Deep learning and computational methods for spatial data analysis

Quentin Blampey

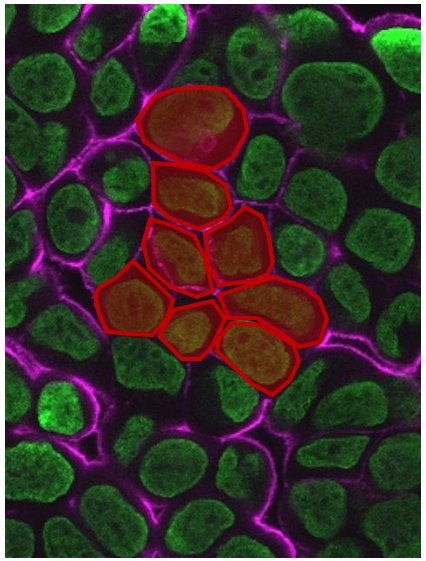
AI & ML in Spatial Single-Cell Transcriptomics Workshop
16 Oct 2025 - Lyon (France)

CentraleSupélec, Paris-Saclay University
Gustave Roussy Institute

Typical tasks in spatial omics analysis

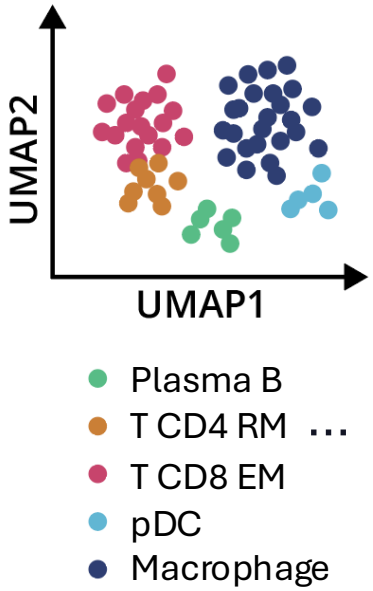
A

Cell segmentation



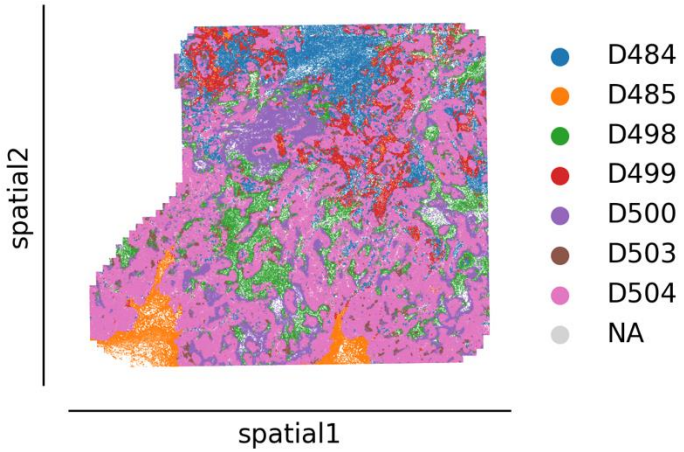
B

Cells embedding and annotation



C

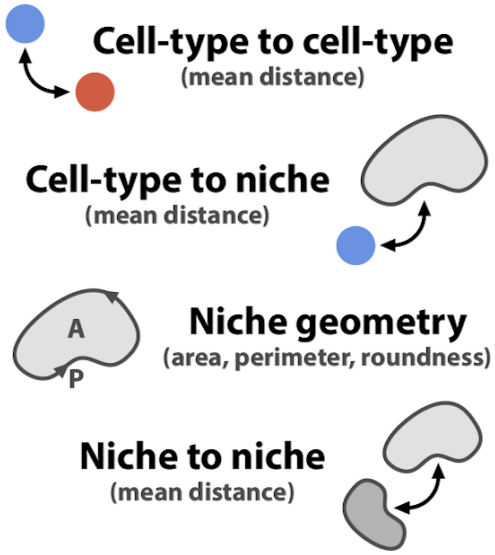
Spatial clustering (also known as spatial domains)



Typical tasks in spatial omics analysis

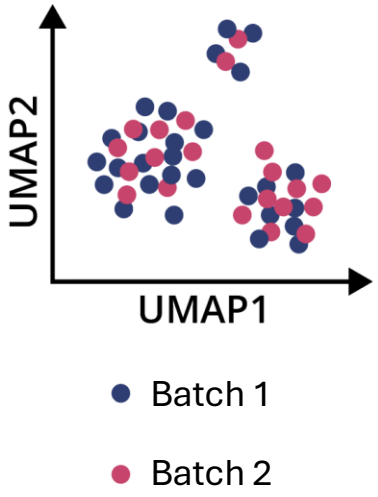
D

Spatial geometrical statistics and distances



E

Domains batch-effect correction



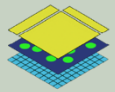
Motivation

Building analysis tools for biological research, within the scverse community

scverse

Core data structures

AnnData



SpatialData

SpatialData is a data structure, not an analysis library

Ecosystem packages

sopa

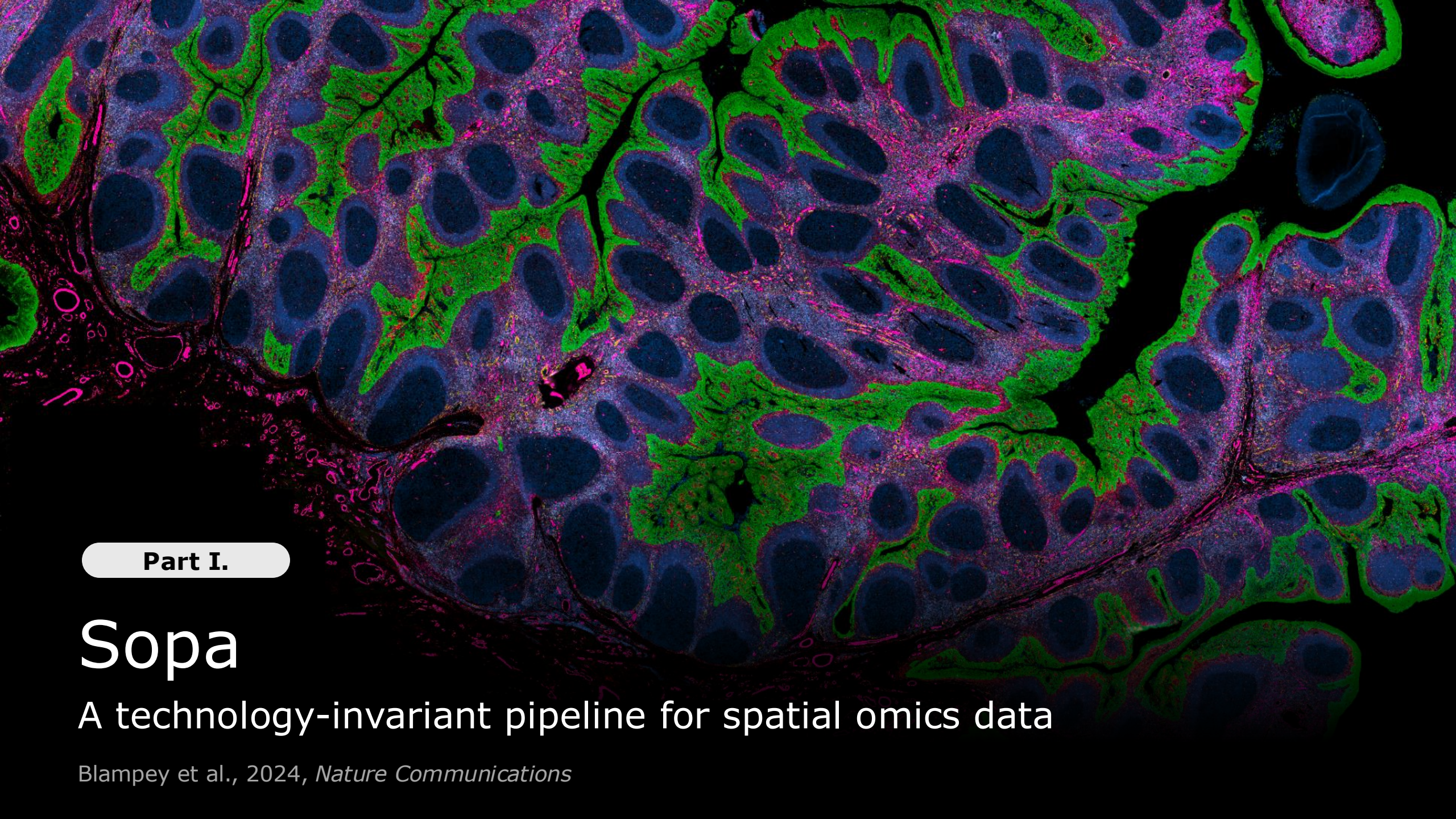
General spatial toolkit

Cell segmentation, tissue segmentation, interactive visualization, efficient aggregation, spatial operations, and many other features

novae

Foundation model

Spatial domains (or niches), hierarchy of domains, domains batch-effect correction, slide architecture analysis, pathway spatial patterns



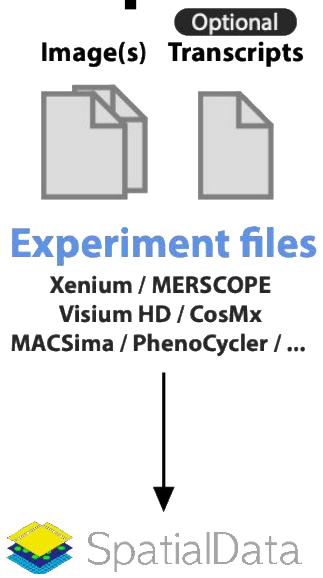
Part I.

Sopa

A technology-invariant pipeline for spatial omics data

Blampey et al., 2024, *Nature Communications*

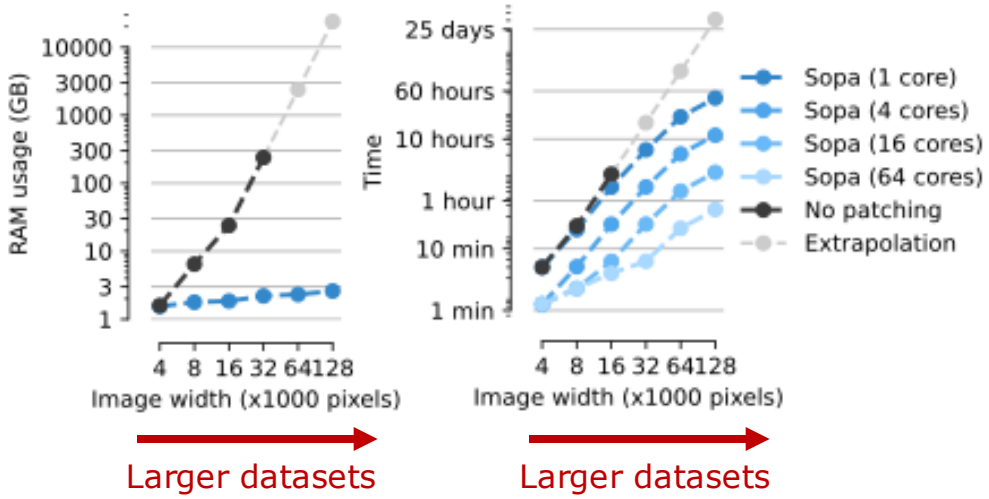
Overview of Sopa



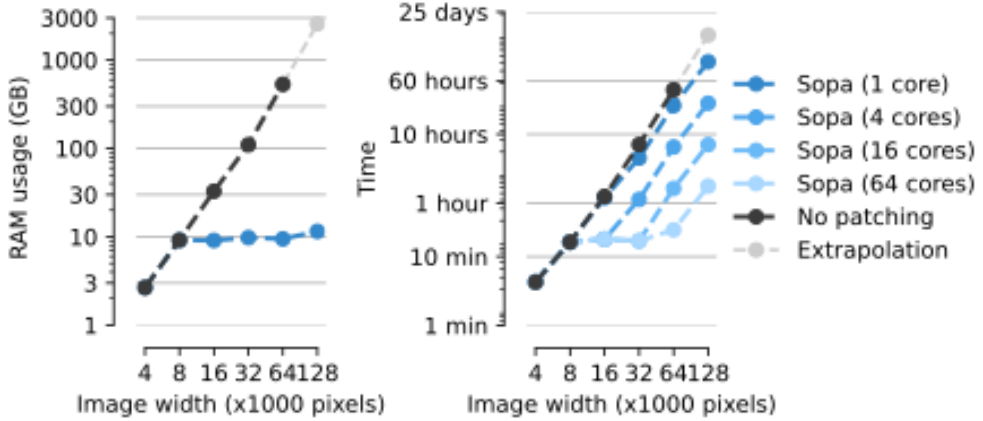
Blampey et al. (2024) | Sopa: a technology-invariant pipeline for analyses of image-based spatial omics (*Nature Communications*)
CellPose (Stringer et al., Nat. Meth.), Proseg (Jones et al., Nat. Meth.), Tangram (Biancalani et al., Nat. Meth.), Squidpy (Palla et al., Nat. Meth.)

RAM and time efficiency, compared to naive approaches

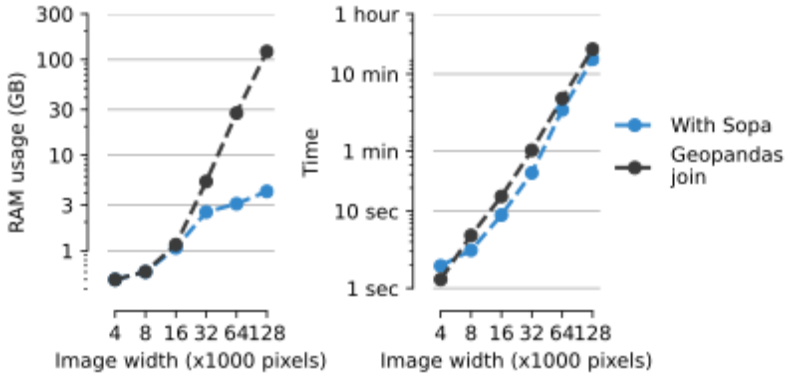
Cellpose segmentation



Baysor segmentation



Transcript aggregation



Channel aggregation

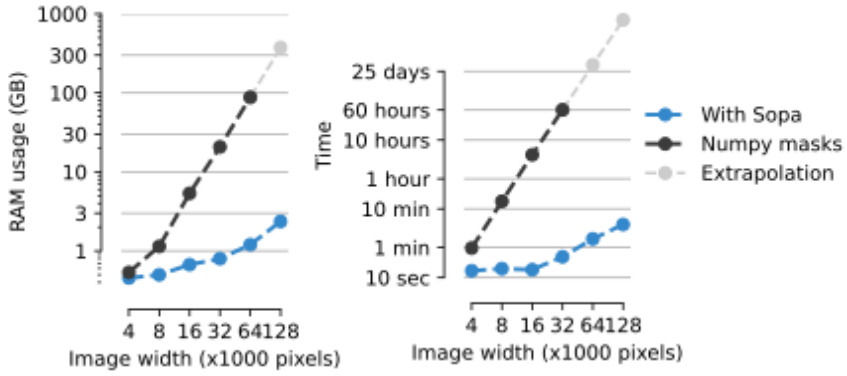
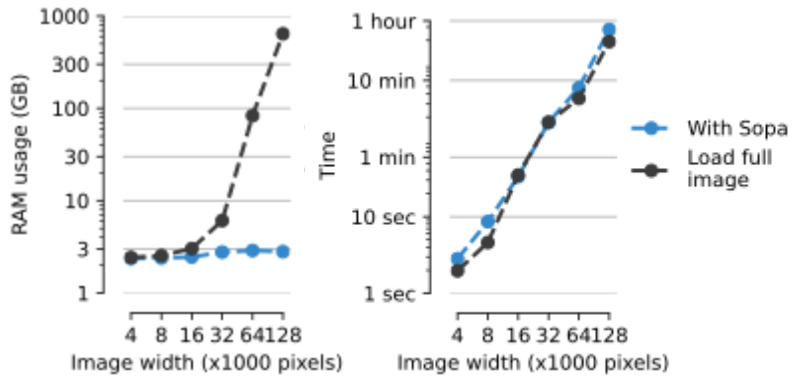


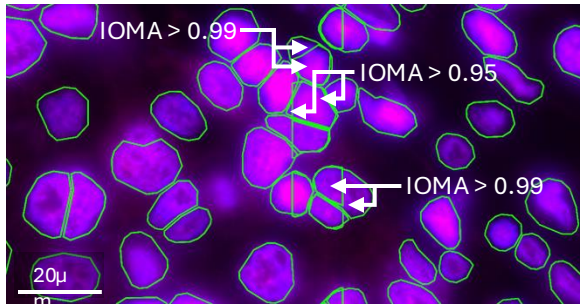
Image on-disk writing



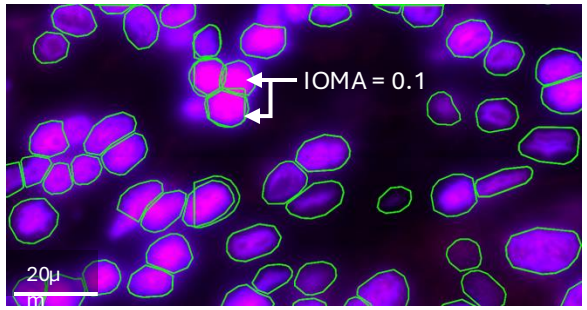
Baysor (Petukhov et al., Nat. Biotechnol.)

Regarding segmentation conflicts

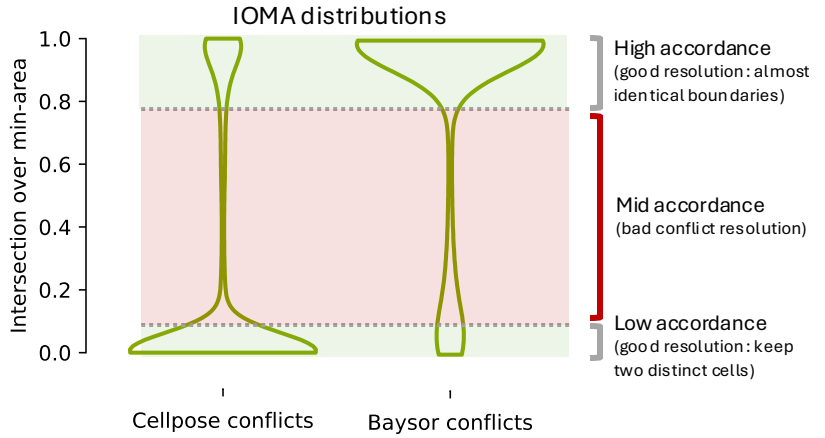
The distribution of the intersection area is close to 0 and 1

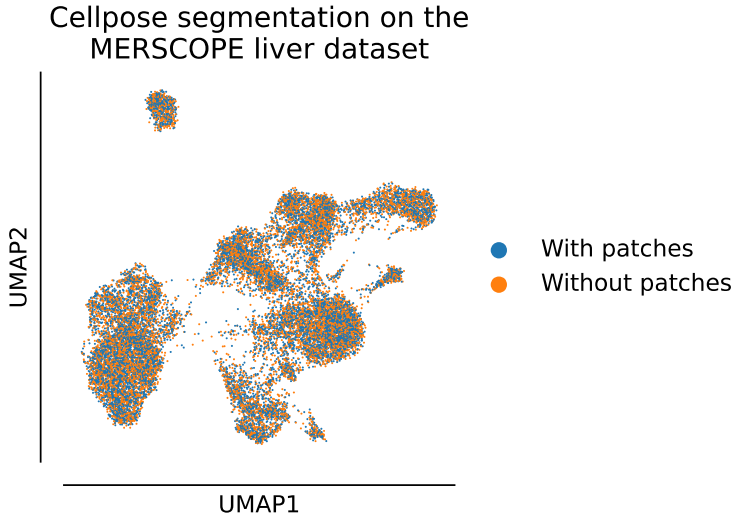


Cells on overlapping patches are segmented twice, the corresponding boundaries are merged



The boundaries of two different cells may slightly overlap (cells kept as distinct)

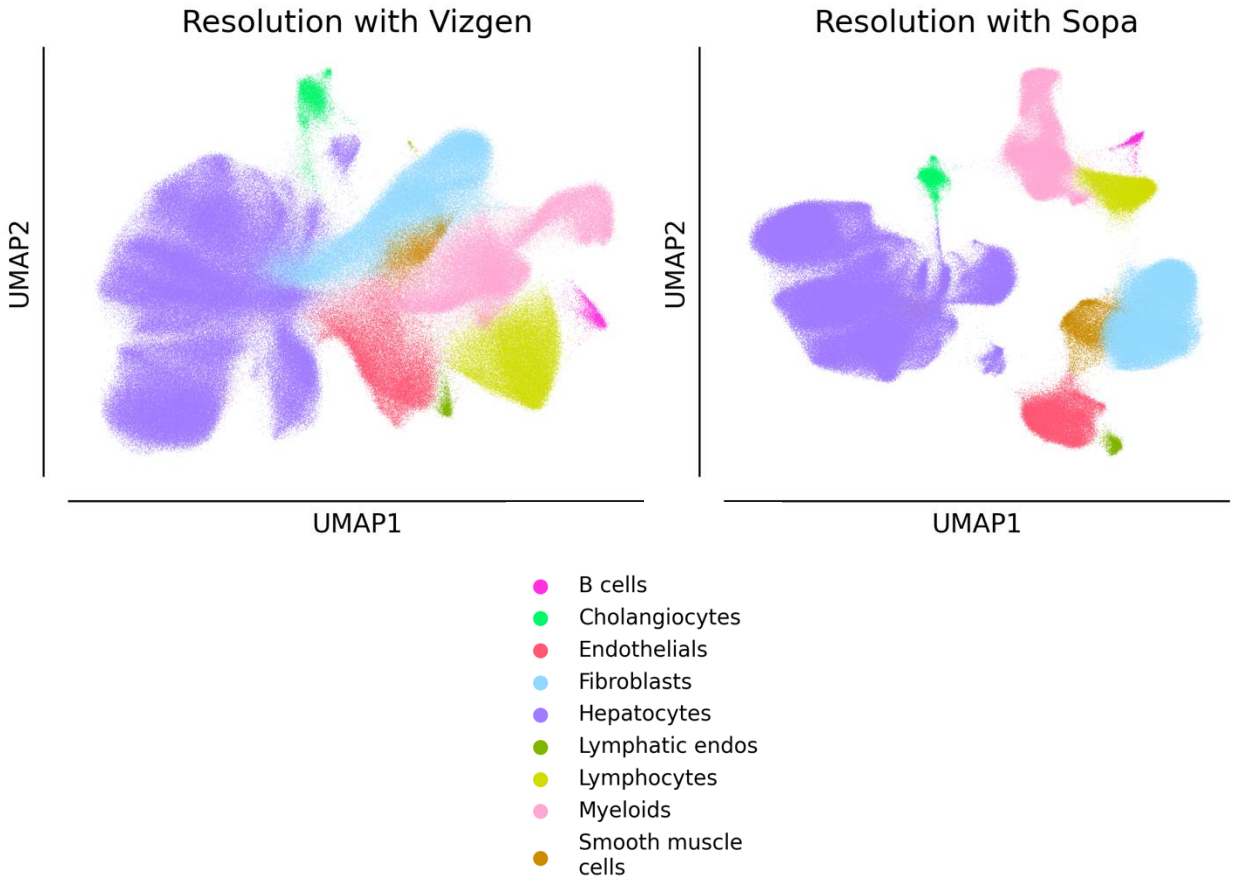


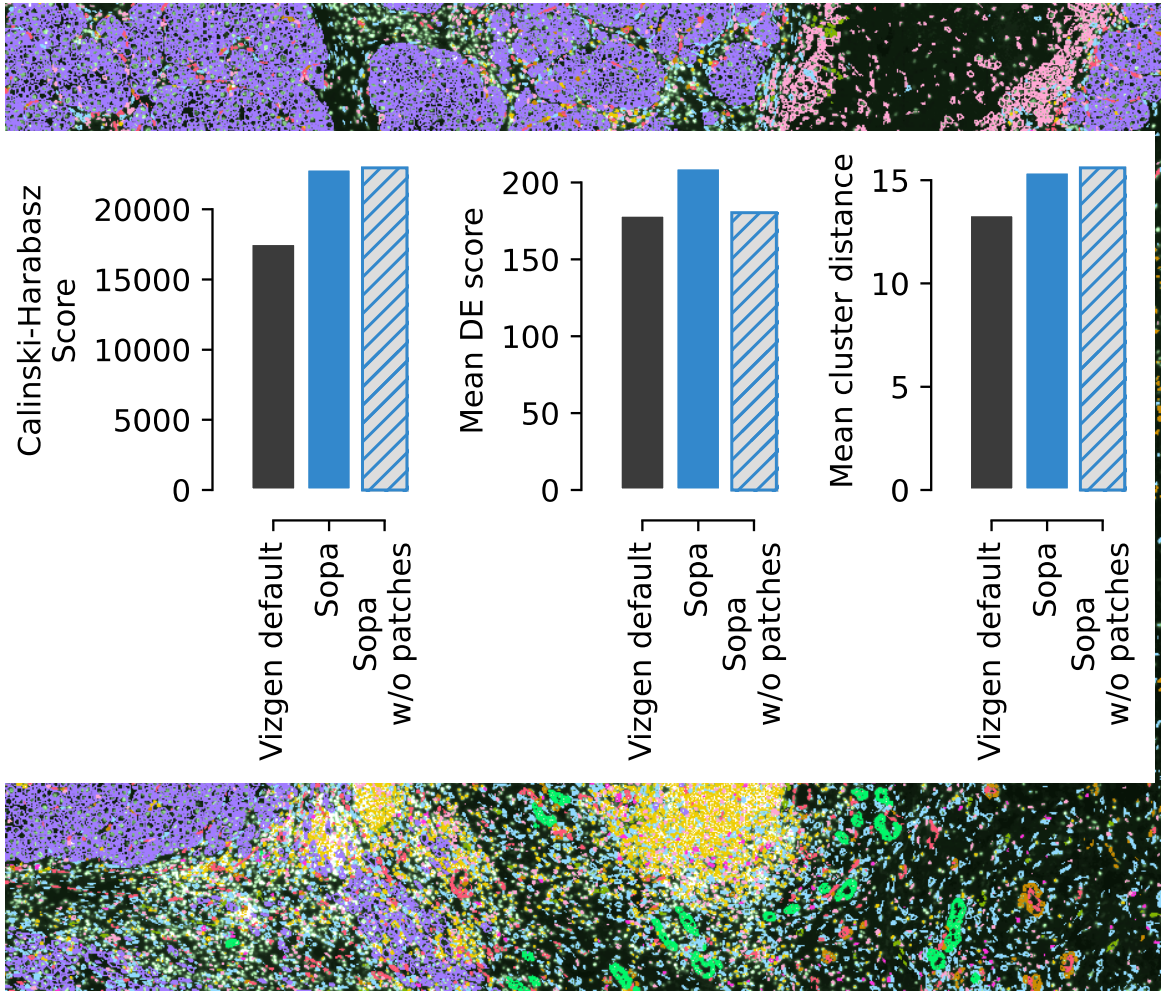


Improved segmentation quality

Comparison: Vizgen segmentation (cellpose-based) and Sopa segmentation (baysor-based)

Note: Baysor can't be run without Sopa due to RAM usage

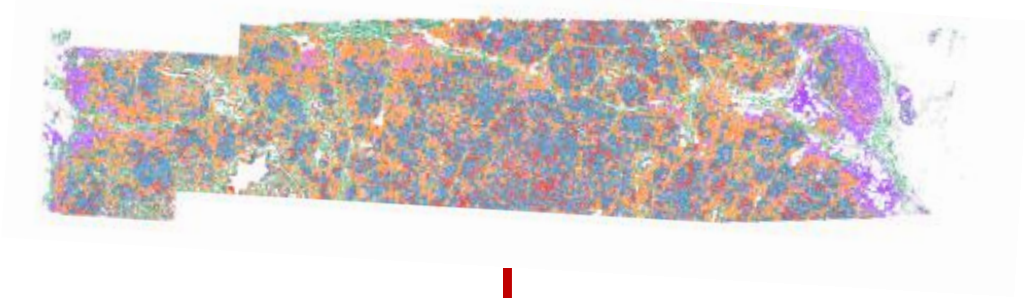




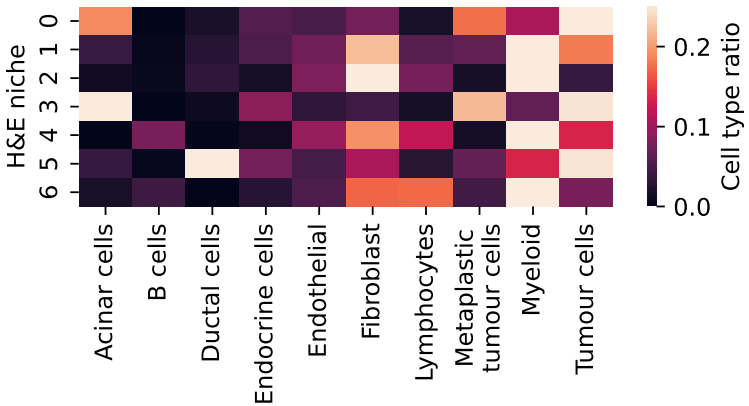
Multi-omics analysis examples

Spatial operations to relate multiple spatial modalities (H&E, RNA, Protein)

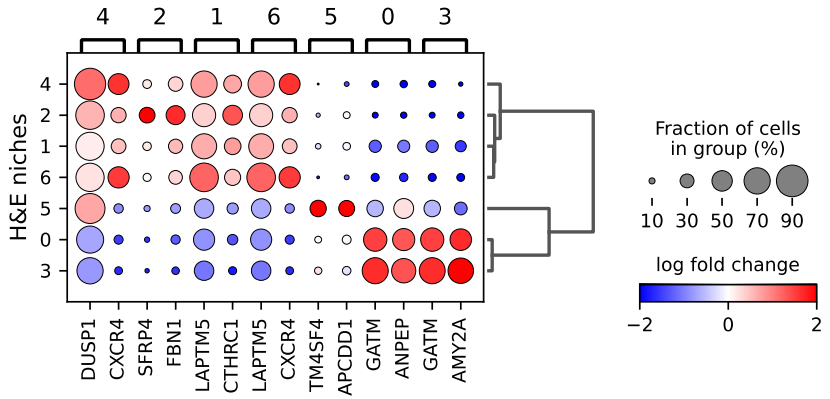
H&E patches clusters (ResNet + Leiden), projected over cells



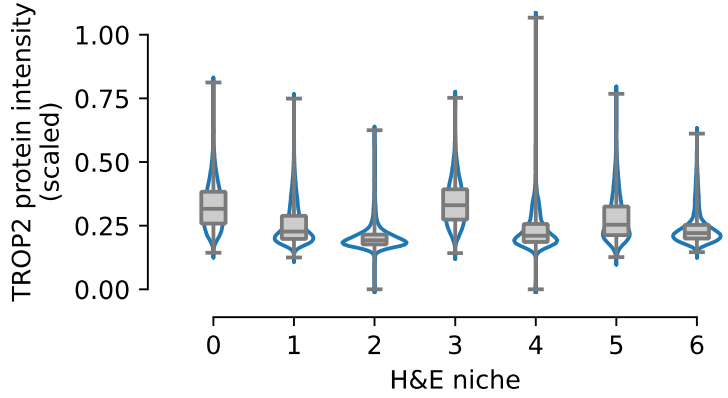
Cell-types



Gene-expression



Proteins



Examples using the API

Usage on Xenium data

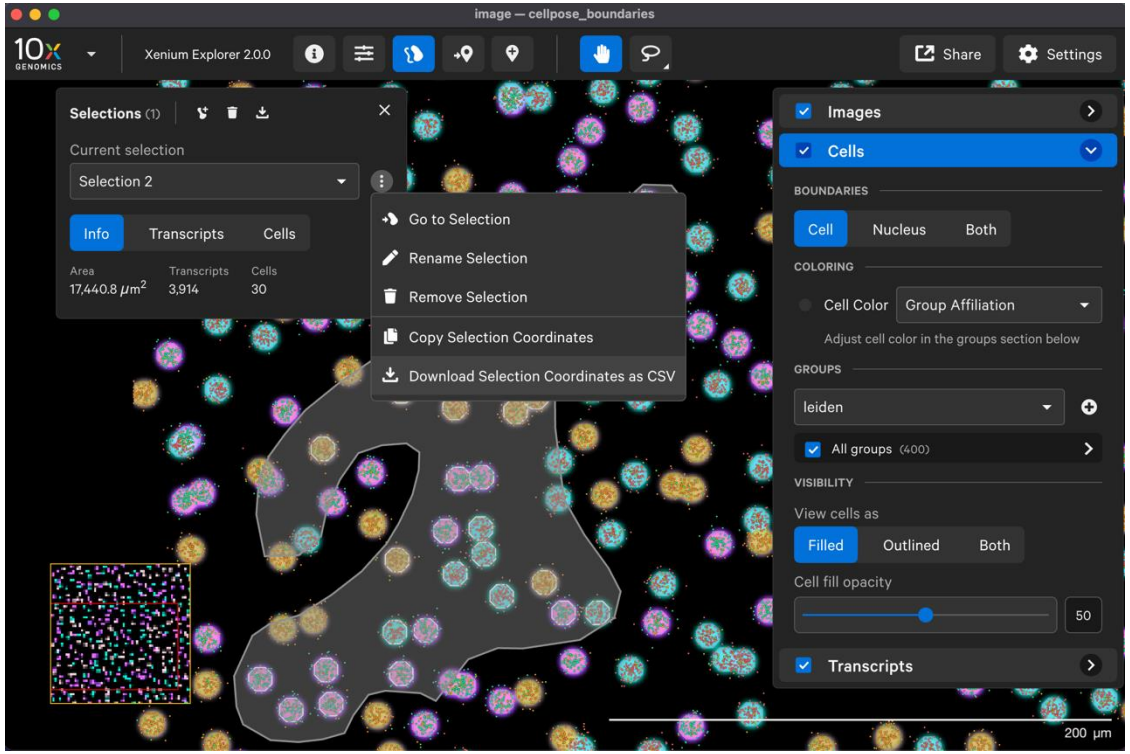
```python
import sopa

### Reading the data

sdata = sopa.io.xenium("/path/to/data_directory")

                    sopa.io.merscope, sopa.io.cosmx, …
### Segmentation

sopa.make_image_patches(sdata)
sopa.segmentation.cellpose(sdata, channels=["DAPI"], diameter=35, gpu=True)

sopa.make_transcript_patches(sdata, patch_width=None, prior_shapes_key="cellpose_boundaries")
sopa.segmentation.proseg(sdata)

### Aggregation

sopa.aggregate(sdata)

### Visualization

sopa.io.explorer.write(sdata, mode="-it") # update the Xenium Explorer

sdata.pl.render_shapes().pl.show() # with spatialdata-plot
```

Xenium data and Xenium Explorer interoperability

SpatialData

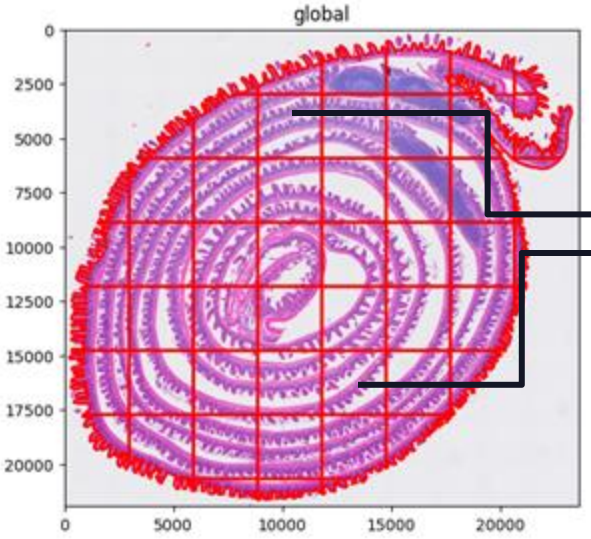
Add the analysis to the explorer

Select cells of interest and analyze them



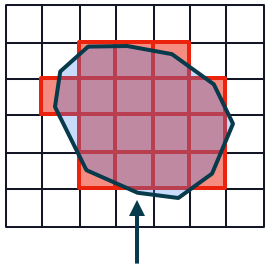
Approach on Visium HD data

Tissue segmentation + patches



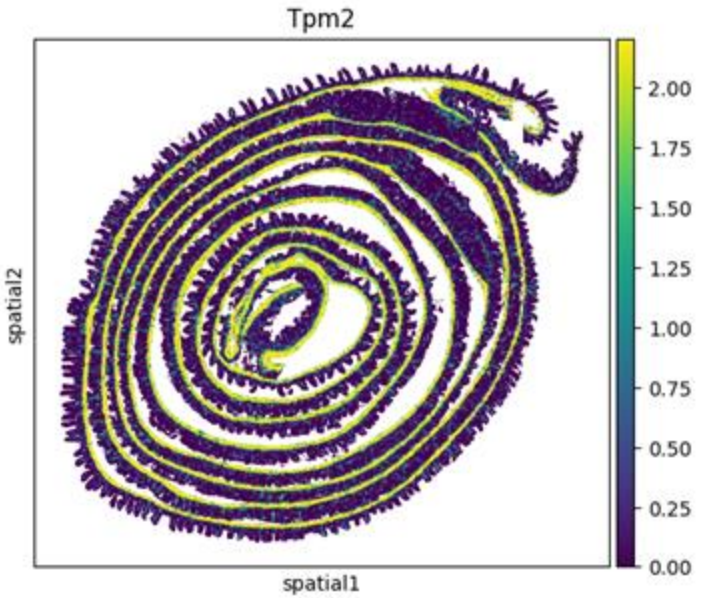
Cellpose

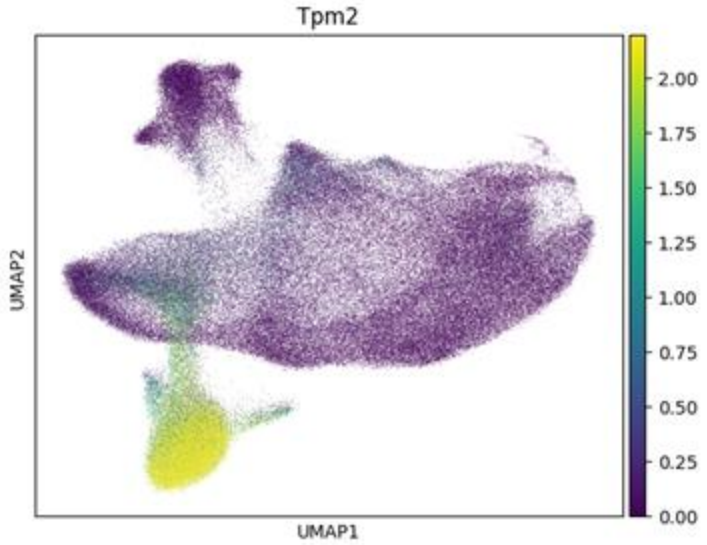
Aggregate the 2-microns bins inside the (expanded) cells



One nucleus boundary

Standard single-cell analysis *(scanpy, squidpy)*





But the API is still very similar

Usage on Visium HD data

```python
import sopa

### Reading the data

sdata = sopa.io.visium_hd("/path/to/data_directory")

### Segmentation

sopa.make_image_patches(sdata)
sopa.segmentation.stardist(sdata)

sopa.segmentation.proseg(sdata, prior_shapes_key="stardist_boundaries")

### Aggregation

sopa.aggregate(sdata)
```

Coming soon

Other usage modes

Command line interface (CLI)

```
> sopa --help # show command names and arguments
> sopa convert merscope_directory --technology merscope # read some data
> sopa patchify image merscope_directory.zarr # make patches for low-memory segmentation
> sopa segmentation cellpose merscope_directory.zarr --diameter 60 --channels DAPI # segmentation
> sopa resolve cellpose merscope_directory.zarr # resolve segmentation conflicts at boundaries
> sopa aggregate merscope_directory.zarr --average-intensities # transcripts/channels aggregation
> sopa explorer write merscope_directory.zarr # convert for interactive vizualisation
```

snakemake CONDA

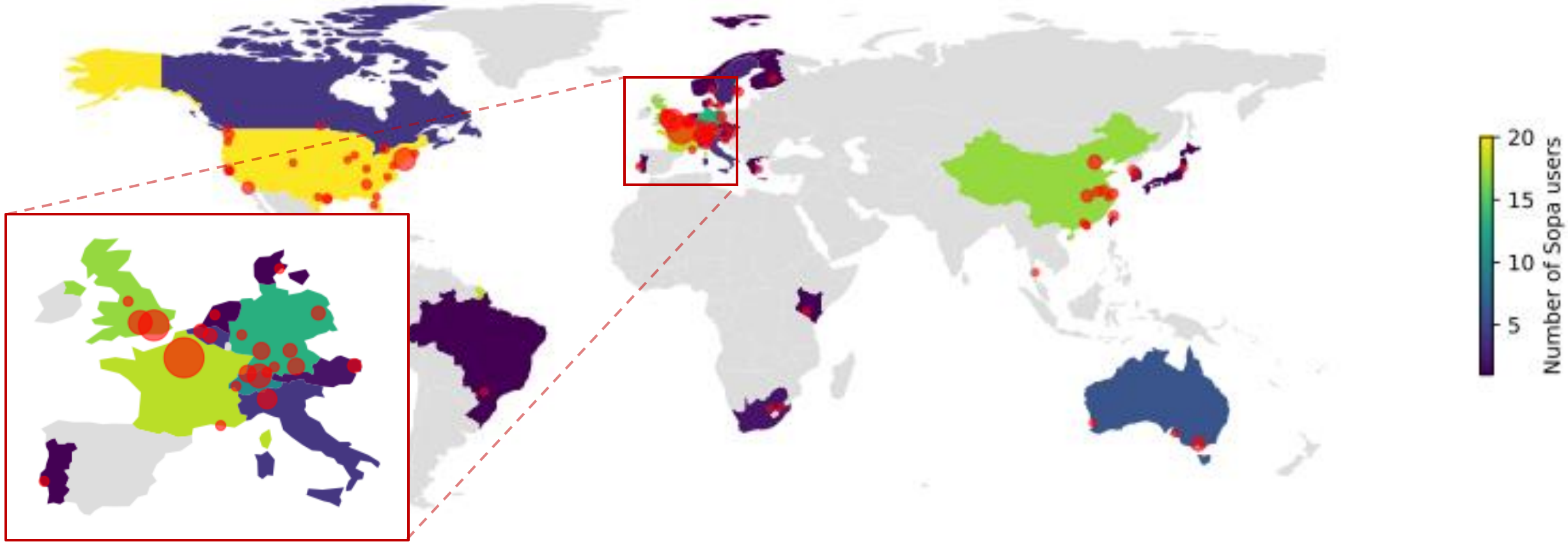
```
> snakemake --configfile=/path/to/config --config data_path=/path/to/data --use-conda
```

Sopa on nf-core docker

```
> nextflow run nf-core/sopa -profile docker --input samplesheet.csv -params-file <PARAMS_FILE>
```

Usage and impact of Sopa *(in Nov 2024, i.e., after 12 months)*

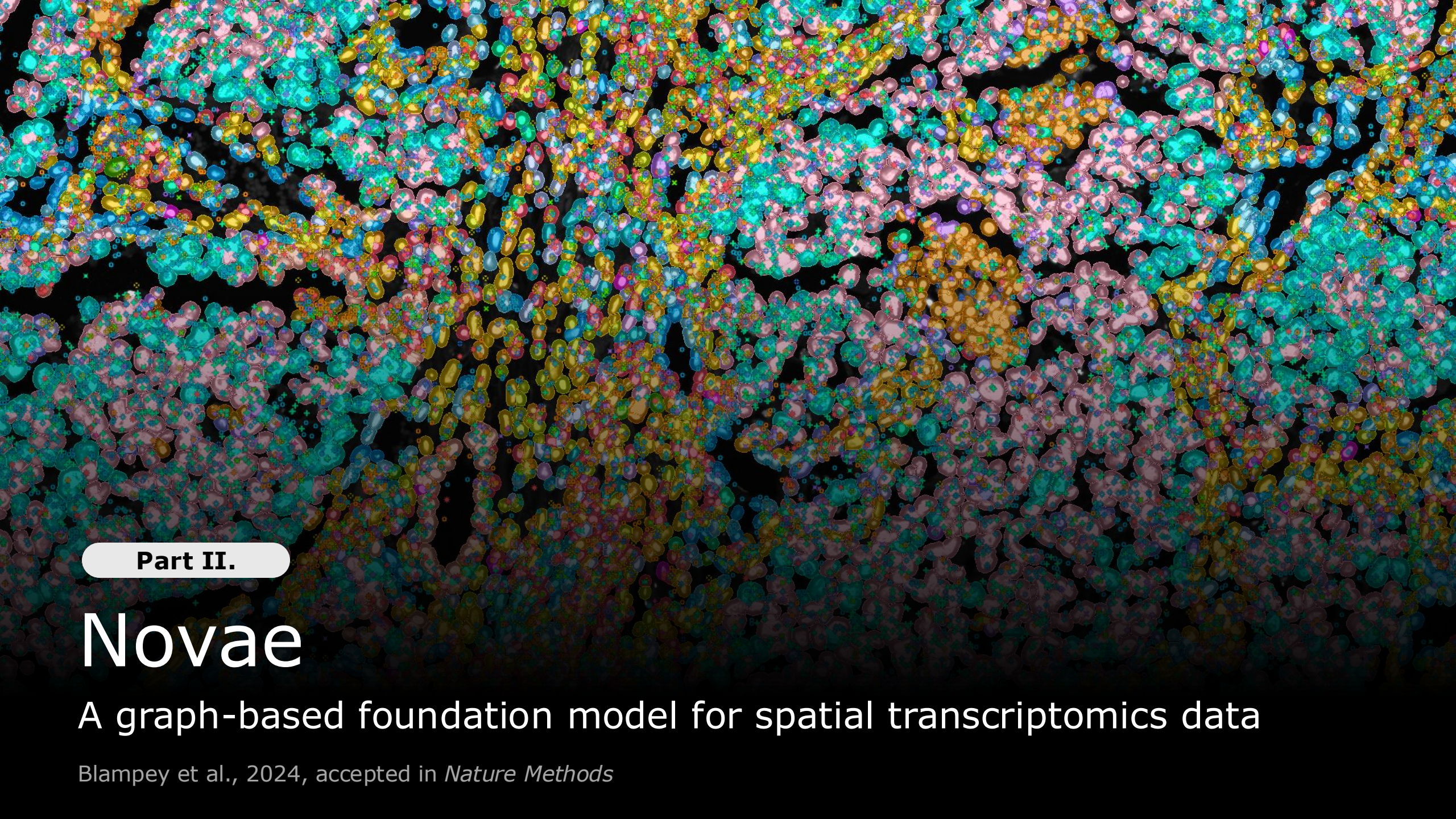
Dots represent researchers who gave a "star" to the repository, or opened an issue/PR



Many other features were not shown, try it out!

https://gustaveroussy.github.io/sopa/

https://github.com/gustaveroussy/sopa



Part II.

Novae

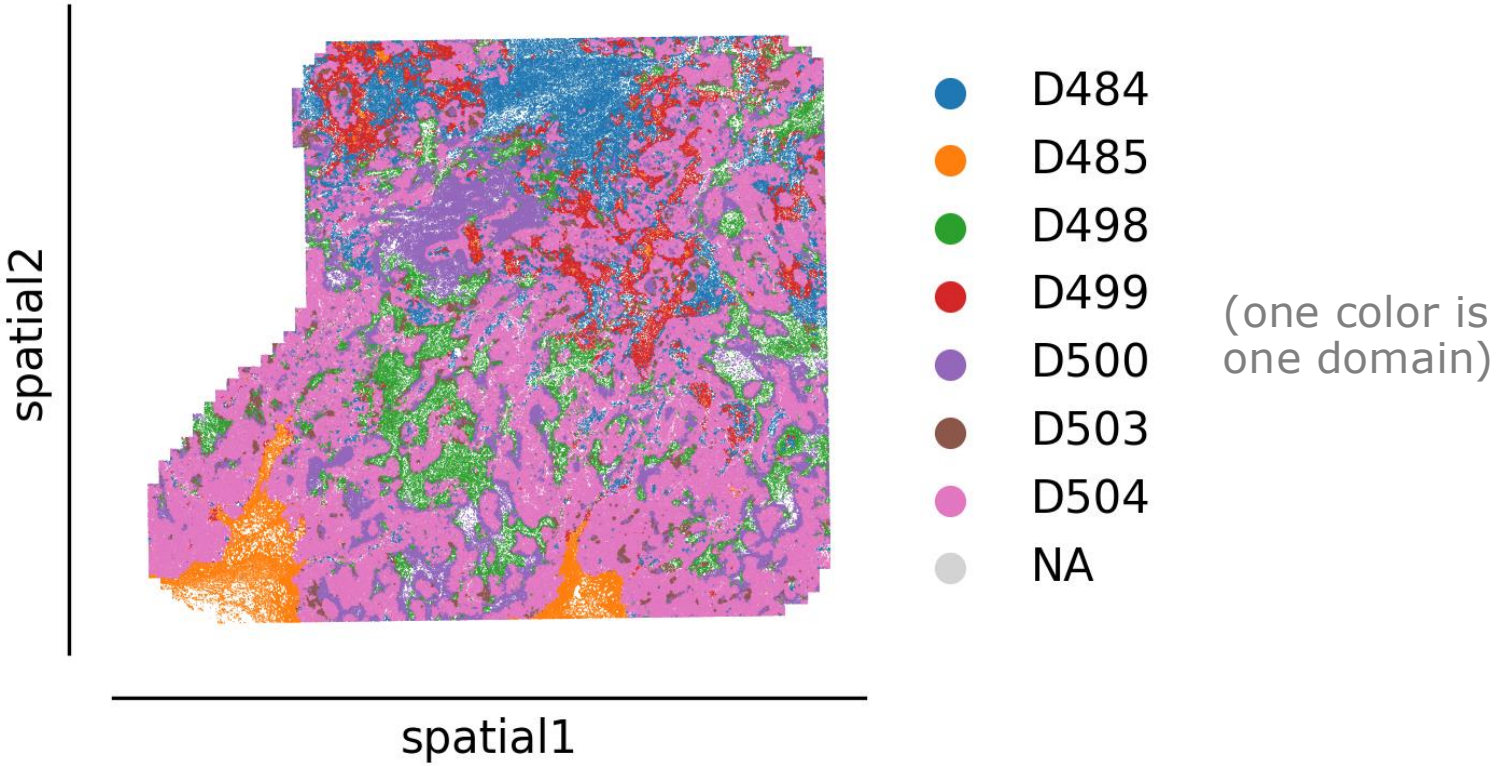
A graph-based foundation model for spatial transcriptomics data

Blampey et al., 2024, accepted in *Nature Methods*

Definition of a spatial domain (or niche)

A spatial domain (a.k.a. niche) is a localized cellular microenvironment within tissues.

> These domains reflect distinct biological functions and interactions



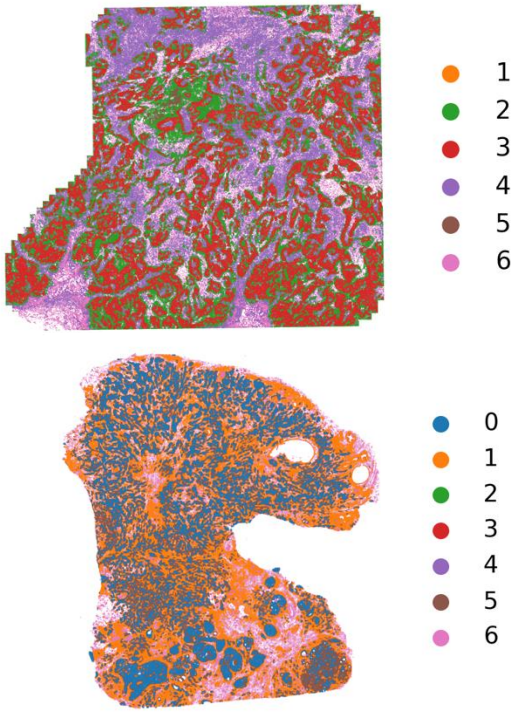
Which implications for tissue function / disease progression?

Main challenges in spatial domains analysis

Existing tools have limitations (STAGATE, SEDR, SpaceFlow, GraphST)

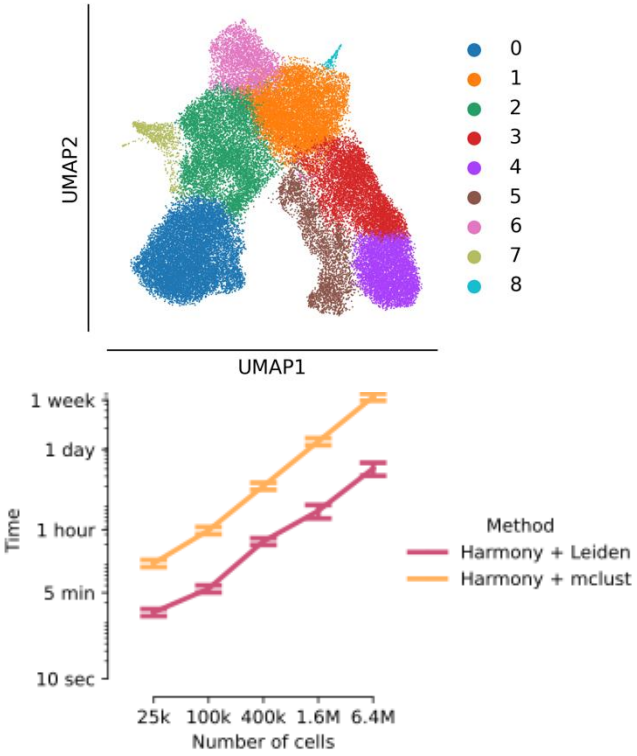
1

Identifying cross-slide spatial domains is challenging



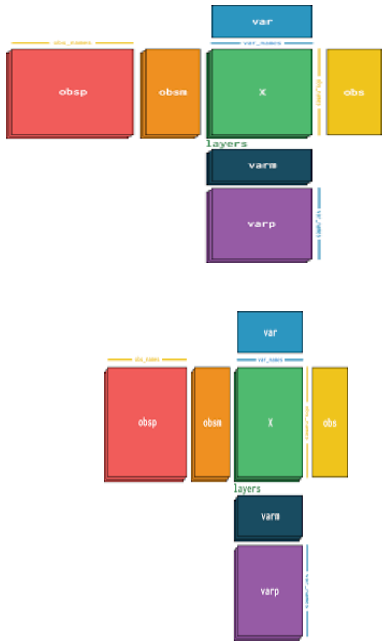
2

Depending on external tools (e.g. Harmony / Leiden) can be slow



3

Many studies are composed of different gene panels (technologies evolve)



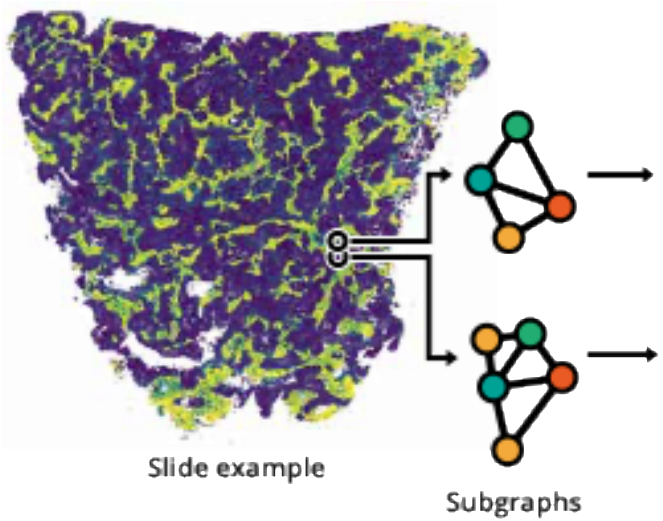
STAGATE: Dong et al. (2022, *Nat. Commun.*), SEDR: Xu et al. (2024, *Genome Med.*), SpaceFlow: Ren et al. (2022, *Nat. Commun.*), GraphST: Long et al. (2023, *Nat. Commun.*)

Method overview

$$\mathbf{Q}^* := \operatorname{argmin}_{\mathbf{Q}\in\mathcal{Q}}\Big(\operatorname{Tr}(\mathbf{Q}\mathbf{C}\mathbf{Z}^\top) - \epsilon H(\mathbf{Q})\Big)$$

Self-supervision model (inspired from the SwAV framework)

$$\mathbf{p}_i = \left(\frac{\exp\left(\frac{1}{\tau}\frac{\mathbf{z}_i^T}{||\mathbf{z}_i||_{L^2}}\mathbf{c}_k\right)}{\sum_{1\leq k'\leq K}\exp\left(\frac{1}{\tau}\frac{\mathbf{z}_i^T}{||\mathbf{z}_i||_{L^2}}\mathbf{c}_{k'}\right)}\right)_{1\leq k\leq K}$$

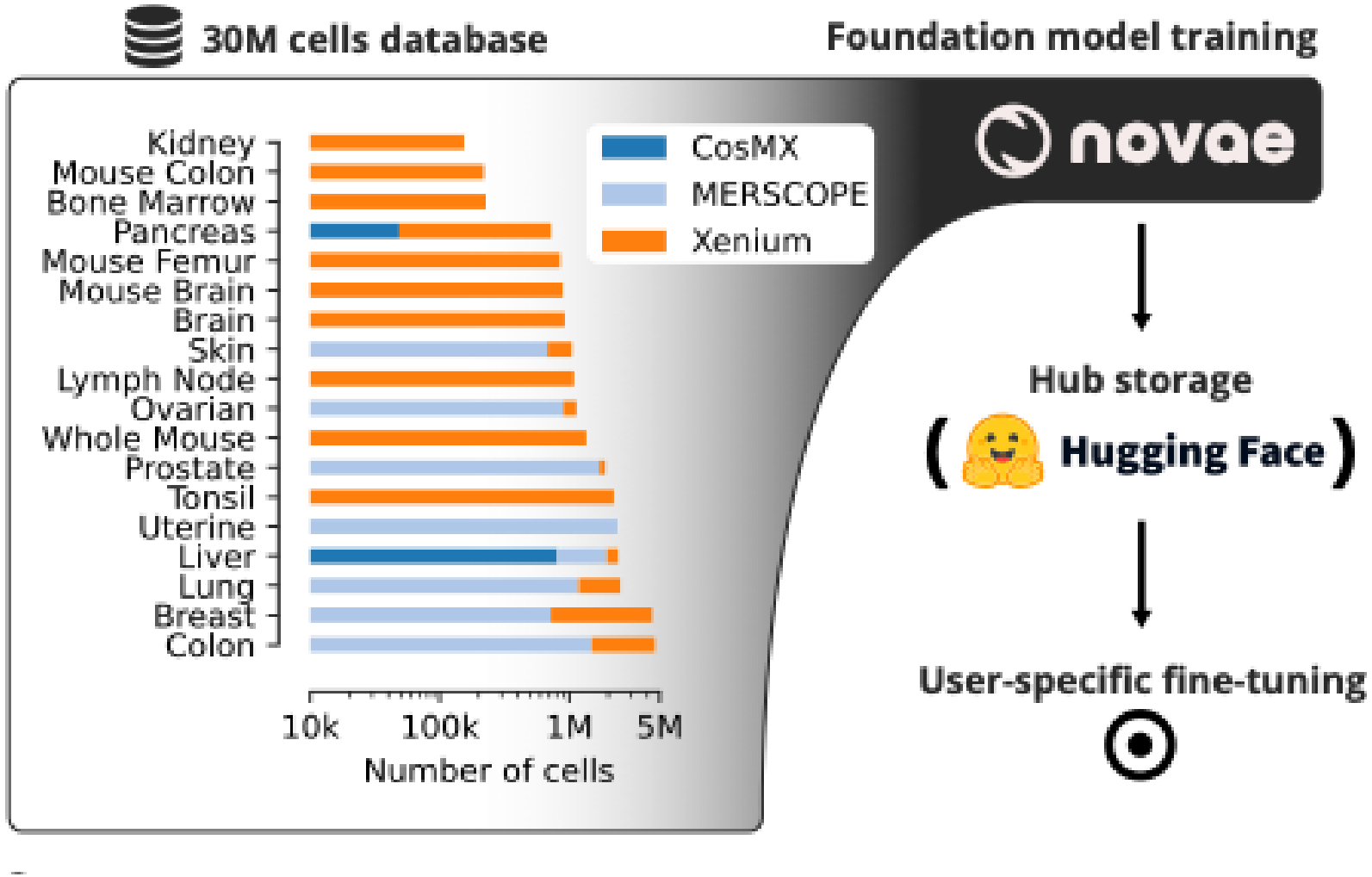


$$\operatorname{embed}(\mathbf{x}_i, \mathcal{P}) = \frac{\sum_{j=1}^{P} x_{ij}(\mathbf{W}\mathbf{v}_{\mathcal{P}[\mathbf{j}]} + \mathbf{b})}{\sqrt{\sum_{j=1}^{P}(\mathbf{W}\mathbf{v}_{\mathcal{P}[\mathbf{j}]} + \mathbf{b})^2}}$$

$$\mathcal{L}(\theta, i, j) = -\sum_{k=1}^{K}\left(q_{ik}\log p_{jk} + q_{jk}\log p_{ik}\right)$$

Training and sharing (single-cell resolution dataset)

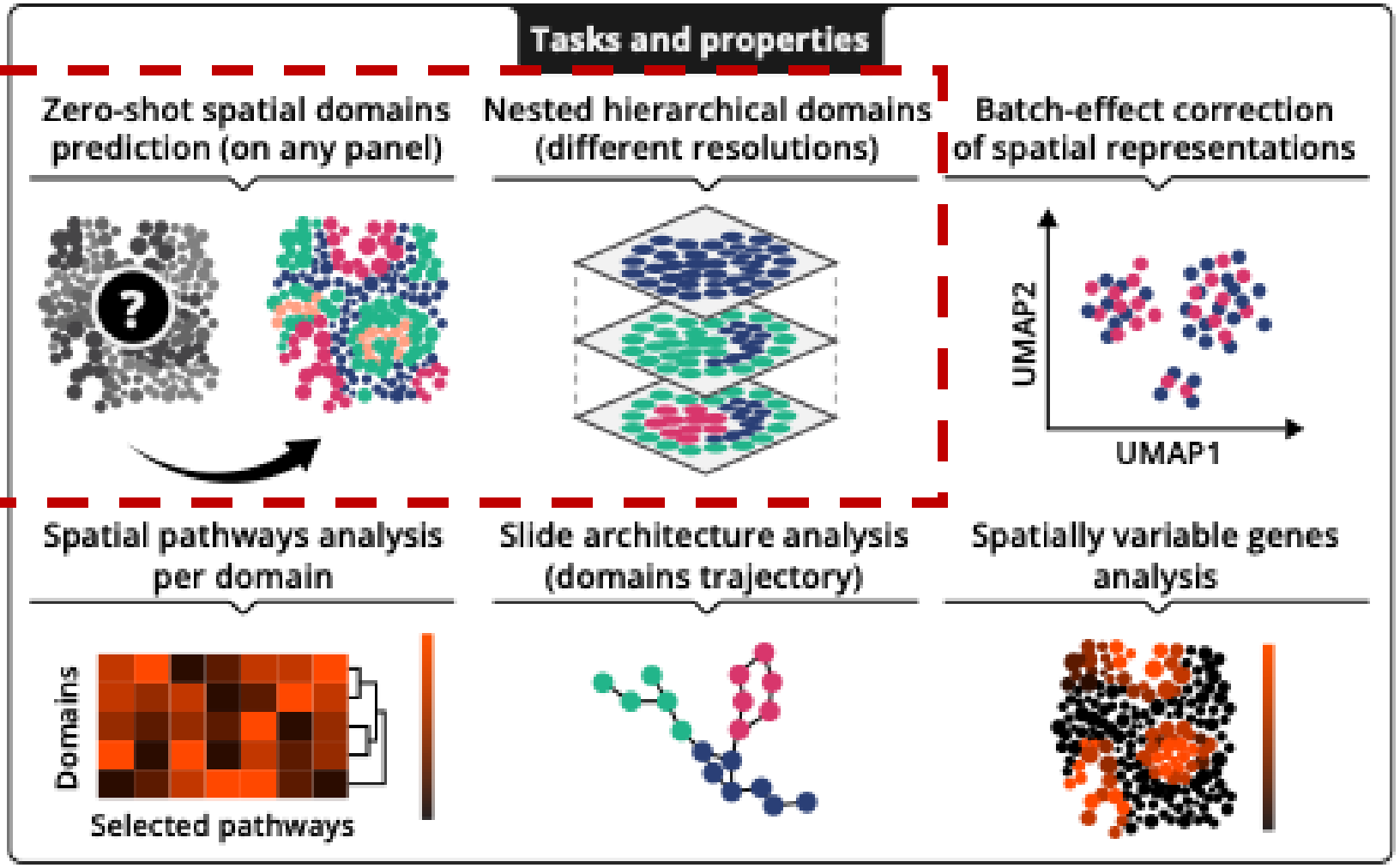
Public database collected from single-cell resolution technologies



Blampey et al. (2024) | Novae: a graph-based foundation model for spatial transcriptomics data (accepted in *Nature Methods*)

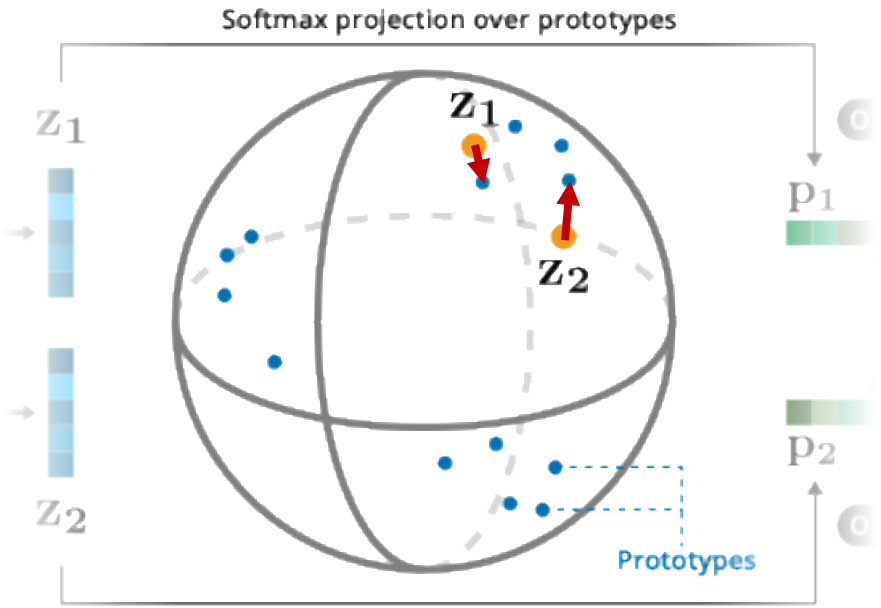
Tasks and properties of Novae

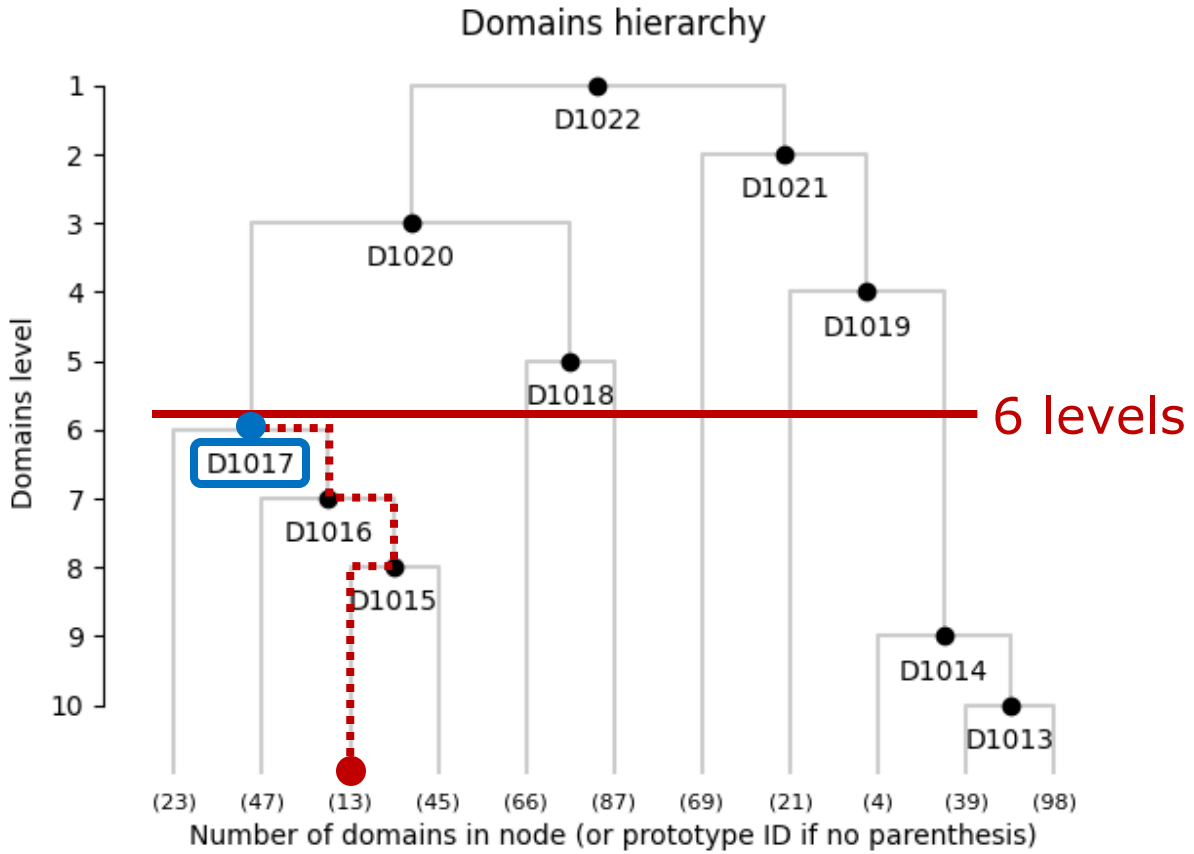
Details



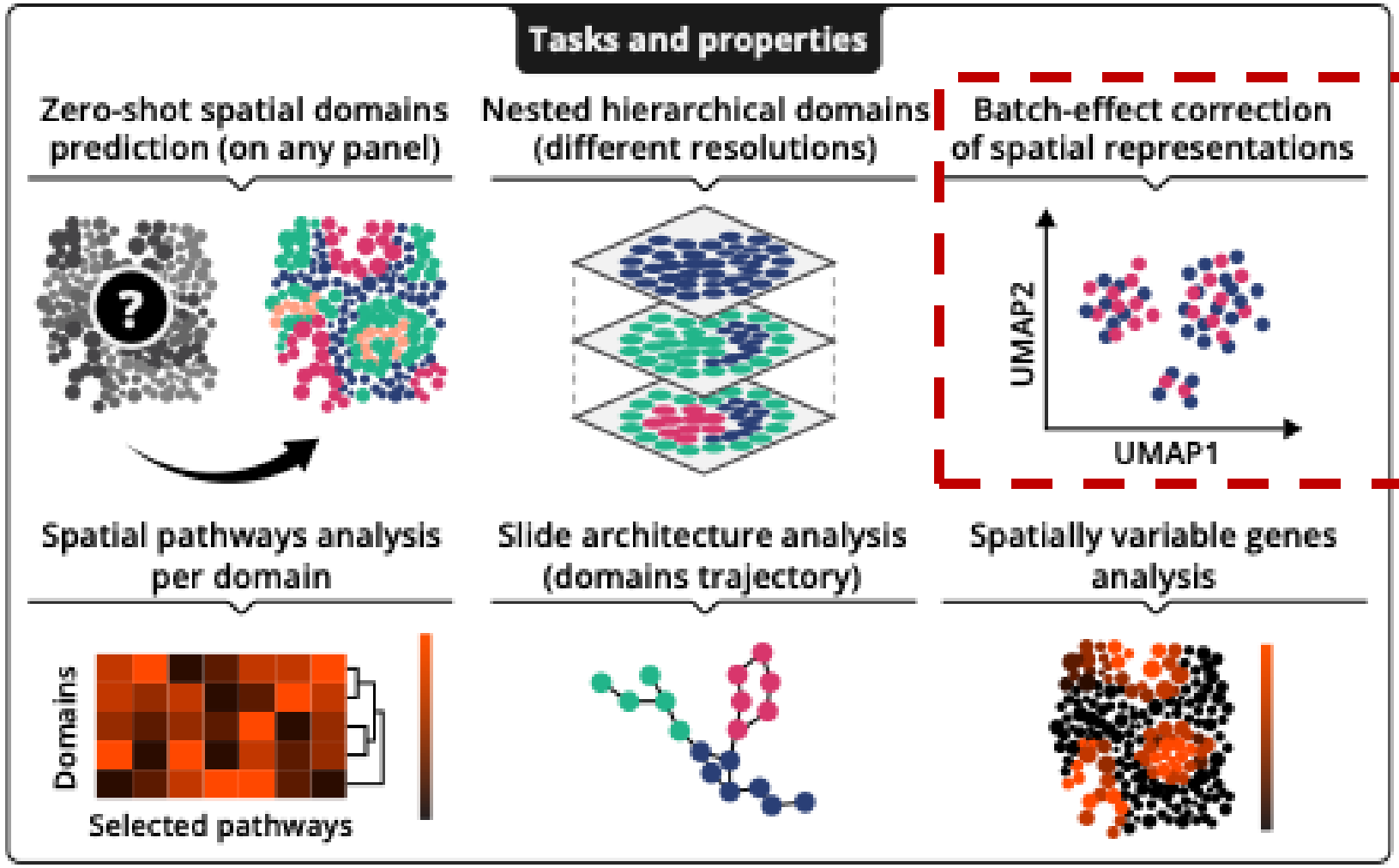
A hierarchy of prototypes

Hierarchical clustering of the prototypes, and assignment to the closest prototype



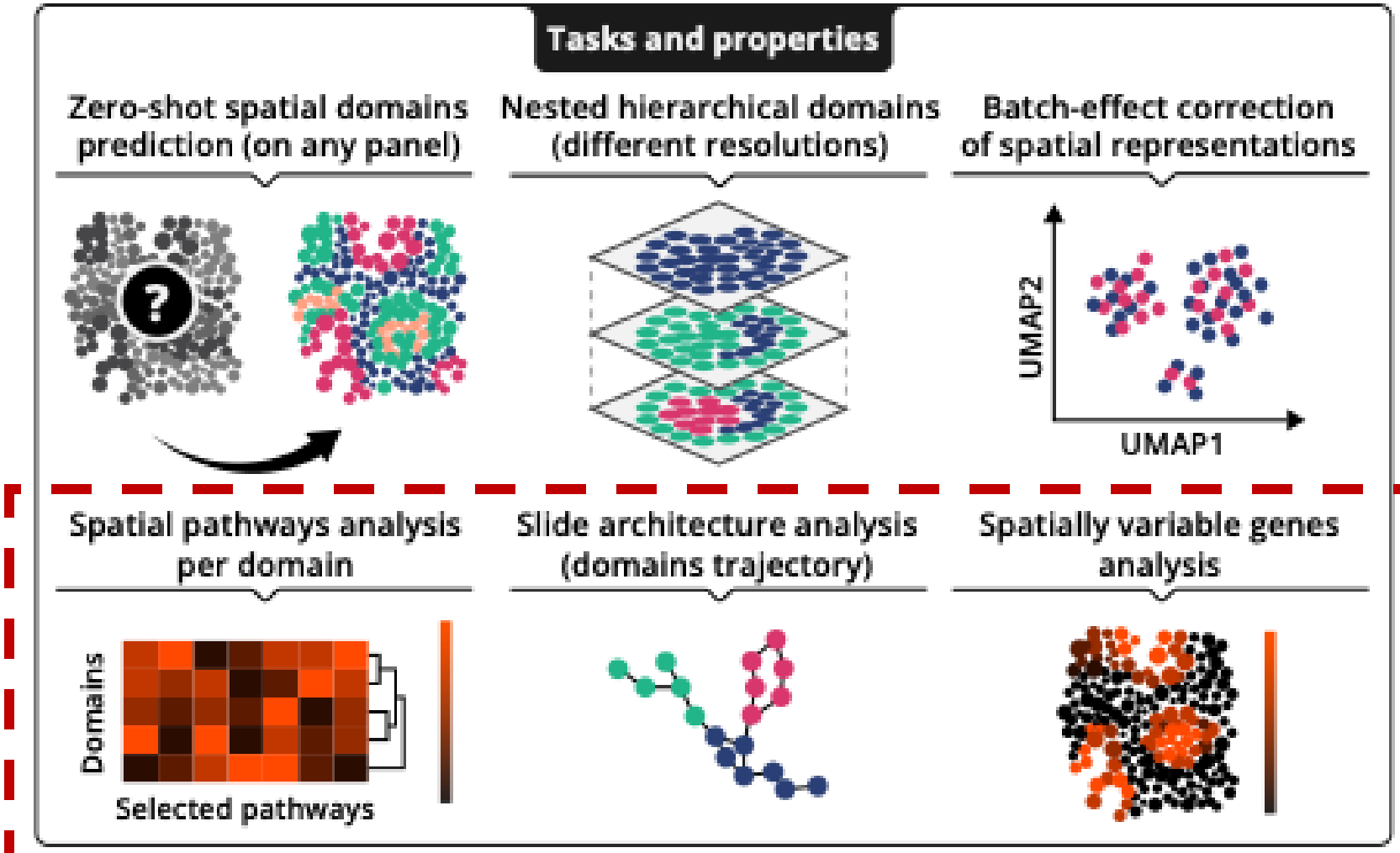


Tasks and properties of Novae



Due to the Optimal Transport

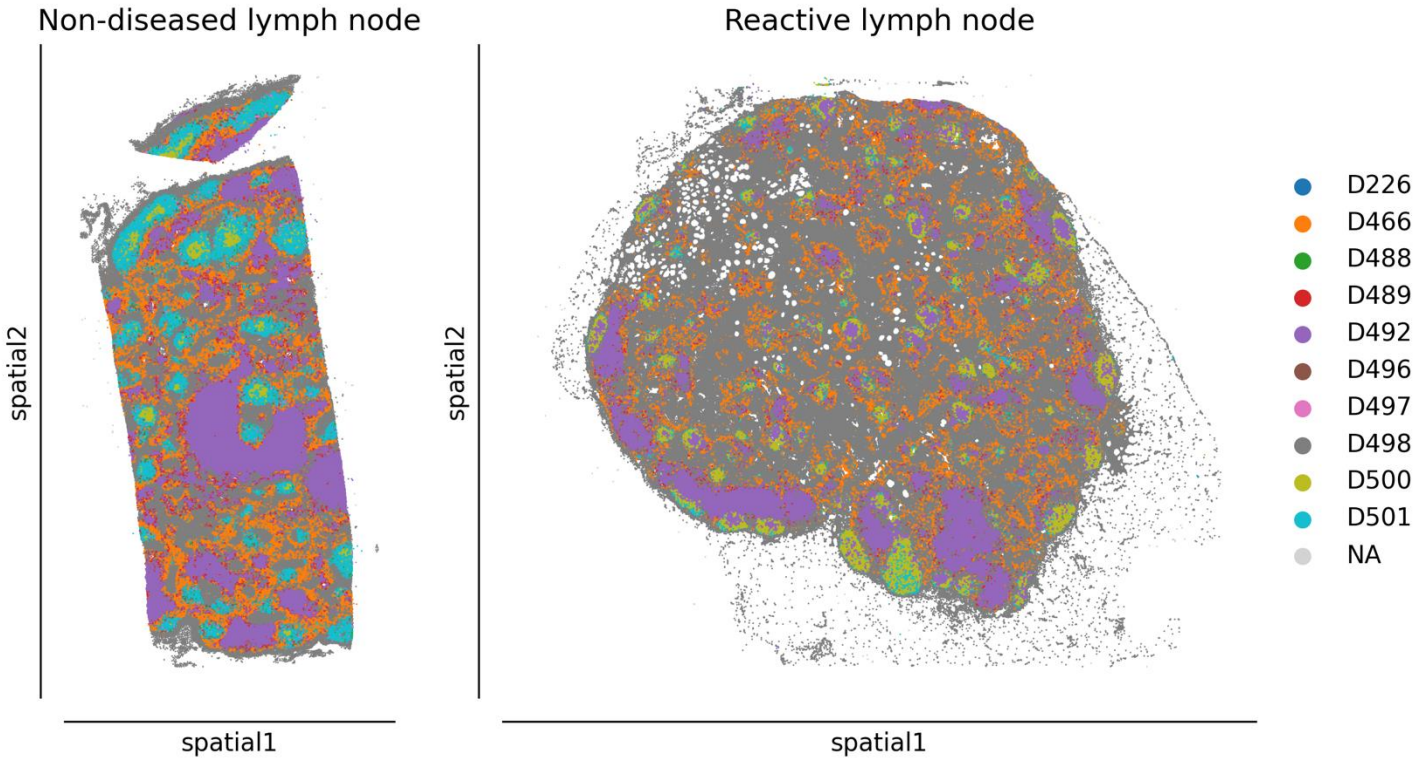
Tasks and properties of Novae

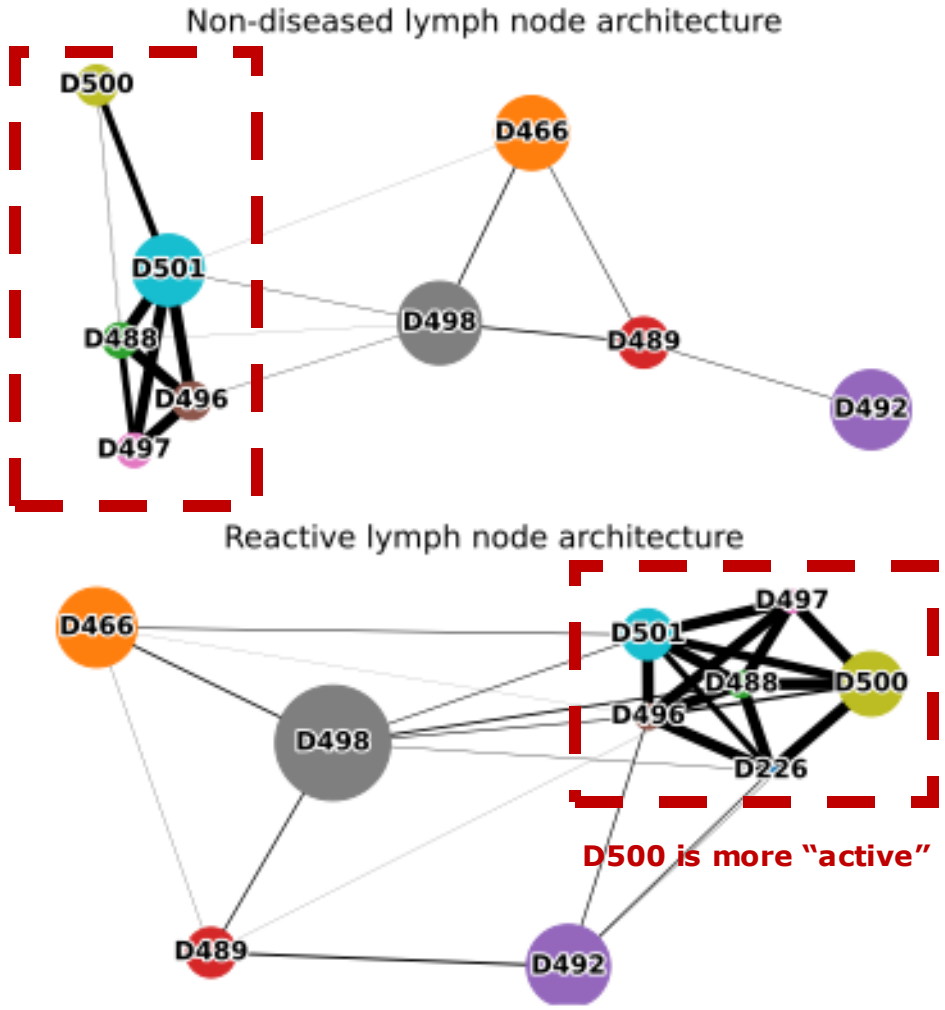


Multiple examples

[Example 1] Architecture of a reactive lymph node

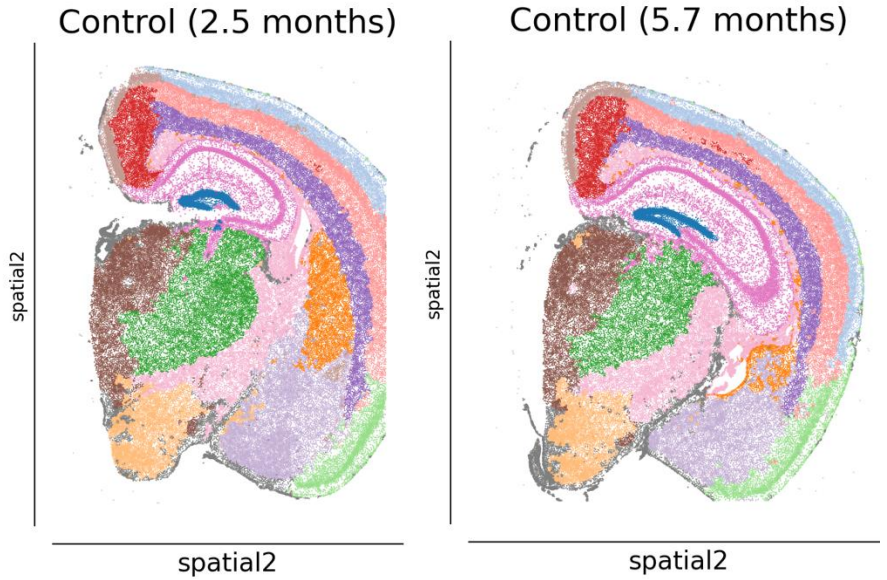
Comparison between a non-diseased and a reactive lymph node

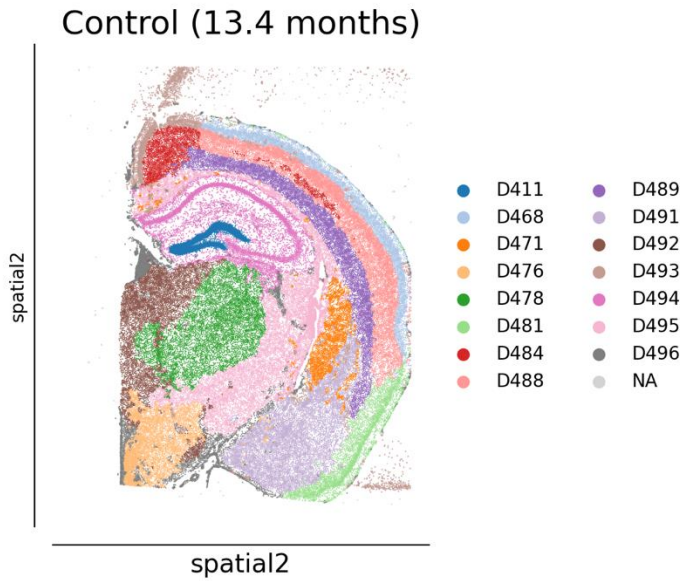




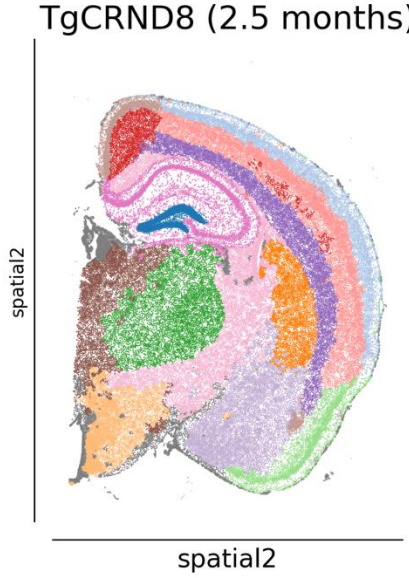
[Example 2] Mice brain slides with Alzheimer-like pathology

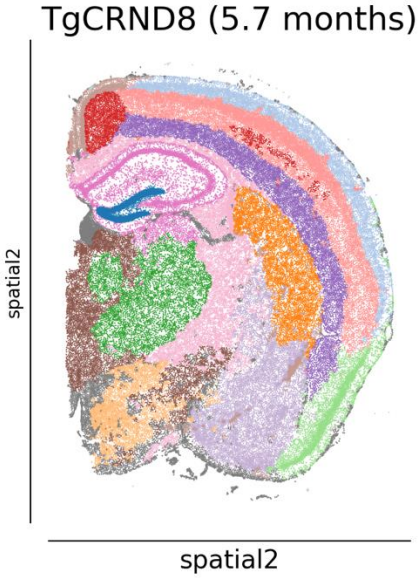
Control Mice (3 time points)

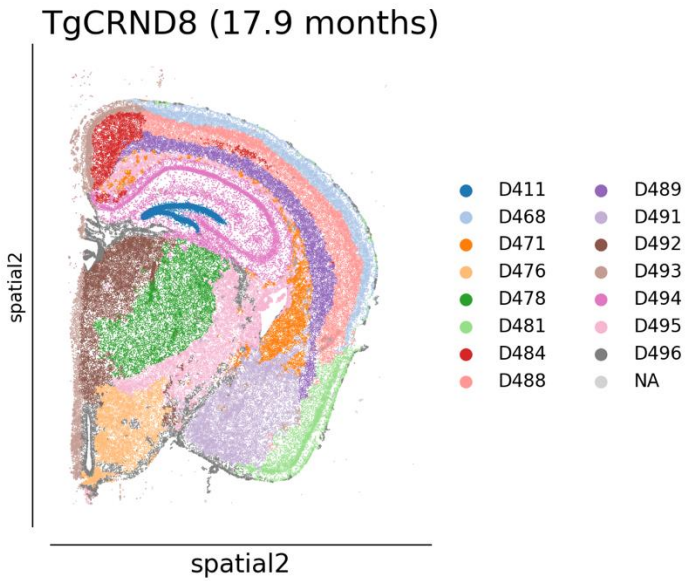




TgCRND8 Mice (3 time points)

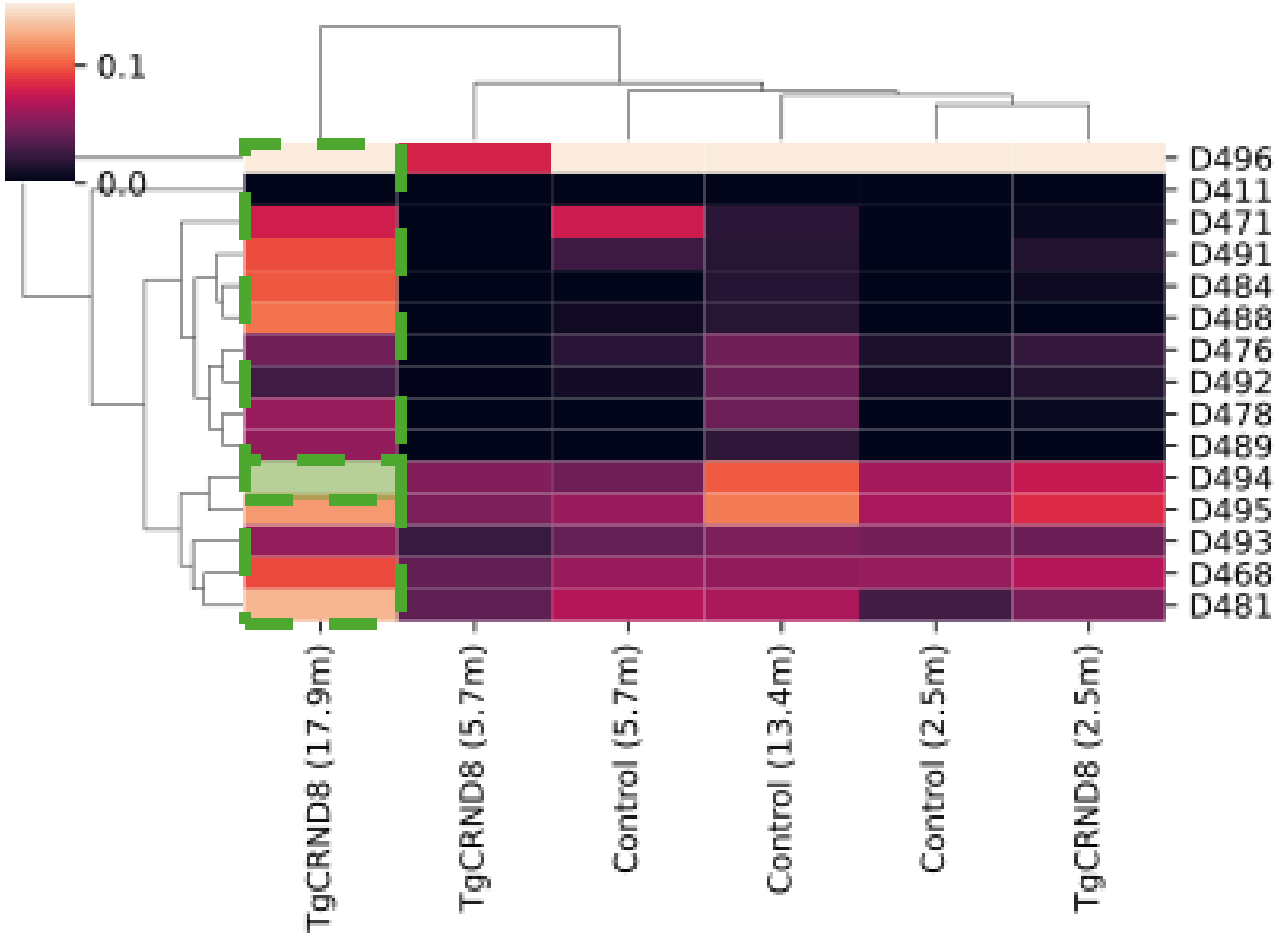


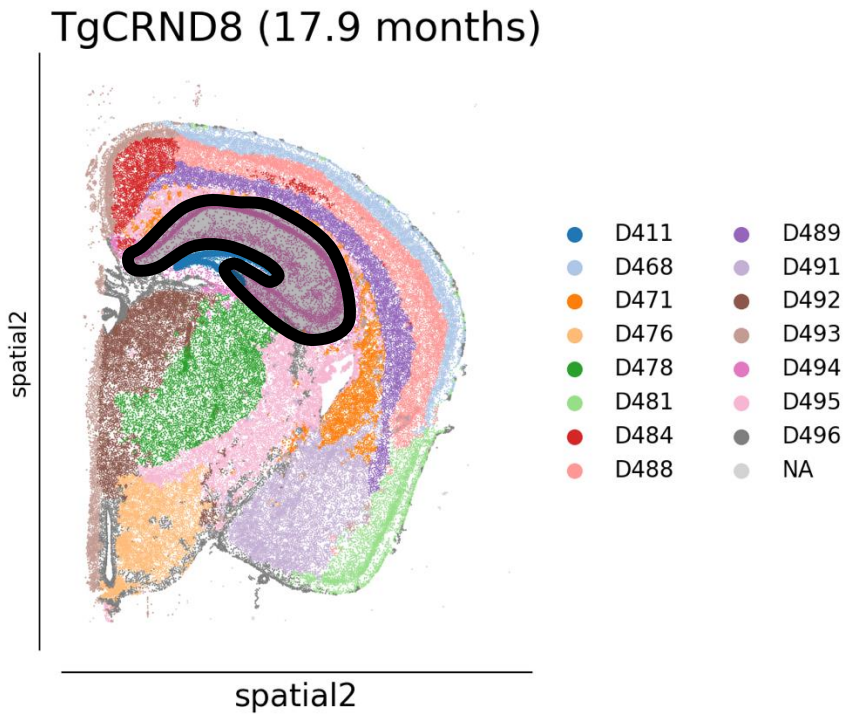




[Example 2] Mice brain slides with Alzheimer-like pathology

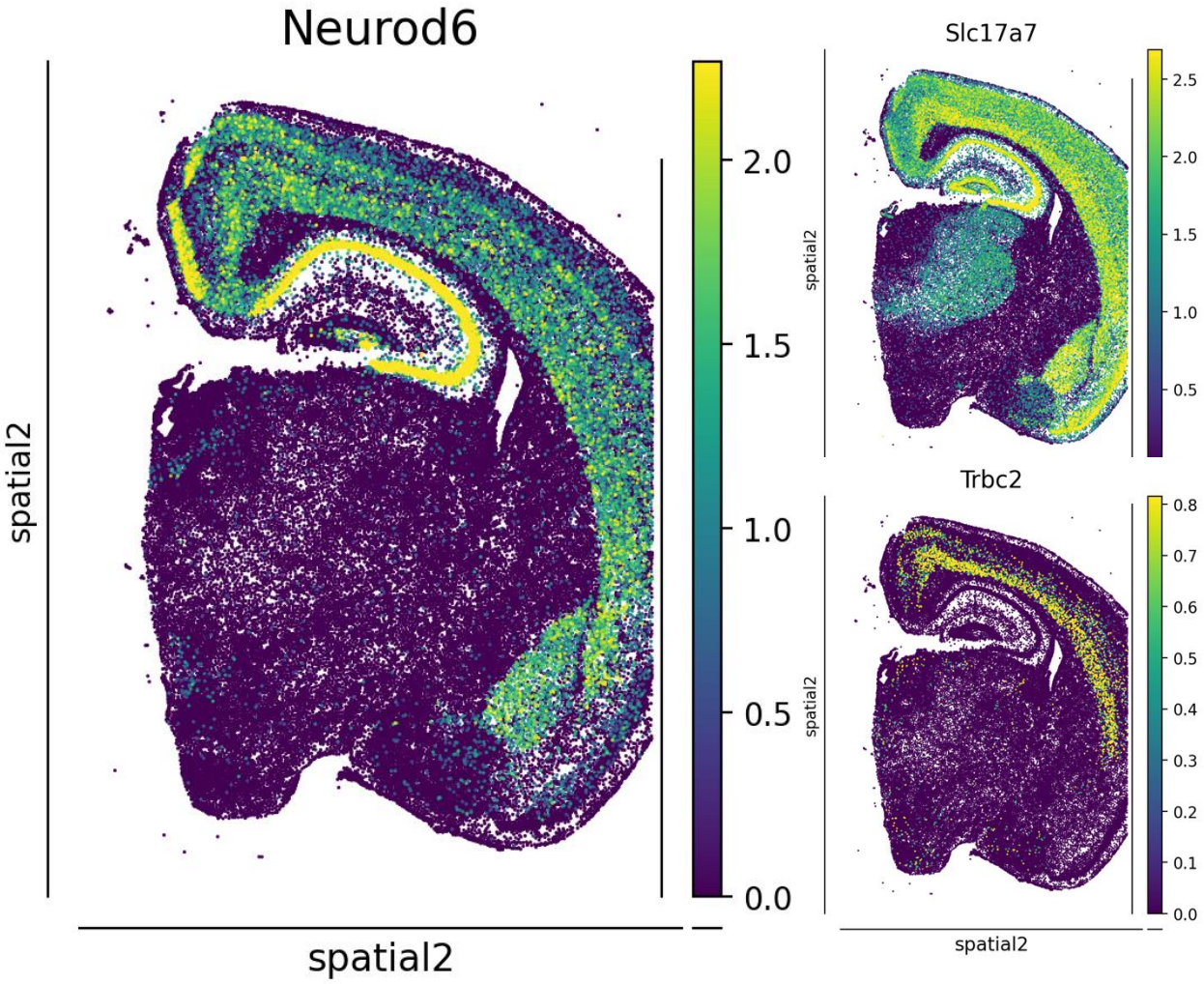
A brain aging pathway is up-regulated in a specific domains of the TgCRND8 17m mouse

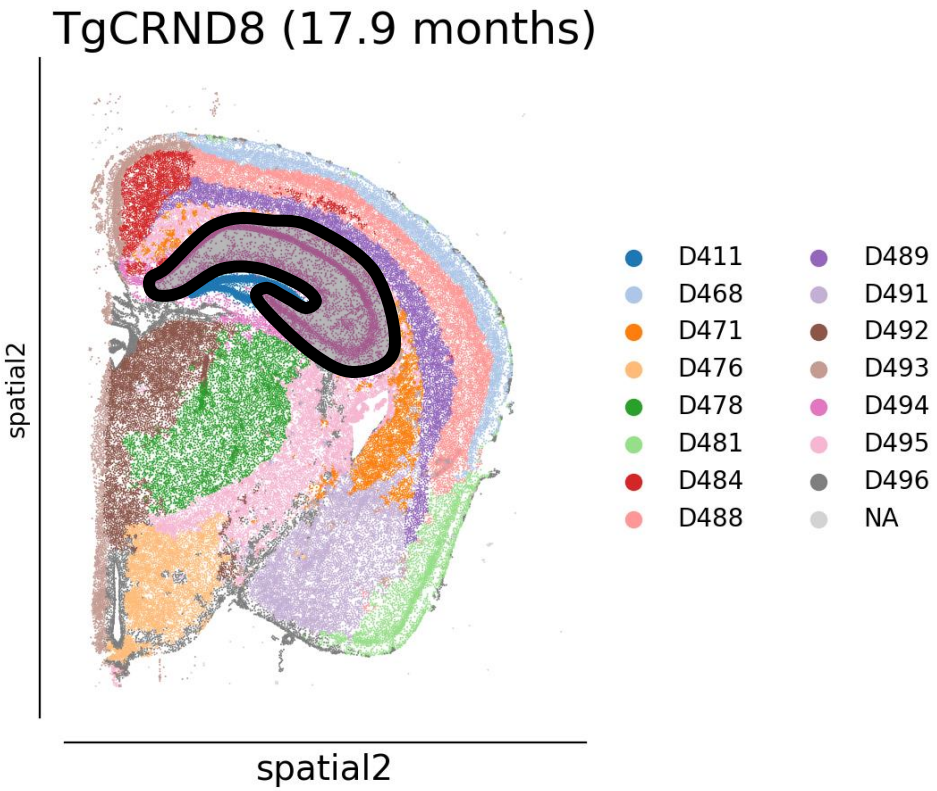




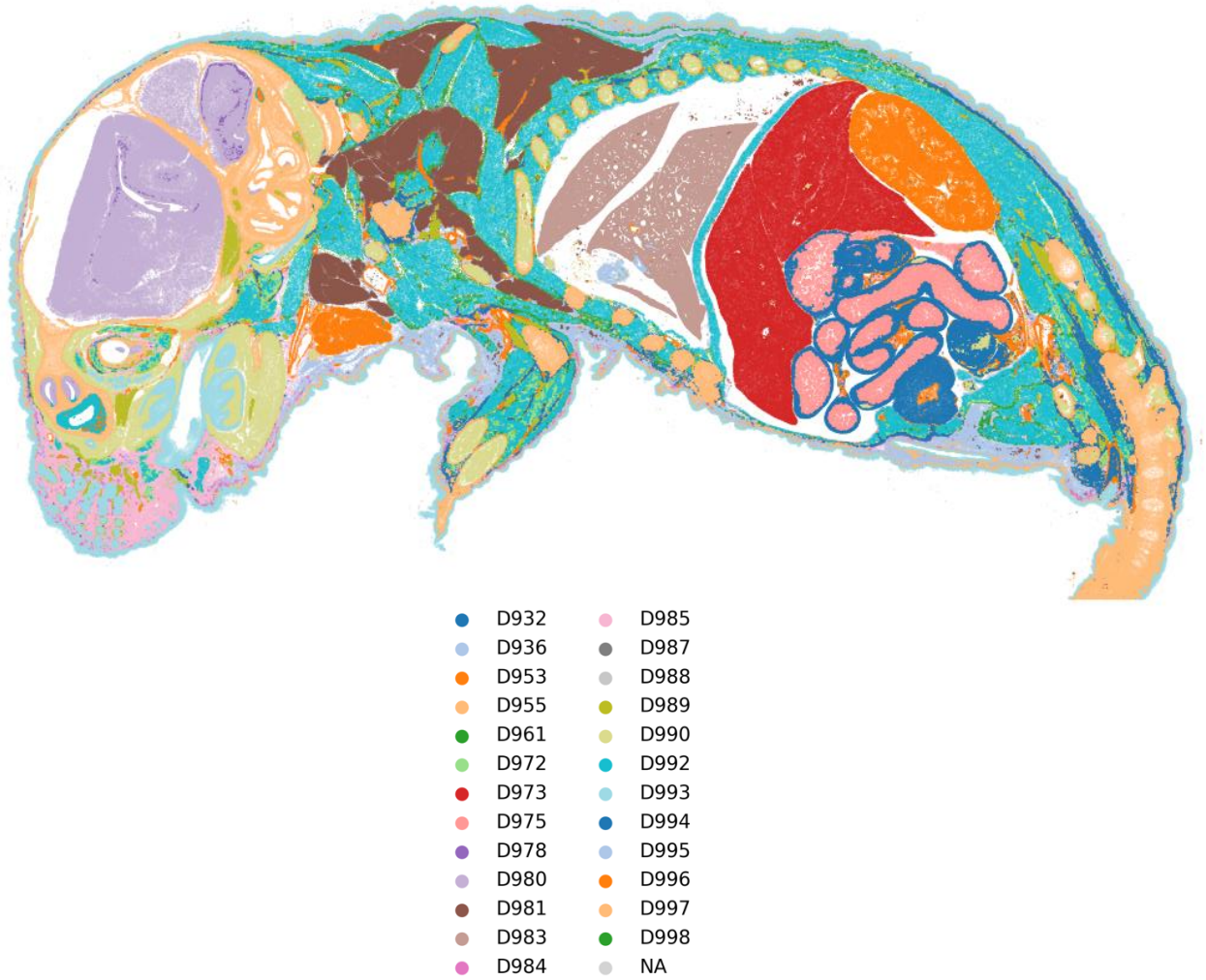
[Example 2] Mice brain slides with Alzheimer-like pathology

Identification of Spatially Variable Genes (DEGs over spatial domains)





[Example 3] Whole mouse spatial domains

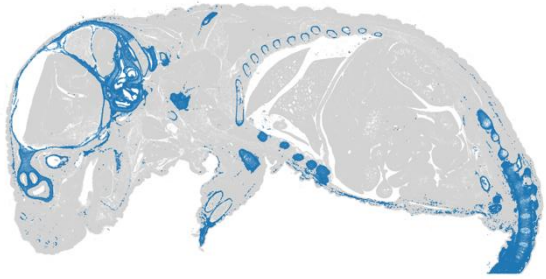


Domains highlights

Intestine

Lung lobes

Liver



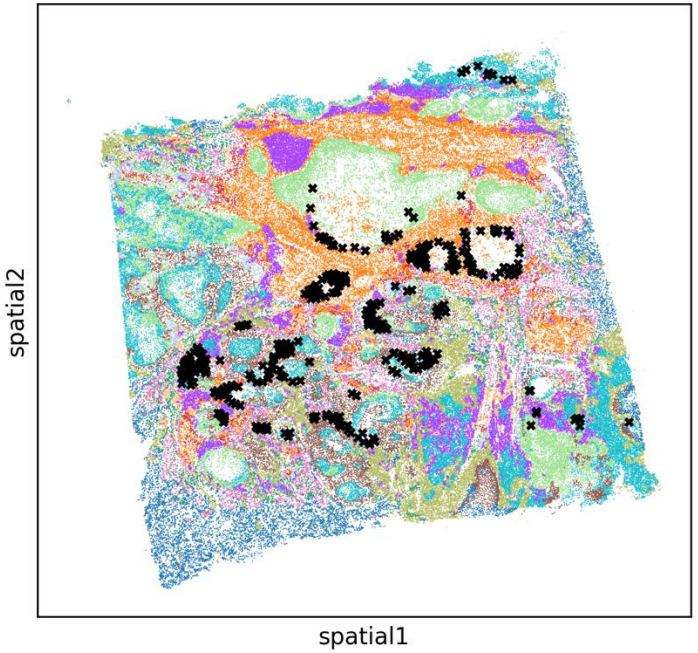
Bones and costal cartilage

[Example 4] Spatial domains of MGCs in head & neck cancer

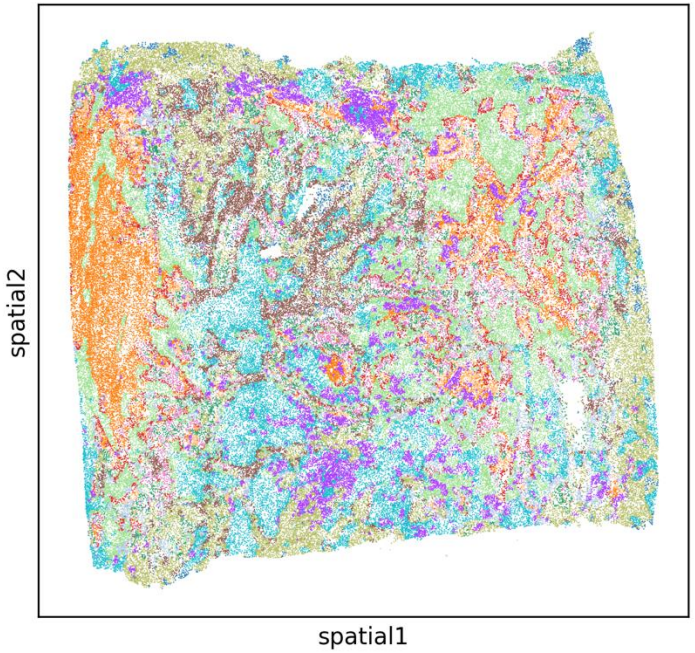
Usage on CosMx **Protein** Assays

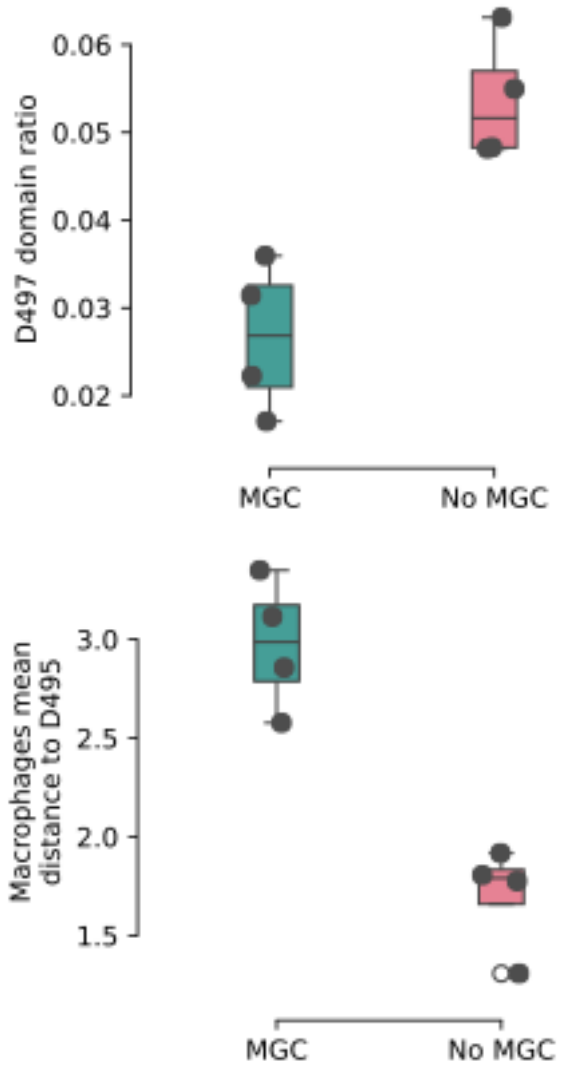
8 patients stratified based on the presence of multinucleated giant cells (MGCs), which are good prognosis

Novae domains (slide with MGCs)



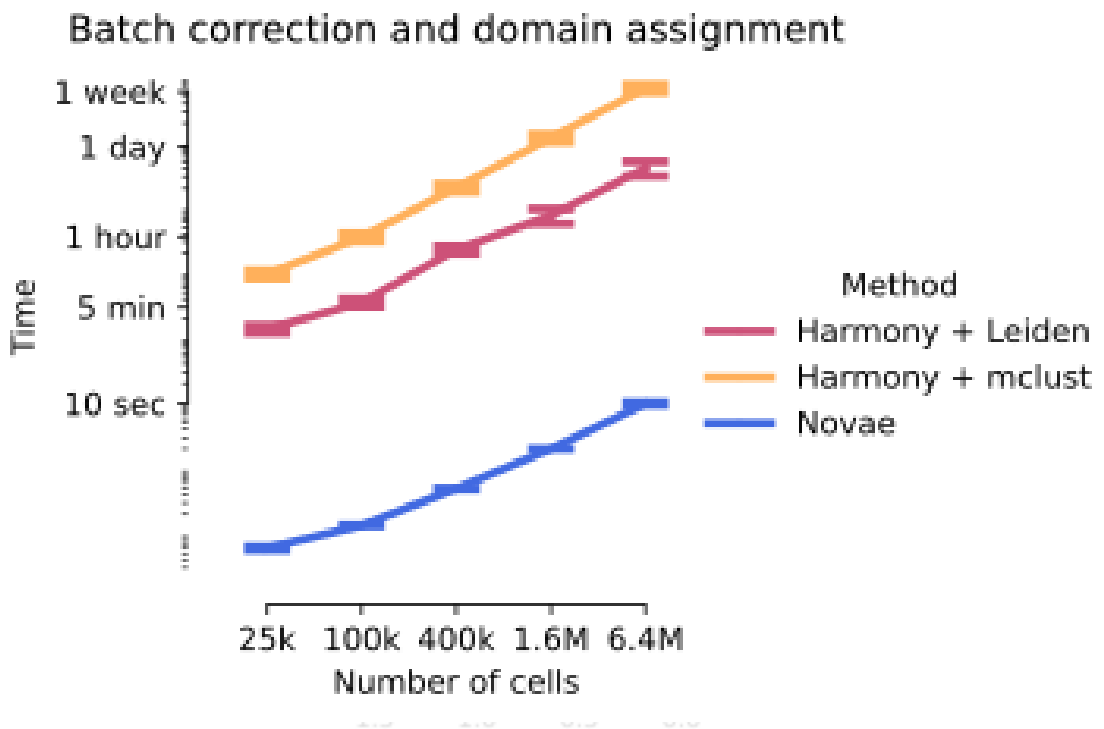
Novae domains (slide without MGCs)





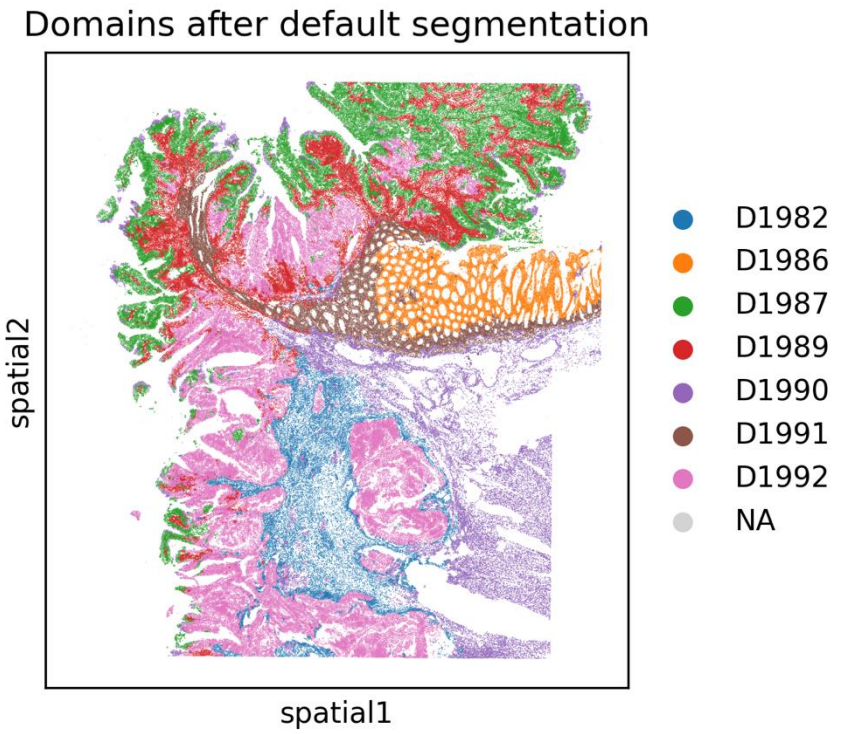
More biological details in the upcoming official publication

Benchmark - Comparison to existing methods

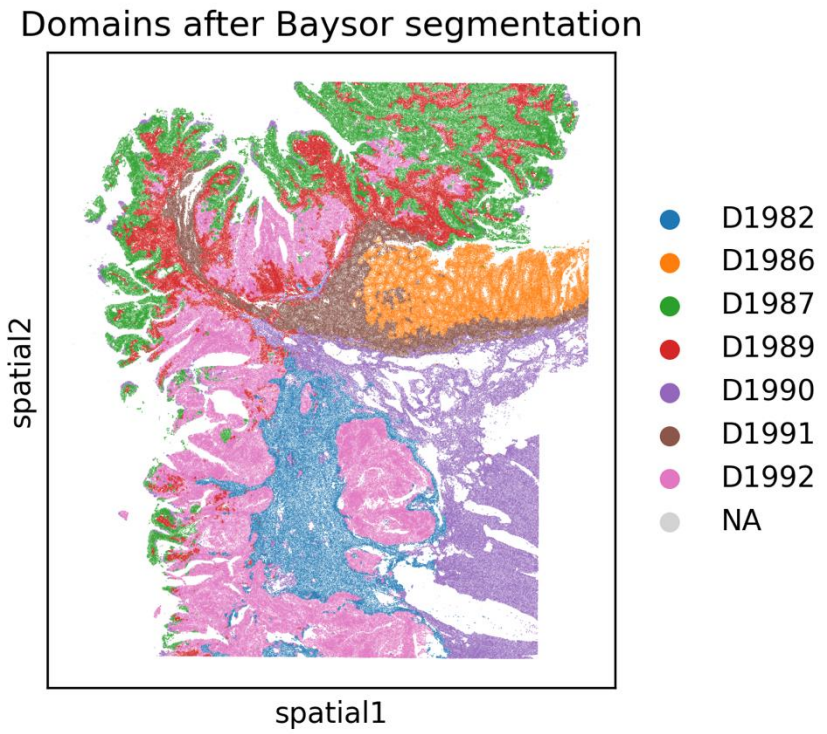


Robustness analysis: sensitivity to segmentation methods

Approach: run two segmentation methods, and show we can retrieve the same domains



(image-based segmentation)

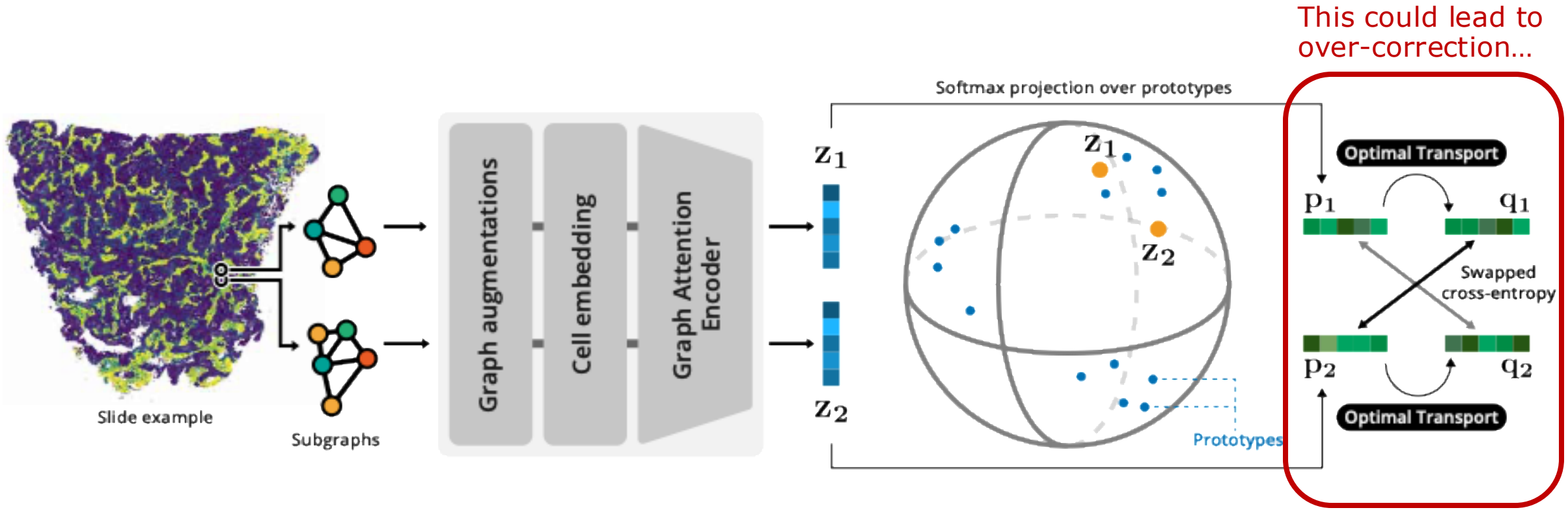


(transcript-based segmentation)

Baysor (Petukhov et al., Nat. Biotechnol.)

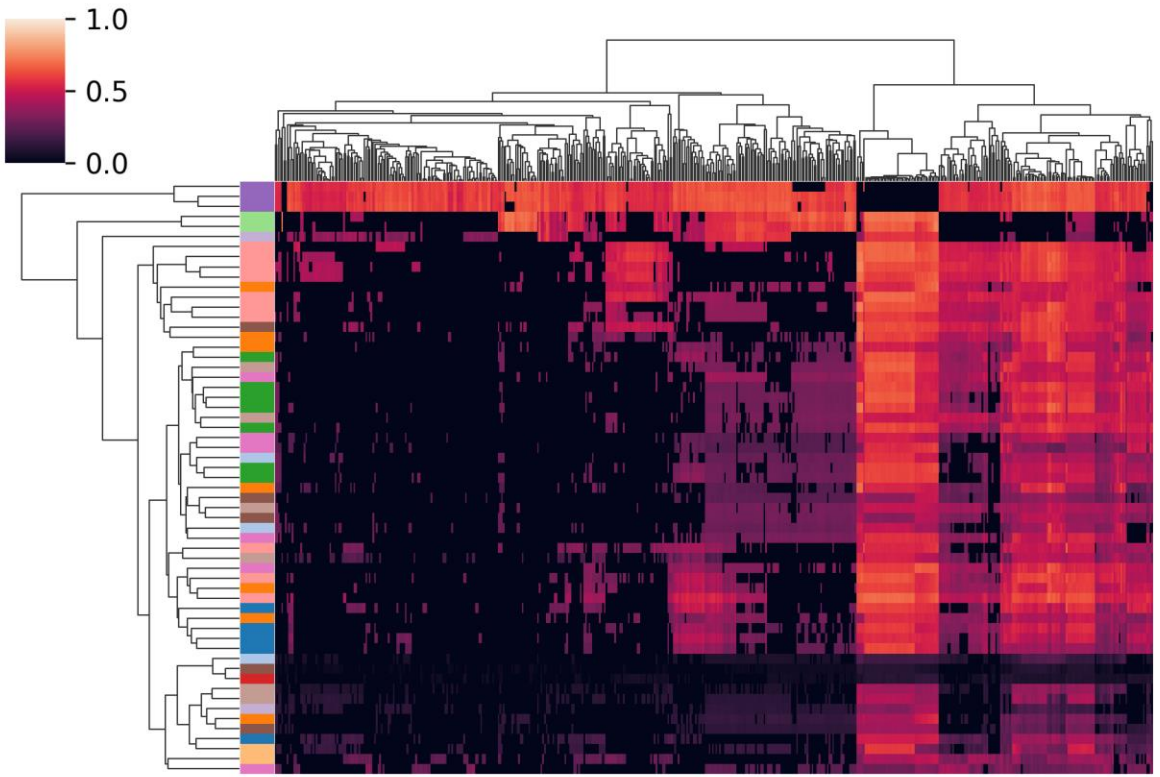
Robustness analysis 2: avoiding over-correction

Reminder: we use optimal transport to correct the batch effect

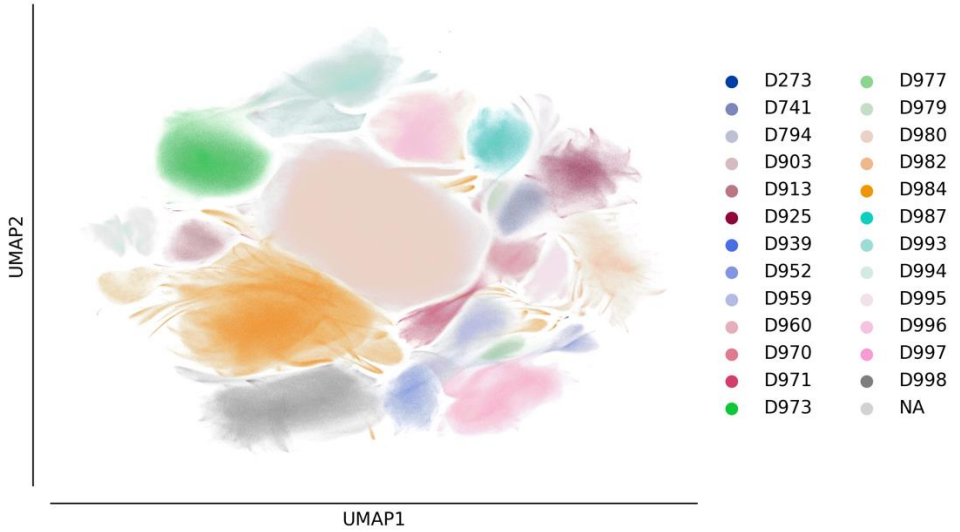


Robustness analysis 2: avoiding over-correction

We allow the prototypes not to be used in every slide

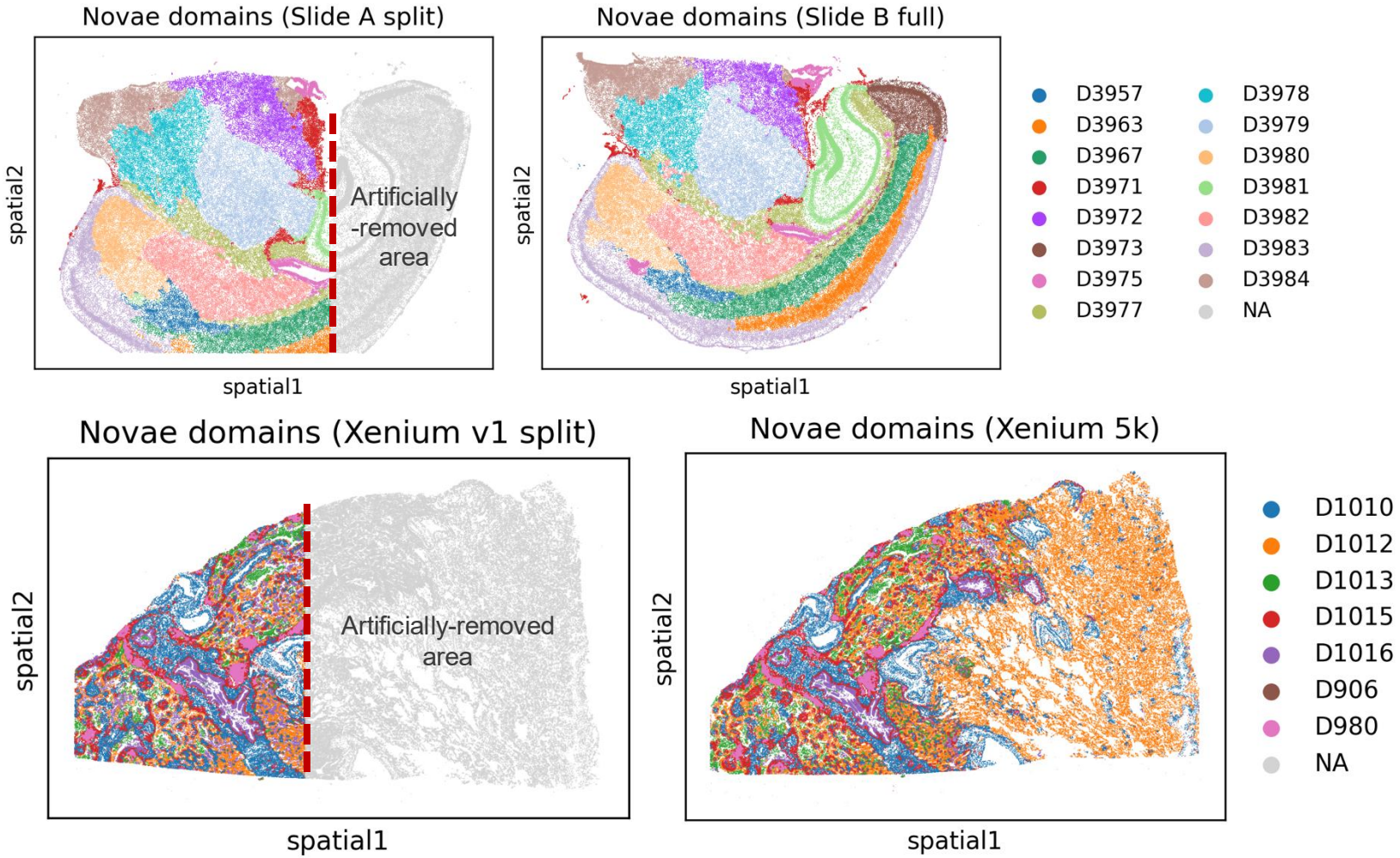


Novae domains (human tissues)



Robustness analysis 2: avoiding over-correction

Removing some parts of the slides to see if it affects the spatial domains assignment



Example usage (API)

Zero-shot assignment of spatial domains on some slides from our dataset

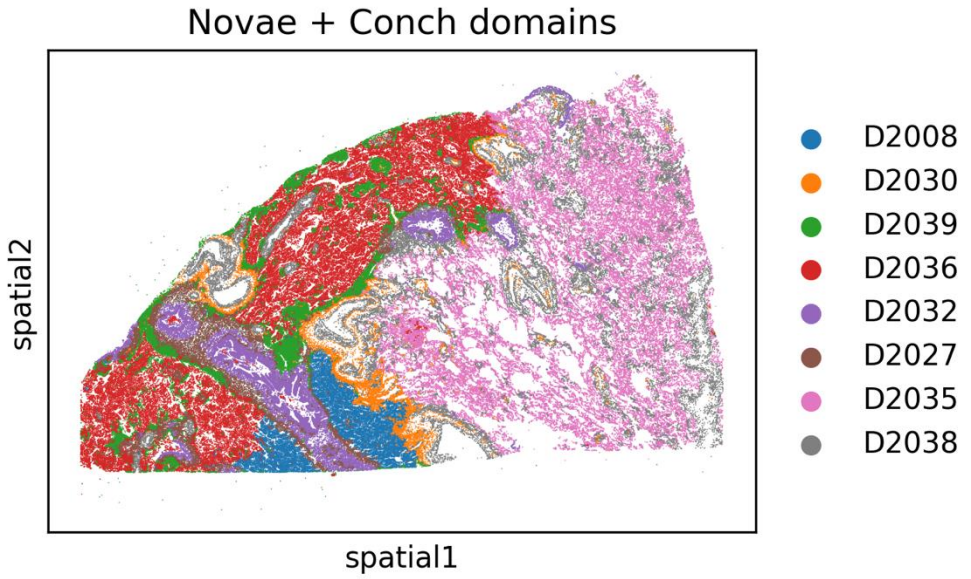
```python
import novae

adatas = novae.load_dataset(tissue="colon", species="human", pattern=".*P2.*")

# Compute delaunay graph
novae.spatial_neighbors(adata, radius=80)

# Load a pretrained model
model = novae.Novae.from_pretrained("MICS-Lab/novae-human-0")

# Compute the spatial domains embeddings
model.compute_representations(adata, zero_shot=True)

# Assign the domains
model.assign_domains(adata)
```

Downloaded automatically from Hugging Face

Combining ST information with H&E data

Use case: aligning modalities, then using an H&E foundation model to get embeddings per subgraph, and train a new Novae model in multimodal mode

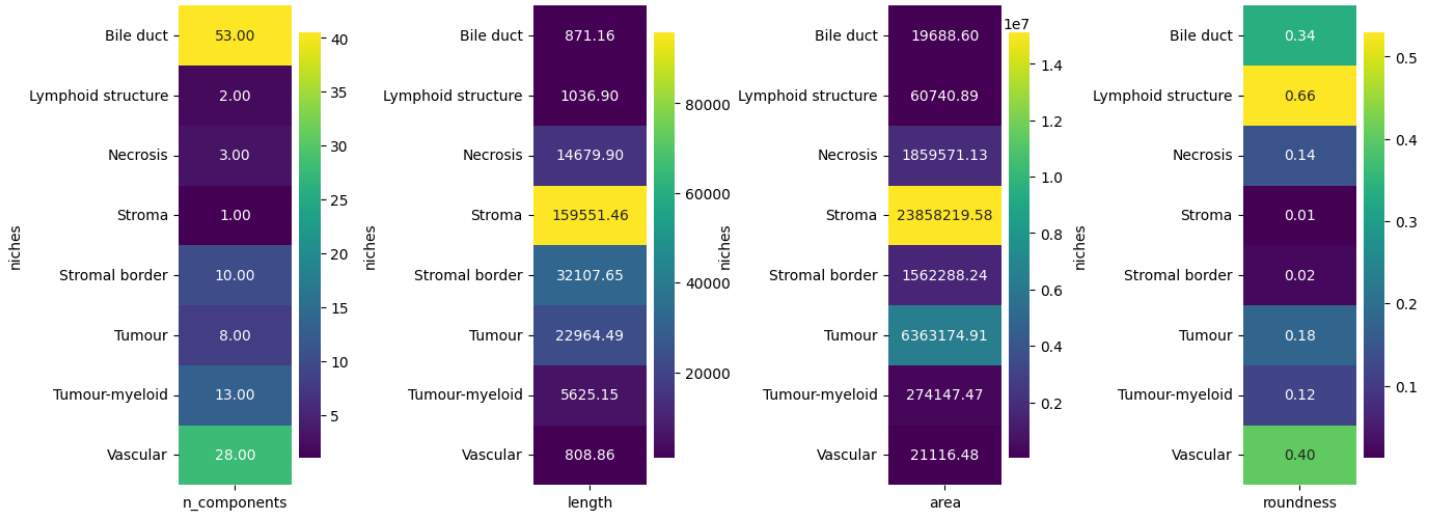
```python
import sopa

import novae

sdata = sopa.io.xenium("/path/to/data", cells_table=True, cells_boundaries=True)

# Run a H&E foundation model to get an embedding per subgraph
novae.compute_histo_embeddings(sdata, model="conch", device="cuda")
novae.compute_histo_pca(sdata)

# Get the AnnData object from the SpatialData object
adata = sdata["table"]

# Train a multimodal Novae model
novae.spatial_neighbors(adata, radius=80)

model.fit(adata, accelerator="cuda", num_workers=4, lr=2e-4, max_epochs=40)

model.compute_representations(adata, accelerator="cuda", num_workers=4)
model.assign_domains(adata, level=8)
```

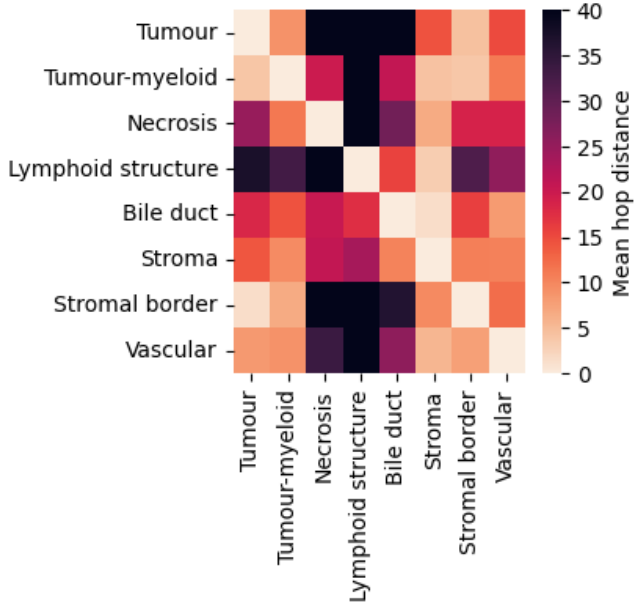


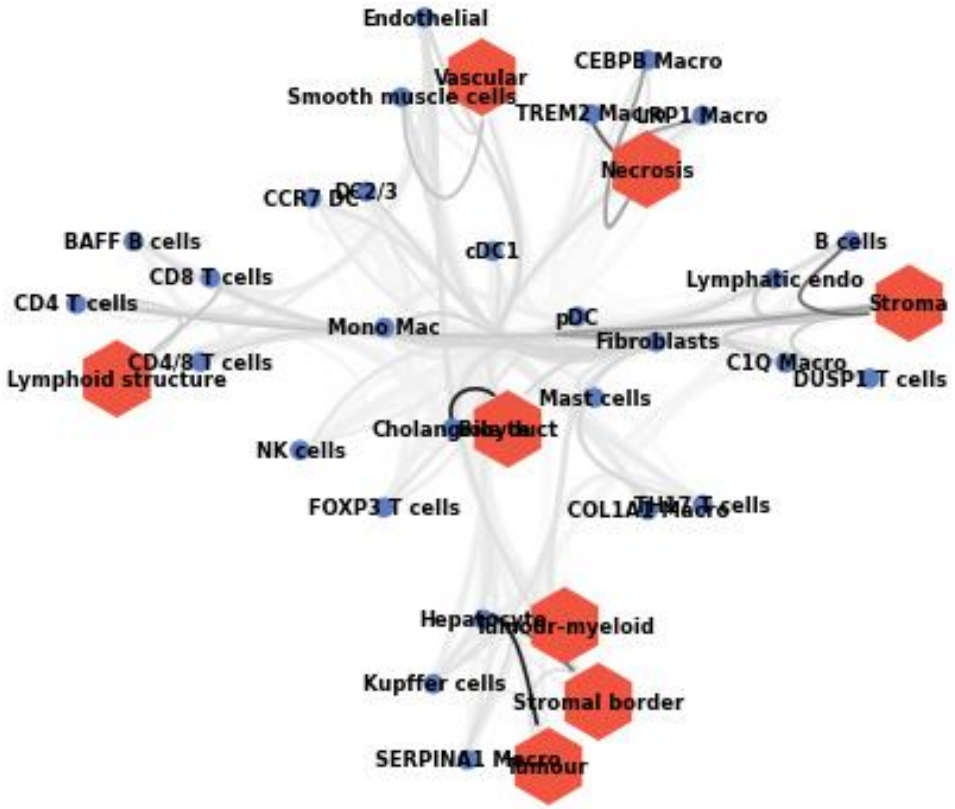
SpatialData + sopa + novae

CONCH - A visual-language foundation model for computational pathology. (Lu et al., Nature Medicine)

Go back to Sopa to compute some spatial statistics







Or other scverse tools...

squidpy

Again, feel free to try it!

https://mics-lab.github.io/novae/

https://github.com/MICS-Lab/novae

Acknowledgments

CentraleSupélec

Paul-Henry Cournède

Hakim Benkirane

Karmen Rabar

Stergios Christodoulidis

Institut Gustave Roussy

Fabrice André

Kevin Mulder

Florent Ginhoux

Charles-Antoine Dutertre

Isabelle Pic

Margaux Gardet

Joana Mourato Ribeiro

Institut Cochin *[External]*

Nadège Bercovici

EMBL / DKFZ *[External]*

Luca Marconato

Wouter-Michiel Vierdag

Helmholtz Munich *[External]*

Giovanni Palla