

Building Foundation Models for Single-Cell Omics and Imaging

Bo Wang ([Twitter/X: @BoWang87](#))

Assistant Professor, University of Toronto

CIFAR AI Chair, Vector Institute

Chief AI Scientist, University Health Network



Before We Start

ScGPT

Preprint:

biorxiv.org/content/10.1101/2023.04.30.538439v2

Code: <https://github.com/bowang-lab/scGPT>

Nature Methods, Published, 2024

MedSAM

Preprint:

<https://arxiv.org/abs/2304.12306>

Code: <https://github.com/bowang-lab/MedSAM>

Nature Comm, Published, 2024

- We welcome your constructive comments and feedback!

What are Foundation Models?

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray O gut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*¹

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

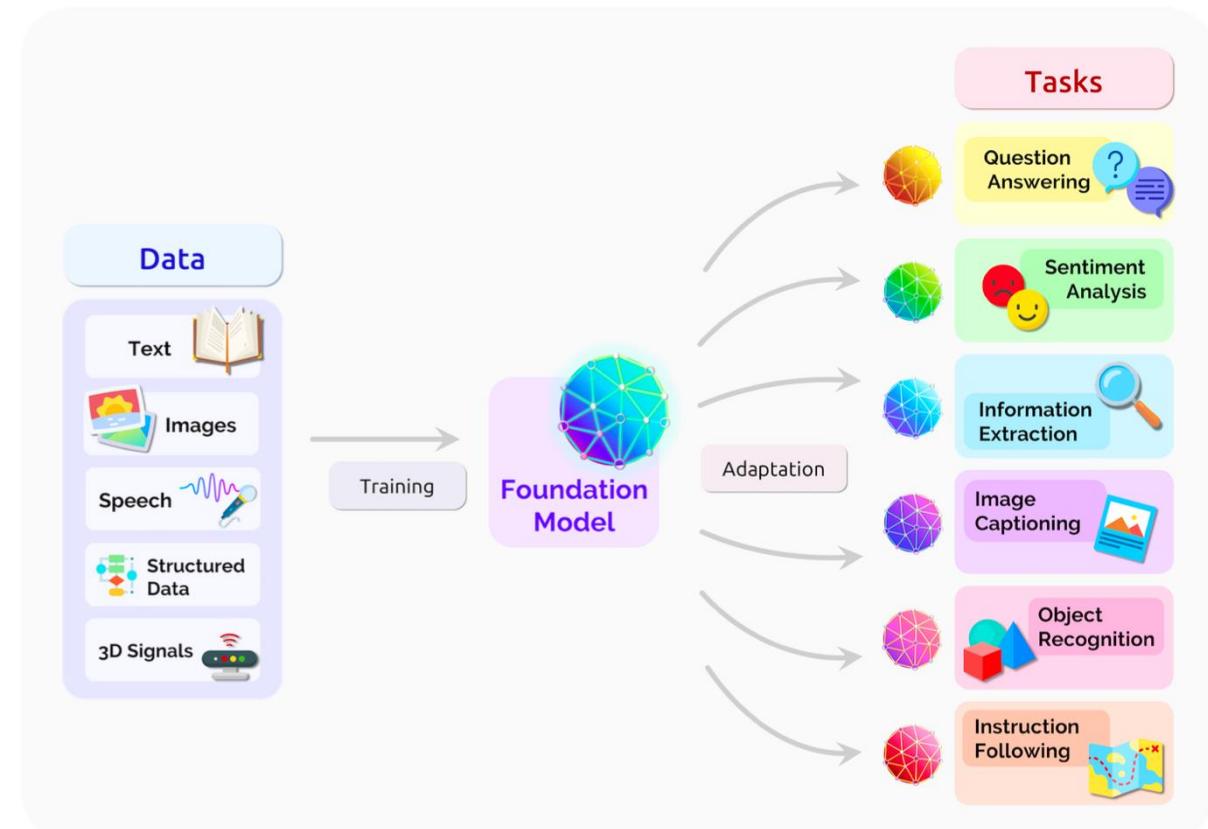


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

What are Foundation Models?

Key Features:

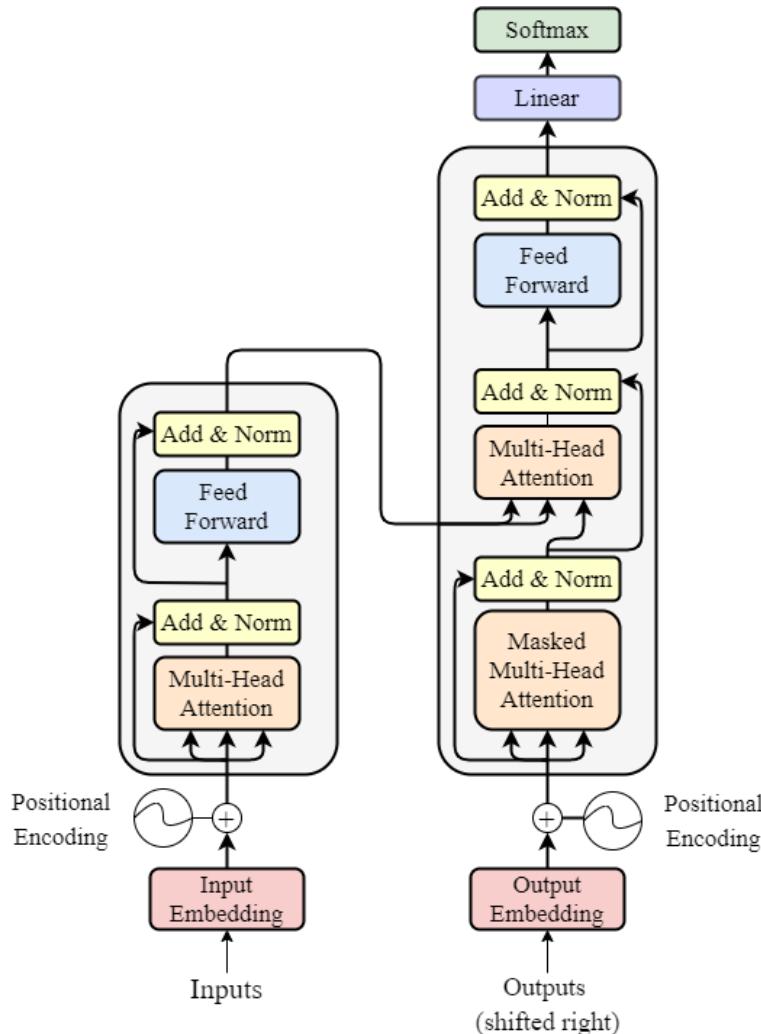
- **Algorithm:** Mostly transformer-based deep learning
- **Business:** One foundation model can enable multiple tasks by transfer learning
- **Computing:** Specialized hardware such as GPUs or TPUs.
- **Data:** Massive amount of data for pre-training

Can we build foundation models for single-cell data?

- **Motivation:** While texts are made up of words, cells can be characterized by genes. This analogy inspires us to explore foundation models for single cell. In the long run, we expect such model has the potential to gain general understanding of the underlying biology.
- Let's see whether we have all the key features (A-B-C-D)!

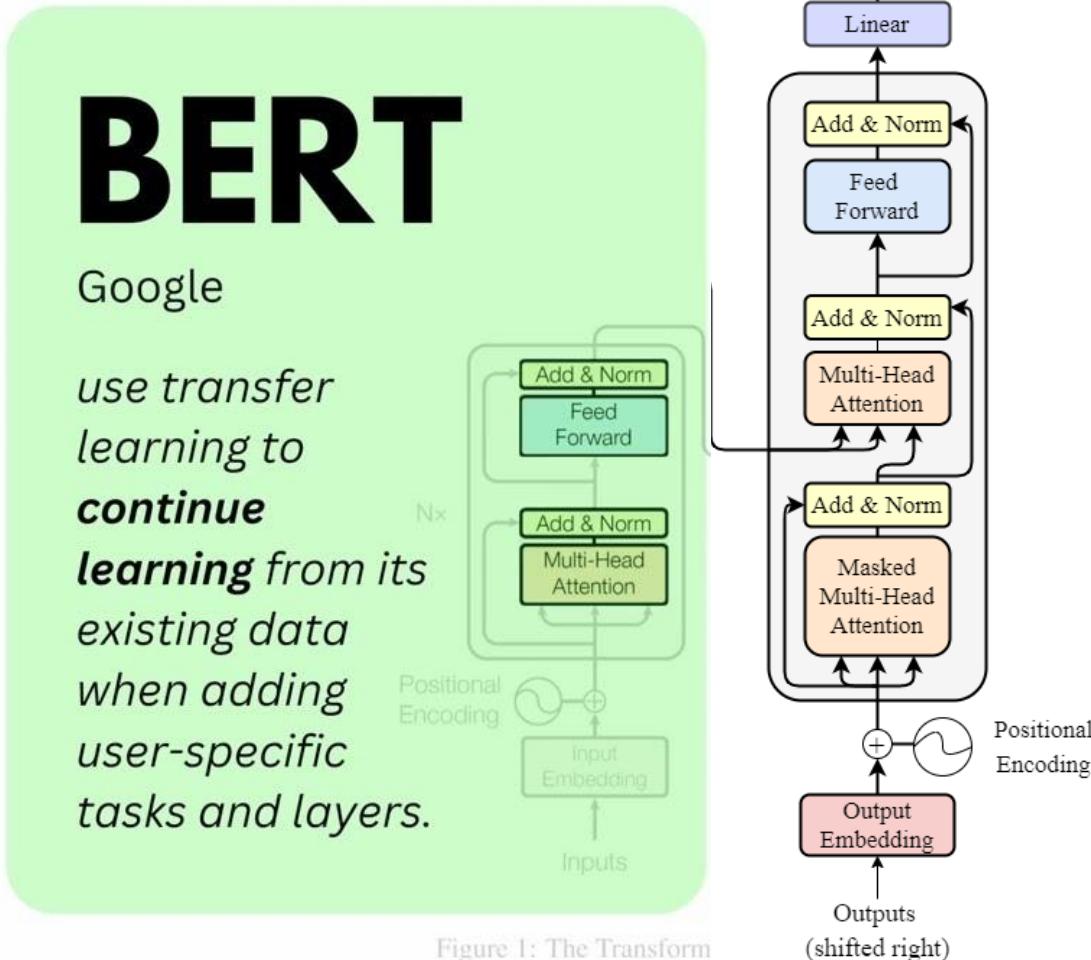
Can we build foundation models for single-cell data?

- **Algorithm:** Attention is all you need!



Can we build foundation models for single-cell data?

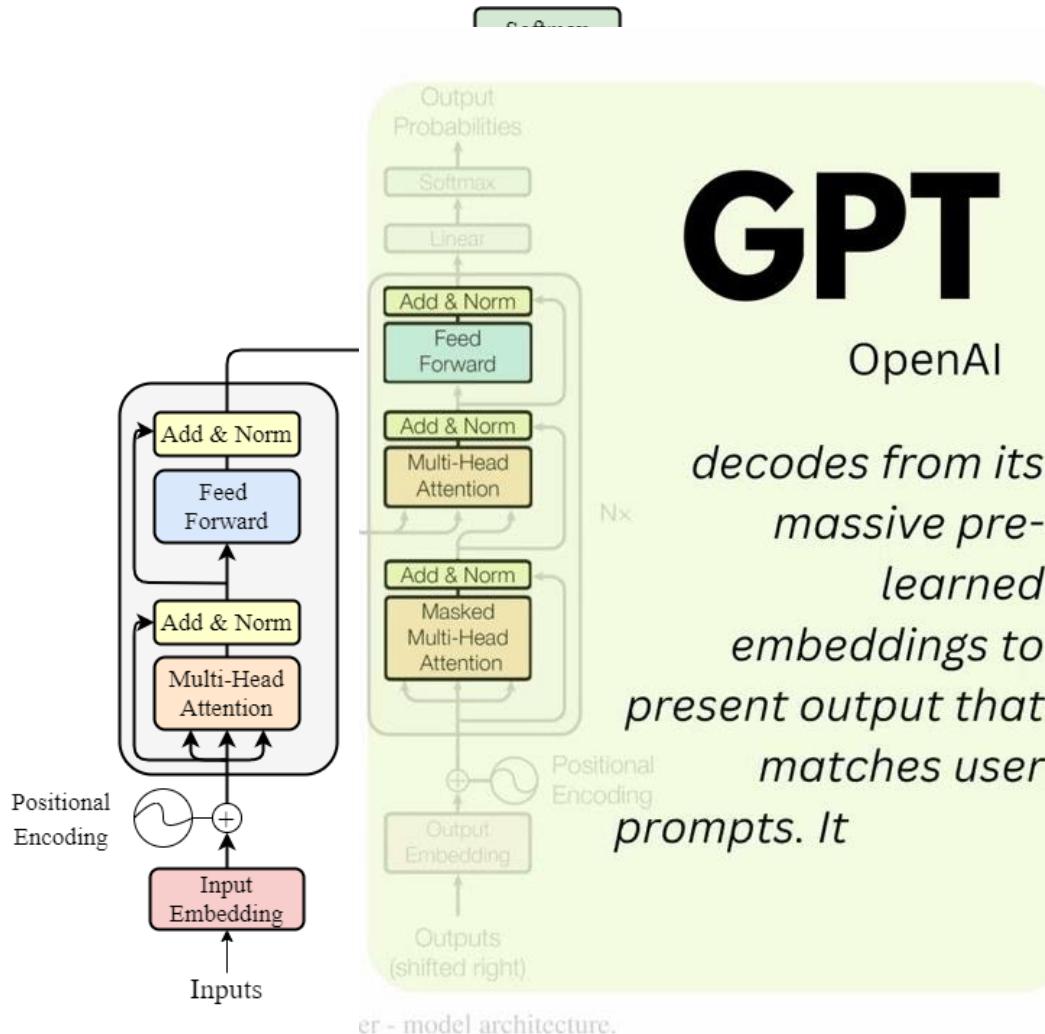
- **Algorithm:** Attention is all you need!



Examples: scBert, scFormer, GeneFormer

Can we build foundation models for single-cell data?

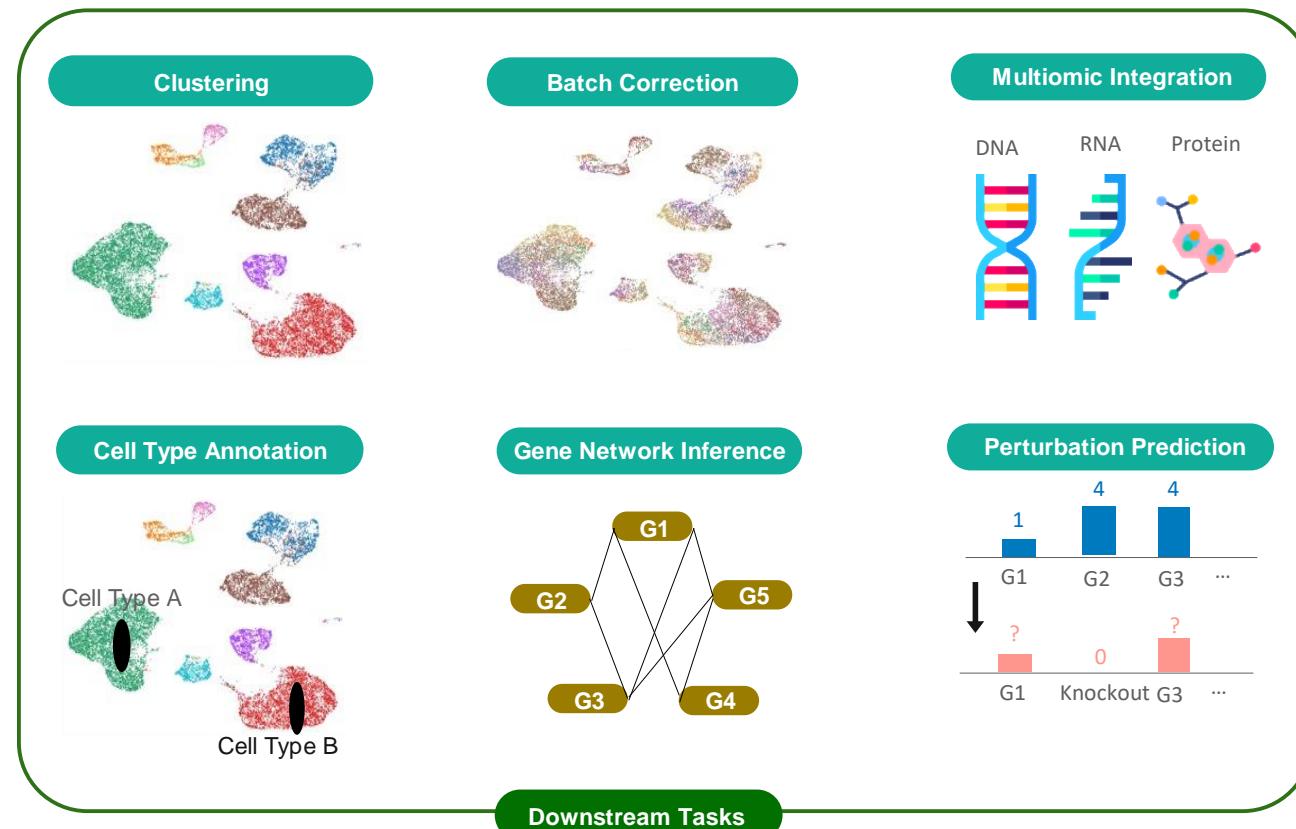
- **Algorithm:** Attention is all you need!



Examples: scGPT

Can we build foundation models for single-cell data?

- **Business:**



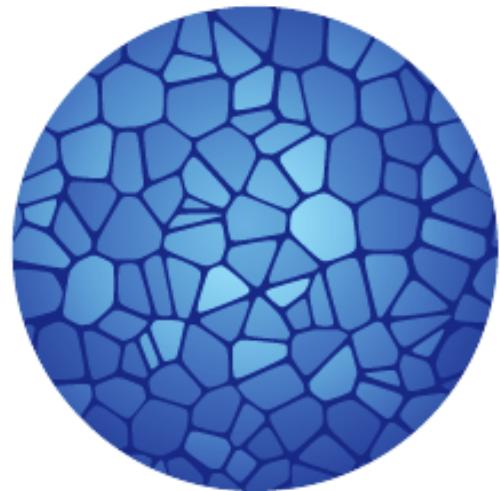
Can we build foundation models for single-cell data?

- Computing:

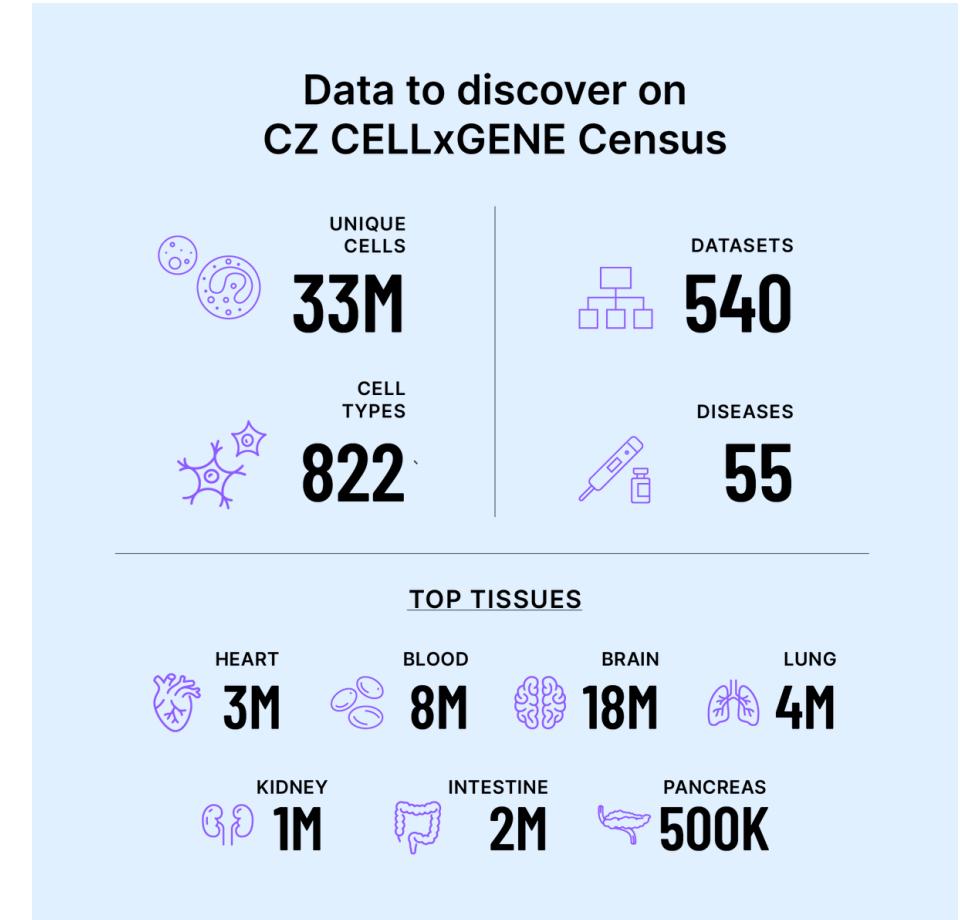


Can we build foundation models for single-cell data?

- Data:



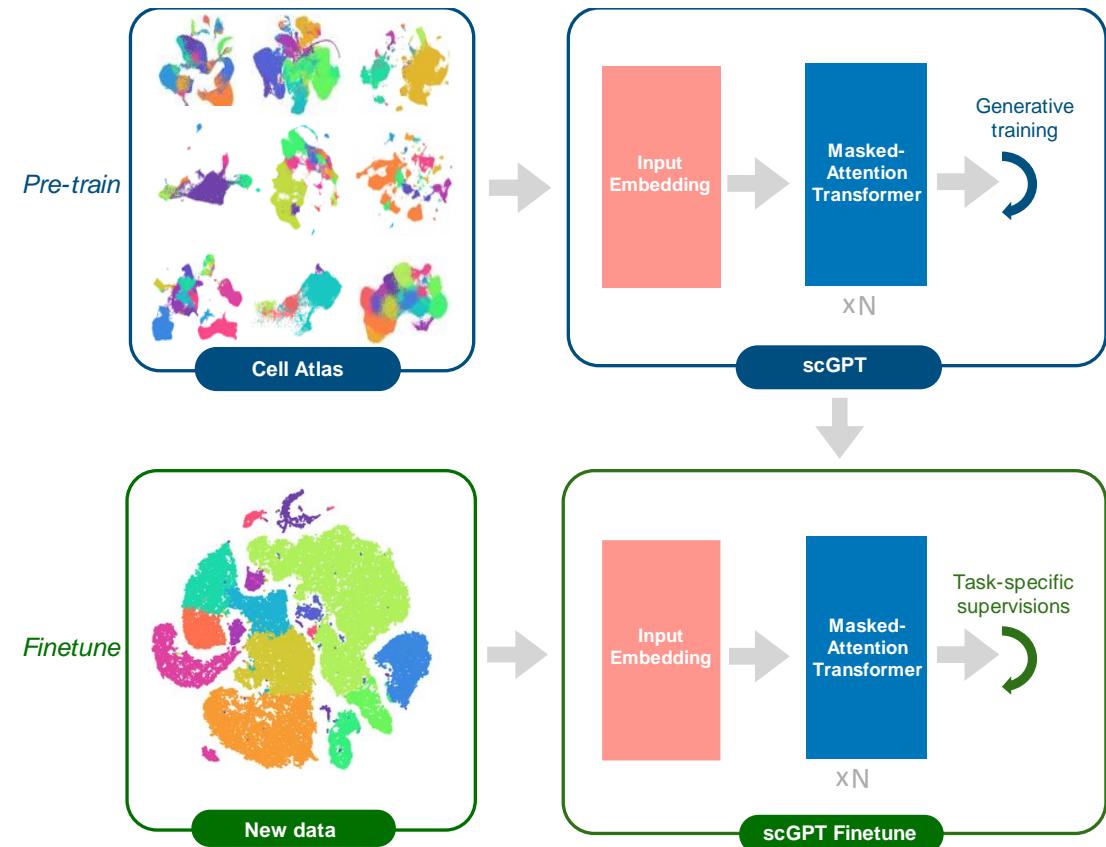
HUMAN
CELL
ATLAS



We are particularly excited about the recent launch of CellxGene census. Figure credit to [@JCoolScience](#)

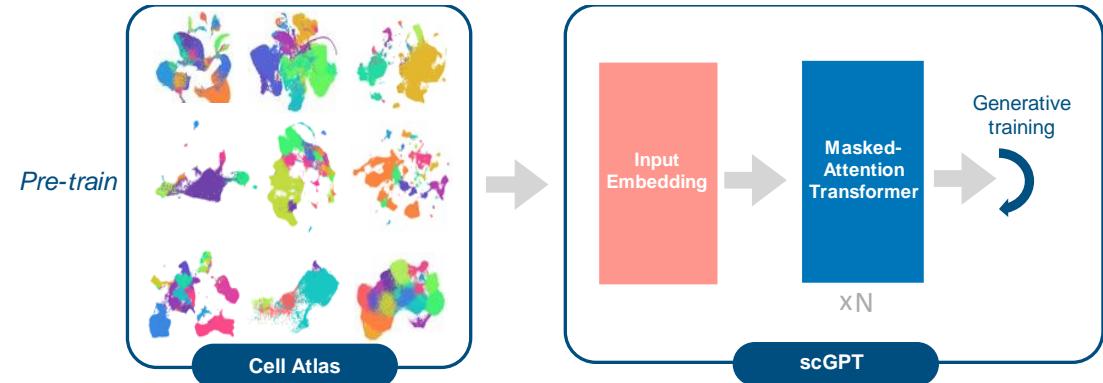
Overview of scGPT

- We present the first attempt to construct a single-cell foundation model. The scGPT workflow contains two steps:
 1. Generative **pre-training** on over 33 million whole-body normal cells
 2. **Fine-tuning** for downstream tasks

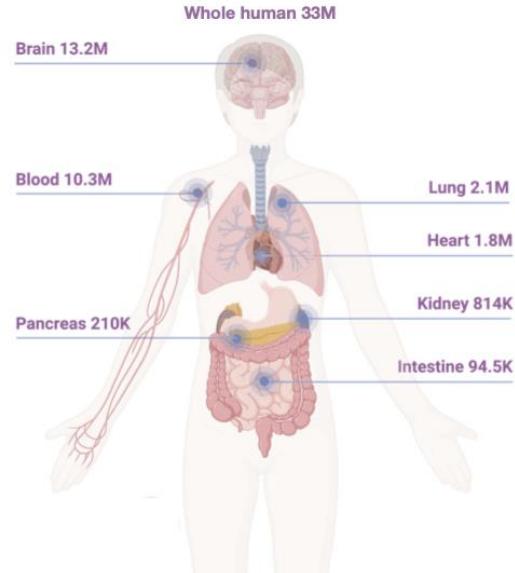


Generative pretraining in scGPT

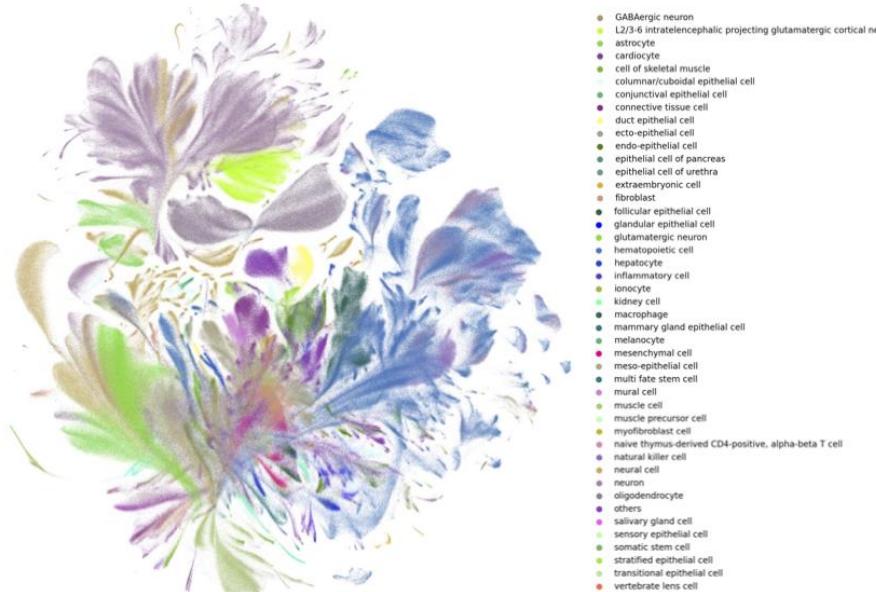
- In the pre-training:
 - Large-scale expression matrices were collected and preprocessed
 - As a first attempt, we filtered around 33 million human normal cells from the CZI CellxGene collection



D Cell numbers and origin tissues included in the pretraining



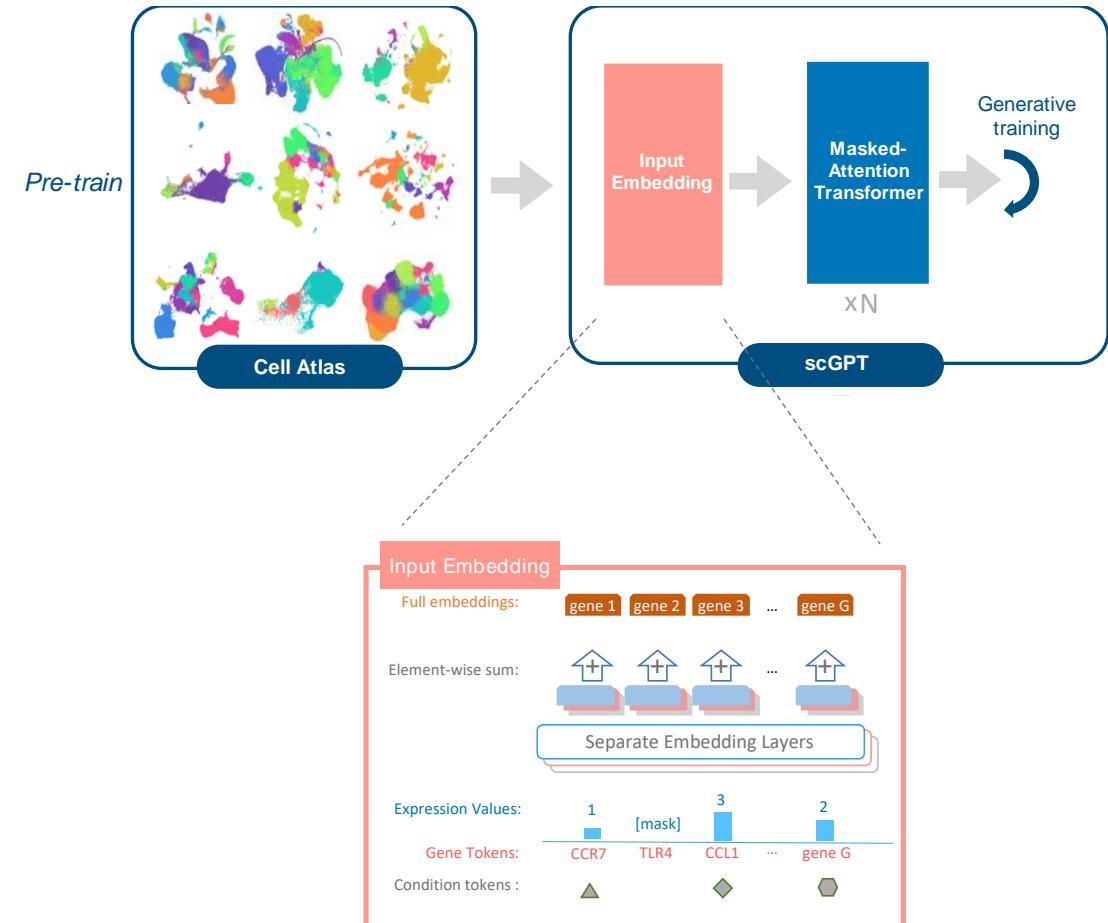
E UMAP of sampled normal human cells using scGPT emb.



Generative pretraining in scGPT

- In the pre-training:
 - Large-scale expression matrix were collected and preprocessed
 - As a first attempt, we filtered around 10.3 million blood and bone marrow cells from the CZ CellxGene collection
 - The input contains three sets of tokens for: (1) gene identifiers (2) expression values (3) condition flags

$$\mathbf{h}^{(i)} = \mathbf{emb}_g(\mathbf{t}_g^{(i)}) + \mathbf{emb}_x(\mathbf{x}^{(i)}) + \mathbf{emb}_c(\mathbf{t}_c^{(i)}).$$

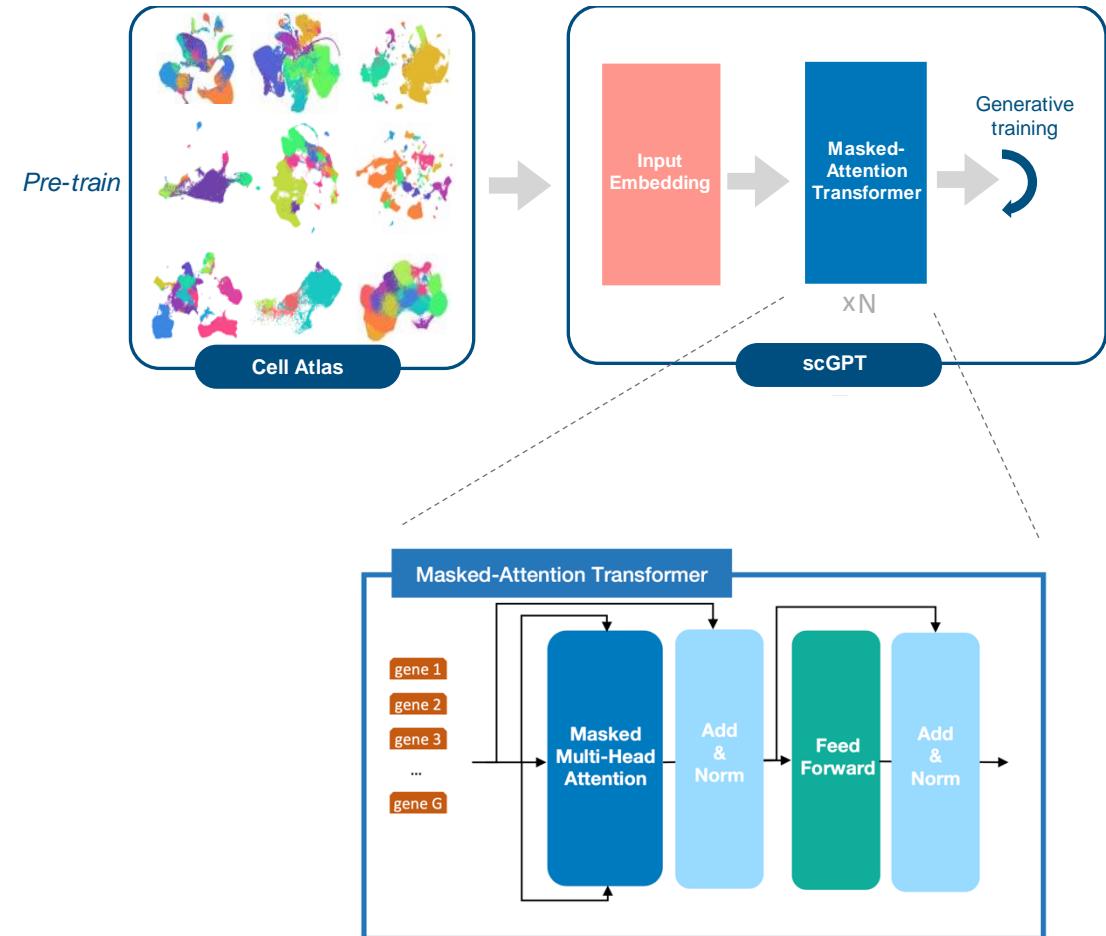


Generative pretraining in scGPT

- In the pre-training:
 - Large-scale expression matrix were collected and preprocessed
 - As a first attempt, we filtered around 10.3 million blood and bone marrow cells from the CZ CellxGene collection
 - The input contains three sets of tokens for: (1) gene identifiers (2) expression values (3) condition flags
 - The core model are 12 stacked transformer blocks, with **customized masked attention for generative training**

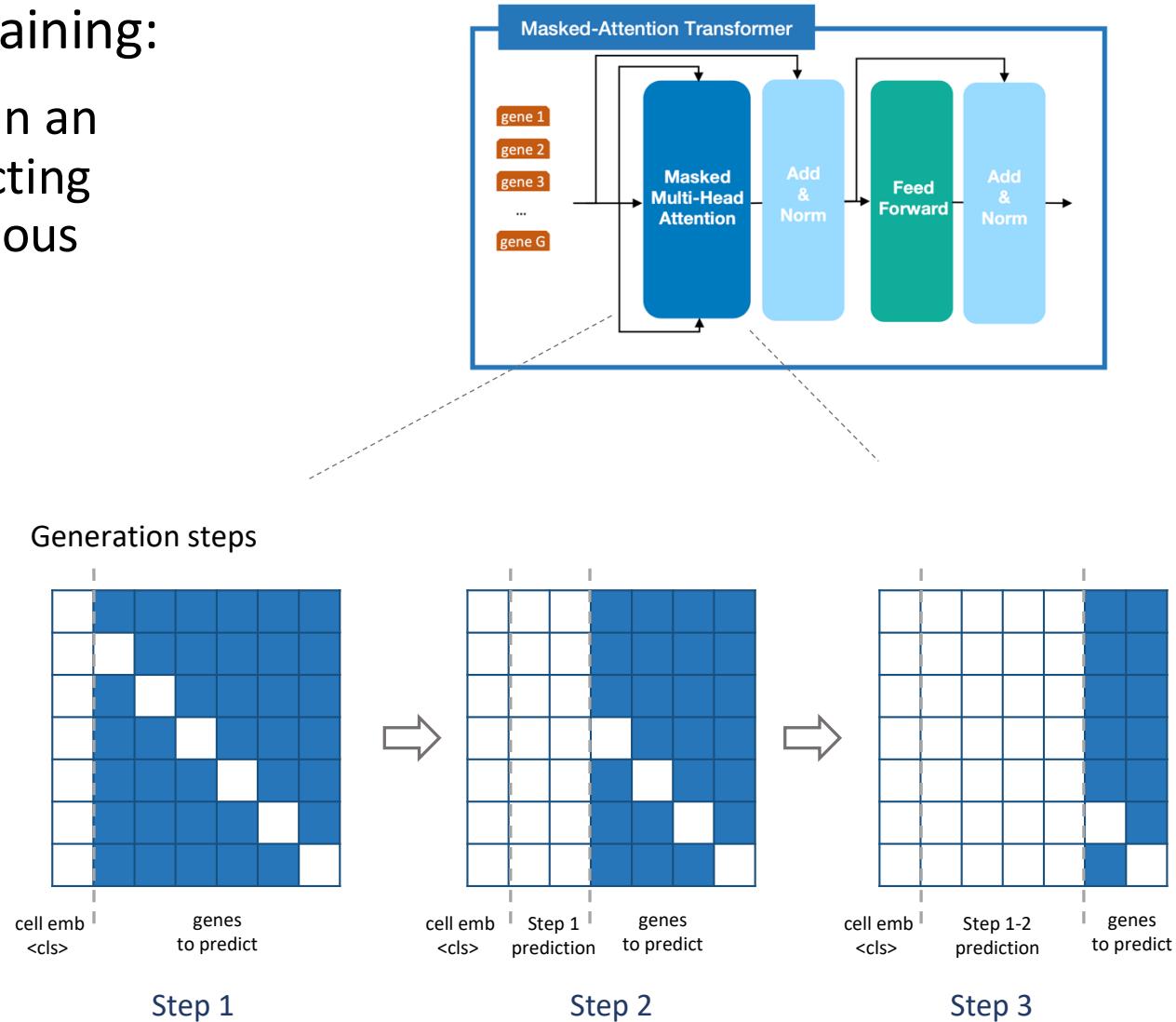
$$\mathbf{h}_0^{(i)} = \mathbf{h}^{(i)}$$

$$\mathbf{h}_l^{(i)} = \text{transformer_block}(\mathbf{h}_{l-1}^{(i)}) \quad \forall l \in [1, n]$$



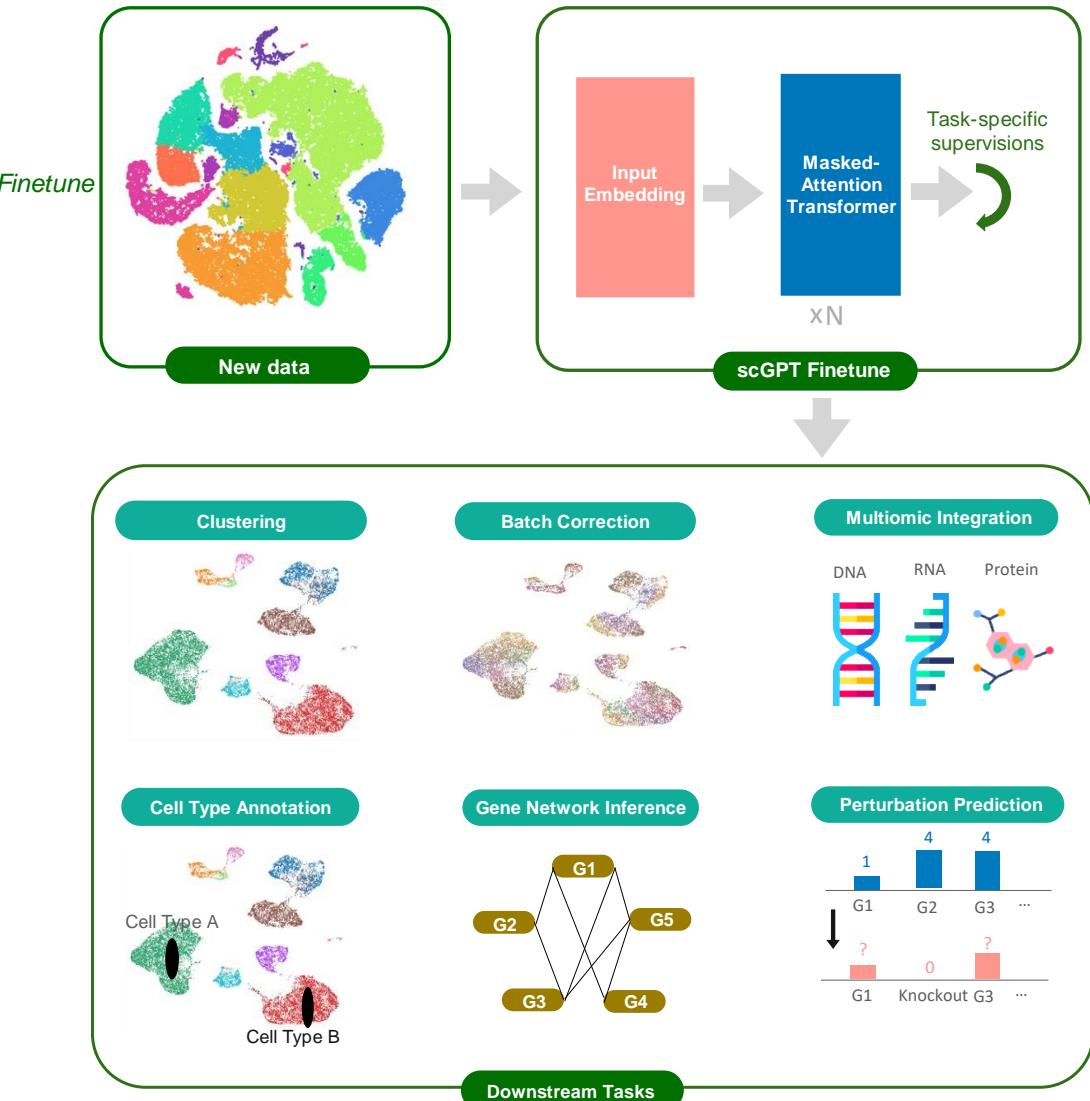
Generative pretraining in scGPT

- Attention mask for generative training:
 - Unknown genes are predicted in an auto-regressive manner: predicting unknown genes based on previous predictions



Overview of fine-tuning in scGPT

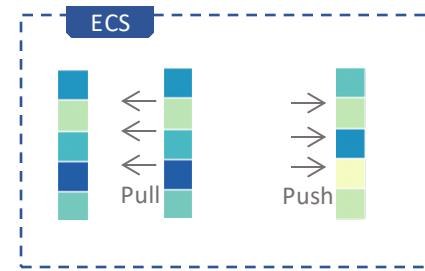
- In fine-tuning:
 - A set of versatile learning objectives are adopted to readily support various applications
 - Self-supervised objectives:
 - Gene Expression Prediction (GEP)
 - Gene Expression for Cell Modelling (GEPC)
 - Elastic Cell Similarity (ECS)
 - Supervised objectives:
 - Domain Adaptation via Reverse Back-propagation (DAR)
 - Cell Type Classification (CLS)



Finetuning scGPT for multi-batch integration

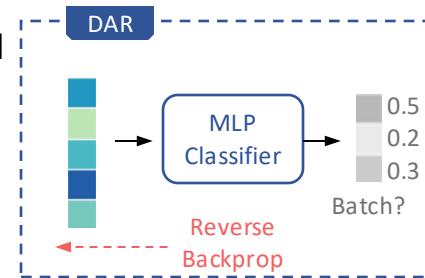
- We apply scGPT for multi-batch integration. To fine-tune the model, we included the objectives of GEP, GEPC, ECS and DAR

ECS[1]: enhance similarity between cell embeddings



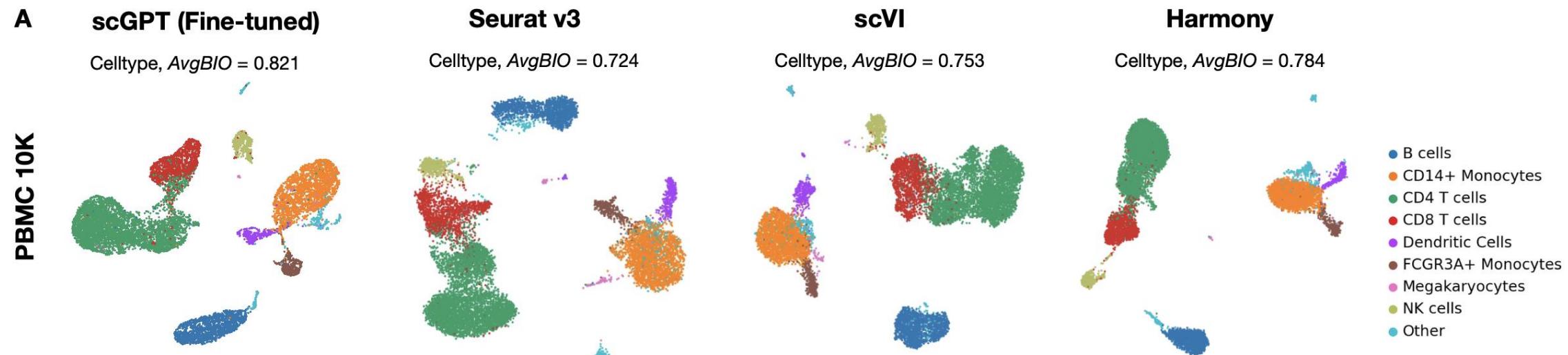
$$\mathcal{L}_{ECS} = -(\text{sim}(\mathbf{h}_c^{(i)}, \mathbf{h}_c^{(i')}) - \beta)^2$$

DAR[2]: predict the batch label and reverse the gradient



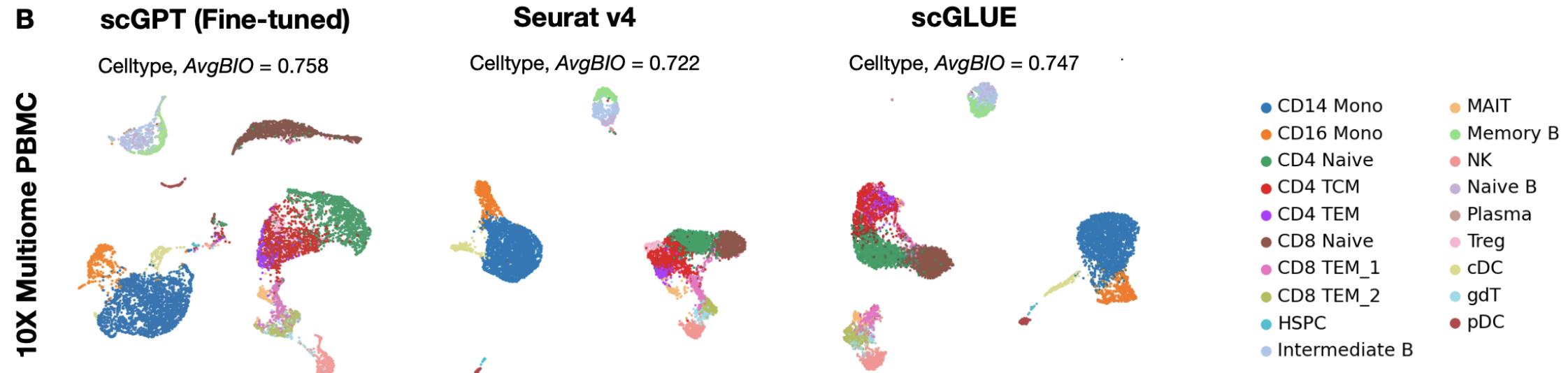
Finetuning scGPT for multi-batch integration

- We apply scGPT for multi-batch integration. To fine-tune the model, we included the objectives of GEP, GEPC, ECS and DAR
- We evaluated using combined scores (AvgBIO) of NMI, ARI and ASW.
- Fine-tuning shows improvements compared with training from-scratch. scGPT also demonstrates reasonable zero-shot ability.



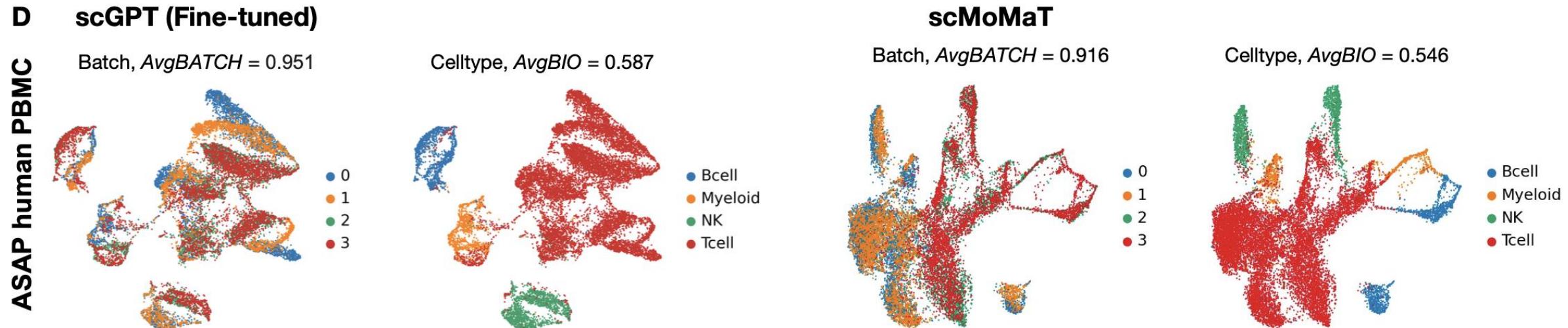
Finetuning for multi-omics integration

- The workflow for multi-omics is similar. The difference is that tokens for new omics (e.g. ATAC-seq) are appended. These tokens need to be trained from-scratch since missing from the pretrained step.
- On a paired sequencing data of 10x Multiome PBMC^[3]



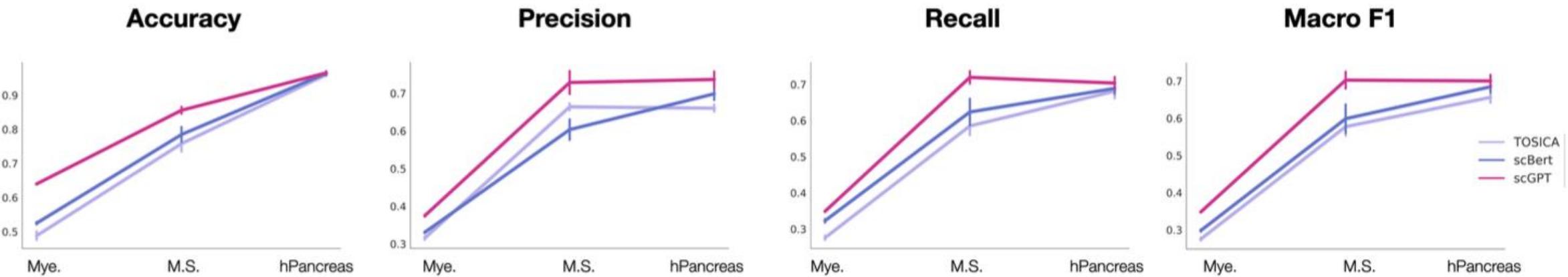
Finetuning for multi-omics integration

- The workflow for multi-omics is similar. The difference is that tokens for new omics (e.g. ATAC-seq) are appended. These tokens need to be trained from-scratch since missing from the pretrained step.
- In a mosaic setting, on ASAP PBMC dataset^[4]



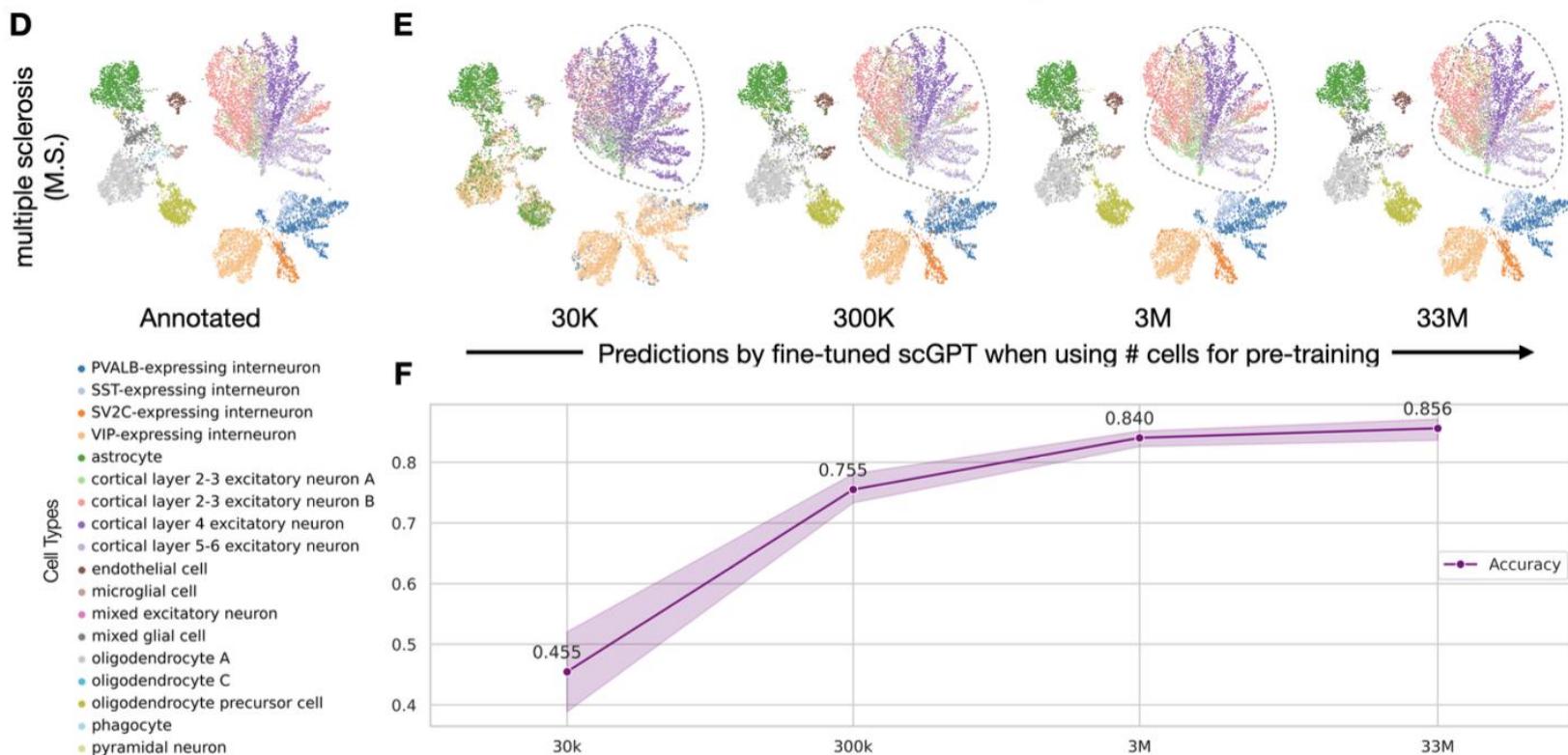
Finetuning for cell type annotation

- For cell type annotation, we used cross-entropy classification loss to train the cell embeddings from scGPT.
- The model is tested on a held-out test set. Accuracy, precision, recall and macro F1 are reported.



Scaling Law of Foundation Models

- For cell type annotation, we used cross-entropy classification loss to train the cell embeddings from scGPT.
- The model is tested on a held-out test set. Accuracy, precision, recall and macro F1 are reported.



Predicting genetic perturbation responses

- To predict the genetic perturbation, we use the Gene Expression Prediction (GEP) objective. Instead of recovering the gene expression, the model is trained to predict the target expression after a perturbation, given the input cell in control setting.
- Evaluation on two datasets Norman et. al.^[5] and Adamson et. al.^[6]. We report the Pearson correlation between the expression values and expression changes (Δ), on top 20 DEs and on all highly variable genes.

Table 1: Results of perturbation prediction

Model	Norman et al.				Adamson et al.			
	DE		ALL		DE		ALL	
	corr	corr(Δ)	corr	corr(Δ)	corr	corr(Δ)	corr	corr(Δ)
MLP	0.909	0.428	0.987	0.408	0.948	0.729	0.991	0.656
GEARS	0.917	0.508	0.986	0.387	0.961	0.726	0.991	0.652
scGPT	0.923	0.546	0.988	0.459	0.971	0.775	0.992	0.647

Limitations

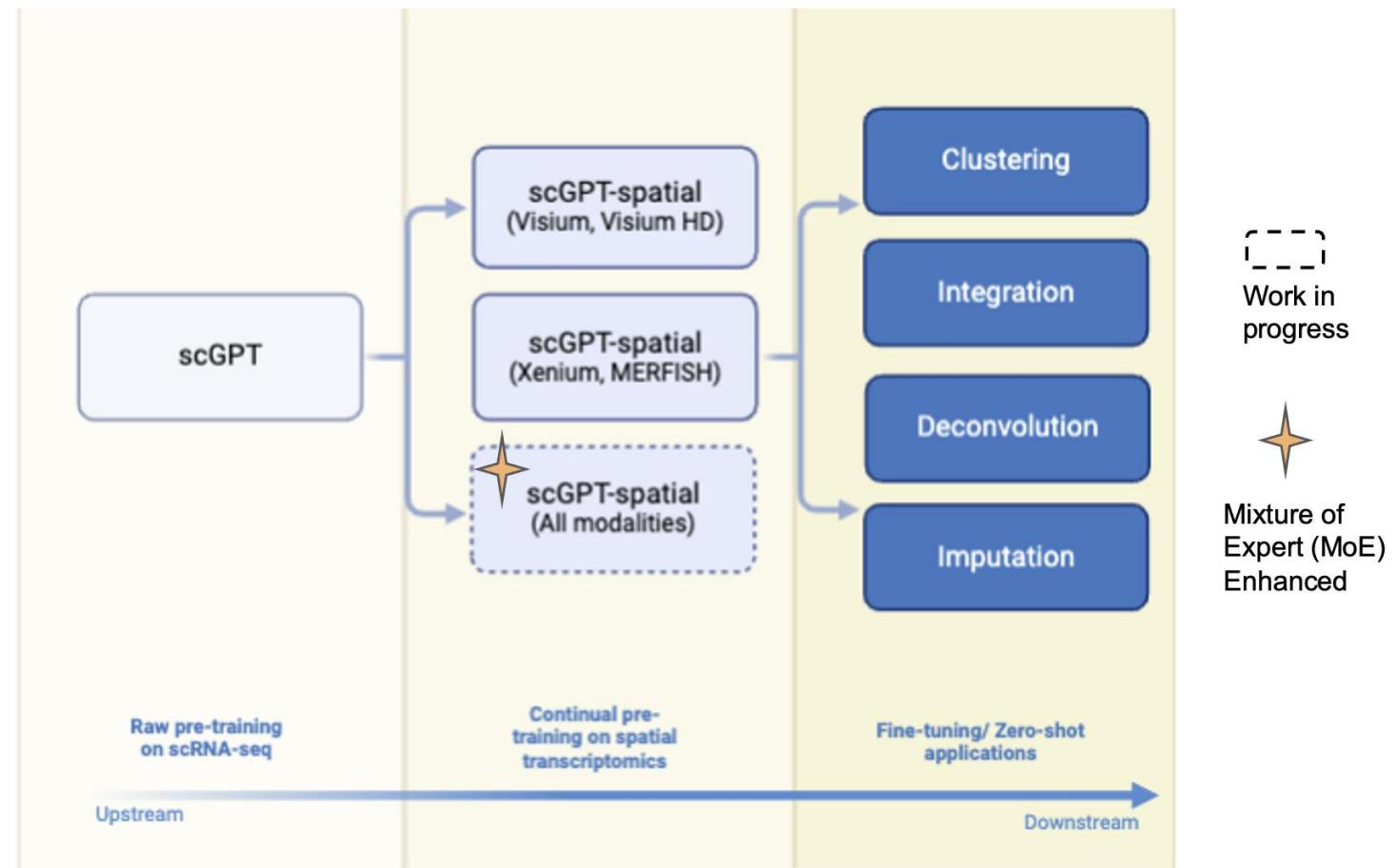
- **scGPT is (mostly) limited in RNAseq.** Despite that scGPT is capable of integrating multi-omic data including scATACseq, we only have a well-trained base model for scRNAseq due to data limitation.
- **guilt by associations.** scGPT, despite of learning meaningful gene programs, is still limited in causal inference.
- **Pre-training is very compute-intensive, though fine-tuning is light-weight.**
- **Zero-Shot capacity is still limited! (GPT2 time for single-cell!)**

Can we extend scGPT to spatial omics?

Continual training for spatial omics workflow

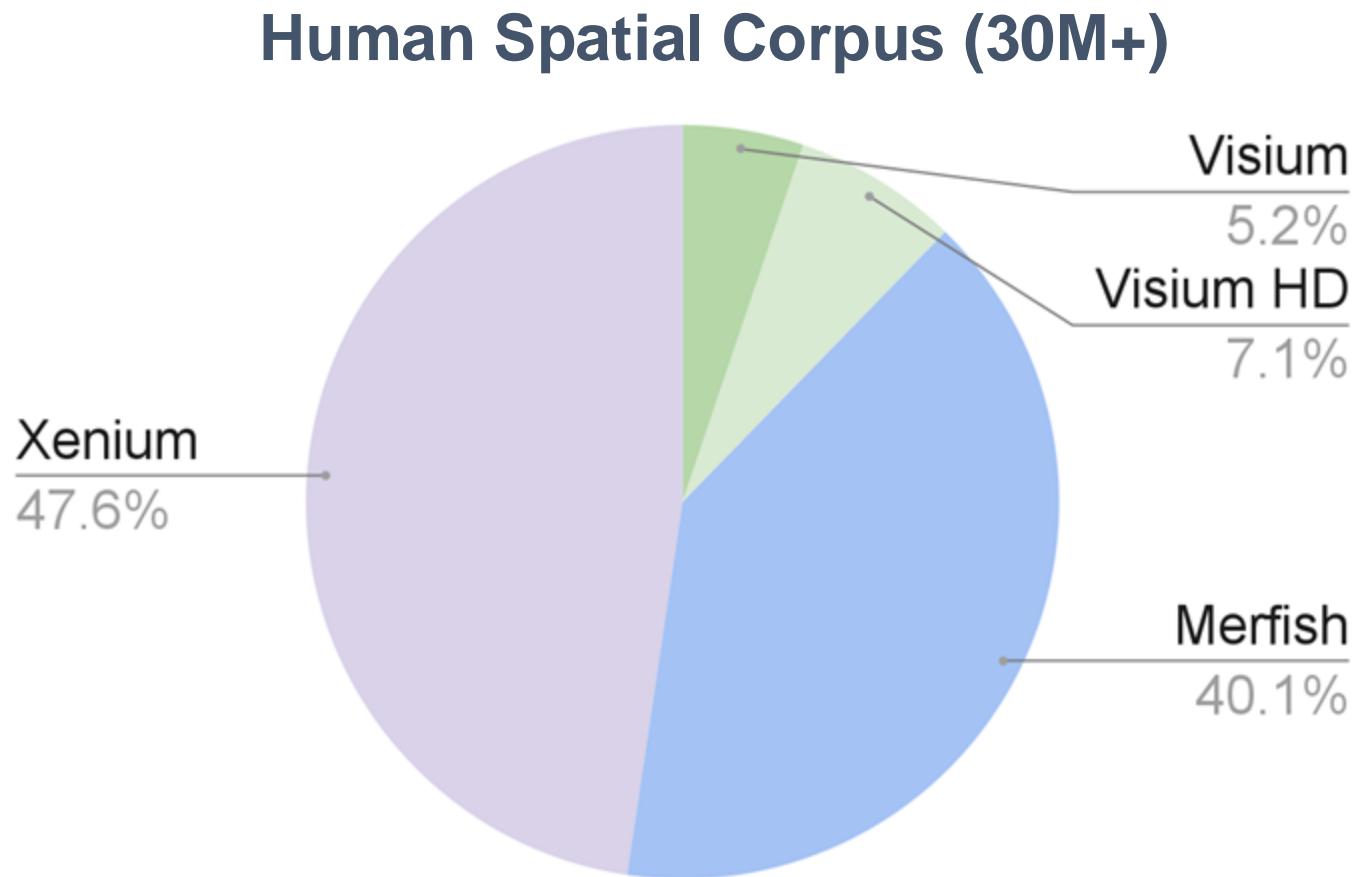
- Start from pre-trained scGPT model on scRNA-seq
- **Continual pre-pretraining** on diverse spatial transcriptomic modalities
- Downstream applications tailored towards each assay

On-going work!



Human Spatial Data Corpus for scGPT-spatial Continuous Pre-training

- Collected **30+ million human single cells/spots**
 - Sequence- and image-based spatial transcriptomics
 - **Visium, Visium HD, Xenium, Merfish**
- Healthy and diseased data (e.g., cancer)
- Major sources include [CellXGene](#), [GEO](#), [10X data release](#), [Vizgen data release](#), and [Allen Brain Cell Atlas](#)

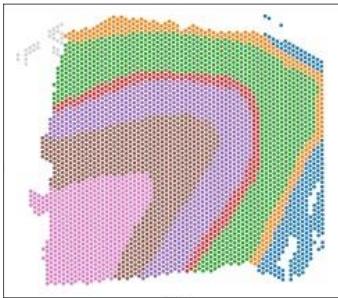


On-going work!

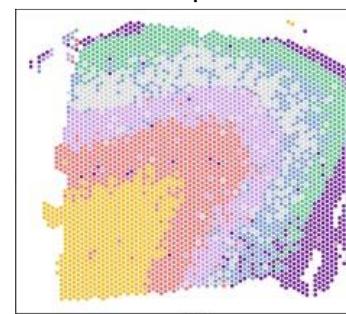
Spatial Domain Clustering on Visium

- Using fine-tuned cell embeddings for spatial domain clustering
 - Apply on the human cortex samples of 10 Visium slides (Maynard et al. 2021)

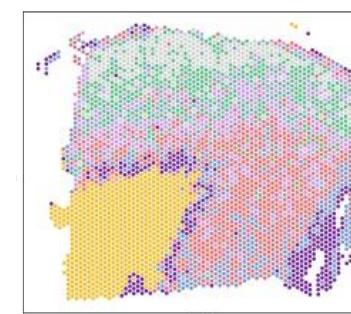
Ground Truth Regions



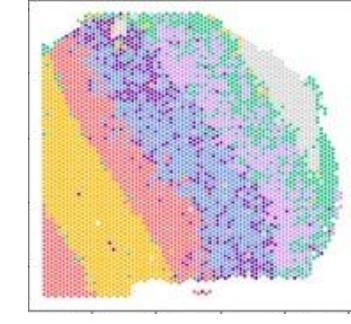
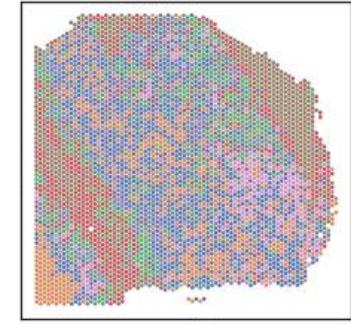
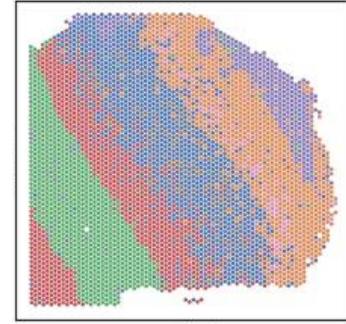
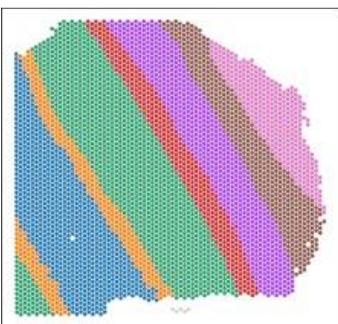
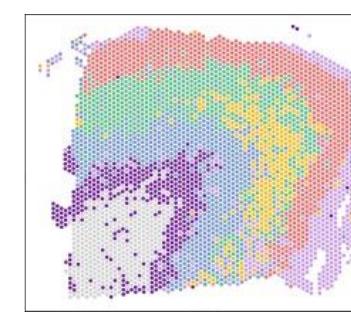
scGPT-Spatial



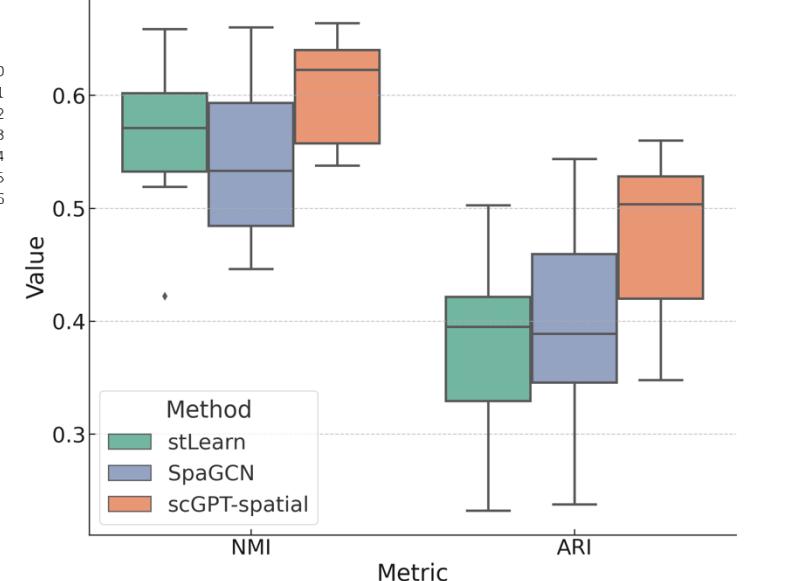
Leiden



spaGCN



Clustering Metrics of stLearn, SpaGCN, and scGPT-spatial



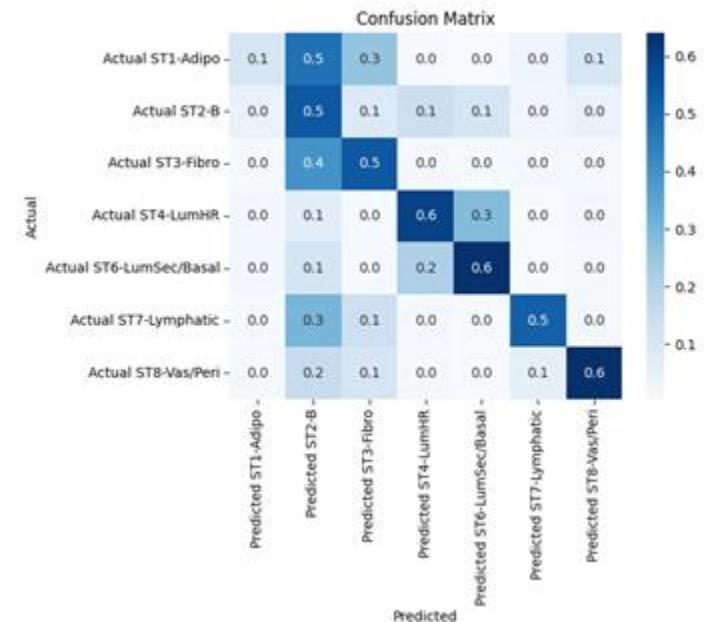
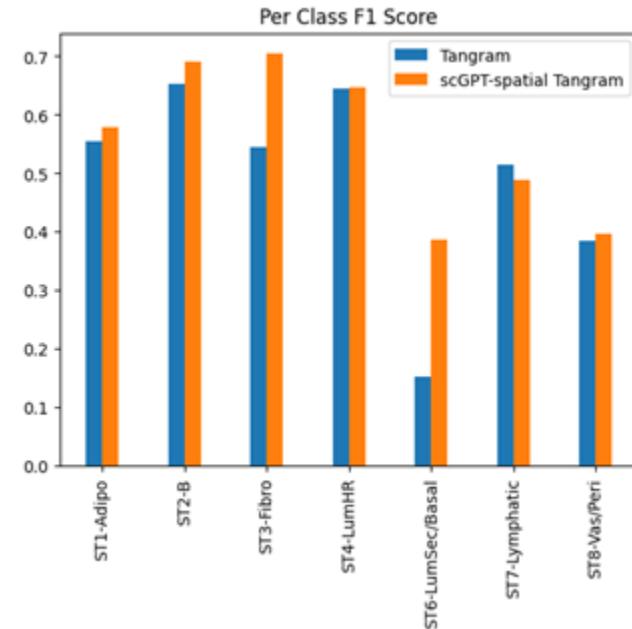
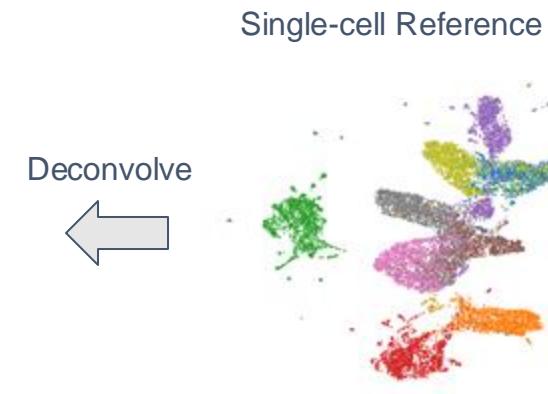
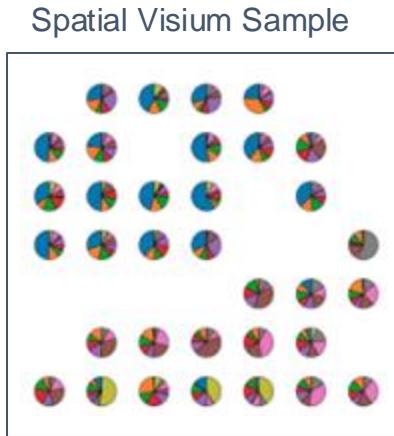
On-going work!

Visium Deconvolution

- Using fine-tuned cell embeddings for deconvolution
 - Estimate cell proportions by aligning spatial embeddings and expression with reference scRNA-seq
- Apply on the human breast samples of 10 visium slides ([Kumar et al. 2023](#))
 - Macro F1: 0.56 by scGPT-spatial enhanced method, compared to 0.49 by the vanilla tangram

On-going work!

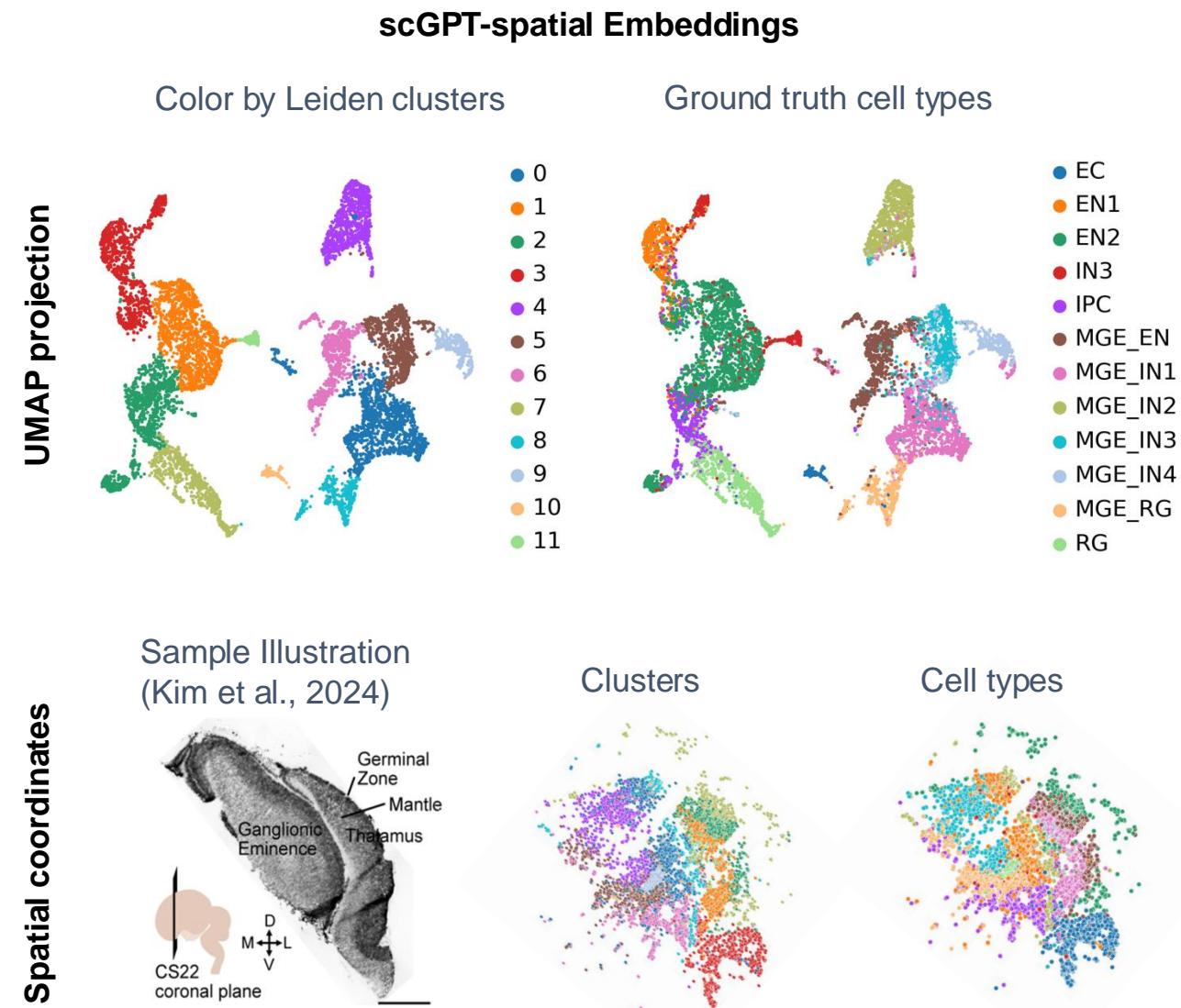
Illustration of Visium Deconvolution



Merfish Clustering

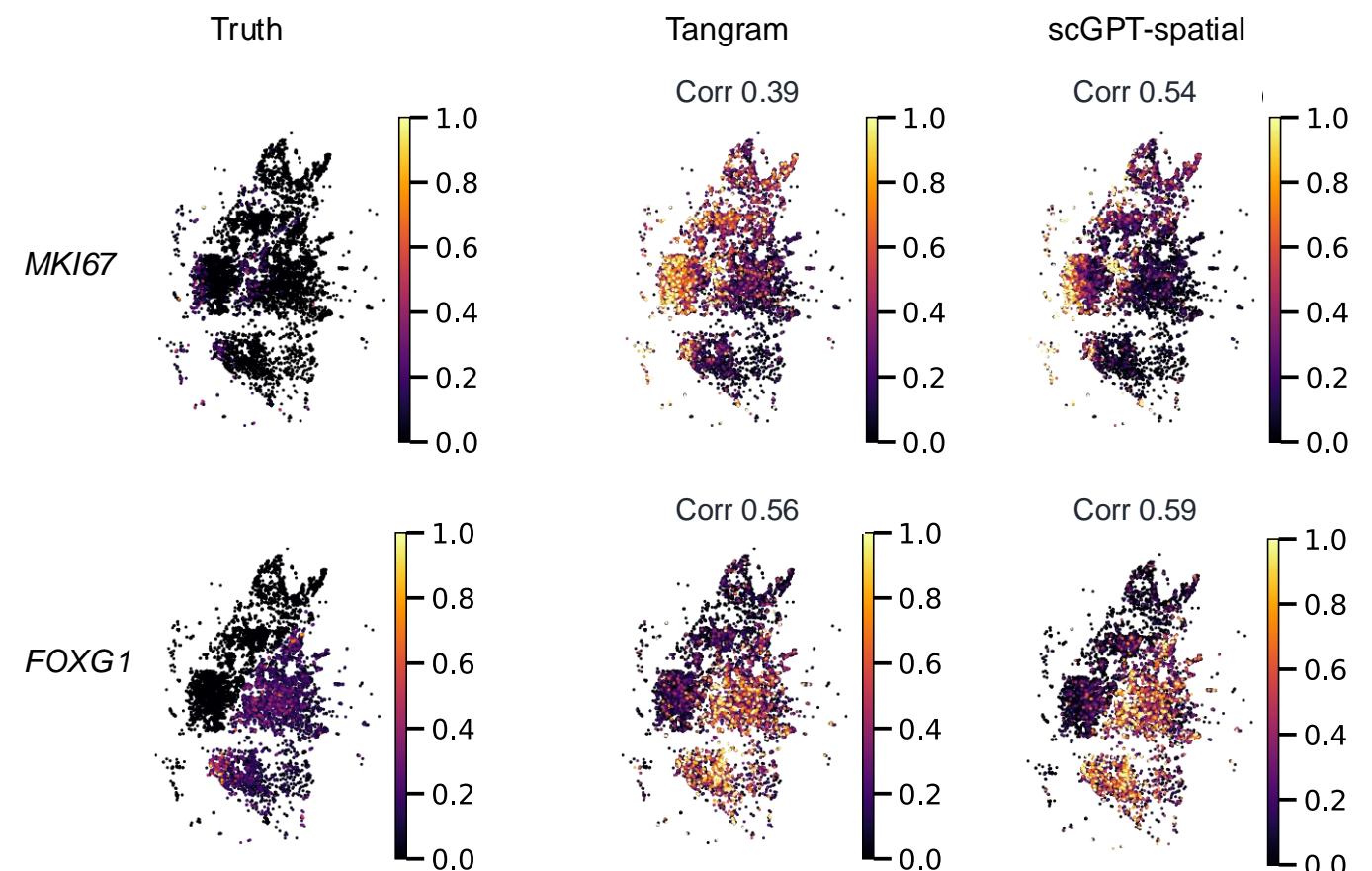
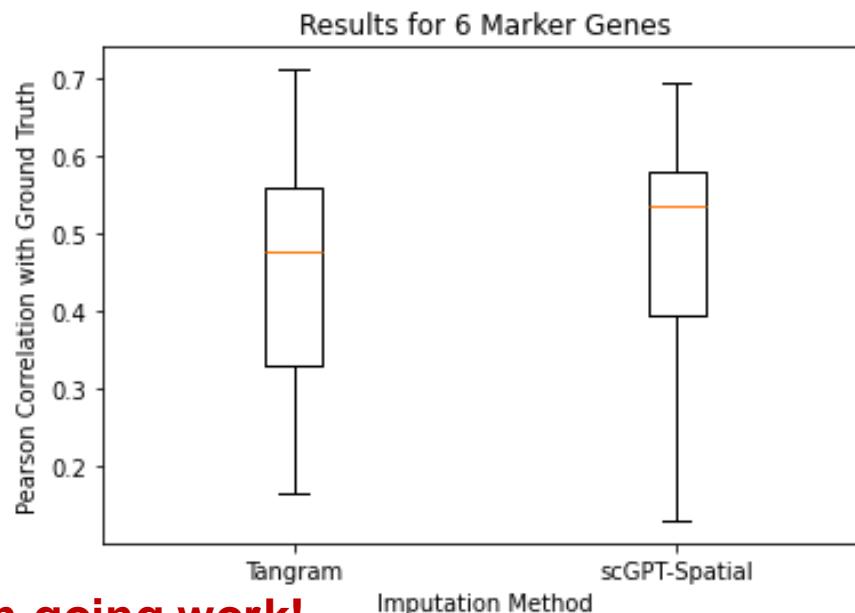
- Light fine-tuning of scGPT-spatial model on a Merfish slide of 8,709 cells from developing human brain (thalamus) ([Kim et al., 2023](#))
- Coordinate-based patch-wise sampling inspired by computer vision models

On-going work!



Imputation of gene expression for MERFISH

- Imputation of missing gene expression by matching reference single-cell data
 - Apply on the thalamus sample from Kim et al. 2023
 - Impute six marker gene expression



Can we add imaging to scGPT (-spatial)?

- The challenge is data scarcity!
- Can we also have a foundation model for biological images?
- The answer is yes!

nature methods

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [nature methods](#) > [comment](#) > [article](#)

Comment | [Published: 11 July 2023](#)

Towards foundation models of biological image segmentation

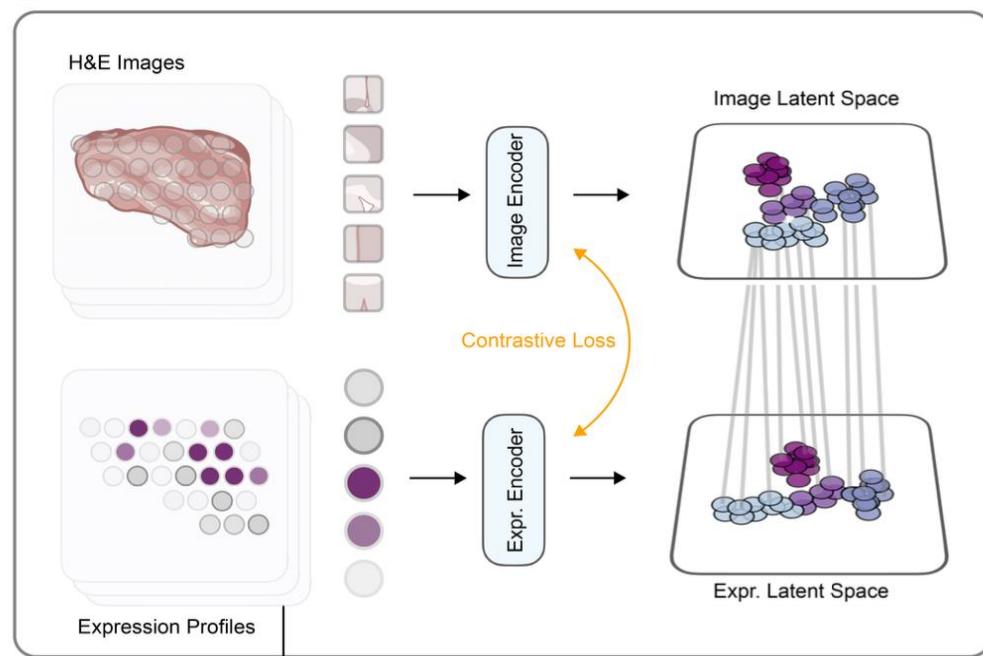
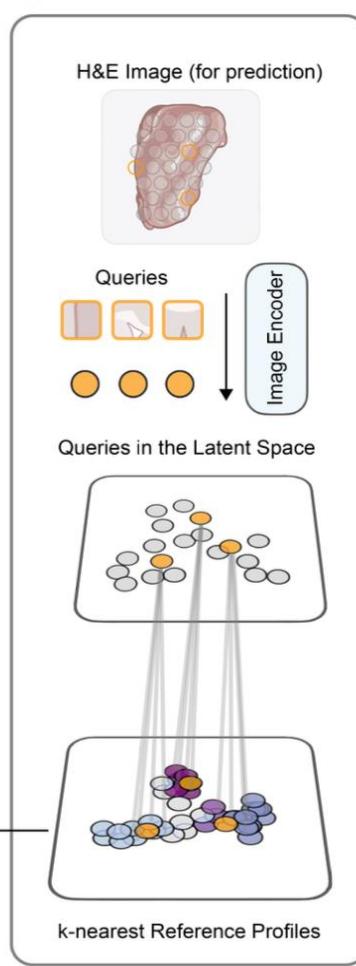
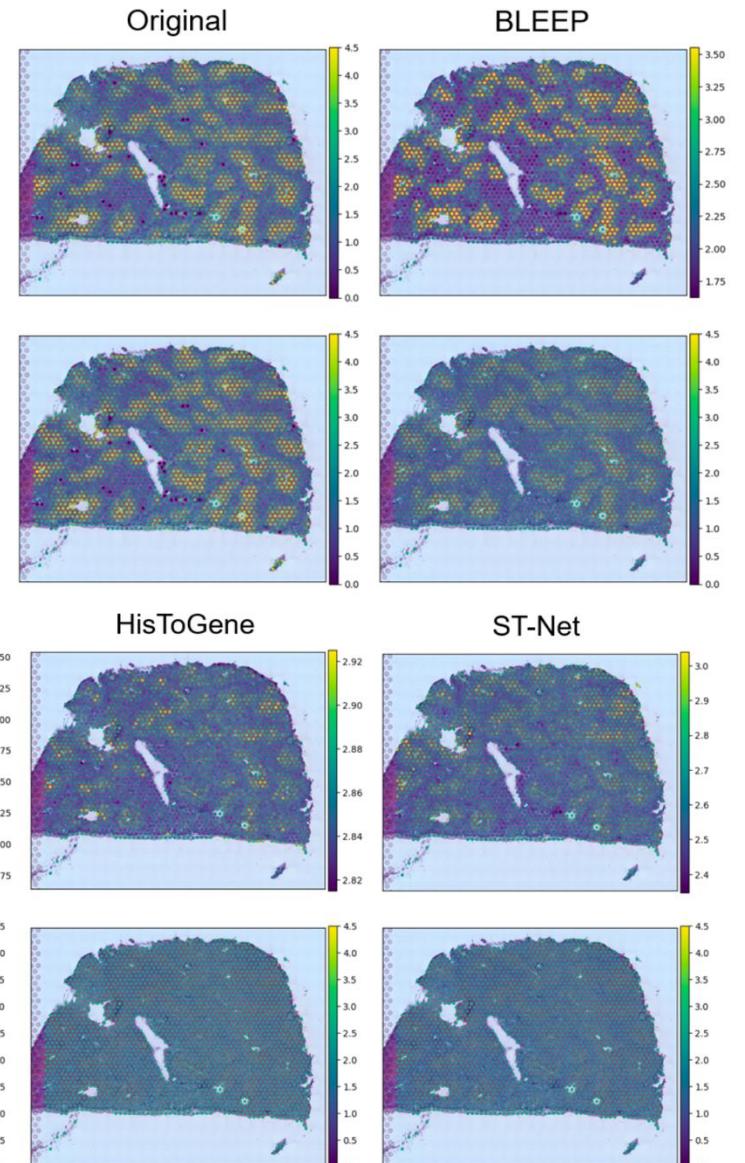
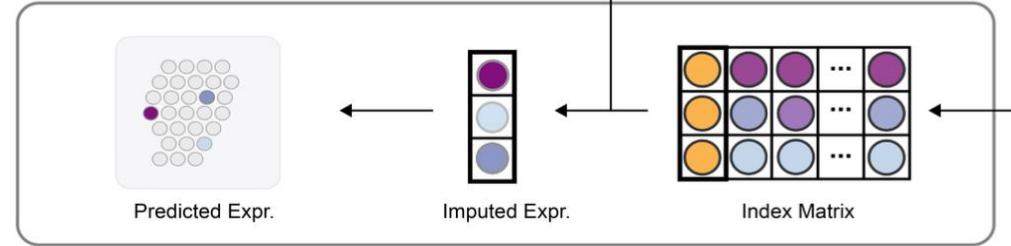
[Jun Ma](#) & [Bo Wang](#) 

[Nature Methods](#) **20**, 953–955 (2023) | [Cite this article](#)

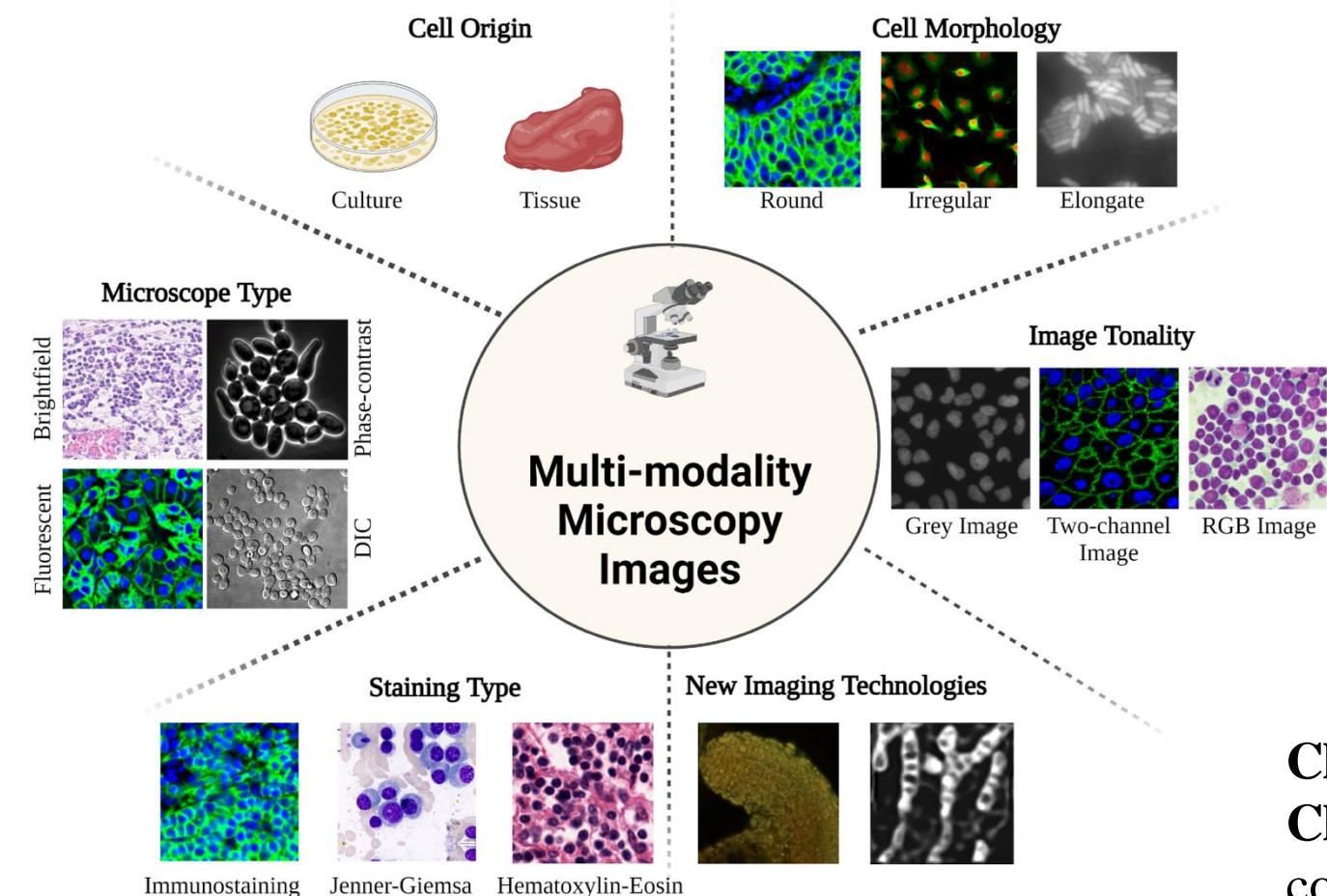
4387 Accesses | 1 Citations | 37 Altmetric | [Metrics](#)

<https://www.nature.com/articles/s41592-023-01885-0>

BLEEP: from H&E Images to Expressions

a**b****c**

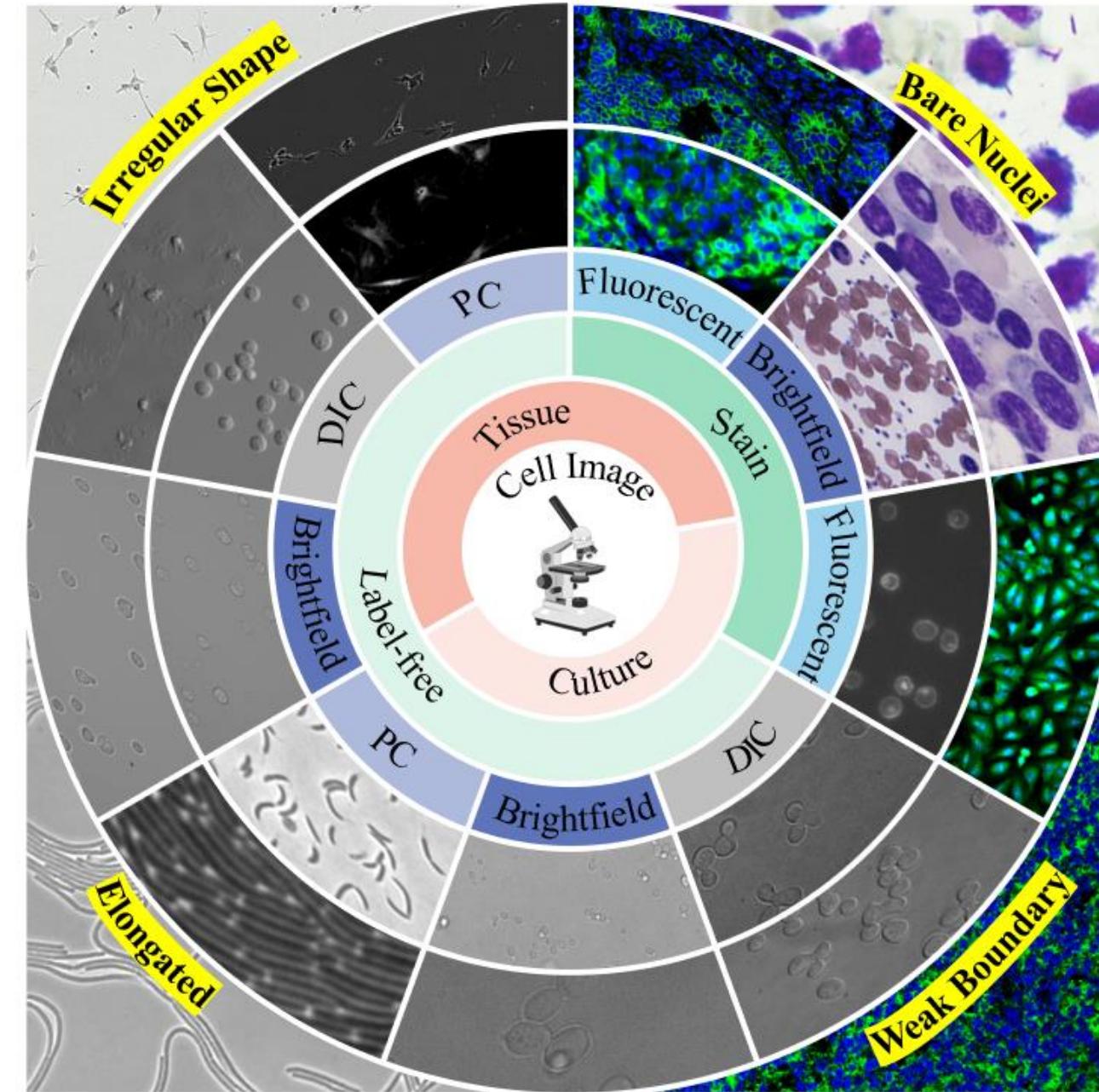
How about biological images?



Universal cell segmentation remains a challenging problem!

- Challenge 1.** Microscopy images are extremely diverse.
- Challenge 2.** Cells touch together and have non-conventional morphology.
- Challenge 3.** Lack of annotations: manually annotating the cells is labor-intensive and time-consuming.

2022 Neurips Competition Dataset

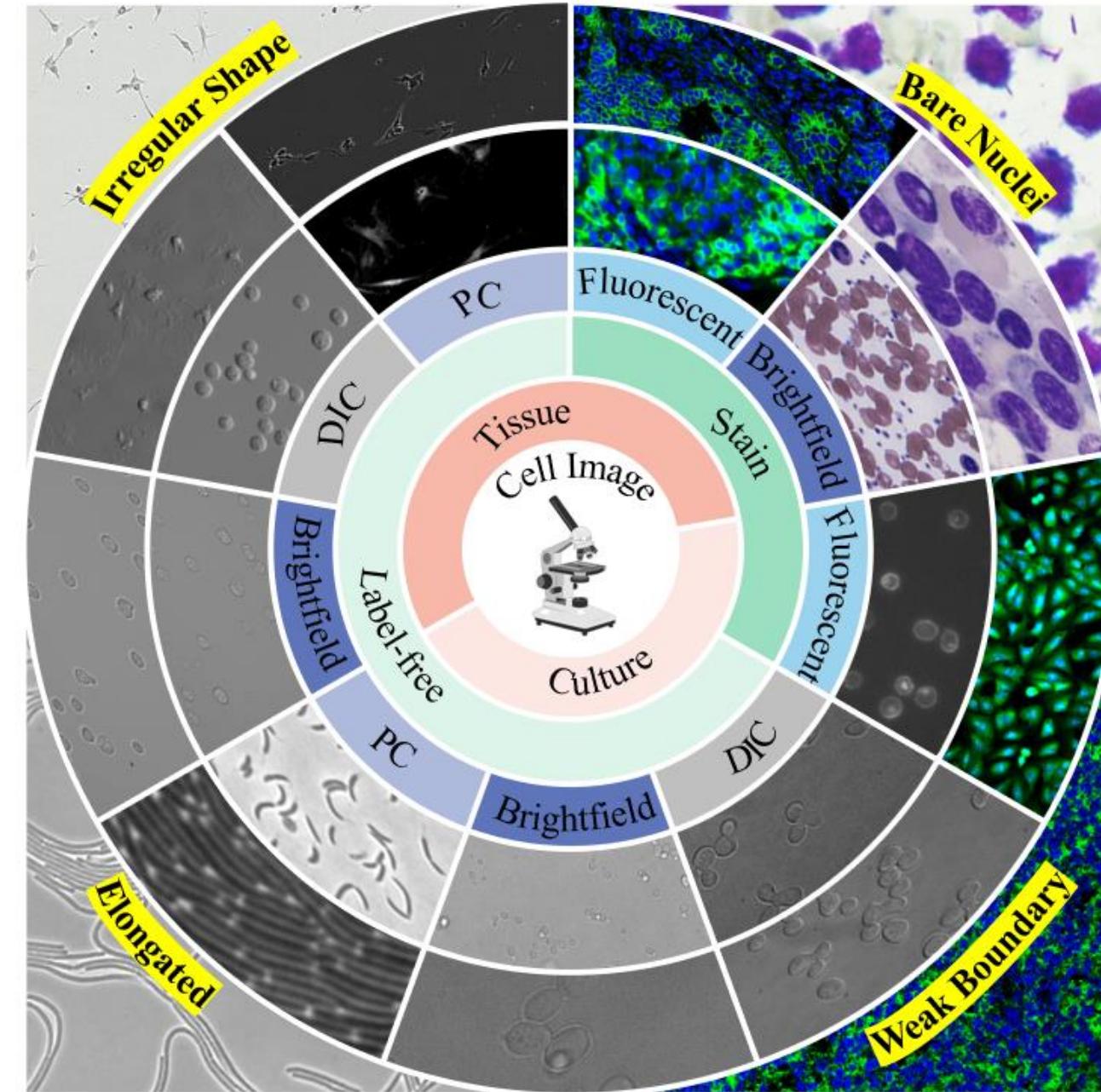


A diverse dataset

- Different cell origins: culture and tissue
- Different microscope modalities:
 - Brightfield
 - Fluorescent
 - Phase-contrast
 - Differential interface contrast (DIC)
- Different staining types: immunostaining, Jenner-Giemsa, label-free...
- Different morphologies: round, elongated, starlike...

<https://www.nature.com/articles/s41592-024-02233-6>
Nature Methods, 2024

2022 Neurips Competition Dataset



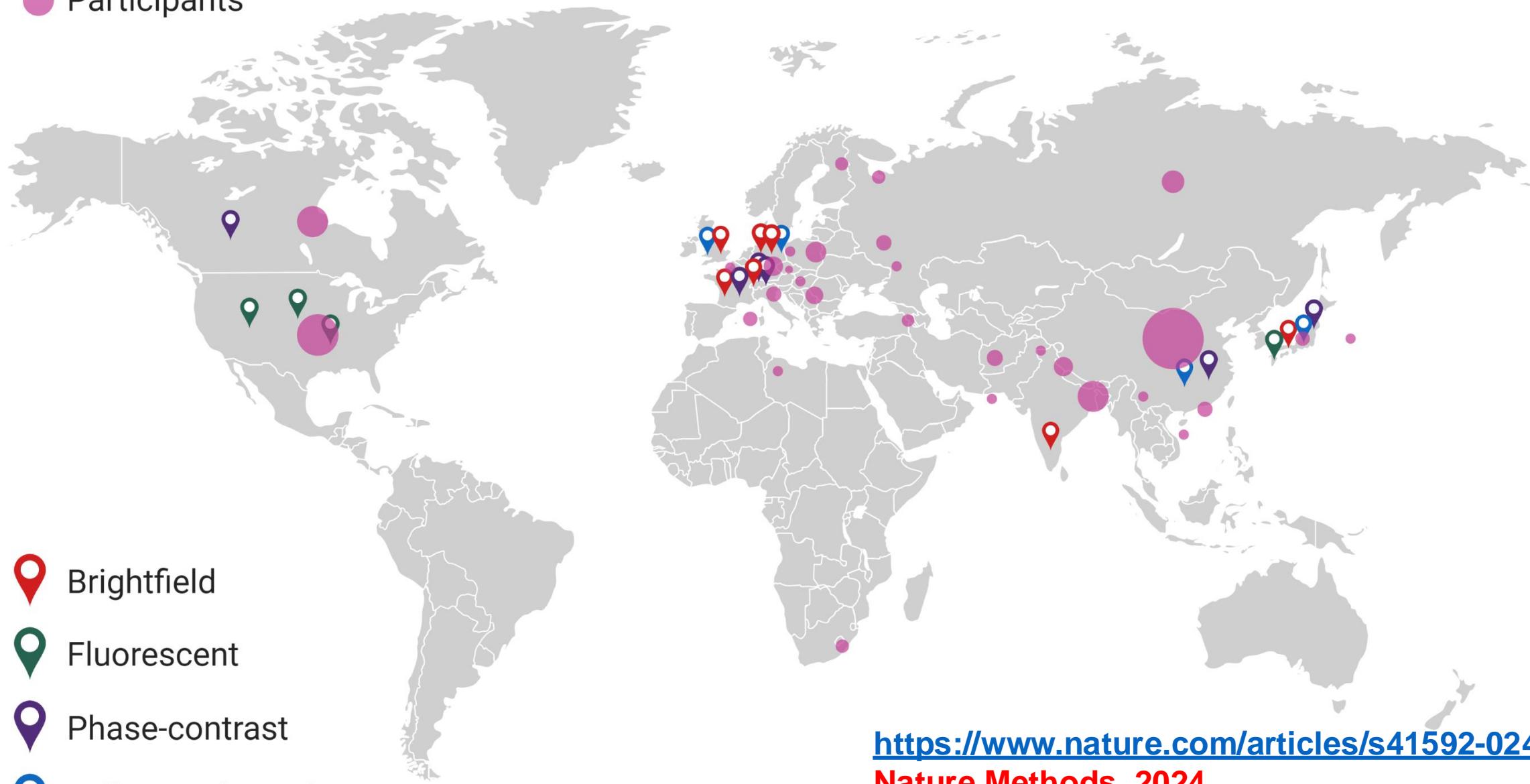
Our Dataset

- 50+ different experiments
- Four microscope types
- Training: 200K+ cells from 1000 labeled images and 1500+ unlabeled images
- Testing: 60K+ cells from 420 images

<https://www.nature.com/articles/s41592-024-02233-6>
Nature Methods, 2024

Dataset Features

● Participants

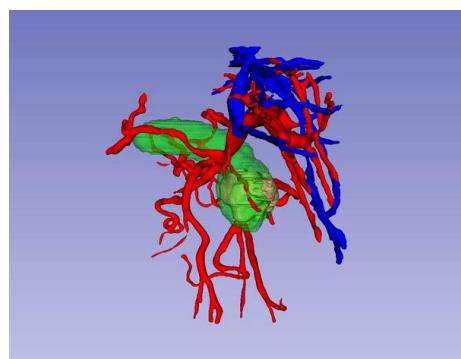
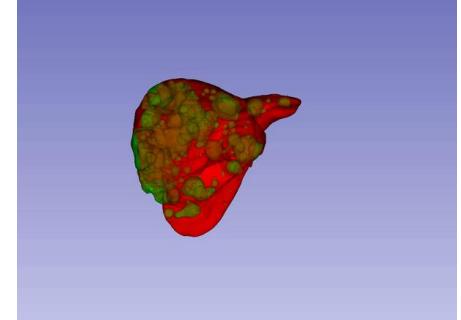
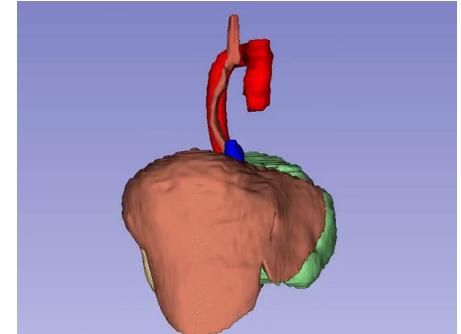
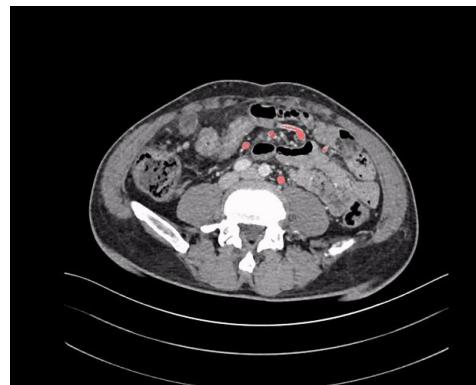
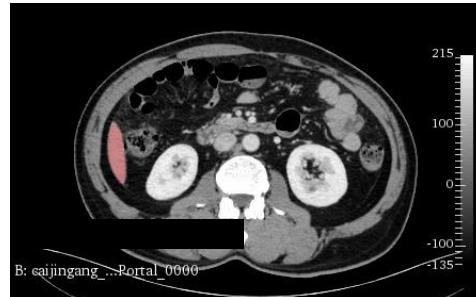


<https://www.nature.com/articles/s41592-024-02233-6>
Nature Methods, 2024

Segmentation Anything in Medical Images (MedSAM)

- Quantification of anatomical structures
- Diagnosis and disease progression
- Treatment planning ...

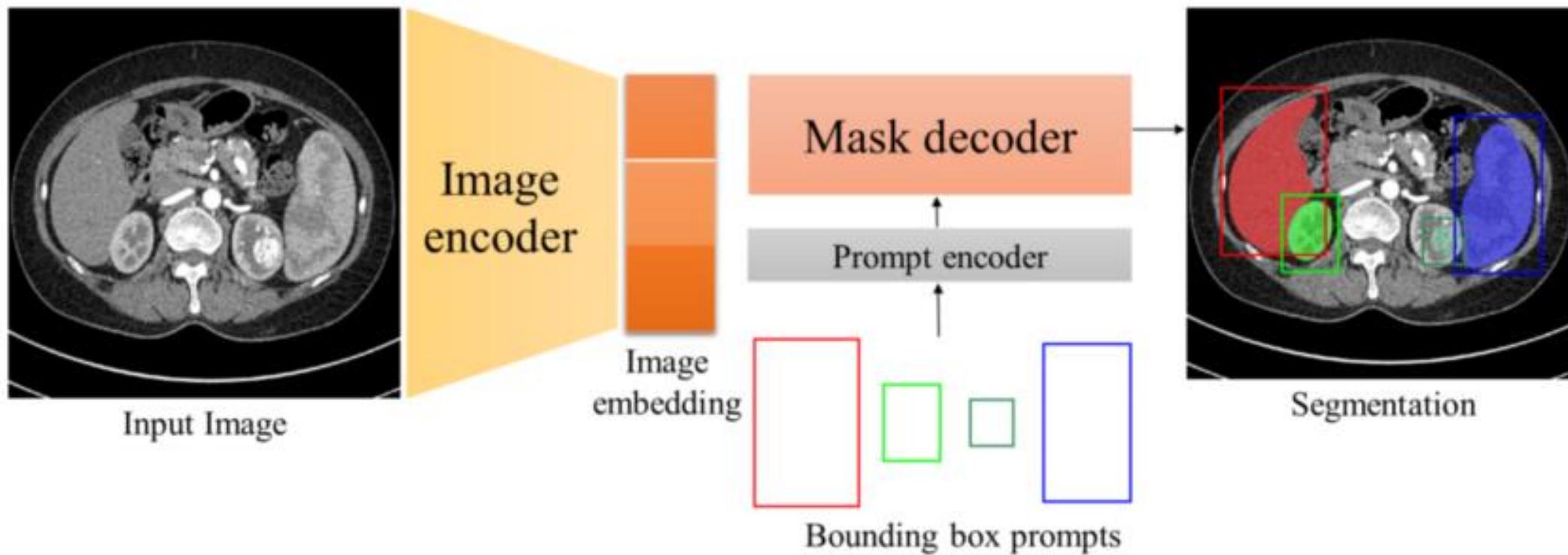
**Segmentation is the core technology
towards precise medical image
analysis!**



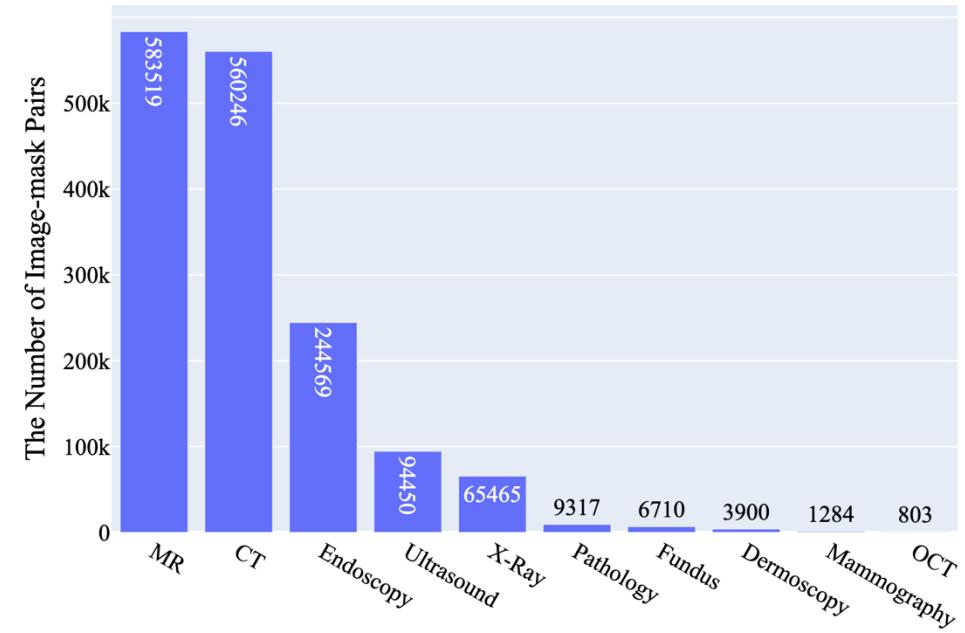
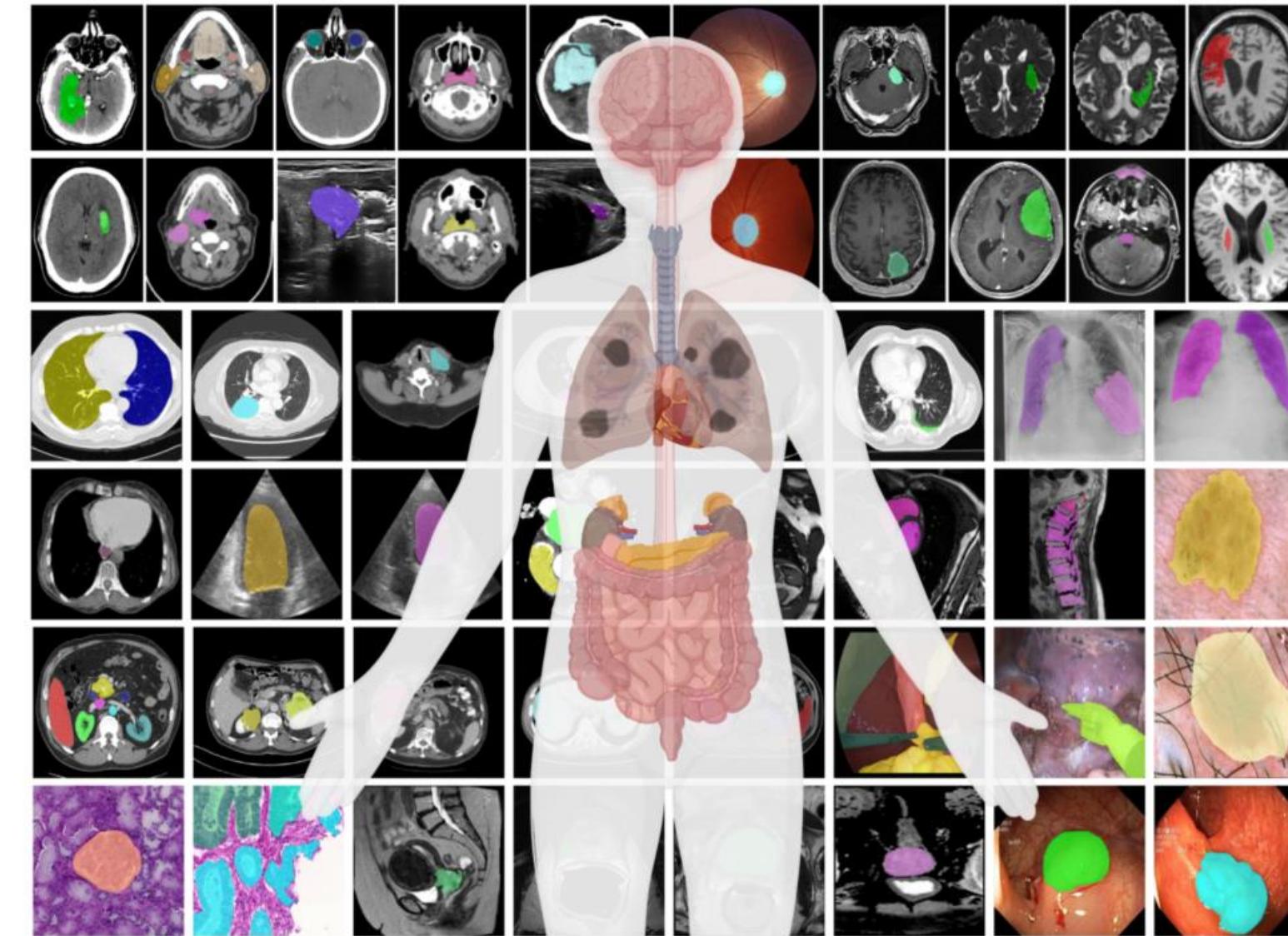
Segmentation Anything in Medical Images (MedSAM)

MedSAM: Pipeline

Fine-tune both image encoder and mask decoder



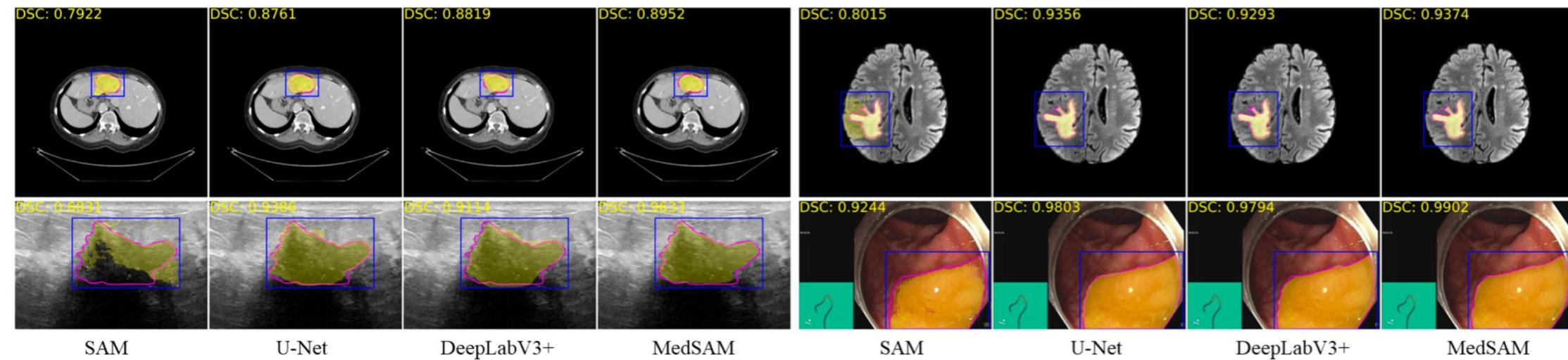
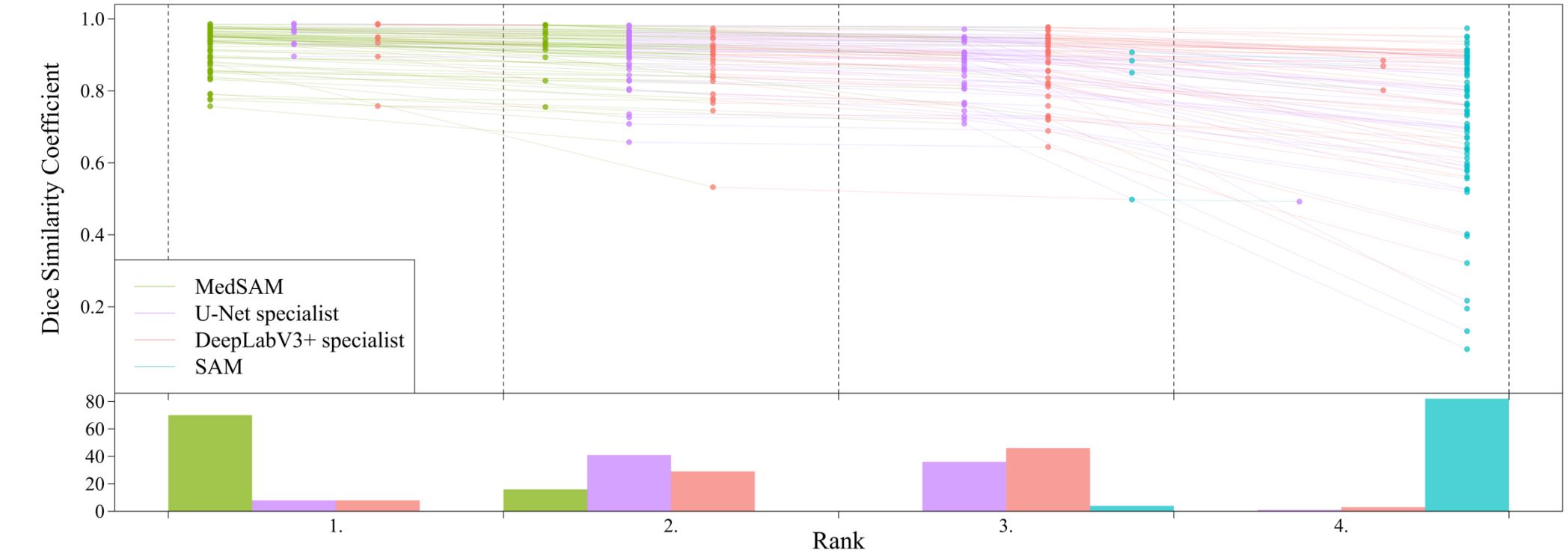
MedSAM: 1M image-mask Pairs



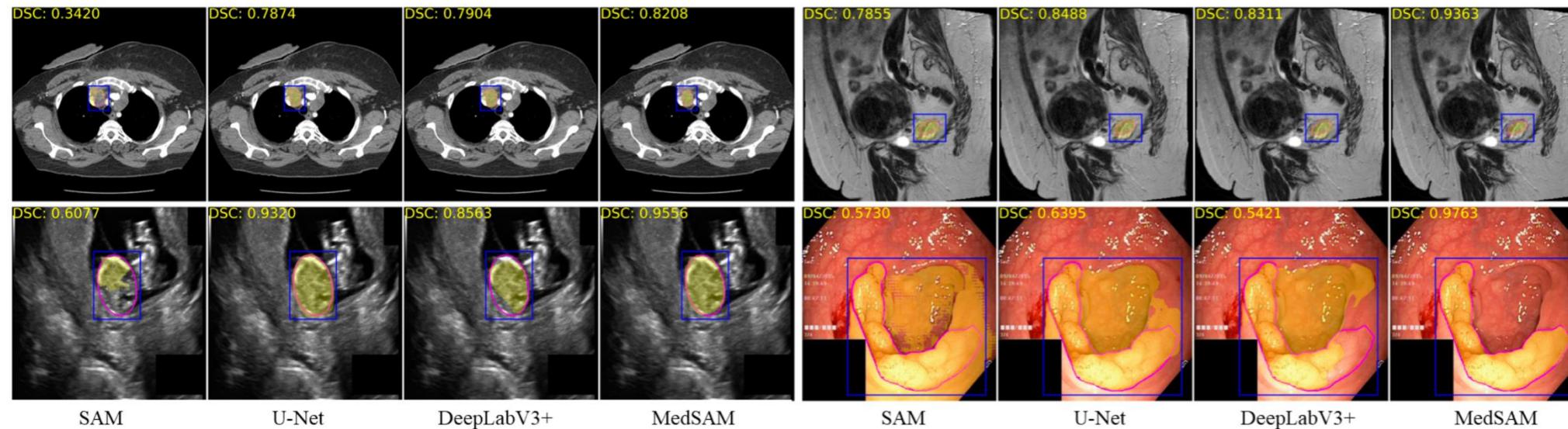
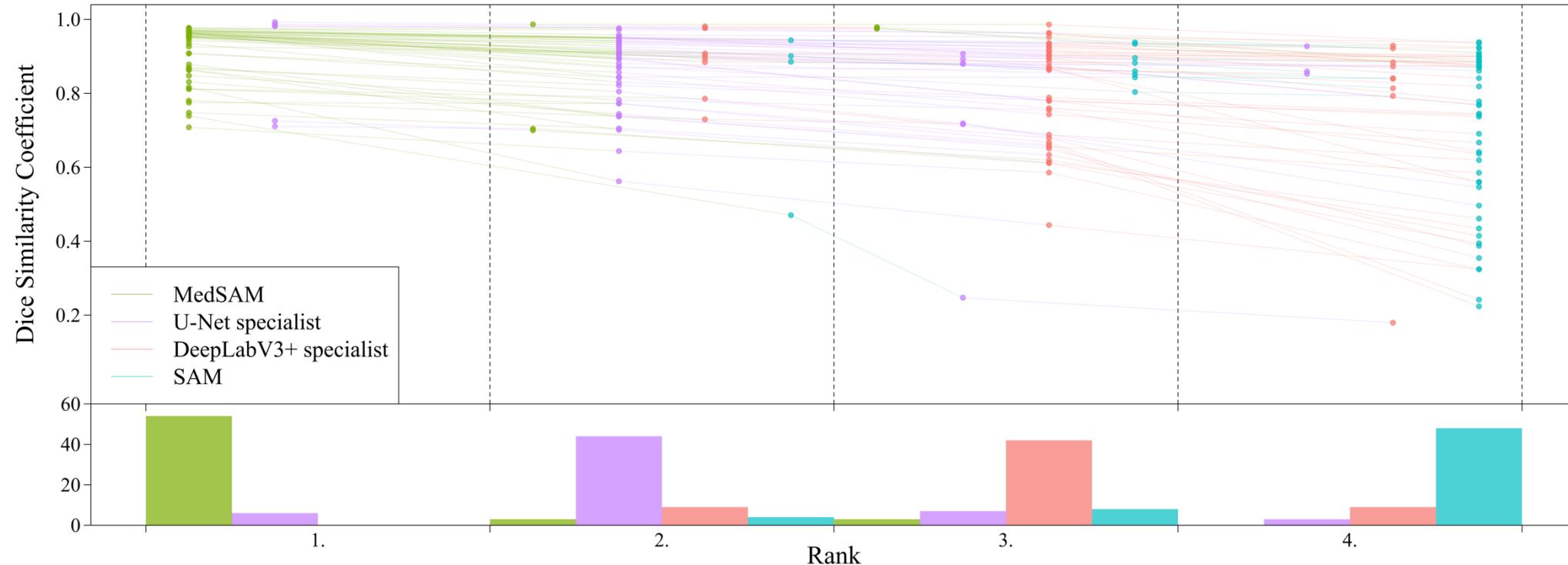
Experimental Settings

- 86 internal validation tasks
- 60 external validation tasks
- Compared to specialist U-Nets and DeepLabV3+ that are trained on each modality

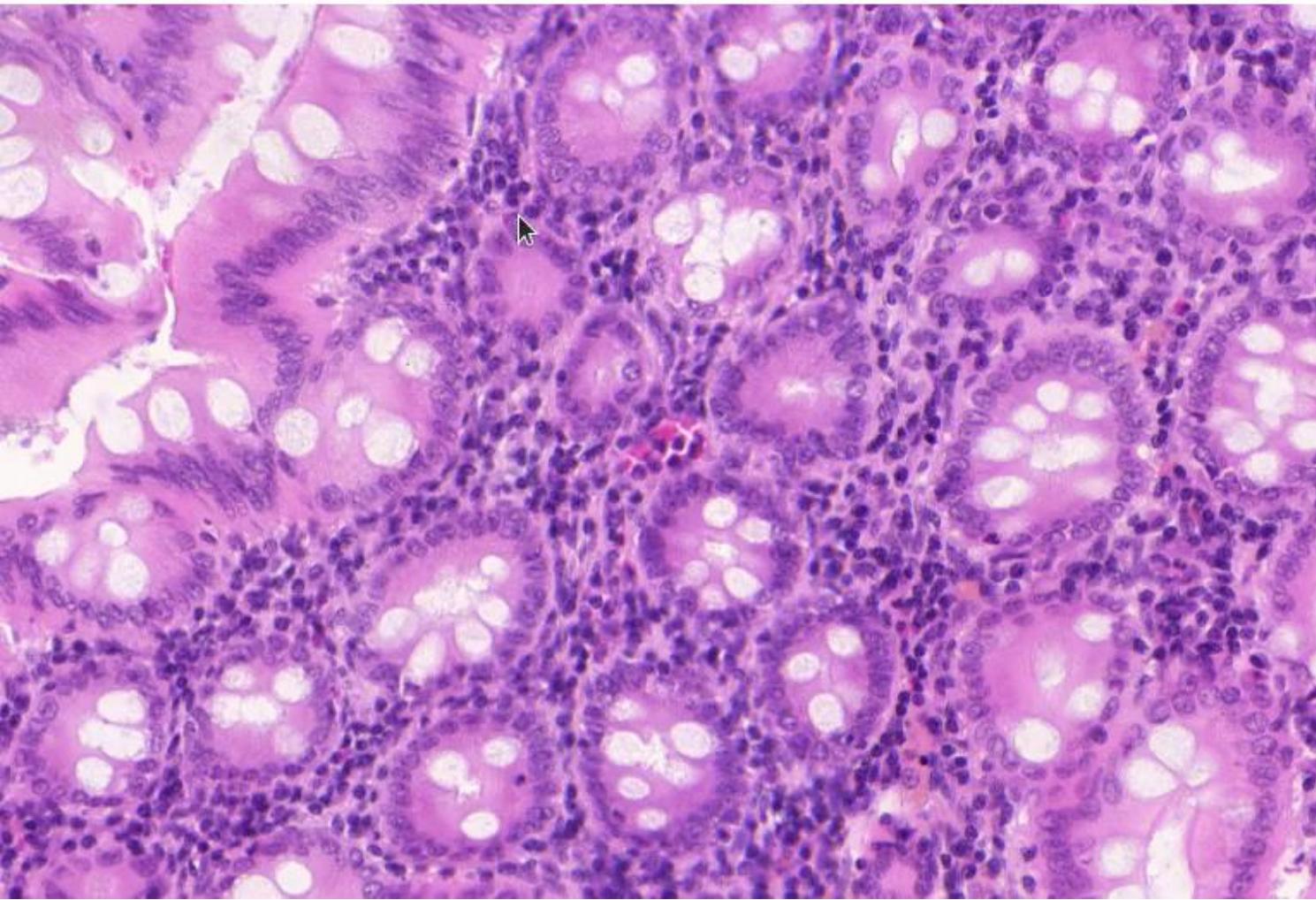
MedSAM: Internal Validation Results



MedSAM: External Validation Results



MedSAM: Demo

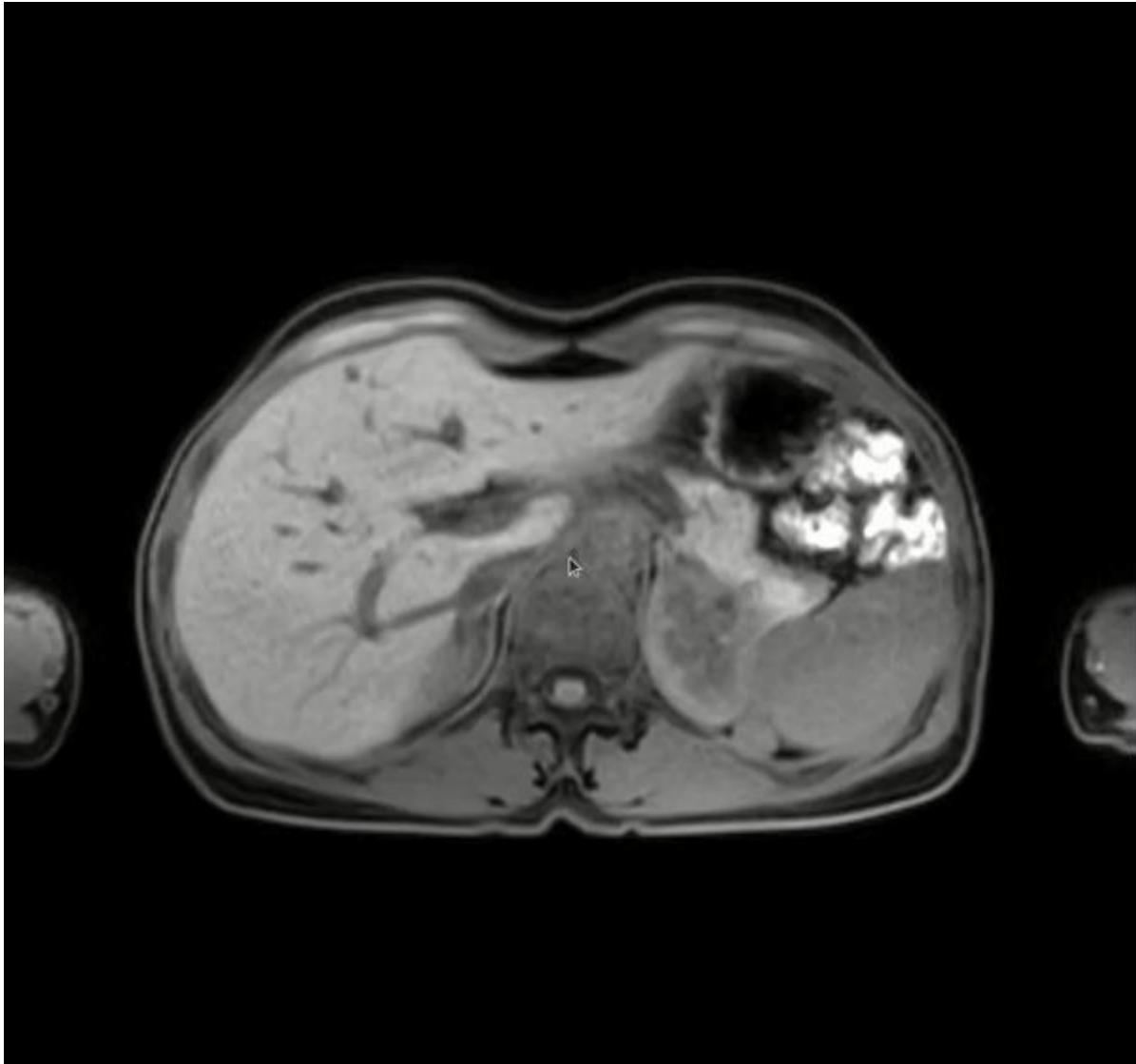


Gland Segmentation in Pathology Images



Liver and Tumor Segmentation in CT

MedSAM: Demo

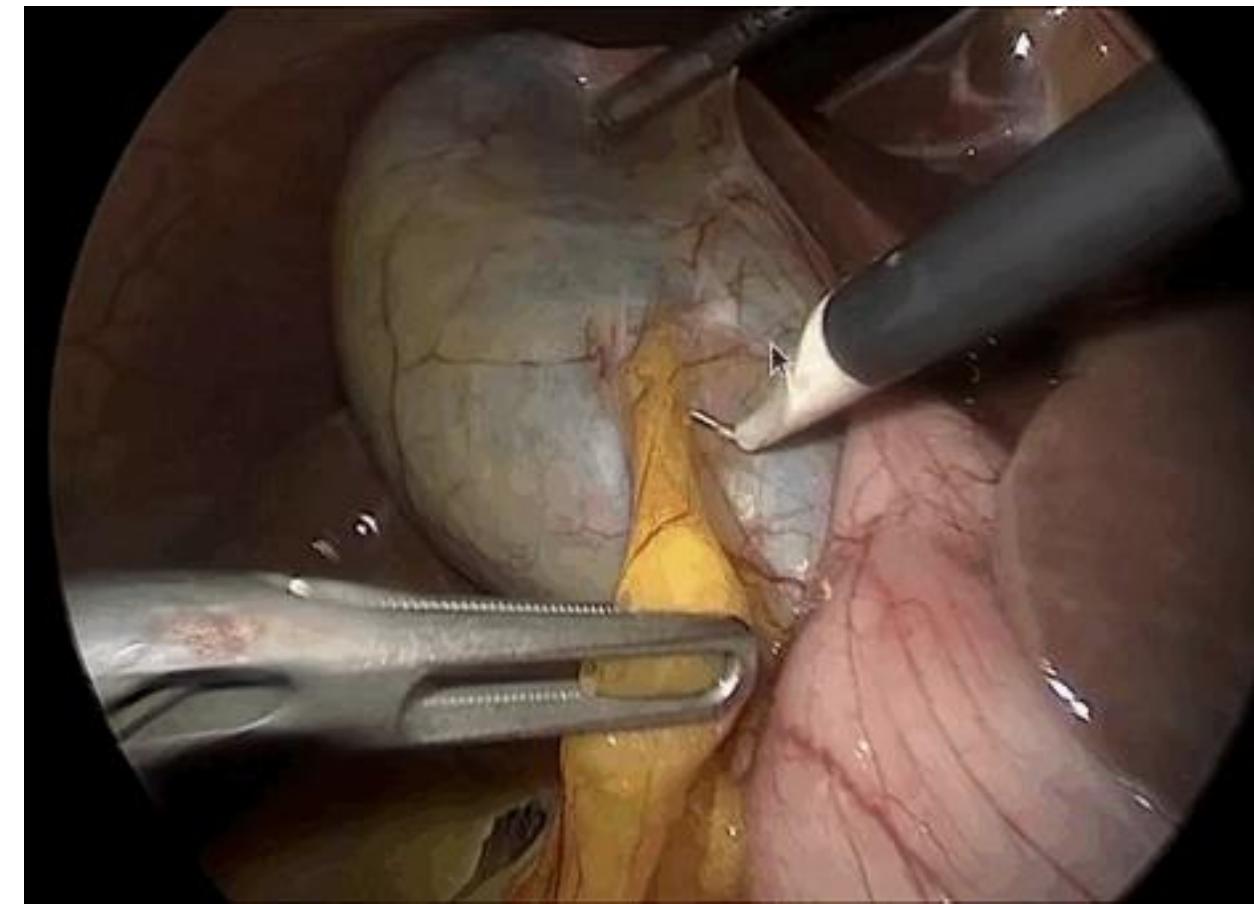


Abdominal Organ Segmentation in MR

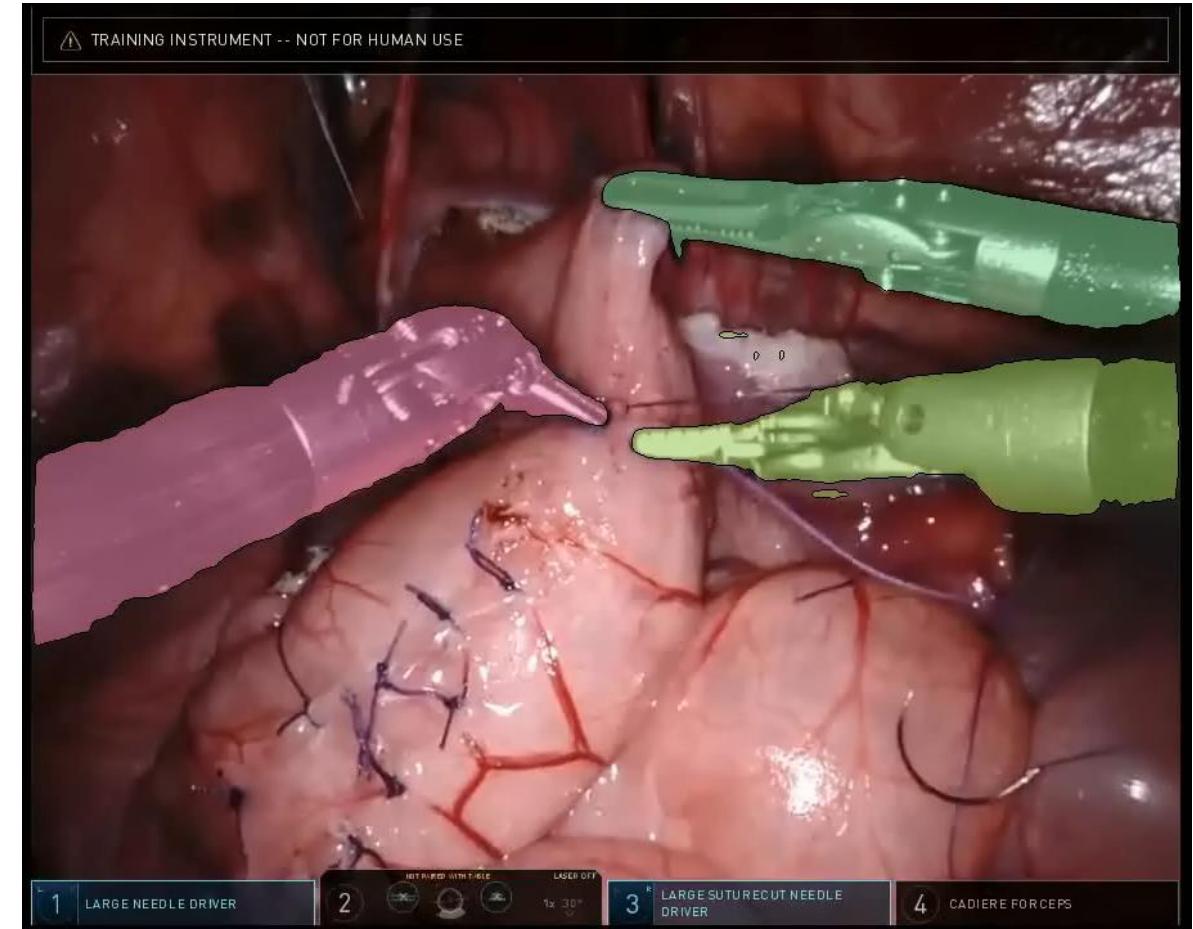


Lungs and Heart Segmentation in X-Ray

MedSAM: Demo

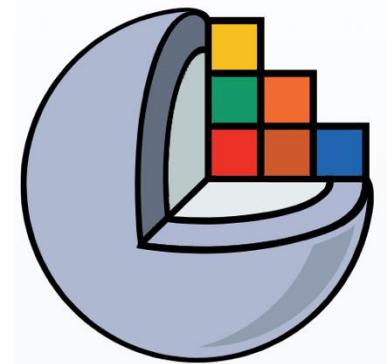
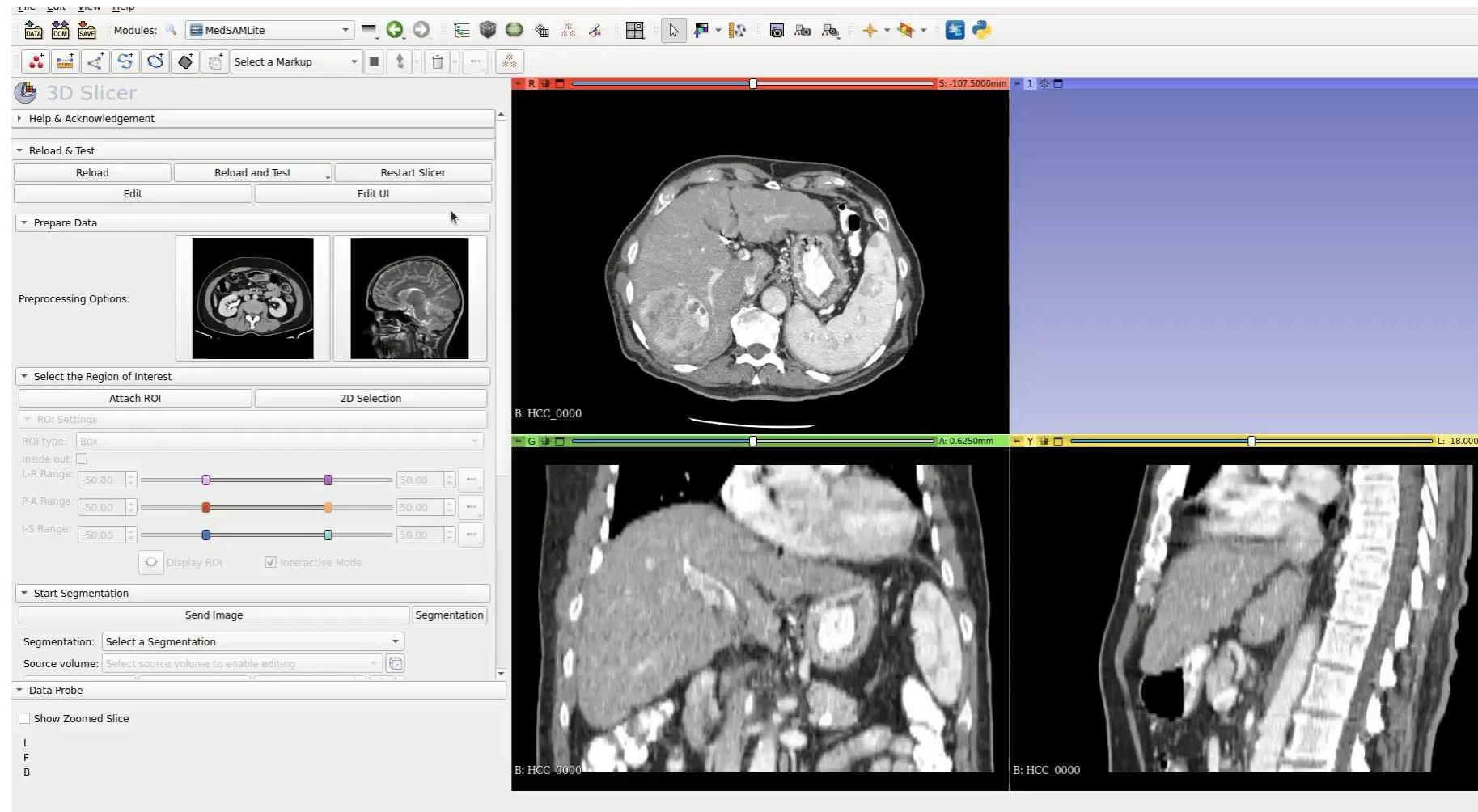


Tissue and Instruments Segmentation in Endoscopy Image



Instruments Segmentation and Tracking in Endoscopy Video

3D Slicer Integration (Open-source Platform)



<https://www.slicer.org/>

MedSAM in the Community

Google Scholar (**800+** citations)

nature communications

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature communications > articles > article

Article | Open access | Published: 22 January 2024

Segment anything in medical images

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You & Bo Wang ✉

MedSAM in Hugging Face

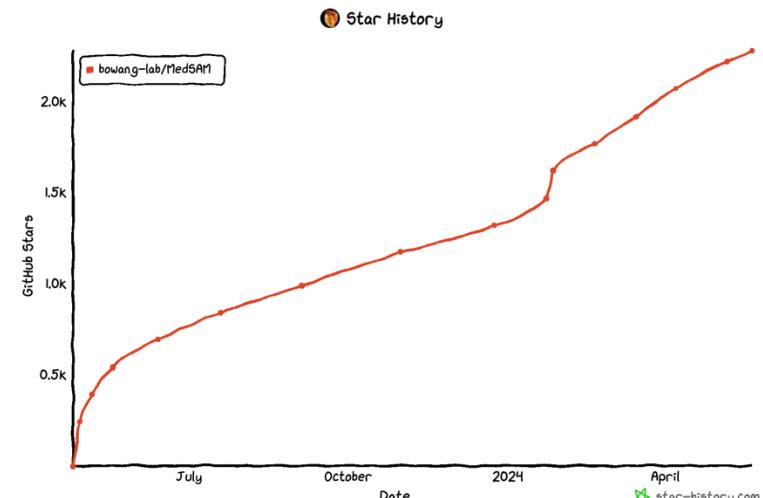
Segment medical images with MedSAM

In this notebook, we're going to perform inference with [MedSAM](#), a fine-tuned version of the SAM (segment-anything model) by Meta AI on the medical domain (thereby greatly improving its performance).

- [Original repo](#)
- [Hugging Face docs](#).

https://github.com/NielsRogge/Transformers-Tutorials/blob/master/SAM/Run_inference_with_MedSAM_using_HuggingFace_Transformers.ipynb

GitHub Stars (**3.0K+**)



<https://star-history.com/#bowang-lab/MedSAM&Date>

MedSAM in napari

I integrated MedSam into napari FYI #36

Closed Karol-G opened this issue on May 5 · 2 comments



Karol-G commented on May 5

Hey,

I just wanted to let you know that I integrated MedSam already into my Napari SAM plugin: <https://github.com/MIC-DKFZ/napari-sam>

So you can check the mark on "3D slicer and napari support" on your todo list if you want ;)

Best,
Karol



<https://github.com/MIC-DKFZ/napari-sam>

MedSAM in the Community

Interview by Computer Vision News

32 Nature Communications Paper

Segment Anything in Medical Images - MedSAM

Computer Vision News



Bo Wang (left) is a tenure-track Assistant Professor in the Departments of Computer Science and Laboratory Medicine & Pathobiology at the University of Toronto. He is the inaugural Temerty Professor in AI Research and Education in Medicine and the Chief AI scientist at the University Health Network, the largest research hospital in Canada. He also holds a CIFAR AI Chair at Vector Institute.

His research focuses on machine learning, computational biology, and computer vision, with a particular emphasis on their applications in biomedicine.



Jun Ma (right) is a Postdoctoral Fellow at the University of Toronto, Vector Institute, and University Health Network (UHN). He will be joining UHN AI Hub as a Machine Learning Lead. His research focuses on using advanced AI technologies to provide accurate and automatic cancer quantification, speed up diagnoses and personalize patient care.

Fresh from being published in the prestigious Nature Communications journal, they are both here to discuss their groundbreaking medical image segmentation method.

The top 50 papers in the field of AI and machine learning

nature communications

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature communications](#) > [focus](#)

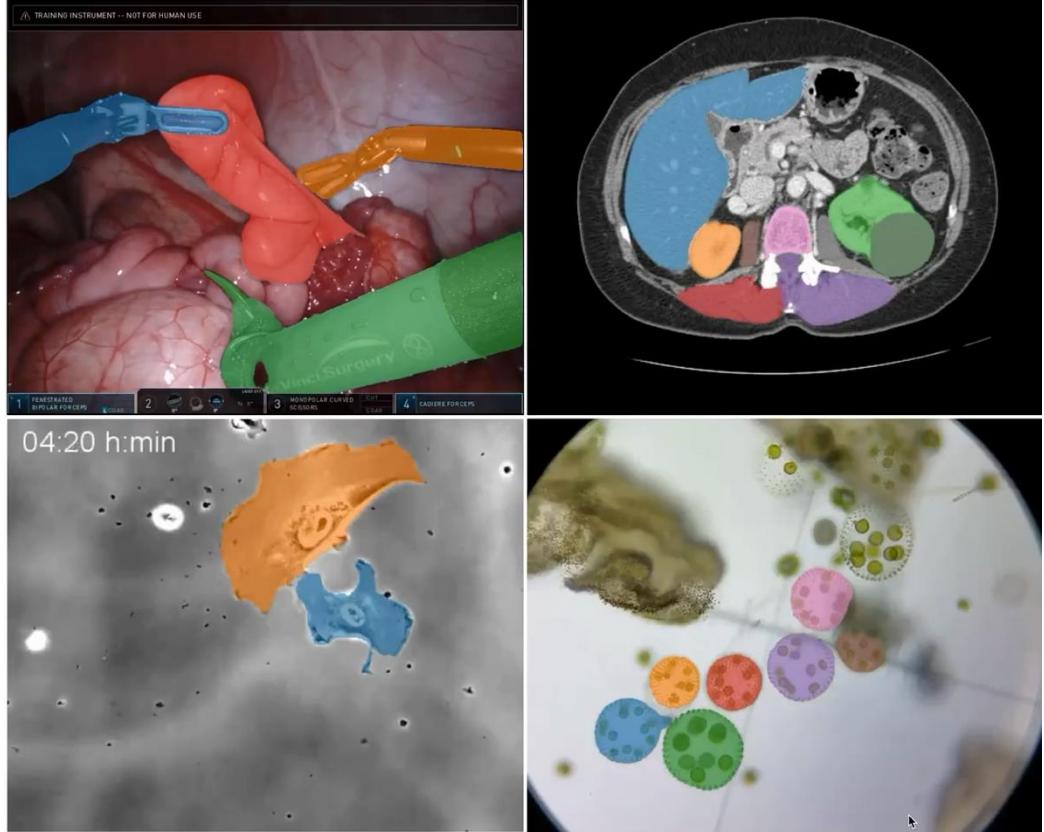
Focus | 26 January 2021

AI and machine learning

Computational advances have enabled the deployment of increasingly complex models, which are applied now to a broad-ranging set of fields. This editorial showcase aims at providing a snapshot of the current tools and challenges that are currently holding the promise to change lives in several ways. Herein, we also highlight research on the underlying pursuit of developing the concept of Artificial Intelligence.



MEDSAM2 for Medical Video Segmentation: Deployment



<https://github.com/bowang-lab/MedSAM/blob/MedSAM2/app.py>
Pay attention to package version: pip install gradio==3.38.0

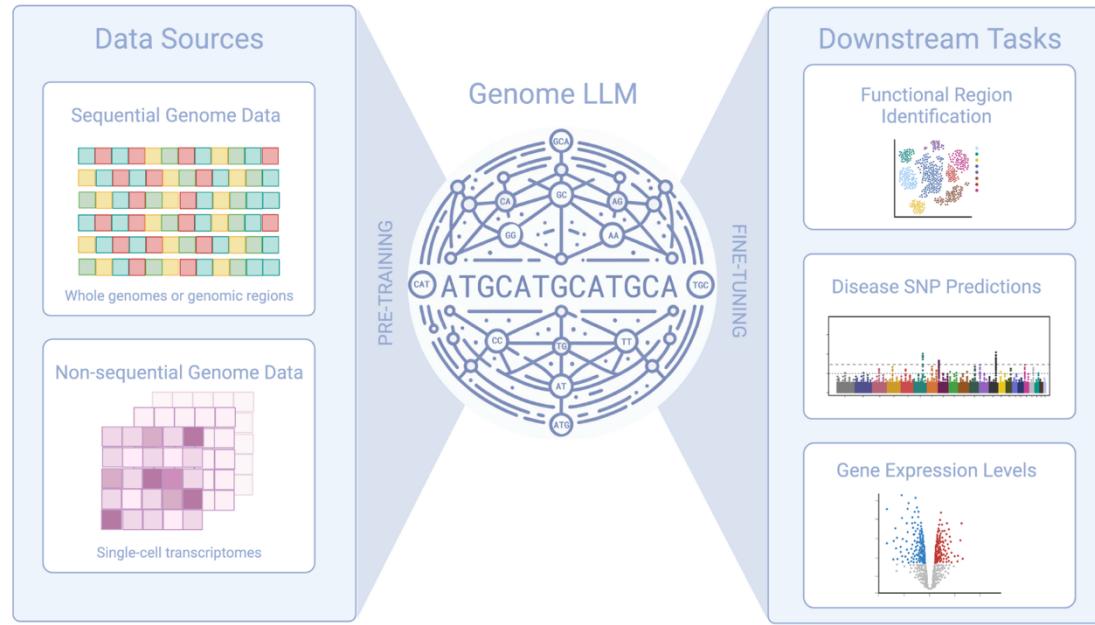
Key advantages:

- upload videos of any length
- multiple tracking objects
- support both box and points prompts ⁵⁰

<https://arxiv.org/abs/2408.03322>

<https://github.com/bowang-lab/MedSAMSlicer/tree/SAM2>

Interested in foundation models for genomics?



Micaela Consens



To Transformers and Beyond: Large Language Models for the Genome

<https://arxiv.org/abs/2311.07621>

Orthrus: Towards Evolutionary and Functional RNA Foundation Models

Philip Fradkin^{1, 2, *}, Ruian Shi^{1, 2, 3, *}, Keren Isaev^{4, 5,}

Brendan Frey^{1, 2, 6}, Quaid Morris^{3, ‡}, Leo J. Lee^{1, 6, ‡}, Bo Wang^{1, 2, 7, ‡}

1. Vector Institute



2. DCS, University of Toronto



3. CSBP, Sloan Kettering Institute



4. NY Genome Center



5. Systems Biology, Columbia



CE, University of Toronto



7. University Health Network



*Indicates Equal Contribution ‡Indicates Equal Advising



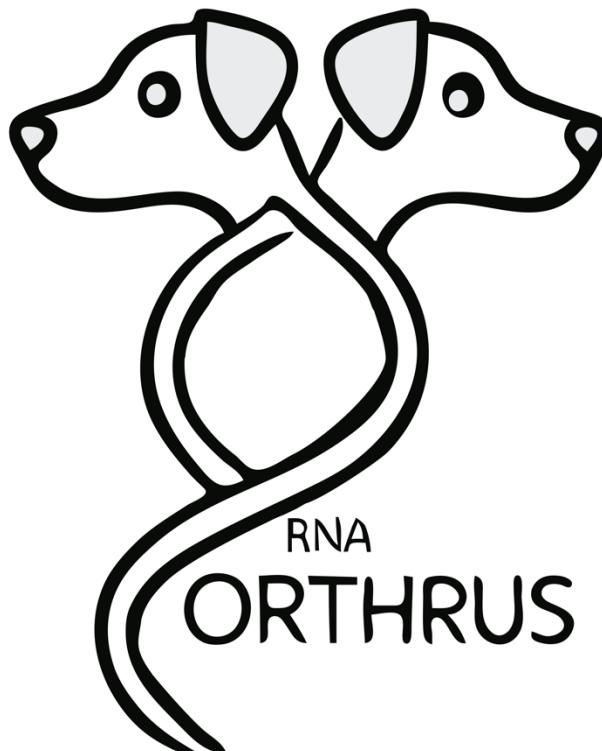
Hugging Face



Code



Rx bioRxiv



Philip Fradkin

<https://www.biorxiv.org/content/10.1101/2024.10.10.617658v1>

Foundation Models with *Care!*

1. How to ***curate*** the large-scale pre-training data beyond scRNAseq?
2. How to ***evaluate*** various foundation models?
3. How to make sure foundation models are ***open and transparent?***
4. How to ***democratize*** the foundation models?

Acknowledgements



Haotian Cui



Chloe Wang



Conor Klamann



Ronald Xie



Sonya MacParland



Gary Bader



Jun Ma

scGPT

Paper: <https://www.biorxiv.org/content/10.1101/2023.04.30.538439v2>

Code: <https://github.com/bowang-lab/scGPT>

BLEEP

Paper: <https://arxiv.org/abs/2306.01859>

Code: <https://github.com/bowang-lab/BLEEP>

MEDSAM

Paper: <https://arxiv.org/abs/2304.12306>

Code: <https://github.com/bowang-lab/MedSAM>



VECTOR INSTITUTE

