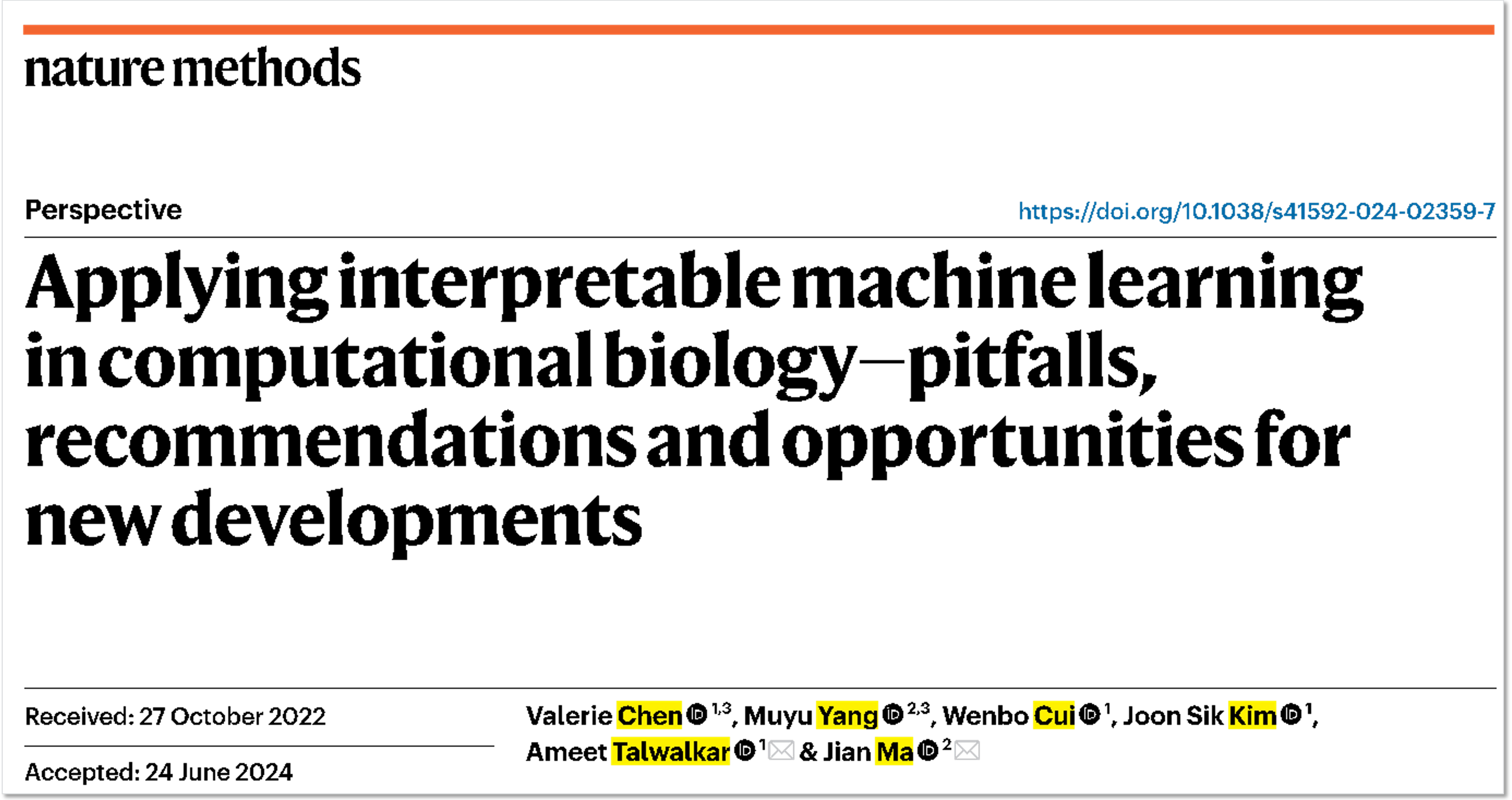
- LLMs trained on massive amounts of texts have shown to have emergent abilities such as reasoning.

- Current FMs have demonstrated promising results in diverse downstream tasks, but *how* such predictions are made remains a black box.

- Can single-cell foundation models explain and reason about the predictions?

  For example:
  - What are the molecular mechanisms that lead to the specific transcriptomic changes due to a genetic perturbation?
  - What are the key molecular pathways that define a cell (sub)type?

Interpretable ML in the era of LLMs



nature methods

Perspective

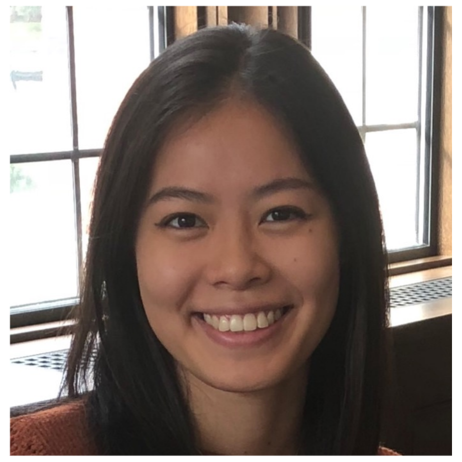
https://doi.org/10.1038/s41592-024-02359-7

Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments

Received: 27 October 2022

Accepted: 24 June 2024

Valerie Chen[1,3], Muyu Yang[2,3], Wenbo Cui[1], Joon Sik Kim[1], Ameet Talwalkar[1] & Jian Ma[2]



Valerie Chen

Wendy Yang

Ameet Talwalkar

Chen[#] and Yang[#] et al. ***Nature Methods,*** in press