SCIENCE MEETS LIFE

Spatial proteomics analysis that makes you happy

Benjamin Rombaut

VIB
GHENT UNIVERSITY
SaeysLab
Data mining and Modelling for Biomedicine
Flanders AI Research Program

16th of May 2025

SaeysLab

Data mining and Modelling for Biomedicine

https://saeyslab.sites.vib.be/en

Trustworthy Models of Cells and Tissues

4 subteams

Our tools integrate well with the scverse

(Imaging) flow cytometry

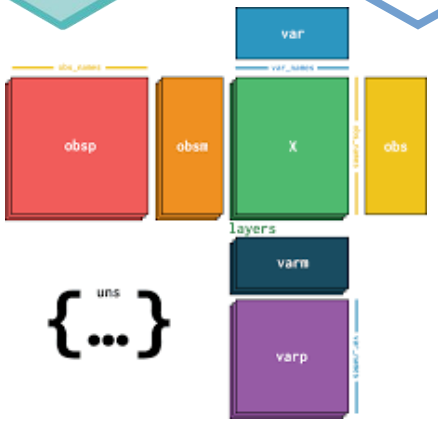
Spatial transcriptomics and proteomics

FlowSOM

SCIP
Scalable cytometry image processing

Arrow

HARPY
GITHUB.COM/SAEYSLAB/HARPY

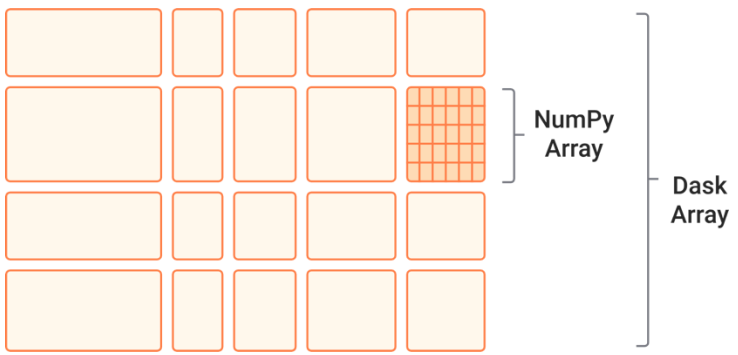


AnnData

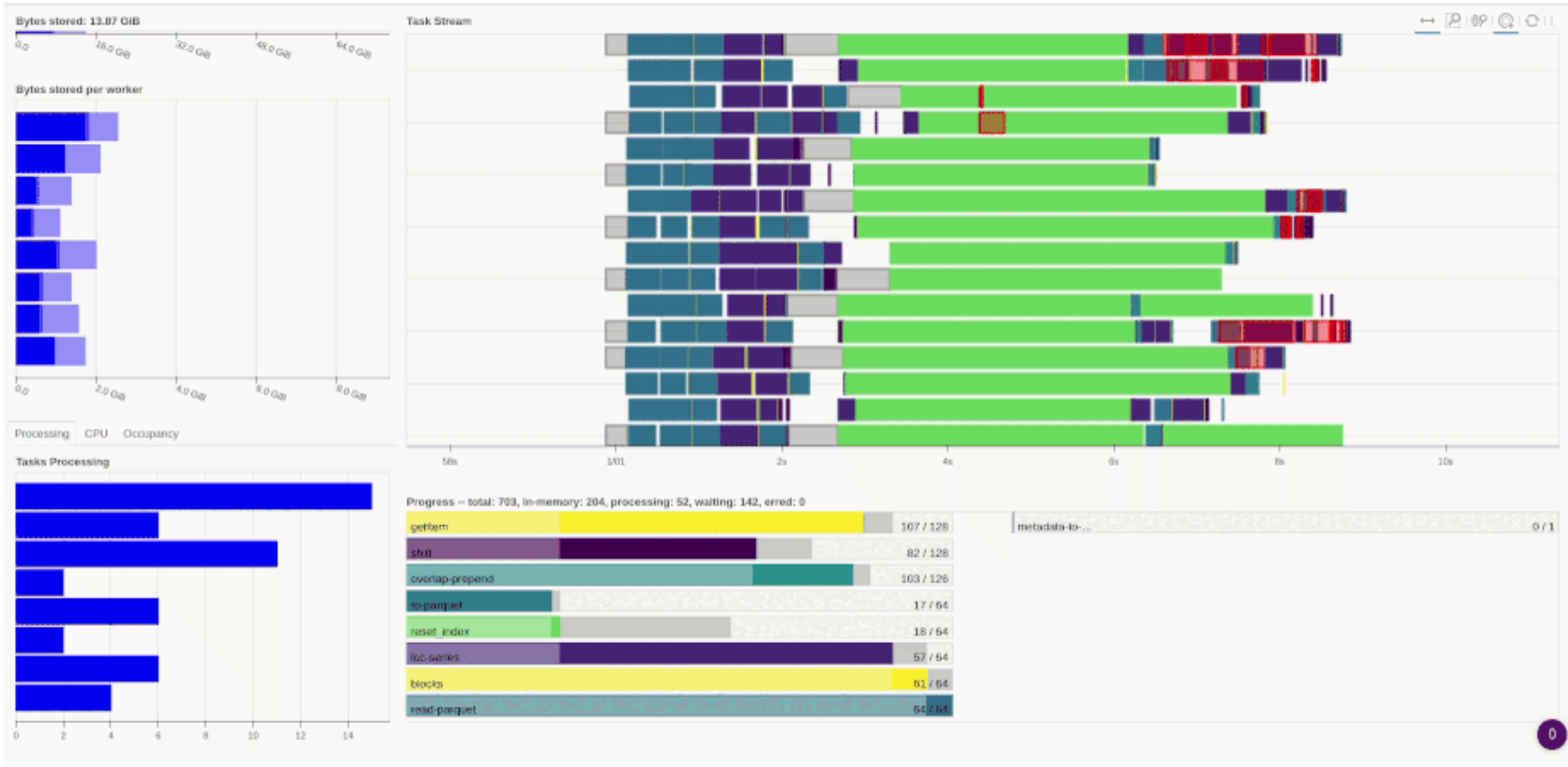
dask

SpatialData

Dask is a flexible parallel computing library



A Dask Array is
just a collection of NumPy Arrays



An analysis run visualized with the Dask Dashboard:
docs.dask.org/en/latest/dashboard.html

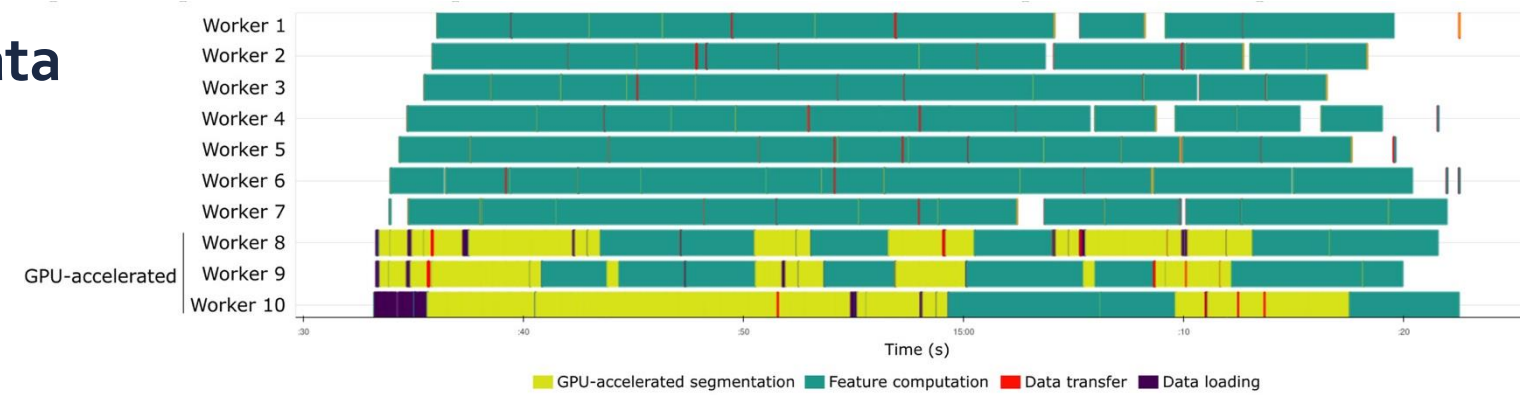
Scalable imaging flow cytometry processing

Morphological profiling of image cytometry and microscopy data

Support for storing IFC in **AnnData**

Support for parallel processing with **Dask**

Example of SCIP workflow execution

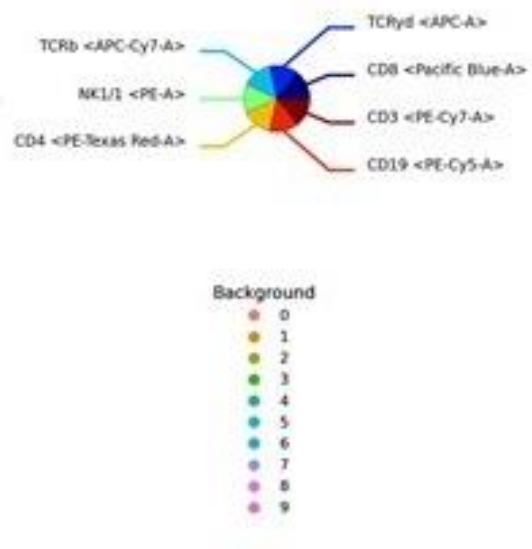


Maxim Lippeveld

Lippeveld, M., Peralta, D., Filby, A., Saeys, Y., 2024. SCIP: A scalable, reproducible and open-source pipeline for morphological profiling of image cytometry and microscopy data. Cytometry Part A 105, 816–828. https://doi.org/10.1002/cyto.a.24896

Unsupervised clustering of flow cytometry

Self-organizing maps for visualization and interpretation of cytometry data



Python and R version

Support for **AnnData**

Support for **batch parallel processing** with Dask and Numba

Artuur Couckuyt

Couckuyt, A., Rombaut, B., Saeys, Y., Van Gassen, S., 2024.
Efficient cytometry analysis with FlowSOM in Python boosts interoperability with other single-cell tools.
https://doi.org/10.1093/bioinformatics/btae179

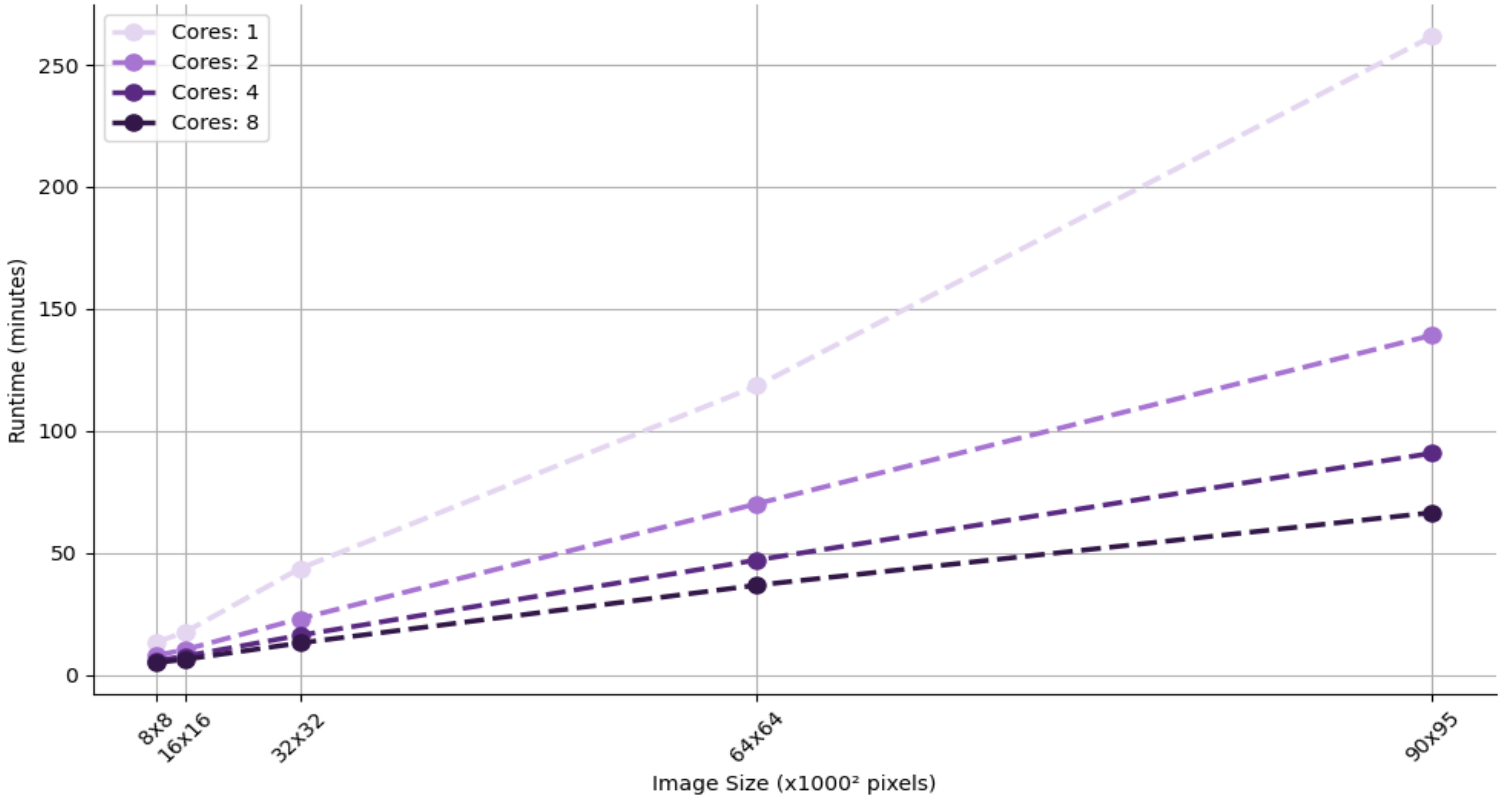
Scalable spatial transcriptomics processing

SPArrOW

A flexible, interactive and scalable pipeline for spatial transcriptomics analysis

SpatialData

Lotte Pollaris



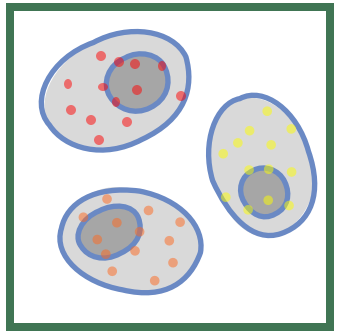
Runtime of whole SPArrOW pipeline on large MERSCOPE dataset with Dask and GPU acceleration

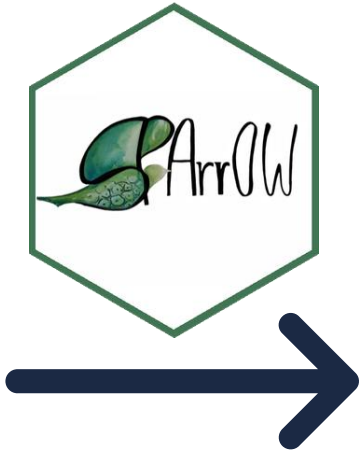
Pollaris, L., Vanneste, B., Rombaut, B., Defauw, A., Vernaillen, F., Mortier, J., Vanhenden, W., Martens, L., Thoné, T., Hastir, J.-F., Bujko, A., Saelens, W., Marine, J.-C., Nelissen, H., Hamme, E.V., Seurinck, R., Scott, C.L., Guilliams, M., Saeys, Y., 2024. SPArrOW: a flexible, interactive and scalable pipeline for spatial transcriptomics analysis. https://doi.org/10.1101/2024.07.04.601829

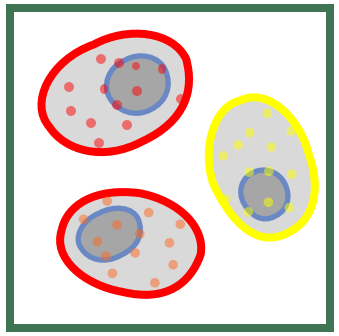
Image-based spatial proteomics analysis is a lot of images

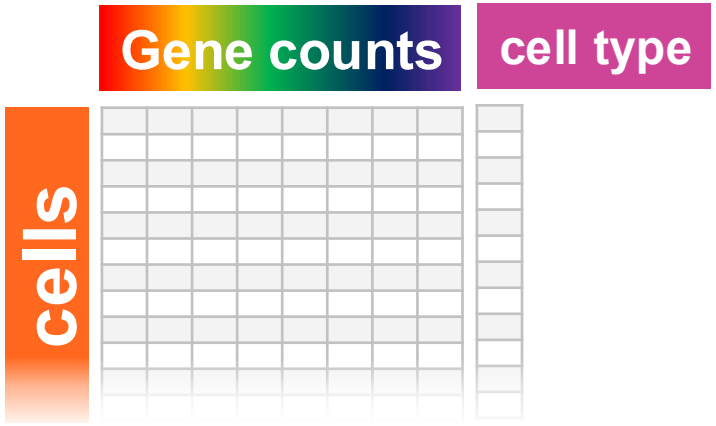
Spatial transcriptomics

Transcript locations +
whole-slide cell image



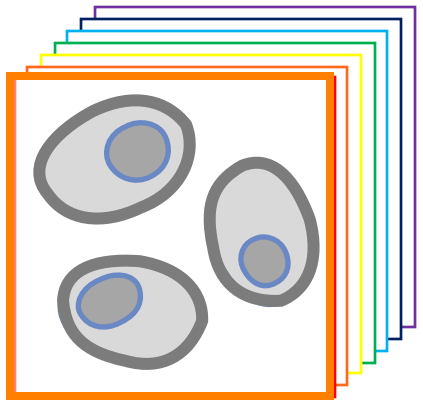




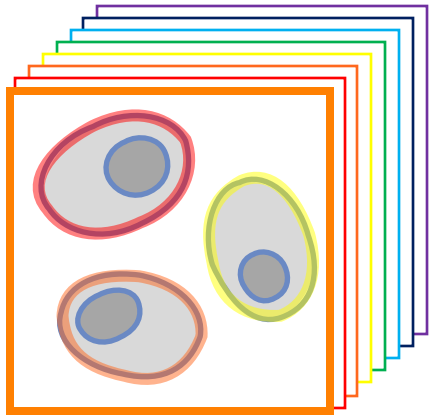


Spatial proteomics

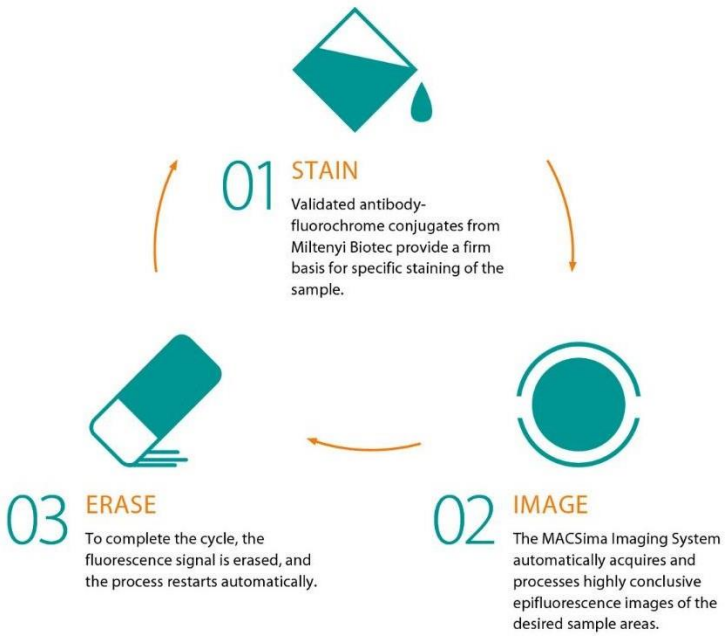
whole-slide
highly-multiplexed images

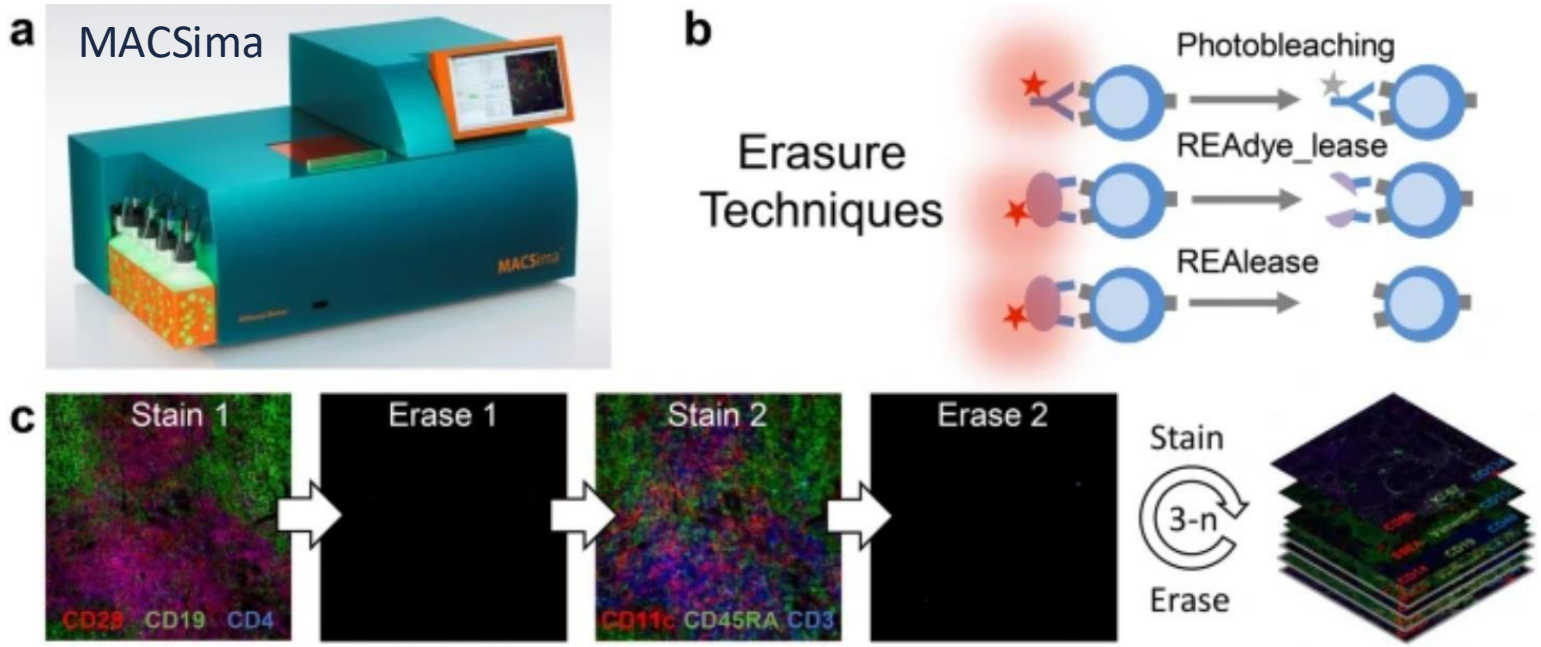


annotated single-cell data with spatial context



Spatial proteomics with MACSima is a type of cyclical immunofluorescence





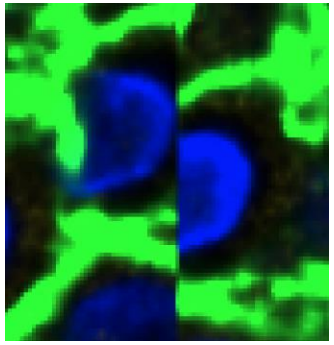
Kinkhabwala *et al*. (2022). https://doi.org/10.1038/s41598-022-05841-4

Some recent reviews and primers:

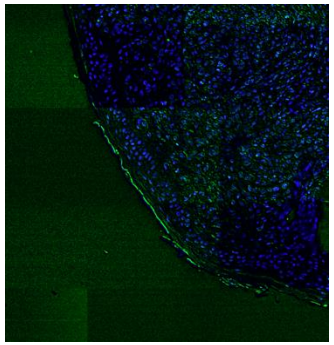
Bussi, Y., Keren, L., 2024. Multiplexed image analysis: what have we achieved and where are we headed? Nat Methods 21, 2212–2215. https://doi.org/10.1038/s41592-024-02539-5

Carstens, J.L., Krishnan, S.N., Rao, A., Sorace, A.G., Seeley, E.H., Ferri-Borgogno, S., Burks, J.K., 2024. Spatial multiplexing and omics. Nat Rev Methods Primers 4, 1–19. https://doi.org/10.1038/s43586-024-00330-6

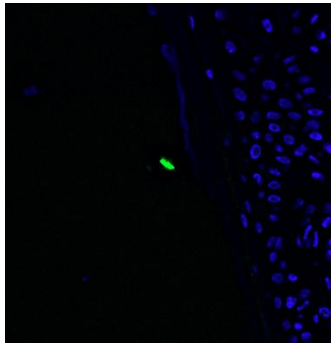
Difficulties of spatial omics



Stitching artefacts

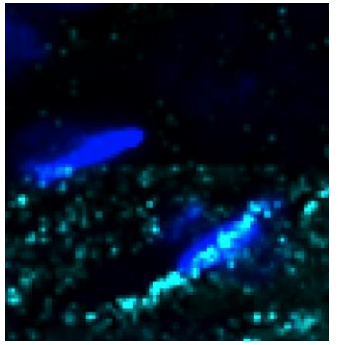


Background noise

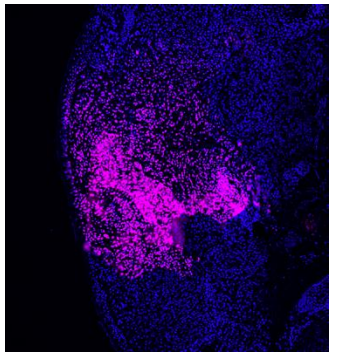


Outliers

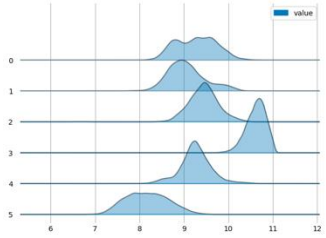
Tiling artefacts



Staining artefacts



Batch effects



Vierdag, W.-M.A.M., Saka, S.K., 2024. A perspective on FAIR **quality control** in multiplexed imaging data processing. Front. Bioinform. 4. https://doi.org/10.3389/fbinf.2024.1336257

Open-source data workflows

MCMICRO: broad focus and uses **Nextflow**

Nextflow/Python https://nf-co.re/mcmicro/dev

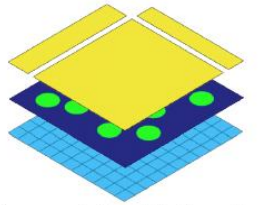
IMCDataAnalysis: focus on **IMC** data

Python/R https://bodenmillergroup.github.io/IMCDataAnalysis/

ark-analysis: focus on **MIBI** data

Python https://github.com/angelolab/ark-analysis

Harpy: focus on **MACSima** data and uses **SpatialData**
Python https://github.com/saeyslab/harpy



4 Harpy highlights

1. Dataset-wide quality control
2. Scalable whole-slide image processing
3. Interactive unsupervised clustering
4. Interoperability with R

https://github.com/saeyslab/harpy

Harpy analysis workflow

Spatial proteomics platforms

MACSima
Phenocycler
Lunaphore
MIBI-TOF
IMC
…

HPC
Cloud storage
Local

Storage options

Imaging data

Image QC + **Optional pixel clustering**

1 Nucleus and cell segmentation

Imaging data
Define implicit image chunks
→ Parallel multi-channel cell segmentation (Cellpose, InstanSeg, *Custom model*)
→ Harmonize cells across chunks
→ Cell masks

Segmentation QC

2 Feature calculation

Imaging data + Cell masks
→ Parallel calculation of features
- Staining-based expression
- Mask-based shape
- Spatial context

→ Cell features

Feature QC

3 Unsupervised clustering

Cell features → Feature transformation → Parallel clustering (FlowSOM) → Annotations

Annotation QC

SpatialData

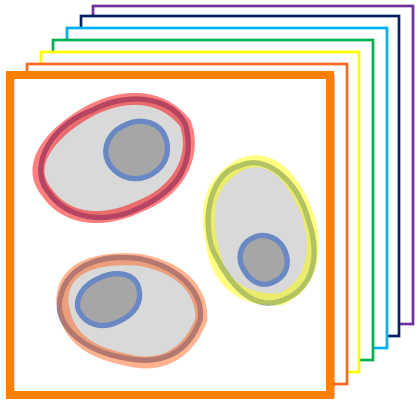
Imaging data + Cell masks + Cell features + Annotations

Interoperability

Python / R interoperability
Visualizations
Downstream analysis

AnnData anndataR napari ITESSCE jupyter squidpy

Download only the data chunk you need from remote object storage

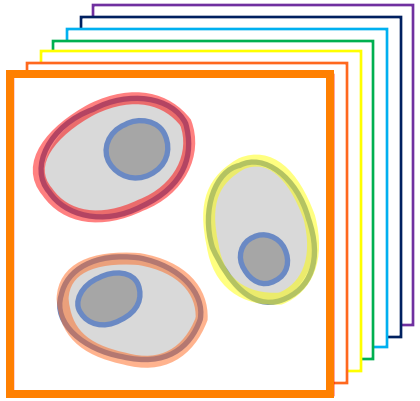


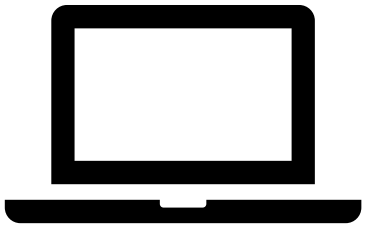
Whole-slide image

One sample: **>10 Gigabytes**!

SpatialData

Zarr



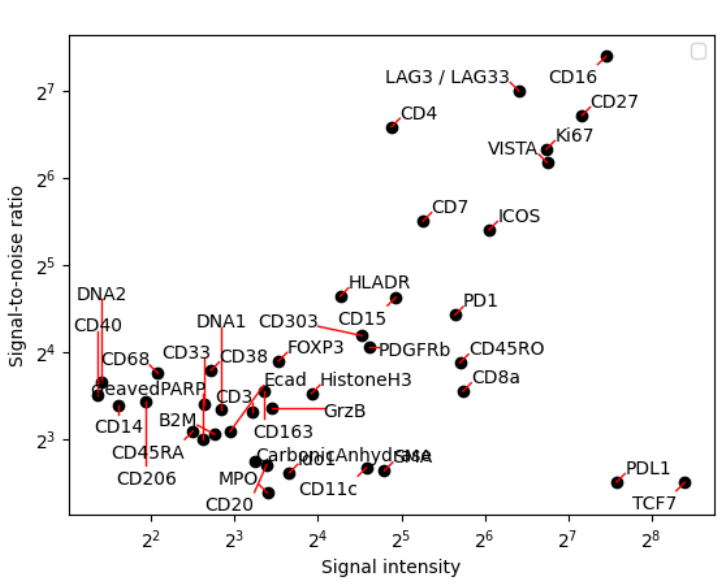


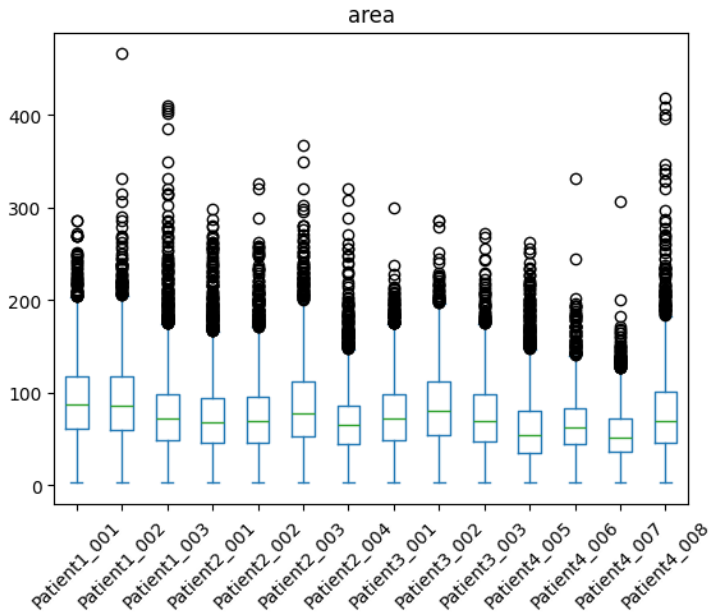
VIB DATA CORE

PR to add native support in SpatialData: https://github.com/scverse/spatialdata/pull/842

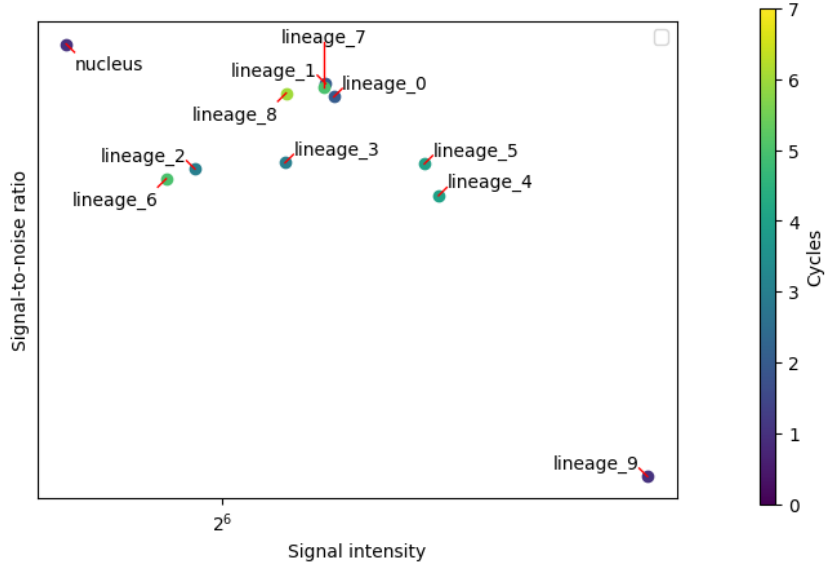
Quality control that scales to your whole dataset

Similar to IMCDataAnalysis, but in **Python**





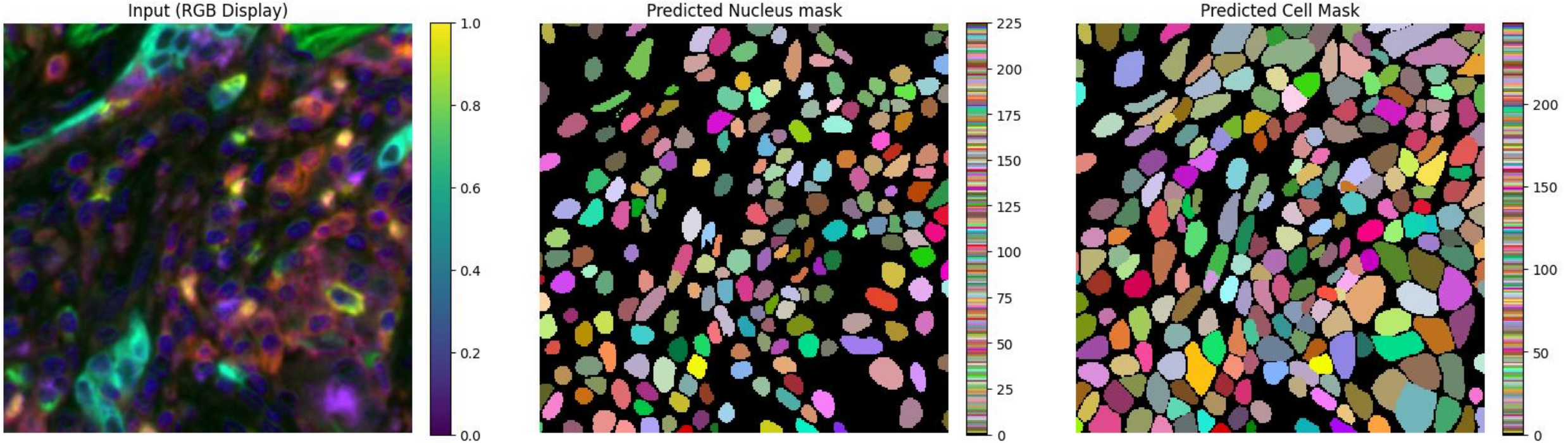
New plots with support for metadata e.g. **cycle order**



***IMCDataAnalysis**: An end-to-end workflow for multiplexed image processing and analysis* https://doi.org/10.1038/s41596-023-00881-0

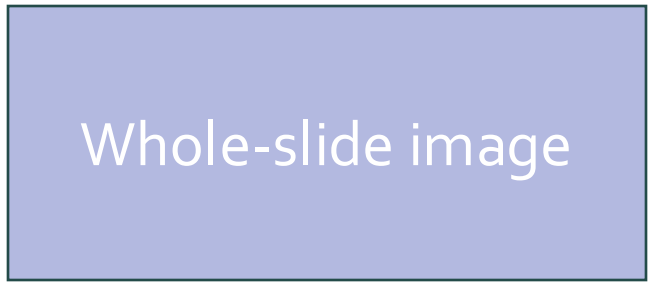
Harpy can apply the latest cell segmentation models

Works on **any combination** of image channels in **any order**
Outputs both predicted **nucleus** and **cell** mask



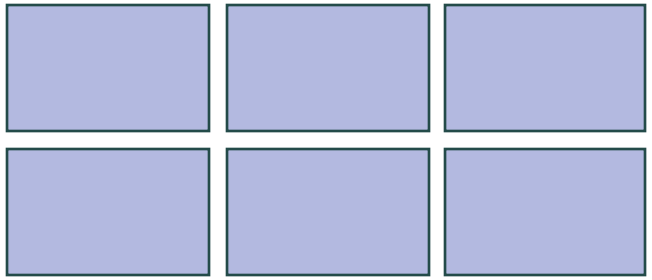
Goldsborough, T. et al. (2024) 'A novel channel invariant architecture for the segmentation of cells and nuclei in multiplexed images using InstanSeg'. *bioRxiv*, p. 2024.09.04.611150. Available at: https://doi.org/10.1101/2024.09.04.611150.

Harpy is fast because of Dask parallelisation



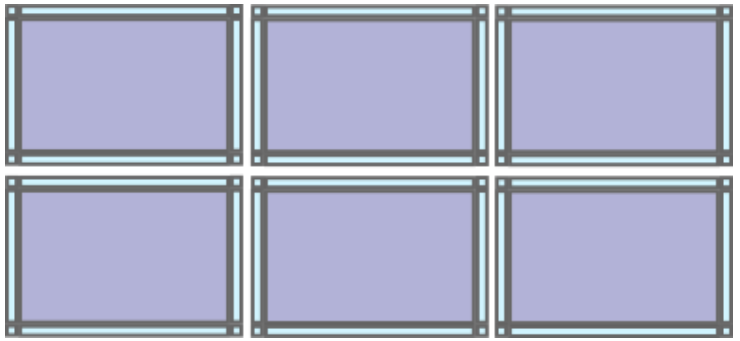
Model apply on WSI

Out-of-memory
Very slow



Model apply on chunks

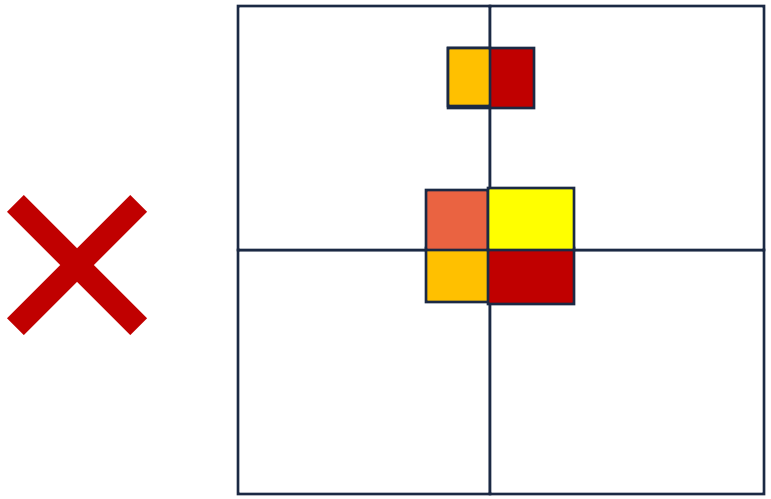
Fast
Border artefacts



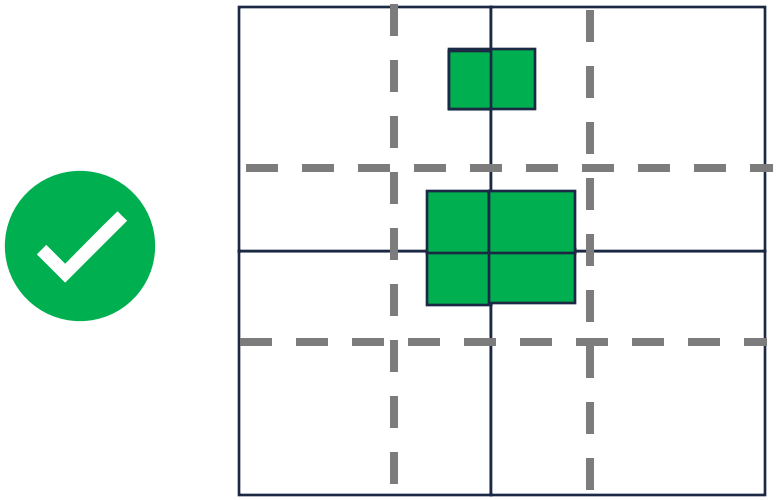
Model apply on chunks via Dask **map_overlap**

Fast, no memory issues
No artefacts
No manual chunks

Overlapping tiles are correct and quick



6 cells segmented using
naive apply over 4 separate tiles
= tile artefacts

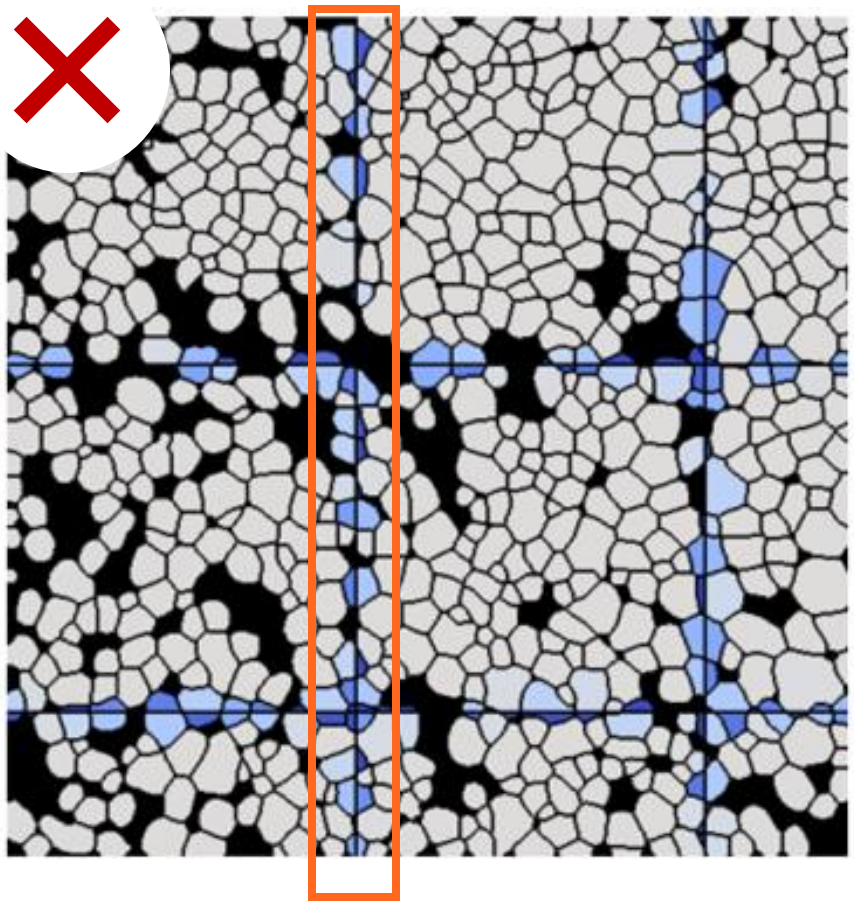


2 cells segmented using
apply over 4 tiles **with Dask map_overlap**
= no tile artefacts… if you merge labels correctly!

More background on tiled processing: https://haesleinhuepf.github.io/BioImageAnalysisNotebooks/32_tiled_image_processing/readme.html

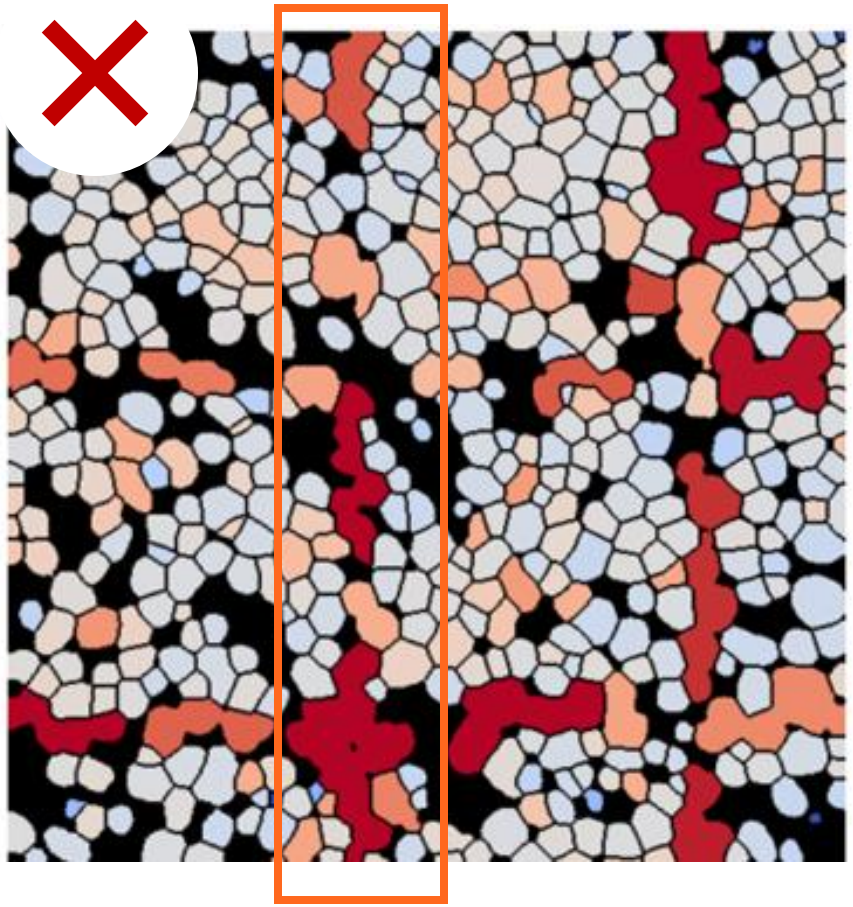
Good label merging is essential

Naive dask.map_overlap



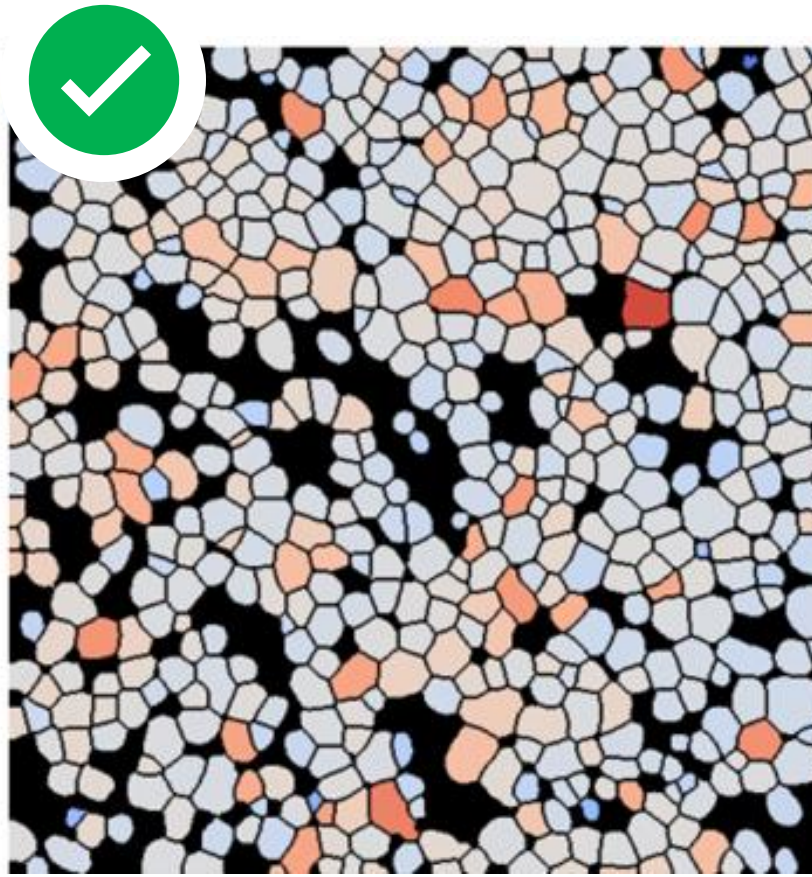
No label merges at tile borders

Naive connected_components or squidpy.segment



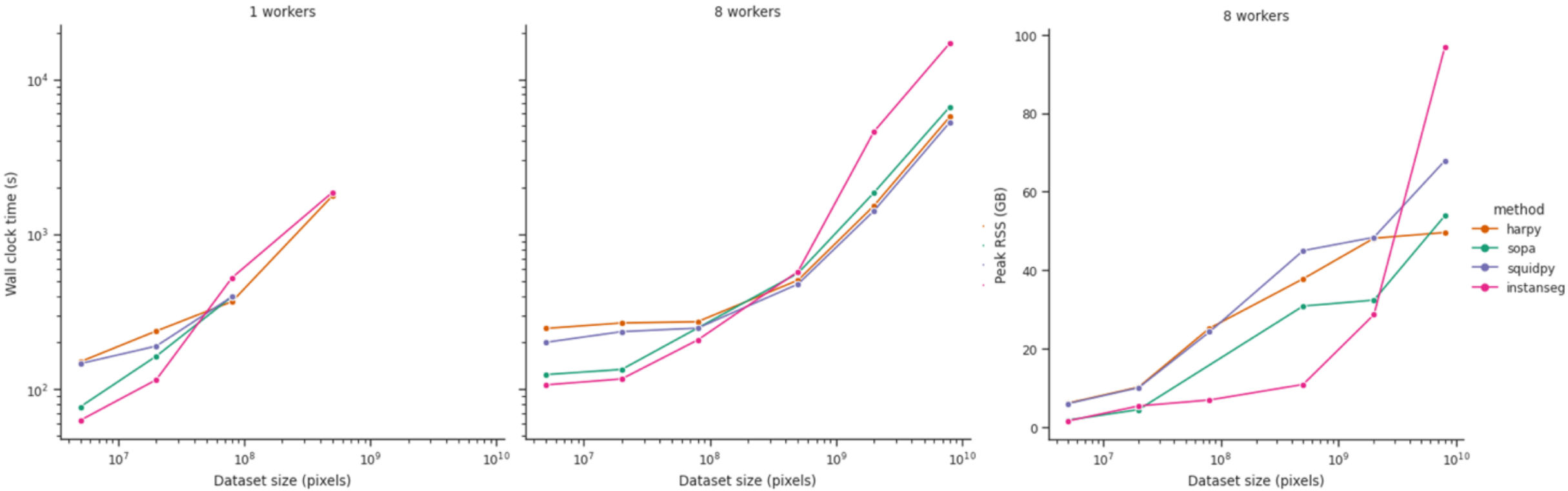
Too many label merges at tile borders

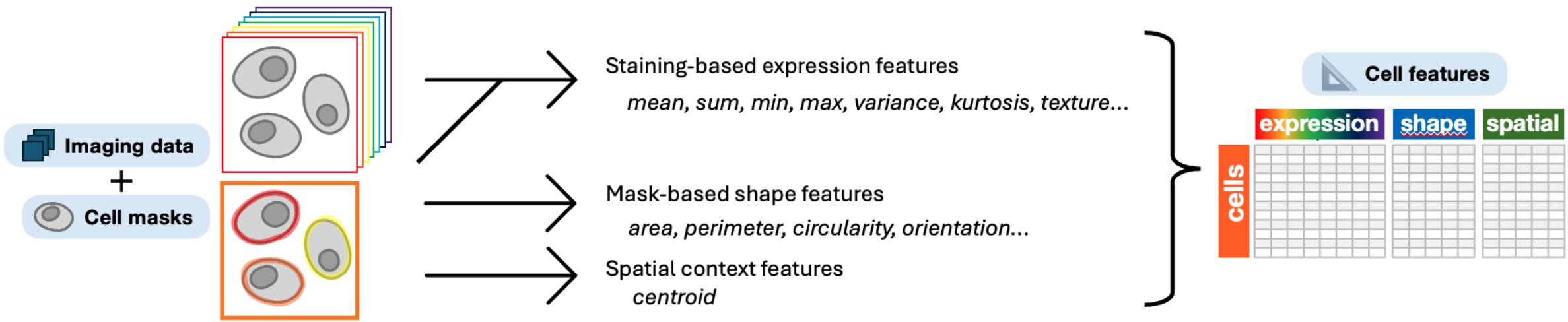
harpy.im.segment

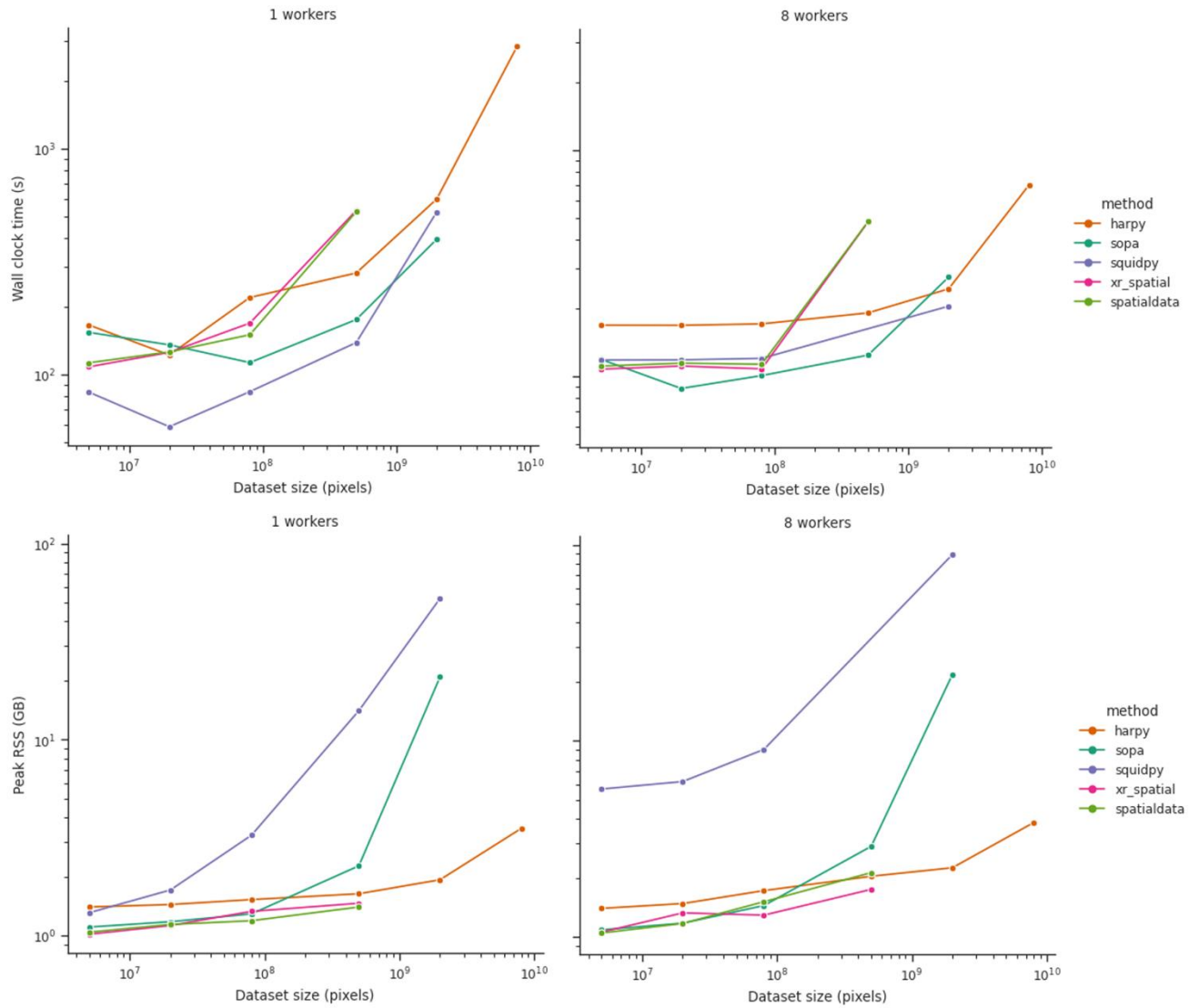


Correct label merges

Discussion on upstreaming this approach https://github.com/scverse/spatialdata/pull/664

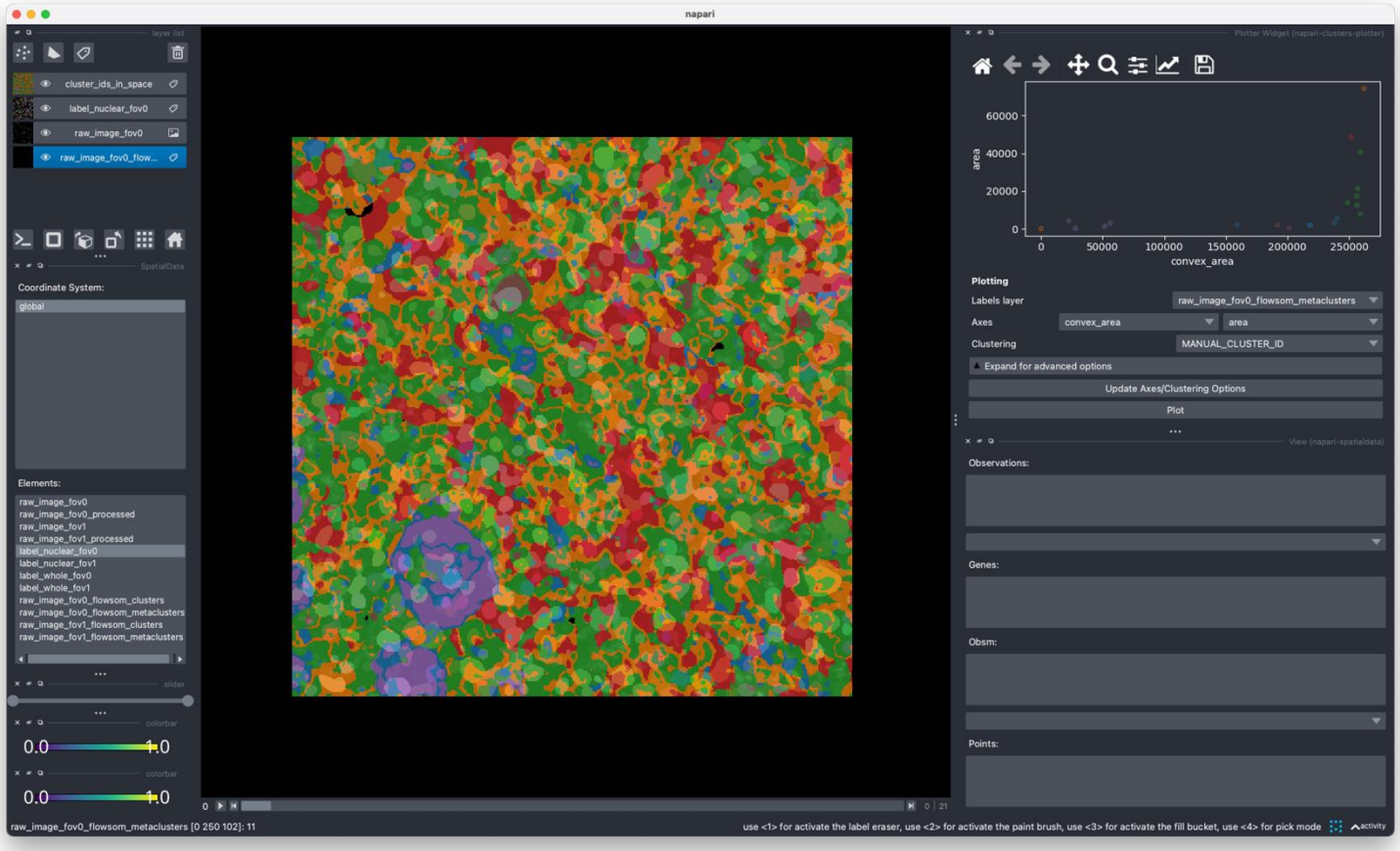

1 workers
8 workers
8 workers
Wall clock time (s)
Peak RSS (GB)
Dataset size (pixels)
Dataset size (pixels)
Dataset size (pixels)
method
harpy
sopa
squidpy
instanseg


Imaging data
+
Cell masks
Staining-based expression features
mean, sum, min, max, variance, kurtosis, texture...
Mask-based shape features
area, perimeter, circularity, orientation...
Spatial context features
centroid
Cell features
expression
shape
spatial
cells


1 workers
8 workers
Wall clock time (s)
Peak RSS (GB)
Dataset size (pixels)
method
harpy
sopa
squidpy
xr_spatial
spatialdata

Interactive unsupervised clustering

Cell and **pixel** clustering workflow using FlowSOM in Python

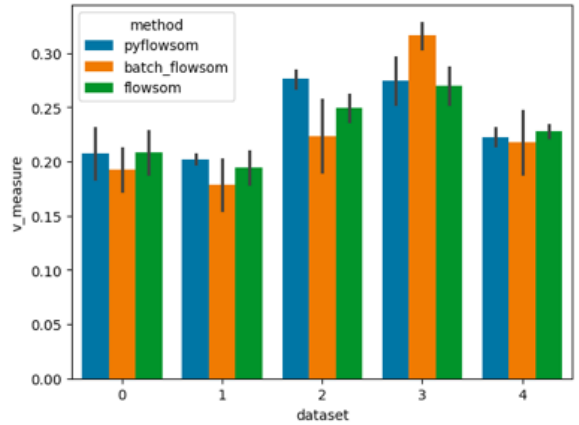


Visualize results in notebook, with napari or with other SpatialData plugins like e.g. napari-spatialdata

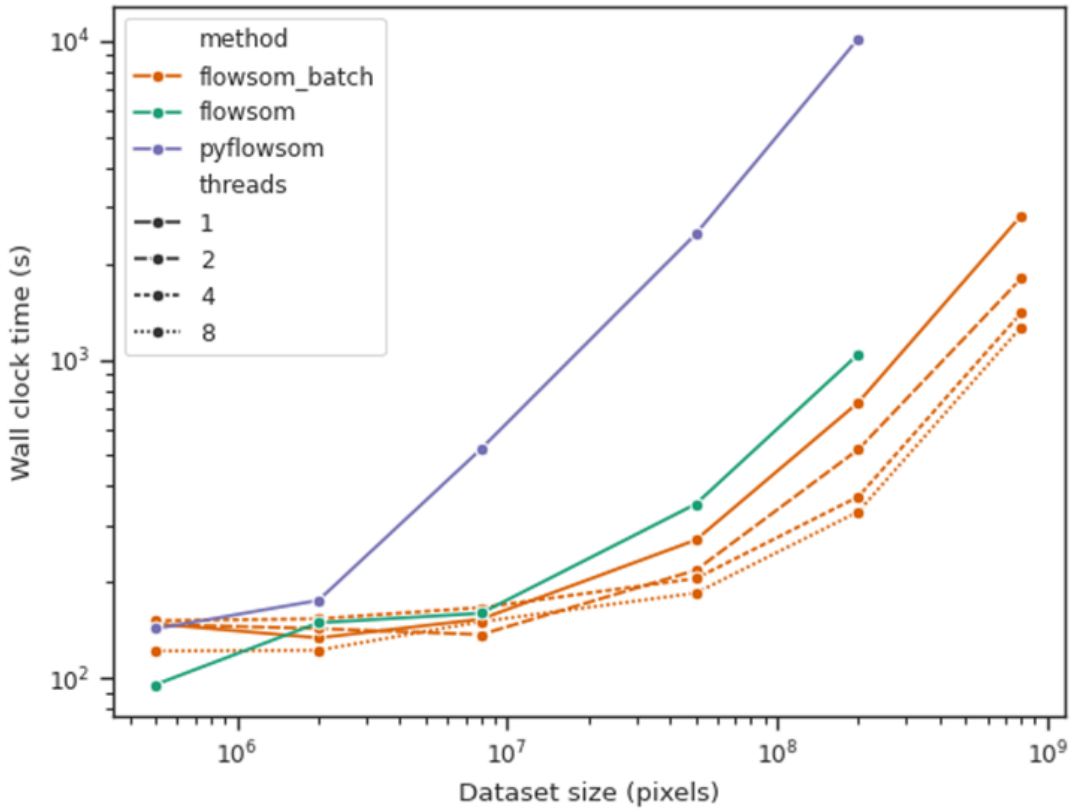
Liu, C.C., Greenwald, N.F., Kong, A., McCaffrey, E.F., Leow, K.X., Mrdjen, D., Angelo, M., 2022. Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering. https://doi.org/10.1101/2022.08.16.504171

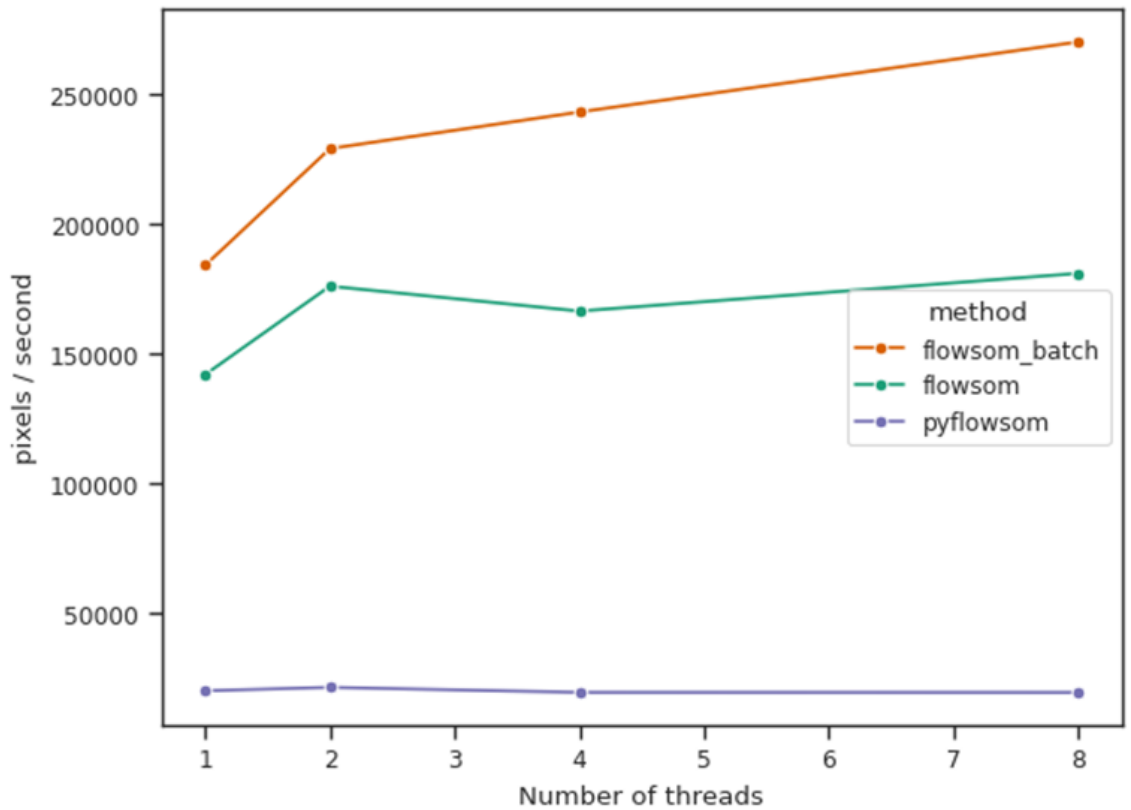
A **Example unsupervised clustering workflow**

B **Quality comparison of SOM cluster models**



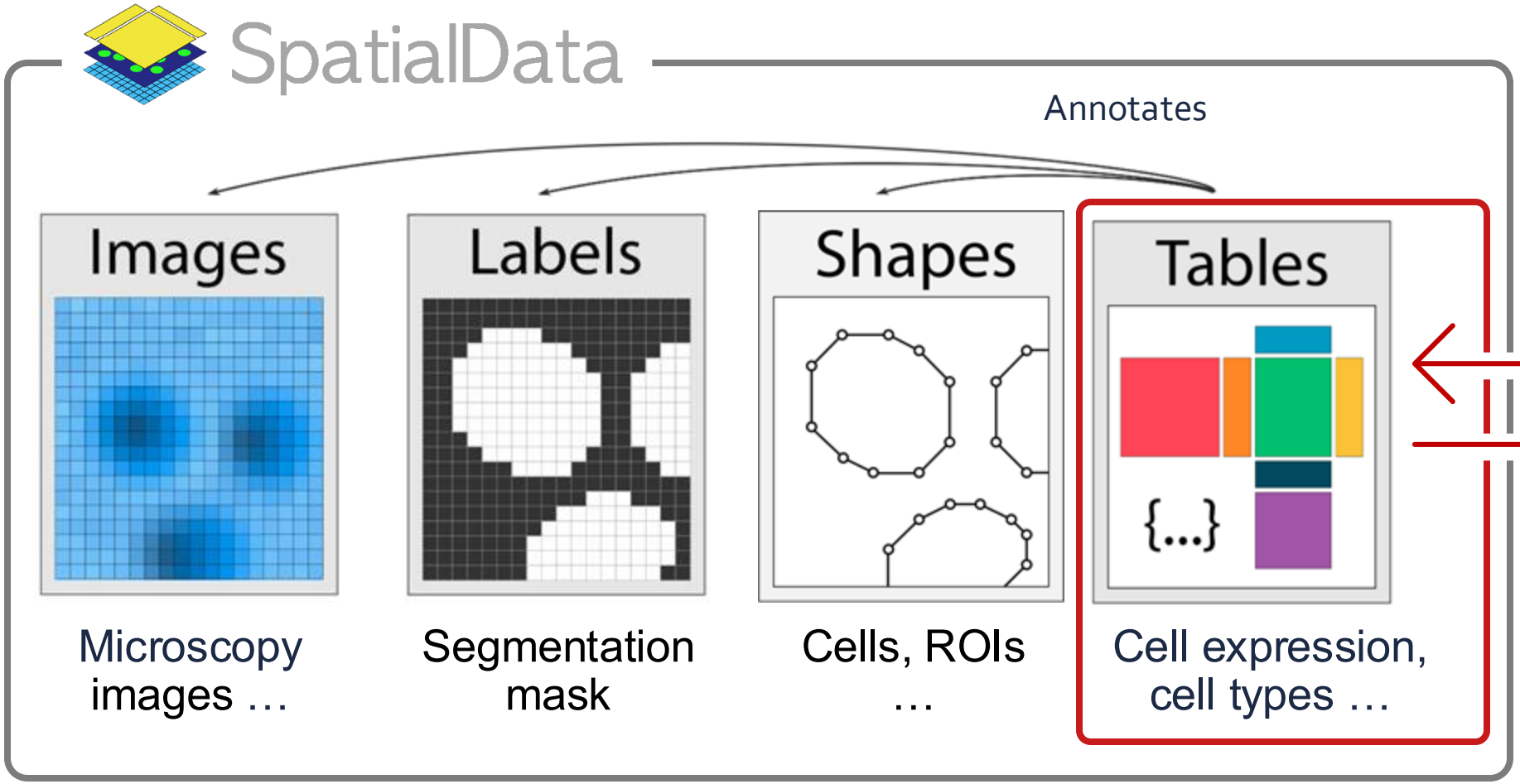
C **Scalability comparison of SOM cluster models**

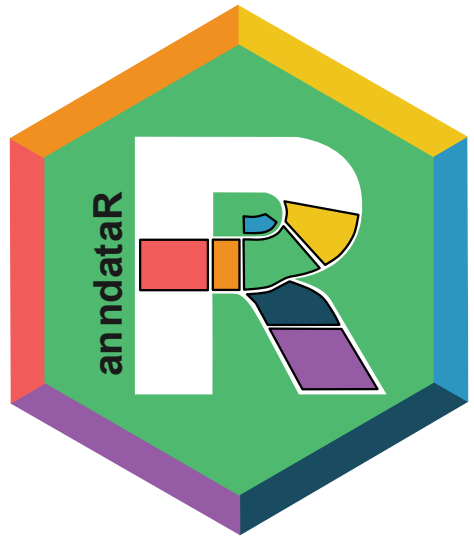




Harpy output is interoperable with R

Louise Deconinck





SeuratObject
SingleCellExperiment

Cannoodt R, Zappia L, Morgan M, Deconinck L (2025). *anndataR: AnnData interoperability in R*. R package version 0.99.0, https://github.com/scverse/anndataR, https://anndatar.data-intuitive.com/.

Future support for complete SpatialData object in R: https://github.com/HelenaLC/SpatialData

More on interoperability in our workshop:

Polyglot programming for single-cell analysis

AUTHORS

Benjamin Rombaut

Louise Deconinck

Robrecht Cannoodt

AFFILIATIONS

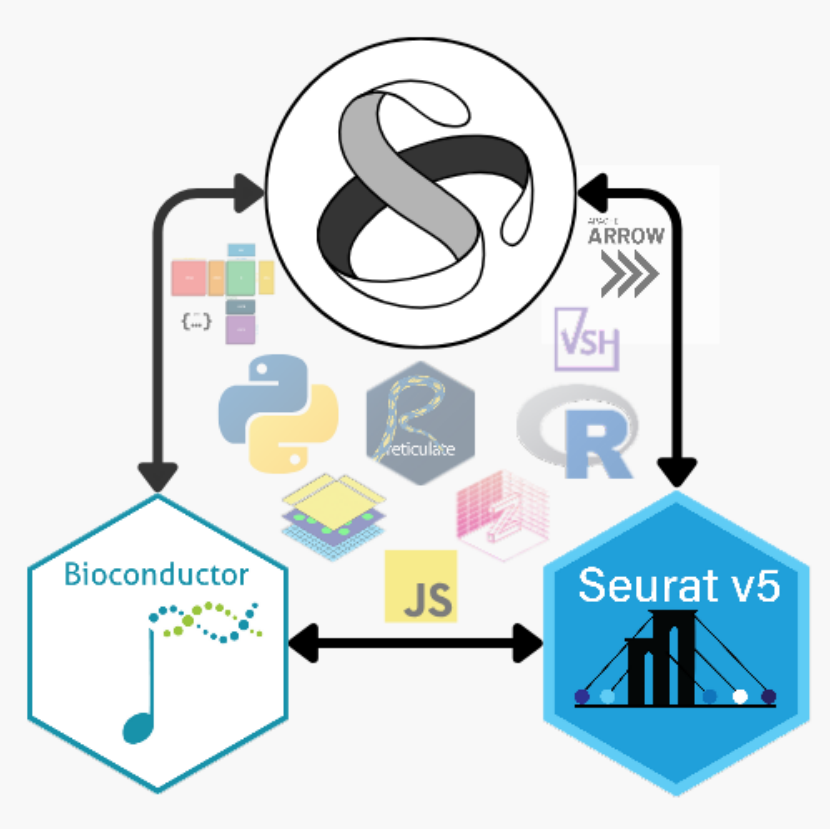
Saeys lab
VIB Center for Inflammation Research
Ghent University

Saeys lab
VIB Center for Inflammation Research
Ghent University

Data Intuitive

PUBLISHED

September 12, 2024



Overview of the **different levels** of interoperability.

Integrating them in a **single workflow**.

https://saeyslab.github.io/polygloty/

Scverse Conference 2024 workshop

Rich parallel abstractions ✅	Channels only
Easy debugging ✅	Hard to debug
Python-only	✅ **Polyglot**
Bio docs can be sparse	✅ Great **nf-core** community

Python packages

Workflows

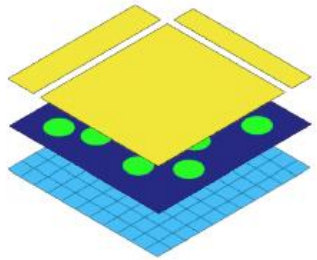
Module
Module
Module

Module
Module
Module

dask

Module
Module
Module

Module
Module
Module



SpatialData

nf-core

Viash

4 Harpy highlights

1. Dataset-wide quality control
2. Scalable whole-slide image processing
3. Interactive unsupervised clustering
4. Interoperability with R

https://github.com/saeyslab/harpy

SCIENCE MEETS LIFE

Acknowledgments

VIB-UGent Center for Inflammation Research
SaeysLab
Lotte Pollaris
Ruth Seurinck
Sofie Van Gassen
…
Yvan Saeys

VIB Spatial Catalyst
Arne Defauw
Frank Vernaillen
Julien Mortier
Evelien Van Hamme

Questions?
Benjamin.Rombaut@UGent.be

or contact the **VIB Spatial Catalyst**
spatial@vib.be

Benjamin Rombaut

3th of March 2024

Use a scheduler to avoid memory errors

Manual solution

Limiting cores, partitions, parallellism…

Write out temporary results to disk

Automated solution

Give all tasks to the scheduler

Spills older objects to disk if approaching memory limit

Can also work for GPU object spilling

Problem now is task granularity and correct memory estimation

Reasons for (not) using proprietary software

e.g. MACS iQ View provided with MACSima platform

Strengths

- End-to-end graphical and user-friendly workflow
- Advanced cell type annotation with visual feedback
- **Python API** for scripts and batch sample processing
- **Good for quality assessment and ground truth annotations**

Weaknesses

- Expensive license and fixed workstation needed
- Not tested on non-MACSima data
- Limited analysis options (lacks state-of the art tools)

Open-source data workflows

MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging

https://doi.org/10.1038/s41592-021-01308-y (more **general**)

Nextflow/Python https://nf-co.re/mcmicro/dev

IMCDataAnalysis: An end-to-end workflow for multiplexed image processing and analysis

https://doi.org/10.1038/s41596-023-00881-0 (more for **IMC** data)

Python/R https://bodenmillergroup.github.io/IMCDataAnalysis/

ark-analysis: Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering

https://doi.org/10.1038/s41467-023-40068-5 (more for **MIBI** data)

Python https://github.com/angelolab/ark-analysis

MIBI workshop 2022 (https://www.youtube.com/playlist?list=PLjNbkEm4vA26o5YvWKeyHXF8HjTJc7yB0)

Spatial Biology workshop 2023 (https://www.angelolab.com/spatial-biology-workshop)

Harpy: https://github.com/saeyslab/harpy