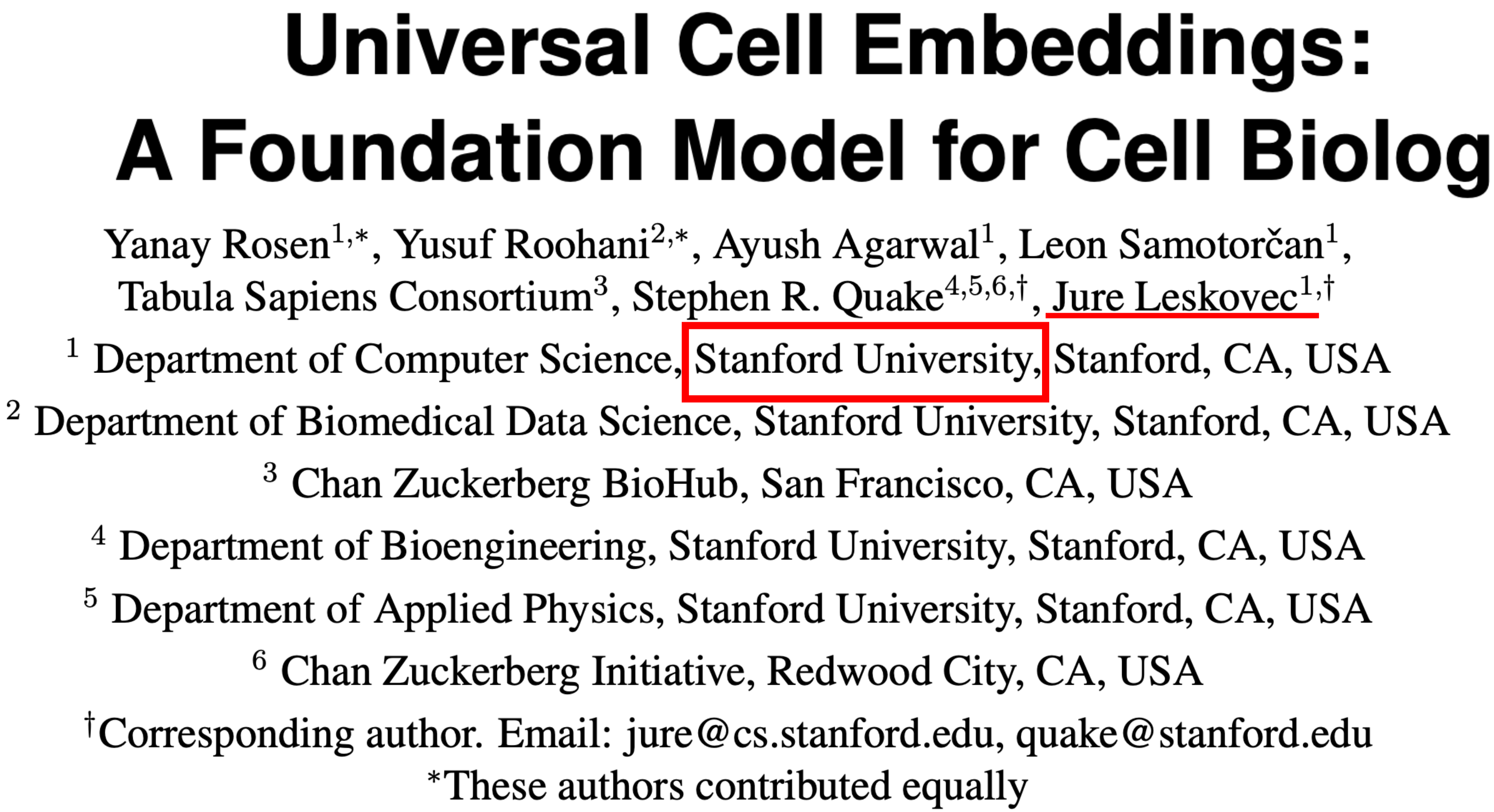
Universal Cell Embeddings:
A Foundation Model for Cell Biology
Wang-lab Journal Club

01/22/2024

Minghao WANG

Author Information



Universal Cell Embeddings: A Foundation Model for Cell Biology

Yanay Rosen[1,*], Yusuf Roohani[2,*], Ayush Agarwal[1], Leon Samotorčan[1],
Tabula Sapiens Consortium[3], Stephen R. Quake[4,5,6,†], Jure Leskovec[1,†]

[1] Department of Computer Science, Stanford University, Stanford, CA, USA

[2] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

[3] Chan Zuckerberg BioHub, San Francisco, CA, USA

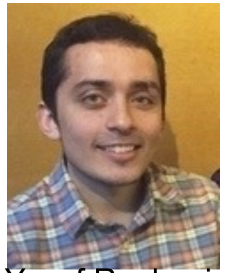
[4] Department of Bioengineering, Stanford University, Stanford, CA, USA

[5] Department of Applied Physics, Stanford University, Stanford, CA, USA

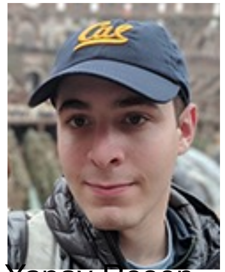
[6] Chan Zuckerberg Initiative, Redwood City, CA, USA

[†]Corresponding author. Email: jure@cs.stanford.edu, quake@stanford.edu

[*]These authors contributed equally



Yusuf Roohani



Yanay Rosen

Publish date: 2023.11.28
Currently on bioRXiv

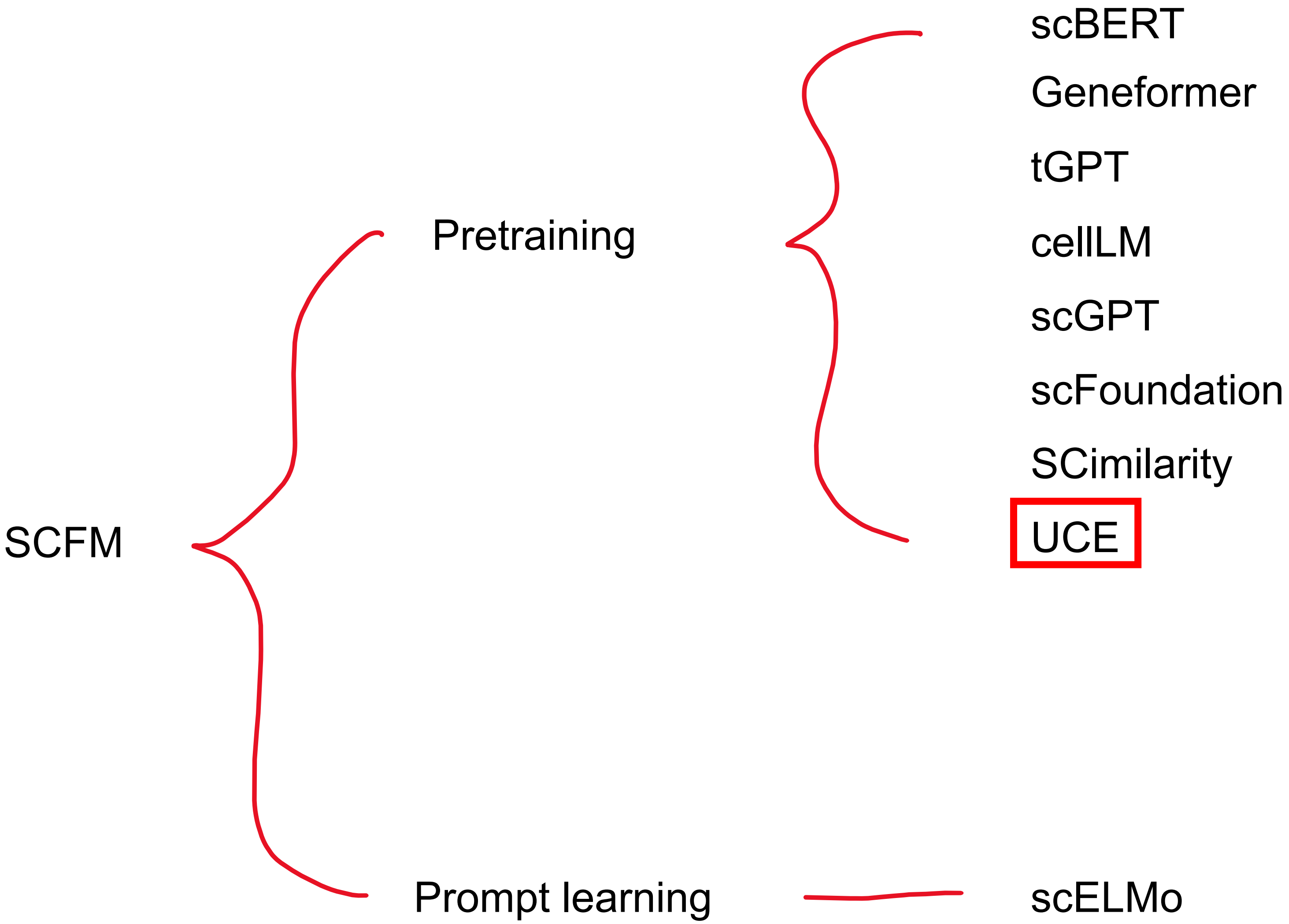
Jure Leskovec

Snap Stanford

Background

1. Taxonomy of Single Cell Foundation Models (SCFM)

- Single cell transcriptome --> single Cell Foundation Models (SCFM) --> Embedding of genes for each cell



SCFM

Pretraining

scBERT

Geneformer

tGPT

cellLM

scGPT

scFoundation

SCimilarity

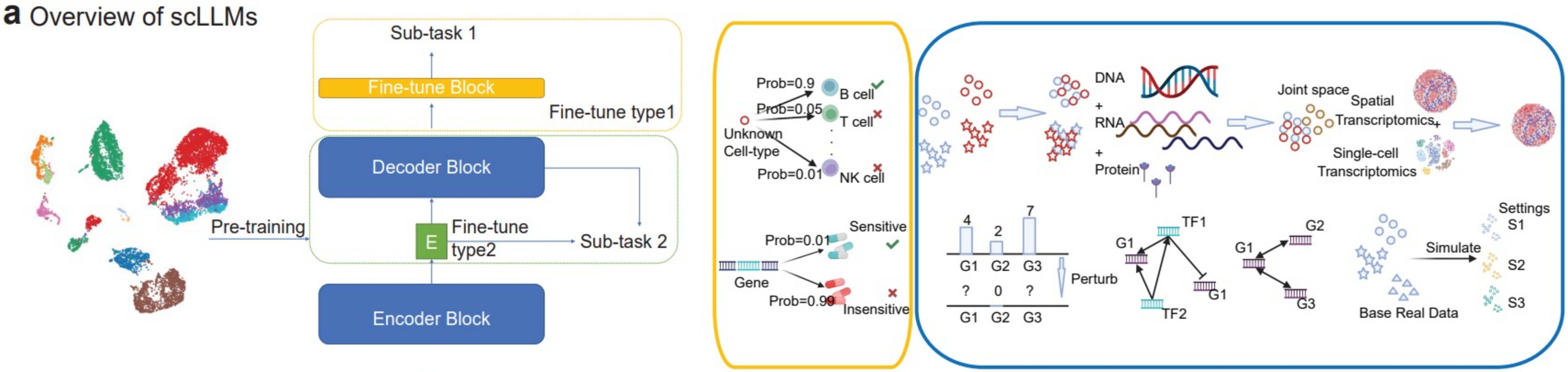
UCE

Prompt learning

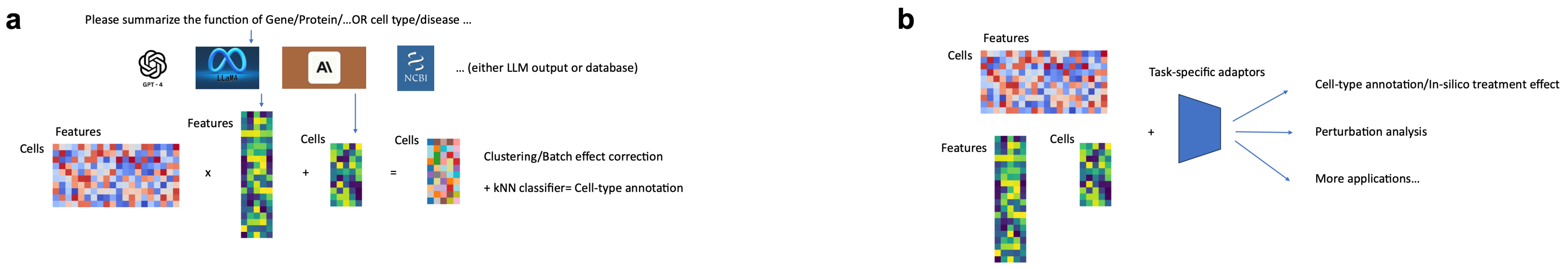
scELMo

Background

2. Overview of SCFM



An example of single-cell LLM


a
Please summarize the function of Gene/Protein/...OR cell type/disease ...
GPT-4
LLaMA
A\
NCBI
... (either LLM output or database)
Features
Cells
x
Features
Cells
+
Cells
=
Clustering/Batch effect correction
+ kNN classifier= Cell-type annotation
b
Features
Cells
Features
Cells
+
Task-specific adaptors
Cell-type annotation/In-silico treatment effect
Perturbation analysis
More applications...

Tianyu Liu et al, Evaluating the Utilities of Large Language Models in Single-cell Data Analysis, bioRxiv 2023.11

Tianyu LIU et al. scELMo: Embeddings from Language Models are Good Learners for Single-cell Data Analysis. bioRXiv 2023.12

Background

- For one gene in different cells, the embedding is different.
- Current sccfm has following issues:

✓ Gene length issue

✓ Species issue (mainly affect gene embedding)

✓ Finetune issue (GPU resources)

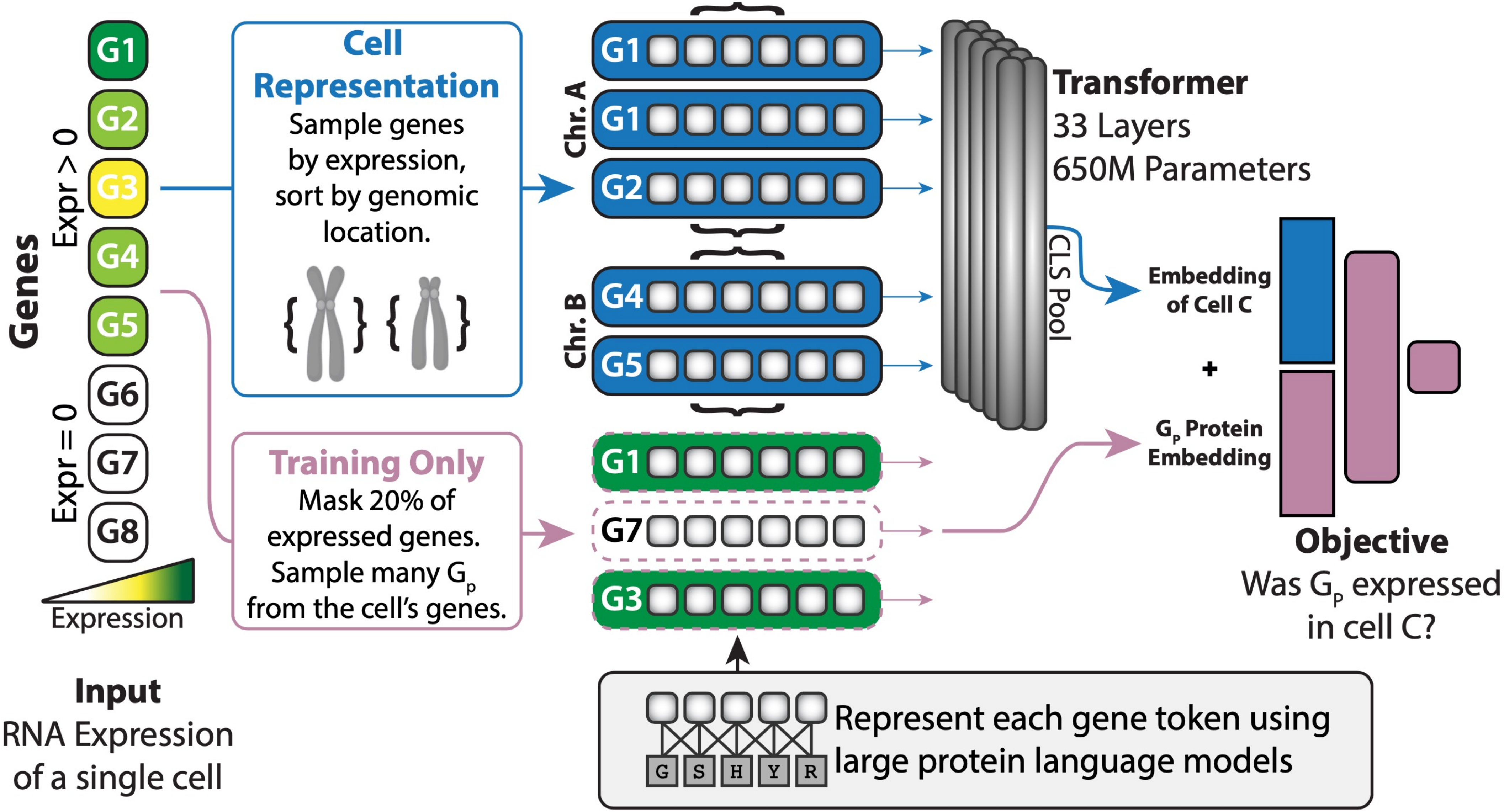
UCE contributions

Contribution

- A **foundation model called UCE** that can generate an embedding of **all species** **without finetuning**

- A dataset called Integrated Mega-scale Atlas (IMA) created by applying UCE with 36M cells, more than 1,000 uniquely named cell types, from hundreds of experiments, dozens of tissues and eight species. (not yet published)

Model



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.

Transformer
33 Layers
650M Parameters

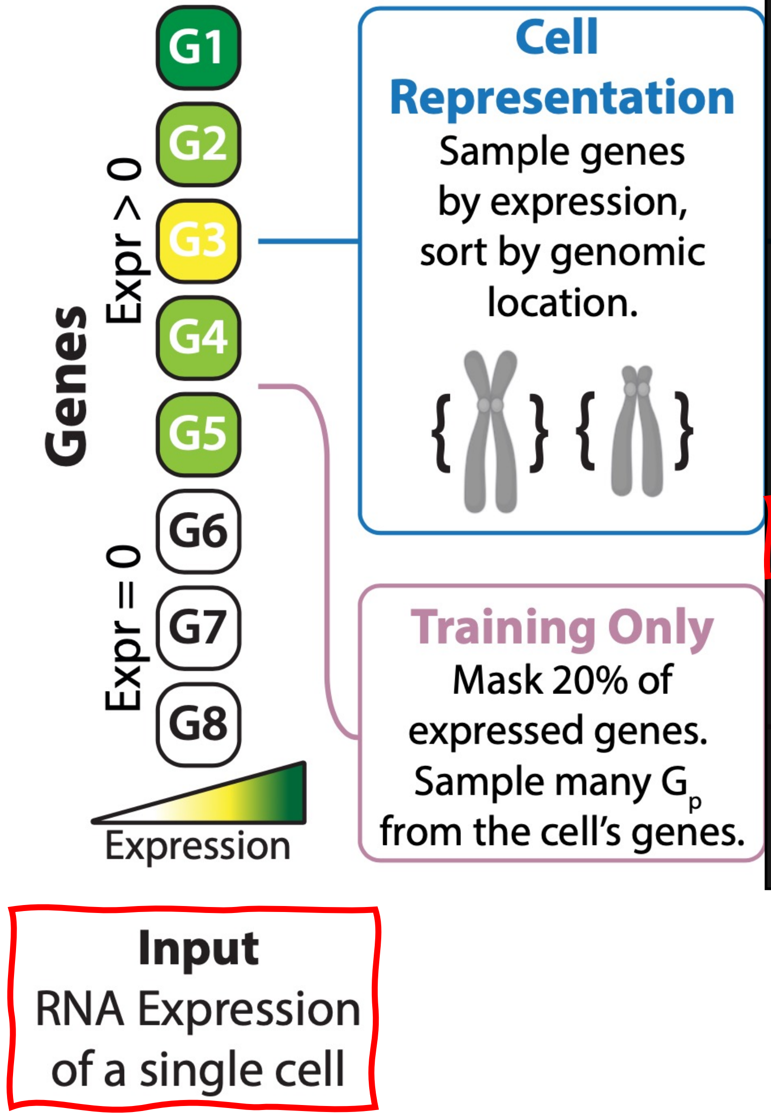
Embedding of Cell C

G_p Protein Embedding

Objective
Was G_p expressed in cell C?

Represent each gene token using large protein language models

Model



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.

Input
RNA Expression of a single cell

1. Input:

with replacement

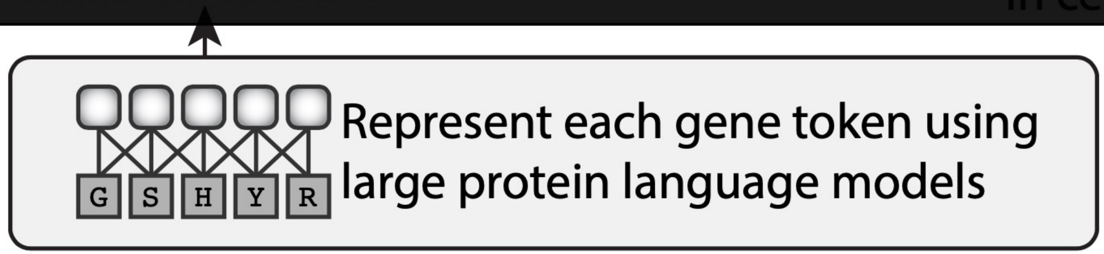
- **Weighted sample** of normalized gene expression, **grouped** by chromosome and **sorted** by genome location --> expression part
- Represent each gene with **protein language models** --> gene part

represent one cell, {} represent one chromosome

Final input:

<CLS> {G2P2 G2P2 G3P3 G1P1 G8P8} {G5P5 G6P6 G6P6 …}, …

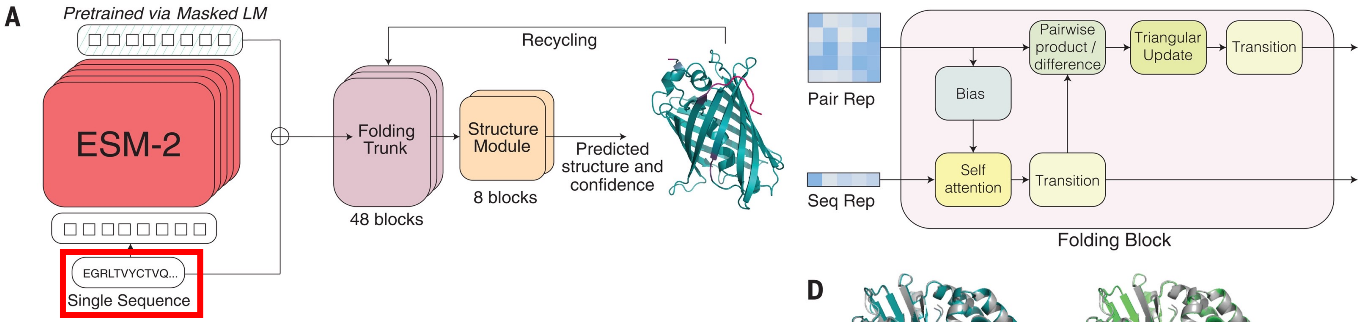
<CLS> is specially designed in BERT pretraining scheme, a randomly initialized vector, used to represent the whole embedding of a cell after passing models.



Represent each gene token using large protein language models

Large Protein Language Models

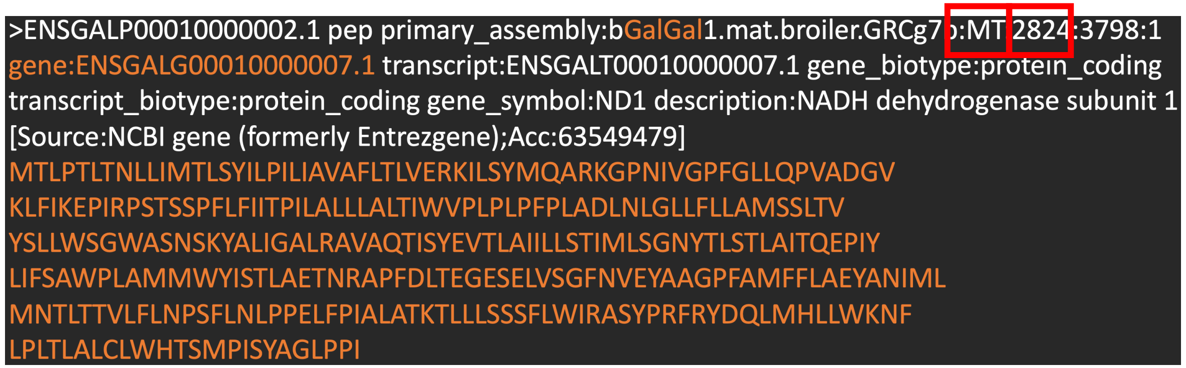
UCE use ESM-2 model to generate protein embedding.



Input: Protein sequence
Output: Protein embedding

Need to convert gene names to protein sequence

In https://www.ensembl.org/, we can download the files to do so.



```
>ENSGALP00010000002.1 pep primary_assembly:bGalGal1.mat.broiler.GRCg7b:MT:2824:3798:1
gene:ENSGALG00010000007.1 transcript:ENSGALT00010000007.1 gene_biotype:protein_coding
transcript_biotype:protein_coding gene_symbol:ND1 description:NADH dehydrogenase subunit 1
[Source:NCBI gene (formerly Entrezgene);Acc:63549479]
MTLPTLTNLLIMTLSYILPILIAVAFLTLVERKILSYMQARKGPNIVGPFGLLQPVADGV
KLFIKEPIRPSTSSPFLFIITPILALLLALTIWVPLPLPFPLADLNLGLLFLLAMSSLTV
YSLLWSGWASNSKYALIGALRAVAQTISYEVTLAIILLSTIMLSGNYTLSTLAITQEPIY
LIFSAWPLAMMWYISTLAETNRAPFDLTEGESELVSGFNVEYAAGPFAMFFLAEYANIML
MNTLTTVLFLNPSFLNLPPELFPIALATKTLLLSSSFLWIRASYPRFRYDQLMHLLWKNF
LPLTLALCLWHTSMPISYAGLPPI
```

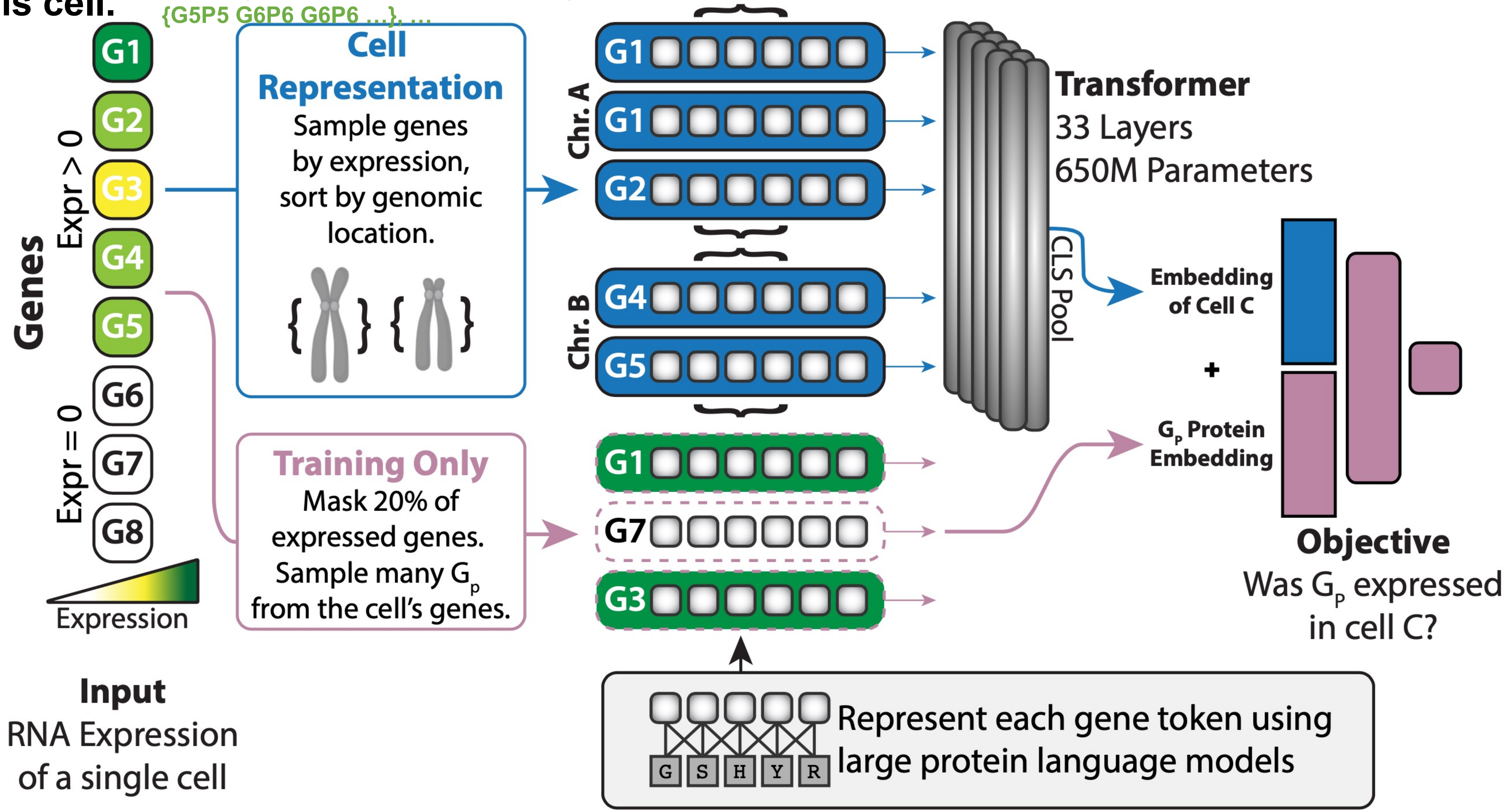
Fig. An example of sample chicken gene and corresponding protein sequences.

Model

2. Pretraining scheme:

- **Mask 20% of expressed genes + sample non-expressed genes**
- **Use the final output of <cls> + 0-exp genes protein embedding to predict whether it was expressed in this cell.**

<CLS> {G2P2 G2P2 G3P3 G1P1 G8P8} {G5P5 G6P6 G6P6 …}, …



Model

3. Pretraining details:

- **33-layer transformer with 650M parameters.**

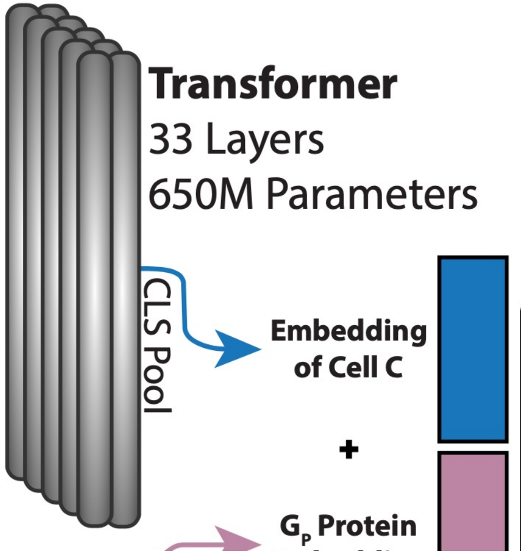
33.9M human + mouse from CxG, 2.3M 8 species

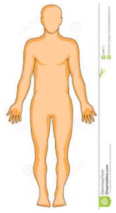
- **Pretrained on more than 300 datasets by CellXGene Corpus, consisting of > 36M cells**

- **Using 24 A100 80GB GPUs for 40 days.**

If use AWS: ~ HK$ 322k

If use other online platform: ~ HK$ 169k




human

mouse

Mouse lemur

Zebrafish

pig

rhesus macaque

Crab eating macaque

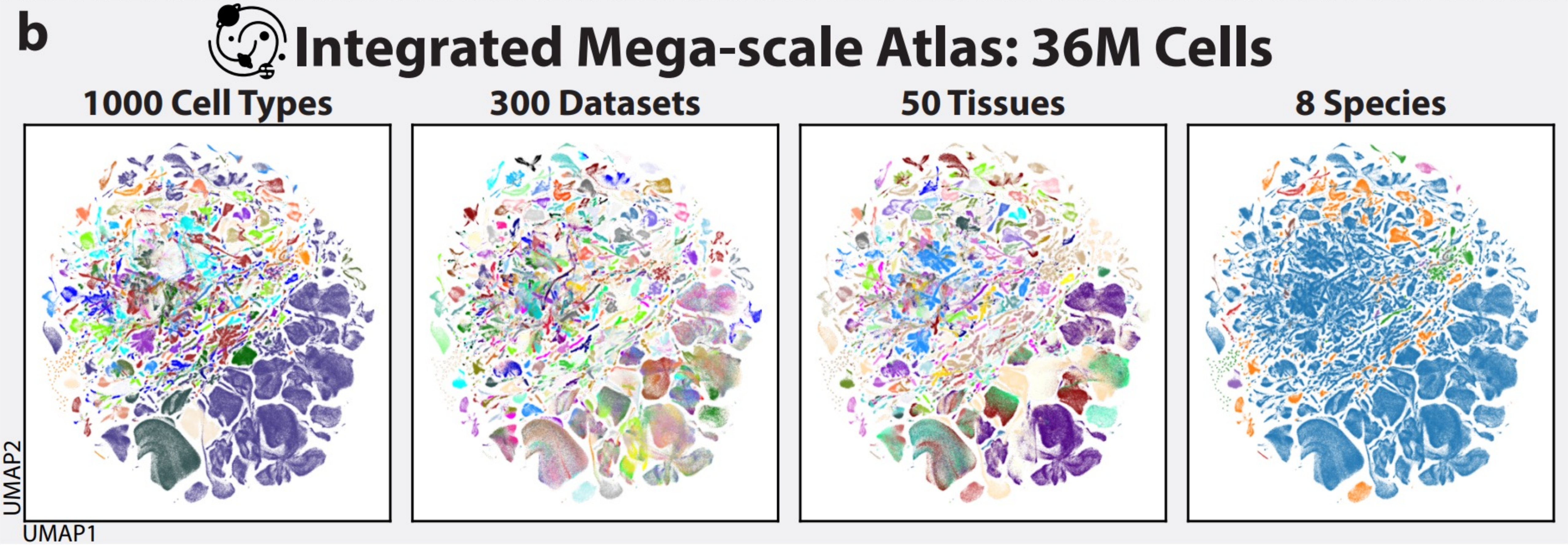
western clawed frog

Results

1. UCE creates an Interated Mega-scale Atlas of 36M cells.

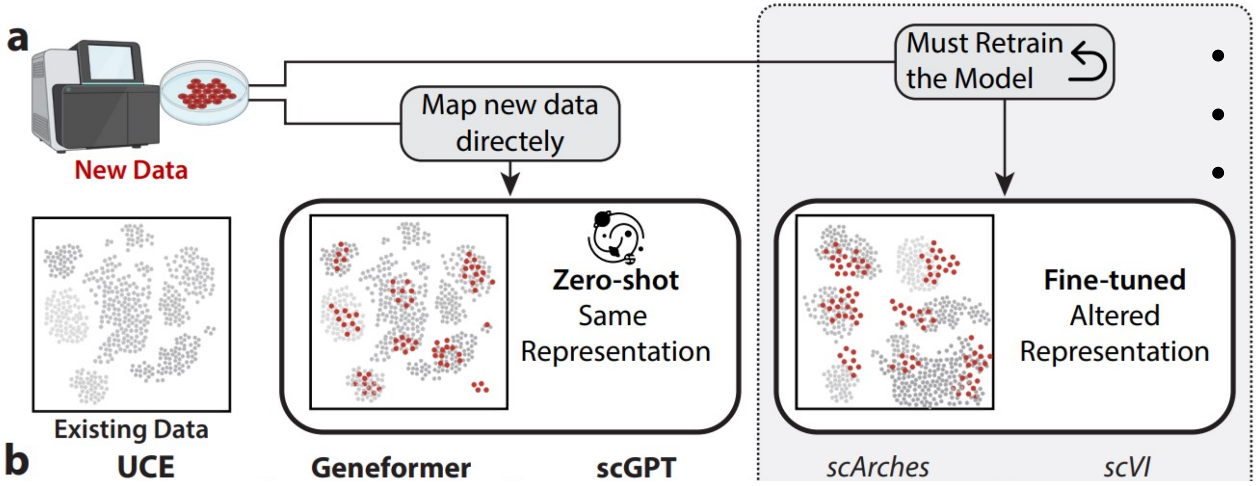
Sampled from diverse biological conditions

- **After pretraining (do not use labels), apply UCE on the same dataset to generate embedding and perform UMAP. Cells within UCE space naturally cluster by biological conditions (cell types, etc.) while mixing among experimental conditions (batch).**



Results

2. UCE embeds new datasets without additional model training



Zero-shot

- Evaluate the universality of UCE rep in 0-shot setting.
- These data are not appeared in training set.
- Also compare with commonly used finetuned methods

Evaluation Dataset: Tabula Sapiens v2 (contribution2)

- Human data from 581k cells, 27 tissues, 167 batches and 162 unique cell types.

Evaluation Metric

- scIB

Malte D. Luecken, et.al Benchmarking atlas-level data integration in single-cell genomics. Nature Methods 2022.1

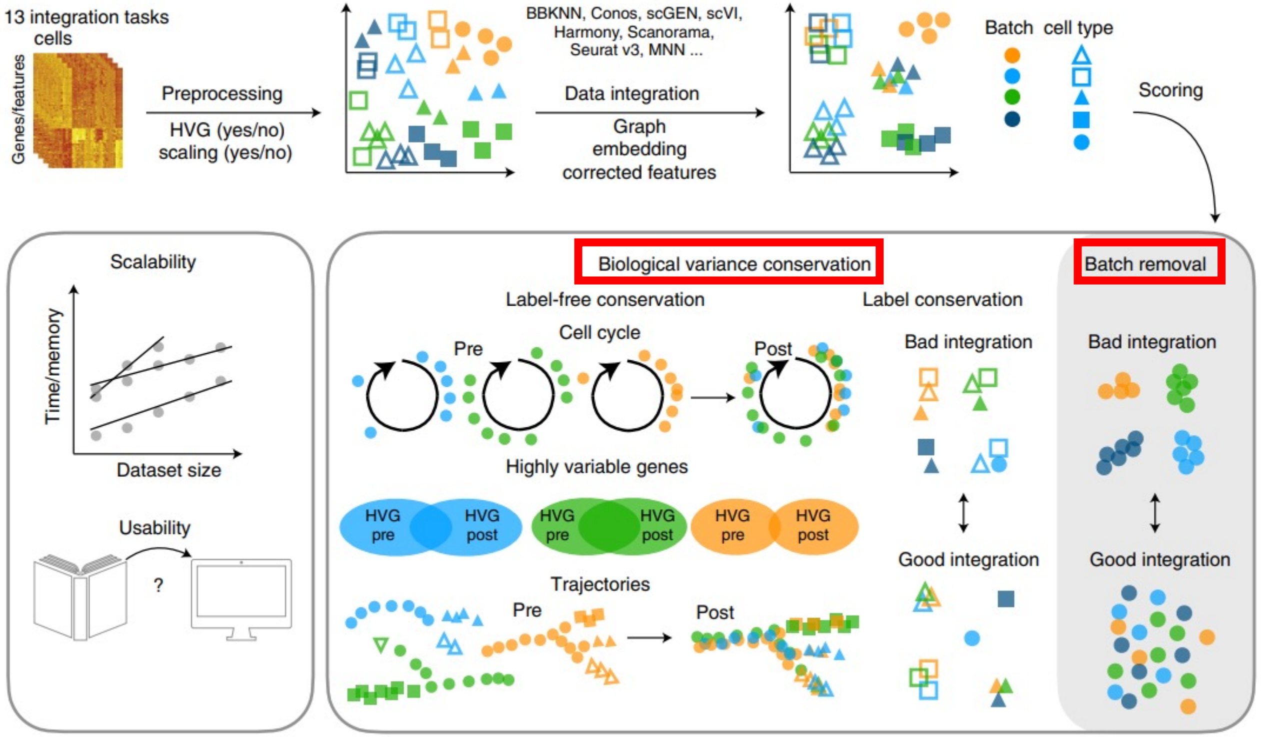
Results

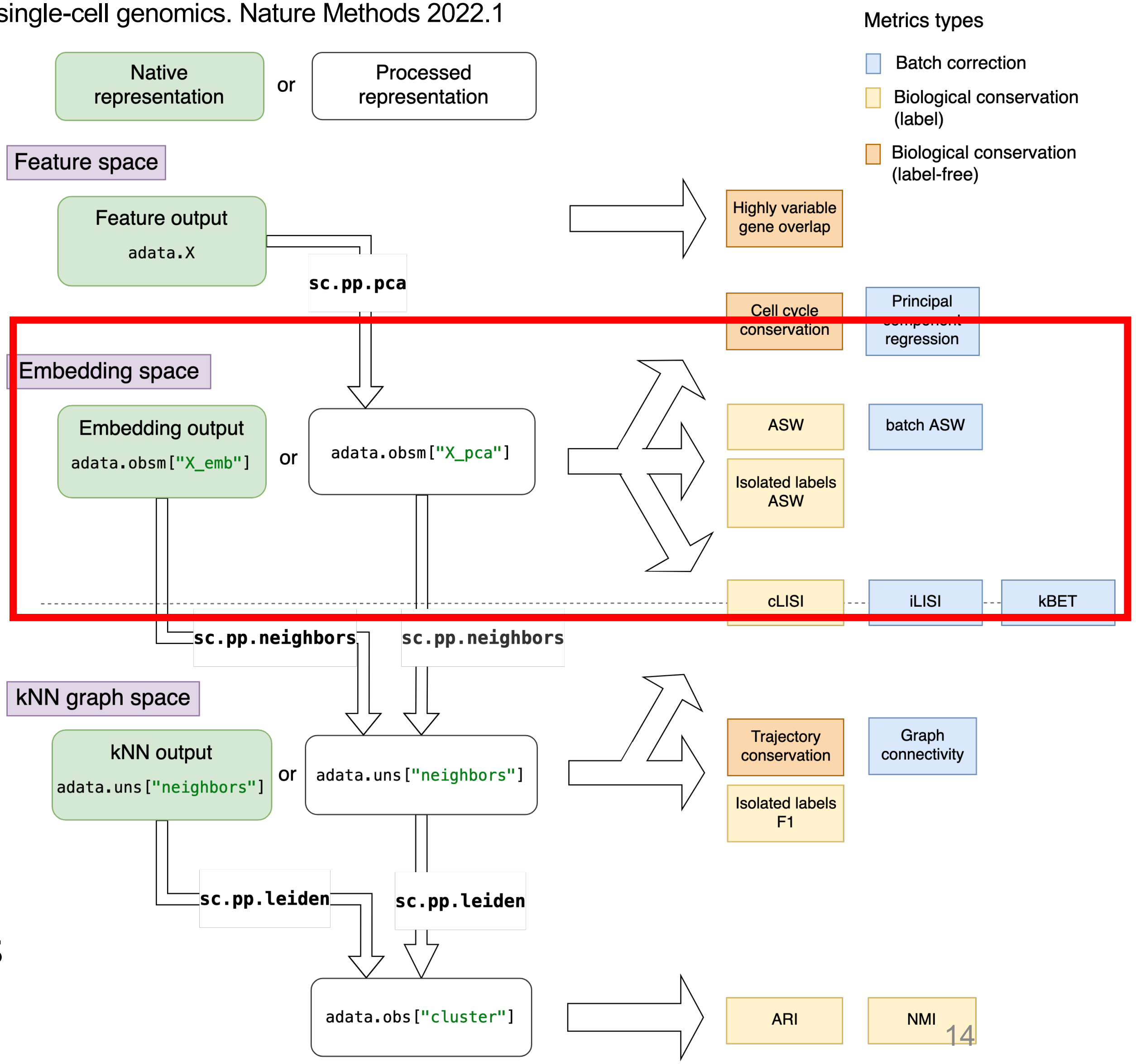
2. UCE embeds new datasets without additional model training

Malte D. Luecken, et.al Benchmarking atlas-level data integration in single-cell genomics. Nature Methods 2022.1

Evaluation Metric

- **The conservation of cell type information & batch correction**





Metrics types

□ Batch correction

□ Biological conservation (label)

□ Biological conservation (label-free)

Native representation or Processed representation

Feature space

Feature output
`adata.X`

`sc.pp.pca`

Embedding space

Embedding output
`adata.obsm["X_emb"]`
or
`adata.obsm["X_pca"]`

`sc.pp.neighbors`
`sc.pp.neighbors`

kNN graph space

kNN output
`adata.uns["neighbors"]`
or
`adata.uns["neighbors"]`

`sc.pp.leiden`
`sc.pp.leiden`

`adata.obs["cluster"]`

Highly variable gene overlap

Cell cycle conservation

Principal component regression

ASW

batch ASW

Isolated labels ASW

cLISI

iLISI

kBET

Trajectory conservation

Graph connectivity

Isolated labels F1

ARI

NMI

Evaluation Dataset: Tabula Sapiens v2 (contribution2)

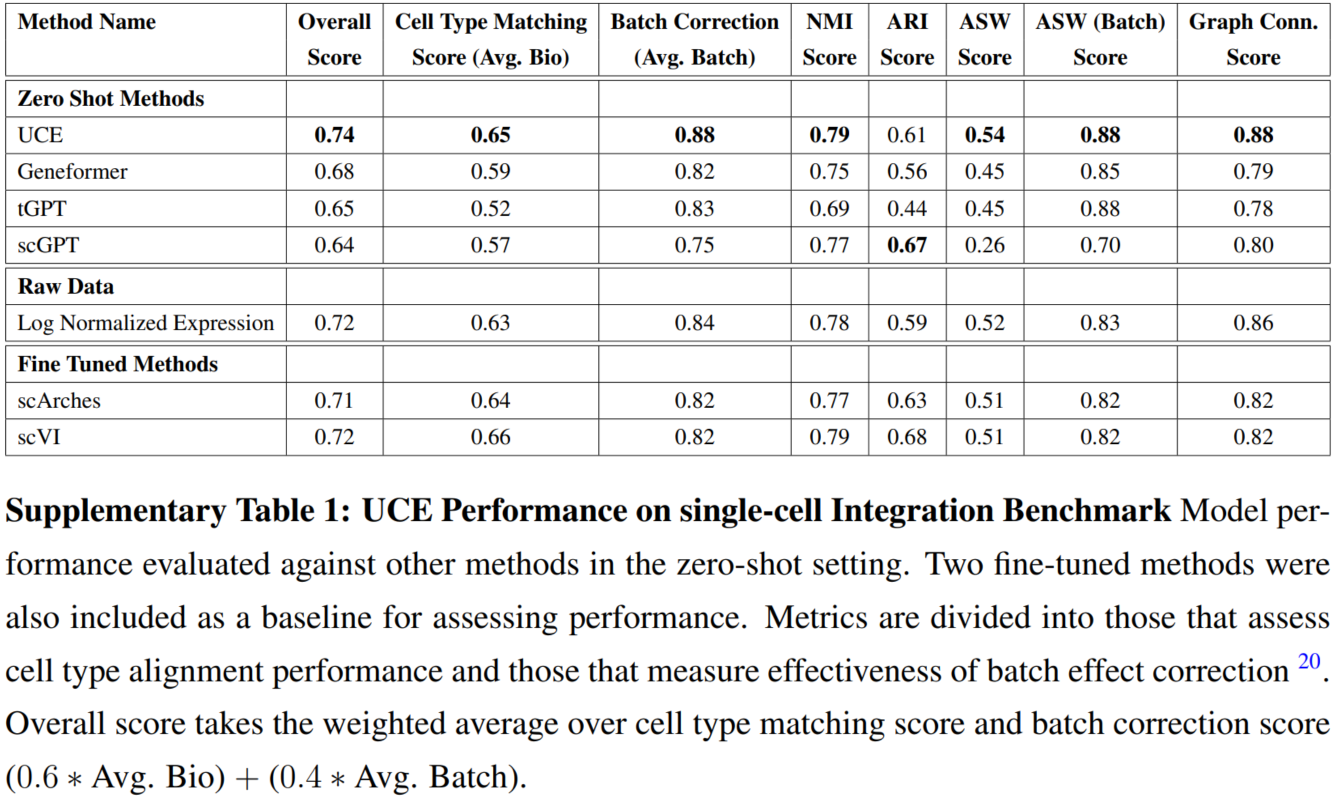
- Human data from 581k cells, 27 tissues, 167 batches and 162 unique cell types.

Results

2. UCE embeds new datasets without additional model training

Evaluation Metric

Malte D. Luecken, et.al Benchmarking atlas-level data integration in single-cell genomics. Nature Methods 2022.1



Method Name	Overall Score	Cell Type Matching Score (Avg. Bio)	Batch Correction (Avg. Batch)	NMI Score	ARI Score	ASW Score	ASW (Batch) Score	Graph Conn. Score
Zero Shot Methods								
UCE	**0.74**	**0.65**	**0.88**	**0.79**	0.61	**0.54**	**0.88**	**0.88**
Geneformer	0.68	0.59	0.82	0.75	0.56	0.45	0.85	0.79
tGPT	0.65	0.52	0.83	0.69	0.44	0.45	0.88	0.78
scGPT	0.64	0.57	0.75	0.77	**0.67**	0.26	0.70	0.80
Raw Data								
Log Normalized Expression	0.72	0.63	0.84	0.78	0.59	0.52	0.83	0.86
Fine Tuned Methods								
scArches	0.71	0.64	0.82	0.77	0.63	0.51	0.82	0.82
scVI	0.72	0.66	0.82	0.79	0.68	0.51	0.82	0.82

Supplementary Table 1: UCE Performance on single-cell Integration Benchmark Model performance evaluated against other methods in the zero-shot setting. Two fine-tuned methods were also included as a baseline for assessing performance. Metrics are divided into those that assess cell type alignment performance and those that measure effectiveness of batch effect correction [20]. Overall score takes the weighted average over cell type matching score and batch correction score $(0.6 * \text{Avg. Bio}) + (0.4 * \text{Avg. Batch})$.

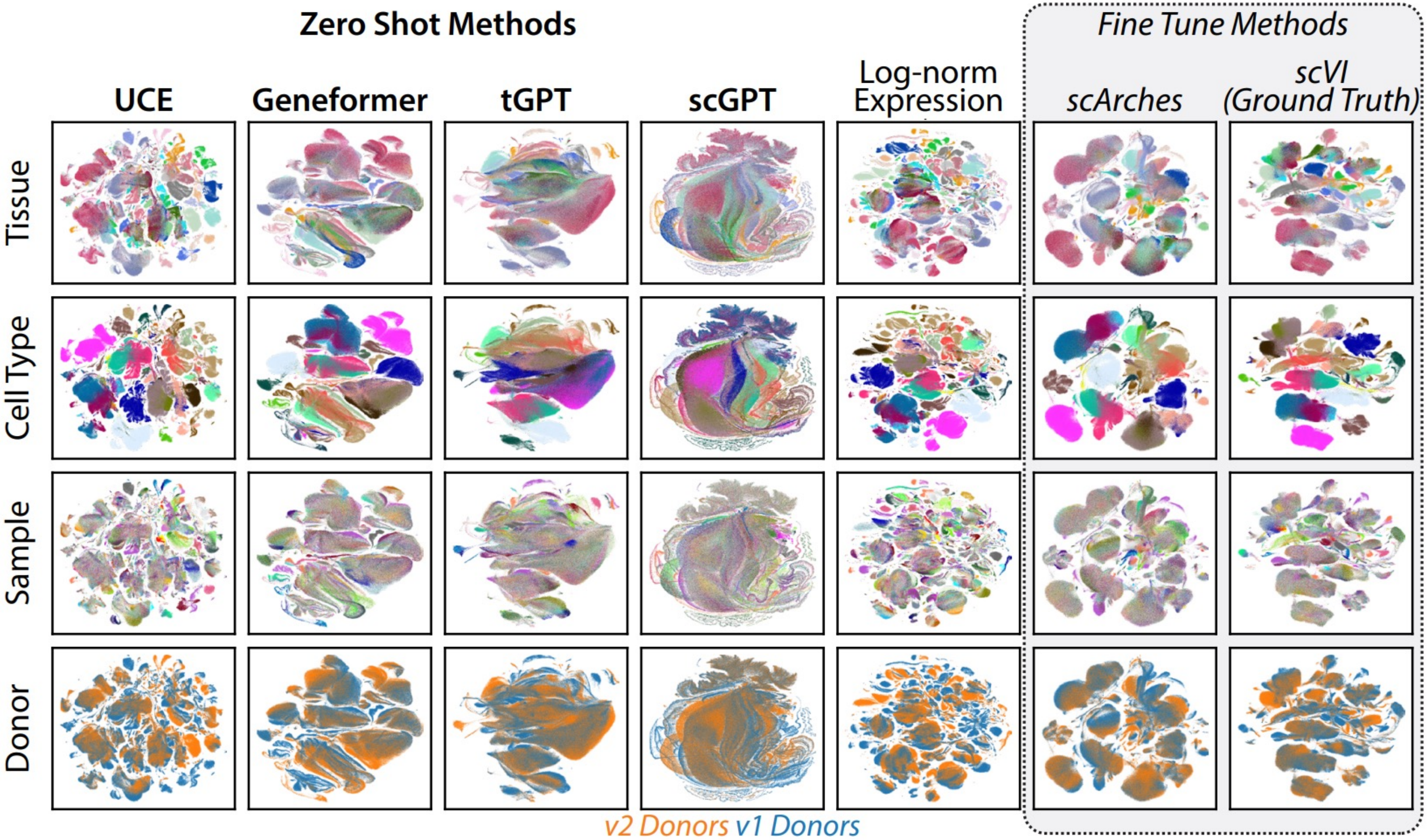
Evaluation Dataset: Tabula Sapiens v2 (contribution2)

- Human data from 581k cells, 27 tissues, 167 batches and 162 unique cell types.

Results

2. UCE embeds new datasets without additional model training

Embeddings on Tabula Sapiens v1 & v2

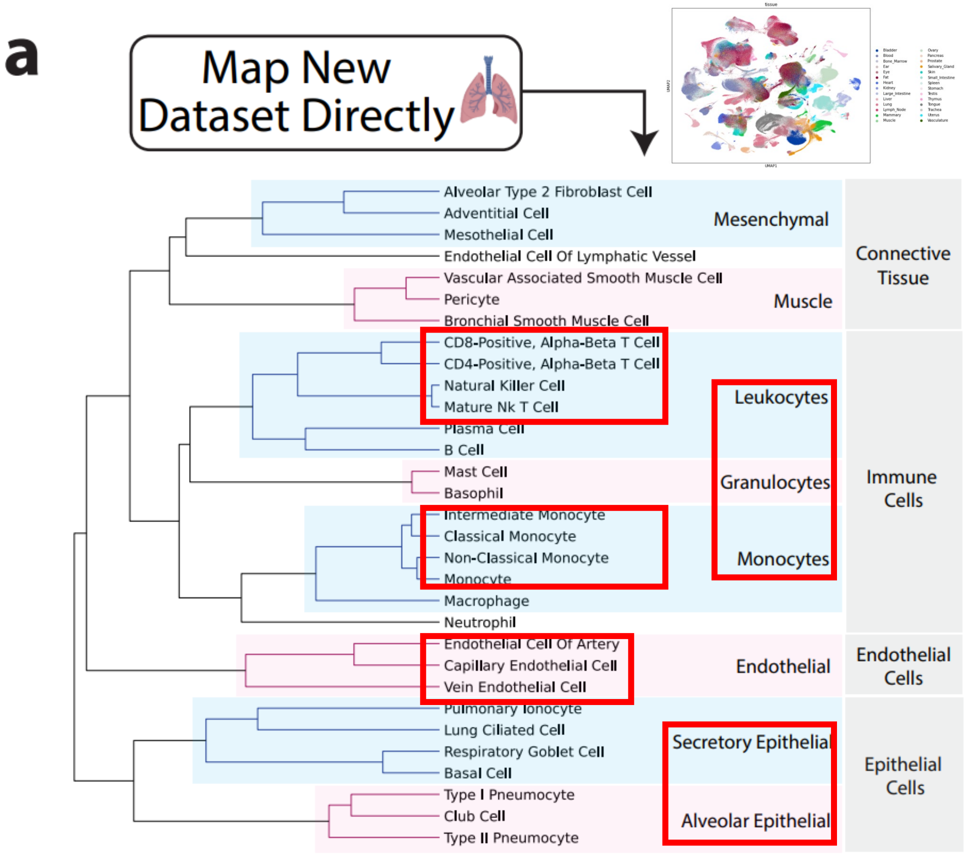


- Can separate cell types more effectively
- UCE emb resembles finetuned models
- Cell types align correctly regardless of whether the data was drawn from new or previously seen donors.

Results

3. UCE learns a meaningful organization of cell types in previously unseen data

Dendrogram for the generated embedding

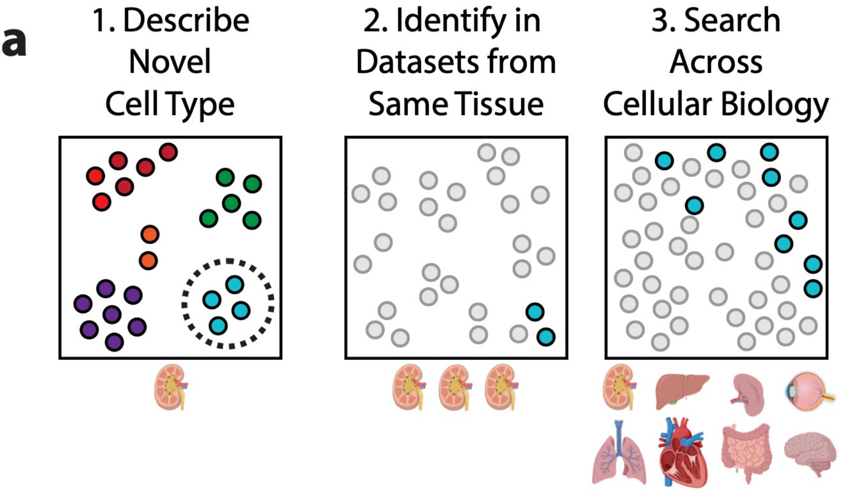


- Lung tissue UCE embedding --> Dendrogram
- Distinct cell types (T, monocytes, endothelial cells), and even high-level categories

Results

4. A workflow for decoding the function of newly discovered cell types

Overview of a novel single cell analysis workflow that UCE facilitates

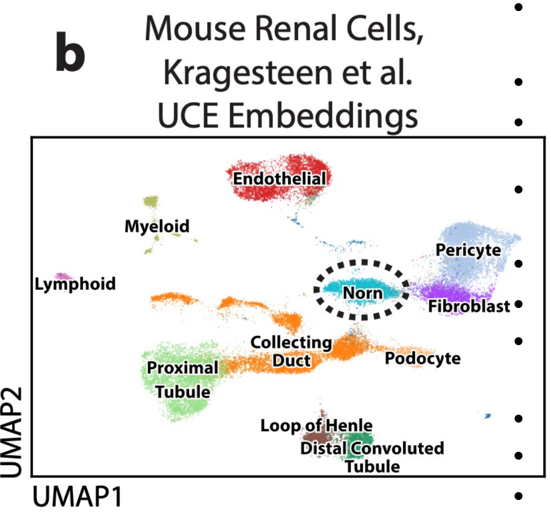

a
1. Describe Novel Cell Type
2. Identify in Datasets from Same Tissue
3. Search Across Cellular Biology

- Step 1: Identification of a novel cell type (circled blue ones) in the embedding space.
  - Using methods such as clustering and confirmation using marker gene analysis.
- Step 2: Identify the novel cell type in other datasets profiled from the same tissue.
  - Train a simple logistic classifier to predict cell types from UCE embedding and then applied to embeddings from other datasets of the same tissue, to confirm the cell type's existence and improve its characterization.
- Step 3: Apply the same classifier to the embeddings of cells from any other tissue to find cell types with similar biological functions or patterns of gene expression.

Kragesteen, B.K. et al. The transcriptional and regulatory identity of erythropoietin producing cells. Nat Med 2023.4

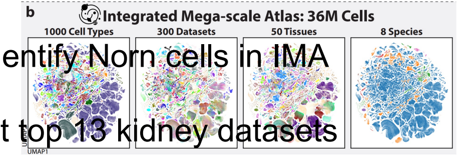
Case study: recently identified kidney Norn cell

促红细胞生成素
에리스로포이에틴

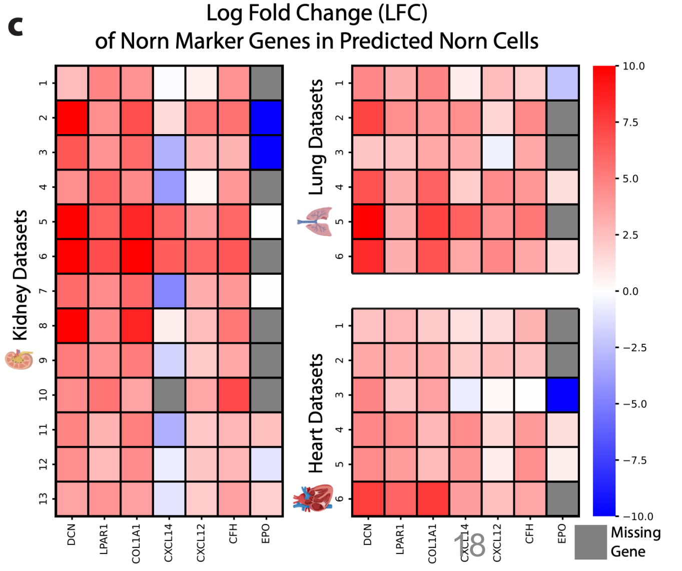


- Kidney Norn cell: the long-sought erythropoietin (Epo) producing cell in the kidney, is characterized as fibroblast-like.
- Use dataset provided by the Nat Med paper, generate an embedding.
- This embedding produces a cluster of cells corresponding to Norn.

- Then use a logistic classifier trained on this embedding to identify Norn cells in IMA datasets (the generated 36M cell embeddings)
- Confirm the Norn identity using marker gene analysis, select top 13 kidney datasets
- Find preferential expression of Norn: Dcn, Lpar1, Colla1, Cxcll2, and Cfh.
- Epo: often missing from datasets and lowly expressed, not typically differentially expressed
- Cxcl14: another marker of Norn cell, mixed expression patterns
- Same pattern also found in cells from other tissues.
- The tissue with highest # of predicted Norn cells: gonad, heart, lung.



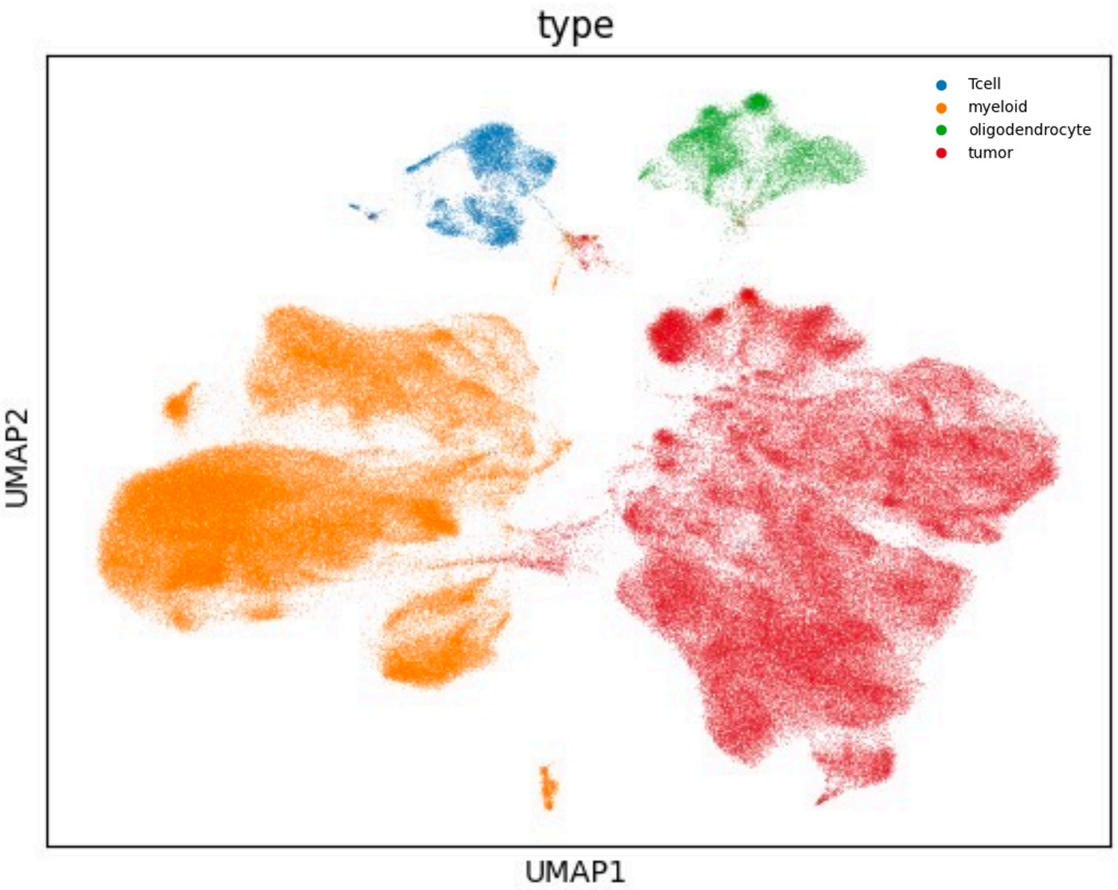
Previously observed

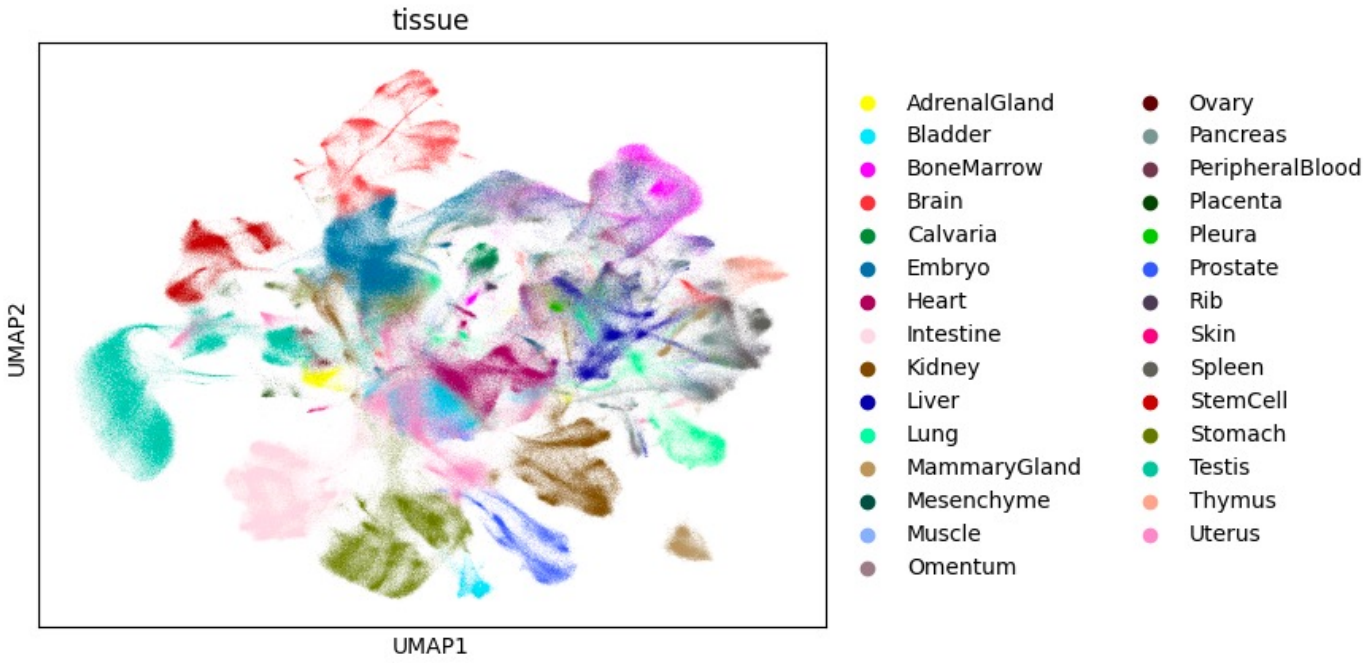


Results on Other Dataset

- Dr. ZHAO Zheng's dataset

- Mouse cell atlas dataset





Conclusion

- Easy to use.
- Don’t need to finetune, can handle multiple species.
- No downstream task demo.
- Slow speed and high demand of memories.
- When it comes to bulk, there are some bugs.

The tutorial will be available after Chinese New Year.

Thank you.