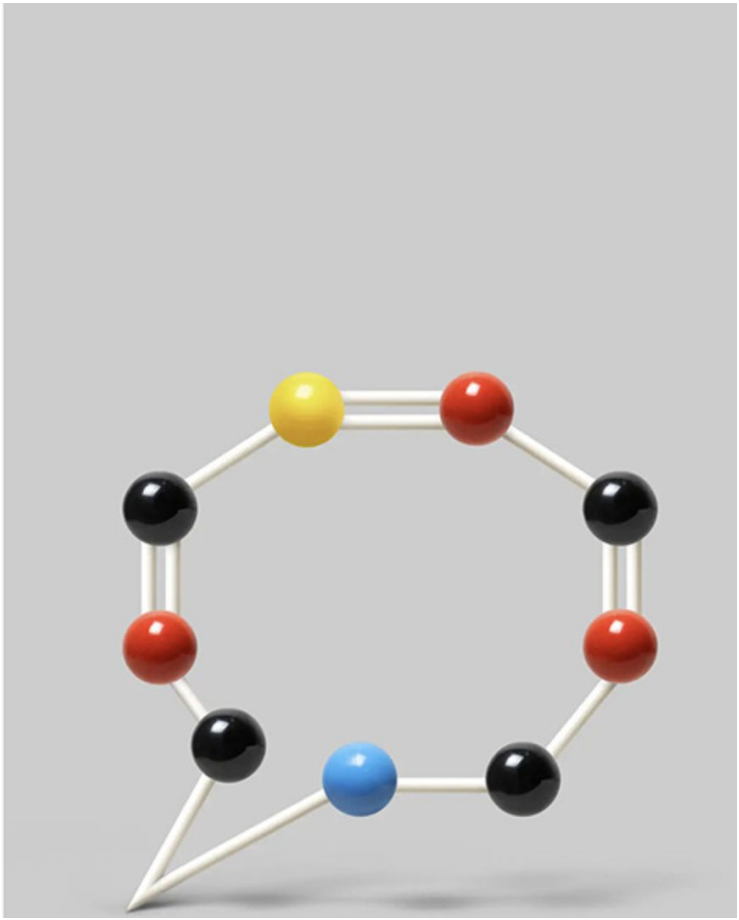
Foundation Models of gene networks

Shreshth Gandhi, Vevo Therapeutics

A.I. Is Learning What It Means to Be Alive

Given troves of data about genes and cells, A.I. models have made some surprising discoveries. What could they teach us someday?



Foundation Model of Cells

Learning gene and cells function in diverse biological contexts

Learn **meaning of words in context**

The ____ of the old ____ was rough and covered in ____.

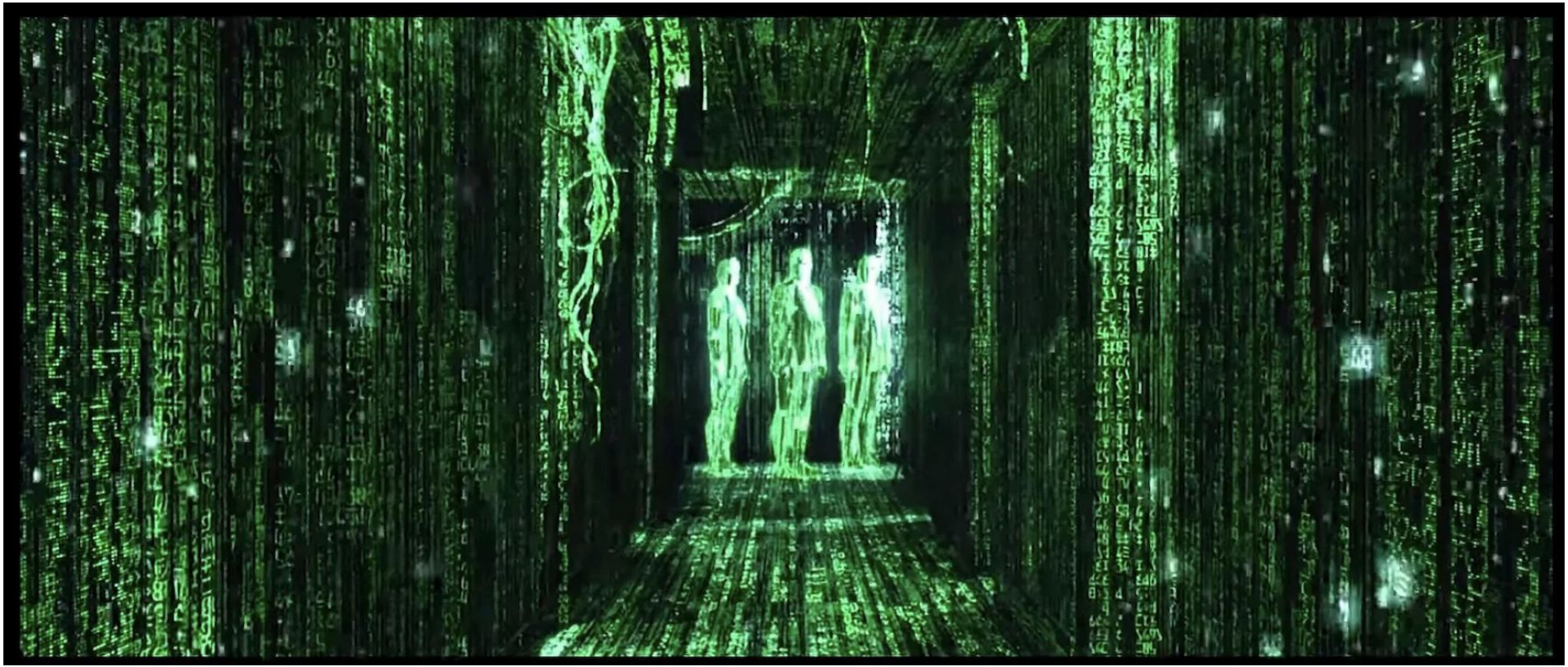
The **bark** of the old **tree** was rough and covered in **moss**.

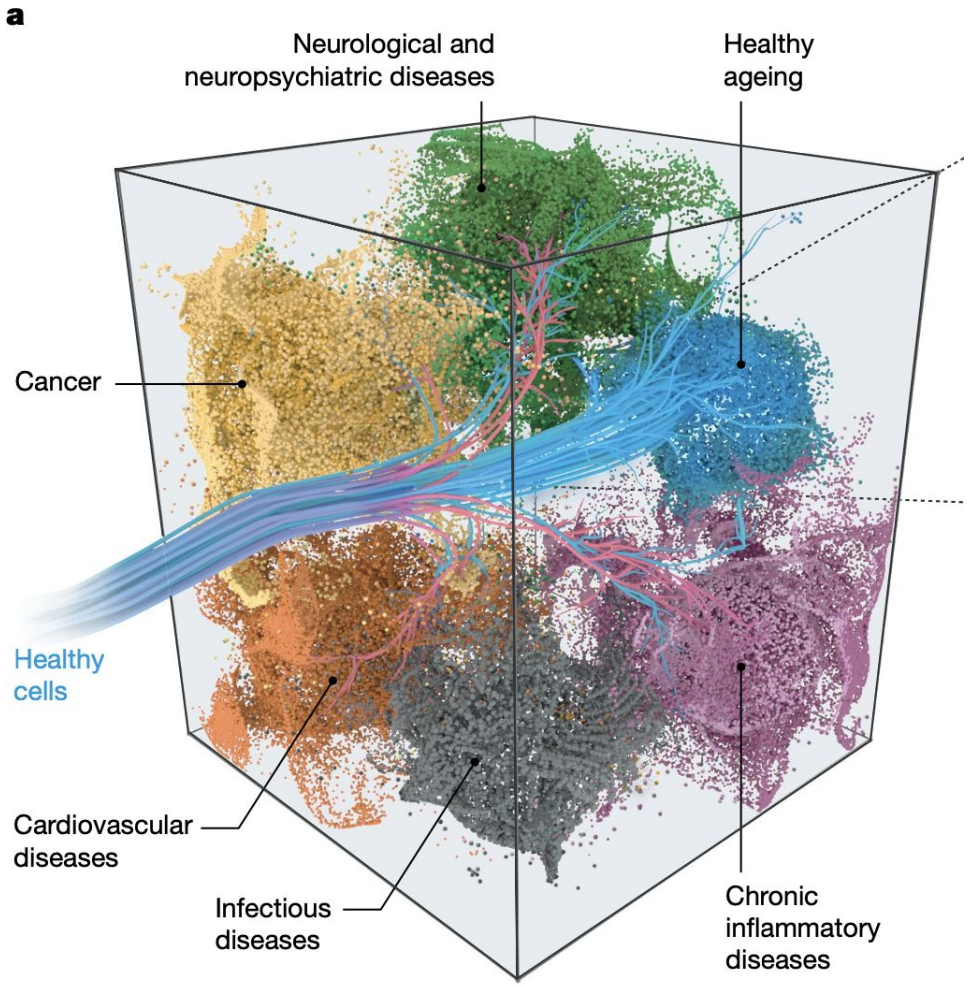
Learn context dependent **gene function**

Self-supervised model trained on
single cell gene expression data

Gene expression

We would like to have models that can “see through the matrix”



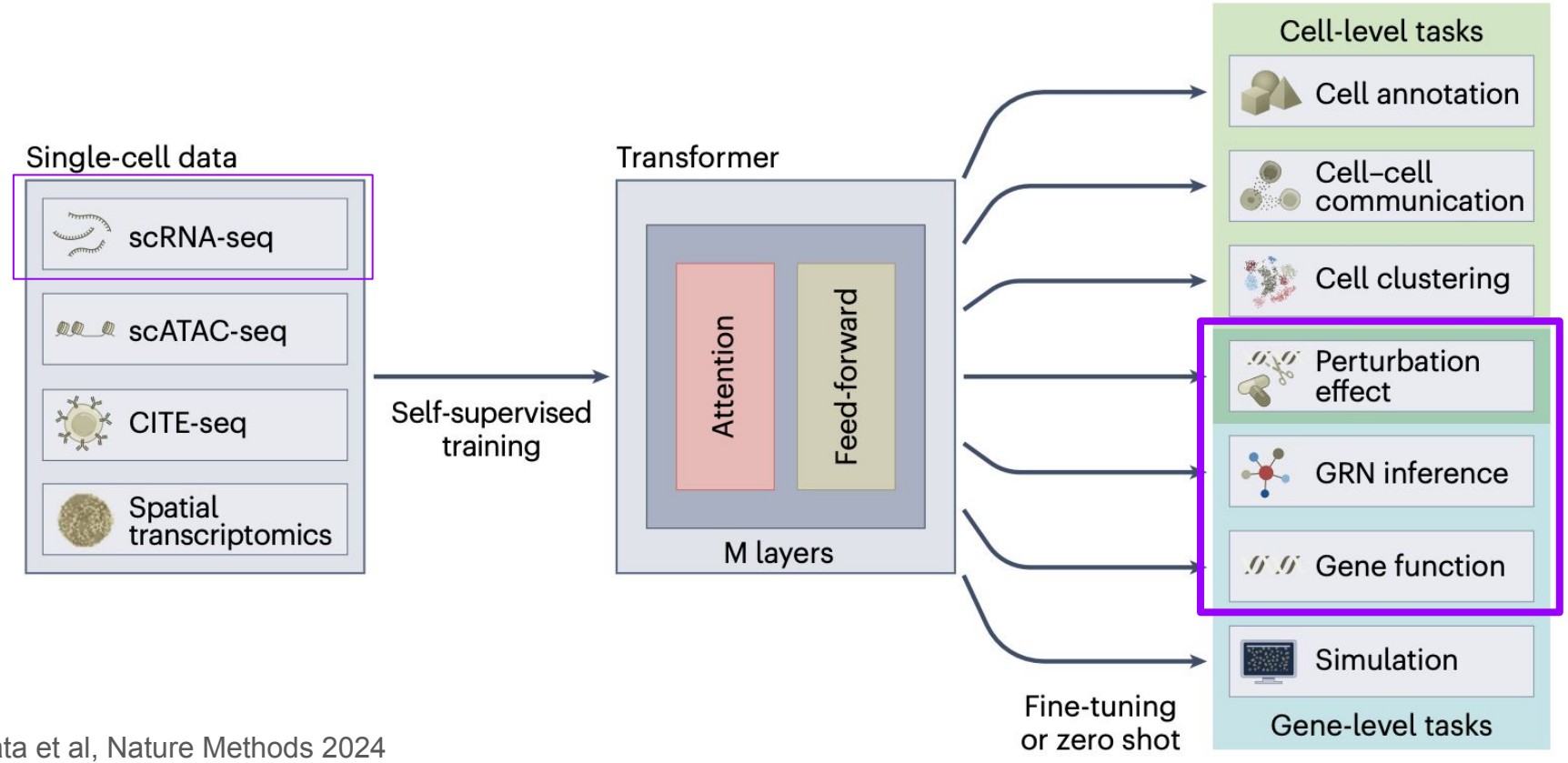


Why look at single cells?

- Capture a diverse distribution of cell states
- Learn grammar of interactions that lead to state transitions
- Reverse engineer disease states to develop interventions

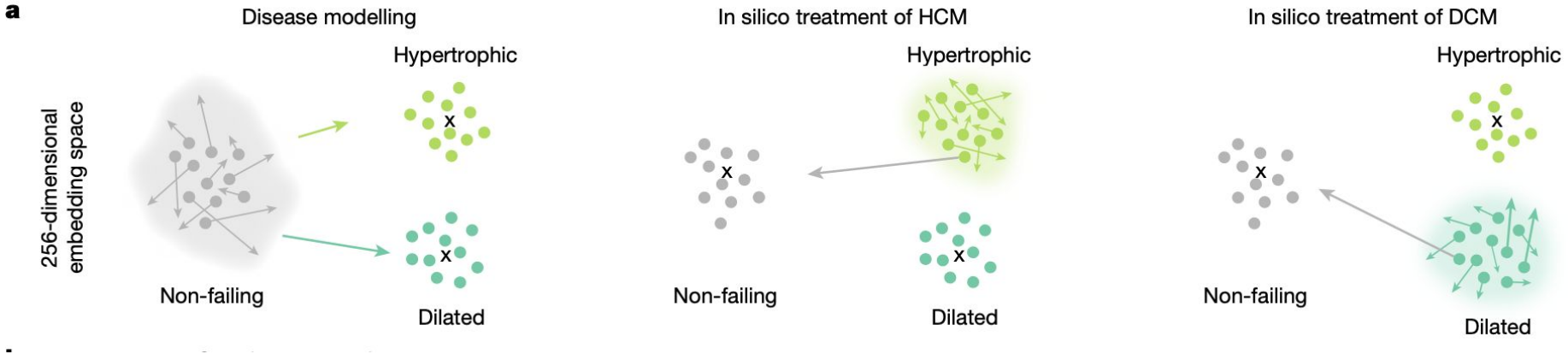
Rajewsky, Almouzni, Gorski, Nature (2020)

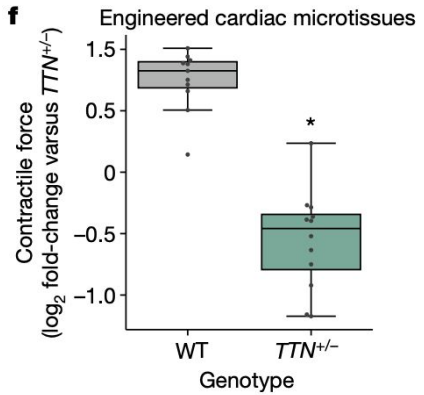
Single cell Foundation Model Applications

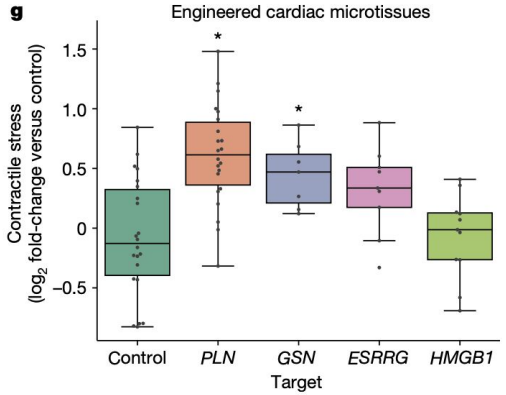


Szalata et al, Nature Methods 2024

A POC drug discovery task - Geneformer



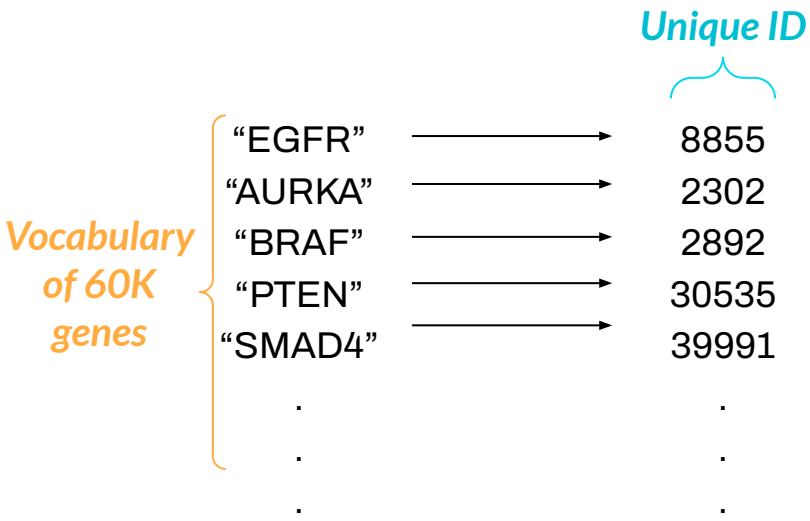




- Single cell gene expression data is used to profile normal cardiomyocytes and diseased ones from hypertrophic or dilated hearts
- TTN heterozygous KO (+/-) is used to create in vitro models of dilated cardiomyopathy (iPSC-derived cardiomyocytes in culture)
- Geneformer is used to predict targets for KD that are beneficial for dilated cardiomyopathy

Theodoris et al, Nature (2023)

Genes as tokens, Cells as sentences

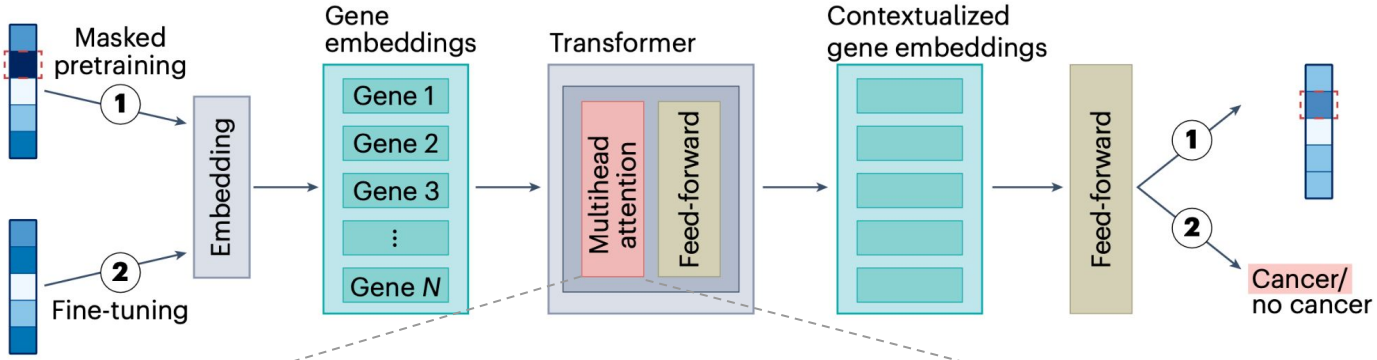


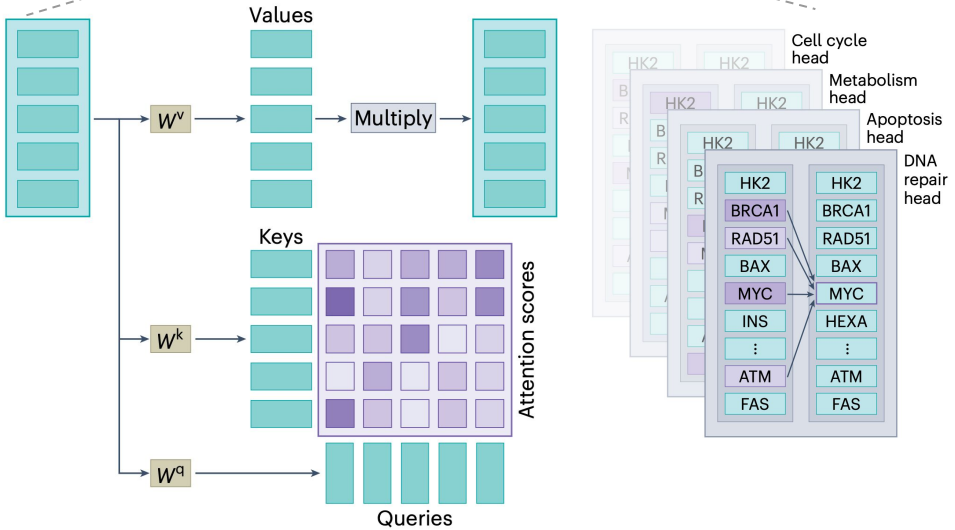
~45M unique single-cell transcriptomes in training set

Each cell has ~1-4k genes expressed (median 1.5k)

~100B tokens

Architecture for single cell gene network models





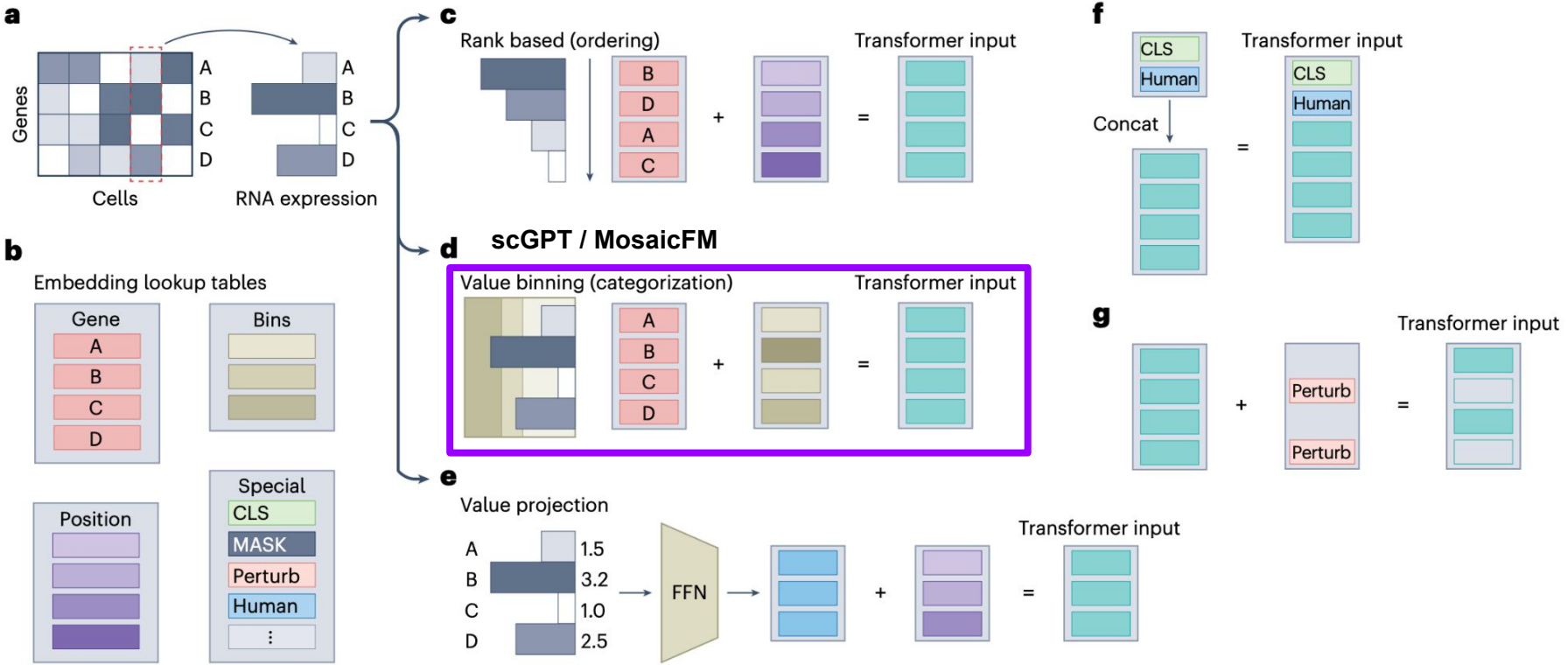
Self supervised learning objective:

Reconstruct masked gene expression signatures encoded in various ways

Attention mechanism:

Attention transforms embeddings based on their similarity, with different attention heads focusing on specific embedding components – this can be interpreted as the embedding of a gene being updated based on the embedding of other genes operating in similar functions / processes, thus capturing gene interactions and biological context

Expression Encoding



Szalata et al, Nature Methods 2024

Transformer Scaling Laws (Chinchilla 20:1 ratio)

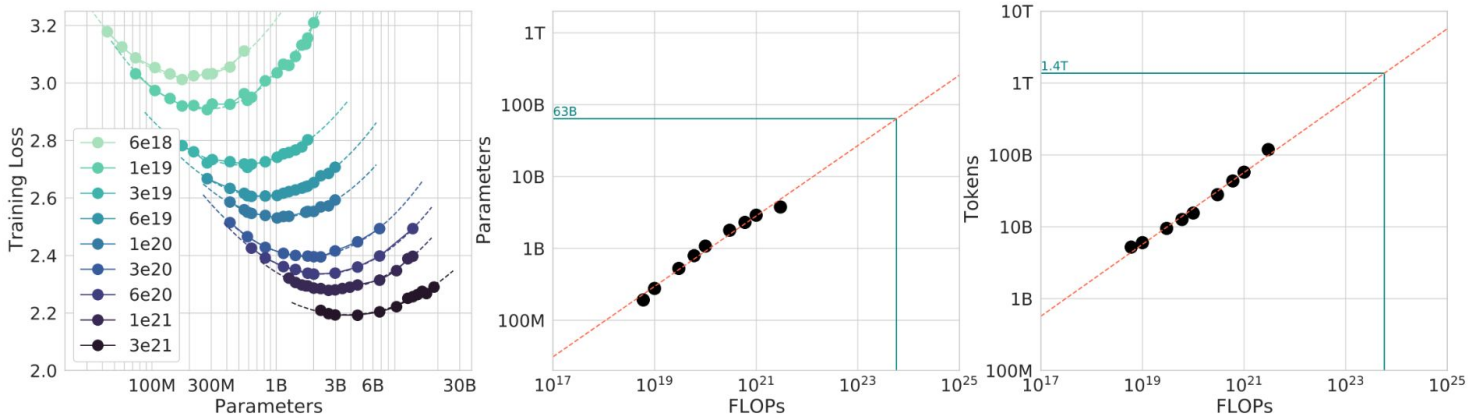
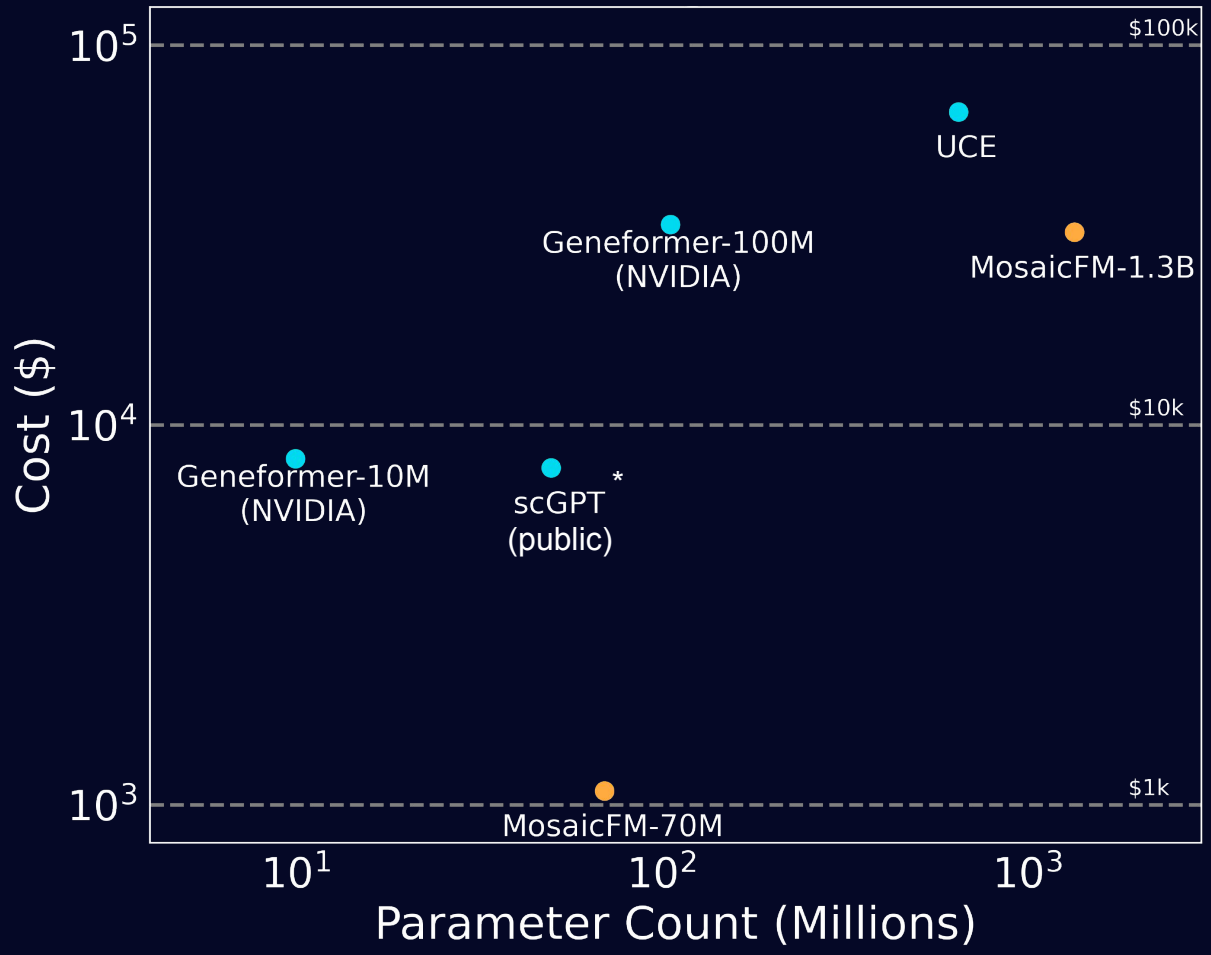


Figure 3 | **IsoFLOP curves.** For various model sizes, we choose the number of training tokens such that the final FLOPs is a constant. The cosine cycle length is set to match the target FLOP count. We find a clear valley in loss, meaning that for a given FLOP budget there is an optimal model to train (**left**). Using the location of these valleys, we project optimal model size and number of tokens for larger models (**center** and **right**). In green, we show the estimated number of parameters and tokens for an *optimal* model trained with the compute budget of *Gopher*.

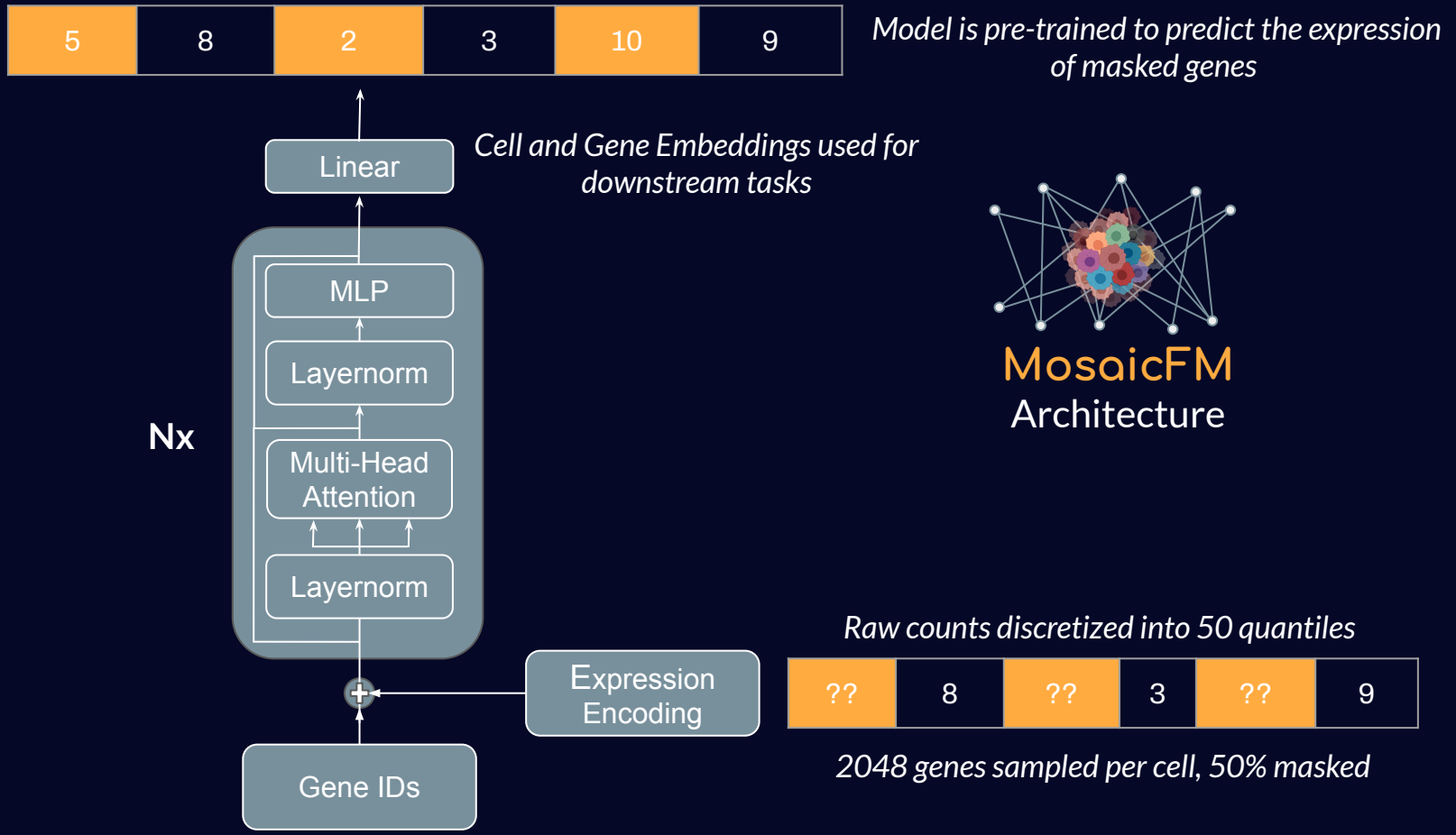
Hoffman, Borgeaud, Mensch et al, Neurips (2022)

We estimated the compute-optimal parameter size for the 50M cell dataset to be around 1.5B parameters.

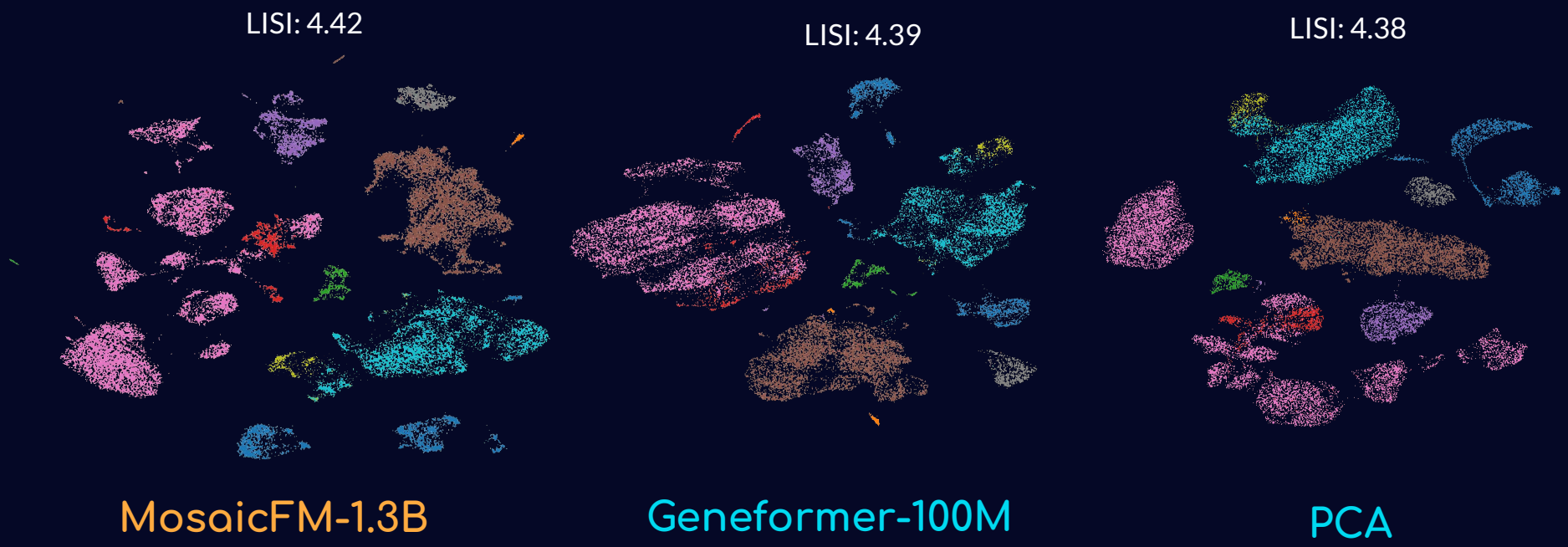
To scale in peace we needed to bring down the training cost



*Base architecture for MosaicFM


5
8
2
3
10
9
Model is pre-trained to predict the expression of masked genes
Linear
Cell and Gene Embeddings used for downstream tasks
MLP
Layernorm
Multi-Head Attention
Layernorm
Nx
MosaicFM
Architecture
Raw counts discretized into 50 quantiles
Expression Encoding
??
8
??
3
??
9
2048 genes sampled per cell, 50% masked
Gene IDs
NRAS
EGFR
PTEN
AURKA
SMAD4
BRAF
45M cells

Zero-shot cell type separation



Data from Kim, Nayoung, et al. Nature communications 11.1 (2020): 2285.

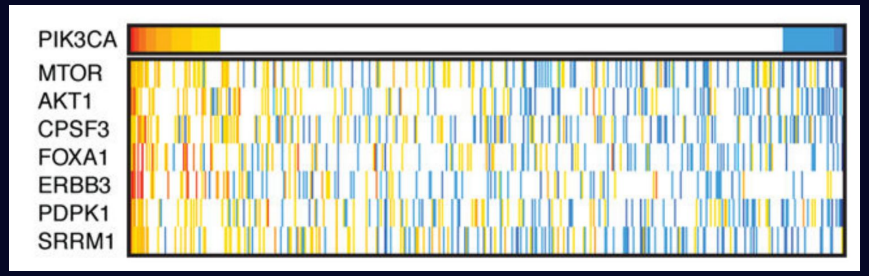
MosaicFM works well on a broad range of biologically meaningful benchmarks

Model	Cancer Essentiality (DepMap)		Gene-Set Membership (MSigDB)		Zero-shot Cell Type Clustering	
	Cell-type specific*	General†	All†	Hallmarks†	Lung Adenocarcinoma (single cell)‡	CCLE (bulk)‡
MosaicFM-1.3B	**0.64**	**0.73**	**0.15**	**0.32**	**4.42**	5.40
scGPT	0.63	0.66	0.13	0.29	4.35	4.30
Geneformer-100M	0.62	0.48	0.10	0.19	4.35	4.59
Baseline (PCA)	0.59	0.66	0.12	0.28	4.38	**6.70**

* auROC
† auPRC
‡ Local Inverse Simpson Index (LISI).

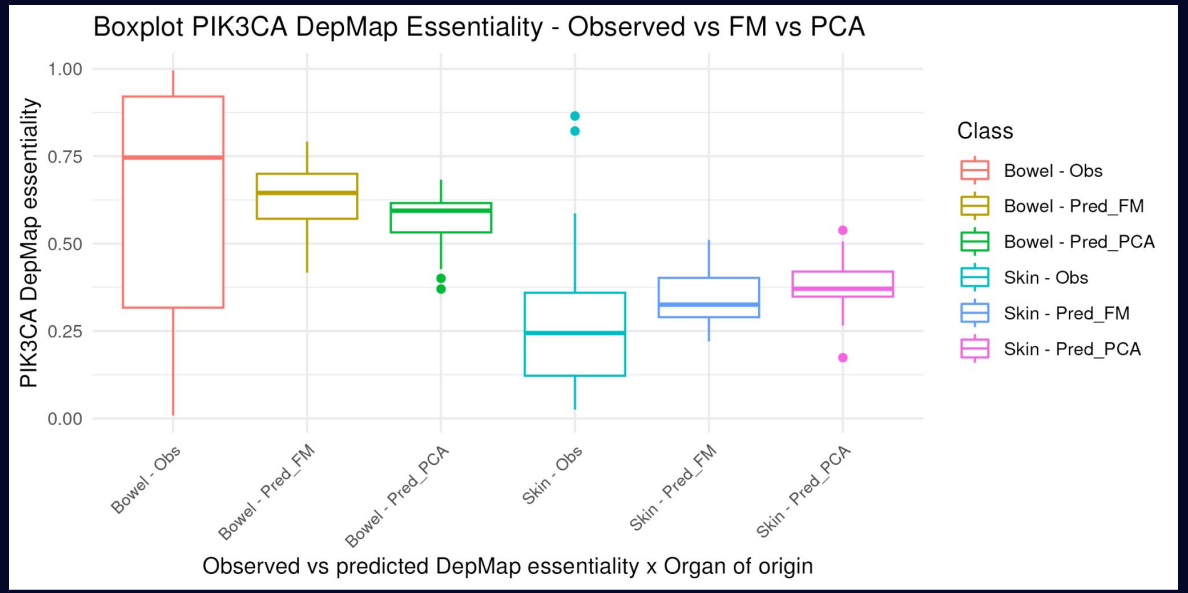
Cancer Dependency Map

Heatmap of CRISPRi essentiality scores (rows) across 501 screened lines (columns) for PIK3CA



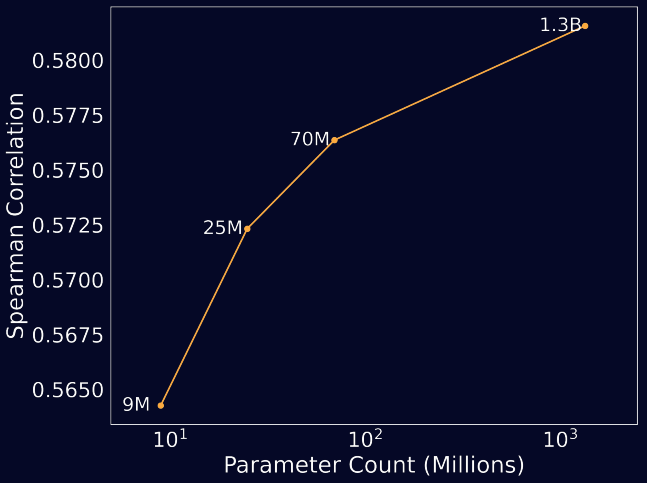
Foundation models should be able to predict:

1) For a given cell line, what genes are essential.
2) What genes are broadly essential across cell-lines

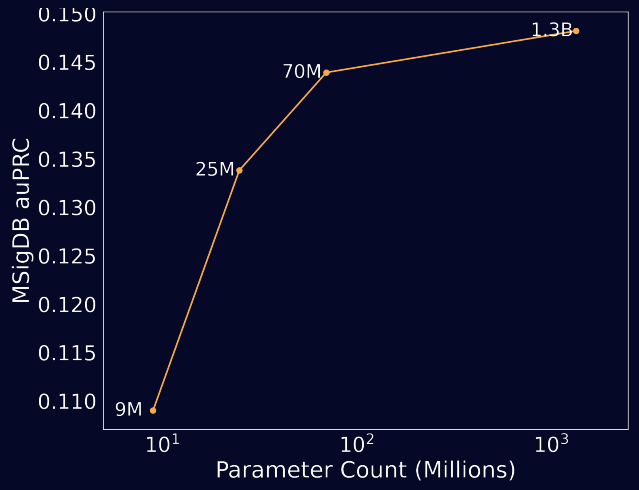


Tsherniak, Aviad, et al, Cell (2017)

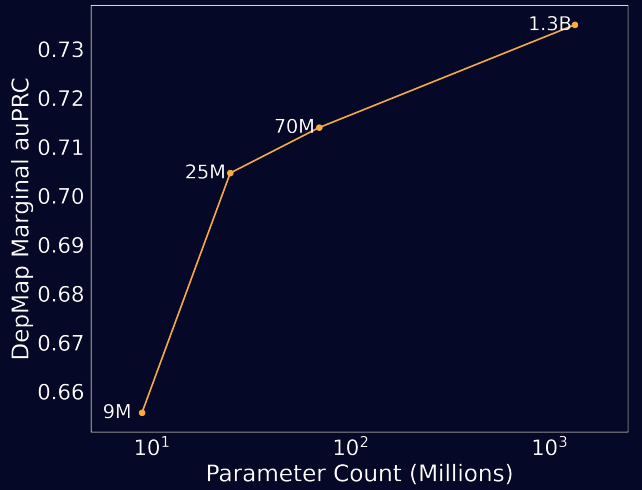
Impact of parameter scaling



Spearman correlation of predicted expression vs raw counts



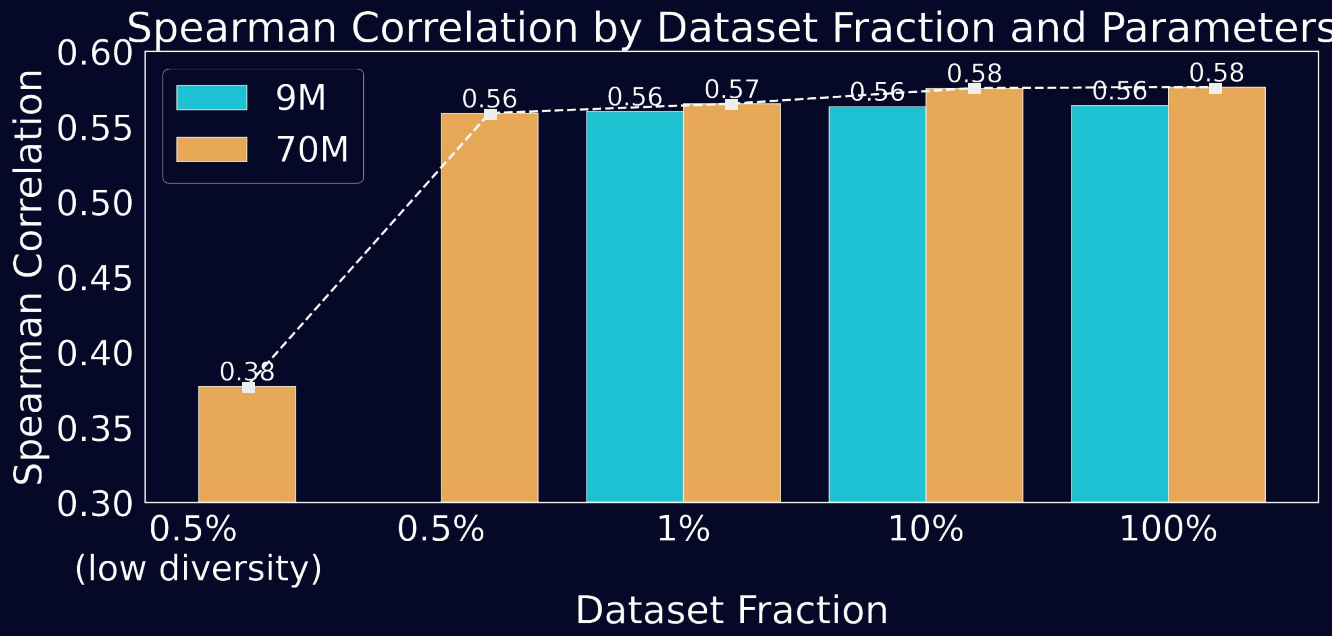
auPRC at predicting gene set membership in MSigDB



auPRC at predicting broadly essential genes in DepMap

Parameter scaling has expected behaviour across benchmarks

Impact of dataset scaling



Reducing cell-type diversity in training data has the largest impact on performance

What’s next?

- Single cell Foundation Models are a promising path towards developing virtual cell models.
- Despite promising results on benchmarks they still have a ways to go before being able to design and discover new drugs. To get there we need:
  - Better architectures and learning objectives - multimodal models!
  - Larger and more diverse datasets - perturbation atlases!
- Resources to check out:
  - https://cellxgene.cziscience.com/
  - https://github.com/bowang-lab/scGPT/tree/main/scgpt
  - https://huggingface.co/ctheodoris/Geneformer

ML/AI Team

Daniele Merico
CDO
U Milan PhD
ex Deep Genomics,
SickKids

Shreshth Gandhi,
Director of ML
IIT BTech, UofT MSc
ex Deep Genomics

Far Javadi,
ML Engineer
Sharif BSc, UBC MSc
ex Huawei

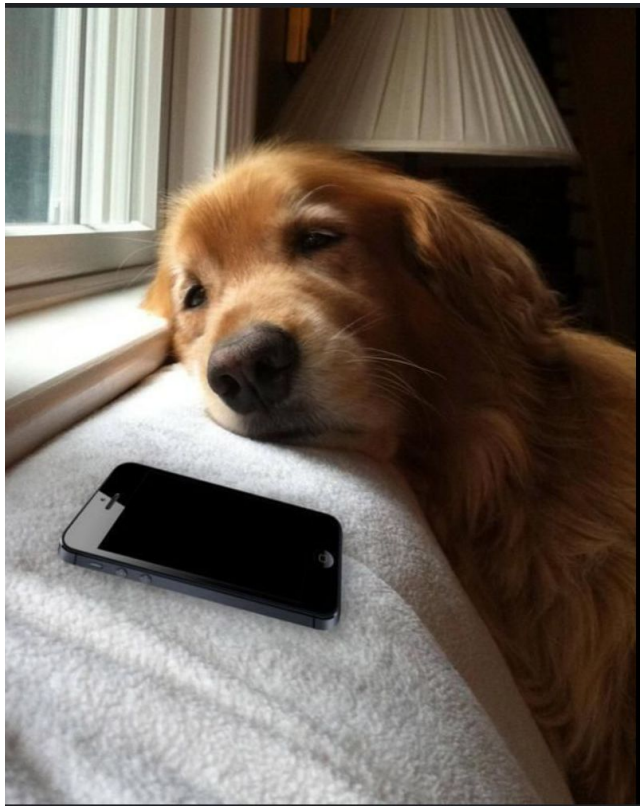
Valentine Svensson,
Principal Comp. Biol.
U Cambridge PhD
ex EMBL, Caltech,
Vesalius, Altos

Advisors

Hani Goodarzi
UCSF, Arc,
Vevo Co-Founder

Bo Wang
UofT, Vector Institute,
UHN
(scGPT lead PI)

Richard Socher
Founder, You.com



Stay tuned for upcoming models and datasets at https://github.com/vevotx