An end-to-end workflow for Cell Painting and time lapse assays

Alán F. Muñoz

2025/04/03

Outline

Introduction

Original expected goals

Original goal

> *We aim to explore and establish methods and best practices in data discovery, interpretation, visualization. We will ... adapt and implement concepts from explainable AI and spatial analysis research fields that could lead to the biological understanding of image-derived data...*

Thus, the main question:

> *"How do we interpret what a given profile (e.g. of a cluster/sample) means?"*

These were general terms

The Data

- Bright field + time series

The Data

- Bright field + time series
    - + Model to obtain virtual staining

The Data

- Bright field + time series
    - + Model to obtain virtual staining
- Traditional Cell Painting assay

The Data

- Bright field + time series
  - + Model to obtain virtual staining
- Traditional Cell Painting assay
- Experiment with viability markers

Defining the task

Some context on the collaborators

- Technically proficient in computational approaches

Some context on the collaborators

- Technically proficient in computational approaches
- Open to ”productionize” deliverables

Some context on the collaborators

- Technically proficient in computational approaches
- Open to ”productionize” deliverables
- Open to suggestions and exploratory new approaches

Some context on the collaborators

- Technically proficient in computational approaches
- Open to ”productionize” deliverables
- Open to suggestions and exploratory new approaches
- Deliverables can be scripts, notebooks and/or software

Some software engineering considerations

- Should we build new pipelines or reuse existing ones? Which is more valuable for the collaboration?

Some software engineering considerations

- Should we build new pipelines or reuse existing ones? Which is more valuable for the collaboration?
- Hacking things together vs building and documenting tools

Some software engineering considerations

- Should we build new pipelines or reuse existing ones? Which is more valuable for the collaboration?
- Hacking things together vs building and documenting tools
- If building, how are we going to integrate it into their workflows?

After further discussion we agreed on a set of more tangible goals

- What are the best practices to facilitate processing and interpretability of microscopy data?

After further discussion we agreed on a set of more tangible goals

- What are the best practices to facilitate processing and interpretability of microscopy data?
- Can we extract dynamic (time) features from single cell measurements?

After further discussion we agreed on a set of more tangible goals

- What are the best practices to facilitate processing and interpretability of microscopy data?
- Can we extract dynamic (time) features from single cell measurements?
- Could they reduce costs/increase throughput by using time series in addition or replacement of Cell Painting?

After further discussion we agreed on a set of more tangible goals

- What are the best practices to facilitate processing and interpretability of microscopy data?
- Can we extract dynamic (time) features from single cell measurements?
- Could they reduce costs/increase throughput by using time series in addition or replacement of Cell Painting?
- How do we interpret biological

After further discussion we agreed on a set of more tangible goals

- What are the best practices to facilitate processing and interpretability of microscopy data?
- Can we extract dynamic (time) features from single cell measurements?
- Could they reduce costs/increase throughput by using time series in addition or replacement of Cell Painting?
- How do we interpret biological
- How do we leverage deep learning features in addition to traditional measurements?

How can we compare the two assays if they are studied in different ways?

If we are to compare bright field time lapses to cell painting data we need a common framework in which both assays are processed

Challenges and constraints

Problem 1: Where is the data? What size is it?
On their GCP storage. ~2.5Tb/experiment, 4 experiments so far

Challenges and constraints

Problem 1: Where is the data? What size is it?
On their GCP storage. ~2.5Tb/experiment, 4 experiments so far

Problem 2: What is "the data"?
Who will/How to process the raw images?

Challenges and constraints

Problem 1: Where is the data? What size is it?
On their GCP storage. ~2.5Tb/experiment, 4 experiments so far

Problem 2: What is ”the data”?
Who will/How to process the raw images?

Problem 3: If we have to process, do we have enough compute power?
GCP & Slurm vs AWS/DGX

Challenges and constraints

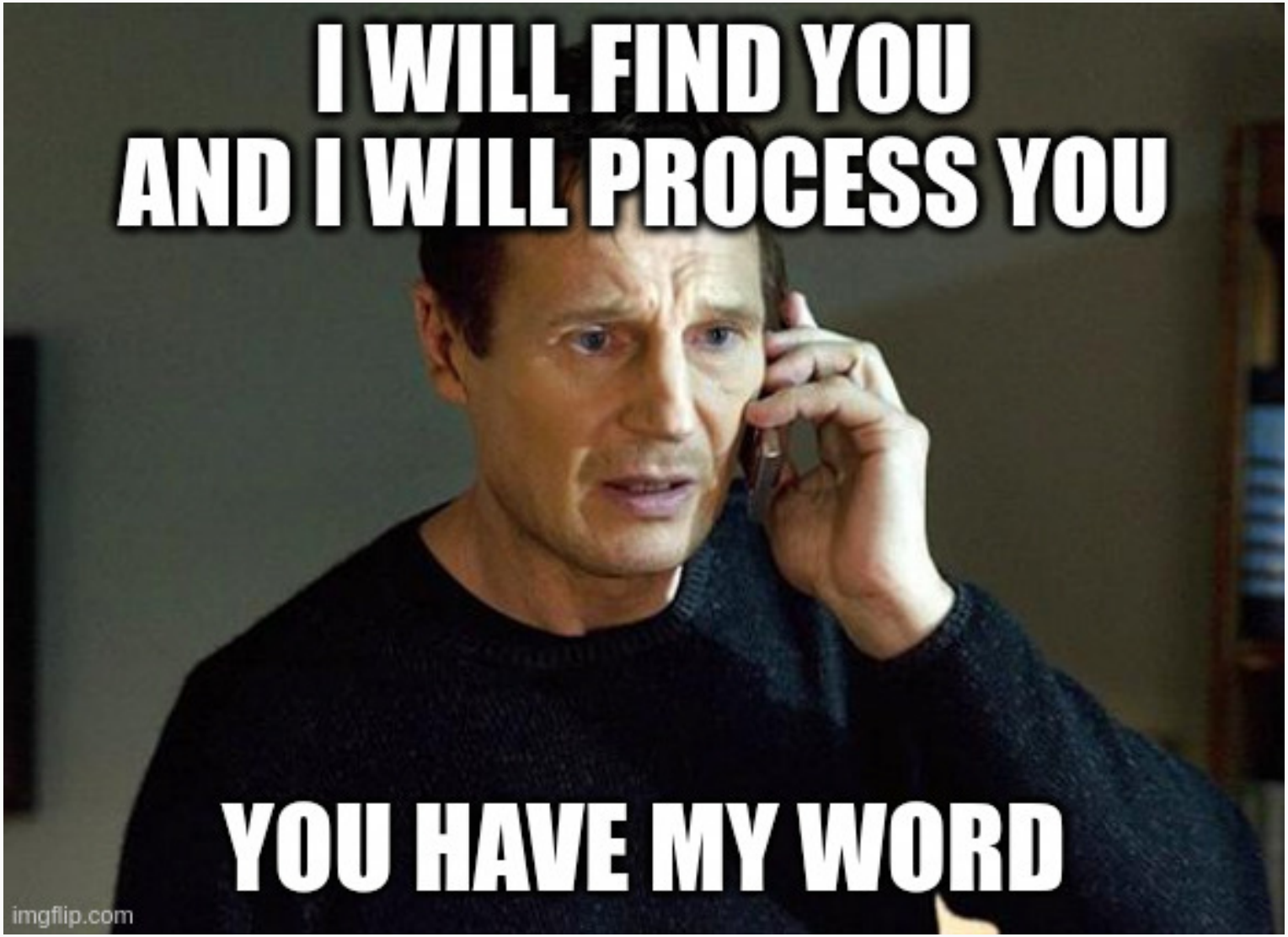
Problem 1: Where is the data? What size is it?
On their GCP storage. ~2.5Tb/experiment, 4 experiments so far

Problem 2: What is "the data"?
Who will/How to process the raw images?

Problem 3: If we have to process, do we have enough compute power?
GCP & Slurm vs AWS/DGX

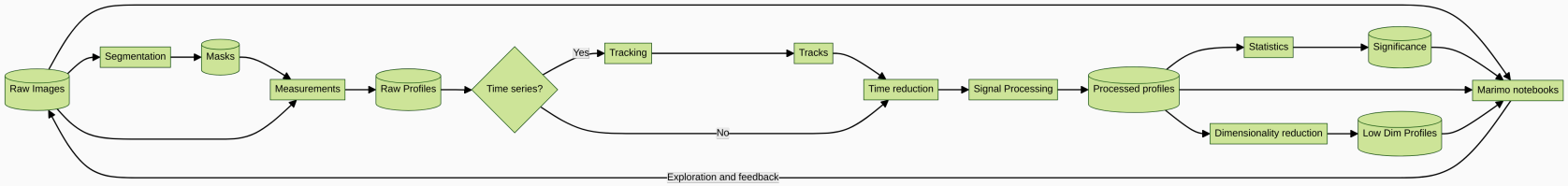
Problem 4: How do we reproduce processing/analysis on their infrastructure?

Even getting the data can be a problem



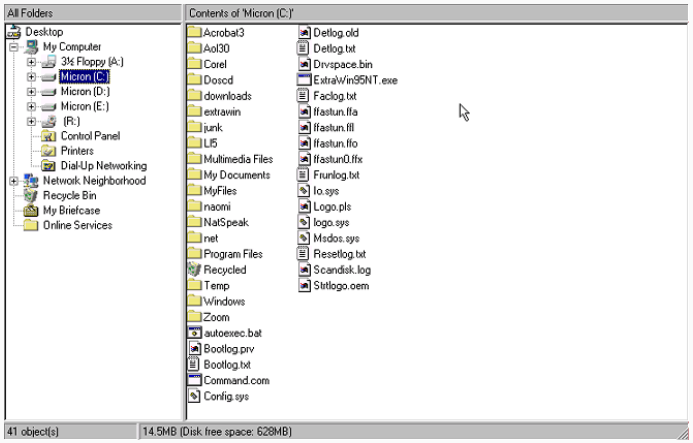
Anatomy of a pipeline

Overview

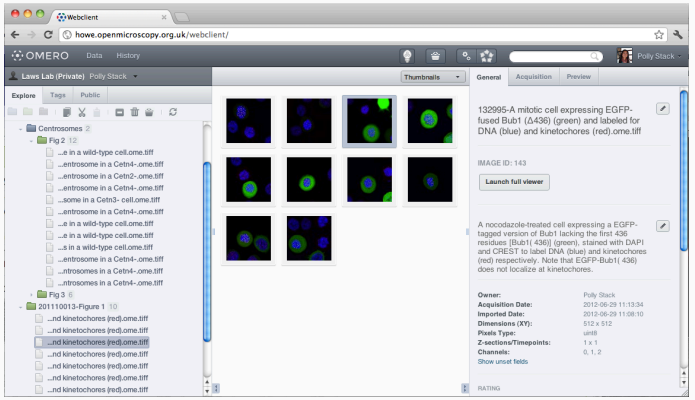

Raw Images
Segmentation
Masks
Measurements
Raw Profiles
Time series?
Yes
Tracking
Tracks
No
Time reduction
Signal Processing
Processed profiles
Statistics
Significance
Dimensionality reduction
Low Dim Profiles
Marimo notebooks
Exploration and feedback

Data ingress: How do we make it easy to access and run the analysis?

Local files



Omero server

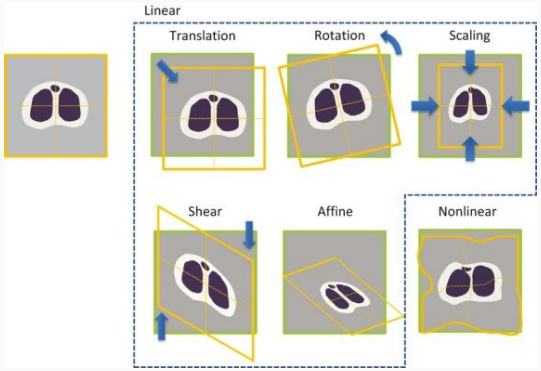


Cloud providers

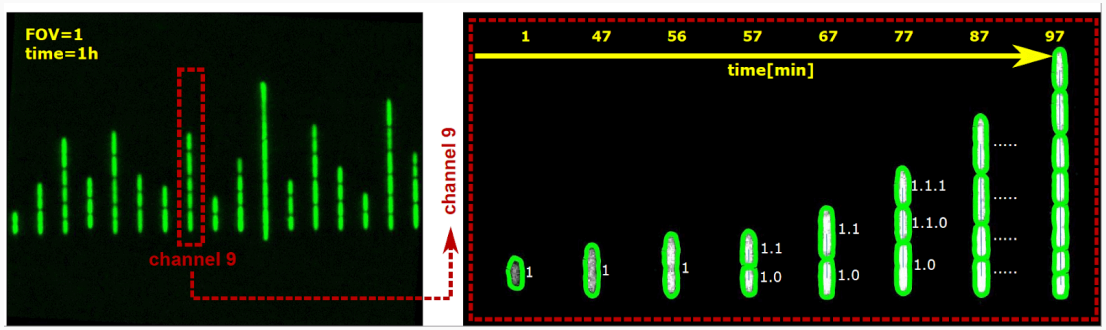
Image registration/corrections: How do we normalize our regions of interest?

Image registration: ”Transforming different sets of data into one coordinate system”

General registration



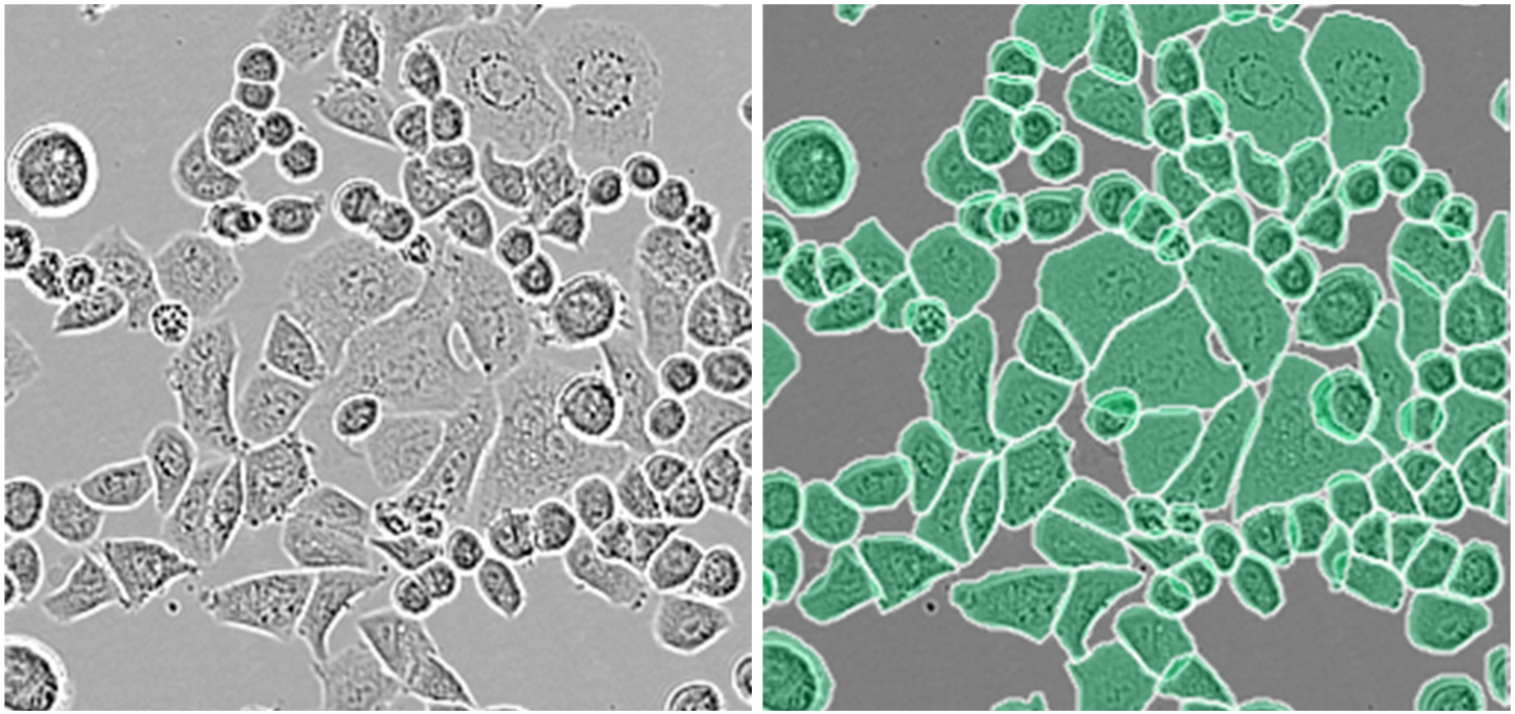
In cell microscopy



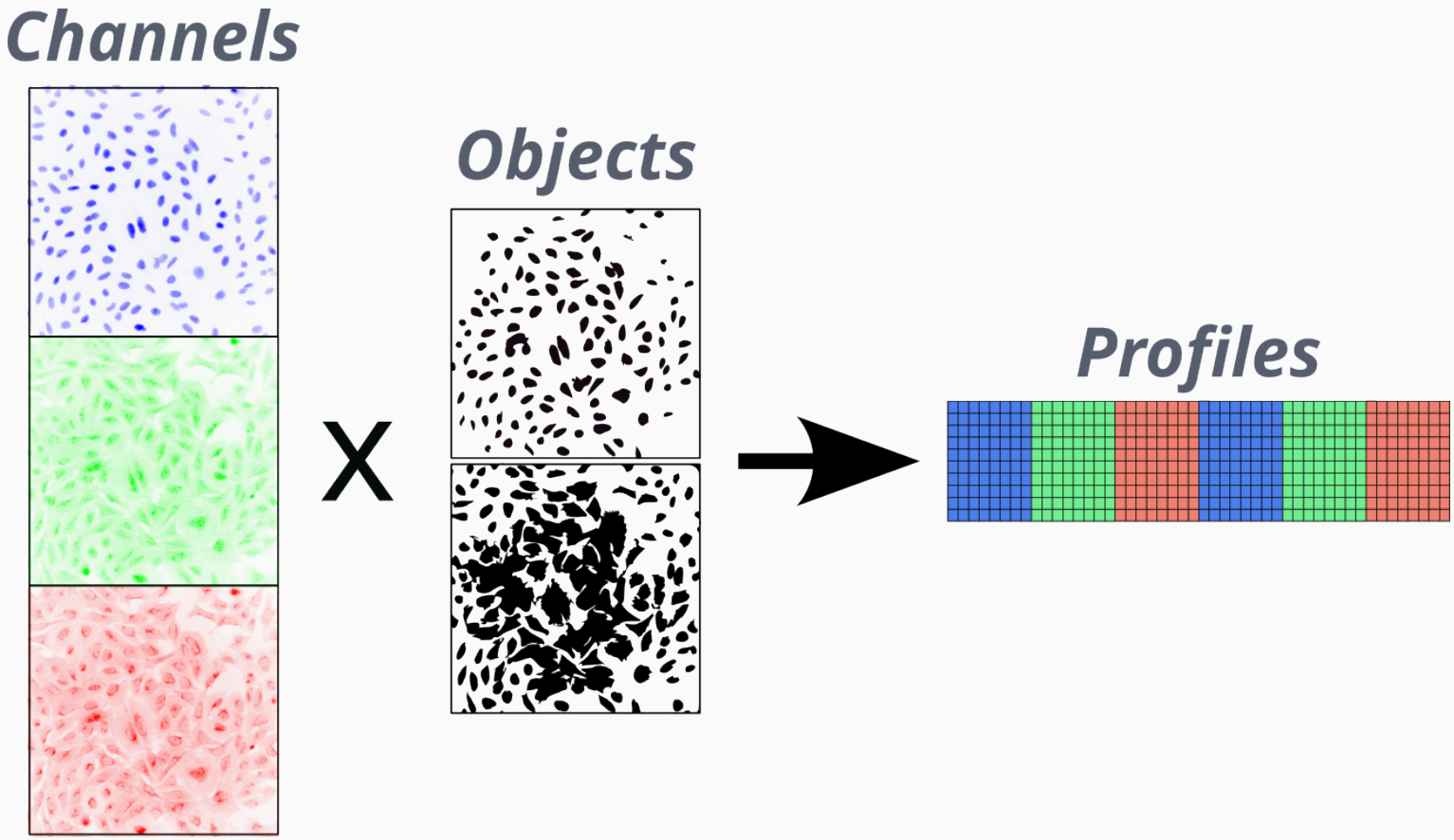
Segmentation: Which pixels do we care about?

Identify the pixels that characterise an object in an image.

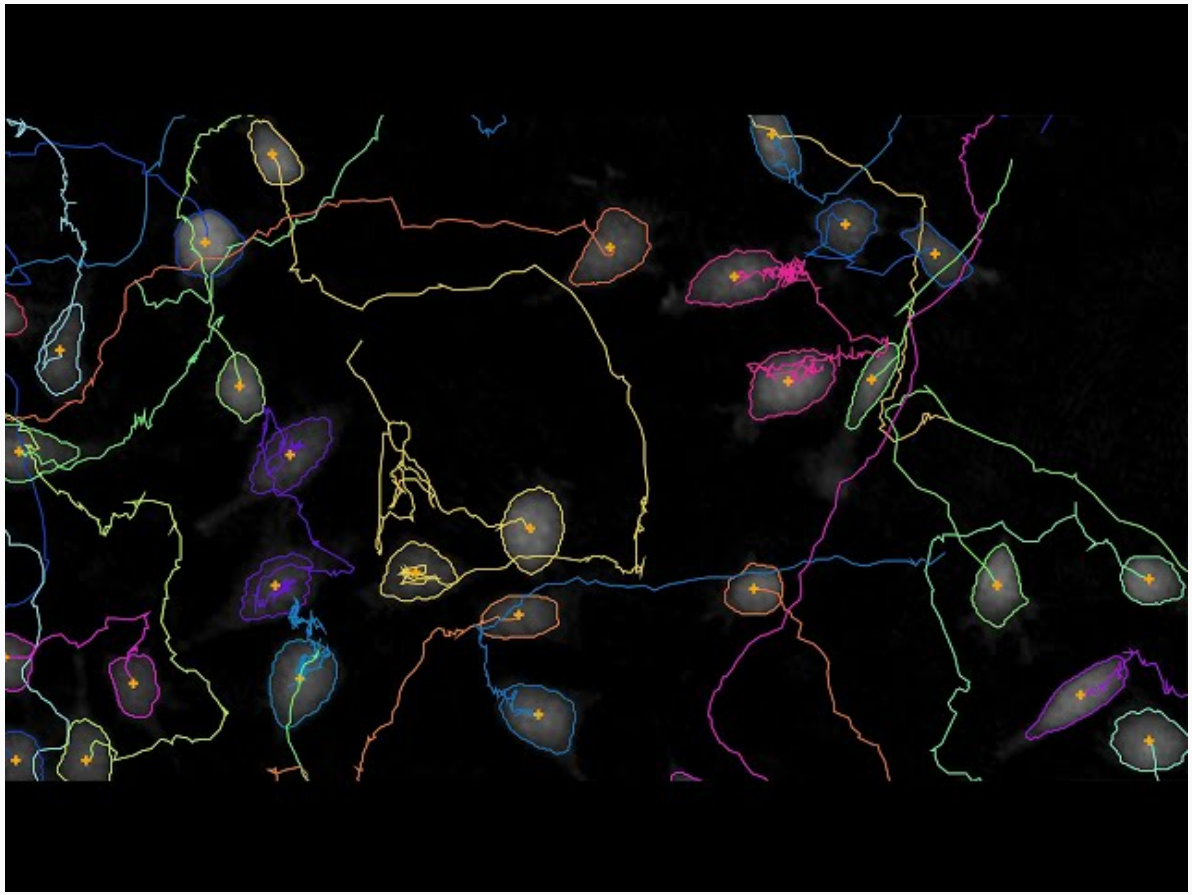
- Traditional computer vision (e.g., Watershed methods)
- Deep Learning (e.g., Convolutional Neural Networks)



Measurements: How do we reduce the dimensionality and size of our data?



Tracking: How do we identify individuals over time?



Tracking provides distinct information from standard Cell Painting: motility, division and growth.

Data egress: How do we format the different results of the pipeline?

Low-stakes decision, but still important to choose wisely:

- profiles: Parquet tables
- Other numerical data: zarr/npy

Orchestration: How do we minimise complexity while wrangling this mess of moving parts?

- Turns a bunch of components into a pipeline.
- Are pipelines actually good?

Signal processing: How do we maximise the information per experiment?

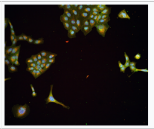
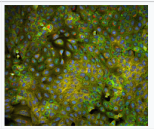
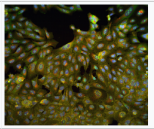
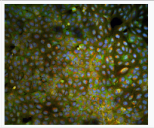
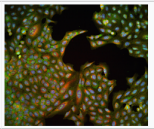
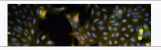
- *catch22*: Aggregate time series data
- *trommel*: Signal processing clean up

Exploration: How do we make sense of the features?

This is an open question.

✎ View and edit SQL

This data as json, CSV (advanced)

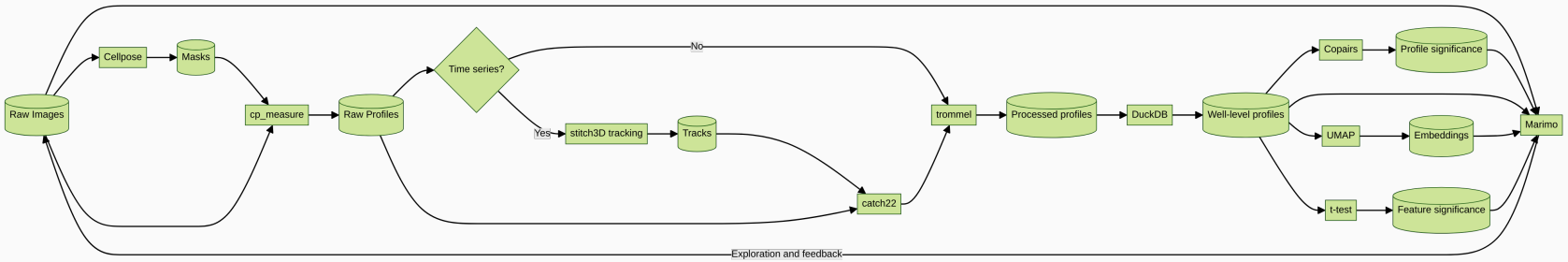
Link	rowid ▼	Compartment	Feature	Channel	Suffix	Feature significance	Perturbation	Corrected p-value	Phenotypic activity	Perturbation example image	Median	Gene Rank	Feature Rank	JCP20
1	1	Cells	Texture_SumVariance	AGP	_3_01_256	0.00946	PLK1				-2.522	999999	16	JCP20
2	2	Nuclei	AreaShapeMajorAxisLength			0.0	PDE8B	0.0618	0.43399		-1.937	999999	2	JCP20
3	3	Cytoplasm	Texture_InfoMeas2	Mito	_5_02_256	0.0	KCNH3	0.05455	0.39131		1.891	999999	10	JCP20
4	4	Cytoplasm	RadialDistributionMeanFrac		_mito_tubeness_1of20	0.0	XIAP	0.87596	0.1708		-1.919	999999	10	JCP20
5	5	Cells	RadialDistributionFracAtD		_mito_tubeness_20of20	0.0	PIAS2	0.01257	0.31734		1.512	999999	11	JCP20
6	6	Cytoplasm	Correlation_K_RNA	DNA		0.0	AIRE	0.02864	0.53388		-2.775	999999	5	JCP20

Results

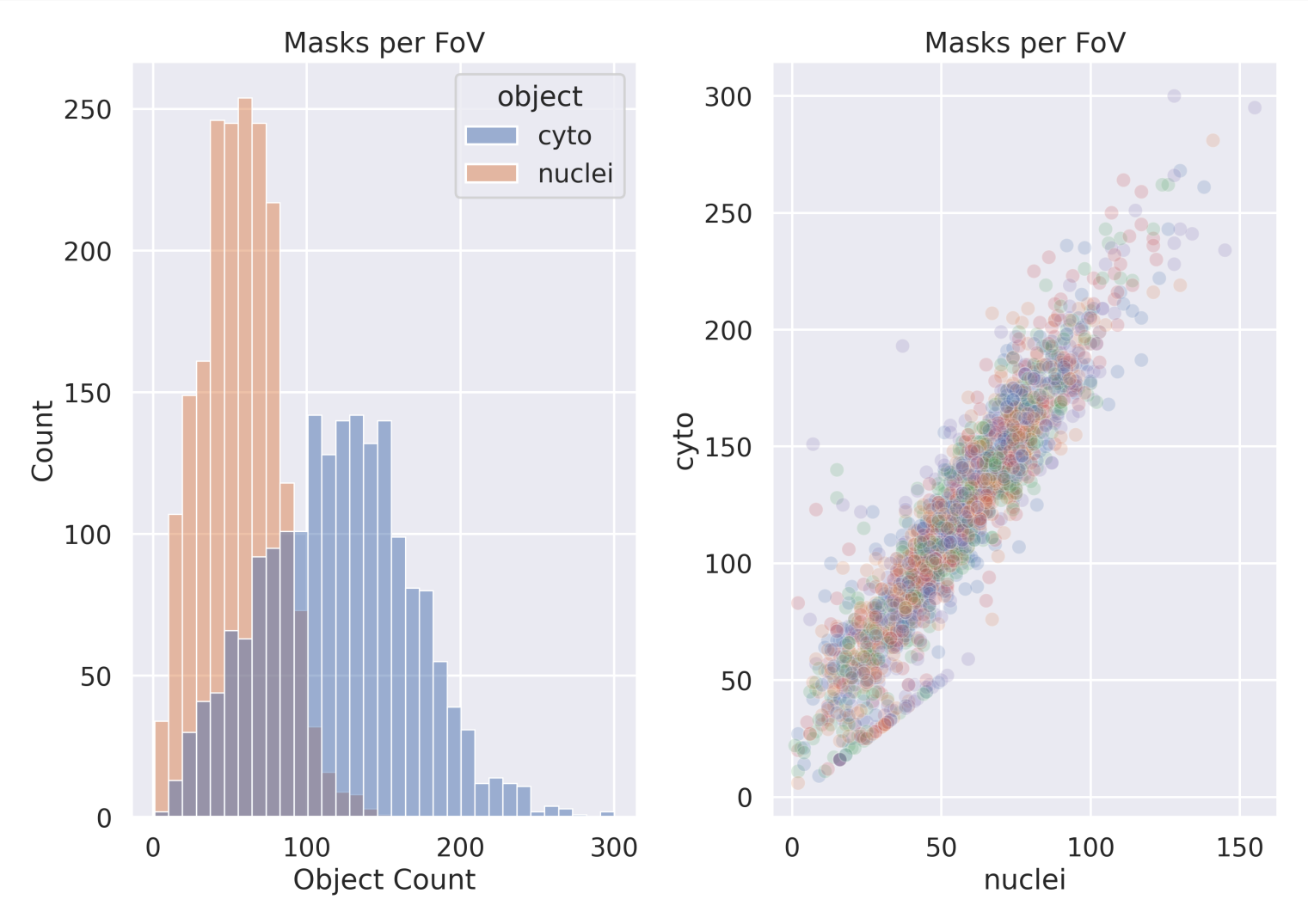
Chosen stack (table)

Step	Tech
Ingress	Local files
Registration	aliby
Segmentation	cellpose
Measurement	cp_measure
Tracking	cellpose's stitch3D
Egress	Parquet+npy
Orchestration	aliby
Signal processing	catch22 (ts) + trommel
Exploration	Marimo

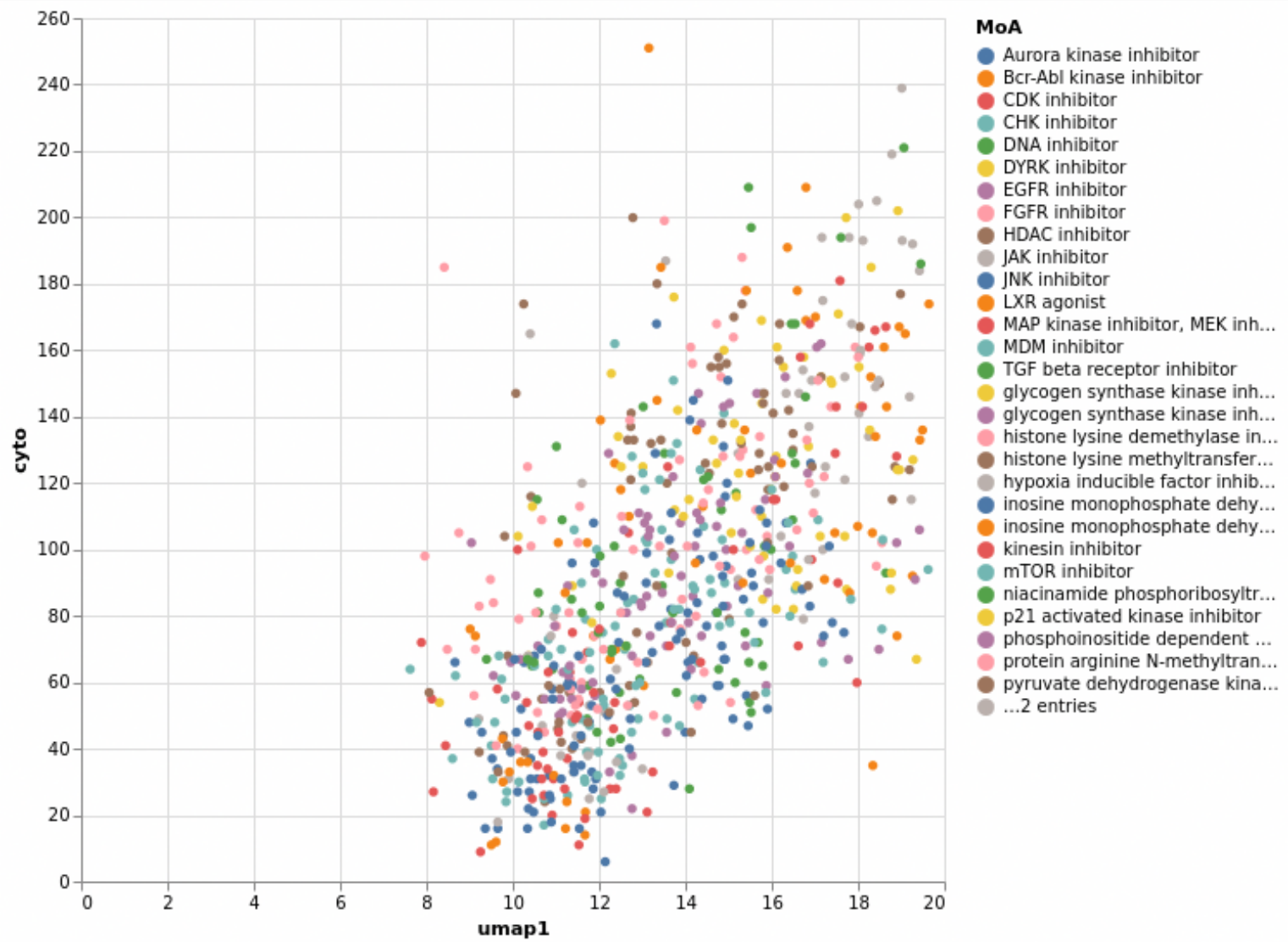
Chosen stack (diagram)



Is vanilla segmentation consistent between cyto and nuclei?



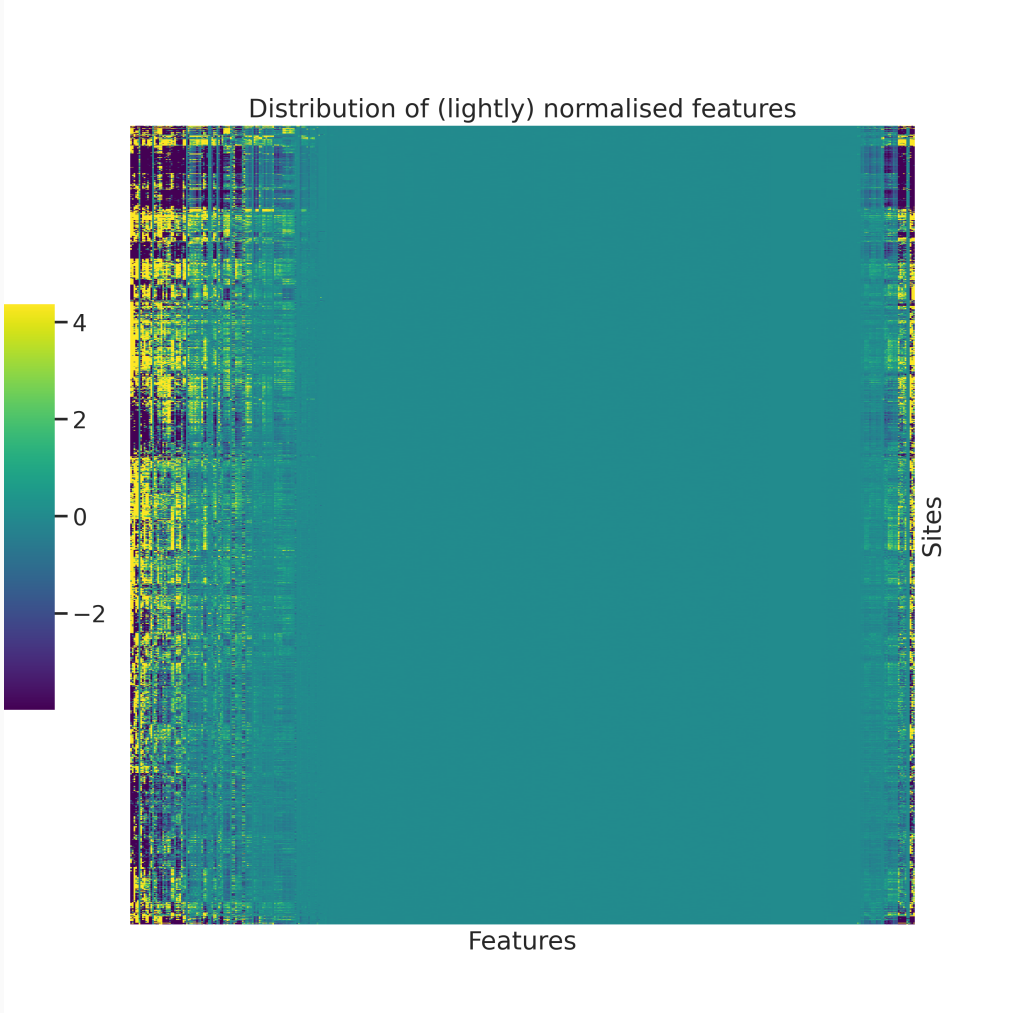
Cell count correlates is a strong determinant of signal



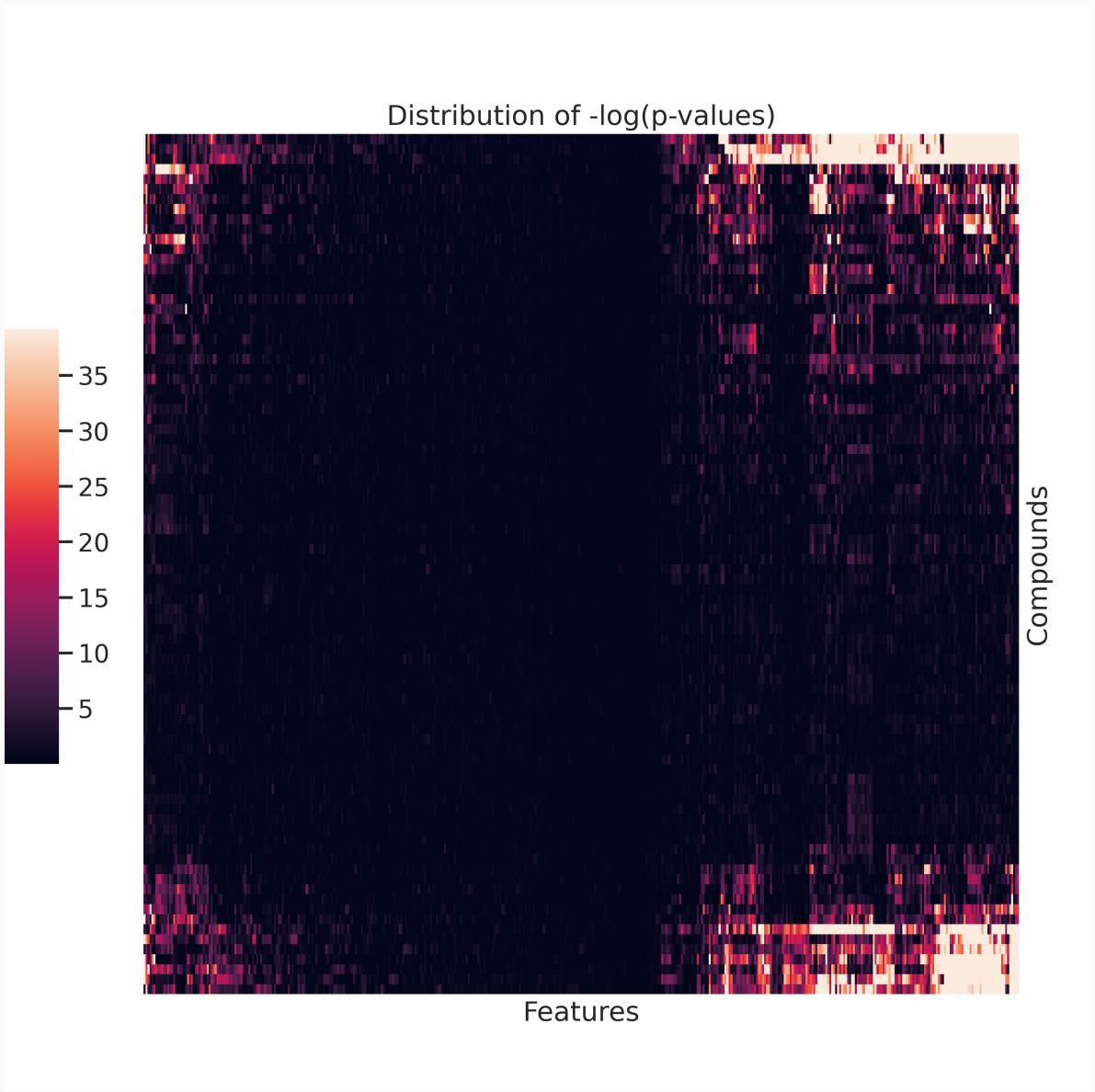
Bringing it all together: A Marimo interface

Key point: Biological interpretation greatly benefits from access to processed and raw images.

The data could benefit from adjustments



Feature selection and stats help, but could be improved upon



Current time estimates (190 cores, ~150 GB RAM):

Assay	Ch	TP	Obj	Time (h)	#FoV completed	FoV/h	(FoV,Tp,Ch)/h
CP	6	1	2	26.6	1920 (all)	72.0	432
TS + VS	20	2	2	49.0	109 (~5%)	2.2	89

(metrics calculated without radial, granularity or zernike)

The current bottleneck are the measurements (and cellpose)

For a given pipeline with 6 channels, 2 objects and 1 time point:

Module	% of time
Granularity	72.0%
Zernike	7.7%
CellPose (Threaded)	6.0%
Radial Distribution	3.4%

Conclusions

Understanding the evolution of the project

- Though not in the initial plans, we had to work out a way to process current and incoming datasets

Understanding the evolution of the project

- Though not in the initial plans, we had to work out a way to process current and incoming datasets
- Marimo seems to be a way to provide both an exploration interface and reproducible notebooks

Understanding the evolution of the project

- Though not in the initial plans, we had to work out a way to process current and incoming datasets
- Marimo seems to be a way to provide both an exploration interface and reproducible notebooks
- Vanilla cellpose, though resulting from a noisier cytosolic segmentation, seems a viable option in the stack

Tools/methods were developed/expanded to further the project

- *aliby*: End-to-end pipeline for both Cell Painting and time series data

Tools/methods were developed/expanded to further the project

- *aliby*: End-to-end pipeline for both Cell Painting and time series data
- *cp_measure*: Cell Profiler measurements one import away

Tools/methods were developed/expanded to further the project

- *aliby*: End-to-end pipeline for both Cell Painting and time series data
- *cp_measure*: Cell Profiler measurements one import away
- *trommel*: Cleans up the data

Tools/methods were developed/expanded to further the project

- *aliby*: End-to-end pipeline for both Cell Painting and time series data
- *cp_measure*: Cell Profiler measurements one import away
- *trommel*: Cleans up the data
- *marimo* interfaces: Explore statistics and images together

Tools/methods were developed/expanded to further the project

- *aliby*: End-to-end pipeline for both Cell Painting and time series data
- *cp_measure*: Cell Profiler measurements one import away
- *trommel*: Cleans up the data
- *marimo* interfaces: Explore statistics and images together
- Significant *copairs* speed up

Tools/methods were developed/expanded to further the project

- *aliby*: End-to-end pipeline for both Cell Painting and time series data
- *cp_measure*: Cell Profiler measurements one import away
- *trommel*: Cleans up the data
- *marimo* interfaces: Explore statistics and images together
- Significant *copairs* speed up
- Fast and scalable per-feature p value calculation

The new(ish) toys that I have found very useful

- *marimo*: Jupyter notebooks the done right

The new(ish) toys that I have found very useful

- *marimo*: Jupyter notebooks the done right
- *duckdb*: SQL on steroids

The new(ish) toys that I have found very useful

- *marimo*: Jupyter notebooks the done right
- *duckdb*: SQL on steroids
- *dask*: Small data, big data? Doesn't make a difference?

The new(ish) toys that I have found very useful

- *marimo*: Jupyter notebooks the done right
- *duckdb*: SQL on steroids
- *dask*: Small data, big data? Doesn't make a difference?
- *ThreadPoolExecutor*: Speed up python code, the easy way

Pending work

Pending work

- Move segmentation to the GPU.

Pending work

- Move segmentation to the GPU.
- Process all time series datasets

Pending work

- Move segmentation to the GPU.
- Process all time series datasets
- Adding masks and tracks to marimo for quality control

Pending work

- Move segmentation to the GPU.
- Process all time series datasets
- Adding masks and tracks to marimo for quality control
- Add port-based steps to avoid dependency bankrupcy

Pending work

- Move segmentation to the GPU.
- Process all time series datasets
- Adding masks and tracks to marimo for quality control
- Add port-based steps to avoid dependency bankrupcy
- Refine workflow for biological exploration

Pending work

- Speed up `cp_measure`?

Pending work

- Speed up `cp_measure`?
- Deeper comparison of Cell Painting and time series datasets

Pending work

- Speed up `cp_measure`?
- Deeper comparison of Cell Painting and time series datasets
- Develop a sensible cell count correction method that works on small datasets

Pending work

- Speed up `cp_measure`?
- Deeper comparison of Cell Painting and time series datasets
- Develop a sensible cell count correction method that works on small datasets
- Combine cytosol and nuclei information to find ”the one true cell”

Pending work

Technical things learned so far

- Local first -> distributed is easier than the other way around

Technical things learned so far

- Local first -> distributed is easier than the other way around
- Pipelines are not just functions stitched together, consider how/who will deal with the output and consciously choose the internals to expose

Technical things learned so far

- Local first -> distributed is easier than the other way around
- Pipelines are not just functions stitched together, consider how/who will deal with the output and consciously choose the internals to expose
- Keep the raw and processed data close to the compute

Technical things learned so far

- Local first -> distributed is easier than the other way around
- Pipelines are not just functions stitched together, consider how/who will deal with the output and consciously choose the internals to expose
- Keep the raw and processed data close to the compute
- Reuse tools as much as possible, it saves time and reduces dull work (if you like tools)

That’s all folks

Thanks for you attention. Questions?