6AM

11AM in London (GMT), 8PM in Tokyo (GMT+9)

Multiscale Foundation Models

Moderator: Yashvardhan Jain, *Indiana University*

Presenters:

- Mo Chen, *Tsinghua University, China*
- Tianhong Zhou, *Tsinghua University, China*
- Yubo Zhao, *Tsinghua University, China*

清華大學
Tsinghua University

Multiscale Foundation Models

scFoundation, scMulan, ROAM and GeST: Powerful Bio Foundation Models

Tianhong Zhou, Mo Chen, Yubo Zhao

Bioinfo Division, BNRIST & Dept. of Automation, Tsinghua University

xglab@mail.tsinghua.edu.cn

Multiscale Foundation Models

scFoundation, scMulan, ROAM and GeST: Powerful Bio Foundation Models

Tianhong Zhou, Mo Chen, Yubo Zhao

Bioinfo Division, BNRIST & Dept. of Automation, Tsinghua University

xglab@mail.tsinghua.edu.cn

Our Team

- **Xuegong Zhang**

  Prof. of Bioinformatics and Pattern Recognition

  Dept. of Automation & Bioinfo Division, BNRIST

  Tsinghua University

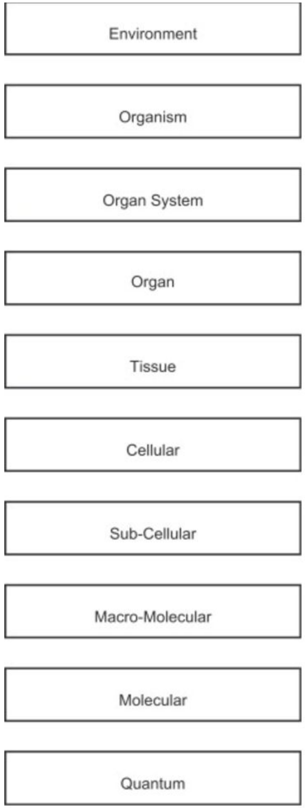
- **XGlab**

  25 members in the interdisciplinary area of life sciences and informatics

Challenges in biological research

Different biological scales work together

- **Molecular Scale**: Protein-protein interactions, gene mutations
- **Cellular Scale:** Signal transduction pathways, cellular dynamics
- **Organ/System Scale:** Integrated regulation in systems like nervous systems



Environment

Organism

Organ System

Organ

Tissue

Cellular

Sub-Cellular

Macro-Molecular

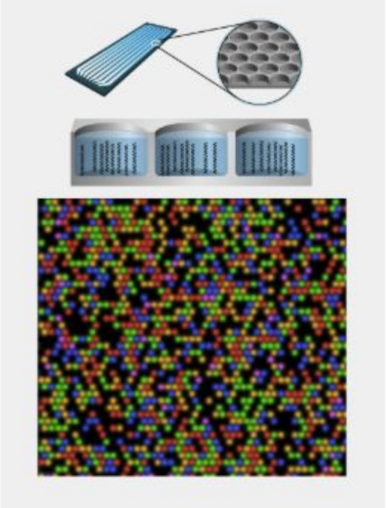
Molecular

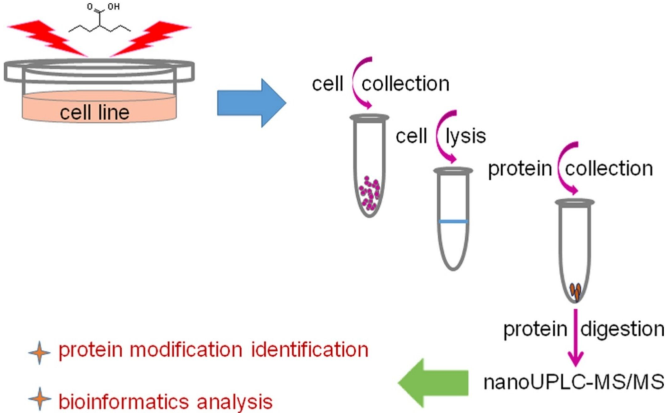
Quantum

Challenges in biological research

Modern bioinformatics technologies produce explosion multi-modal data

- sequencing
- mass spectrometry
- Medical image





cell line

cell collection

cell lysis

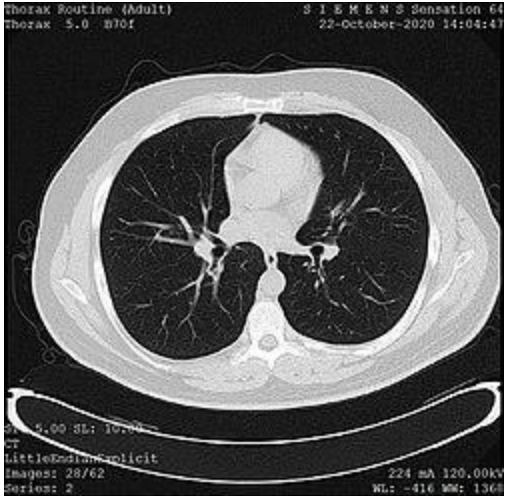
protein collection

protein digestion

nanoUPLC-MS/MS

protein modification identification

bioinformatics analysis



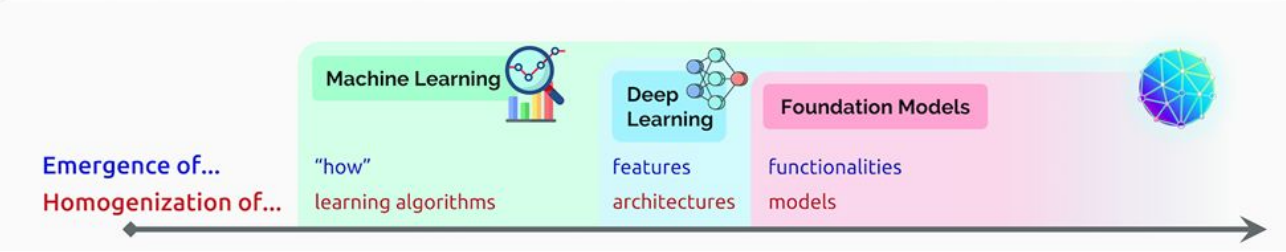
We need efficient model

Limitations of Traditional Methods:

- Rely heavily on expert knowledge, hard to manage large-scale data.
- Statistical models capture only local relationships

We need tools that can handle multi-scale, multi-modal data comprehensively and efficiently.

What are foundation models?



Foundation models homogenizes the model itself.

Foundation models are enabled by transfer learning and scale.

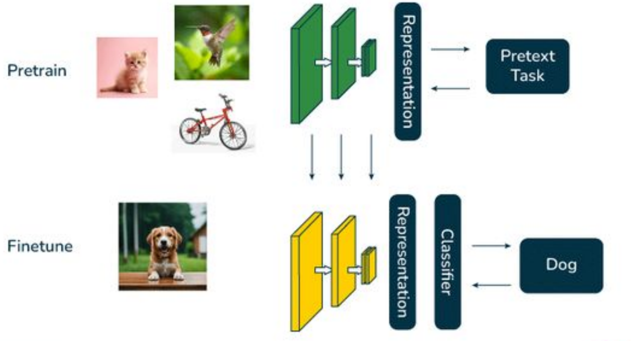
A model is trained on a surrogate task (often just as a means to an end) and then adapted to the downstream task of interest via fine-tuning.

Why foundation models can work?

Self-supervised learning: pretraining task is derived automatically from unannotated data.

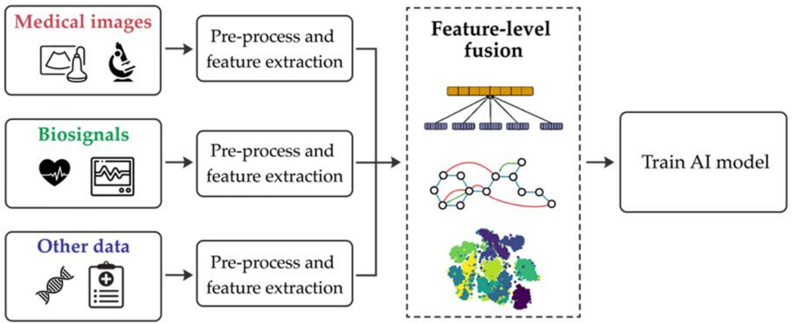
Reduce reliance on domain-specific tweaks, offering a more generalized solution.

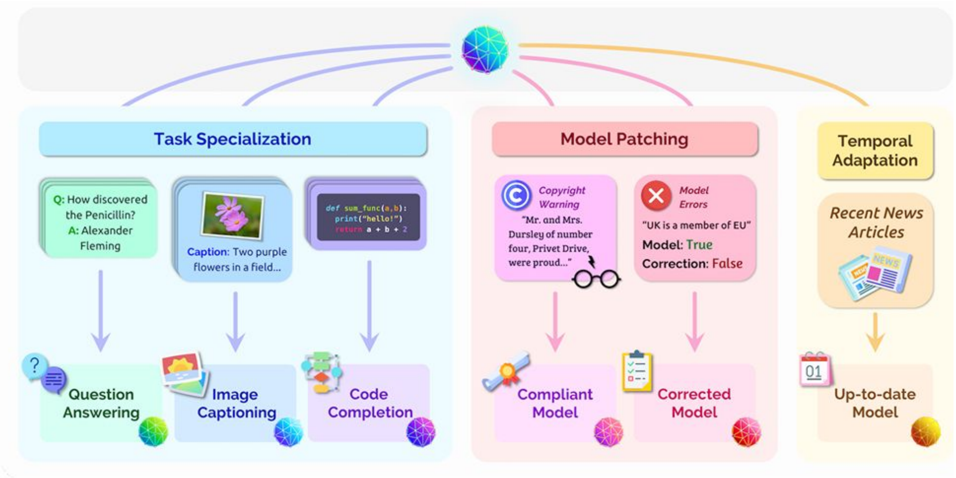
During fine-tuning, the pretrained model is exposed to a smaller, task-specific dataset with labeled examples, allowing it to adapt its general knowledge to particular requirements.



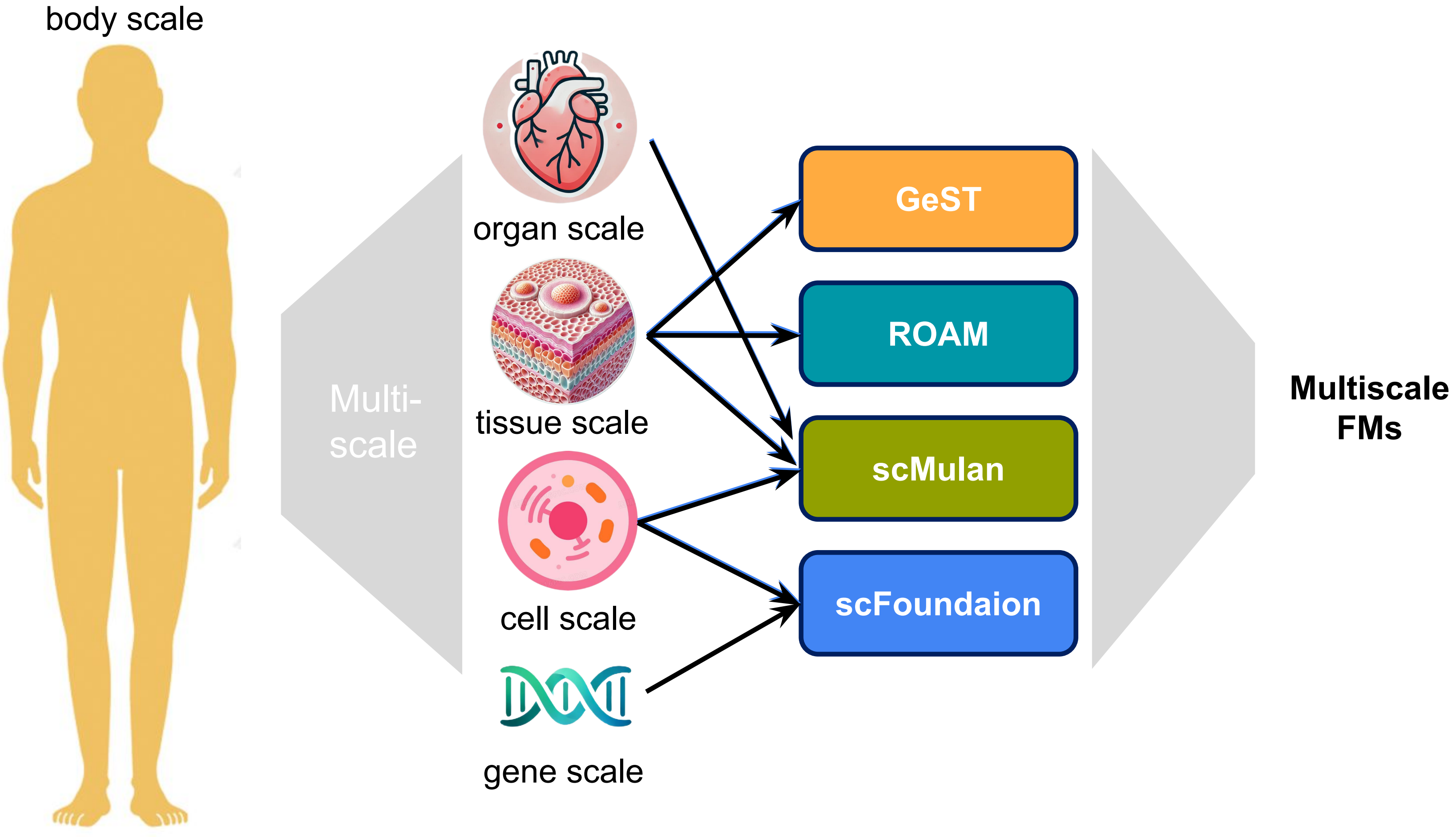
Why foundation models can work?

Multi-Modal Integration: Combine genetic, imaging, and textual data into one framework





How FMs Work in Multiscale Biological Scenarios



body scale

Multi-scale

organ scale

tissue scale

cell scale

gene scale

GeST

ROAM

scMulan

scFoundaion

Multiscale FMs

scFoundation: FM on single-cell transcriptomics

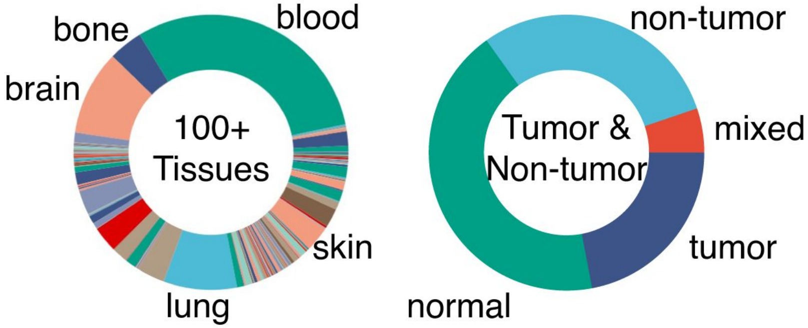
50 million multi-scale single cell data

Gene Expression Omnibus

HUMAN CELL ATLAS

EMBL-EBI

……



xTrimoGene model

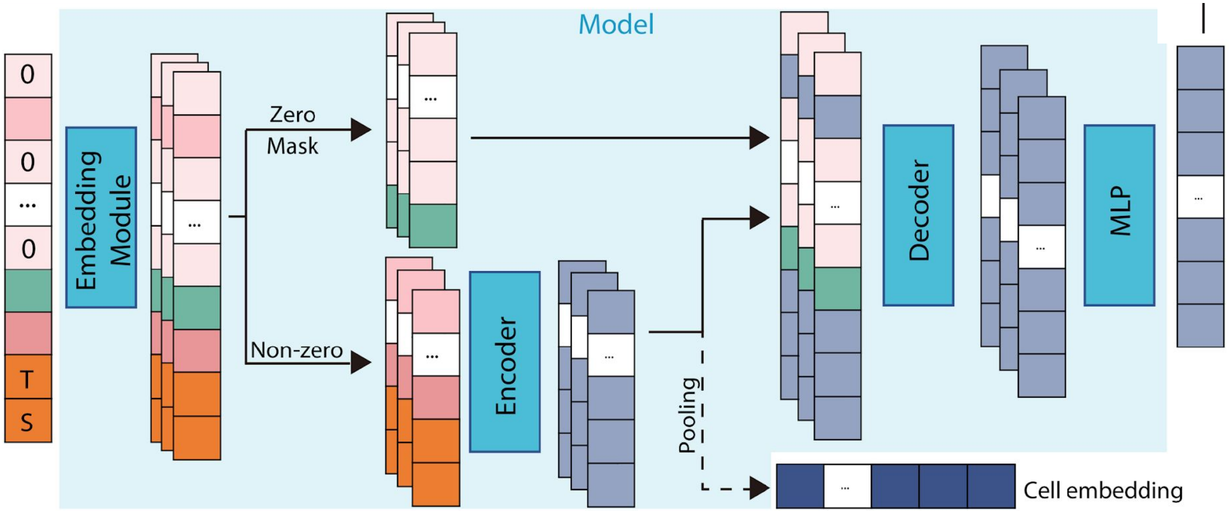
~100 million parameters

embedding module

converted gene expression into learnable high-dimensional vectors

asymmetric architecture

reduce the computational and memory challenges

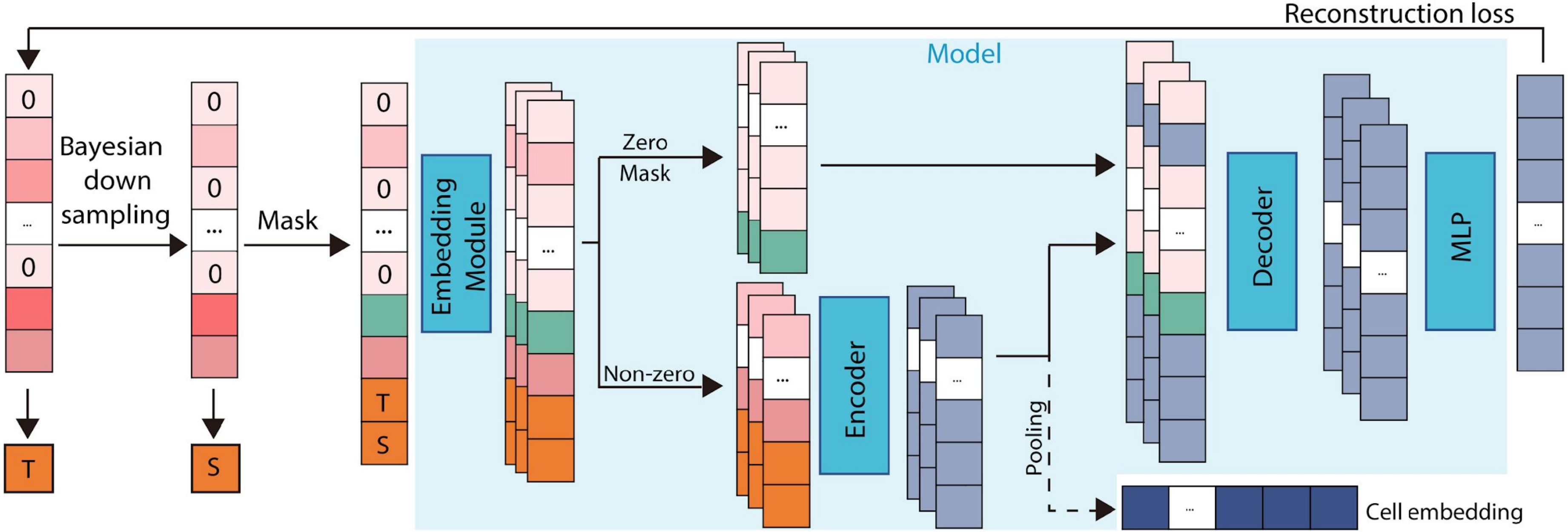


scFoundation: FM on single-cell transcriptomics

Training task: Sequencing Depth-Aware Expression Recovery

Input a gene expression vector sequenced at depth S to recover a target gene expression vector at depth $T (S \leq T)$

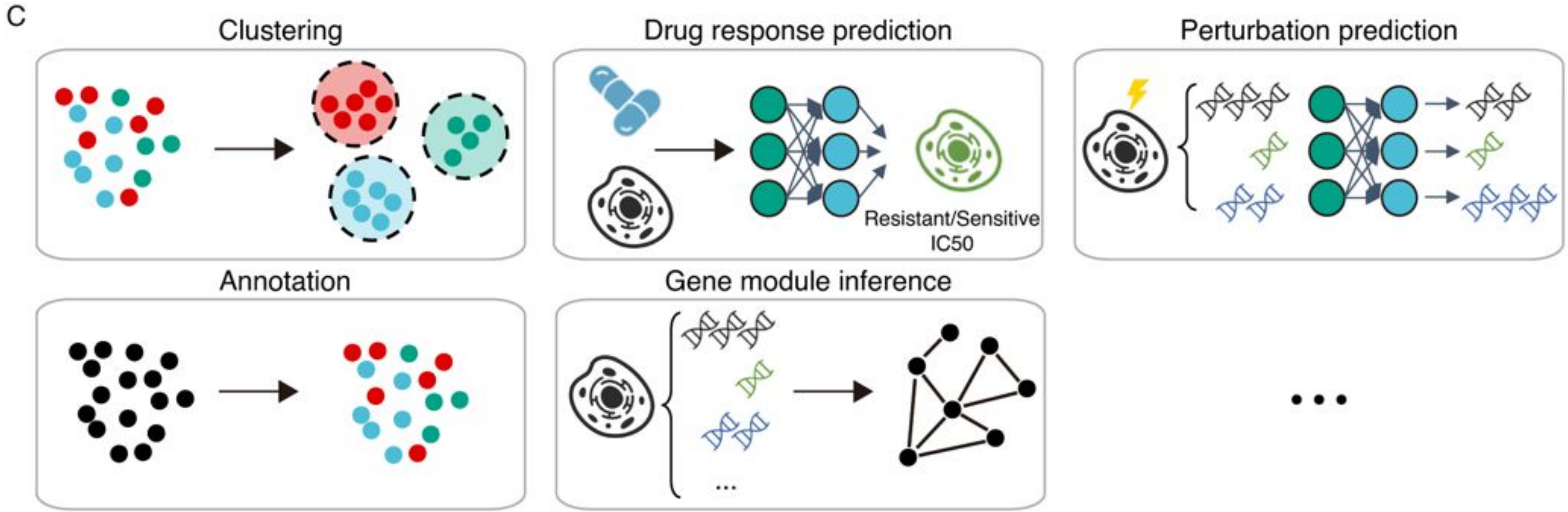
- Target vector: The **original** gene expression vector.
- Input vector: A vector obtained by **Bayesian sampling** from the original expression values.



scFoundation: FM on single-cell transcriptomics

Application format:

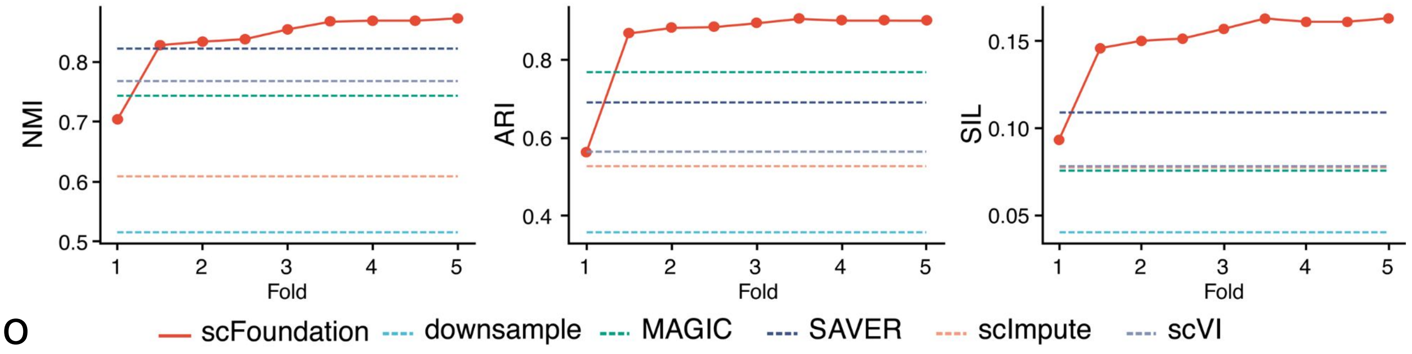
- Cell Representation(Sentence)
- Gene Representation(Word)

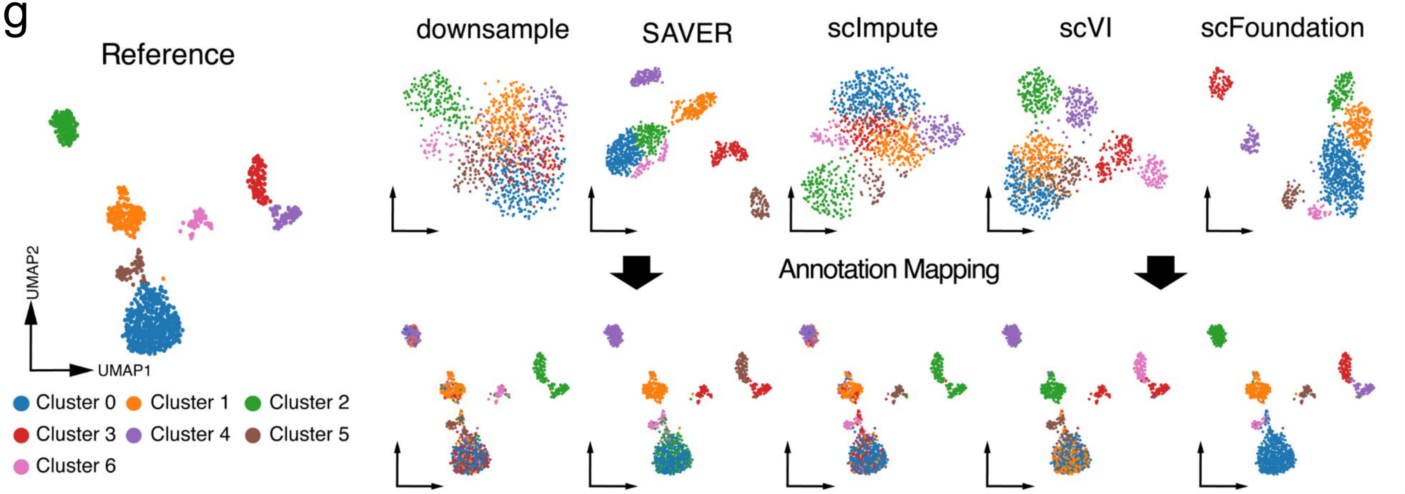


scFoundation: FM on single-cell transcriptomics

Better cell representation

Increasing the T:S ratio improves the clustering performance of cell embeddings



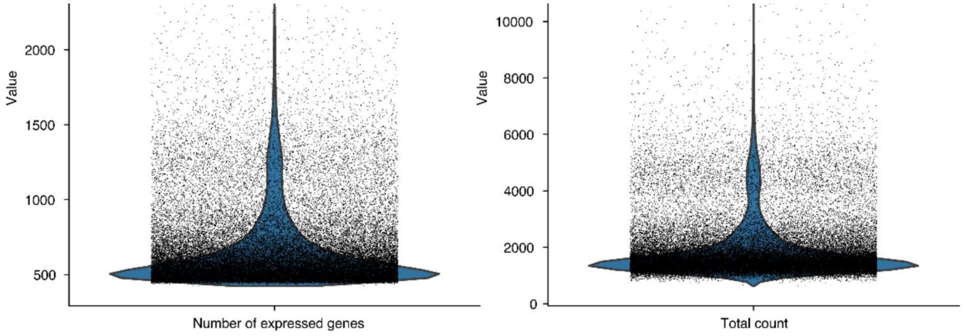


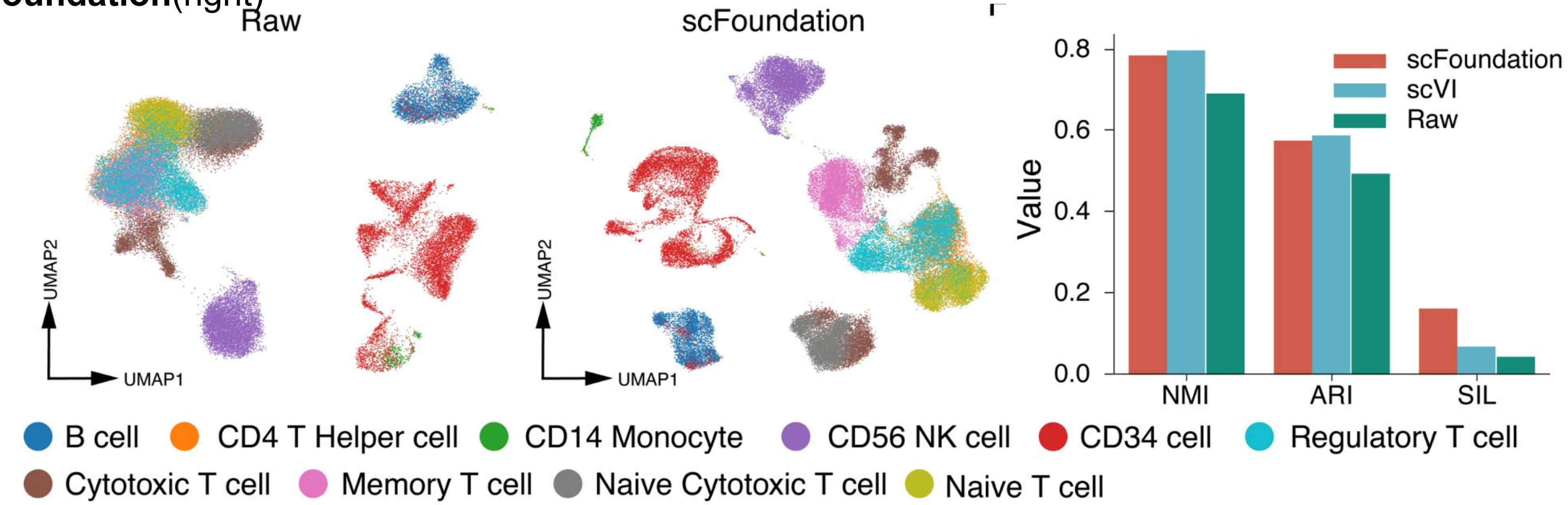
scFoundation: FM on single-cell transcriptomics

Sequencing Depth Enhancement:

Zheng68K (left) with low sequencing depth

Sequencing depth enhanced by **scFoundation**(right)

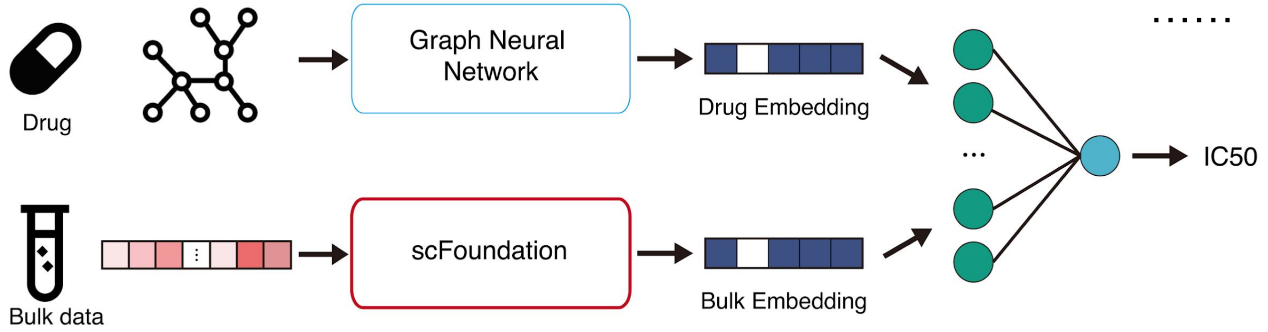




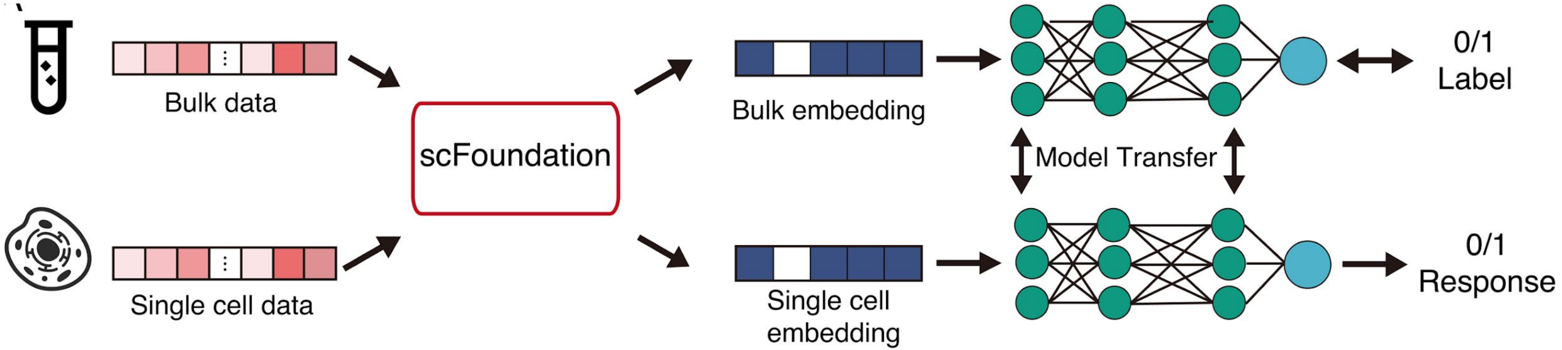
scFoundation: FM on single-cell transcriptomics

Various downstream tasks

- Drug-Cancer Effective Concentration Prediction



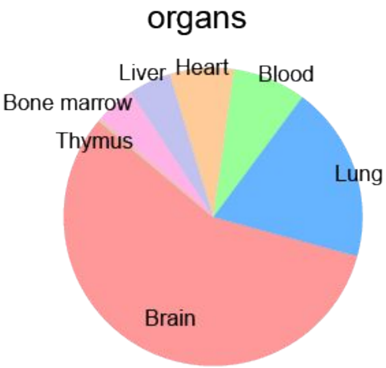
- Drug-Sensitive/Non-Sensitive Single-Cell Classification



- Cellular Perturbation Prediction
- Gene Network Inference

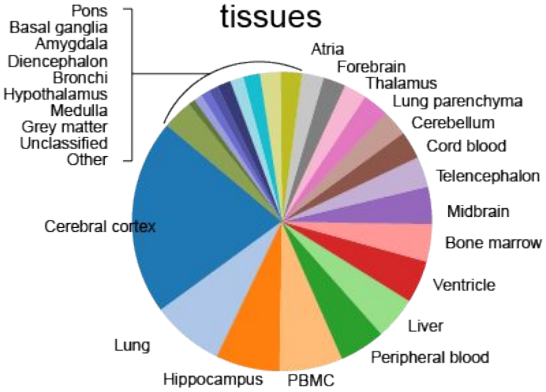
……

scMulan: A Multitask Generative FM for Sc-analysis

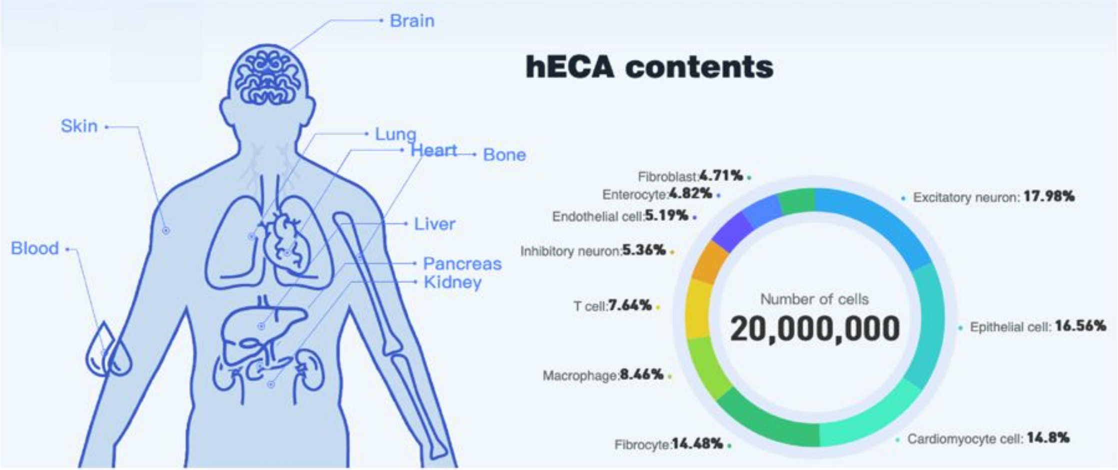


Training data

- 10M multiscale single-cell data
- 20K genes
- ~ 160G natural language corpus

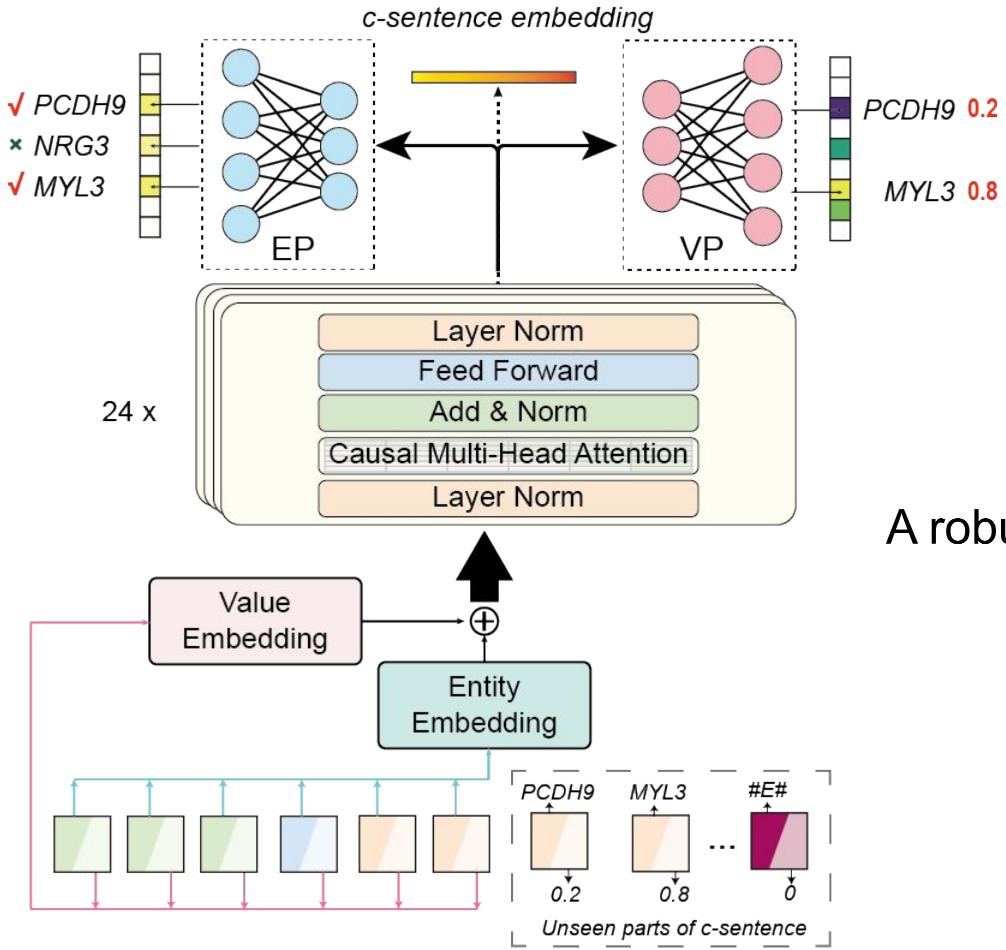


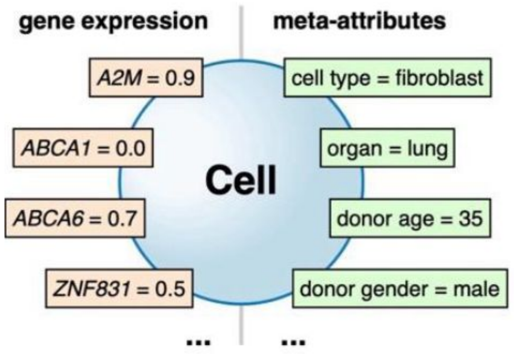
hECA-10M-version1



http://eca.xglab.tech/

scMulan: A Multitask Generative FM for Sc-analysis





A robust model: 368 M parameters

- c-sentence transferring
- cell language modeling
- multiple pre-training tasks

scMulan: A Multitask Generative FM for Sc-analysis

Multiple pre-training tasks

Task: Cell generation

Input: ***Heart, Cardiomyocyte cell, <CG>***

Output: ***MYL2, CCDC3, NEGR1, PCDH9, MYL3 …***

Task: Cell type annotation

Input: ***Heart, PCDH9, MYL3, MYL2, NEGR1, CCDC3 …, <PCT>***

Output: ***Endothelial cell, Vascular endothelial cell***

Task: Organ region prediction

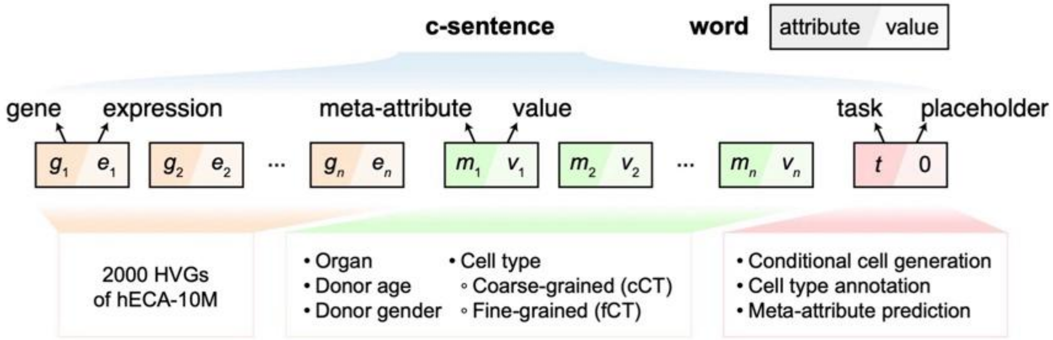
Input: ***PCDH9, MYL3, MYL2, GJR5, CCDC3 …, <POR>***

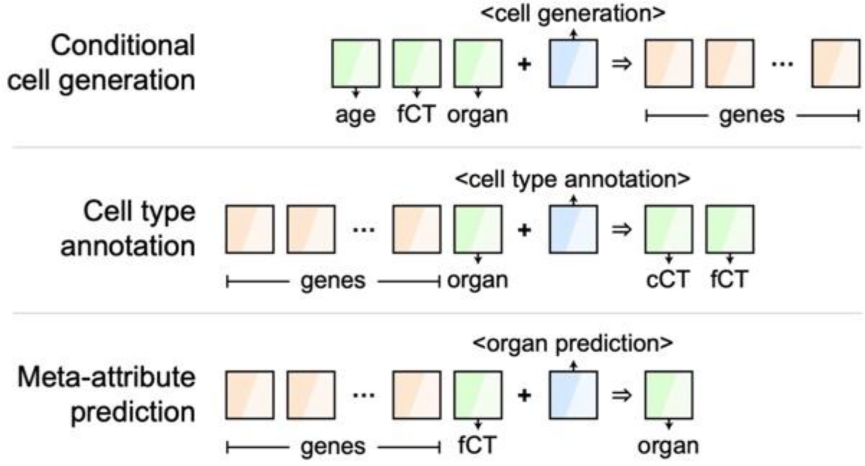
Output: ***Heart, Atria***

Task: Time series generation

Input: ***Bone marrow, HSC, GATA1, NEGR1, CD3D … , <NS>***

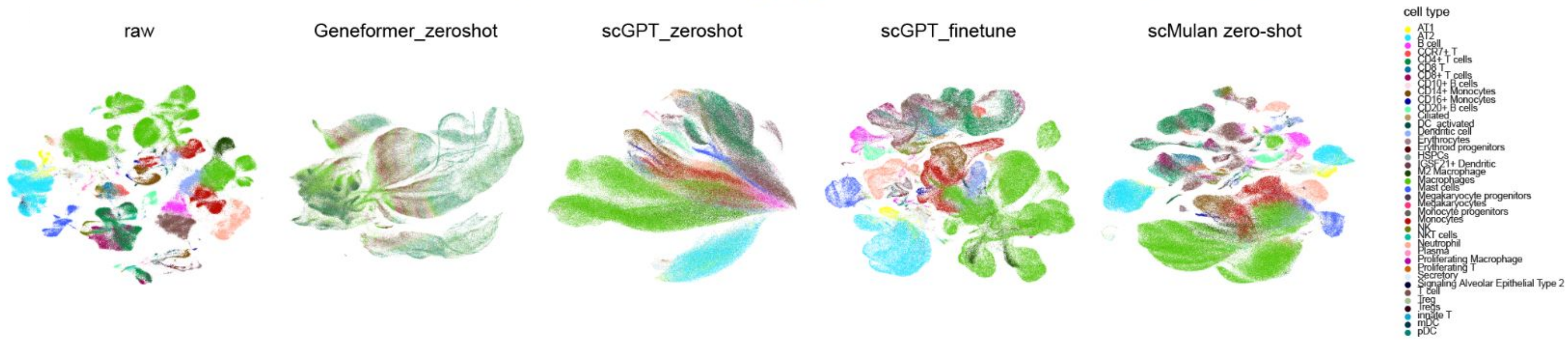
Output***: Bone marrow, HSC, GATA1, GATA2, NEGR1 …***

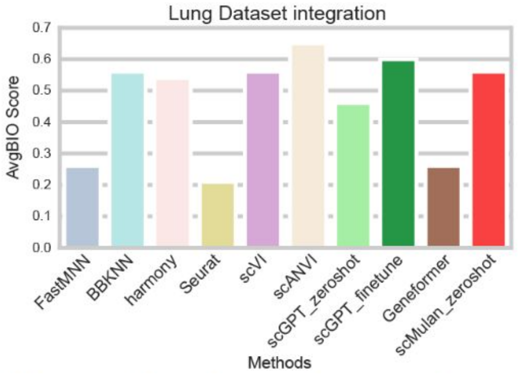


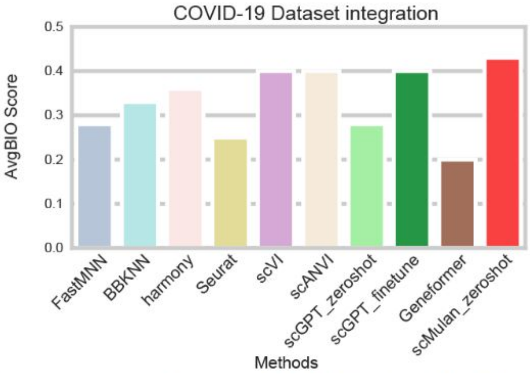


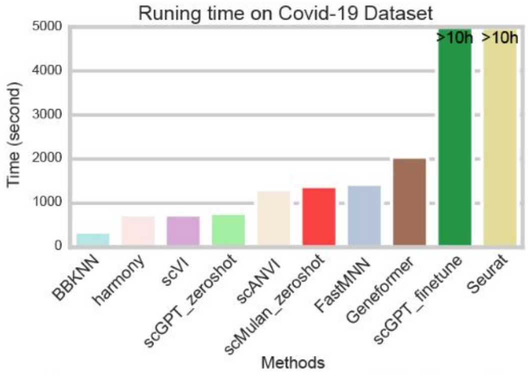
scMulan: Zero-shot integration

scMulan enables fast and accurate integration across multiple batches without fine-tuning

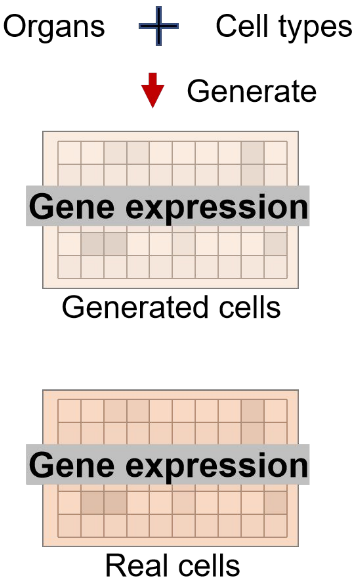








scMulan: Conditional generation



Organs + Cell types

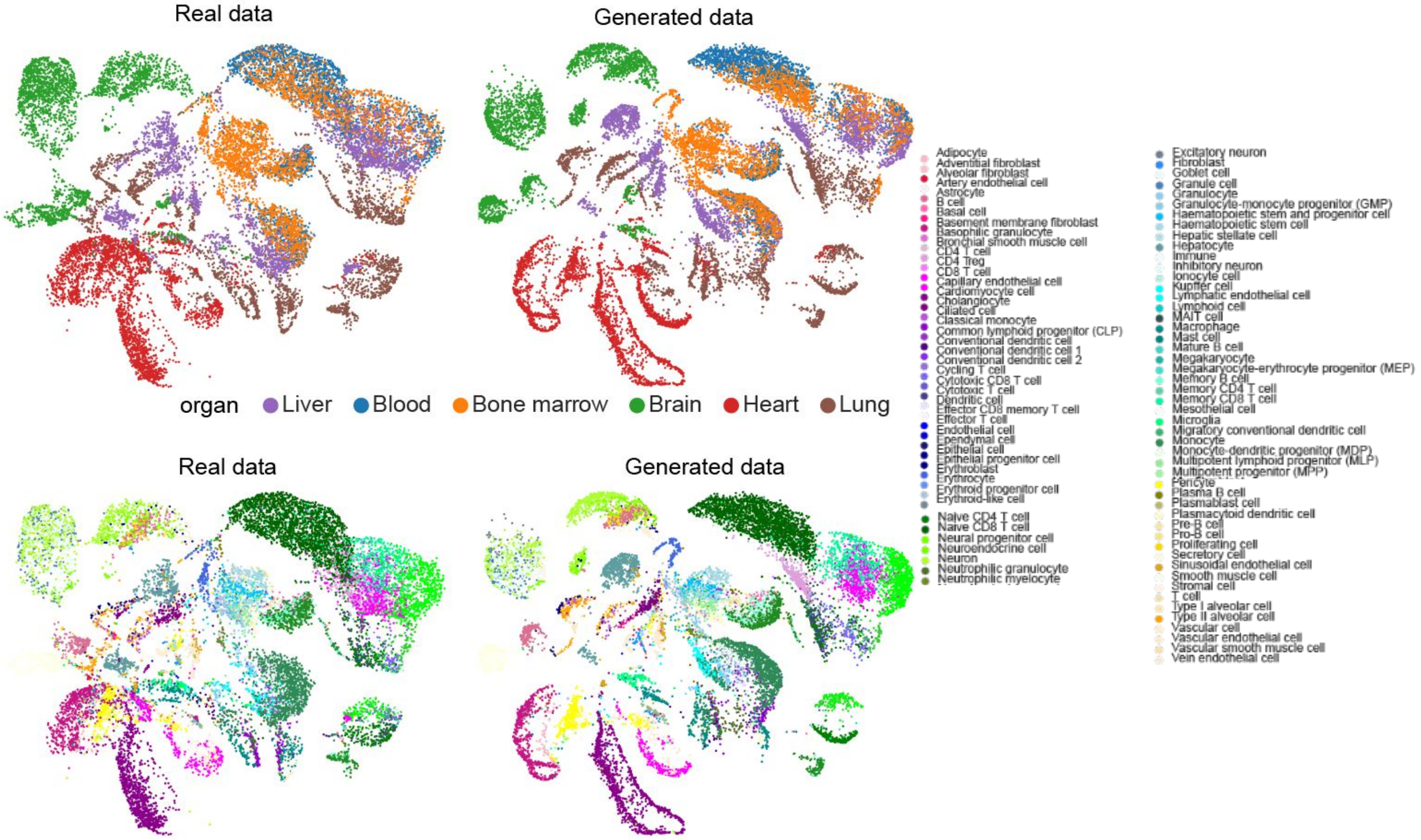
Generate

Gene expression

Generated cells

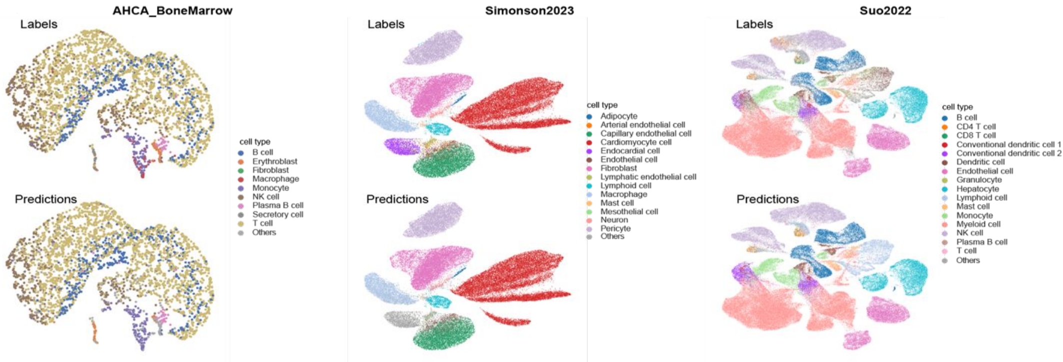
Gene expression

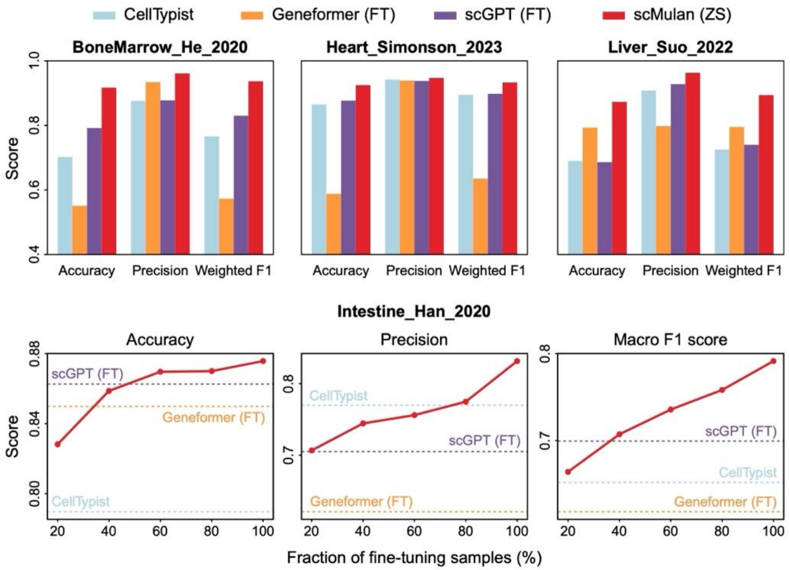
Real cells



scMulan: Zero-shot cell type annotation on specific organ

- **Outperforms SOTA in Included Organs**
- **Shows Significant Improvement After Fine-Tuning on Unseen Organs**



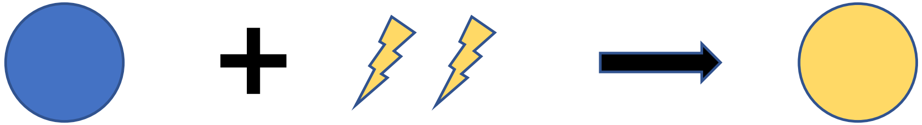


	AHCA_BoneMarrow			Simonson2023			Suo2022		
	Acc.	Prec.	F1.	Acc.	Prec.	F1.	Acc.	Prec.	F1.
scMulan	**0.917**	**0.961**	**0.937**	**0.927**	**0.947**	**0.934**	**0.873**	**0.963**	**0.894**
CellTypist	0.600	0.859	0.665	0.865	0.942	0.895	0.344	0.909	0.420
scGPT	0.686	0.928	0.740	0.877	0.938	0.898	0.792	0.878	0.830
Geneformer	0.793	0.798	0.796	0.588	0.939	0.635	0.742	0.911	0.760

scMulan: In silico perturbation on genes

Prompt: cell type A + <CG> + (PGene A, expression level) + (PGene B, expression level)

Output: cell type A + <CG> + (PGene A, expression level) + (PGene B, expression level) +
(Gene X, expression level) + (Gene Y, expression level) +, …, + (Gene …, expression level) + <End>

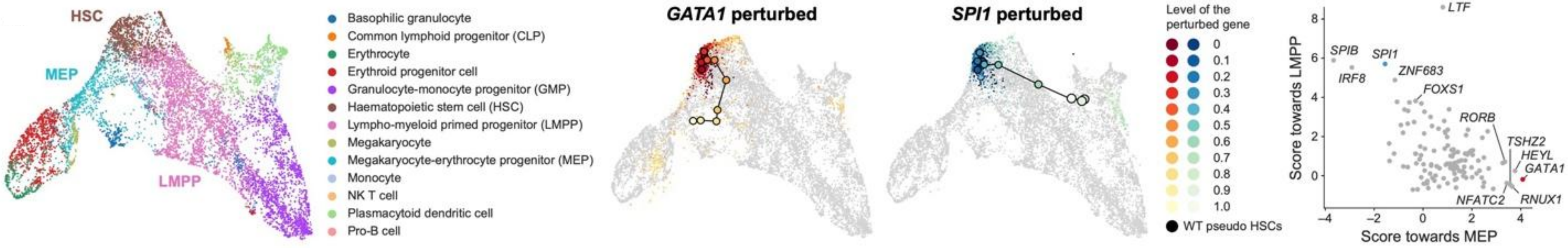


cell type A

Perturbed Genes

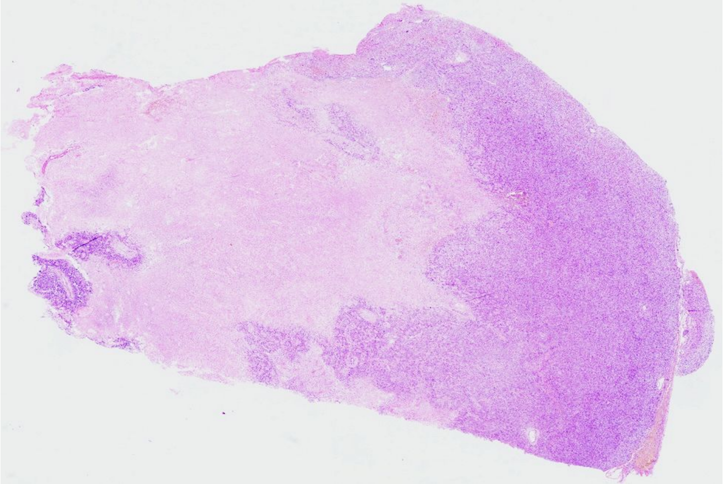
Expression
after perturb

Over-express GATA1 and SPI1 gene

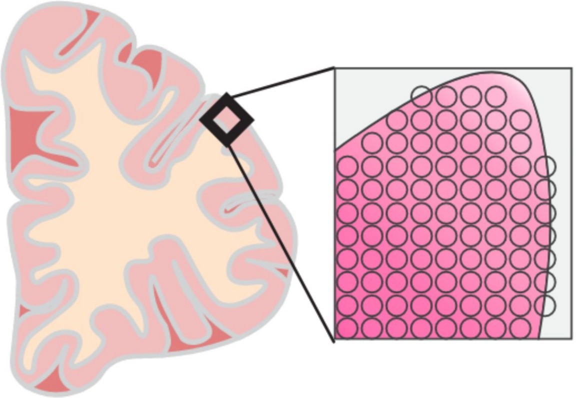


In-situ Tissue Level Modeling

- **Image**

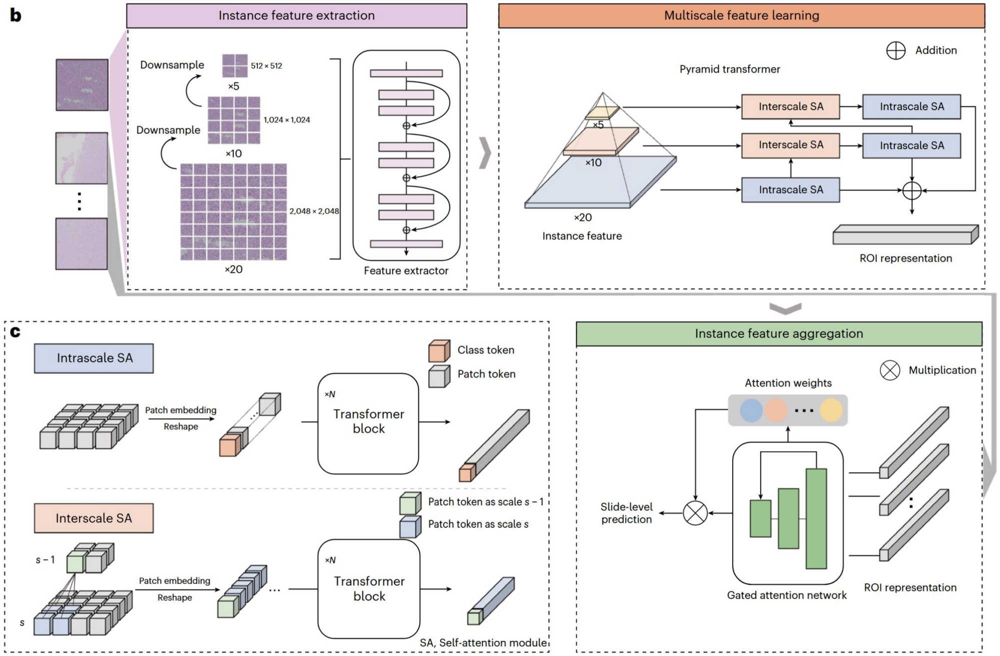


- **Spatial Transcriptomics**



ROAM: Image Foundation Model for Glioma Diagnosis

- **Transformer-based model**
  - **Image as a sentence of patches**
- **Large-scale pretraining**
  - **1109 whole slide images (WSIs)**
- **Multitasking**
  - **Tumor diagnosis**
  - **Molecular status prediction**
  - **…**

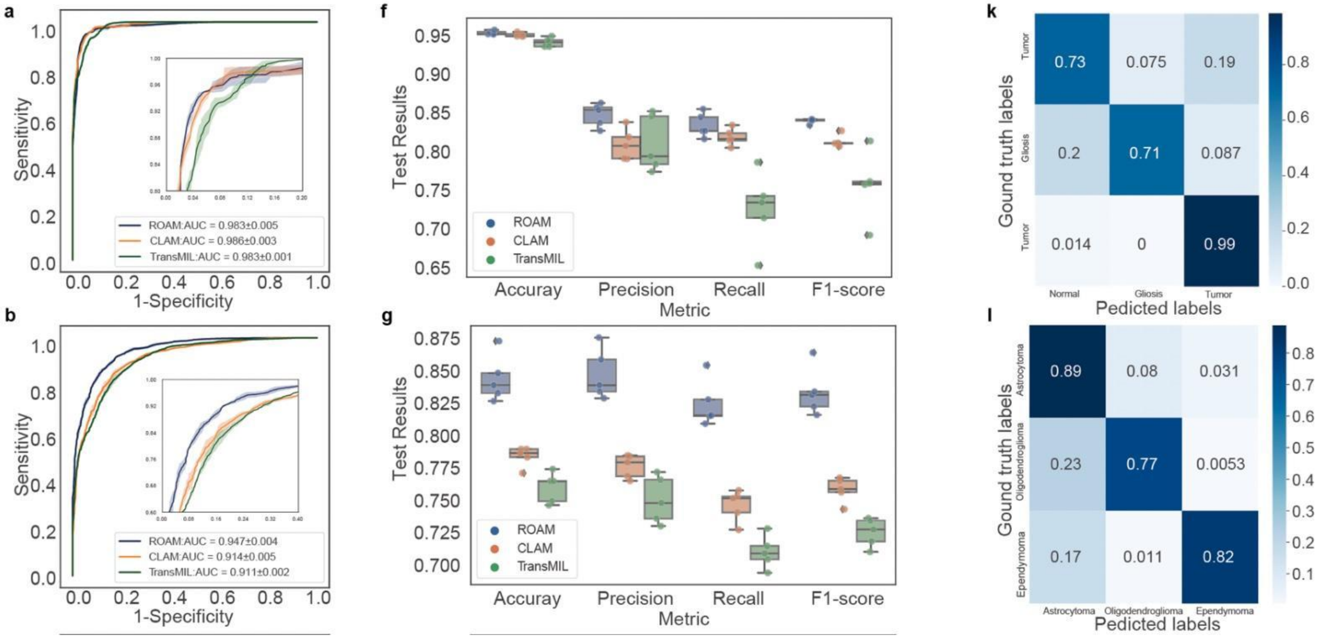


Jiang et al., *Nature Machine Intelligence*, 2024

ROAM accurately diagnoses gliomas

Glioma detection

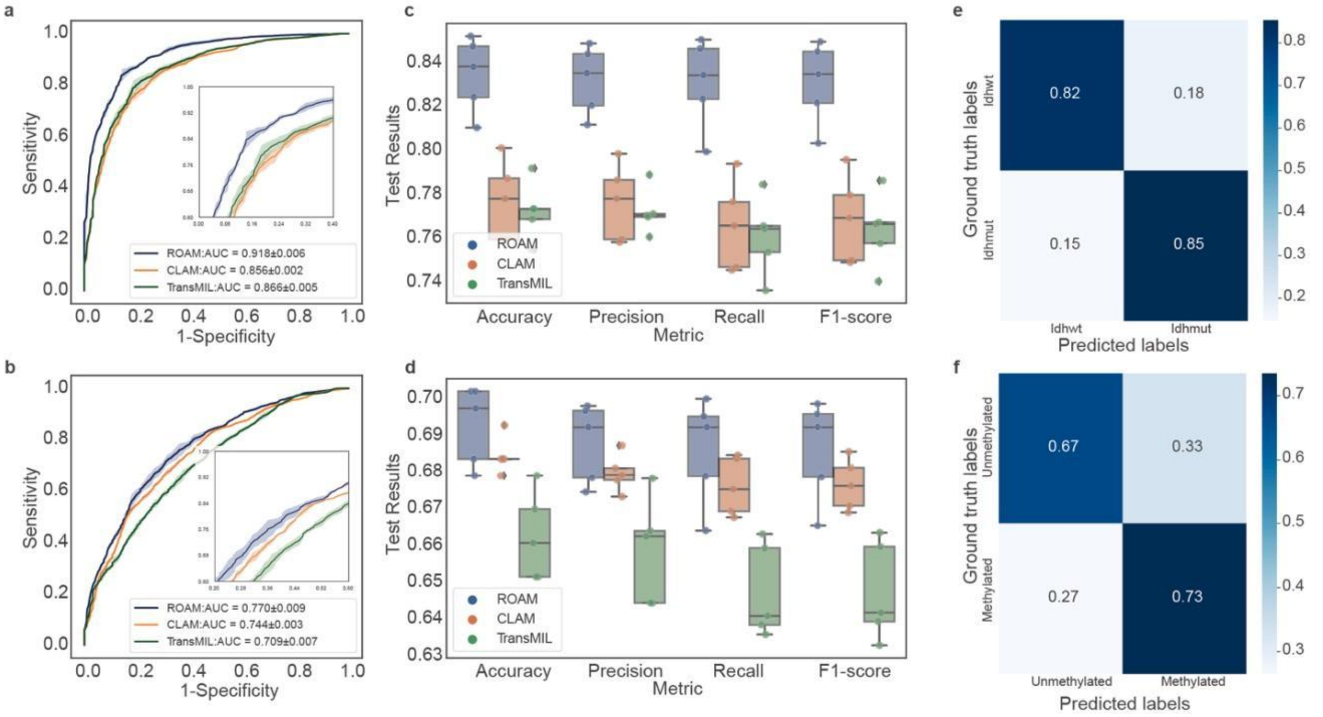
Glioma subtyping



ROAM predicts molecular status

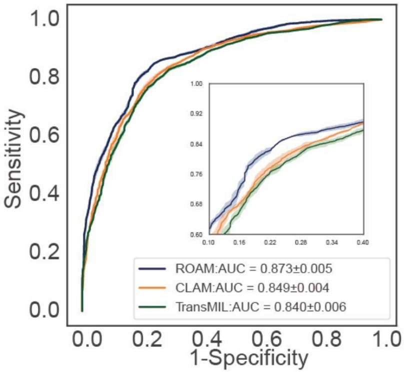
IDH

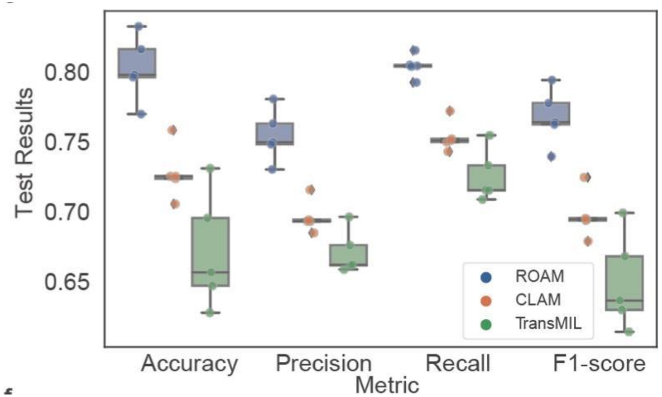
MGMT



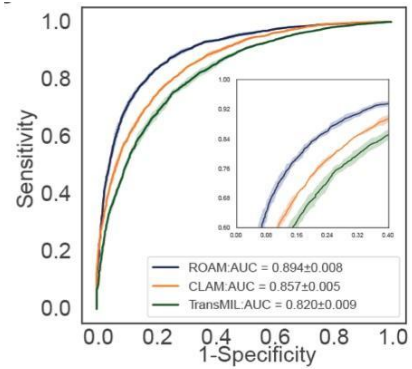
ROAM generalizes to external independent test datasets

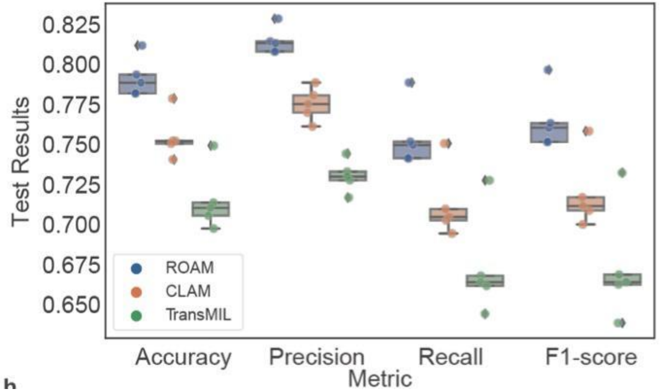
Glioma subtyping



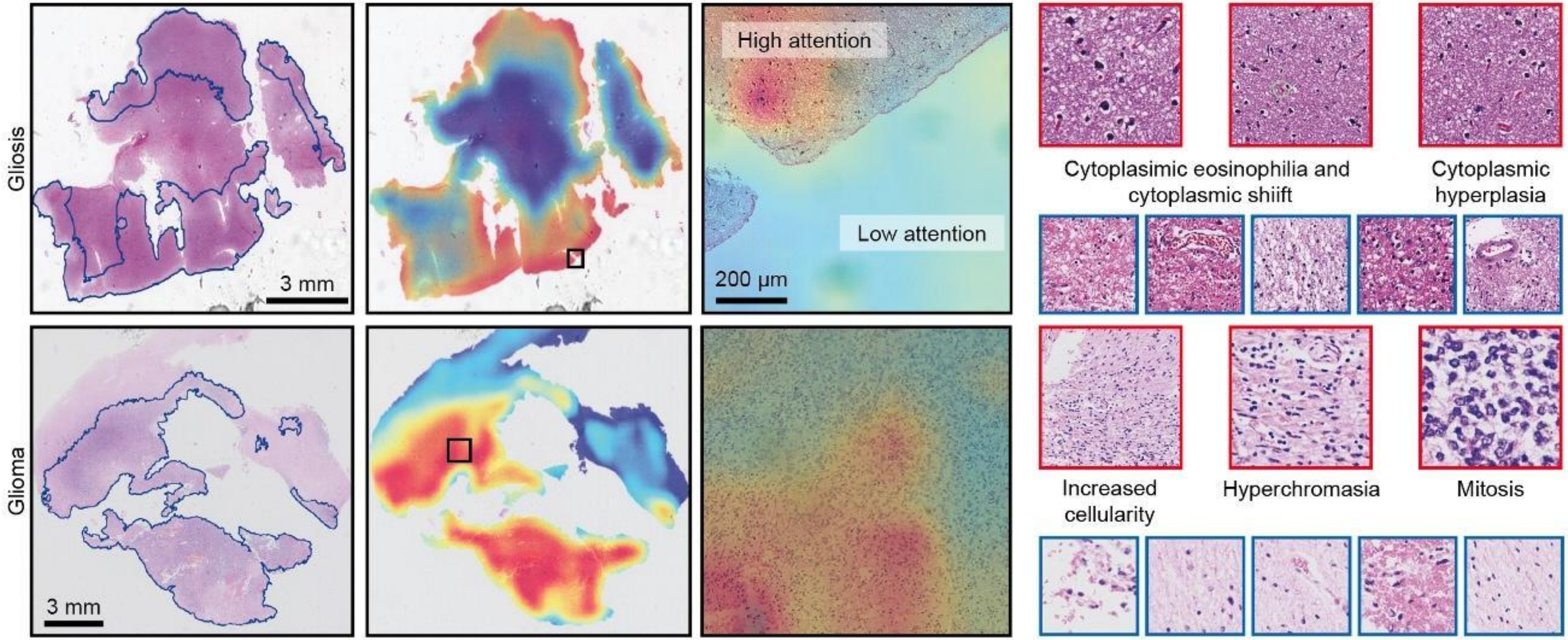


IDH status prediction



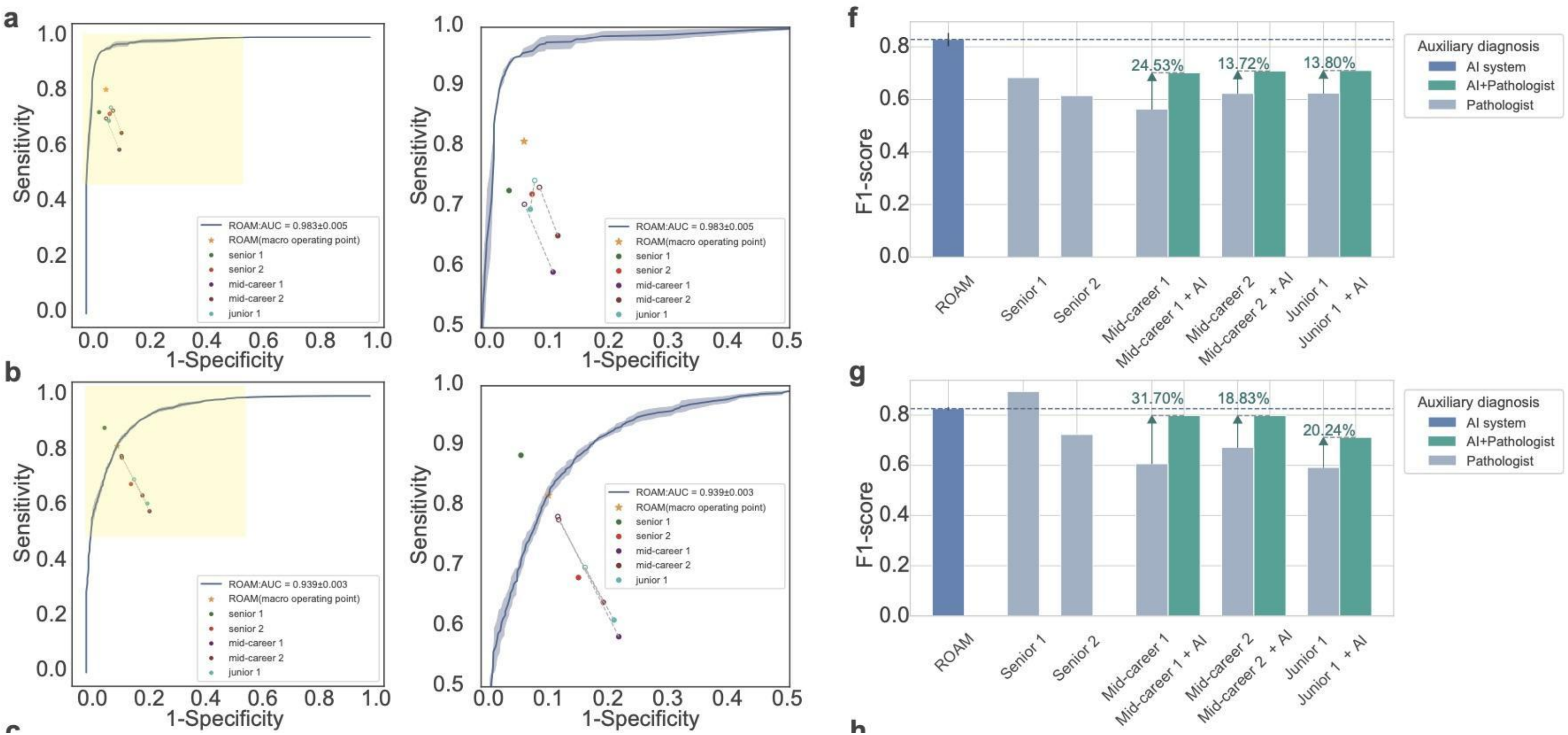


ROAM’s results are well interpretable

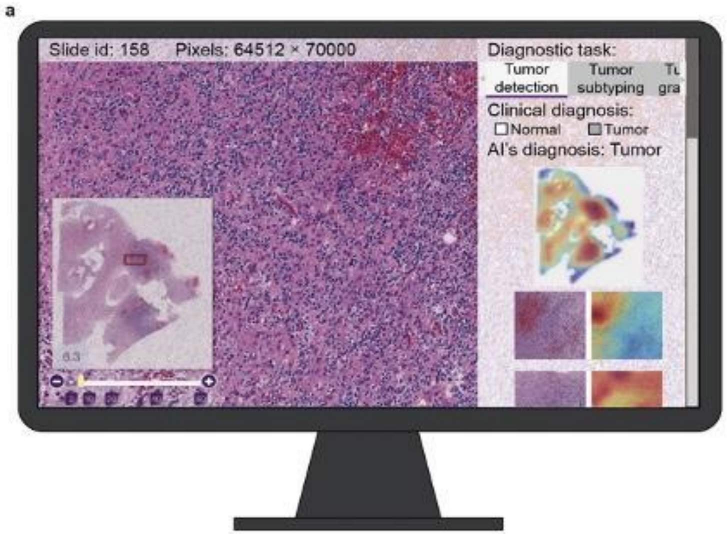


Slide-level visualization ROI-level visualization

ROAM serves as auxiliary diagnosis



ROAM discovers molecular-morphological biomarkers



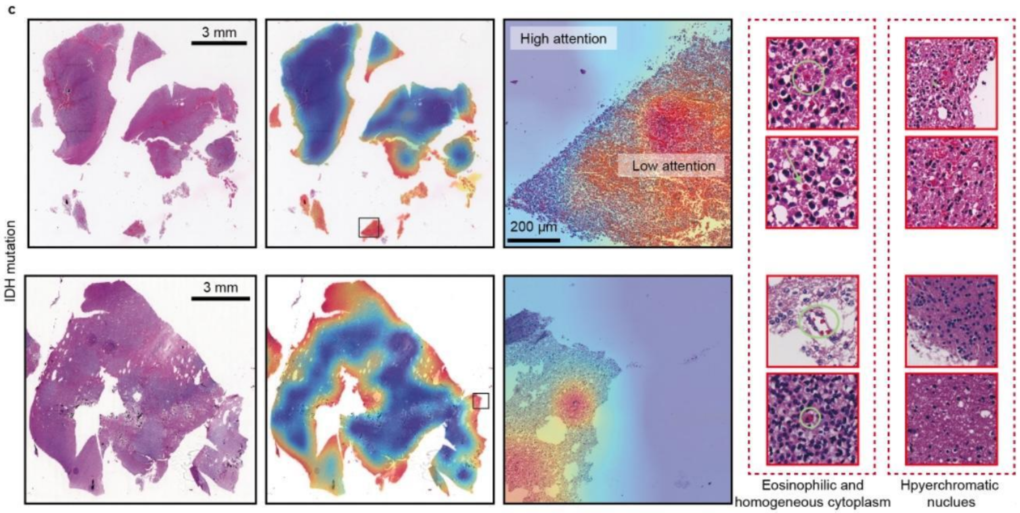
Slide id: 158 Pixels: 64512 × 70000

Diagnostic task:

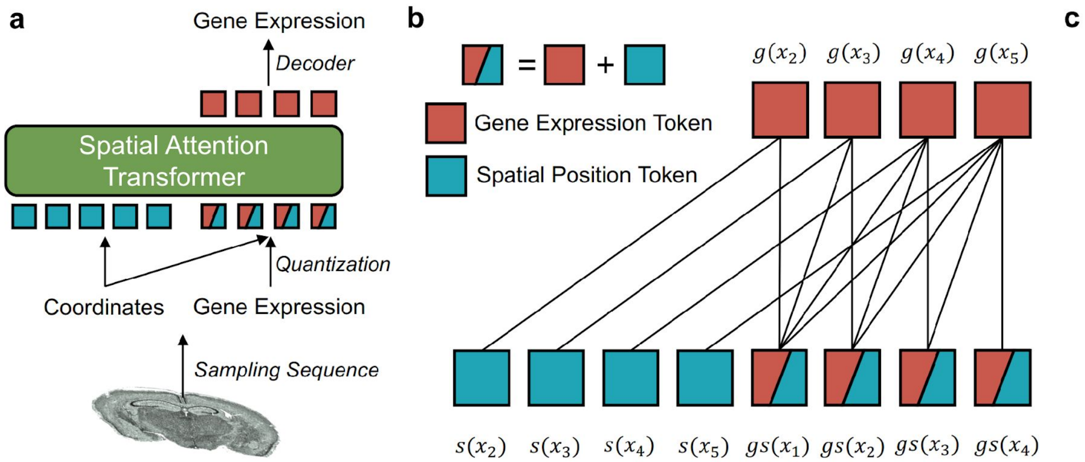
Clinical diagnosis:

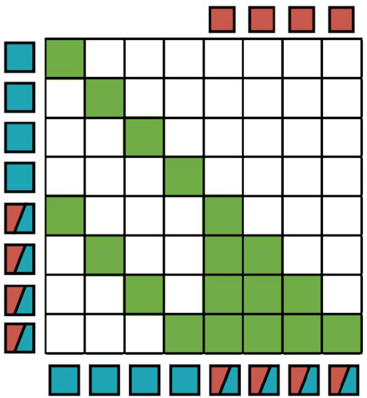
☐Normal ☒Tumor

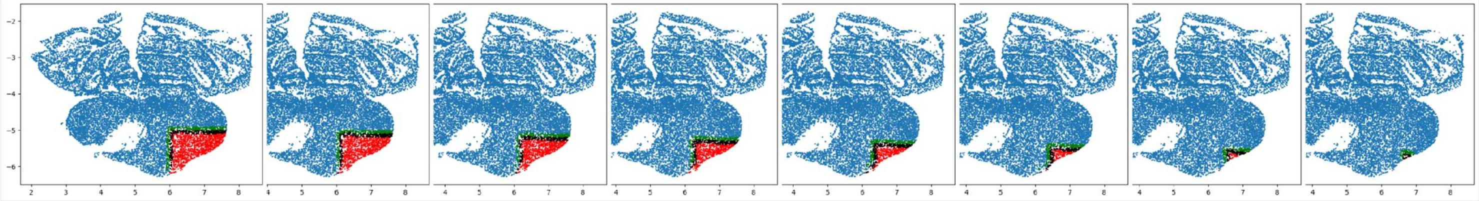
AI's diagnosis: Tumor



GeST: Generate Spatial Transcriptomics like GPT!



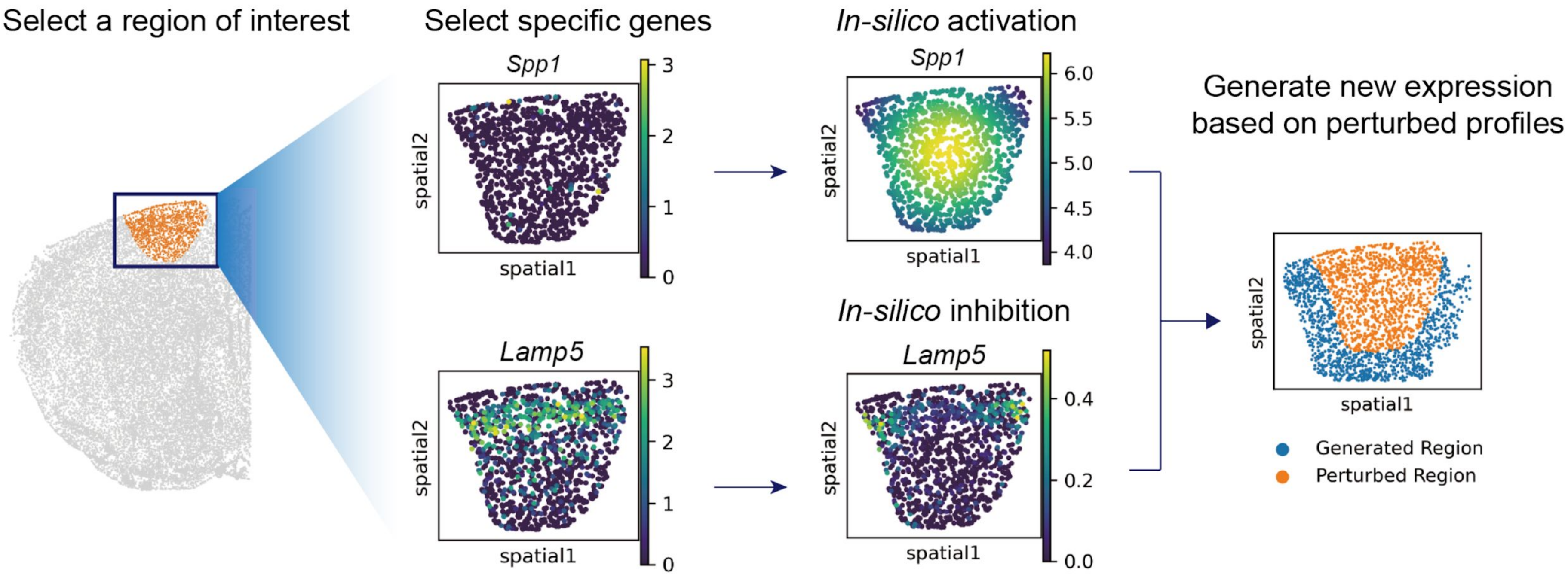




Auto-regressive gene expression generation

Hao et al., Under Review

GeST predicts spatial perturbation


Select a region of interest

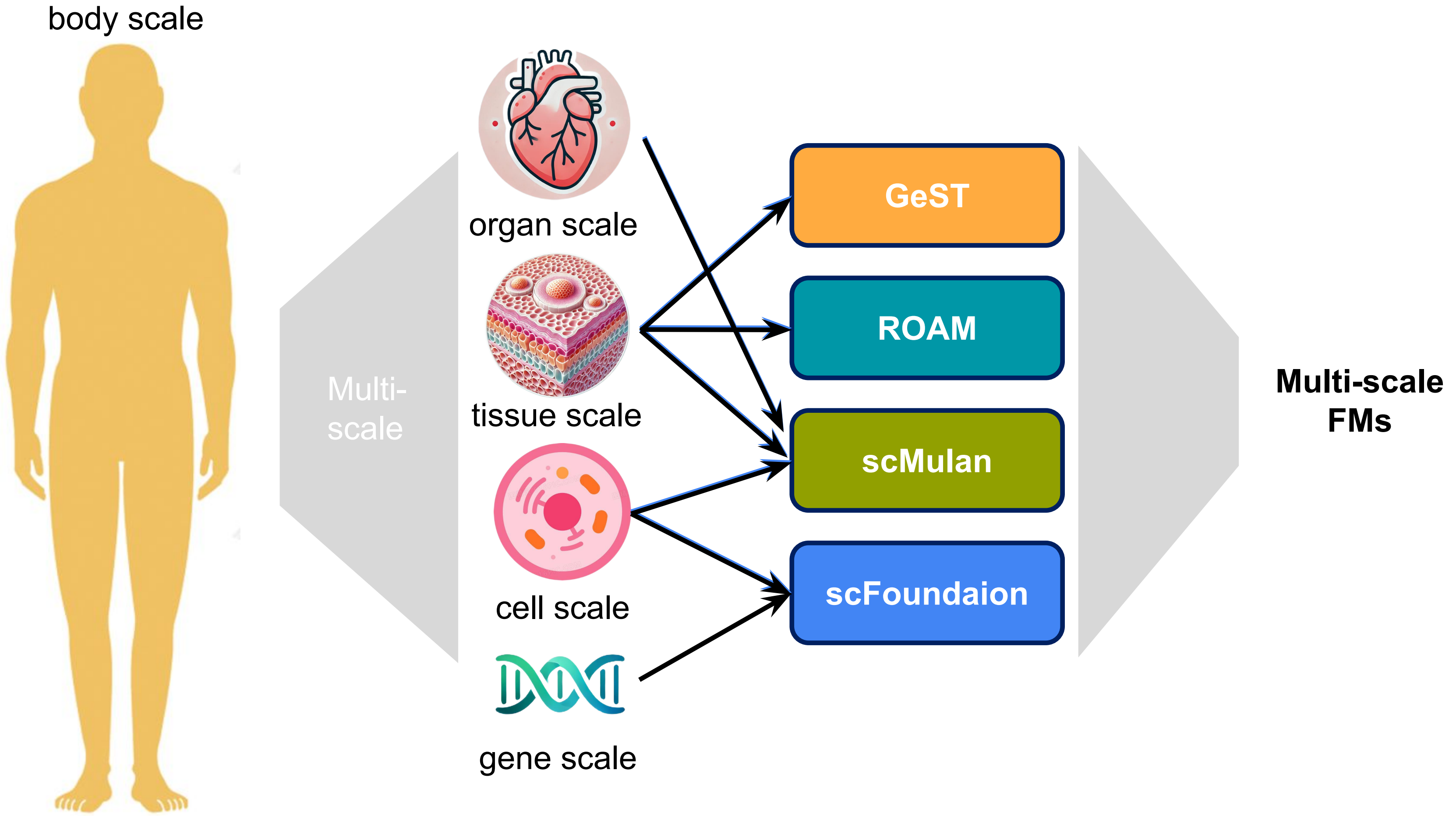
Select specific genes

In-silico activation

In-silico inhibition

Generate new expression based on perturbed profiles

Summary: Present and Future of Multi-scale FMs



body scale

Multi-scale

organ scale

tissue scale

cell scale

gene scale

GeST

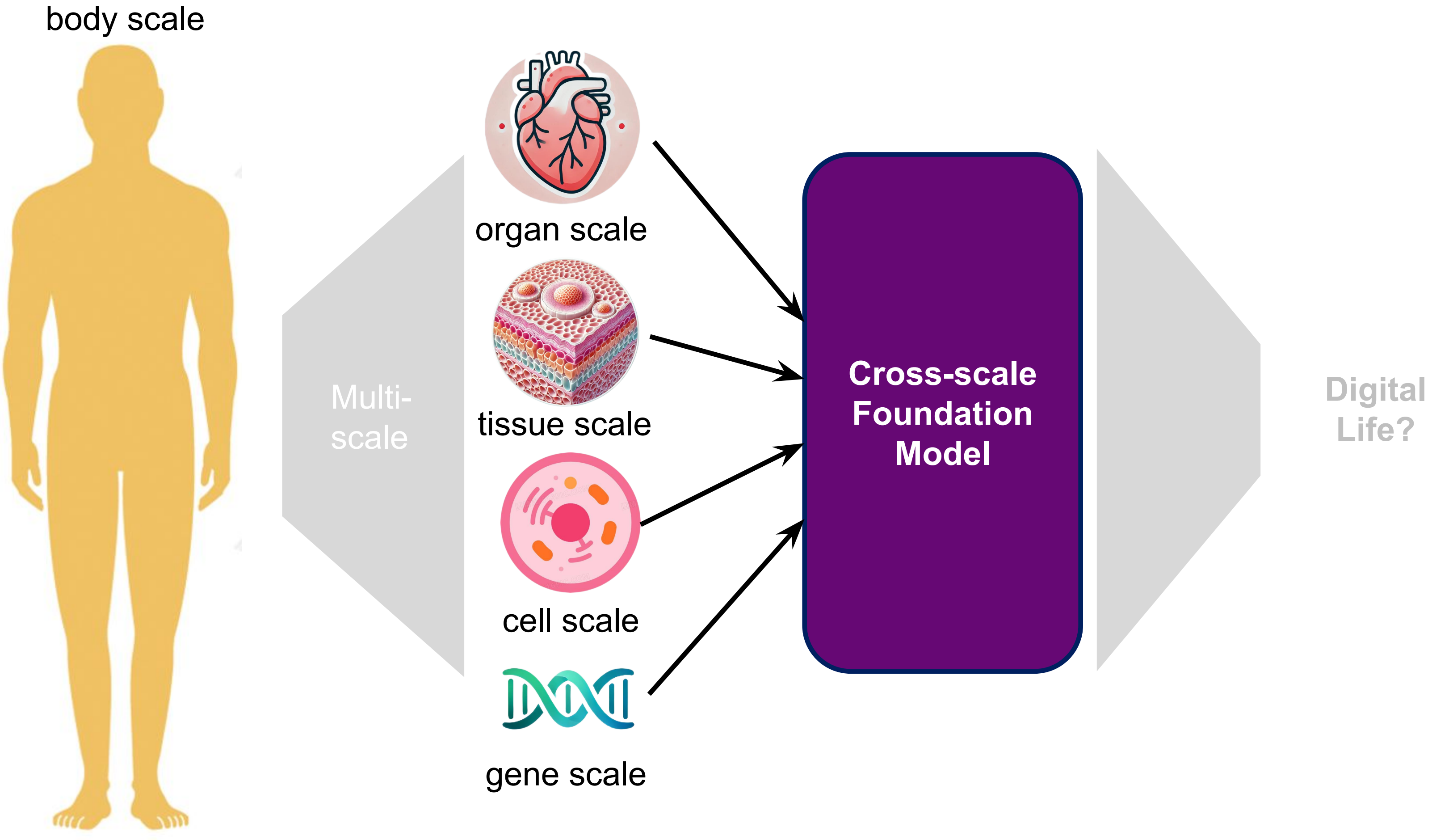
ROAM

scMulan

scFoundaion

Multi-scale FMs

Summary: Present and Future of Multi-scale FMs



body scale

Multi-scale

organ scale

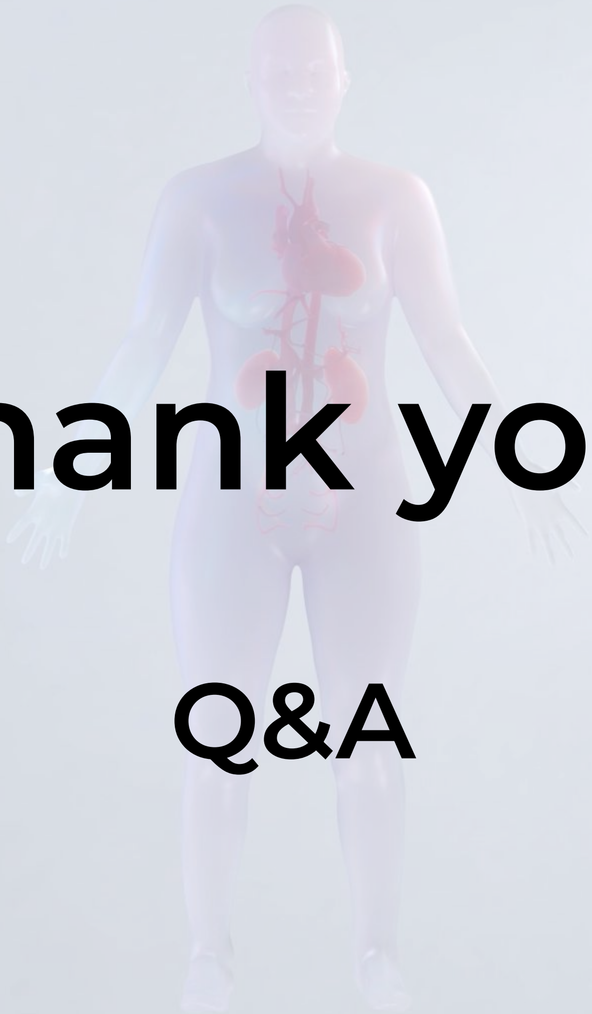
tissue scale

cell scale

gene scale

Cross-scale Foundation Model

Digital Life?



Thank you!

Q&A

Questions

How do we define a Multiscale Human?

How do we map a Multiscale Human?

How do we model a Multiscale Human?

How can LLMs or RAGs be used to advance science and clinical practice?

Thank you