Under review

TissueFormer: Extending single-cell foundation models to predict population-level phenotypes

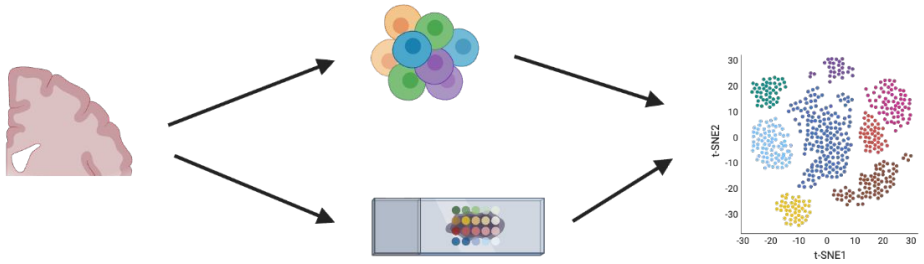
Ari S. Benjamin[1] **and Anthony Zador**[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, United States

Corresponding author and email address: Ari Benjamin - benjami@cshl.edu

Janie Oberhauser & Bivas Nag
5/19/26

Intro to single cell and spatial transcriptomics



Single Cell	Both	Spatial
> Cells/nuclei dissociated from tissue context > Whole-genome options > Higher throughput, cheaper > Many analysis tools	> Provide insights into changes in gene expression, cell-cell interactions, and cell type composition* driven by development, disease, regional variability, etc.	> Probe-based > Tissue morphology and physiological context maintained > More expensive > Less sophisticated analysis tools

Mapping tools: STEM, CytoSPACE, Cell2Spatial, NicheFormer, etc.

What kinds of questions can we use transcriptomic data to answer?

1. How does cellular composition change across physiological conditions, individuals, or disease states?
2. How does gene expression within cell types change across development, aging, and disease?
3. Which cell types drive specific biological processes or phenomena?
4. For spatial data: How are cell types organized within tissues, and how do they interact locally?
5. Can cellular and transcriptional signatures predict clinical outcomes or distinguish patient subgroups?

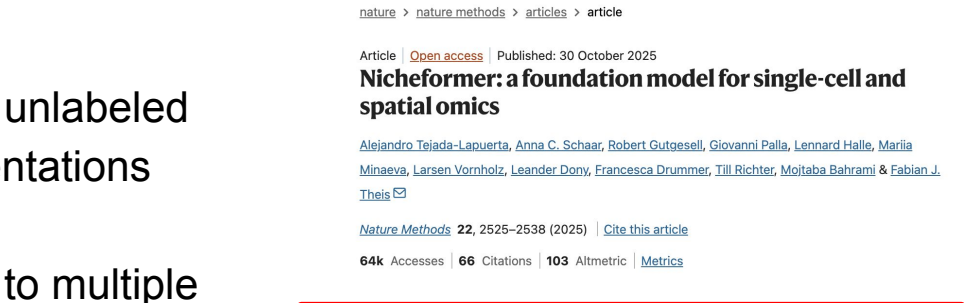
Transcriptomic data and foundation models

A foundation model requires:

1. Self-supervised pre-training on large, unlabeled datasets (learns patterns and representations without labeling)
2. After training, weights can be applied to multiple tasks without restarting training from scratch

Non-biology examples: GPT and BERT

“Self-attention”: ML awareness of how much an individual word, datapoint, etc. relates to its context.



nature › nature methods › articles › article

Article | Open access | Published: 30 October 2025

Nicheformer: a foundation model for single-cell and spatial omics

Alejandro Tejada-Lapuerta, Anna C. Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, Mojtaba Bahrami & Fabian J. Theis

Nature Methods **22**, 2525–2538 (2025) | Cite this article

64k Accesses | **66** Citations | **103** Altmetric | Metrics

nature › articles › article

Article | Published: 31 May 2023

Transfer learning enables predictions in network biology

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu & Patrick T. Ellinor

Nature **618**, 616–624 (2023) | Cite this article

202k Accesses | **1038** Citations | **587** Altmetric | Metrics

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow this preprint

scGPT-spatial: Continual Pretraining of Single-Cell Foundation Model for Spatial Transcriptomics

Chloe Wang, Haotian Cui, Andrew Zhang, Ronald Xie, Hani Goodarzi, Bo Wang

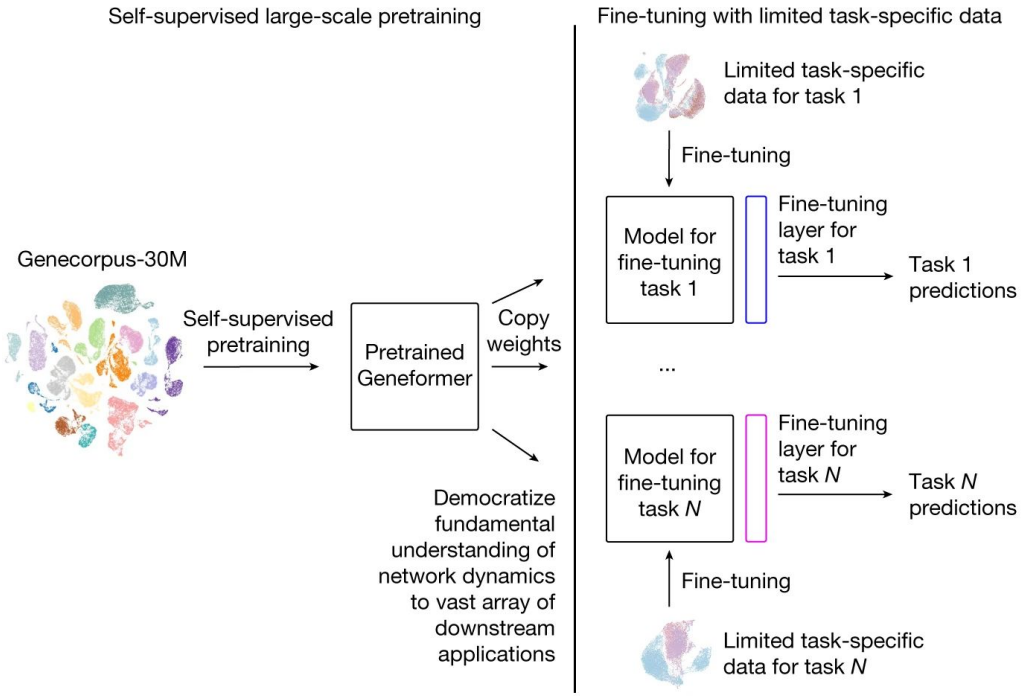
doi: https://doi.org/10.1101/2025.02.05.636714

This article is a preprint and has not been certified by peer review [what does this mean?].

Transfer learning with GeneFormer: TissueFormer's first building block

Transfer learning = applying previously acquired knowledge to new contexts

GeneFormer* is a "context-aware, attention-based deep learning model, Geneformer, pretrained on large-scale transcriptomic data to enable predictions in settings with limited data."



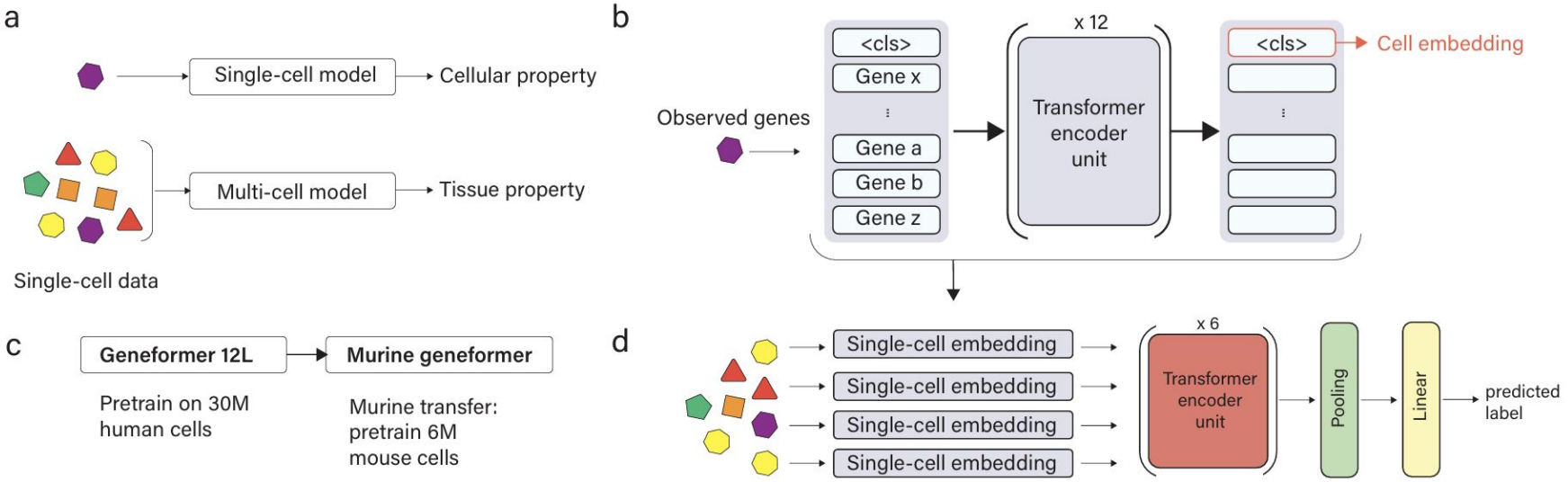
*Development led by Christina Theodoris before she started her lab here at UCSF!

How are GeneFormer and TissueFormer similar/different?

GeneFormer	Both	TissueFormer
> **Input:** ranked gene list > Works on single cells > Self-supervised pre-training > No need for labeled training data > **Architecture:** 12-layer BERT-style transformer > **Output:** cell-level info (cell state, gene dosage sensitivity, etc.) > Broadly reusable > **Limitation:** Cannot capture population-level information	> Use rank-based gene expression inputs > Trained on mouse/human transcriptomic data > Classifier	> **Input:** multiple ranked gene lists > Works on cell populations > No pre-training; uses GeneFormer's pre-set weights > Requires labeled training data > **Architecture:** Geneformer + 6 transformer layers + linear classifier > **Output:** sample-level info (disease severity, brain region, etc.) Reuse requires targeted retraining > **Limitation:** Requires large labeled datasets, limited generalizability

TissueFormer's overall architecture

Figure 1



Major points the authors hope to address with their model

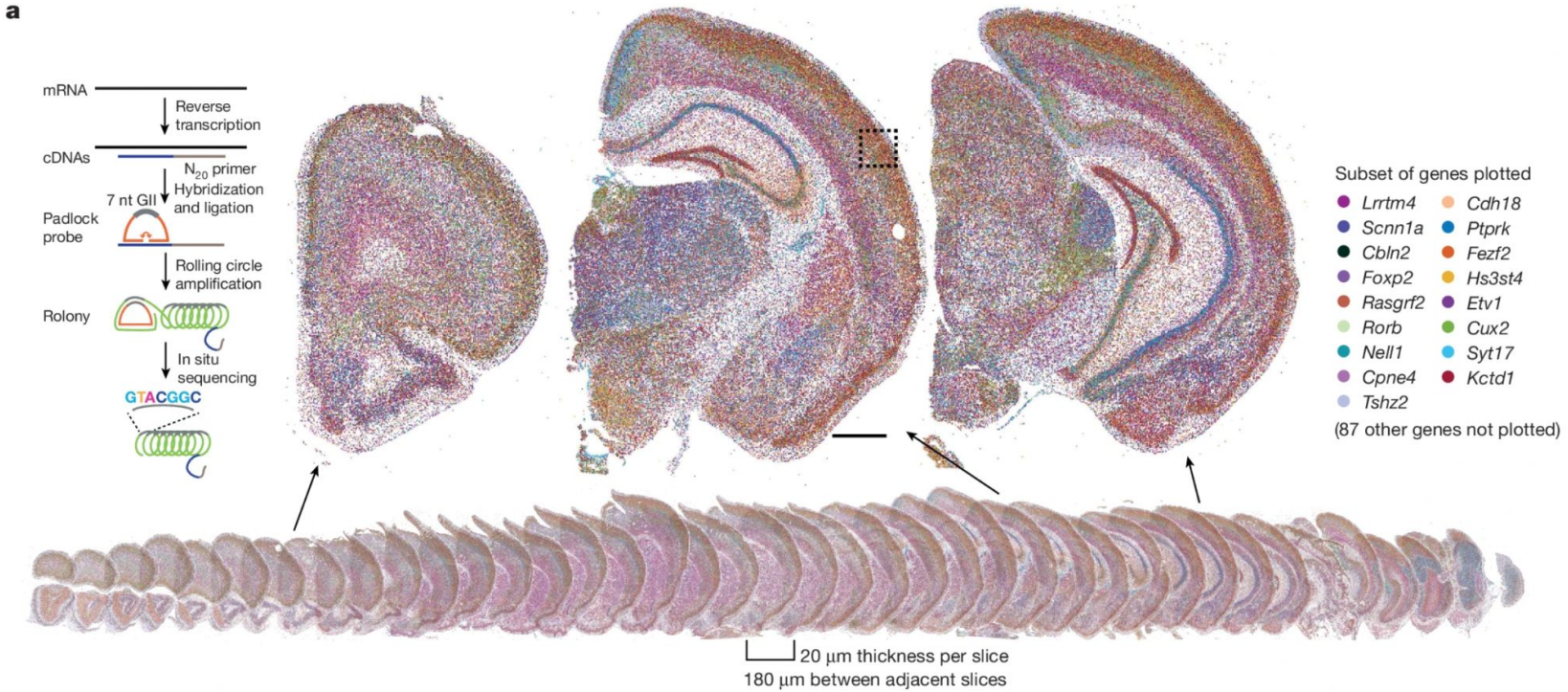
1. Mapping cortical areas using transcriptomic data
    a. Natural inter-individual variability in cell type composition/brain structure.
2. Identifying disease severity using model predictions

Important: Cell aggregation through dimensionality reduction using (pseudobulk) or summary statistic-based methods lose important information on cell-cell variability.

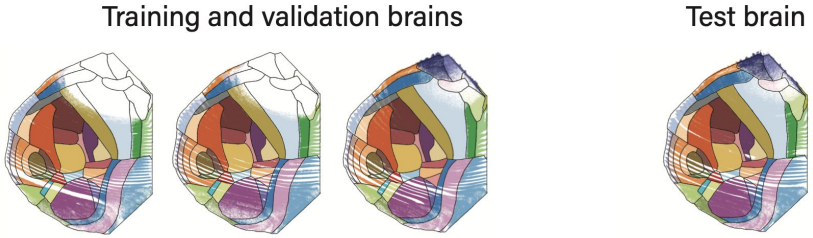
Need access to full transcriptional info for each cell at a population-wide level

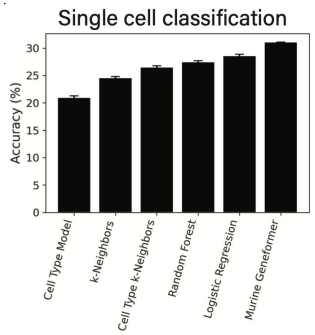
TissueFormer training dataset

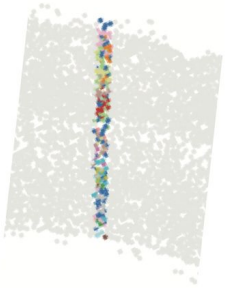
Chen et al., 2024 mouse “Whole-cortex in situ sequencing” with BAR-seq



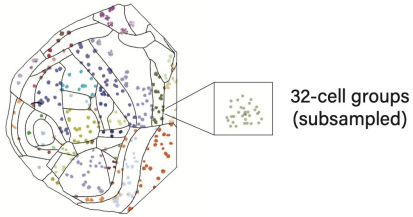
Can we figure out which cortical area a piece of brain tissue belongs to, just from gene expression?

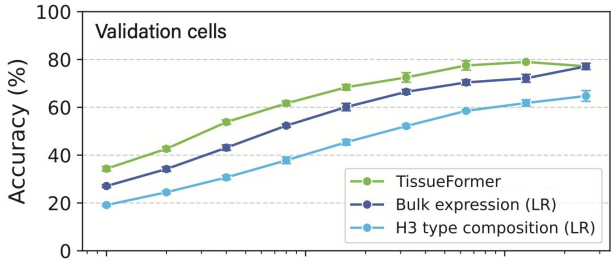






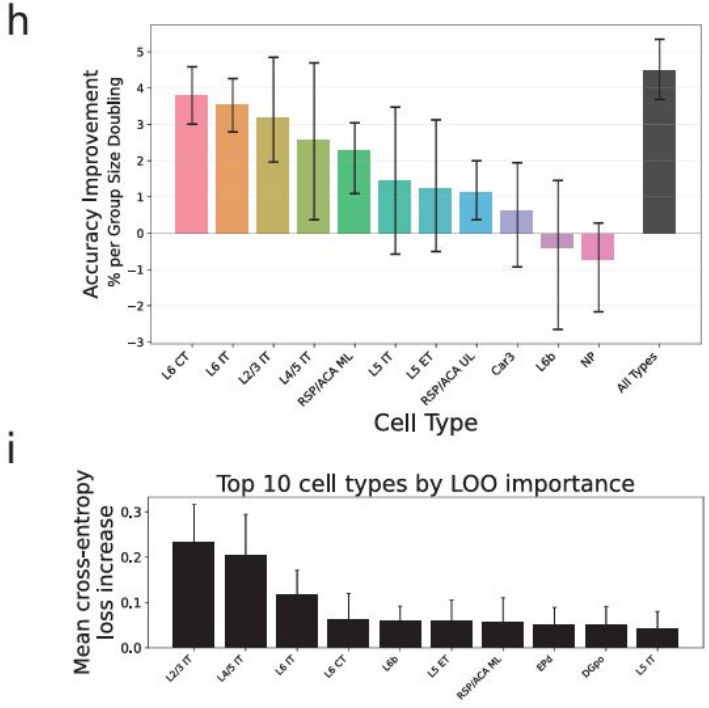
Group cells in columns



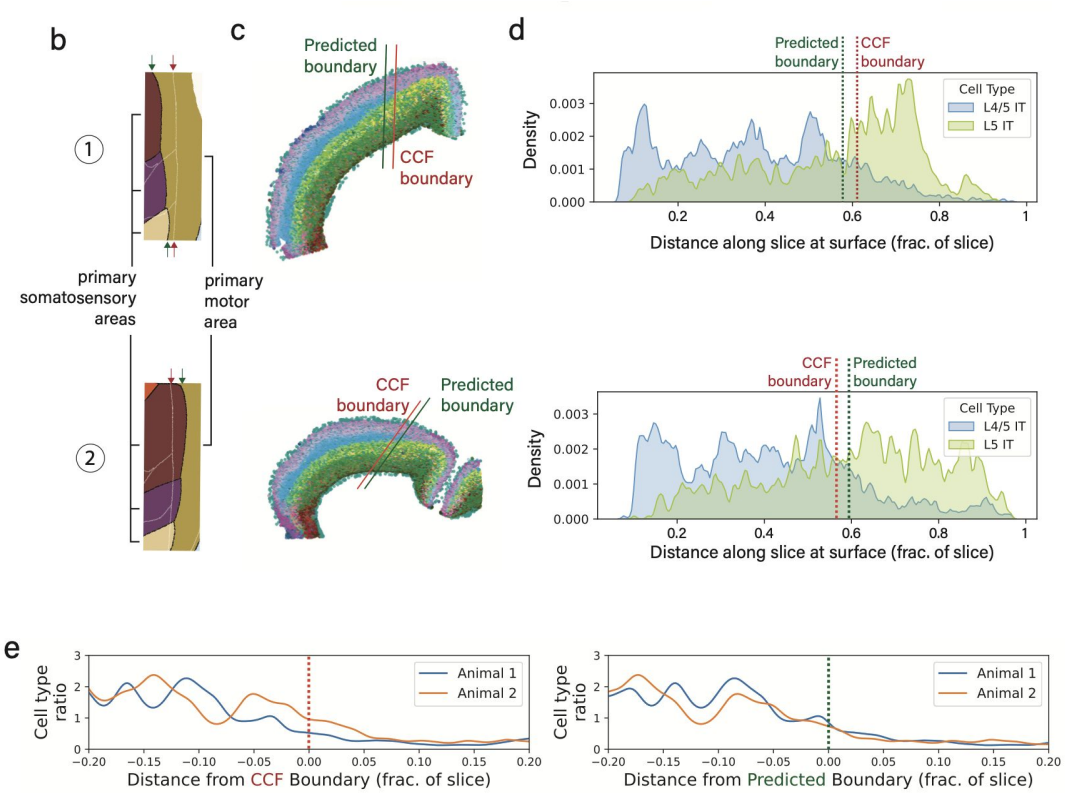


Accuracy increased to 80%
with increased sample size

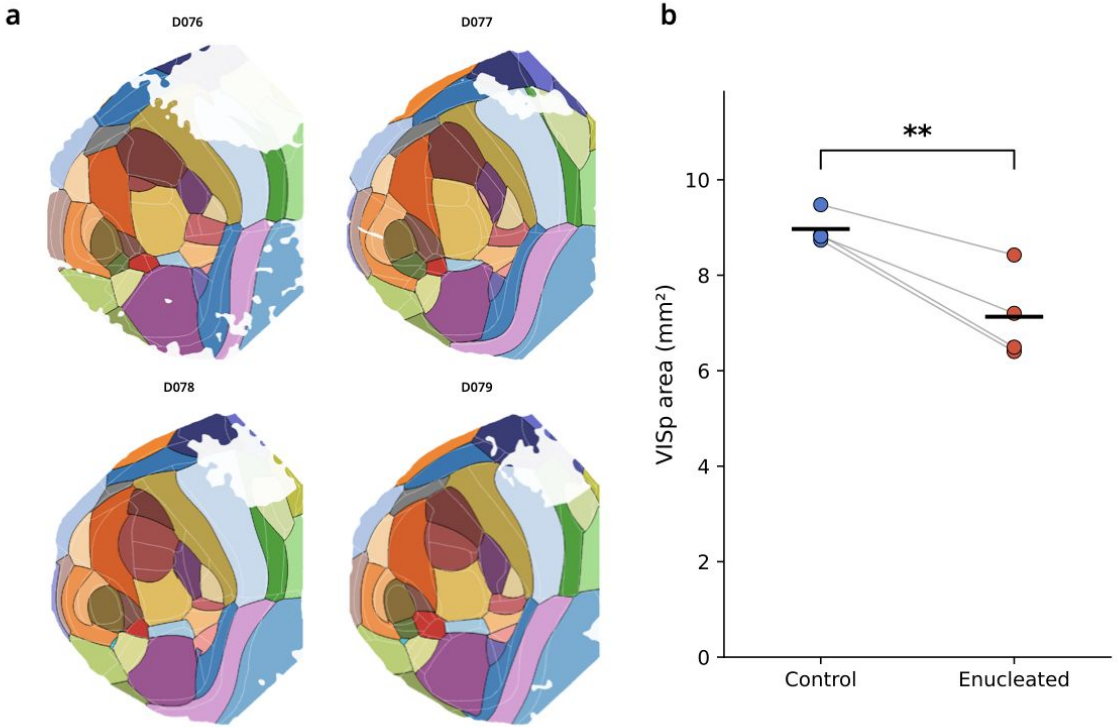
“Leave-one-cell-type-out” analysis reveals most important cell types for accurate TissueFormer area prediction



Predicted boundaries match real biology better than the standard atlas



TissueFormer detects a smaller visual cortex in mice raised without vision



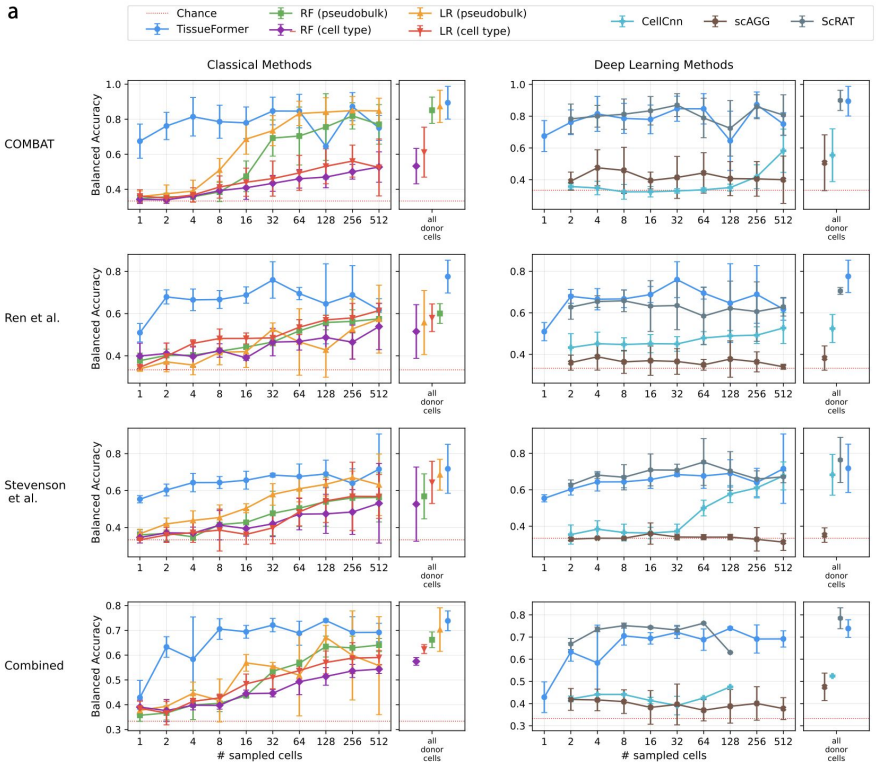
COVID testing data

CellXGene data from three independent datasets:

1. COMBAT Consortium dataset (637,266 cells from 85 donors)
    a. 10 healthy, 29 mild COVID-19, and 46 severe COVID-19
    b. 10x 5’ v1
2. Ren et al., 2021 (1,456,806 cells from 185 donors)
    a. 25 healthy, 77 mild COVID-19, and 83 severe COVID-19
    b. 10x 3’ v2 and v3
3. Stephenson et al., 2021 (585,153 cells from 106 donors)
    a. 23 healthy, 55 mild COVID-19, and 28 severe COVID-19
    b. 10x 3’

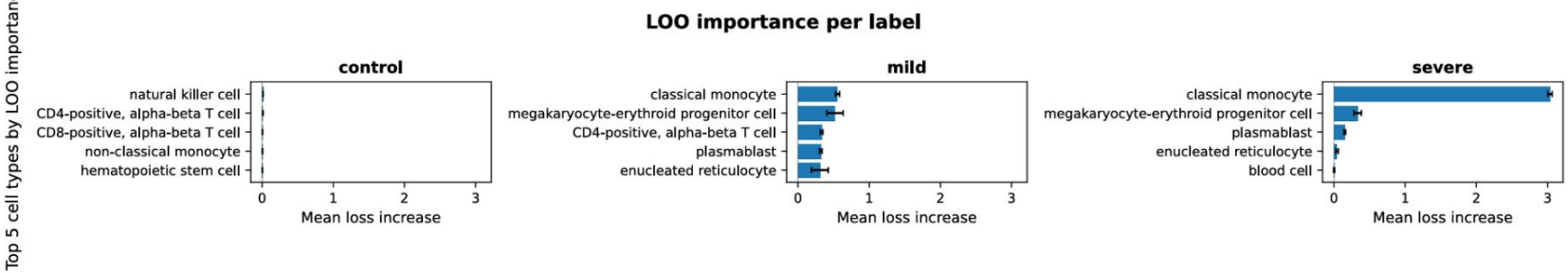
Does it work beyond the brain? Predicting COVID severity from human blood samples as a “generalizability” measure

- Human patients, not mice
- Blood sample, not brain tissue
- No spatial coordinates
- Label is a disease state: healthy, mild, or severe



“Leave-one-cell-type-out” analysis reveals cell type drivers of COVID severity

b



Major takeaways from this preprint

1. Modeling grouped cells outperforms modeling individual cells for tissue and sample-level predictions.
2. TissueFormer's cortical mapping captures inter-individual variability lost with direct brain atlas mapping.
3. TissueFormer performs slightly better than other computational tools for predicting sample-level phenotypes from groups of single-cell gene expression profiles.
4. Overall sample/tissue state can be predicted directly from the transcriptional profiles of a random subset of cells.

Limitations of the study

- Unclear practical value over existing tools and atlases
- As presented, the authors do not use TissueFormer to its full potential for identifying biological insights (i.e. going beyond cell types driving brain region or disease severity predictions)
- Technical limitations:
  - No comparison to atlases in Figure 4
  - What is the value of pretraining?
  - Given need for large training dataset, how could this tool be generally applied to smaller datasets or poorly characterized diseases, organisms, tissues, etc. without pre-labeled

Limitations of the study - the value of pre-training?