Towards scalable and interactive pipelines for spatial biology

Yvan Saeys

yvan.saeys@ugent.be

VIB

SaeysLab
Data mining and Modelling for Biomedicine

saeyslab.sites.vib.be

Lyon, France

SaeysLab

Data mining and Modelling for Biomedicine

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

VIB CENTER FOR INFLAMMATION RESEARCH

Trustworthy Models of Cells and Tissues

HARPY

Arrow

dynverse

CytoNorm

MACHINE LEARNING

SPATIAL OMICS

CELL-CELL COMMUNICATION

GENE REGULATION

CLINICAL APPLICATIONS

SCIP

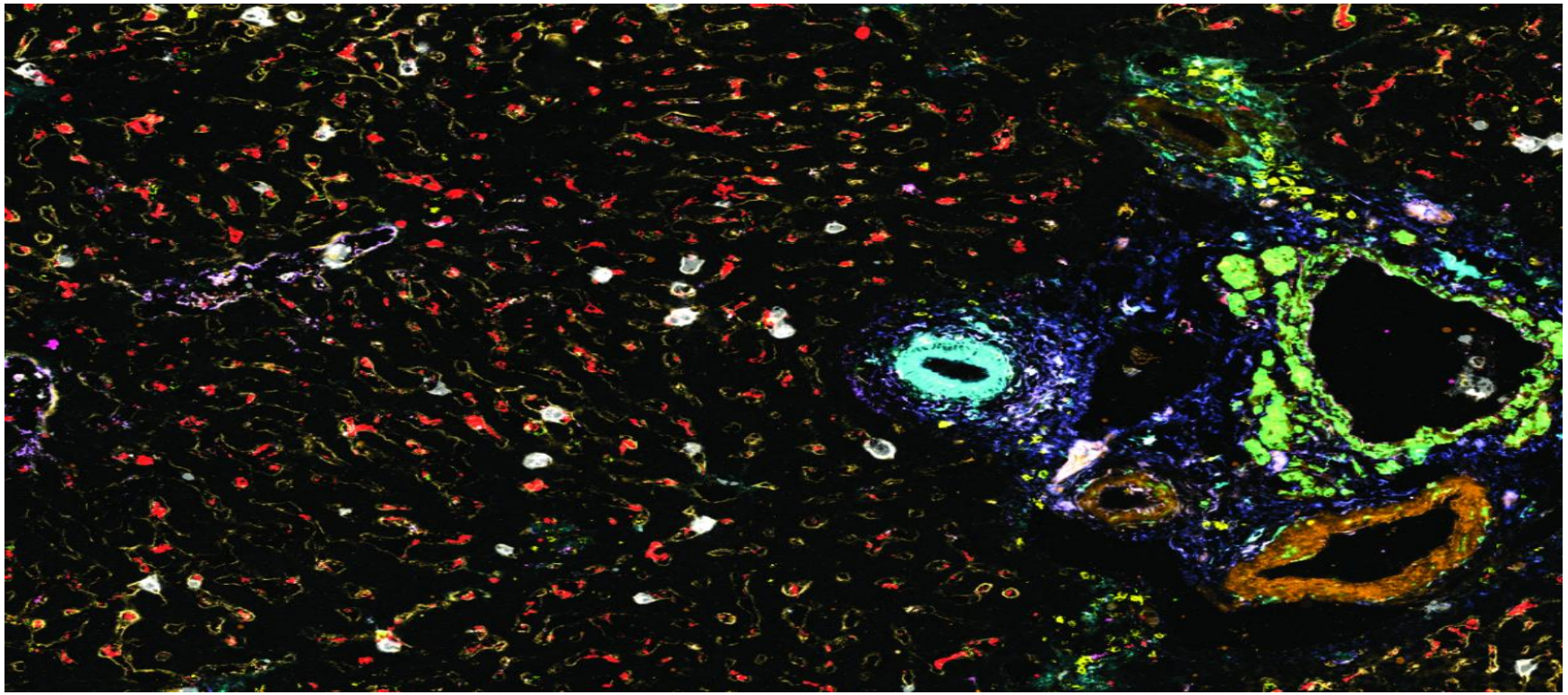
Scalable cytometry image processing

NicheNet

FlowSOM

funkyheatmap

Next generation microscopes…again



Guilliams M et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Cell. 2022 Jan 20;185(2):379-396.e38

Tissue architecture

Spatial niches
(e.g. tumor microenvironment)

Find new types of (spatial) biomarkers

Why do we need spatial omics ?

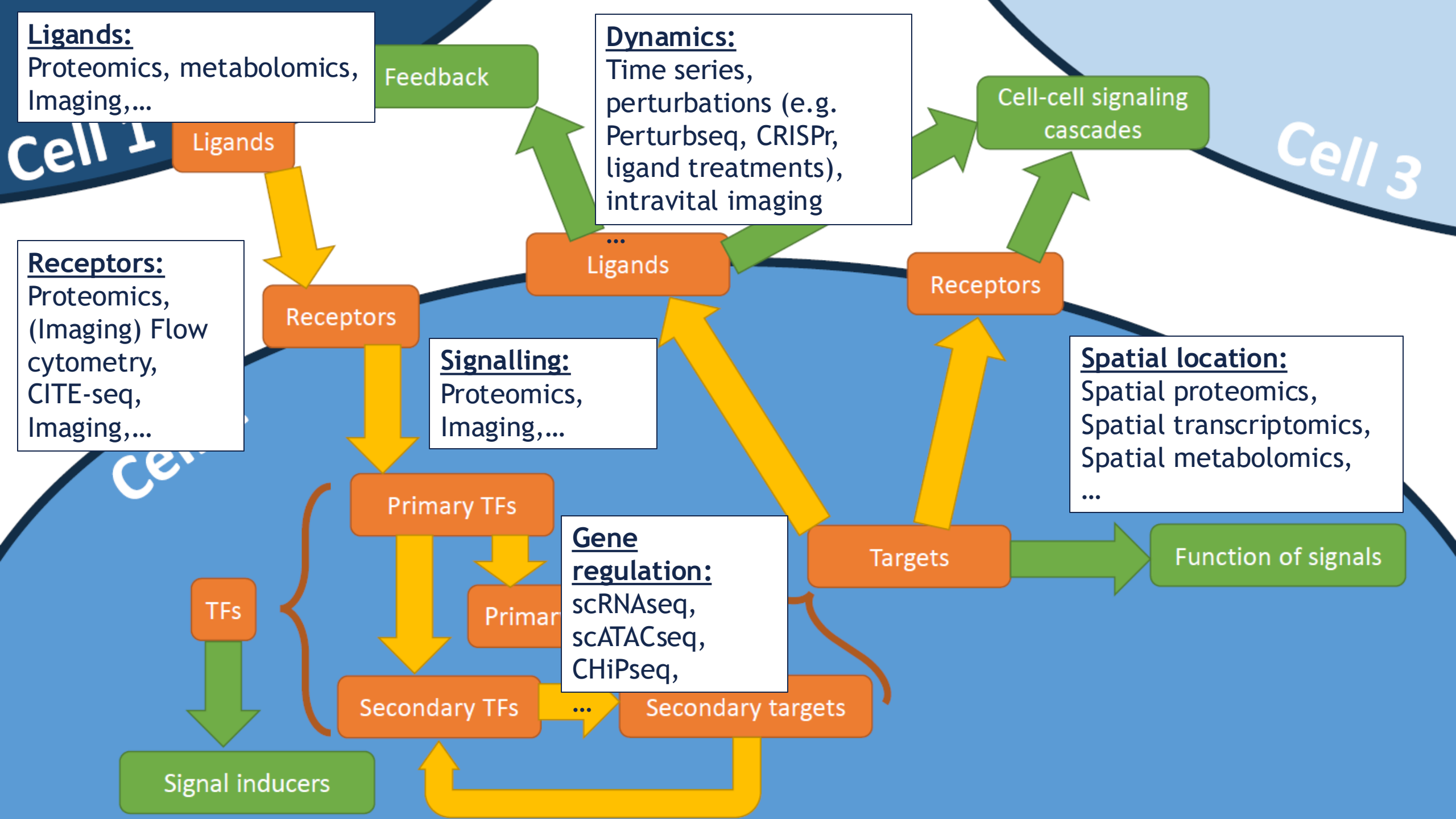
Cellular heterogeneity

Spatial gradients

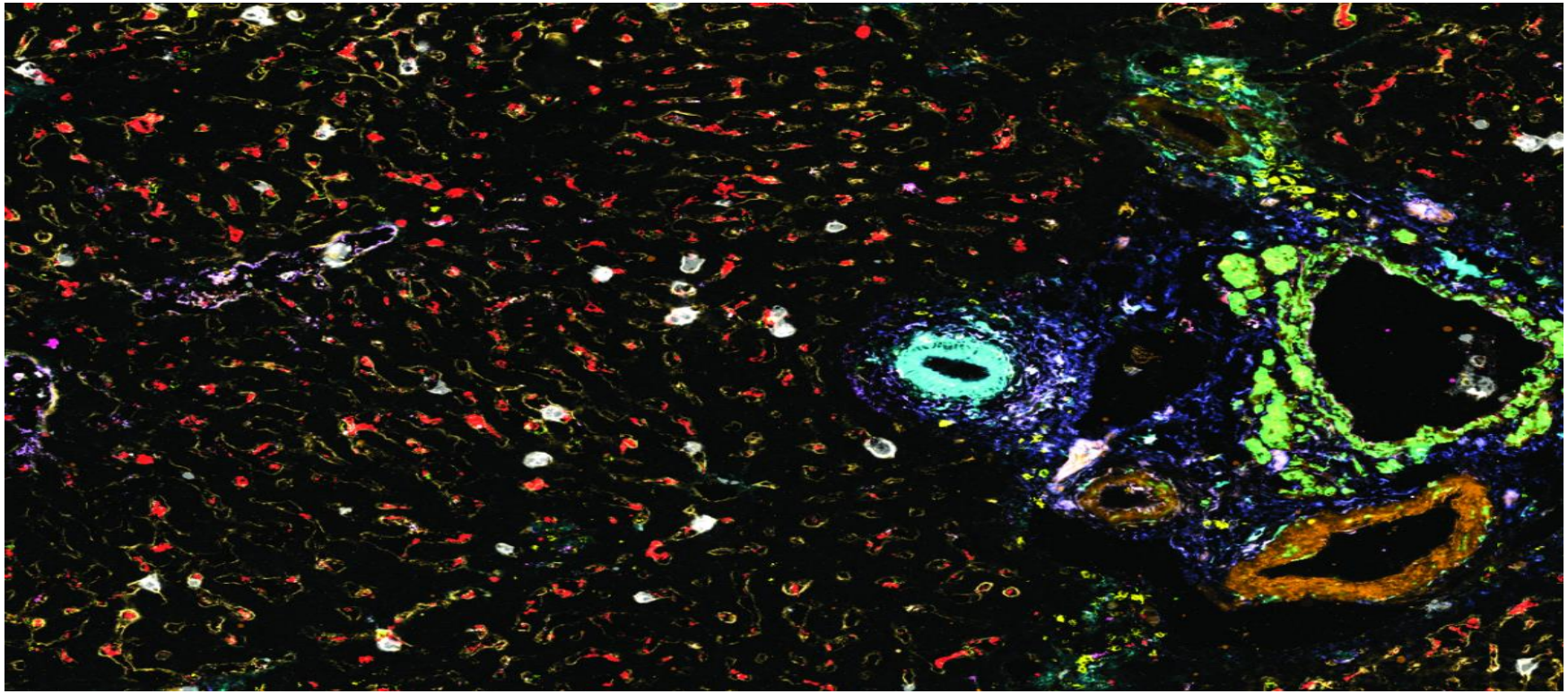
Cell-cell interactions

Computational models of cells and tissues

What types of data do we need to study *functional* cell-cell communication ?


Ligands:
Proteomics, metabolomics, Imaging,…
Feedback
Dynamics:
Time series, perturbations (e.g. Perturbseq, CRISPr, ligand treatments), intravital imaging
…
Cell-cell signaling cascades
Cell 1
Ligands
Cell 3
Receptors:
Proteomics, (Imaging) Flow cytometry, CITE-seq, Imaging,…
Receptors
Ligands
Receptors
Signalling:
Proteomics, Imaging,…
Spatial location:
Spatial proteomics, Spatial transcriptomics, Spatial metabolomics, …
Primary TFs
Gene regulation:
scRNAseq, scATACseq, CHiPseq, …
Targets
Function of signals
TFs
Secondary TFs
Secondary targets
Signal inducers

Towards functional spatial “omics”

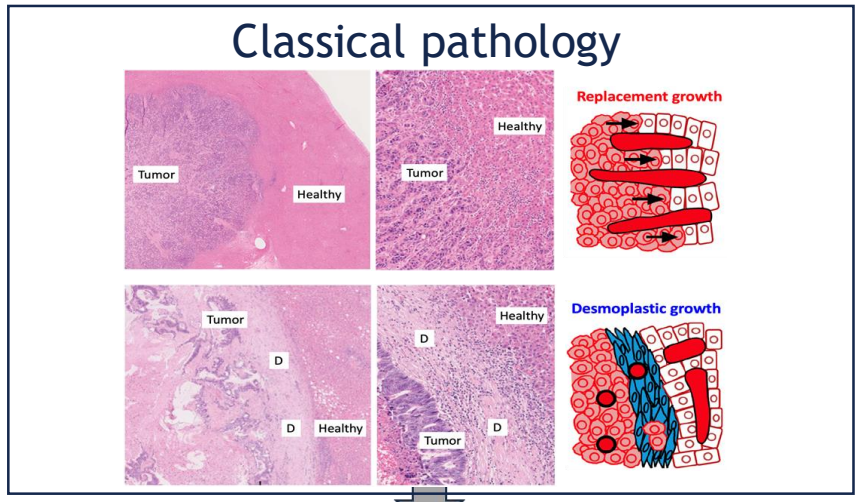


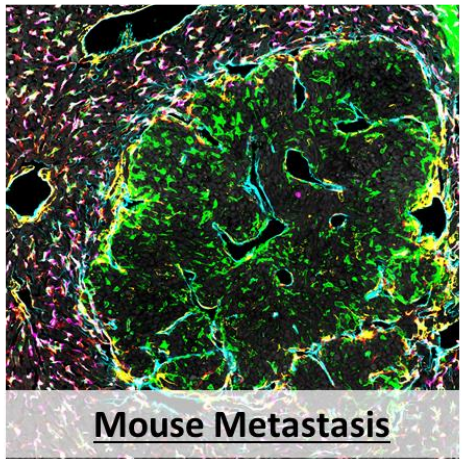
multiple modalities

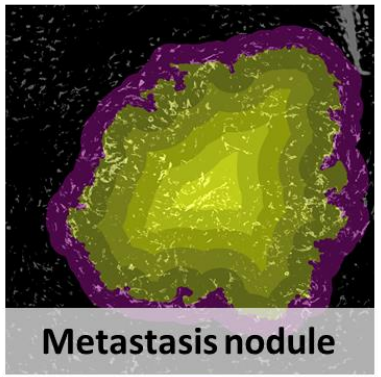
+ spatial context

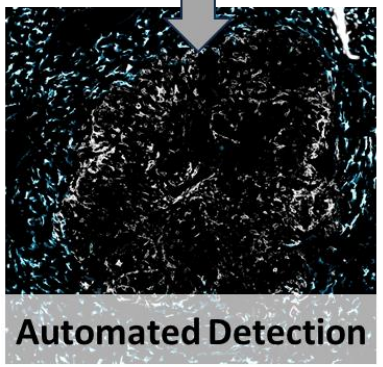
+ AI/computational models of cellular interactions and gene regulation

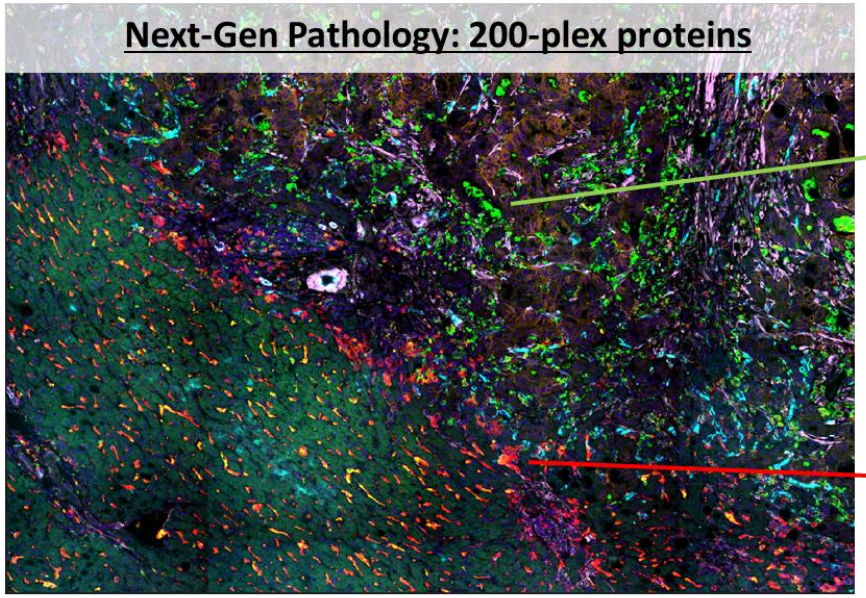
Building the foundations for next-generation pathology











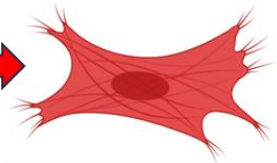
Suppressive Macrophages

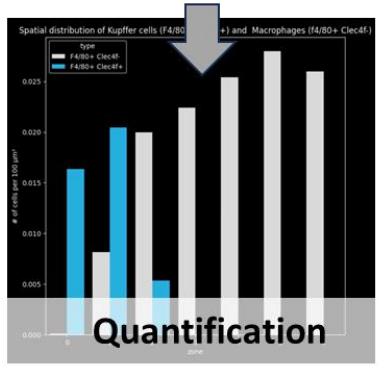
Same Macrophages found in Mouse

Pro-fibrotic Macrophages

TGF-b

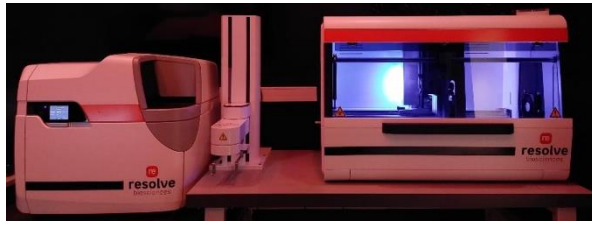
Activated Fibroblasts





The reality check of spatial omics:
#1 A variety of platforms calls for unified pipelines

resolve biosciences



vizgen

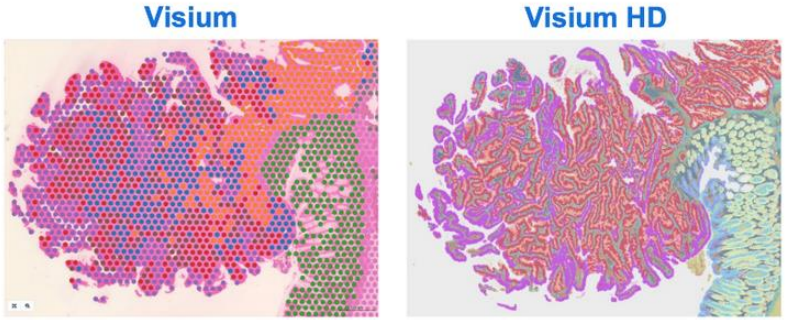
nanoString

10x GENOMICS

BGI 华大

StereoSeq
STOmics

MC1	Merscope	CosMx	Xenium
FF	FF/FFPE	FF/FFPE	FF/FFPE
100 plex	140, 300, 500 plex	1000 plex	350 plex (100 custom)
Flexible panel	Flexible panel	Fixed panel	Fixed and flexible panel
RNA only	RNA + limited IHC	RNA + IHC	RNA + IHC
ROI small	ROI medium	ROI big	ROI big



Sang-aram, C. et al. (2024) **Spotless, a reproducible pipeline for benchmarking cell type deconvolution in spatial transcriptomics.** *eLife* **12**:RP88431.

We don’t want a separate pipeline for each platform -> Unified pipelines

The reality check of spatial omics:
#2 The size of the data calls for scalable solutions

Raw images

Name	Size
micron_to_mosaic_pixel_transfor…	225 B
mosaic_Cellbound1_z0.tif	23.6 GB
mosaic_Cellbound1_z1.tif	23.6 GB
mosaic_Cellbound1_z2.tif	23.6 GB
mosaic_Cellbound1_z3.tif	23.6 GB
mosaic_Cellbound1_z4.tif	23.6 GB
mosaic_Cellbound1_z5.tif	23.6 GB
mosaic_Cellbound1_z6.tif	23.6 GB
mosaic_Cellbound2_z0.tif	23.6 GB
mosaic_Cellbound2_z1.tif	23.6 GB
mosaic_Cellbound2_z2.tif	23.6 GB
mosaic_Cellbound2_z3.tif	23.6 GB
mosaic_Cellbound2_z4.tif	23.6 GB
mosaic_Cellbound2_z5.tif	23.6 GB

Images: 23,6 GB* 7 z-stacks * 5 stainings=
822 GB of images!

Sometimes processed data is already available

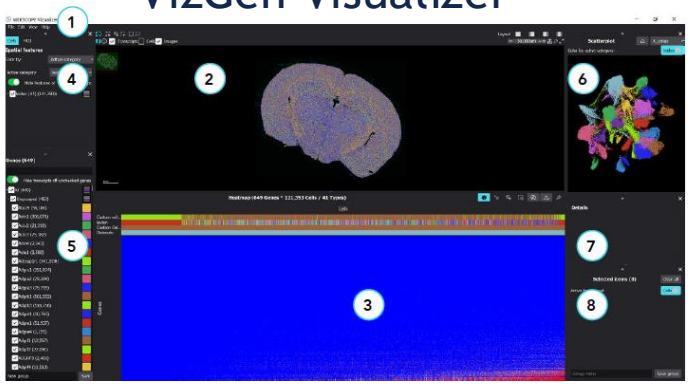
Filter by name prefix only | Filter | Filter objects and folders

Name	Size
cell_boundaries/	–
cell_by_gene.csv	987.1 MB
cell_metadata.csv	61.7 MB
detected_transcripts.csv	16.1 GB
images/	–

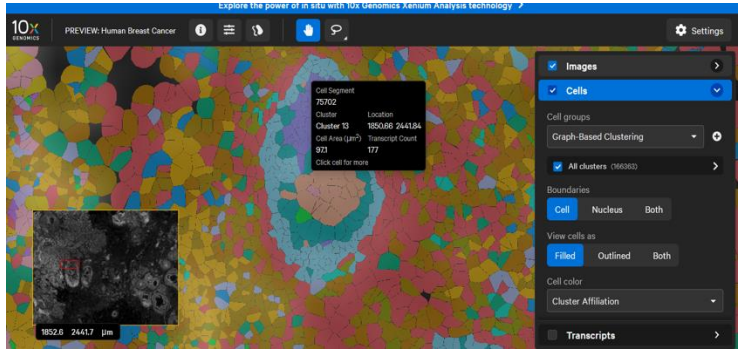
We need scalable solutions that work on many different computing infrastructures

The reality check of spatial omics: #3 Limited functionality of the vendor tools

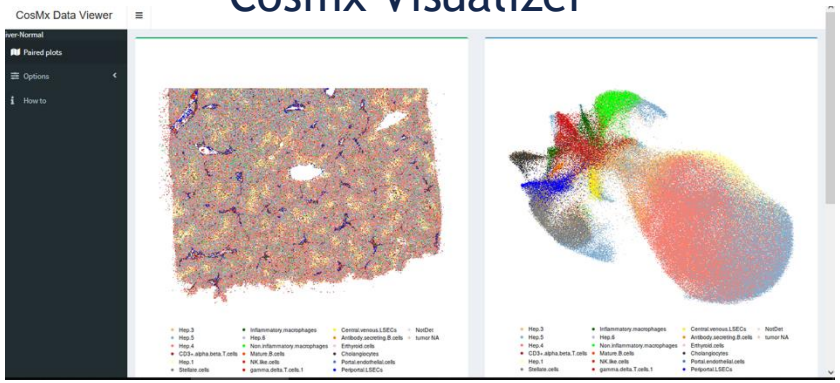
VizGen Visualizer



Xenium Visualizer



Cosmx Visualizer



Pro's

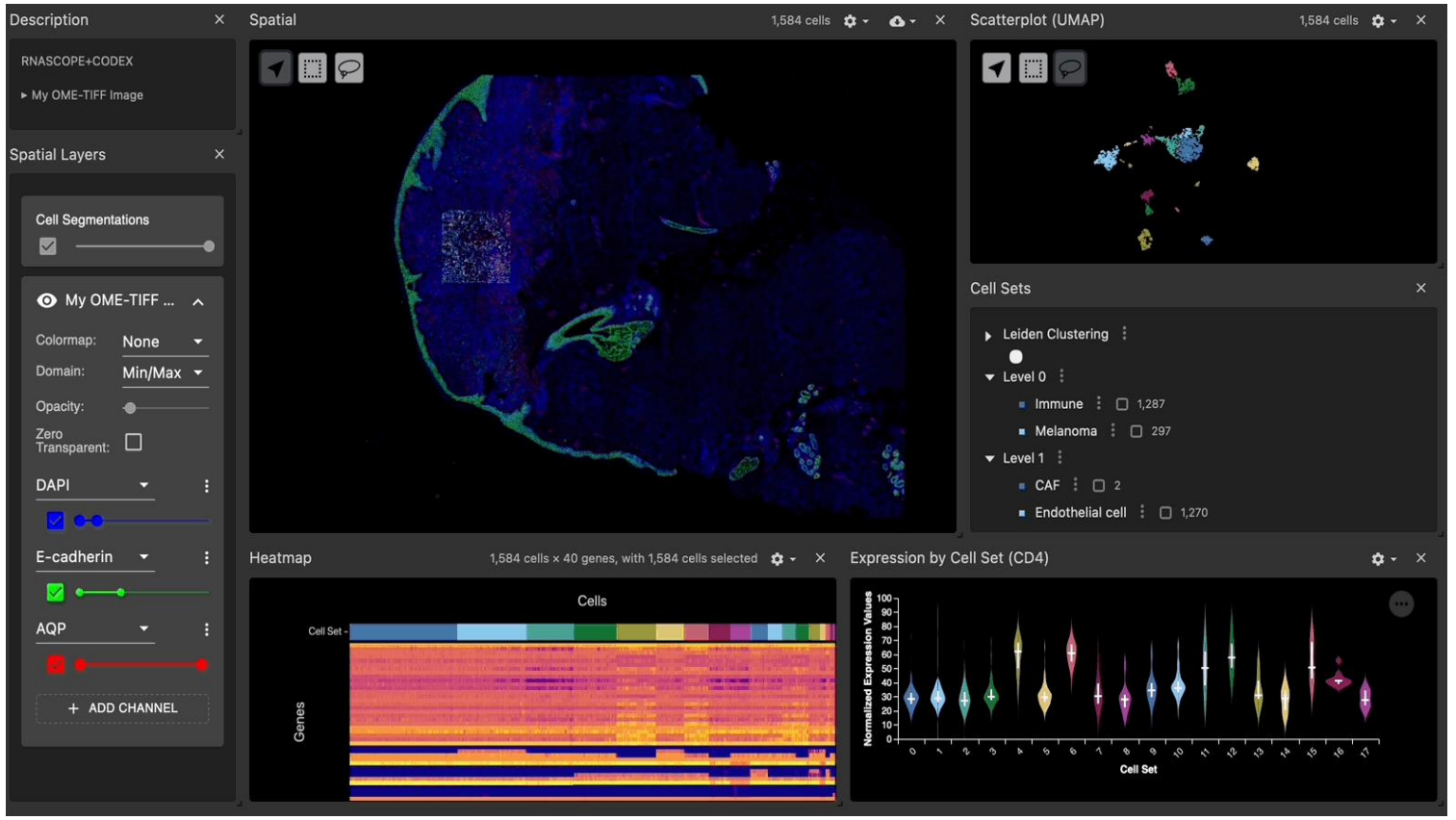
- Quick analysis
- Xenium tool runs on normal laptop
- PI wants to have a quick look at the data
- You can reload your analysed data back in the tools
- Qualitative view of transcripts

Con's

- Different platform for every tool: learning curve
- Limited functionality, dependent on the platform for functionality
- Not optimized for your tissue and research questions
- Lots of back and forth

We need more powerful, open and reproducible tools with better functionality

The reality check of spatial omics:
#4 Best results require the biologist in the loop



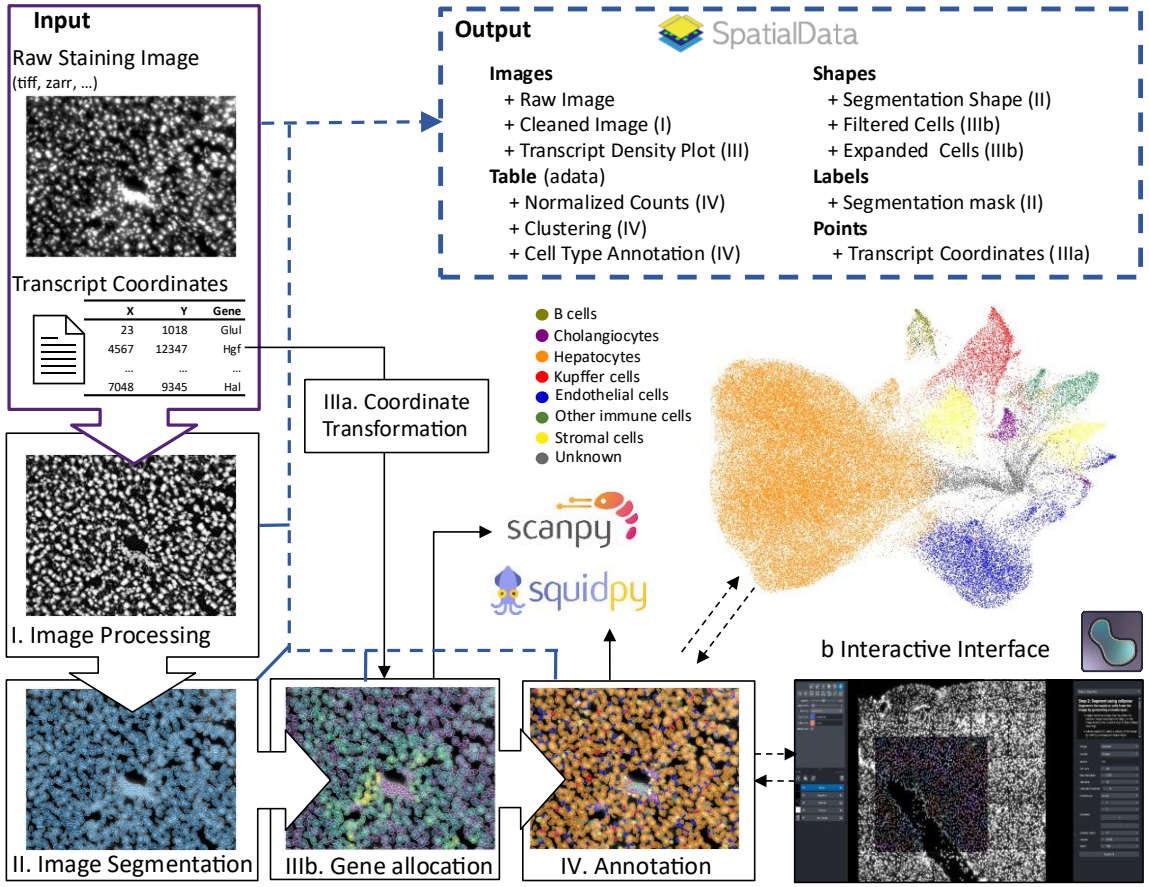
- We need lots of quality controls to check the data to ensure robust biological interpretation

We need interactive tools that allow biologists to play with the data and explore

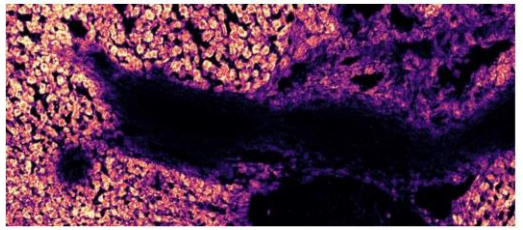
Lotte Pollaris

SPArrOW: a flexible, scalable, modular and interactive Spatial Omics Workflow

SPArrOW: a scalable, modular and interactive workflow for spatial omics with improved quality control

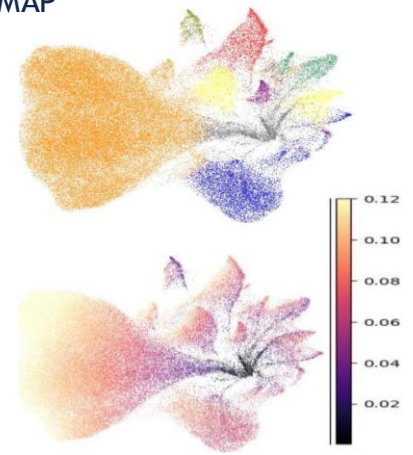


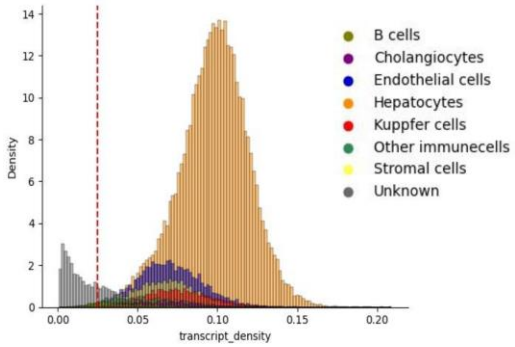
QC metrics and plots



gene	proportion_kept	raw_counts
Dpt	0.717878	110555
Nes	0.770773	22144
Alas2	0.778949	144541
Lhfp	0.783647	112201
Sfrp1	0.784296	40092
Sdc3	0.789354	253928
Cald1	0.793574	1616618
Cybb	0.801593	78828
Bpgm	0.801965	104204
Myh10	0.802167	59813

UMAP

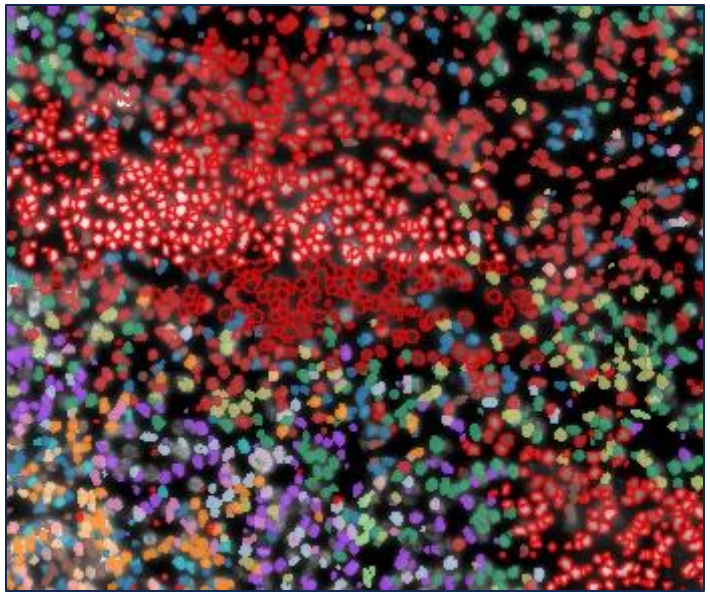




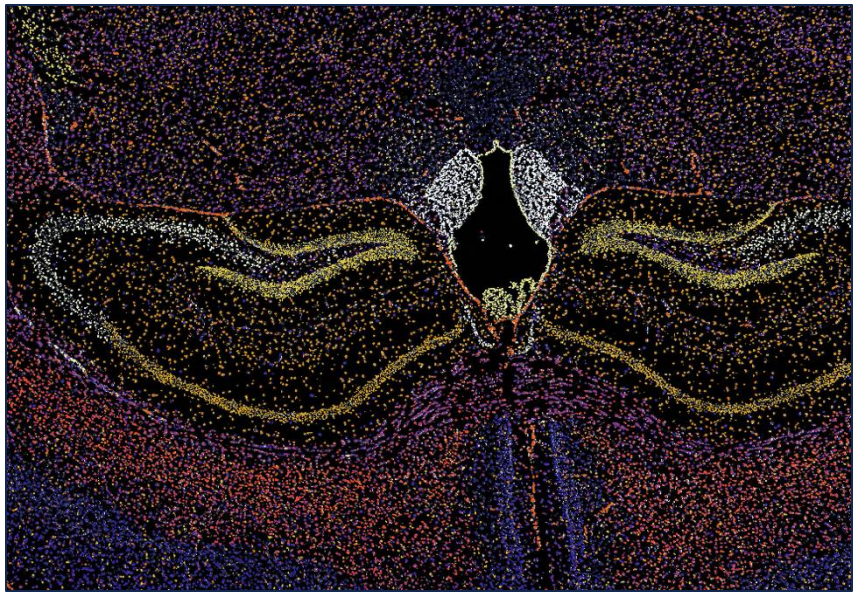
Pollaris L. et al. SPArrOW: a flexible, interactive and scalable pipeline for spatial transcriptomics analysis. bioRxiv 2024.07.04.601829

SPArrOW is currently benchmarked on 80 spatial transcriptomics datasets

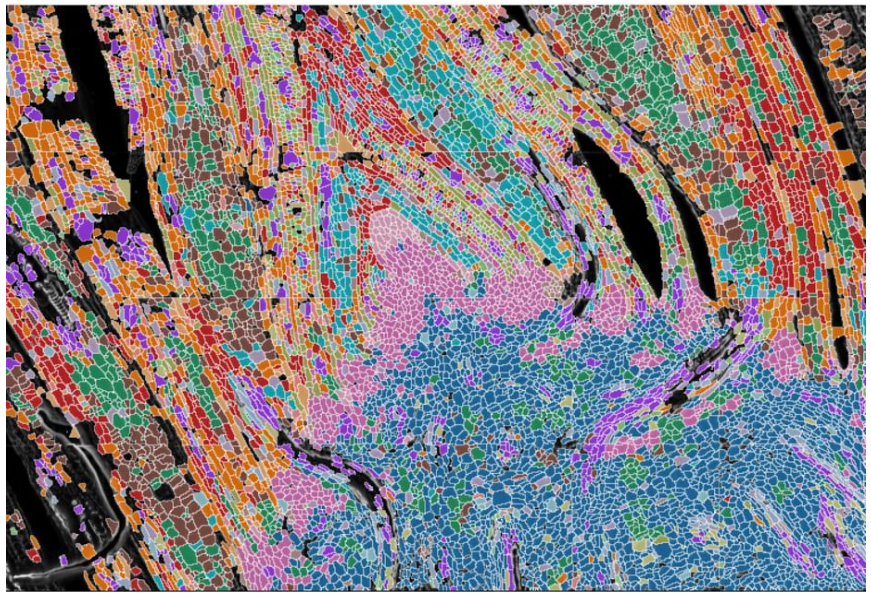
Human Melanoma



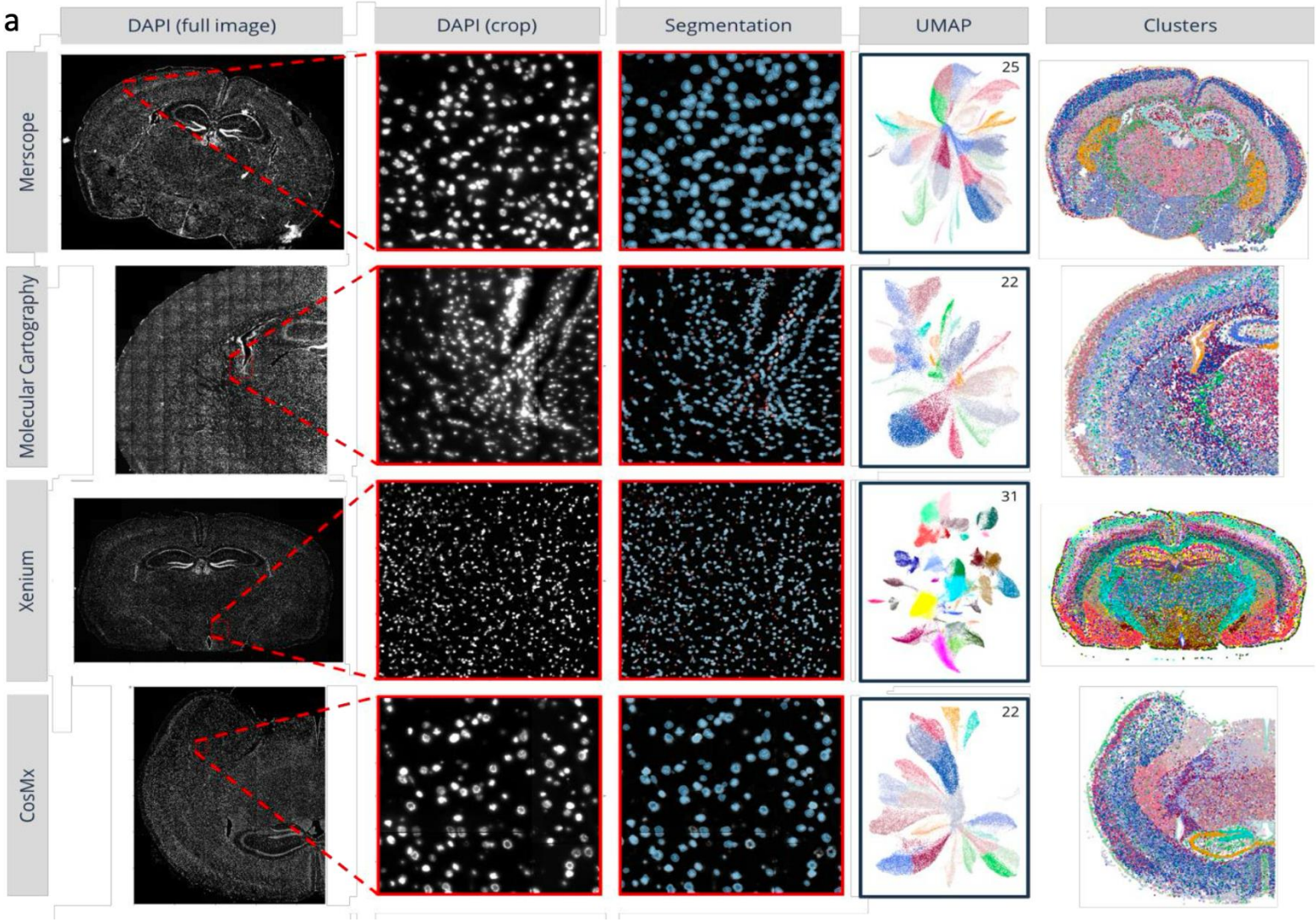
Mouse Brain



Maize Shoot apical meristem



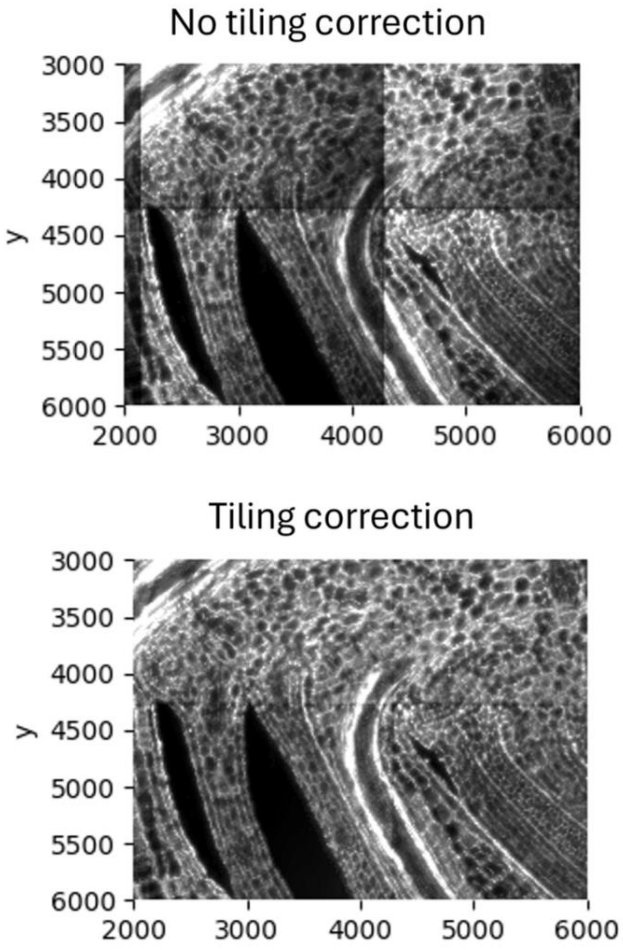
SPArrOW is a flexible framework for spatial transcriptomics



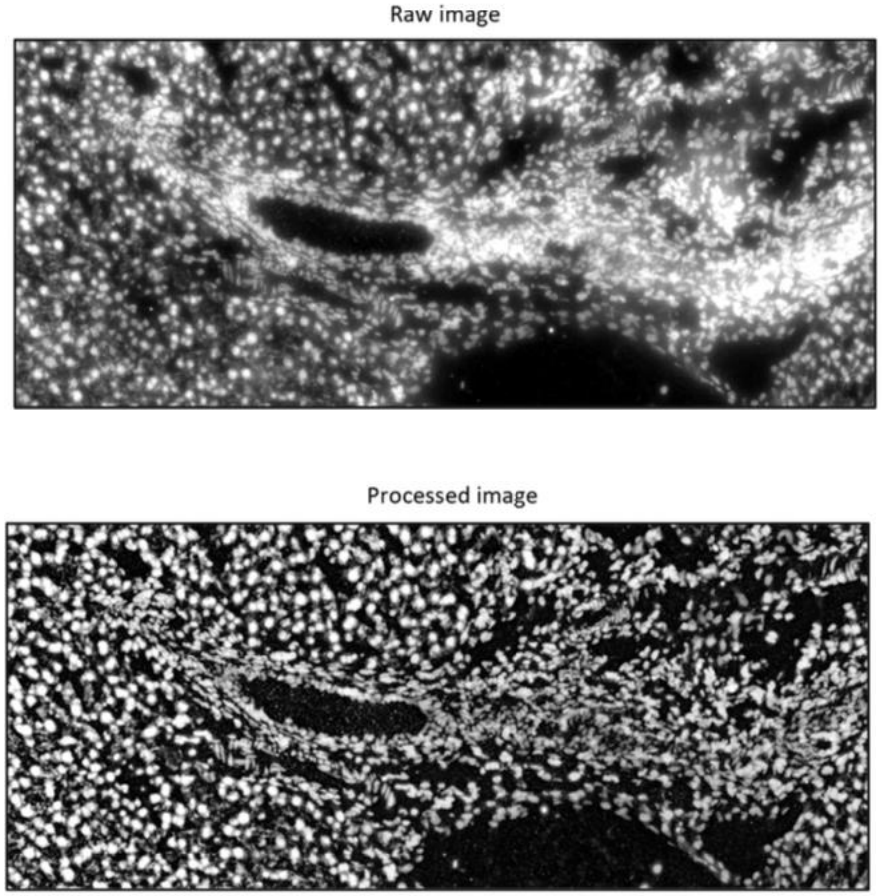
SPArrOW currently supports:

- MERSCOPE (Vizgen),
- Xenium (10x Genomics)
- Molecular Cartography (Resolve Biosciences)
- Stereo-Seq (STOmics)
- CosMx SMI (Nanostring)
- GenePS (SpatialGenomics)
- Pyxa (Stellaromics)

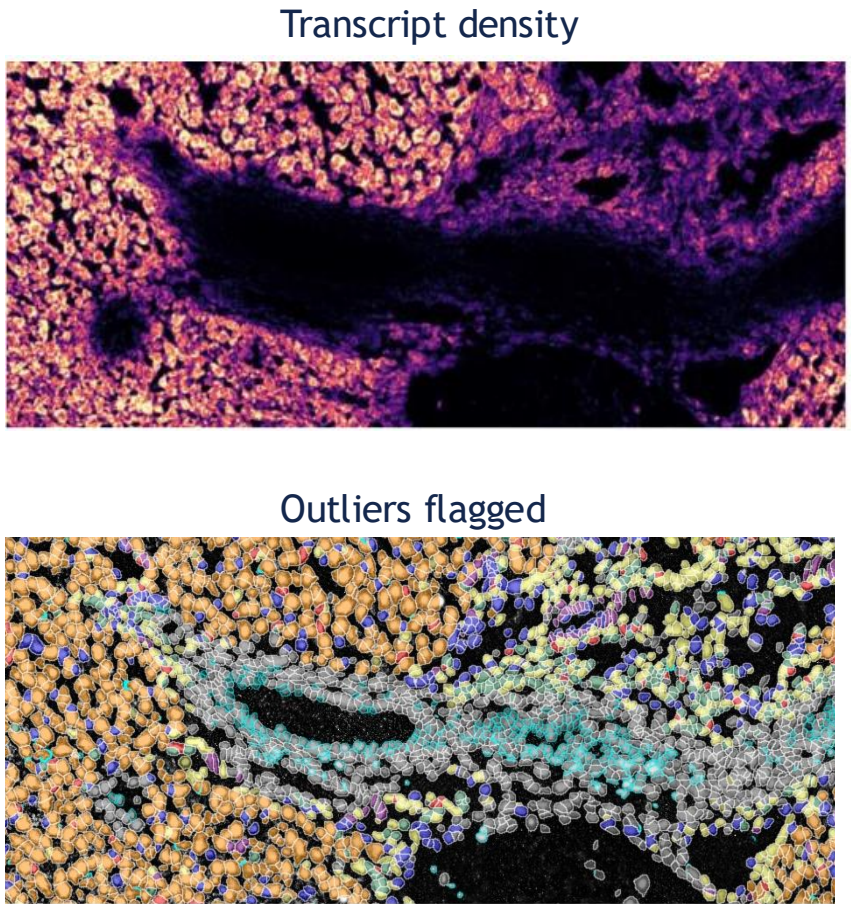
QC and preprocessing … and nothing else matters



Molecular Cartography, Maize tissue



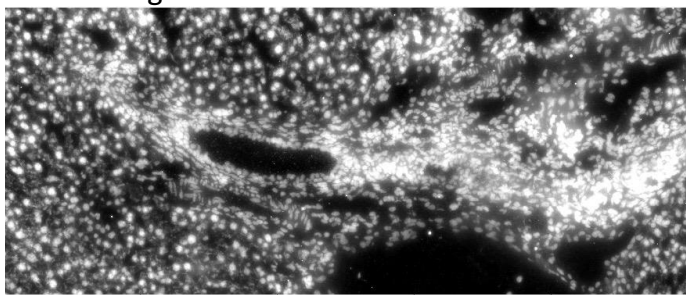
VizGen MERSCOPE, liver tissue



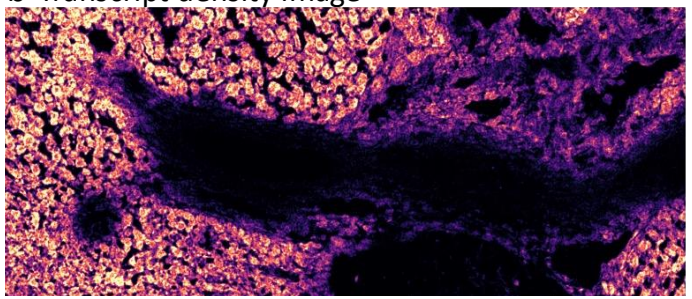
VizGen MERSCOPE, liver tissue

SPArrOW facilitates improved quality control, resulting in more robust downstream analysis

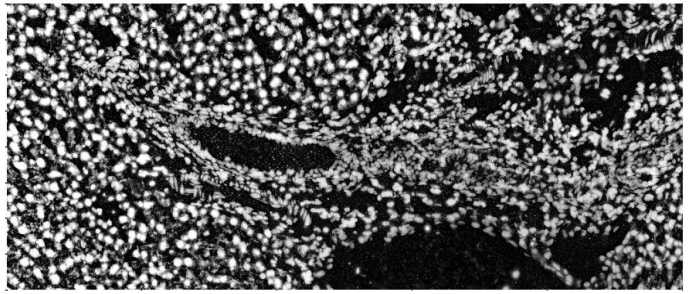
a Raw image



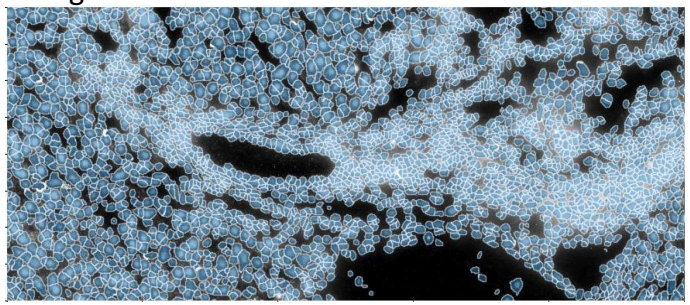
b Transcript density image



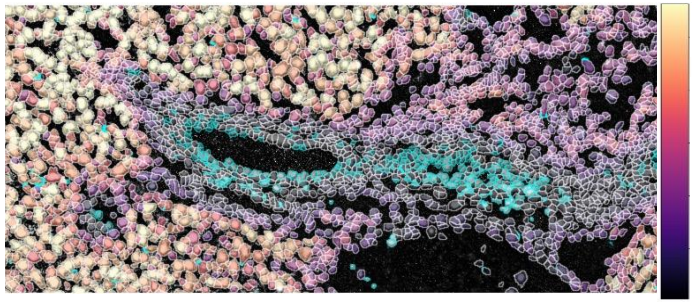
c Processed image



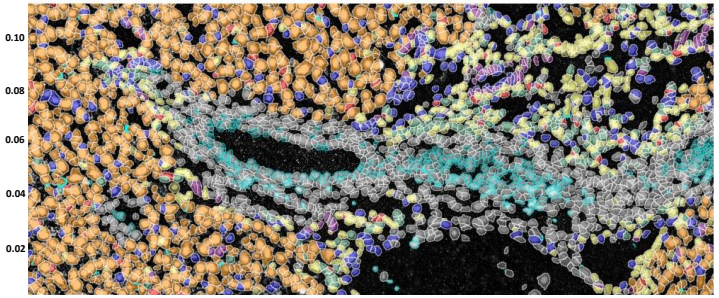
d Segmentation masks



e Cells colored by total counts

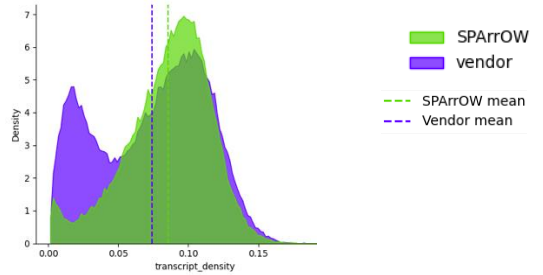


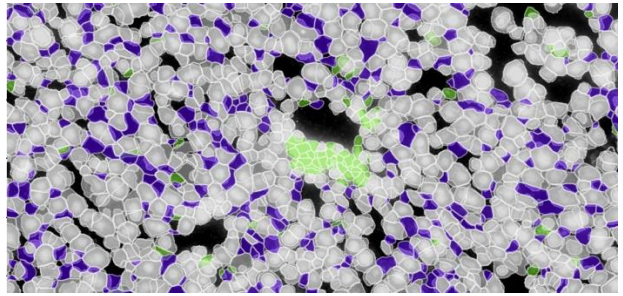
f Cells colored by celltype annotation



B cells
Hepatocytes
Cholangiocytes
Kupffer cells
Endothelialcells
Stromalcells
Other immunecells
Unknown
Filtered cells

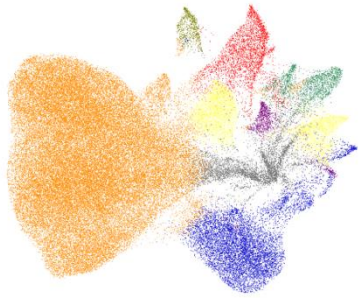
Transcript density





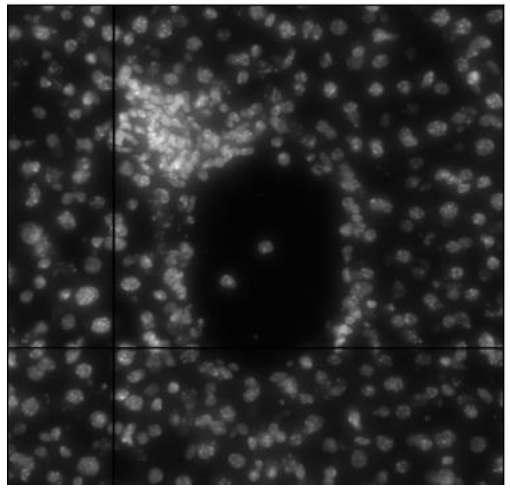
SPArrOW

Vendor

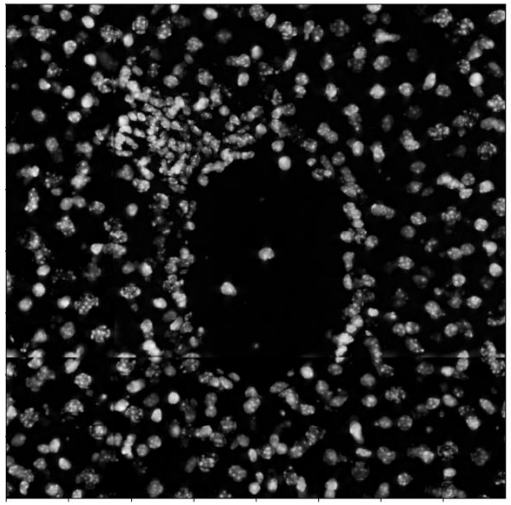


QC and prepocessing enhances cell segmentation

Raw image

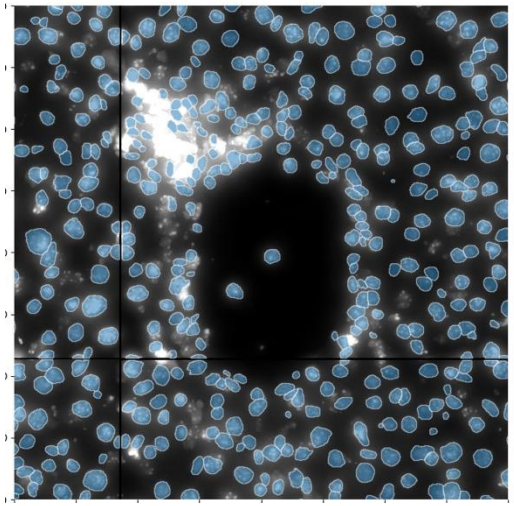


Illumination correction
Inpainting
Tophat filter
Contrast enhancing
Parameter tuning



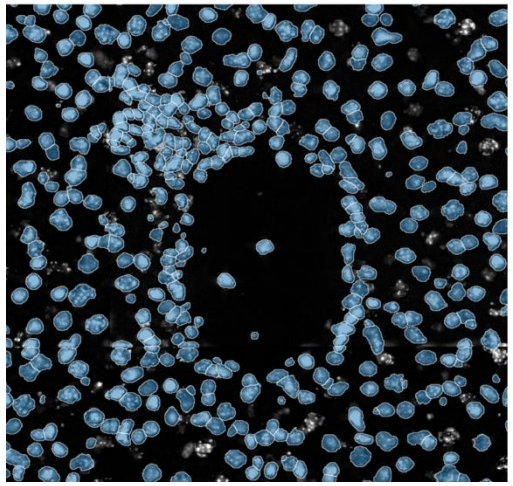
Cleaned image

Cellpose



Low quality segmentation

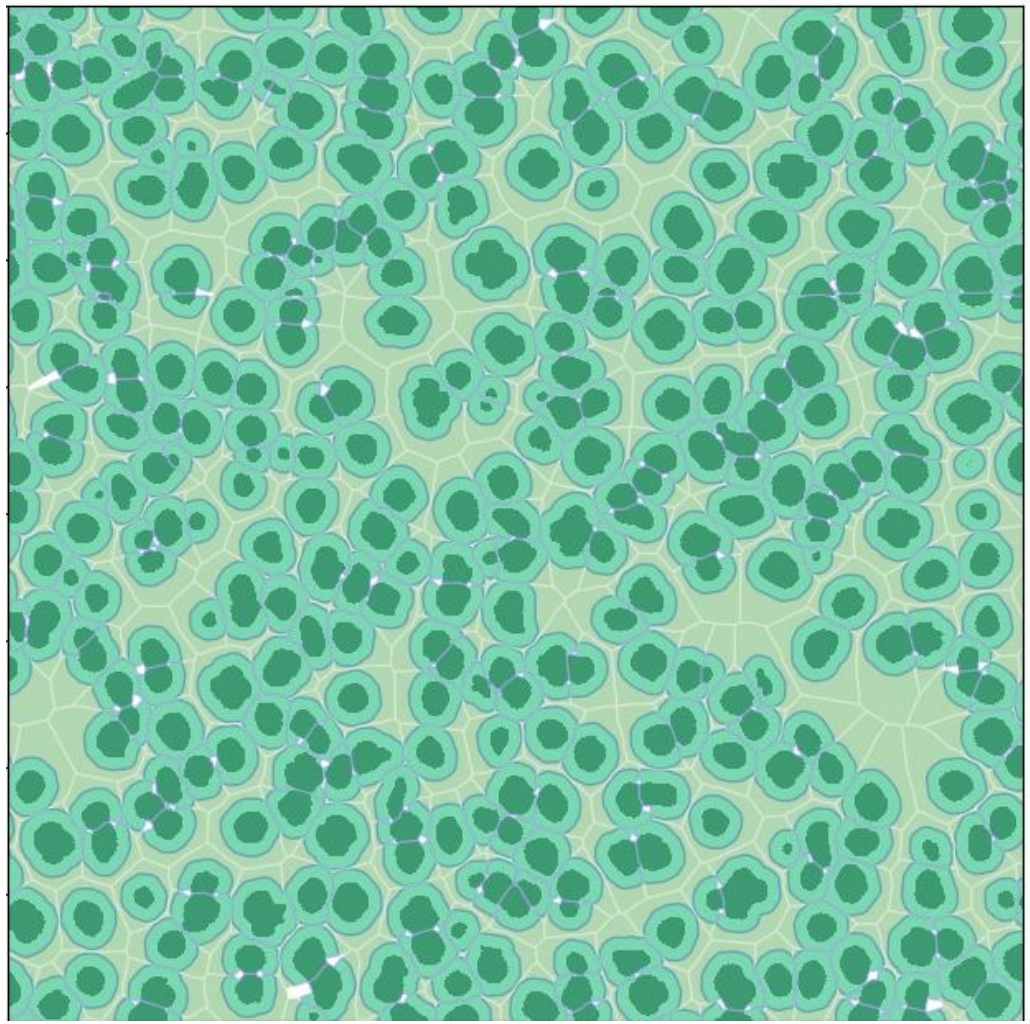
Cellpose



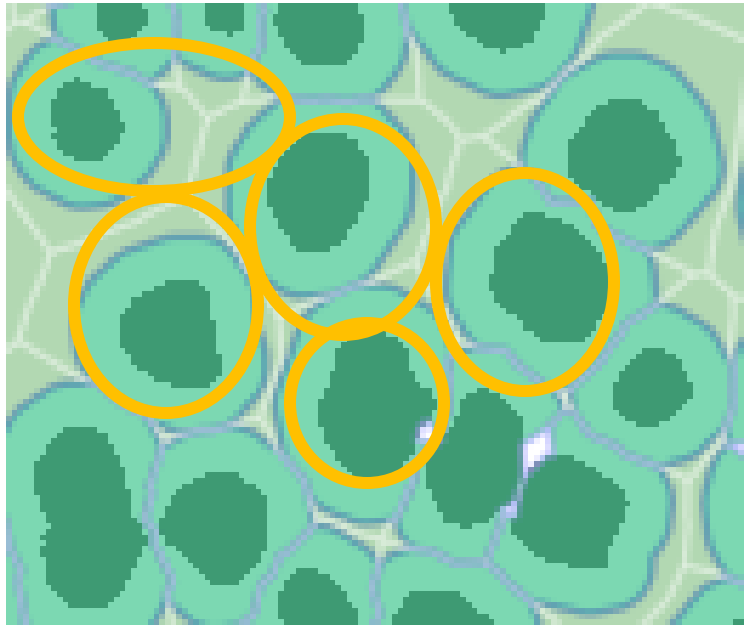
High quality segmentation

Stringer, Carsen, et al. "Cellpose: a generalist algorithm for cellular segmentation." *Nature methods* 18.1 (2021): 100-106.

Nuclear segmentation using DAPI stain results in the cleanest results



Full cell segmentation



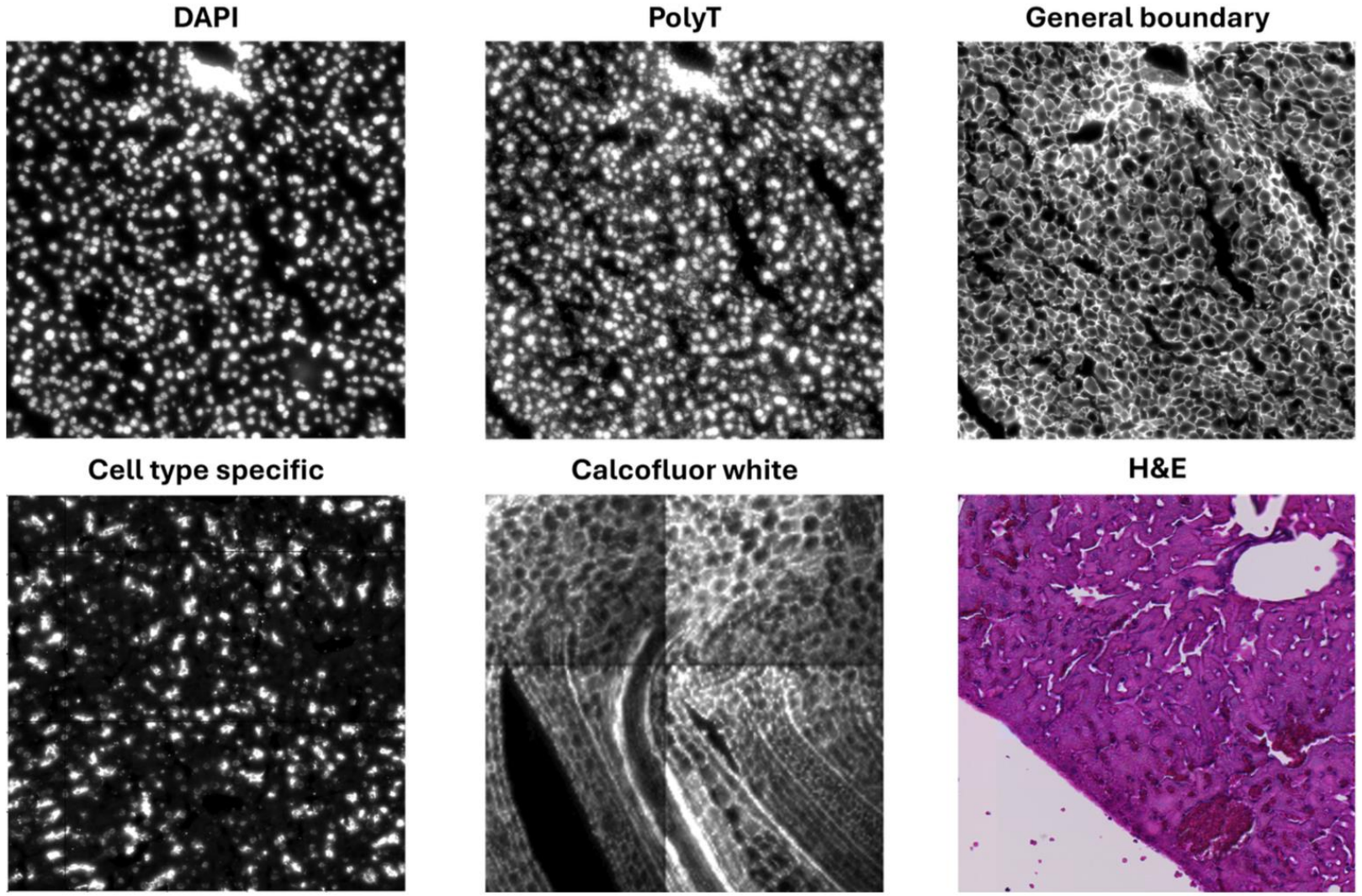
Nucleus

Expanded Nucleus

Voronoi

'ground truth' cellshape

Improving cell segmentation using multiple stains

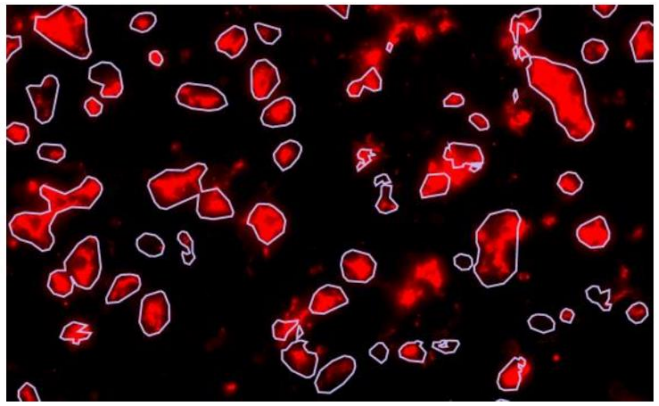


These can improve cell segmentation, but:

- Cell type specific
- Tissue specific
- They don’t always work

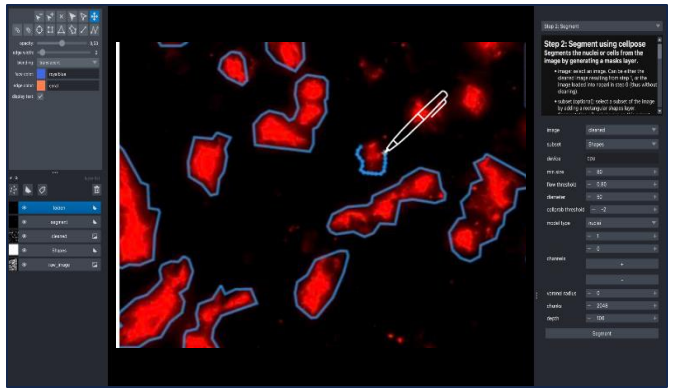
A human-in-the-loop model drastically improves cell segmentation and annotation

Pretrained Cellpose model (cyto2)

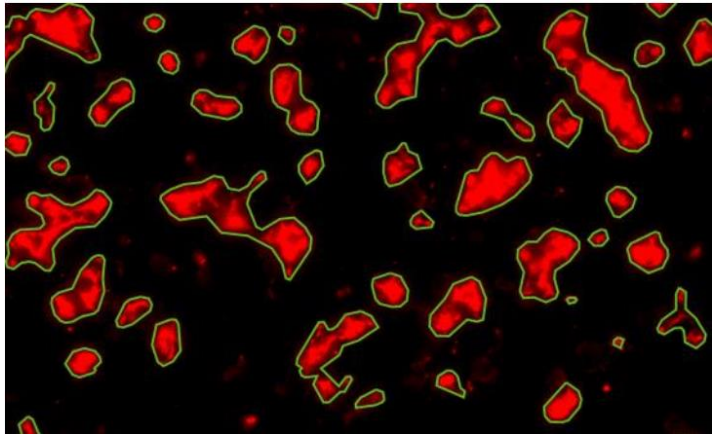


+

Sparrow-Napari interactive interface

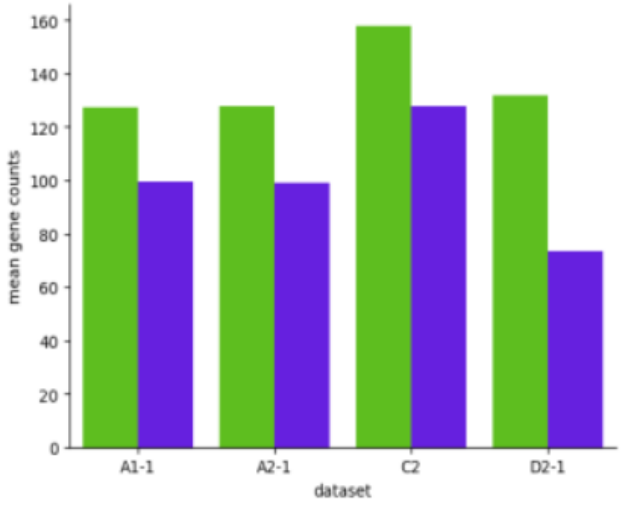


Manual annotation of 116 cells

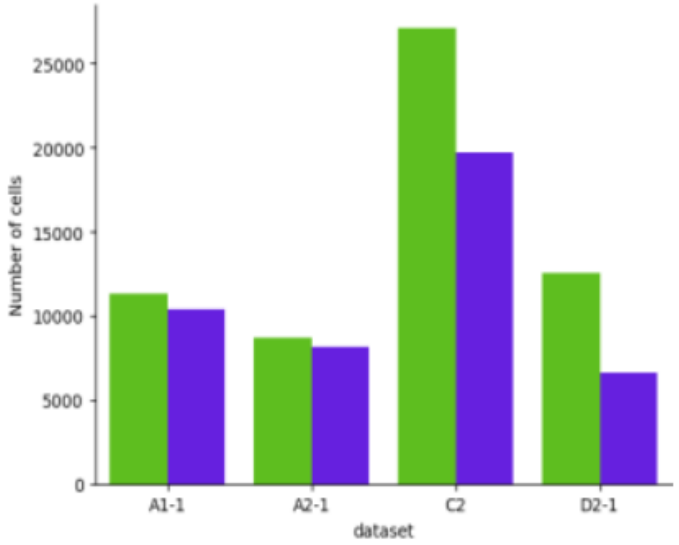


A human-in-the-loop model drastically improves cell segmentation and annotation

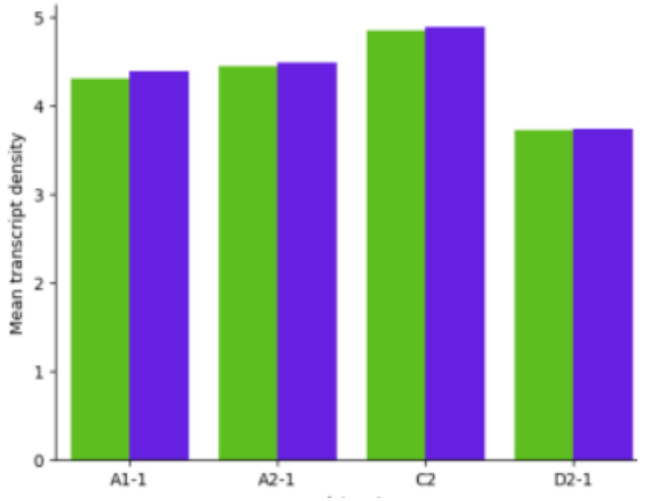
Average transcript counts



Number of cells



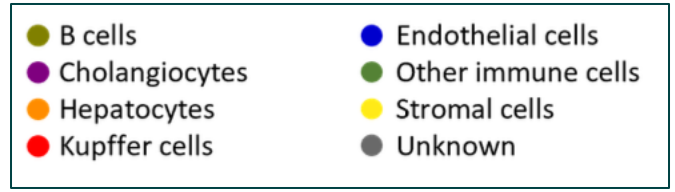
Transcript density



Human-improved
Default

- Human-improved *SPArrOW* identified 33% more cells across all cell types with high transcript densities
- Many of these additional cells are region and cell-type-specific
- Most strikingly, cholangiocyte detection increased by 50%, impacting any downstream data interpretation

SPArrOW improves cell type annotation from spatial transcriptomics data



Cluster purity

Tangram

Score_genes (Scanpy)

Tacco

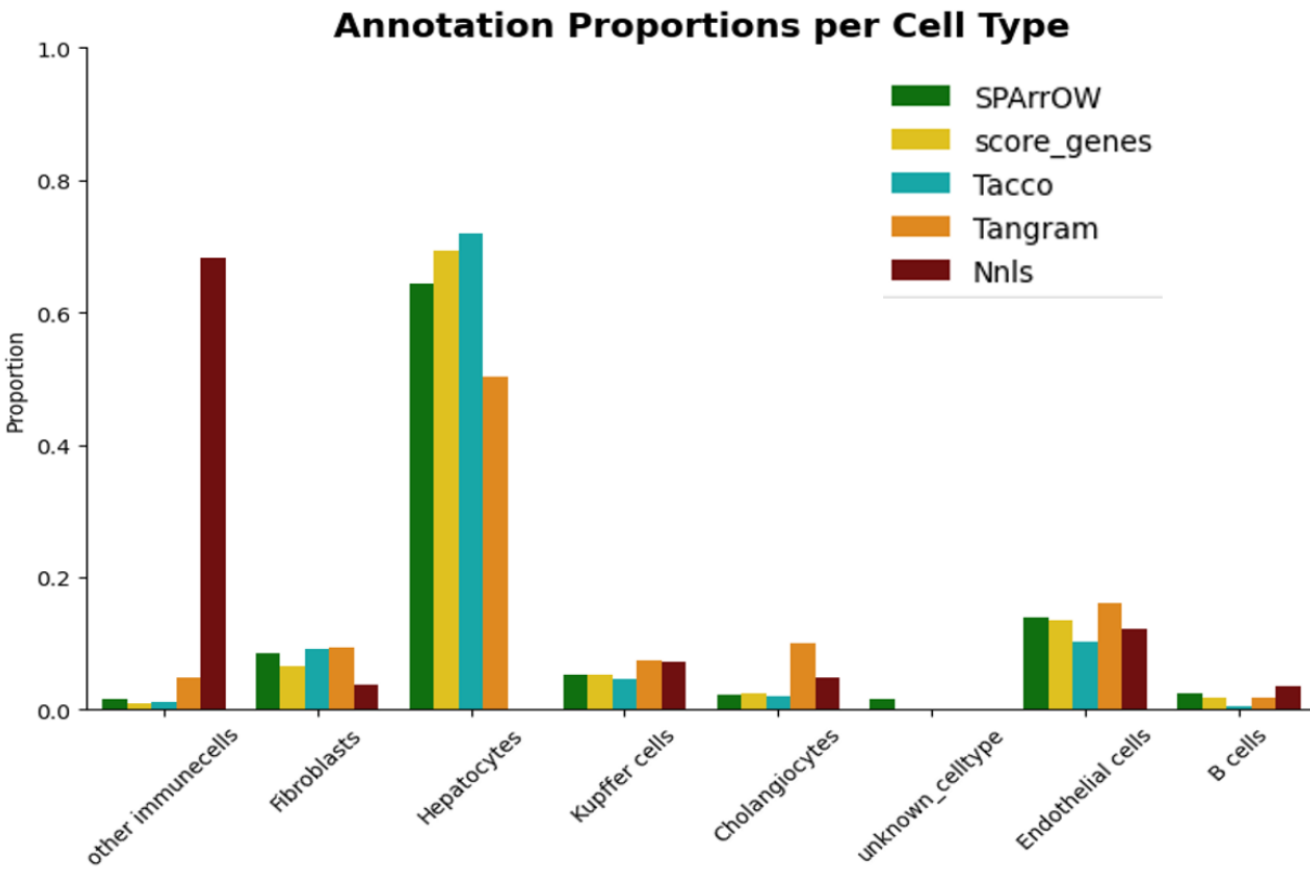
NNLS

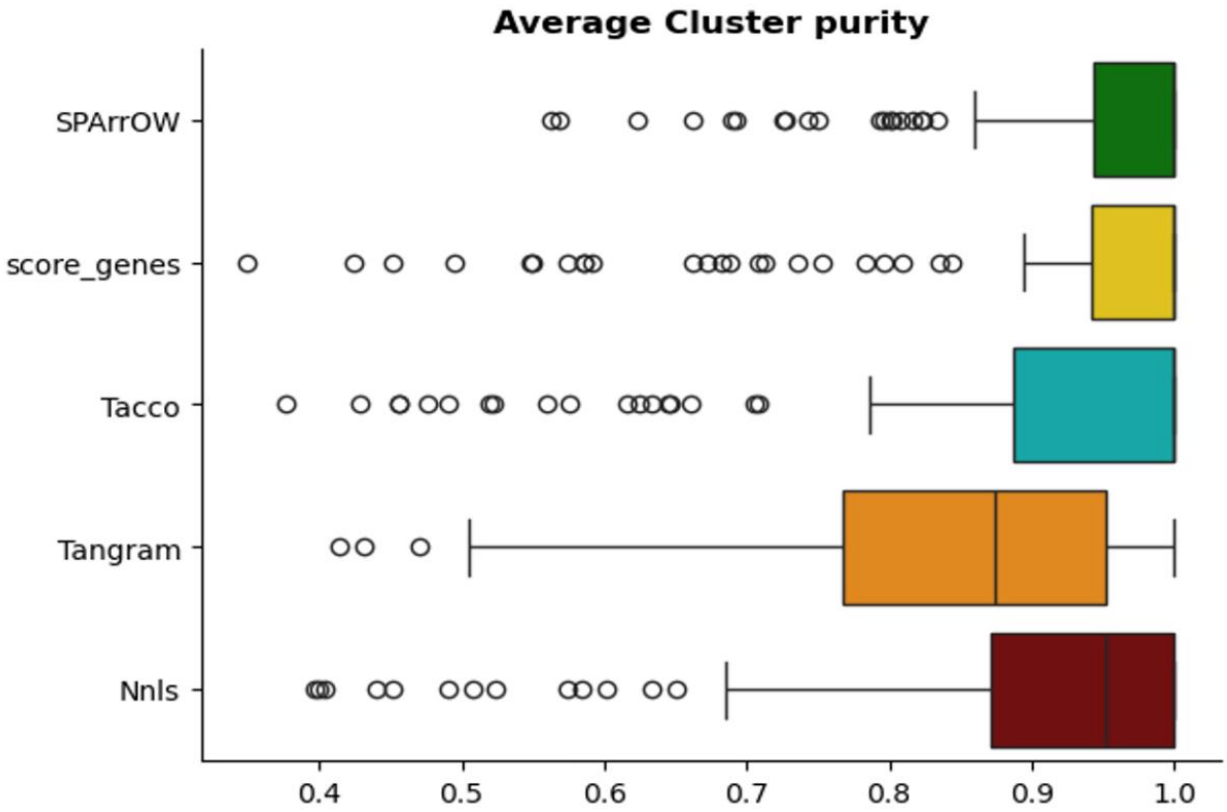
SPArrOW

Metrics to evaluate the annotation algorithms

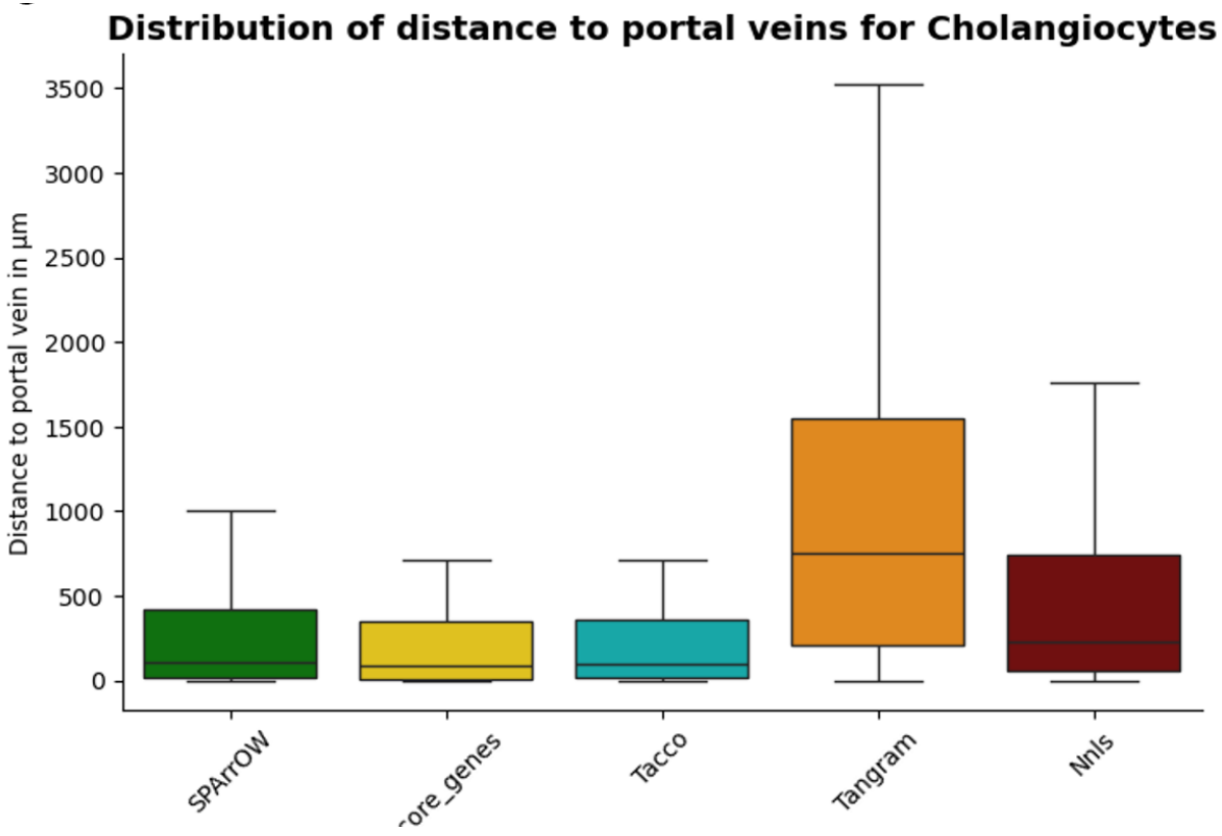
1. **Cluster purity assessment**
   Leiden clustering with a resolution of 10 to overcluster the data, resulting in 105 clusters. For each cluster, the dominant cell type was identified, and its percentage used as purity measure
2. **Biologically informed distance**
   For each cell, the distance to the closest vein was calculated using the distance function in GeoPandas. As cholangiocytes only occur around portal veins, the distance between cholangiocytes and the closest portal vein was used as a quality metric.
3. **Robustness to Expansion**
   As non-informed expansion might mix and match the cytoplasm of different cell types, methods need to be able to handle background gene expression and retain the original annotation.
4. **Robustness to reference atlas**
   The three methods that were dependent on a reference atlas (Tacco, Tangram and NNLS) were run on the same dataset, once with the complete mouse liver dataset as reference, and once with the subset of only the scNucseq dataset as reference

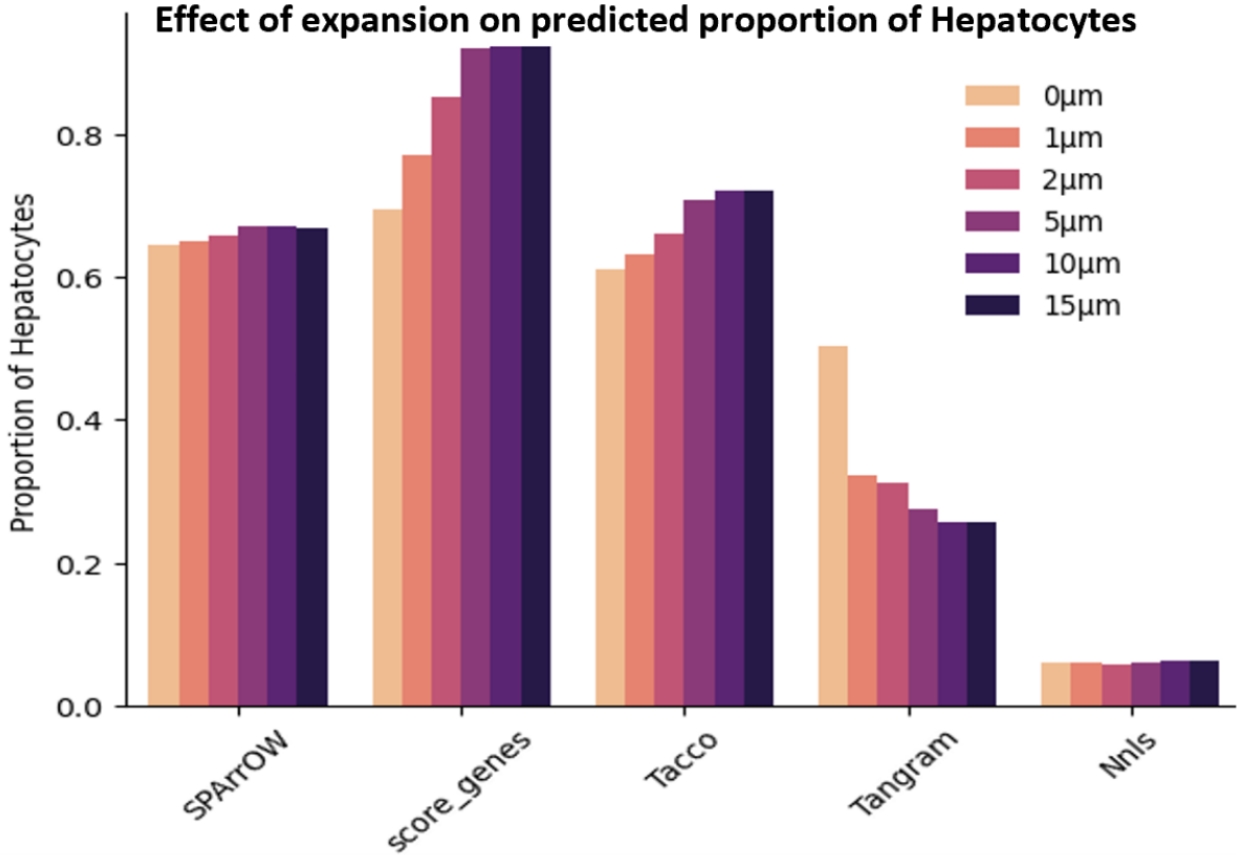
Multi-objective cell type annotation evaluation



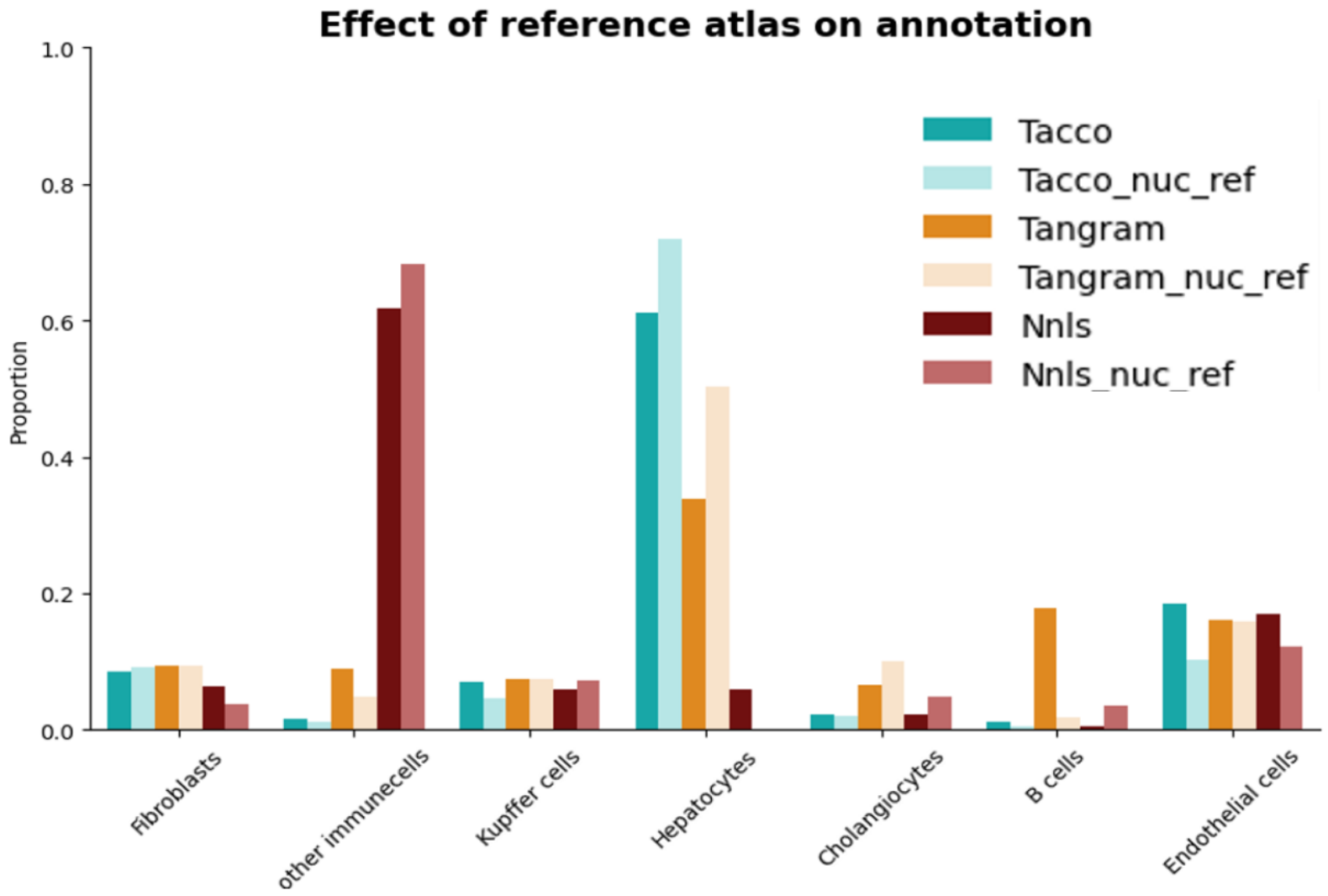


Multi-objective cell type annotation evaluation





Multi-objective cell type annotation evaluation

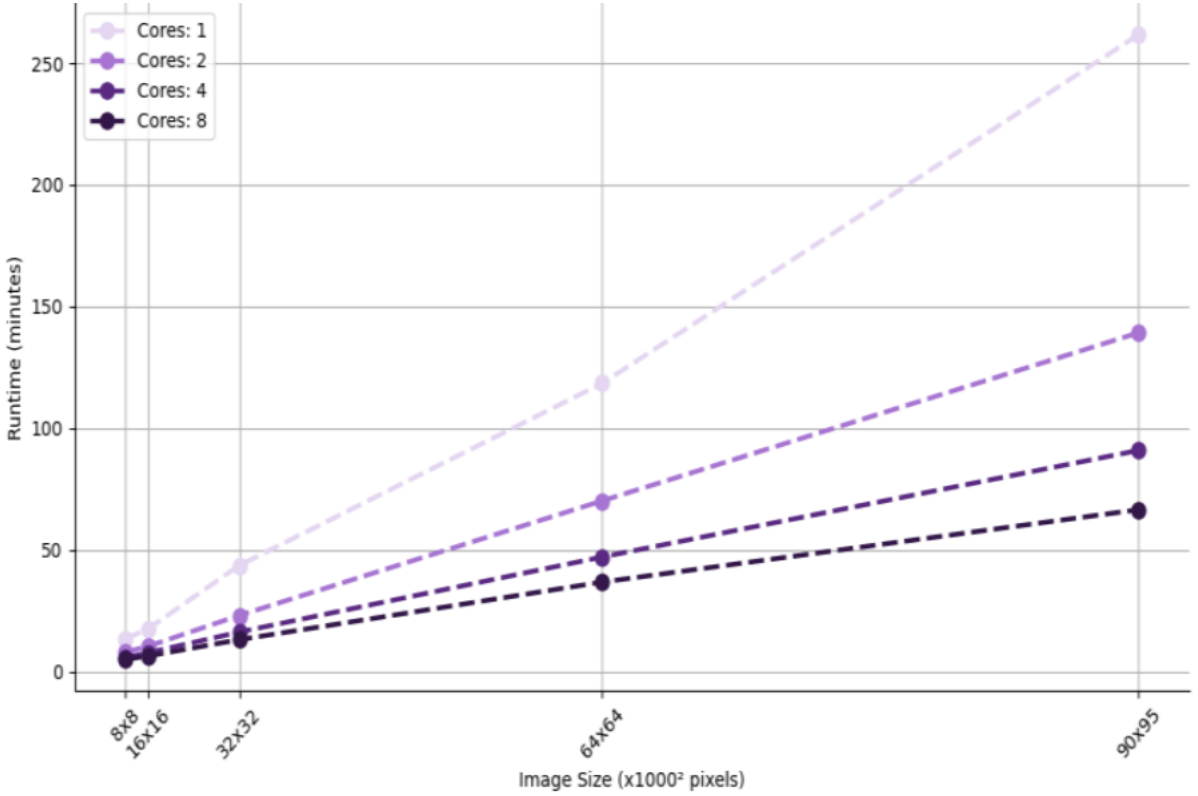


Take-away messages

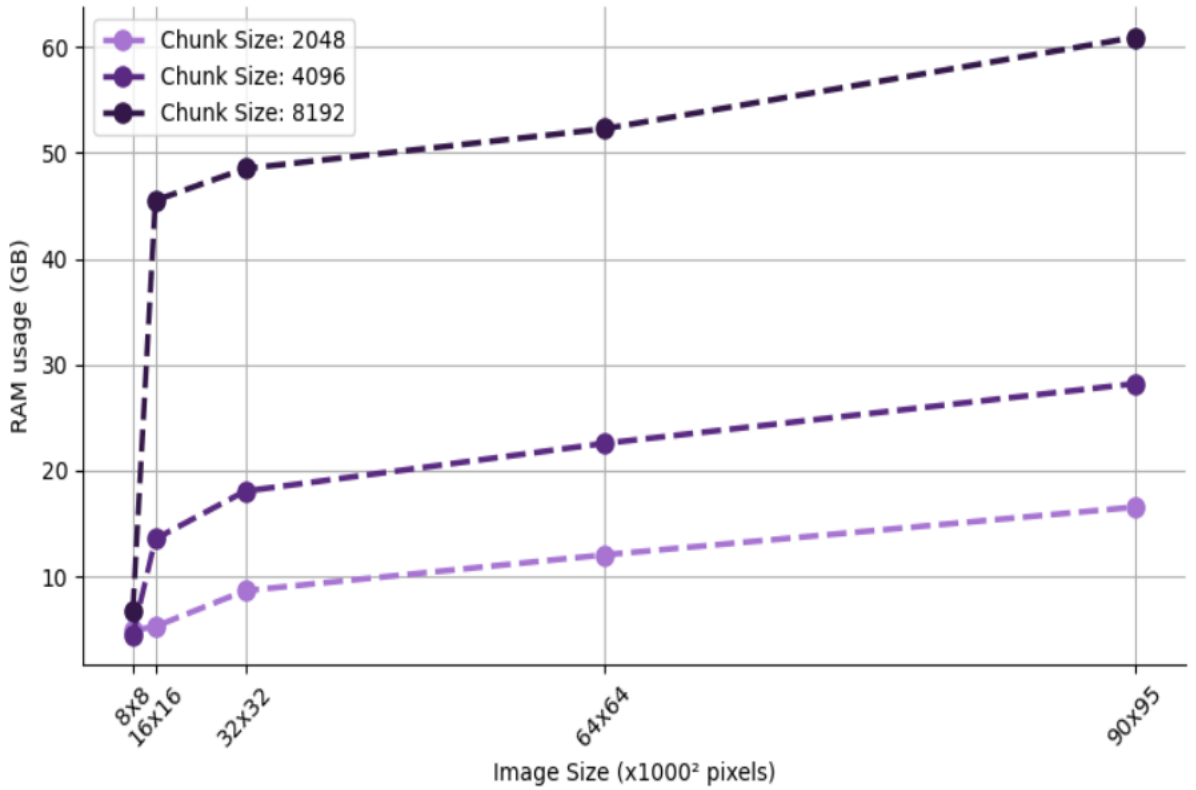
- SPArrOW offers:
  - Improved quality control and preprocessing, including interactive parameter tuning and optimization
  - Flexibility in platforms, segmentation and annotation algorithms
  - Improved segmentation and annotation
  - Increased interactivity promoting a biologist-in-the-loop approach:
    - Parameter tuning
    - Retraining of segmentation models to make them dataset specific

SPArrOW flexibly scales to gigapixel images

Runtime

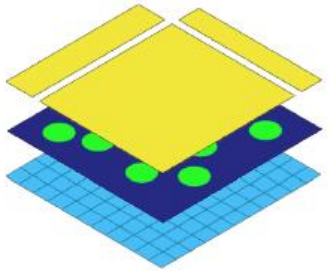


Memory usage

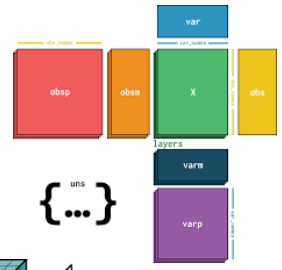


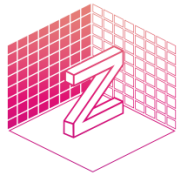
Scalable tooling, interoperability and acceleration

Big Data file formats

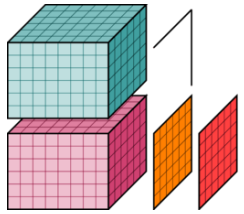


SpatialData





Zarr



xarray

GeoPandas

High-performance computing

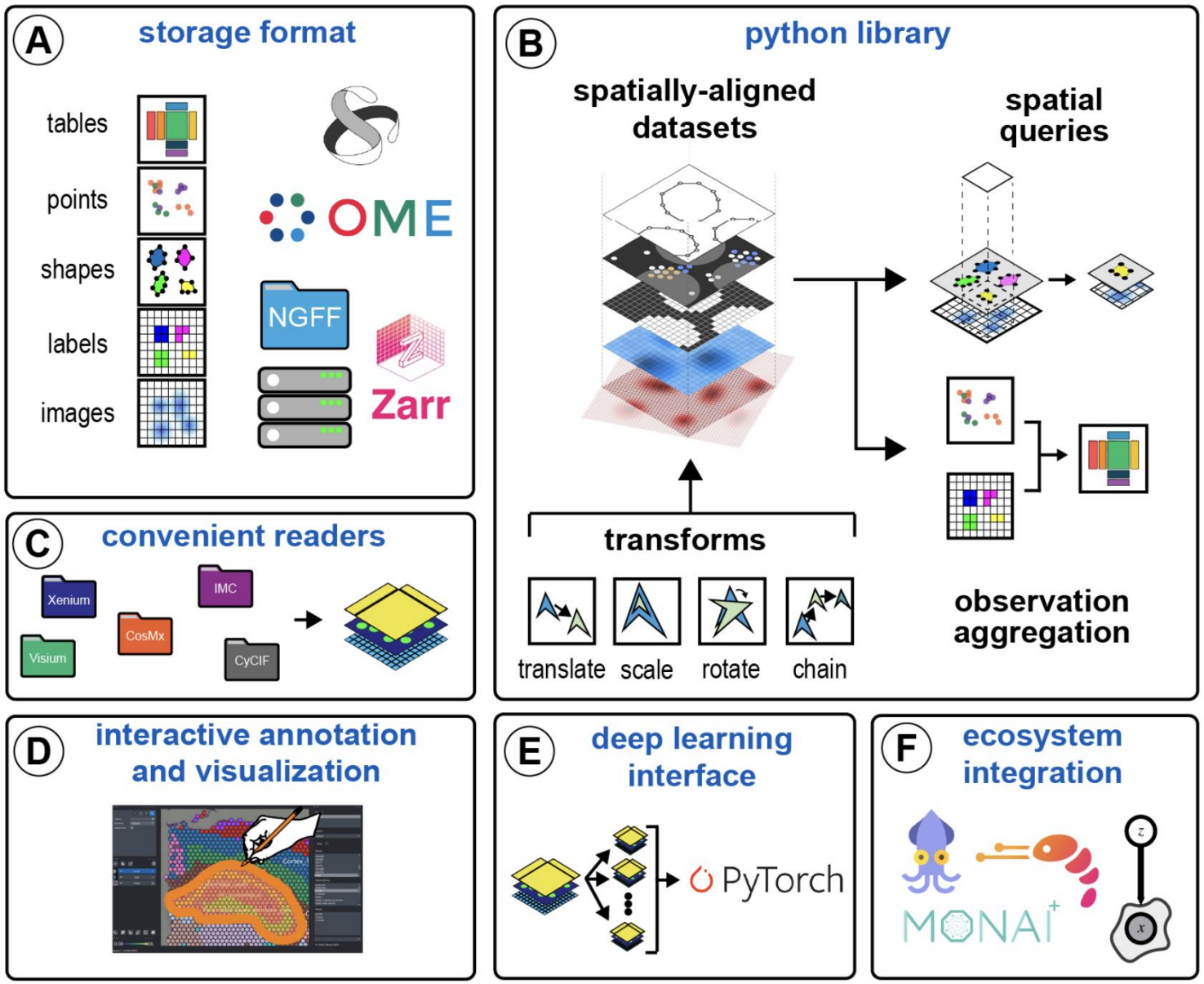
dask

PyTorch

VLAAMS SUPERCOMPUTER CENTRUM

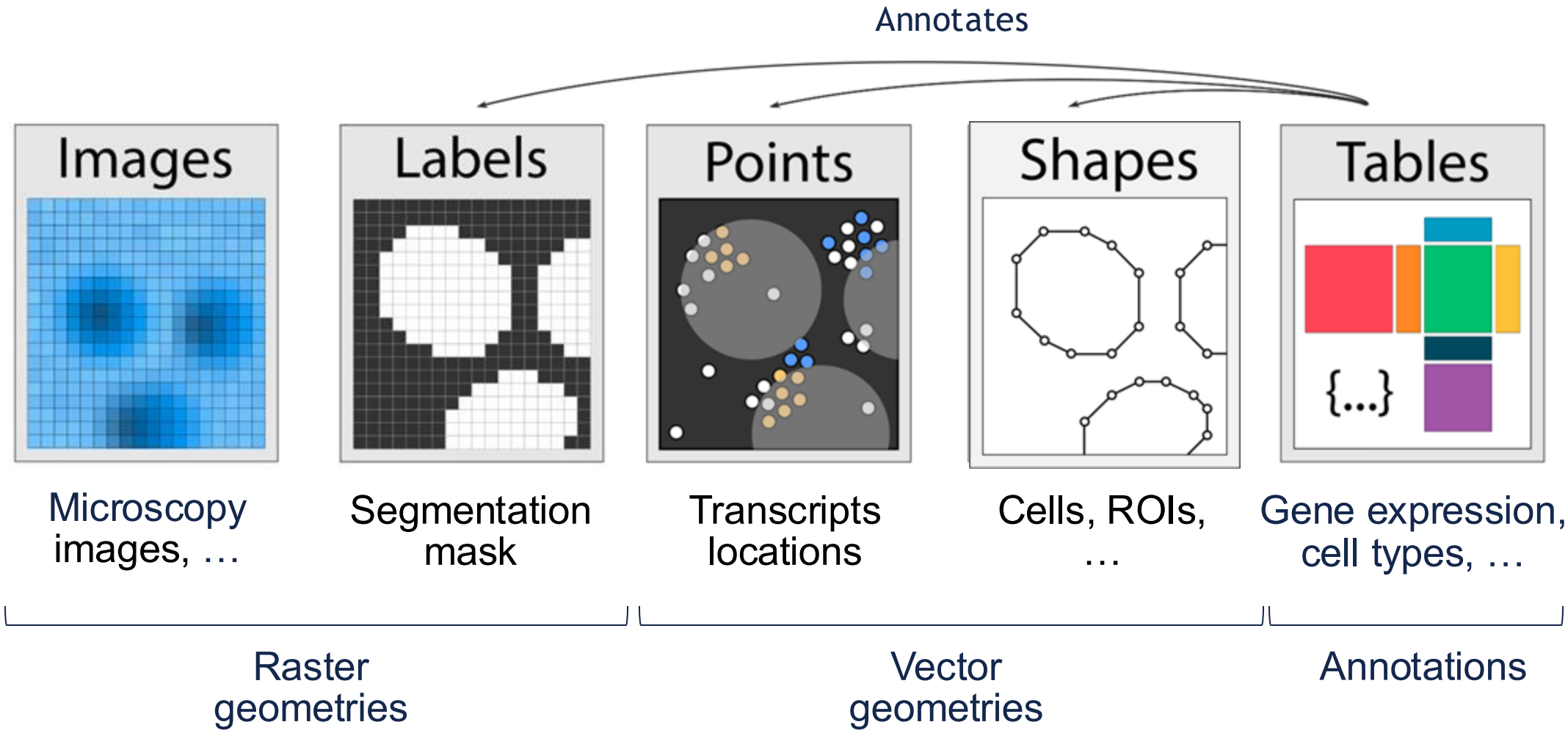
Vlaanderen
is supercomputing

SpatialData Framework

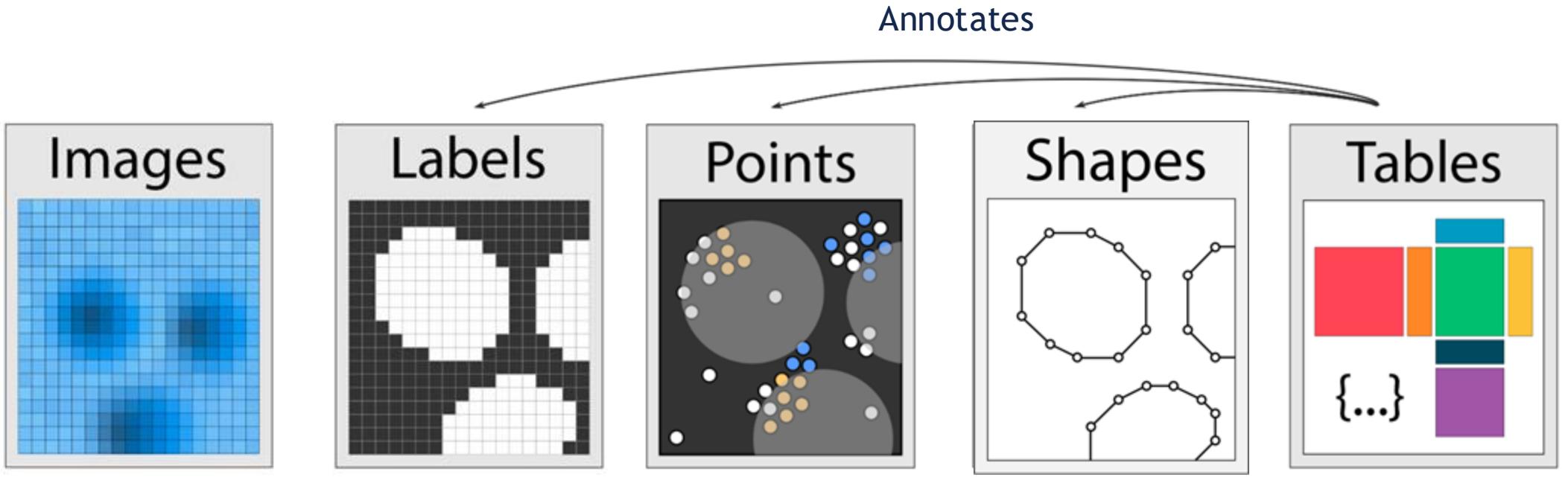


Marconato, L., Palla, G., Yamauchi, K.A. et al. SpatialData: an open and universal data framework for spatial omics. Nat Methods 22, 58–62 (2025)

Data representation is abstracted as a modular combination of reusable elements

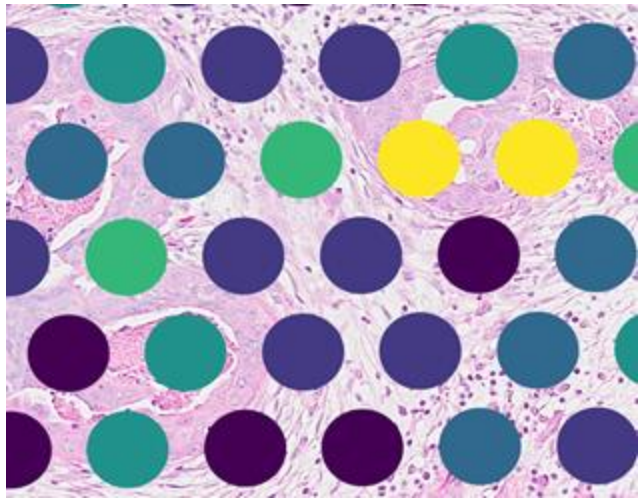


Data representation is abstracted as a modular combination of reusable elements



SpatialData unifies the representation of spatial omics across technologies

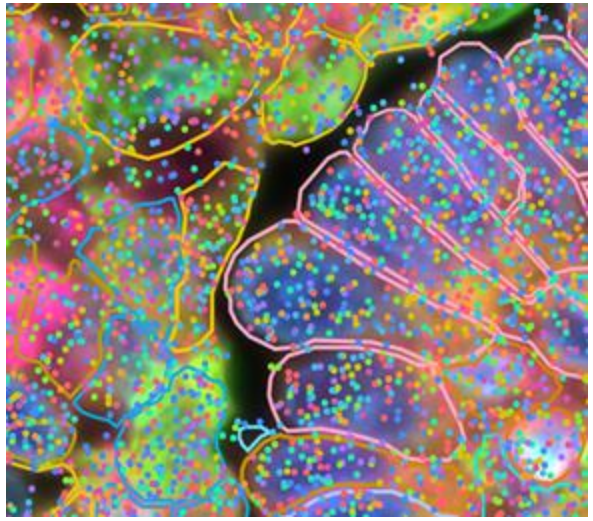
Visium



Data: https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast

Resolution: 55µm
Transcriptome-wide

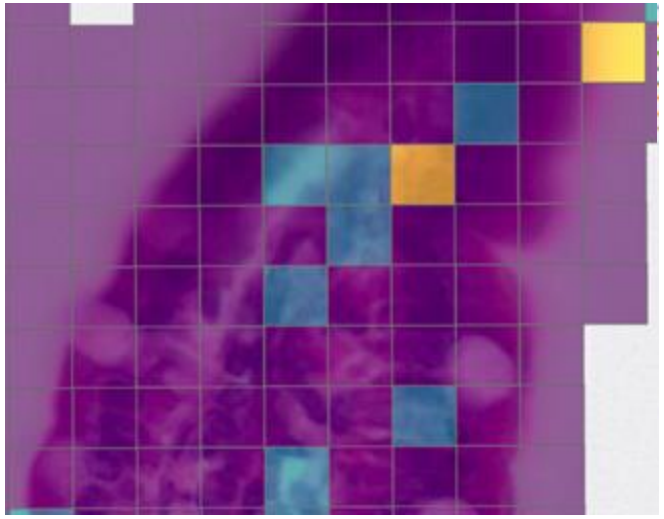
Xenium



Preview Data: FFPE Human Lung Cancer with Xenium Multimodal Cell Segmentation

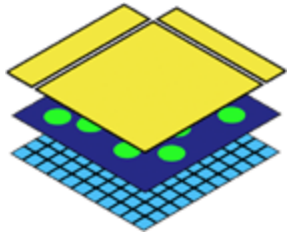
Resolution: single-molecule
Up to 5K genes

Visium HD



Visium HD Spatial Gene Expression Library, Mouse Small Intestine (FFPE)

Resolution: 2µm, 8µm, 16µm, …
Transcriptome-wide



SpatialData

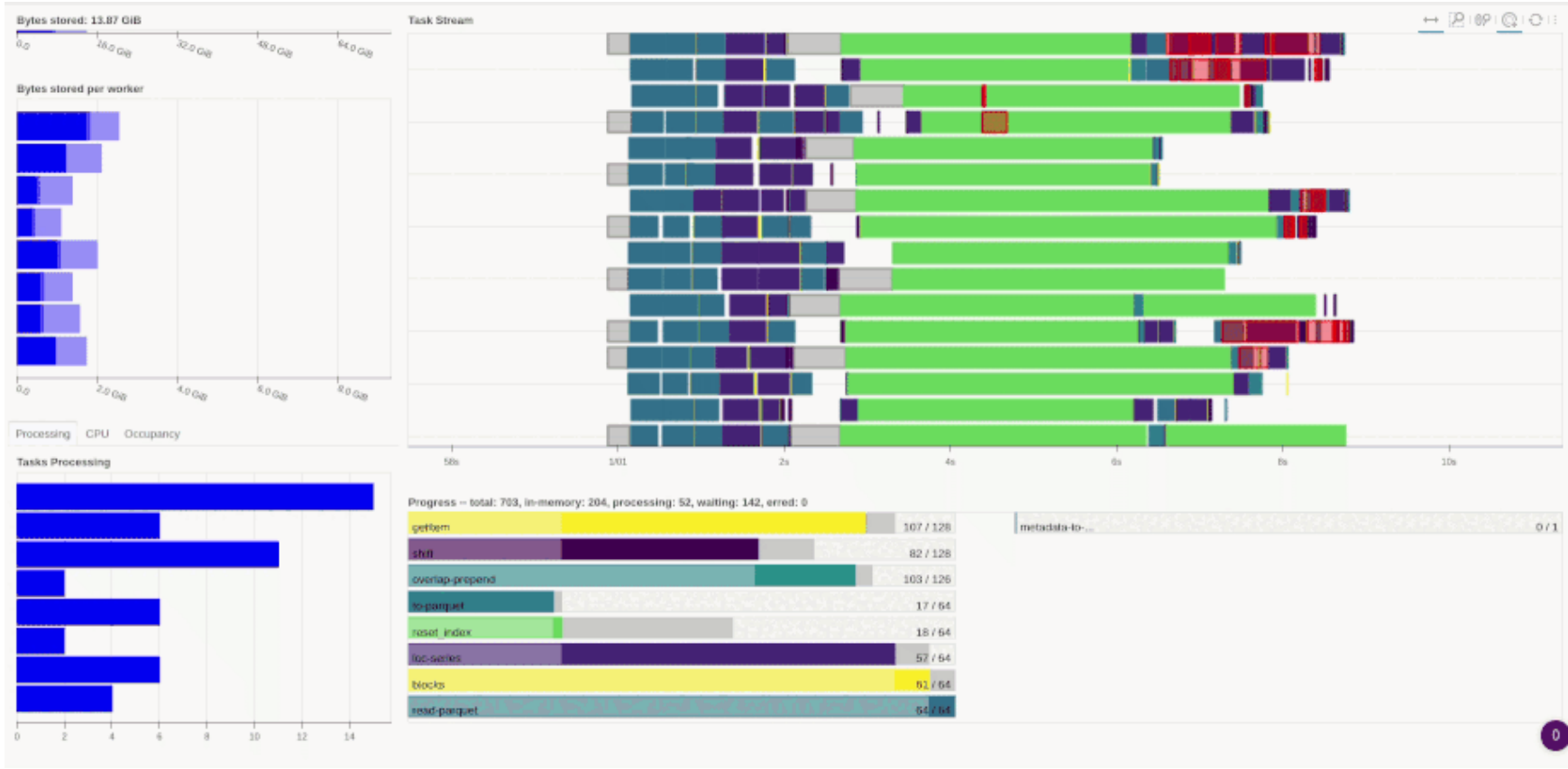
- Simple read/write
- Flexible representation
- Object manipulation

- Interoperability across analysis methods

scanpy squidpy

Scalable and resilient parallel and distributed computation using Dask

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

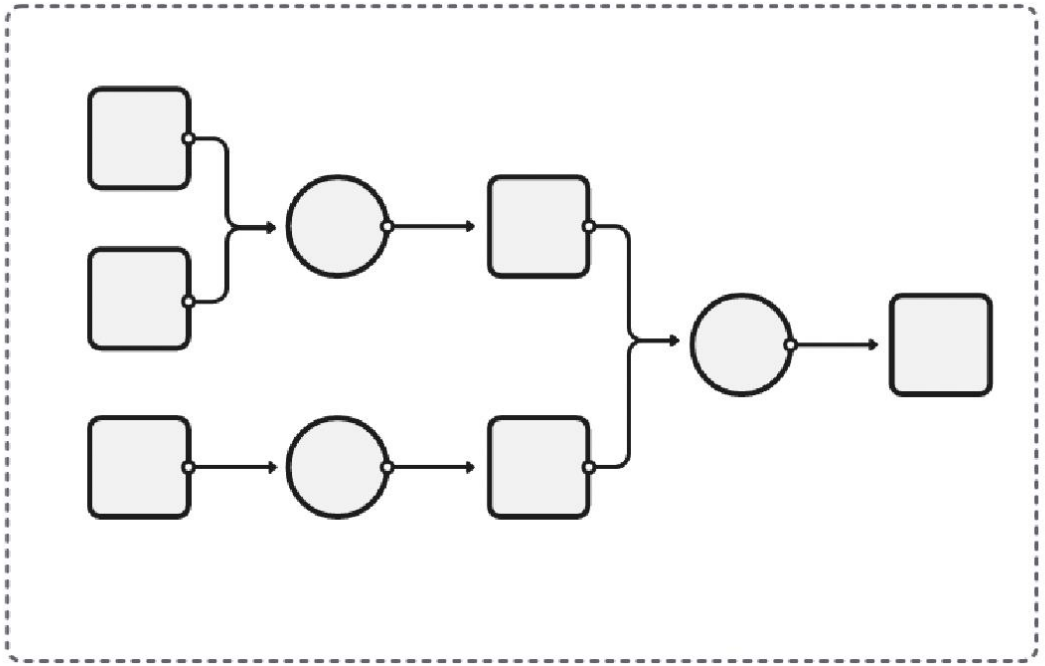
Scalable and resilient parallel and distributed computation using Dask

Collections

(create task graphs)

- Dask Array
- Dask DataFrame
- Dask Bag
- Dask Delayed
- Futures

Task Graph



Schedulers

(execute task graphs)

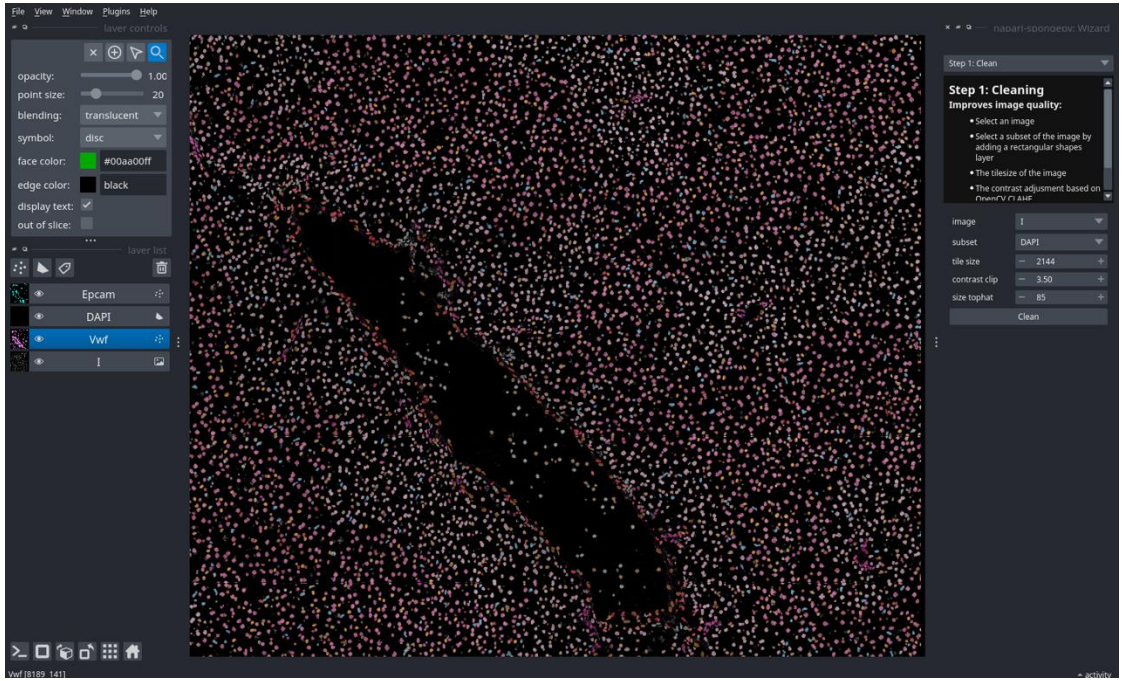
- Single-machine (threads, processes, synchronous)
- Distributed

Characteristics	SPArrOW	SOPA	Steinbock	Molkart	SMINT	STARFISH	PIPEFISH	Squidpy	MEGA-FISH	stereopy	FISH-quant
Input data	Image + mRNA coord	Image+ mRNA coord	Image+ mRNA coord	Image+ mRNA coord	Image+coordinates	Fluorescent images	Fluorescent images	Cell*gene matrix	Fluorescent images	GEM/GEF+ opt image	Fluorescent images
Image processing	Yes, tunable	No	Yes	Yes	No	Yes, tunable	Yes	No	Yes	No	Yes
Segmentation	Yes, tunable	Yes	Yes	Yes	Yes	Yes, watershed	Yes	Yes, external	Basic	Yes	Yes
Allocation	Yes, tunable	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Celltype Annotation	Yes, tunable	Tangram	For IMC	No	SingleR	No	No	Yes, external	No	yes	No
Interactive visualization	Yes, Napari	10x Xenium visualizer	Yes, Napari	No	No	Yes, Old napari	No	No	Napari (limited)	Yes, in notebook	imJoy
Platform versatility	Yes	Yes	Yes, created for IMC	Molecular Cartography	Yes	Not really	Yes	Yes	Yes	stereoSeq	limited
GPU accelaration	Yes	Yes	No	No	Yes	No	Yes	External	Yes	yes	yes
Data Format	SpatialData	SpatialData	Spatial Experiment	Anndata	Rds	spaceTx	spaceTx	Anndata	Zarr (no spatialdata)	GEM/GEF+ adata	Tif/csv/npz
Programming language	Python	Python	Python/R	Python/ nextflow	R/Python	Python	Python	Python	Python	Python	Python
Image-level QC	Yes	No	For IMC	No	No	Yes	Yes	No	No	no	Yes
Cell-level QC	Yes	Yes	For Imc	Yes	limited	No	limited	Yes	No	yes	No
Gene-level QC	Yes	No	For IMC	limited	Yes	No	Yes (at spot level)	No	No	no	No
Dimensionality reduction + clustering	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	yes	No
Tunability	Yes	Limited	No	No	no	Yes	No	Yes	limited	No	yes
Usability	CLI, API, GUI	CLI, API	API,CLI	Nextflow CLI	Not really	API	CLI (CWL?)	API	API	API	API
Normalization	tunable	Lib size	For IMC	No	logNorm	intensities	No	Lib size	No	Lib size	no
Interoperability	scVerse	scVerse	Bioconductor	no	Bioconductor	None	None	scVerse	Napari	no	no
Last commit	2025	2025	2024	2024	2024	2025	2024	2025	2025	2025	2022

SPArrOW allows flexible usage

- 3 ways to interact with SPArrOW
  1. No code in napari
  2. Jupyter notebooks: constant feedback plots
  3. Commandline tool using Hydra

Example: Interactive parameter tuning in napari



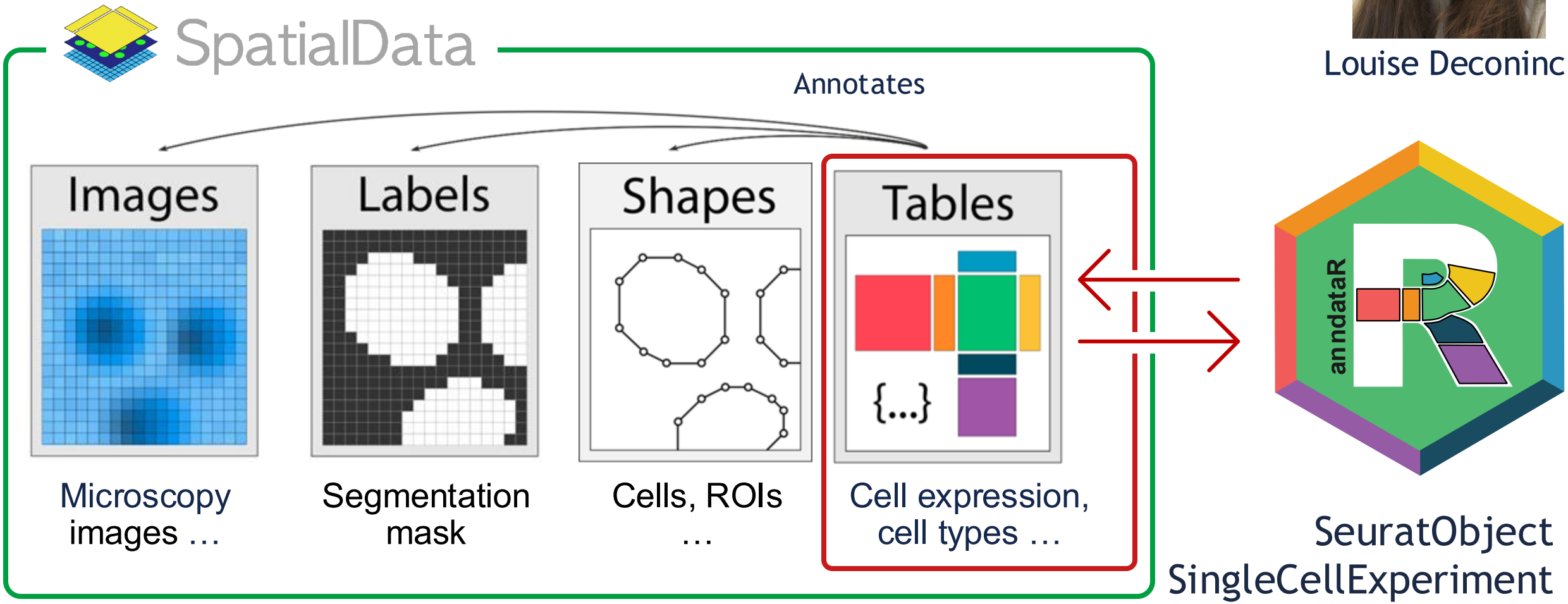
Visual, direct feedback, but very slow on large datasets: subset selection possible

Less flexible

Note: you can run napari from notebooks too!

SPArrOW output is interoperable with R

Louise Deconinck



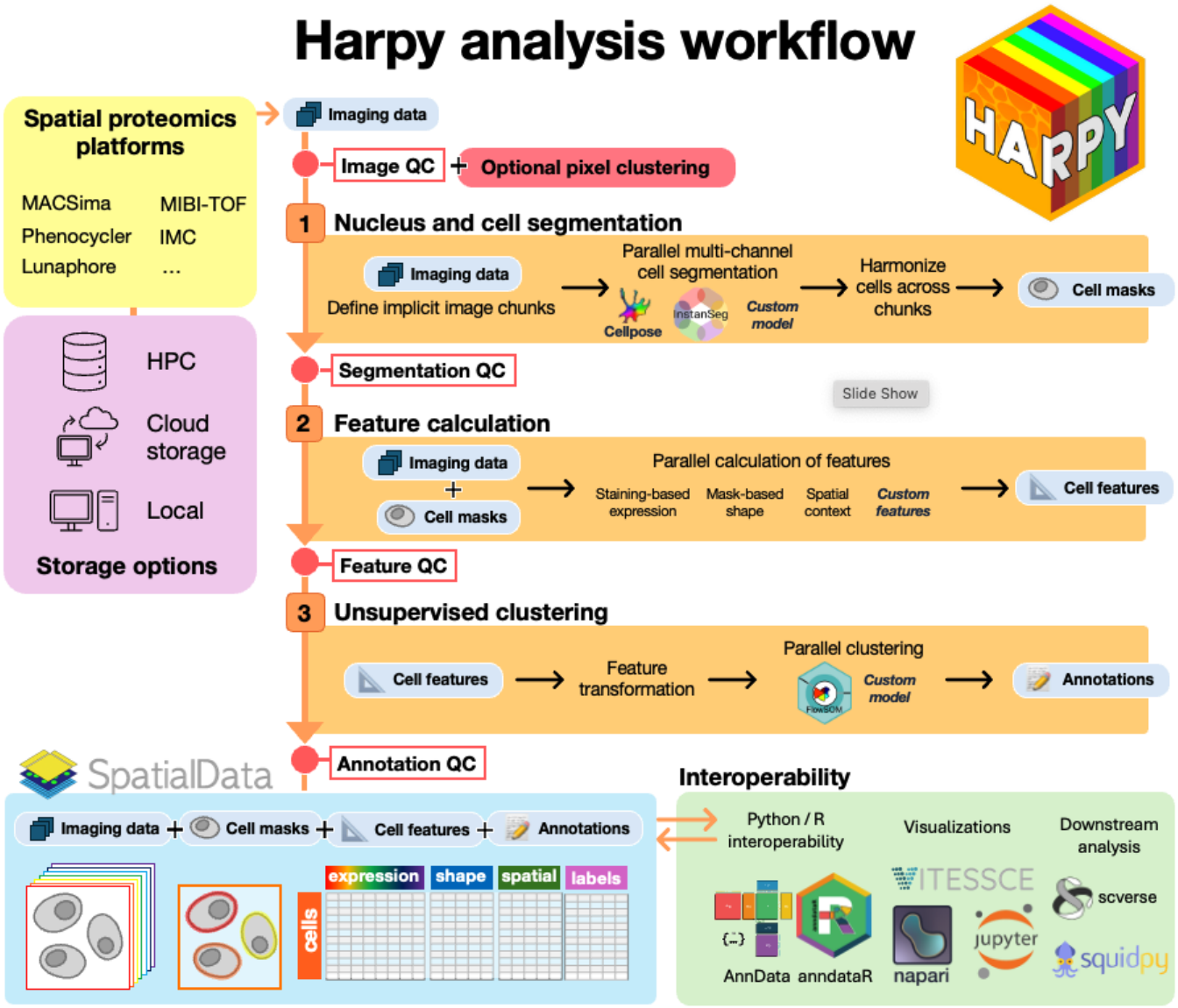
Cannoodt R, Zappia L, Morgan M, Deconinck L (2025). *anndataR: AnnData interoperability in R*.
R package version 0.99.0, https://github.com/scverse/anndataR, https://anndatar.data-intuitive.com/.

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

Harpy: Spatial proteomics analysis that makes you happy

Benjamin Rombaut

https://github.com/saeyslab/harpy



Acknowledgements

Lotte Pollaris, Benjamin Rombaut, Robin Browaeys, Chananchida Sang-Aram, Louise Deconinck

Charlotte Scott
Martin Guilliams
Jean-Christophe Marine
Oliver Stegle
Fabian Theis
Julio Saez-Rodriguez

Spatial Catalyst
Arne De Fauw
Julien Mortier
Frank Vernaillen
Evelien Van Hamme