How AI is reshaping computational biology

Part I. Pre-training for biological research: What does a biological foundation model actually learn?

EMBL Visualising Life Summer School

Maren Hackenberg

22 July 2026

EMBL

AI is changing how we do science

Reliable results: Discerning high-quality from low-quality analysis will be a major challenge in the era of AI-produced science.
ILLUSTRATION BY MAKSYM FILIPENKO

PERSPECTIVES / **ARTIFICIAL INTELLIGENCE**

Add us as a Preferred Source on Google

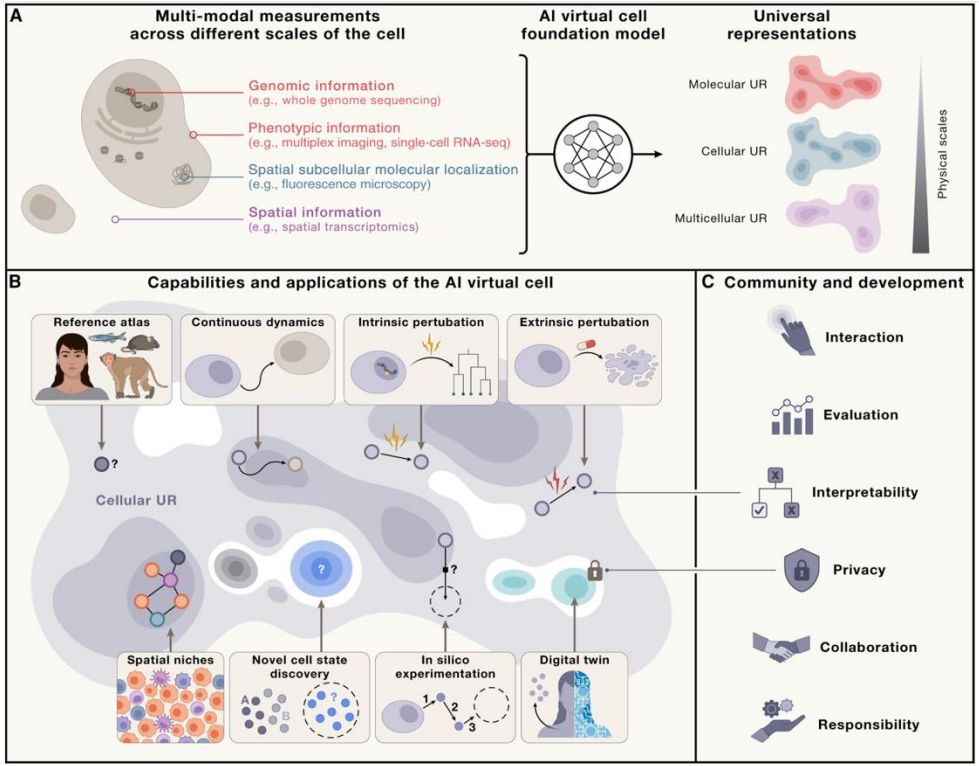
Mass-produced science is coming. What happens to scientists?

Artificial intelligence may soon enable researchers to generate high-quality science at a previously unimaginable speed. For science consumers—the public, medical patients, technology users—the likely effects will be positive. For scientists, the effects will be as disruptive as industrial mass production was for artisan manufacturers.

BY **KENNETH HARRIS**
9 JULY 2026 | 9 MIN READ

EMBL

A vision for computational biology: the virtual cell



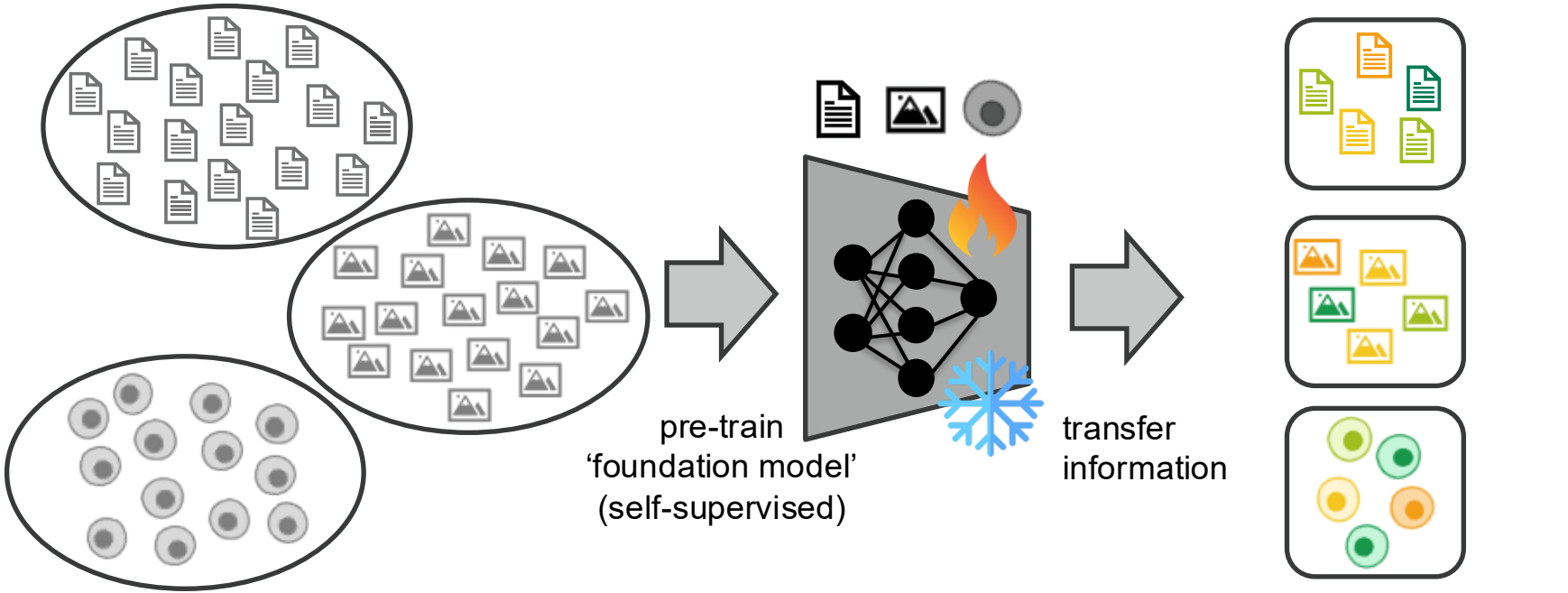
Bunne et al. Cell 2024

Multiple levels:

- **Representations**
  Pre-train models to turn large datasets into reusable embeddings
- **Predictions**
  Use those representations to predict, e.g., cell types, perturbation responses, spatial interaction patterns,…
- **Scientific workflows**
  Language models and scientific agents can help automate analysis workflows or hypothesis generation

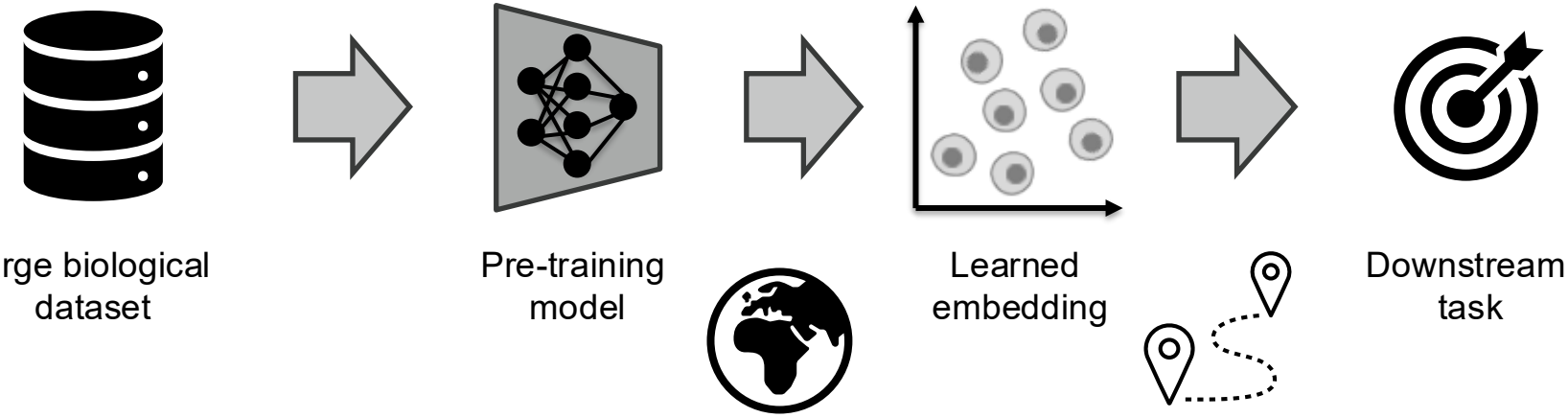
EMBL

What is pre-training and what is the point?



EMBL

How does it work conceptually?



Large biological dataset

Pre-training model

Learned embedding

Downstream task

1. Which organisms, modalities, technologies, conditions,… were observed?

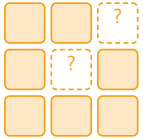
2. What was the model asked to do during pre-training?

3. What question do we want to answer using the model?

EMBL

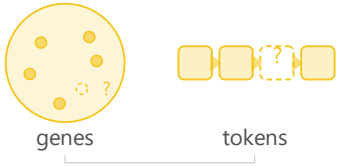
Common pre-training objectives

Language:
predict the next or masked word

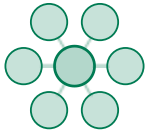


Images:
reconstruct masked patches

Protein sequences:
predict masked amino acids



Single-cell gene expression:
reconstruct masked genes
(random or in an ordered sequence)



Spatial omics:
reconstruct based on neighbourhood context

Biology is not natural language
Variables are often proxies, there is no single universal natural biological token,…

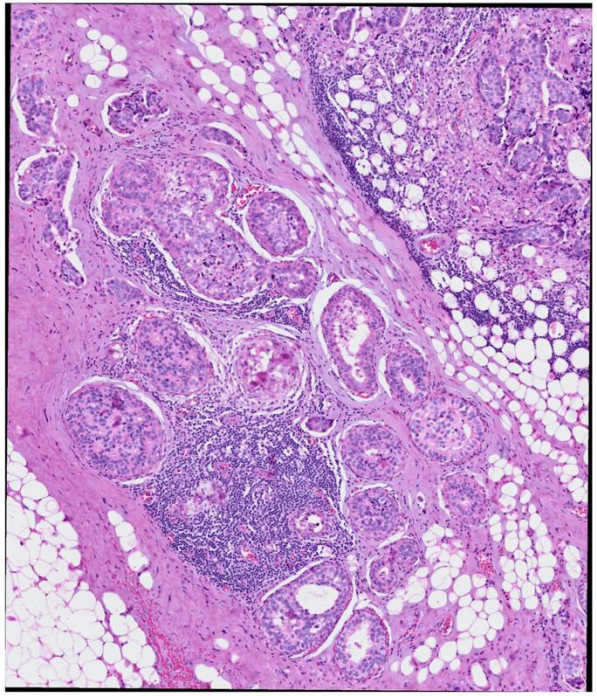
EMBL

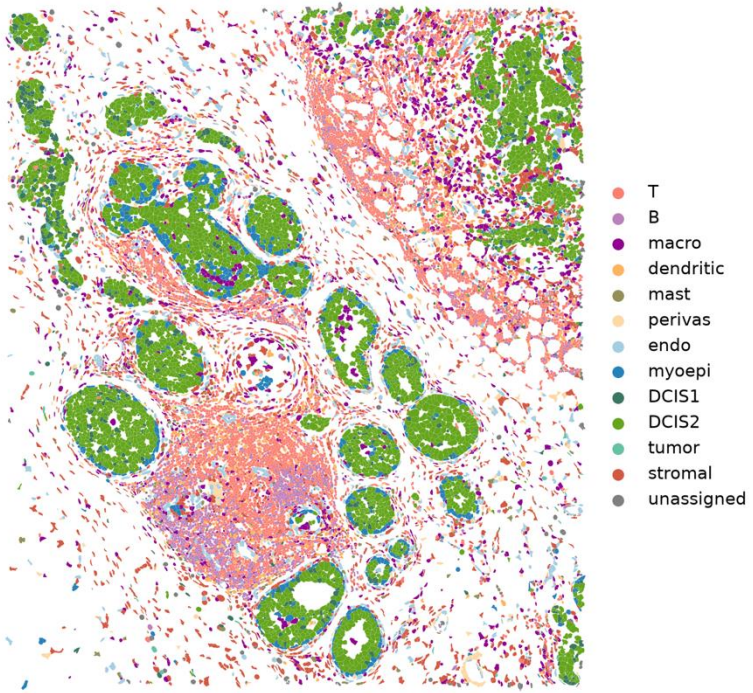
How to evaluate whether we've learned anything?

- Did the model optimize its training objective?
- Can we recover biological signal in the representation, like cell type, tissue, morphology, stage?
- Does it improve over training from scratch or simpler alternatives?
- Does it generalise to a new donor, study, tissue, species, or perturbation?
- Is it scientifically useful? Does it help generate hypotheses, prioritize experiments, or reveal something new?

Let’s look at an example

Data: Xenium spatial transcriptomics breast cancer sample (cropped tissue section image)





Data from Janesick et al. Nat Comm 2023

EMBL

Let’s embed the H&E image with CONCH

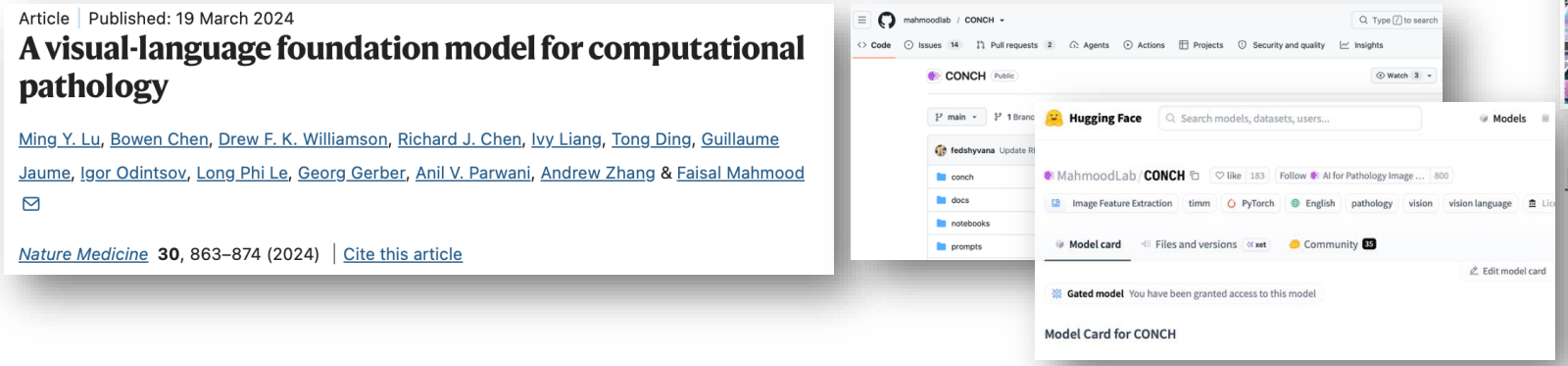


Image taken from https://github.com/mahmoodlab/CONCH

- Download the code from GitHub and the model weights from HuggingFace
- Figure out how to process data, set hyperparameters, set prompts…
- Run one model forward pass on the example image:
  - Tile image in patches and embed with CONCH image encoder
  - Embed query text prompts with text encoder
  - Compute image-text similarity in embedding space
  - → get tiles that are most similar to a prompt like ‘invasive breast carcinoma’

EMBL

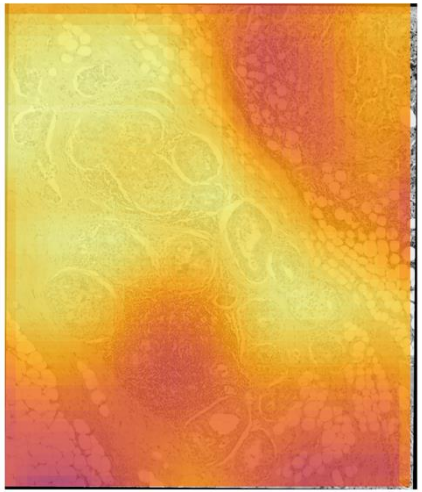
CONCH embeddings for a few text prompts

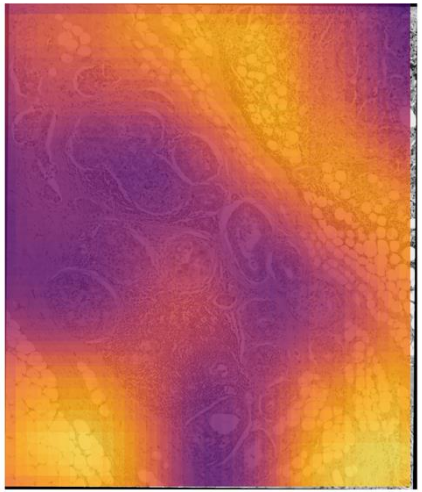
‘Invasive breast carcinoma’

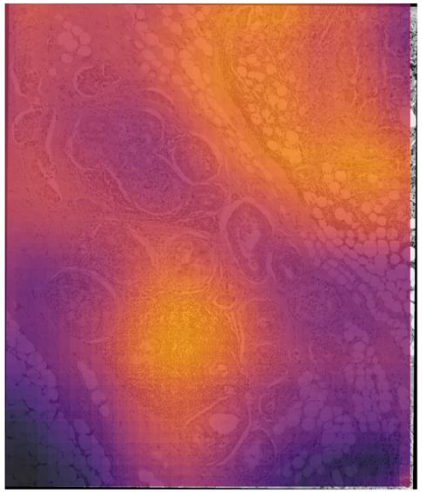
‘Adipose tissue’

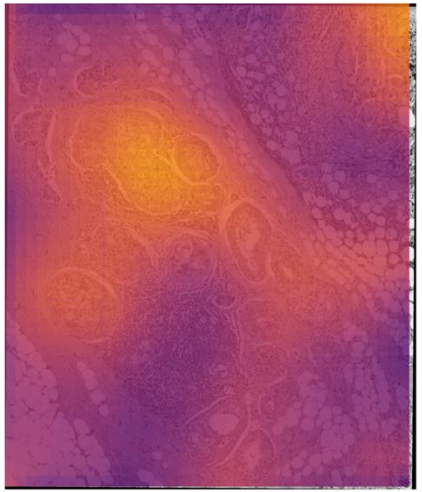
‘Lymphocyte-rich immune infiltrate’

‘Tissue that enjoys Shakespeare’

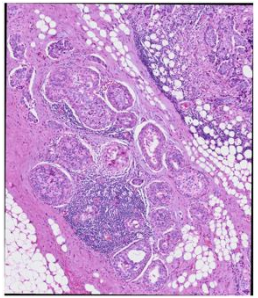


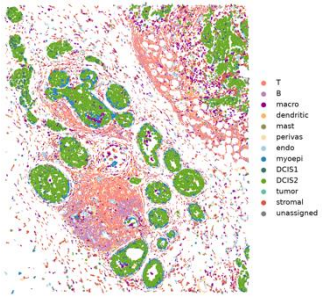






Brighter colours indicate higher similarity to the prompt





EMBL

Let’s embed the gene expression data with CellWhisperer

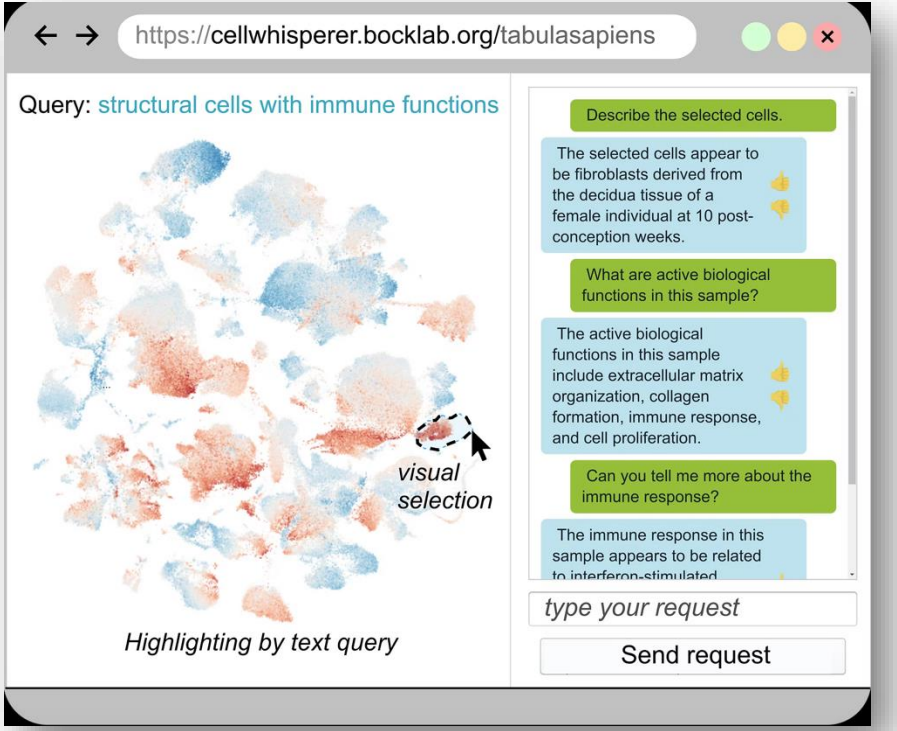
Article | Open access | Published: 11 November 2025

Multimodal learning enables chat-based exploration of single-cell data

Moritz Schaefer, Peter Peneder, Daniel Malzl, Salvo Danilo Lombardo, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Celine Sin, Jörg Menche, Eleni M. Tomazou & Christoph Bock

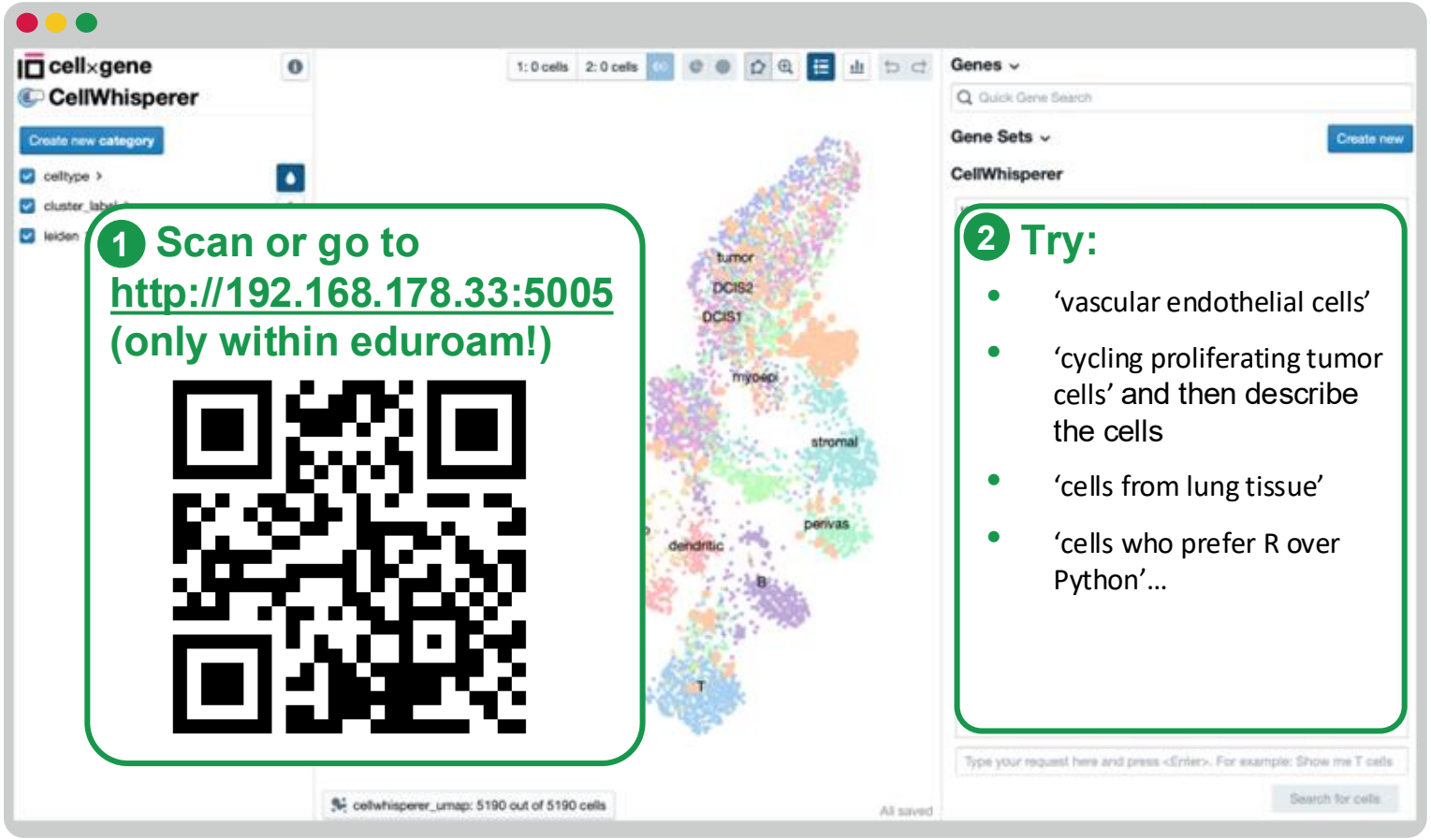
Nature Biotechnology (2025) | Cite this article

Image taken from
https://cellwhisperer.bocklab.org



EMBL

CellWhisperer demo on the example data



→ should map to orange 'endo' cluster

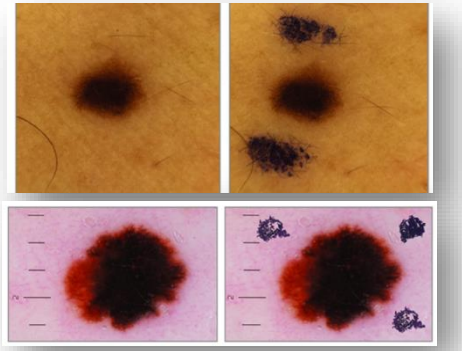
→ should highlight subset of tumor-like cells with markers like MKI67, TOP2A, STMN1

→ should return nothing because this is a breast cancer sample

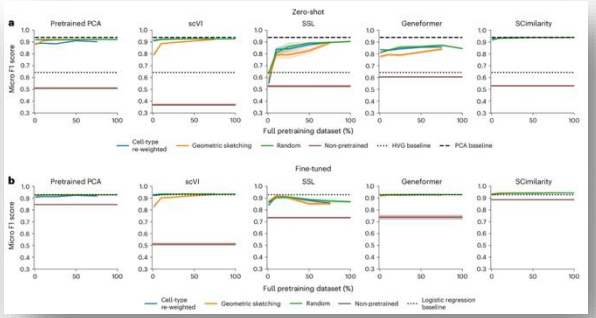
EMBL

Common pitfalls

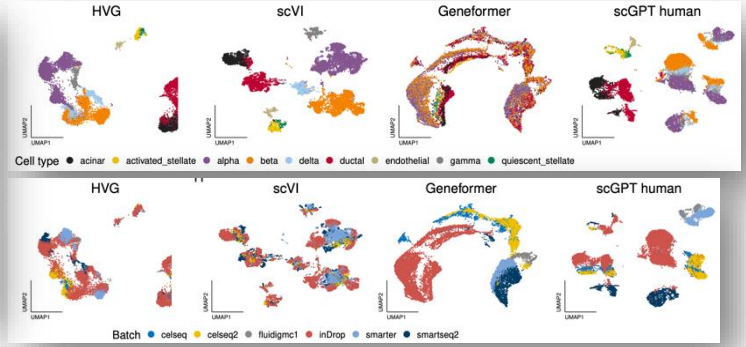
- Shortcut learning
- Data leakage
- Overconfidence (especially out of distribution)
- Larger training data doesn’t necessarily lead to better generalisation
- Evaluation too close to pre-training / no true zero-shot evaluation



Winkler et al. JAMA Dermatology 2019



DenAdel et al. Nat Methods 2026



Kedzierska et al. Genome Biology 2025

EMBL

Summary

- AI is changing computational biology (and science in general) on multiple levels: representations, predictions, workflows
- Vision of the ‘virtual cell’: universal representations
- Pre-training: train models on large datasets using self-supervised objectives, to learn representations that generalise and enrich analysis of smaller target data
- Think of a learned embedding like a map:
how useful it is depends on the terrain in which you’re using it
- Models can capture biological signal but typically don’t flag when they don’t know
- Keep in mind common failure modes and
check whether models improve over simple baselines

Thanks for listening!

Questions?

EMBL