Topological Data Analysis: extracting insights from the “shape” of data

55e Journées de statistique de la SFdS
27 May 2024

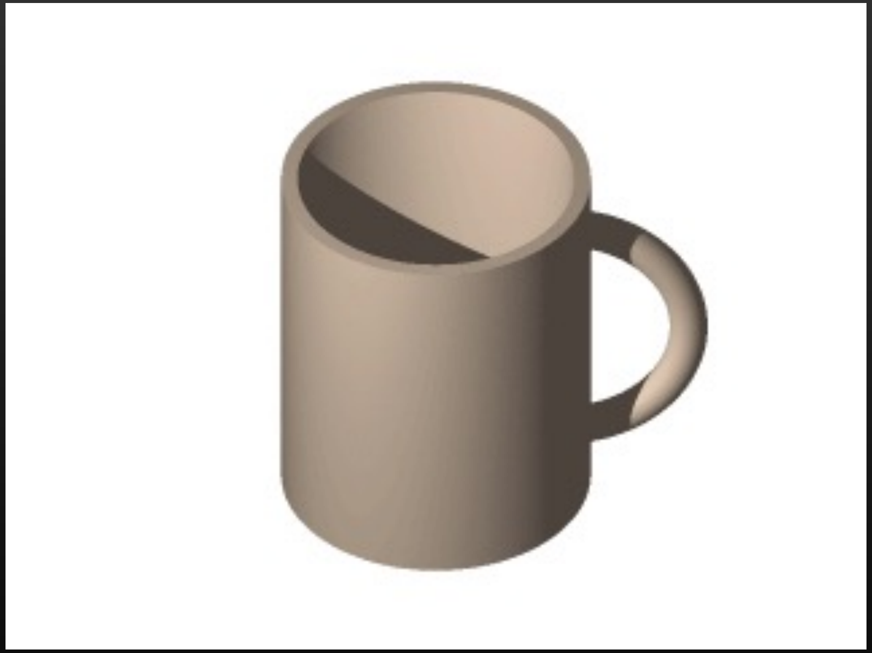
EPFL

What is topology?


shape
deformation
connectivity
continuous
path
cavity
geometry
invariants
connected
donut
classification
simplex
open
topology
proximity
mug
homology
equivalence
algebra
closed
continuity
complex

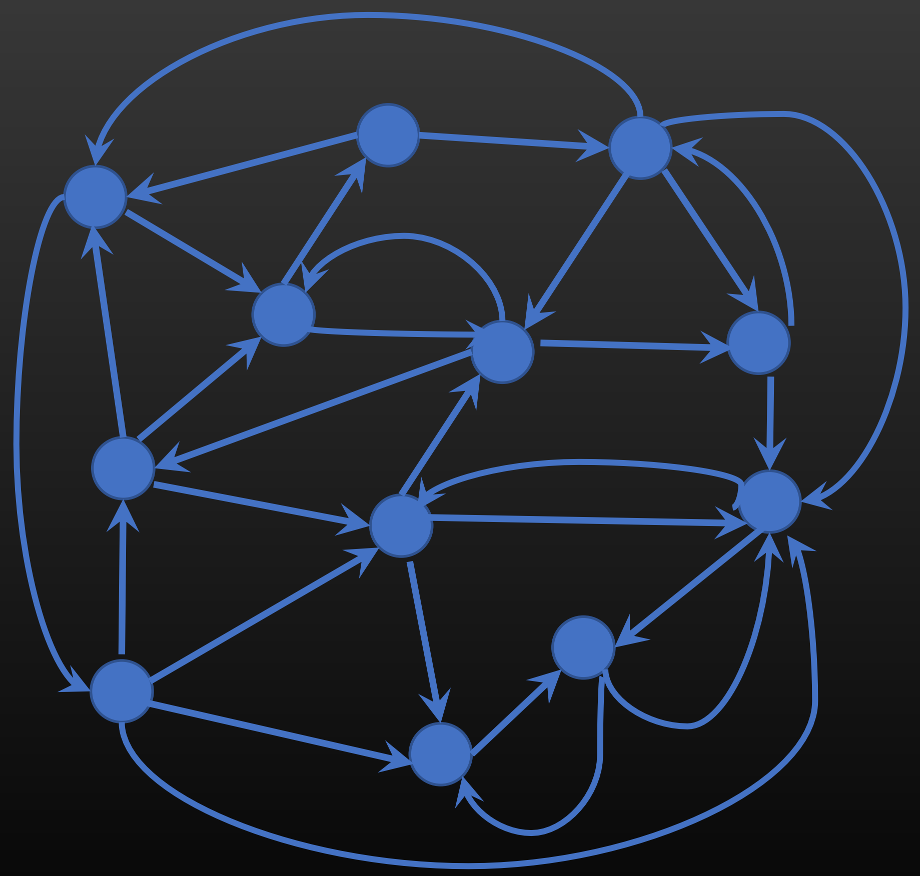
Topology is…

- the mathematics of shape;



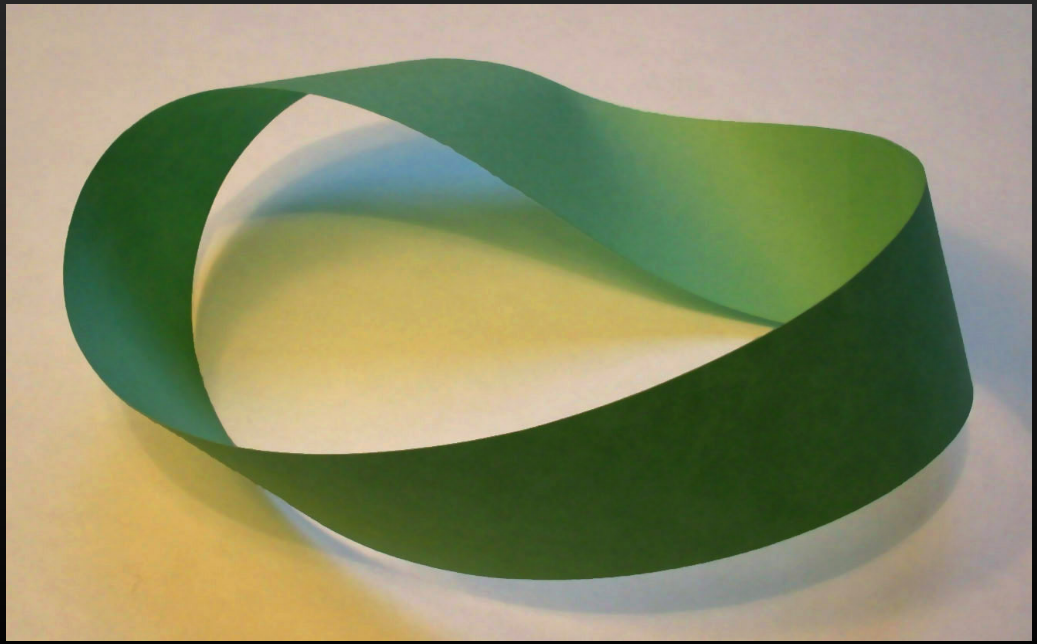
Topology is…

- the mathematics of shape;
- the mathematics of connectivity;



Topology is…

- the mathematics of shape;
- the mathematics of connectivity;
- the mathematics of emergence of global structure from local constraints.



Application to Data Science

The shape of a data set,

described by a topological signature

encoding its multi-scale structure,

can reveal important relations among the data points,

with the help of machine learning.

Topological Data Analysis (TDA)

Topological analytical tools

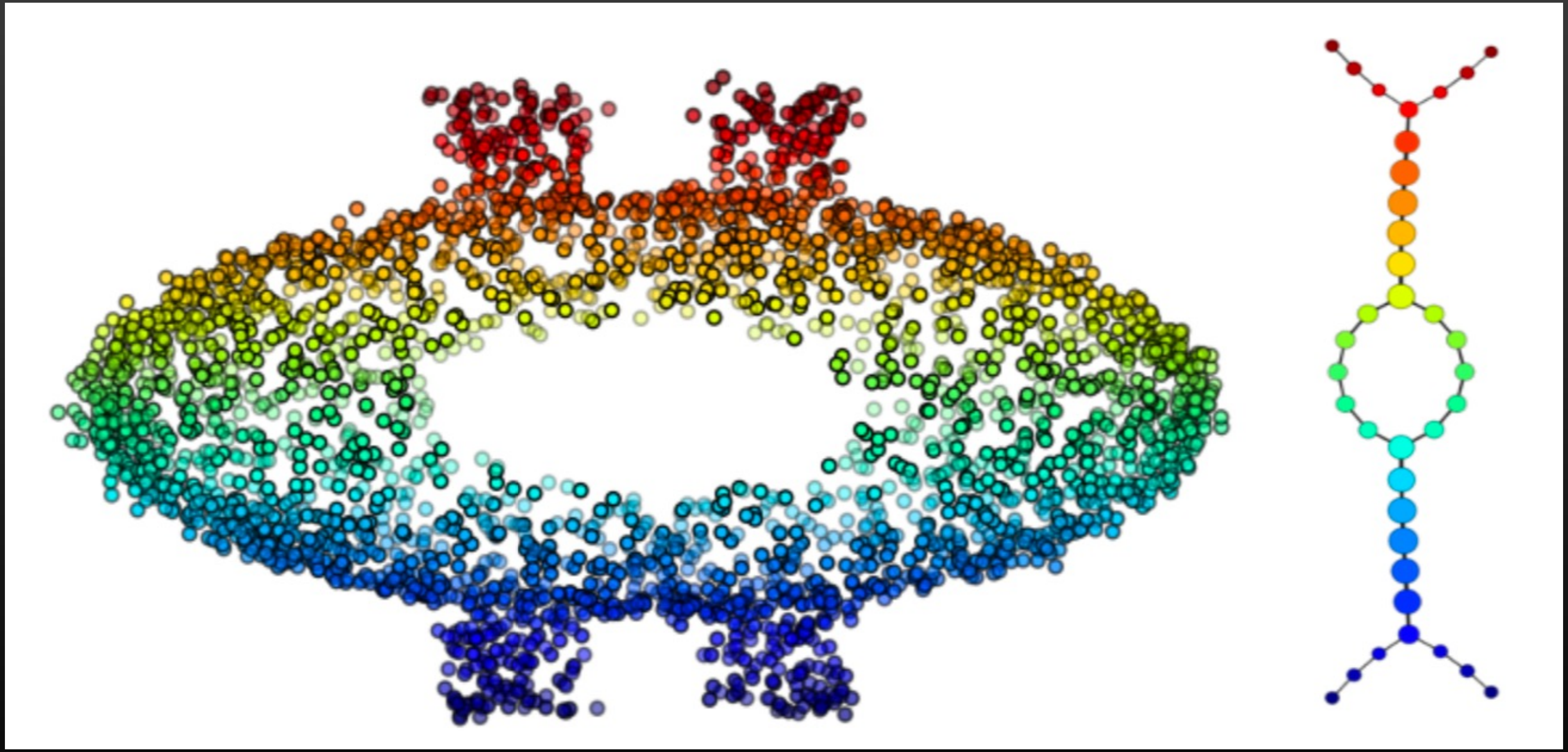
Method	Appropriate data types
Mapper	Clinical data, metabolomics, genomics, etc.
Two-tier Mapper	Gene expression data, single-cell transcriptomics
Persistent homology	Connectivity data, high-dimensional point cloud data
Graph signal processing	Connectivity data + “signal”

Mapper

Overview

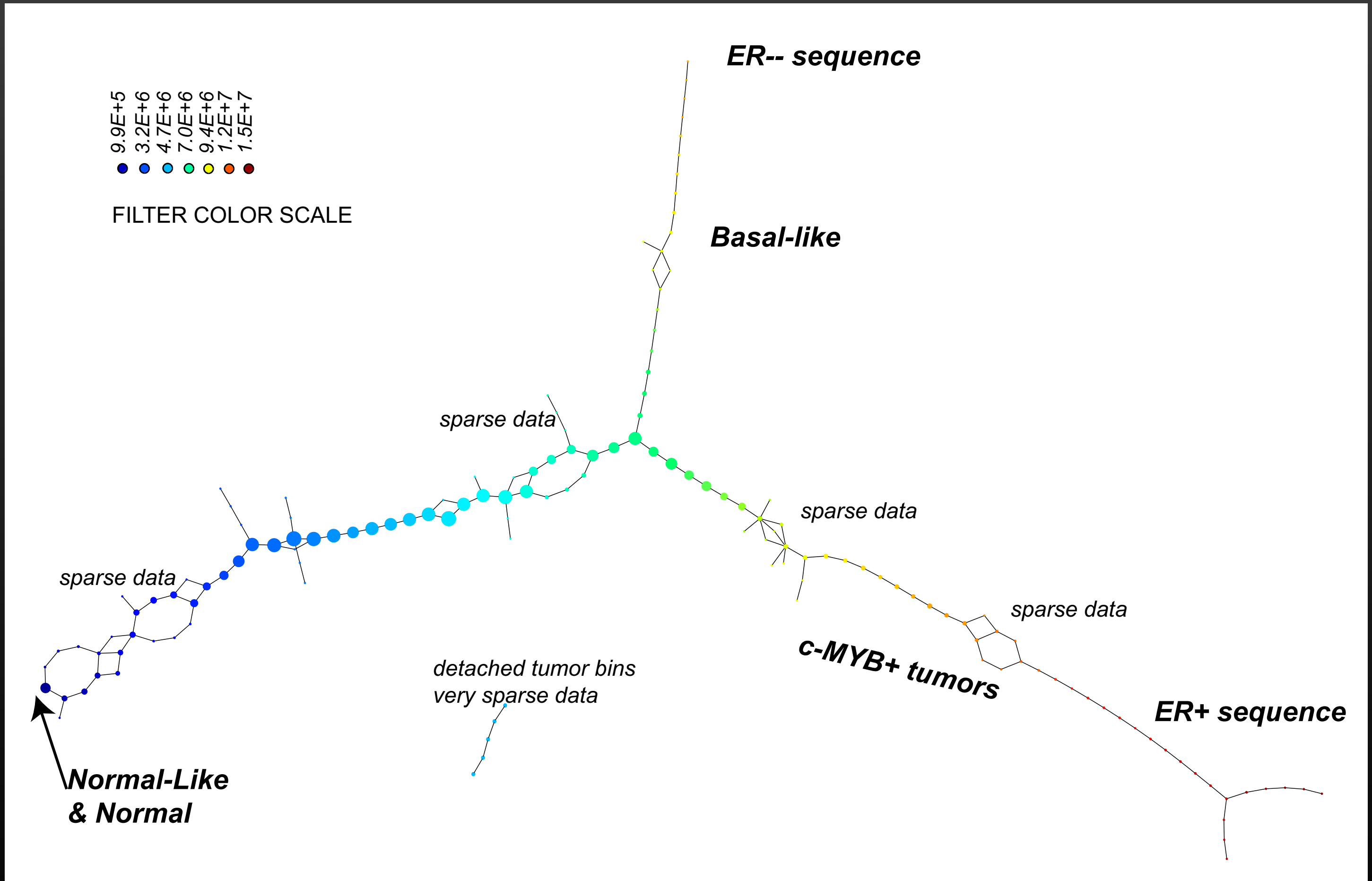
- (Mostly) unsupervised mutivariate pattern analysis of high-dimensional data, retains more information than PCA
- Produces a compressed visual representation of the data, providing a strong indication of where to look for meaningful clustering and encoding relations between clusters
- Numerous remarkably successful applications
- Input:
  - Data set X equipped with notion of “distance” between points
  - Real-valued “measurements” on X
  - Decomposition of the real line into overlapping subsets
  - Choice of clustering algorithm

Mapper output: synthetic data



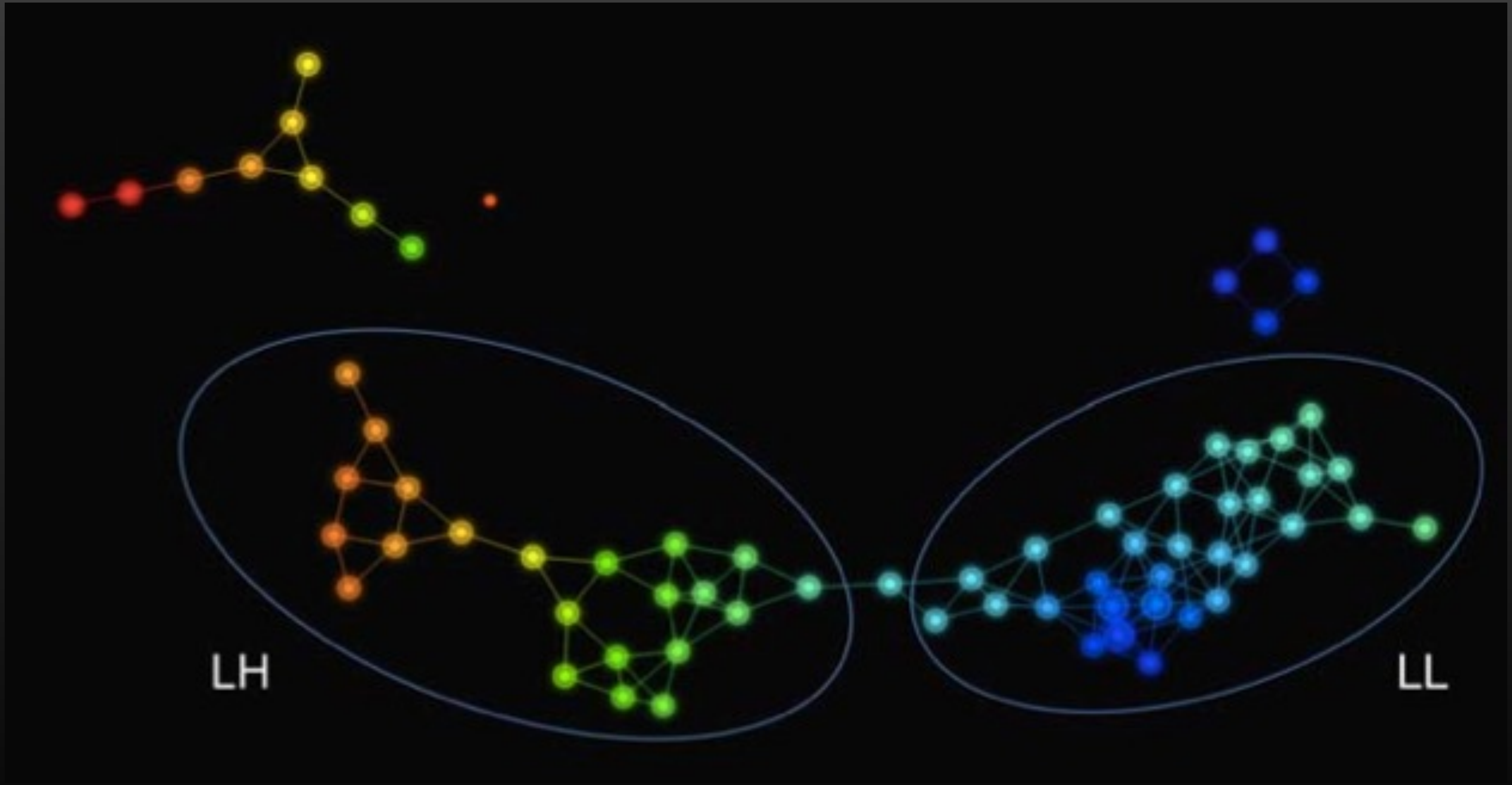
Masters thesis, F. Palma, 2018

Mapper output: gene expression data



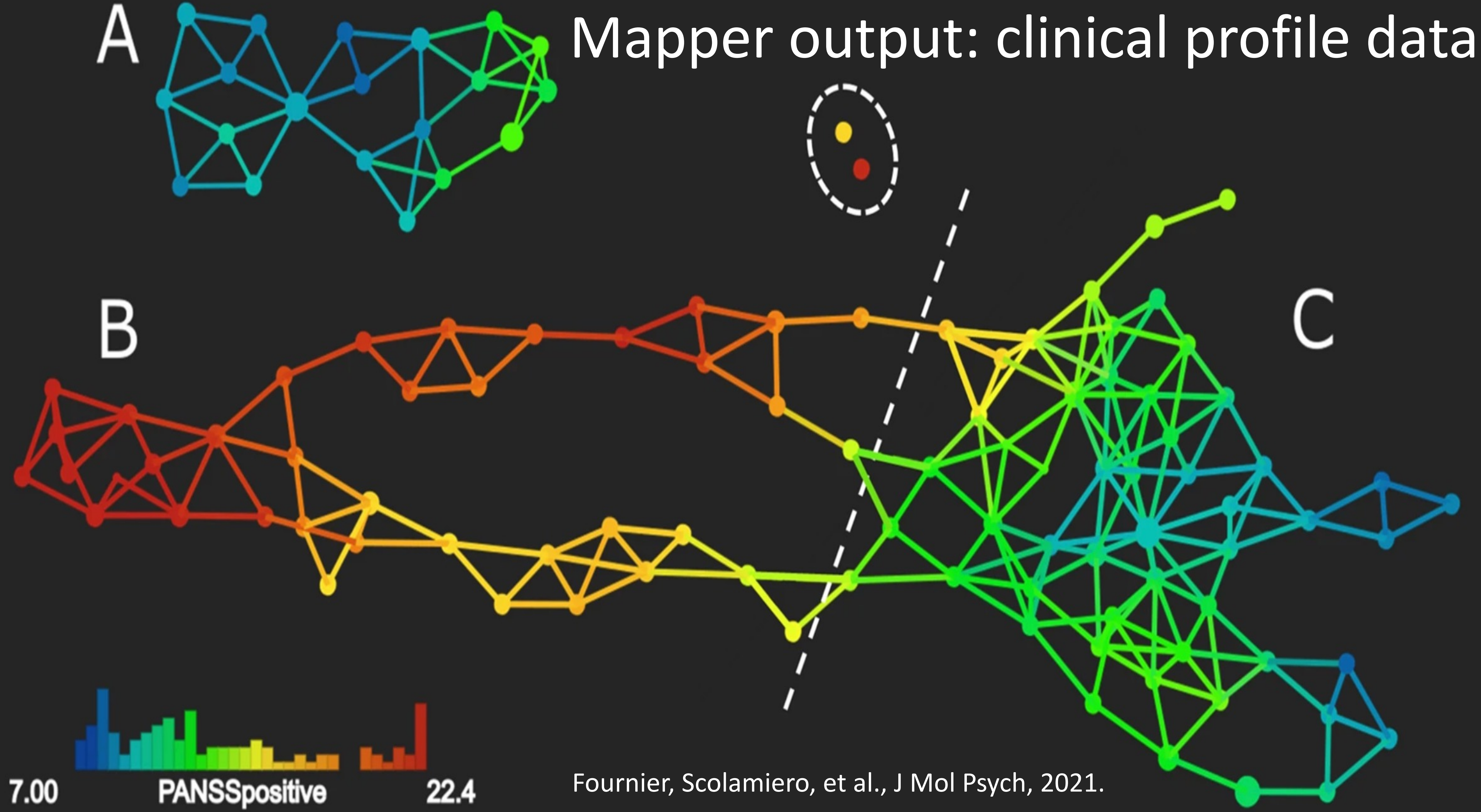
Levine, et al. PNAS 2011.

Mapper output: fMRI data

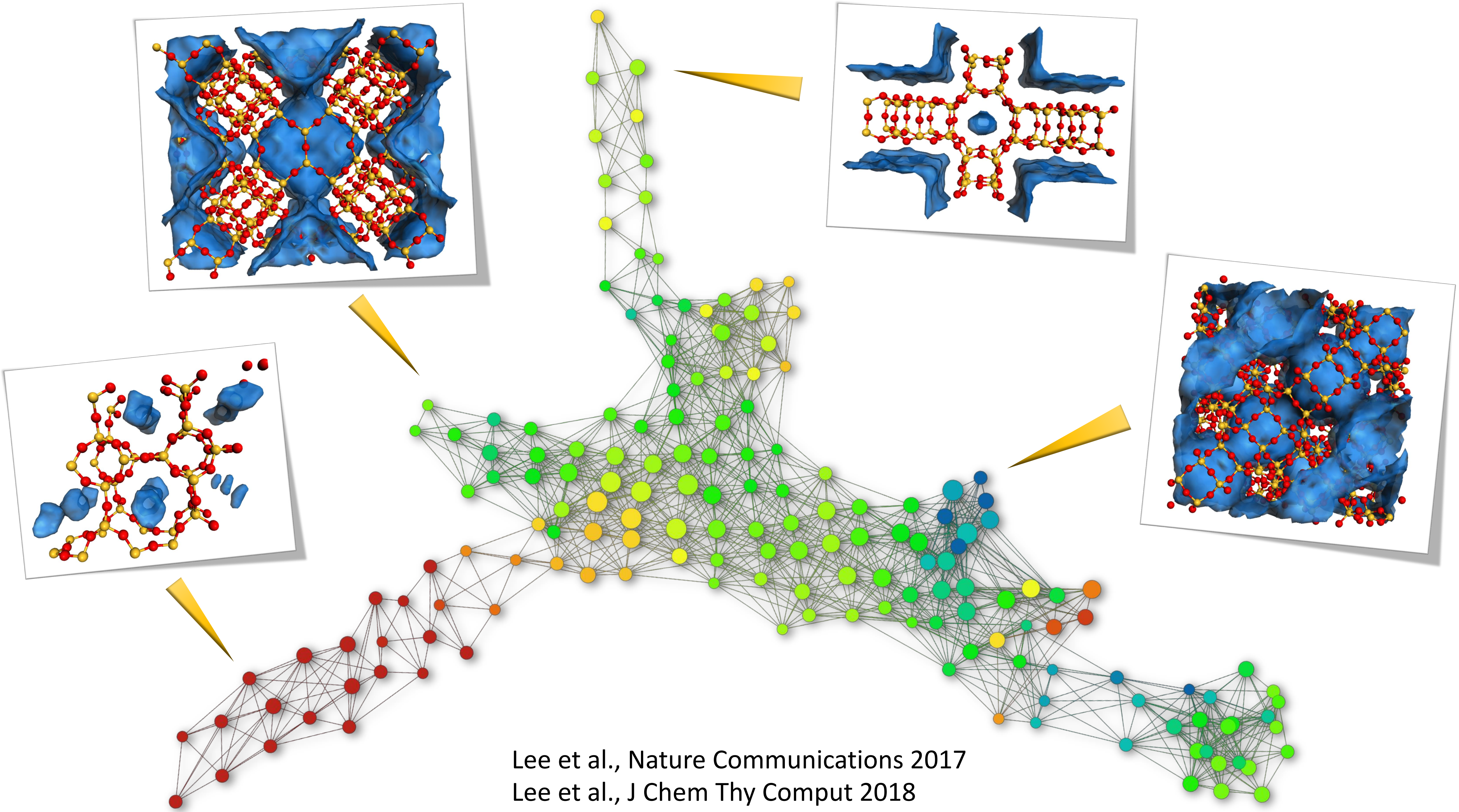


D. Romano et al, Human Brain Mapping 35:4904-4915 (2014)

Mapper output: clinical profile data



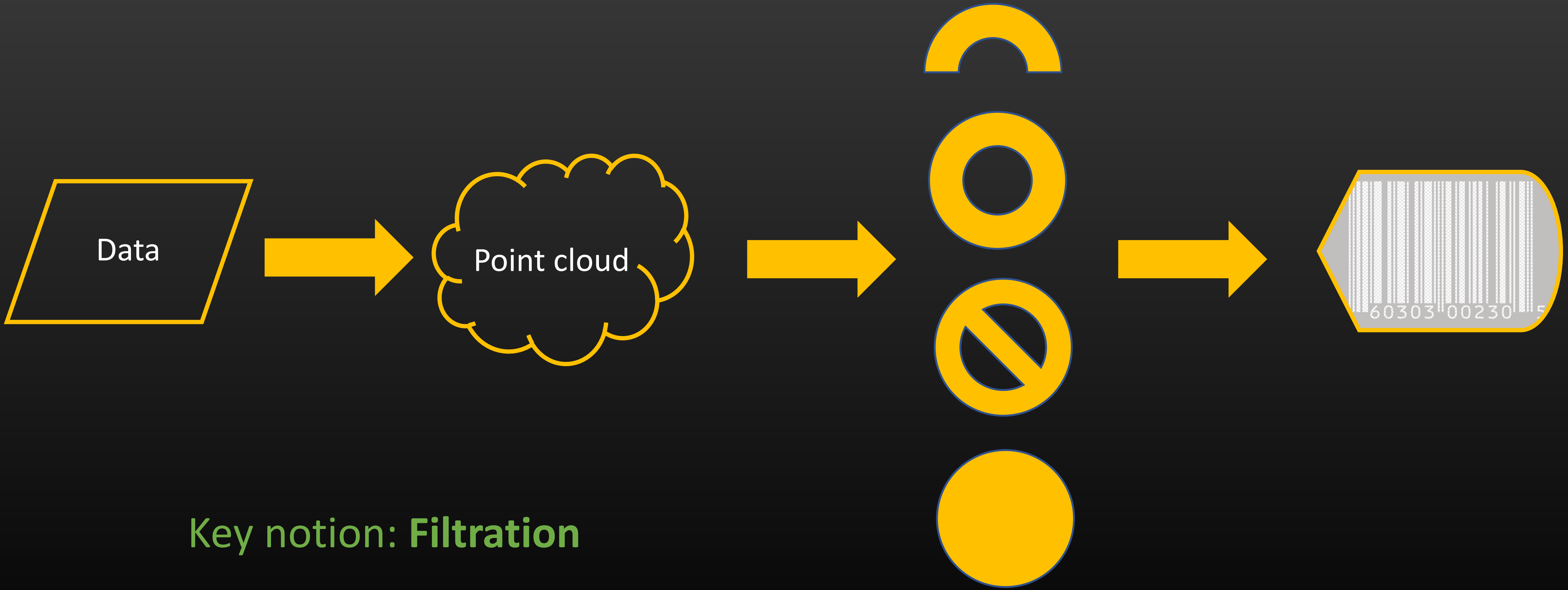
Fournier, Scolamiero, et al., J Mol Psych, 2021.



Lee et al., Nature Communications 2017
Lee et al., J Chem Thy Comput 2018

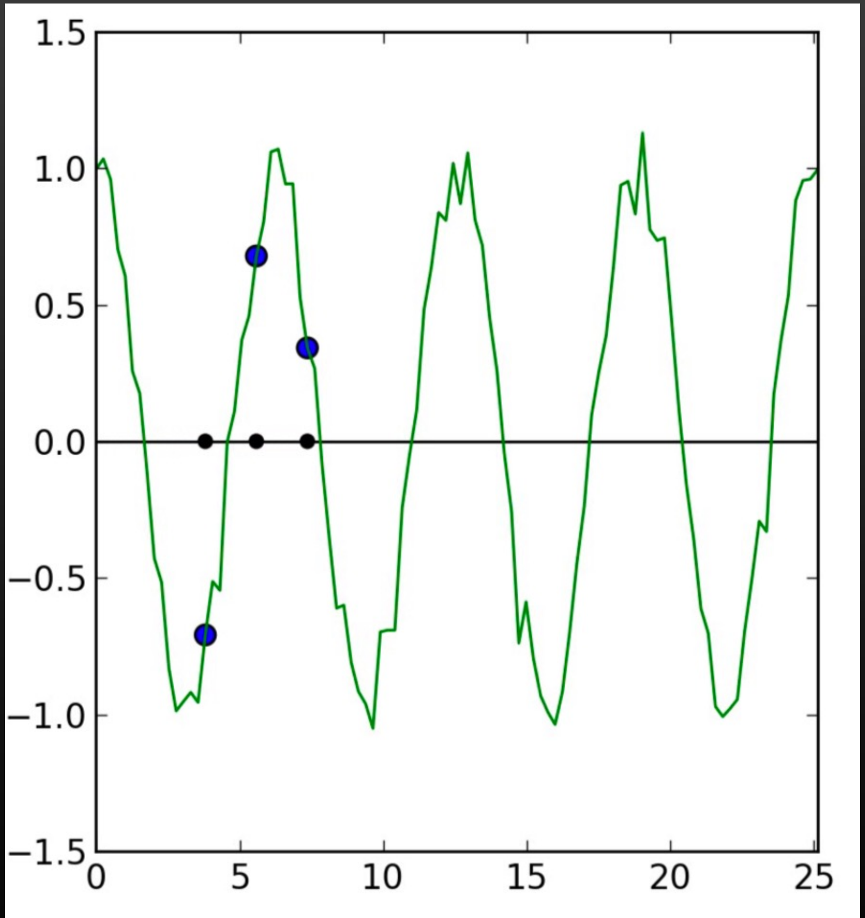
Persistent homology

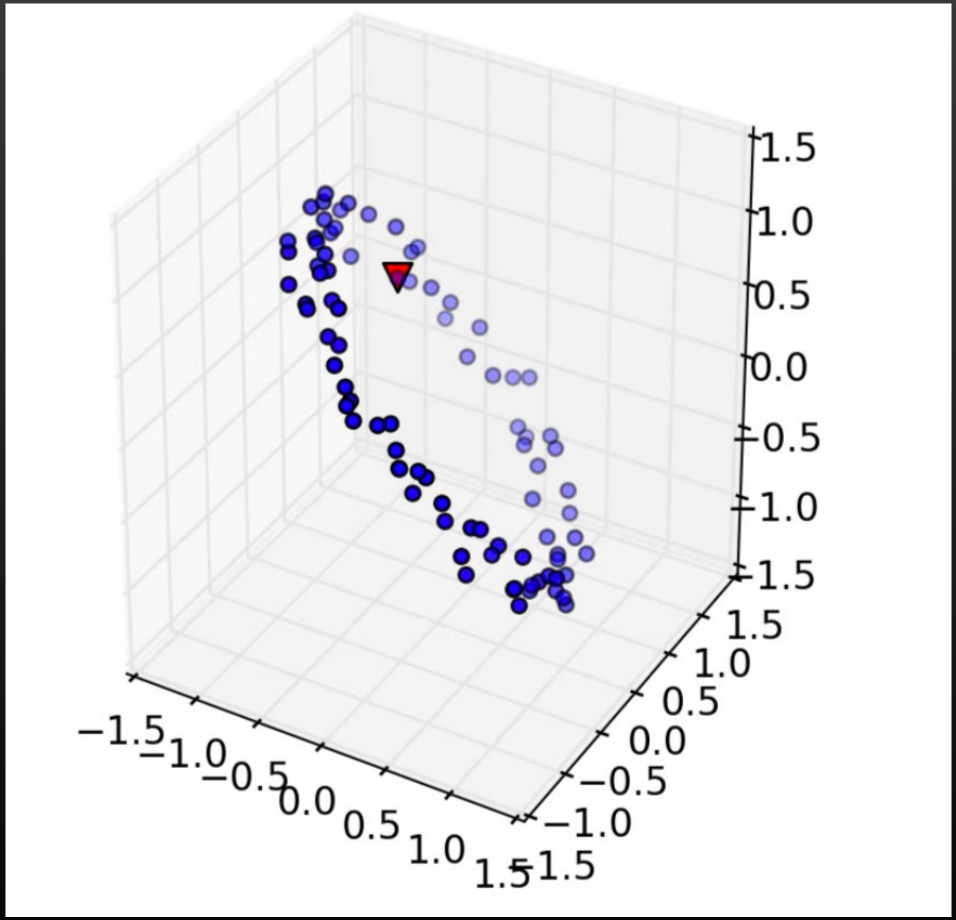
The basic persistence workflow



Key notion: **Filtration**

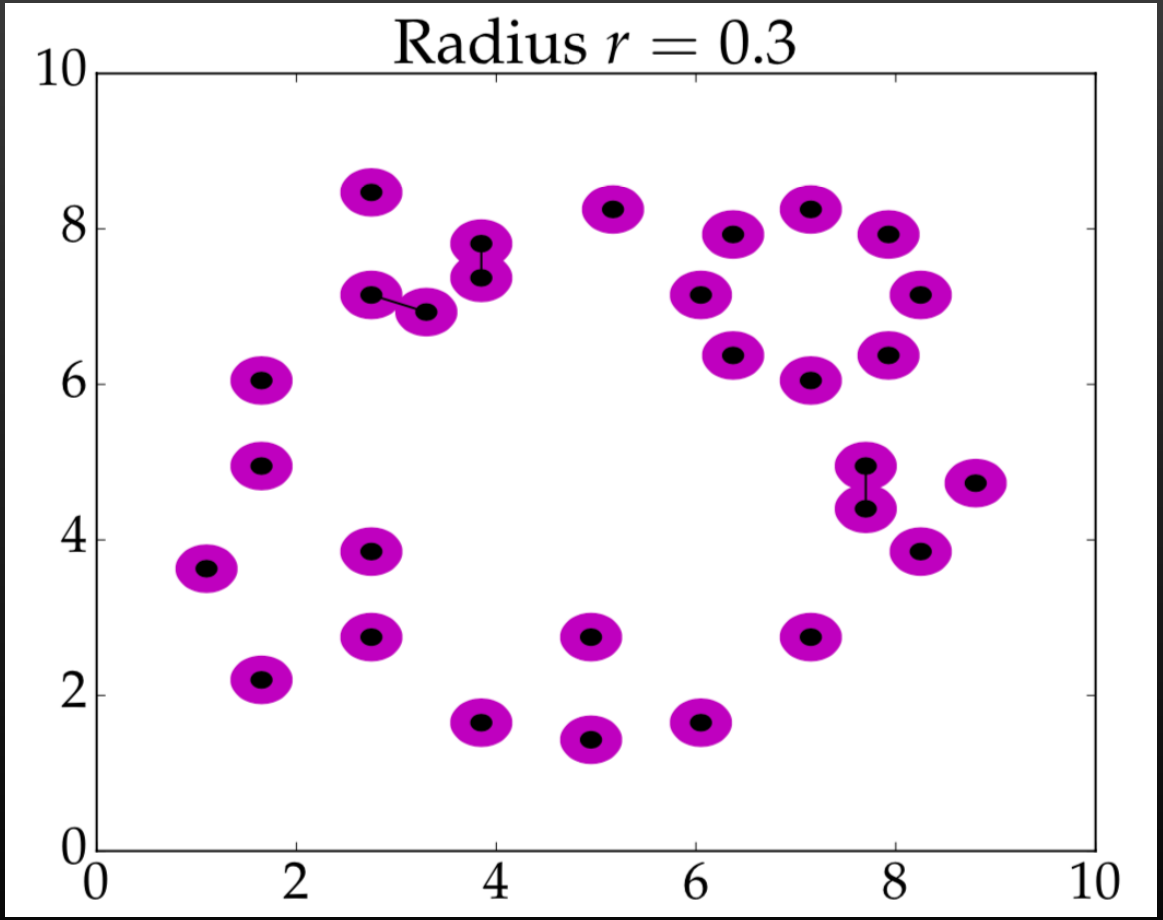
Step 1: Data to Point Cloud





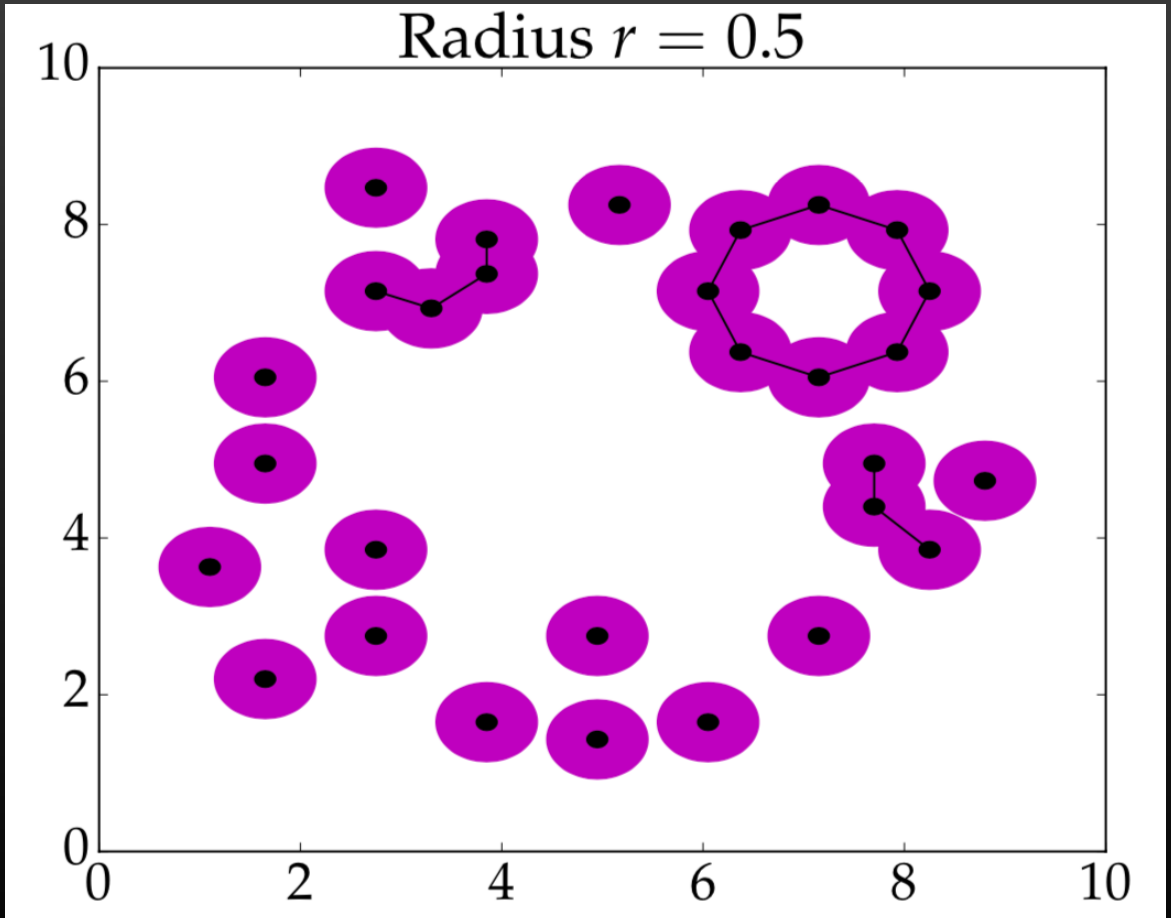
L. Munch, 2019.

Step 2: Point cloud to nested complexes



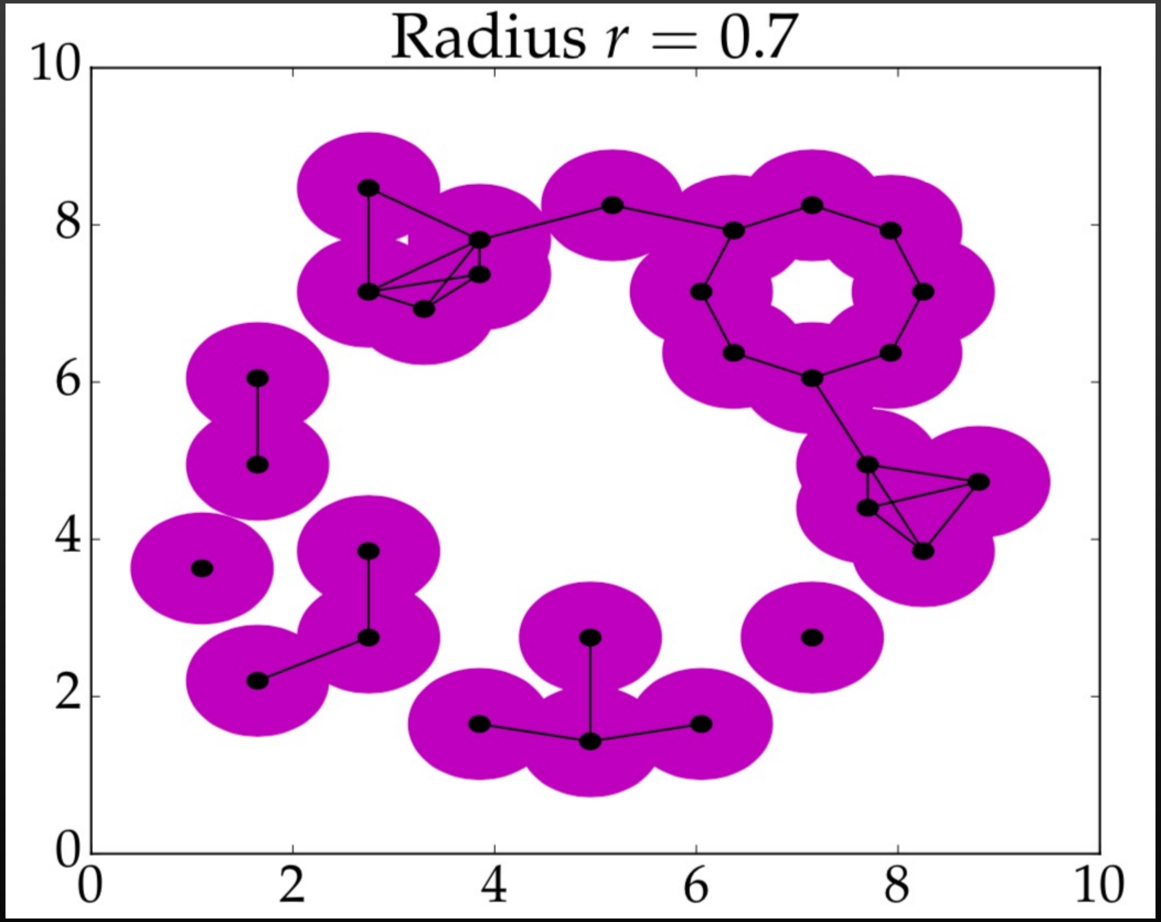
L. Munch, 2019.

Step 2: Point cloud to nested complexes



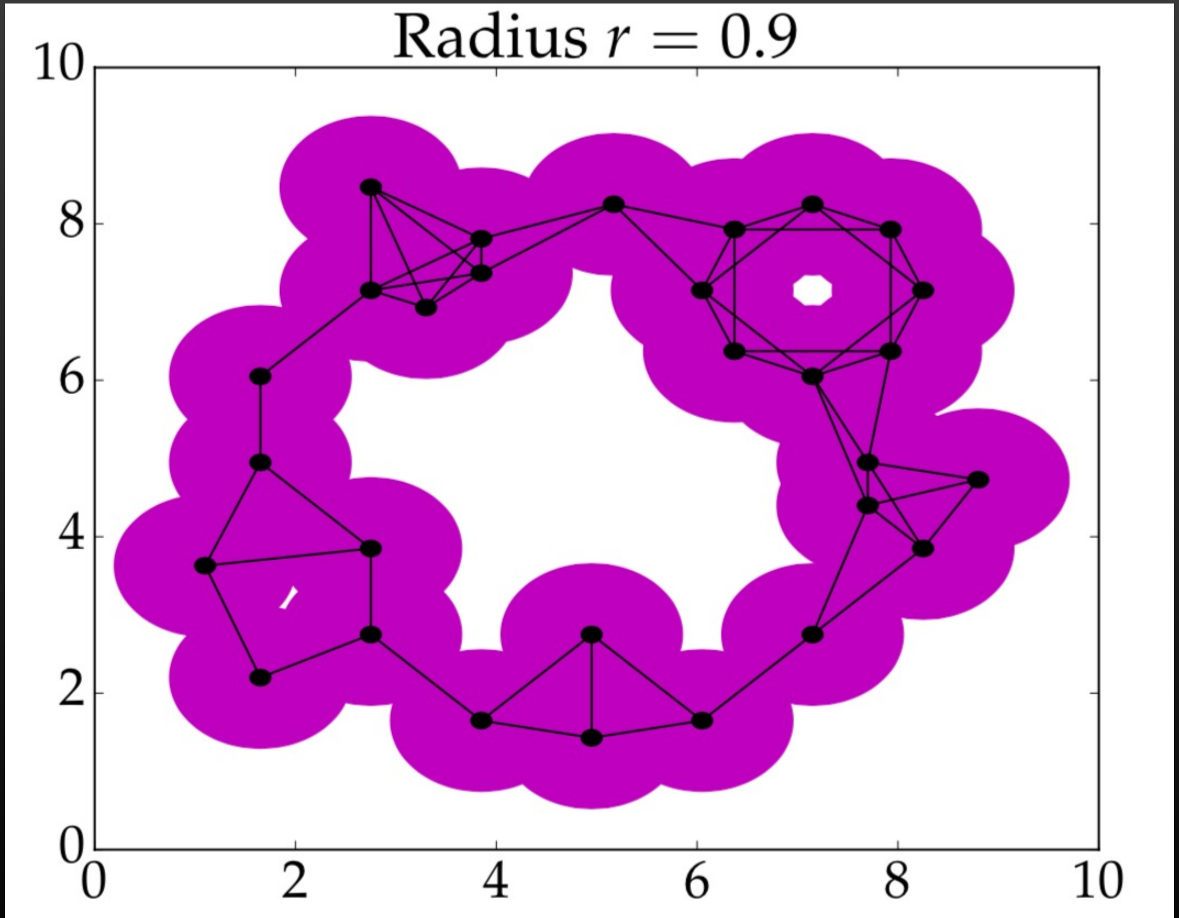
L. Munch, 2019.

Step 2: Point cloud to nested complexes



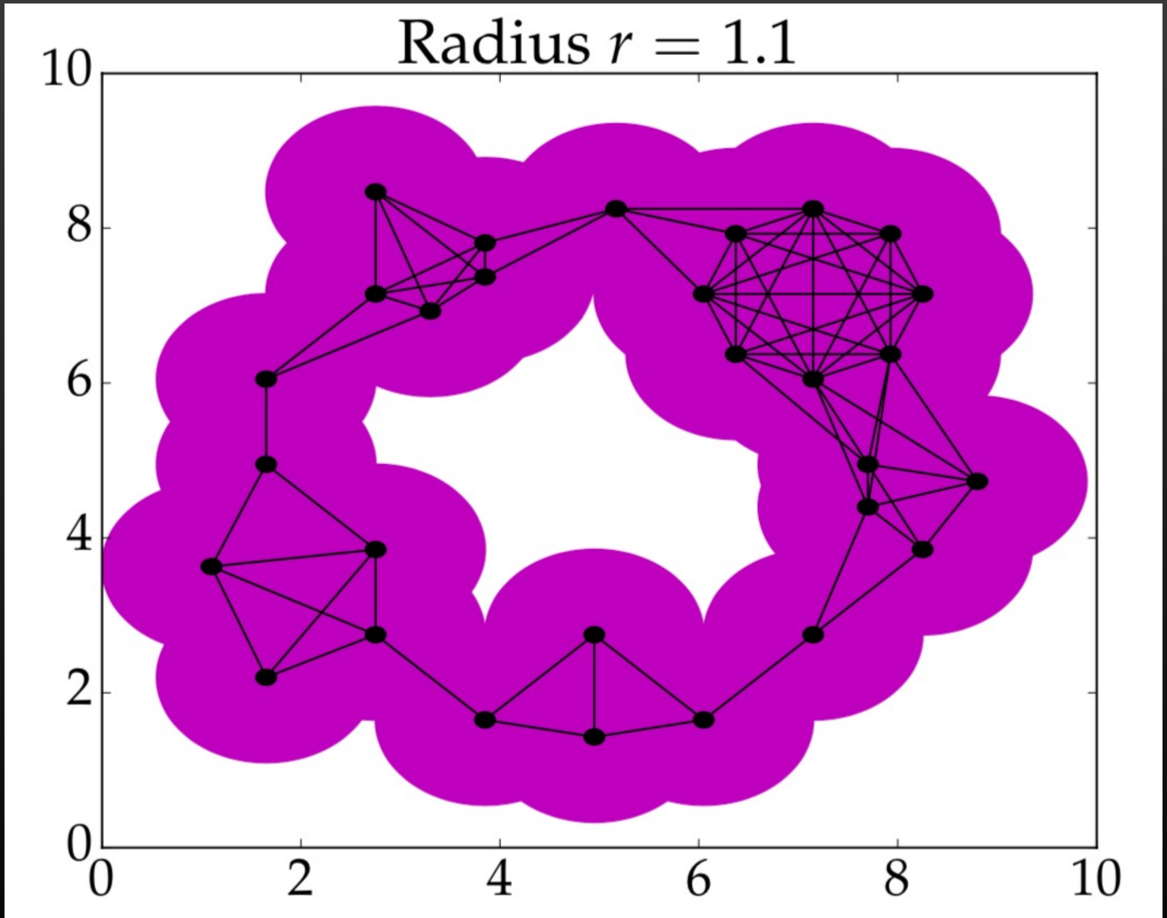
L. Munch, 2019.

Step 2: Point cloud to nested complexes



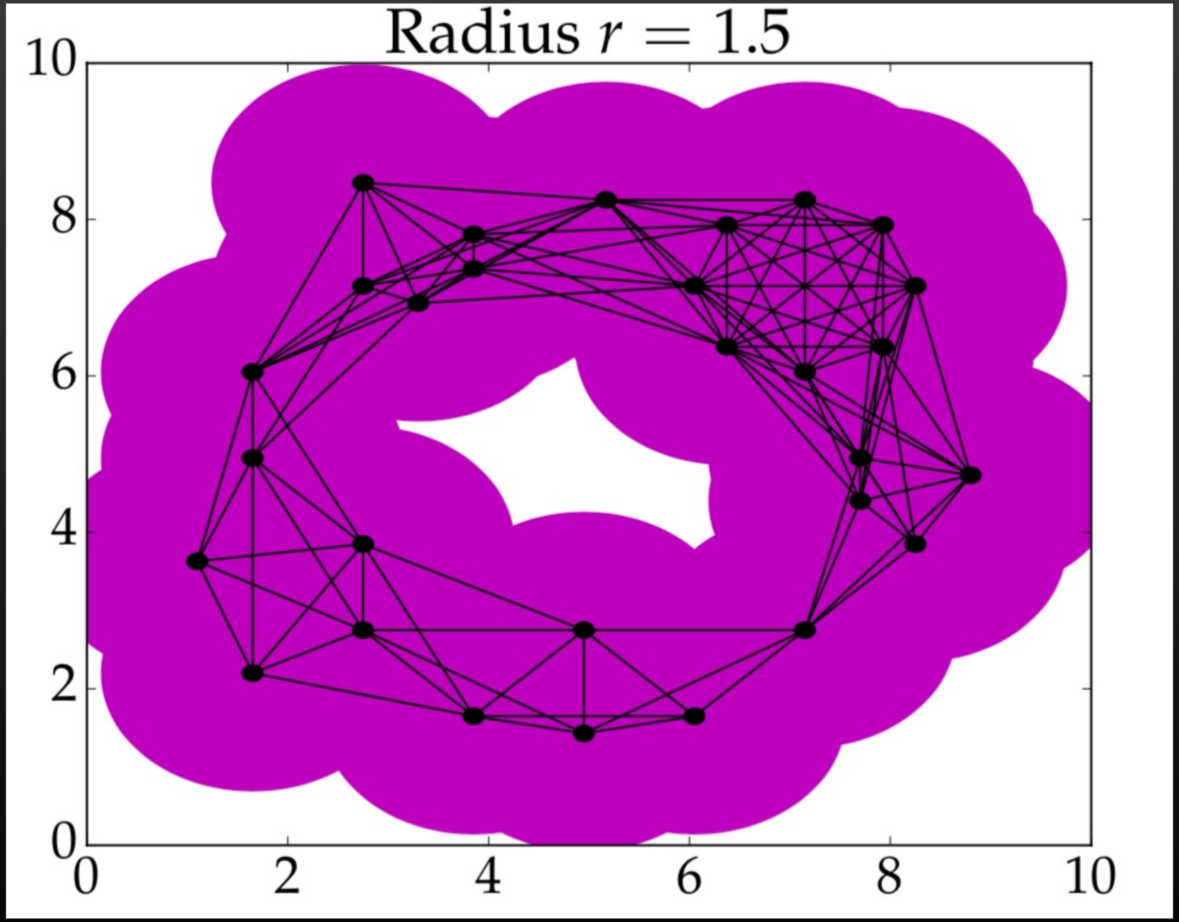
L. Munch, 2019.

Step 2: Point cloud to nested complexes



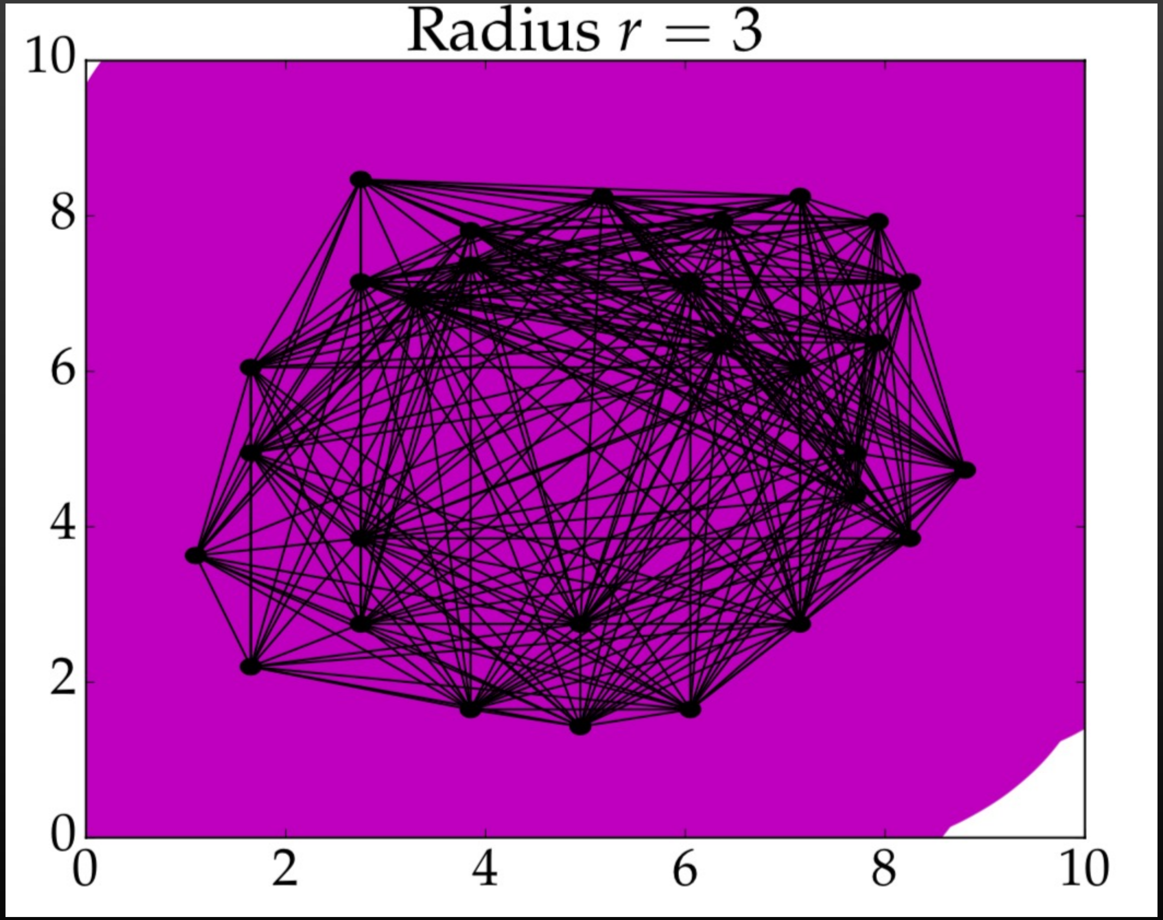
L. Munch, 2019.

Step 2: Point cloud to nested complexes



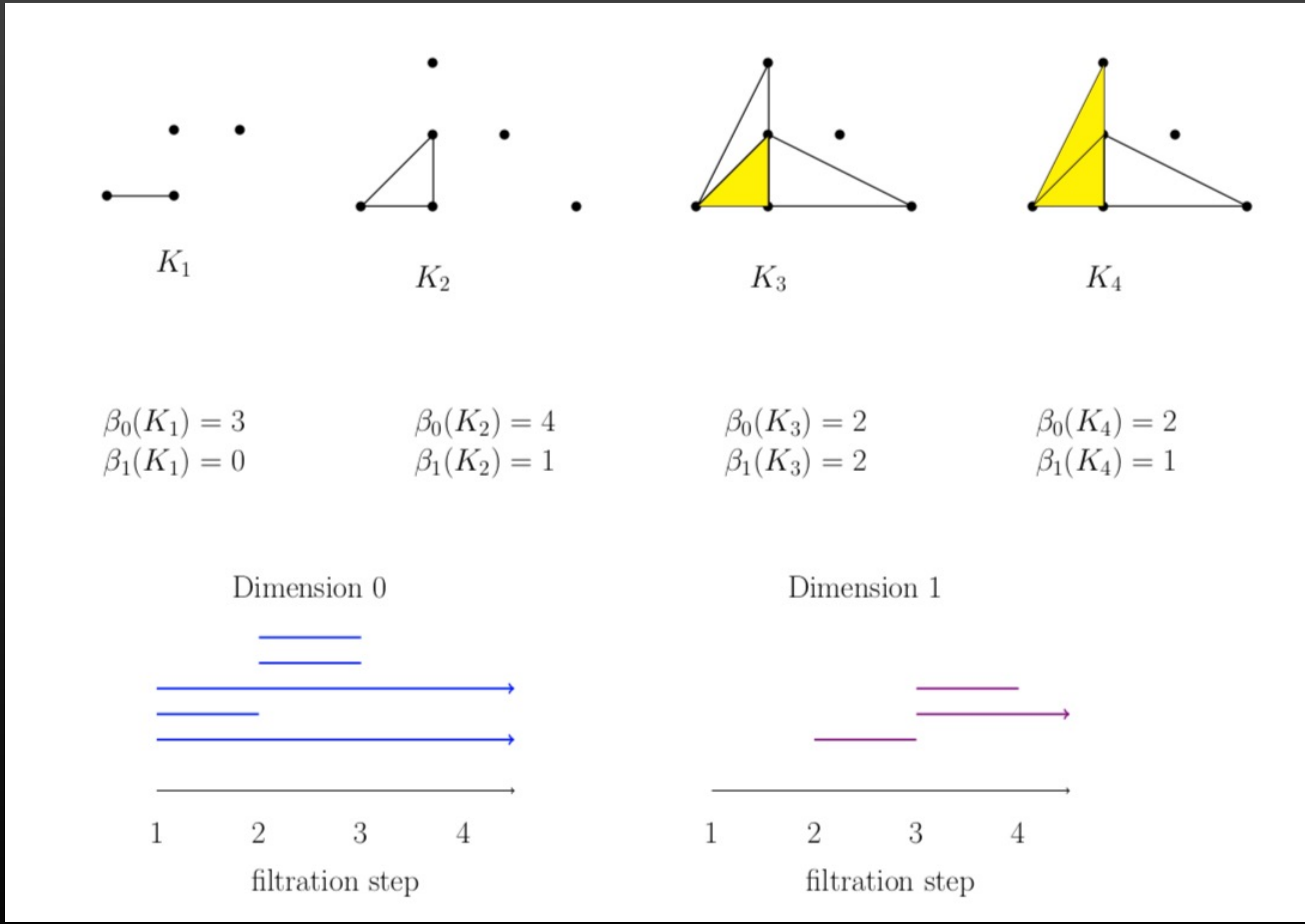
L. Munch, 2019.

Step 2: Point cloud to nested complexes



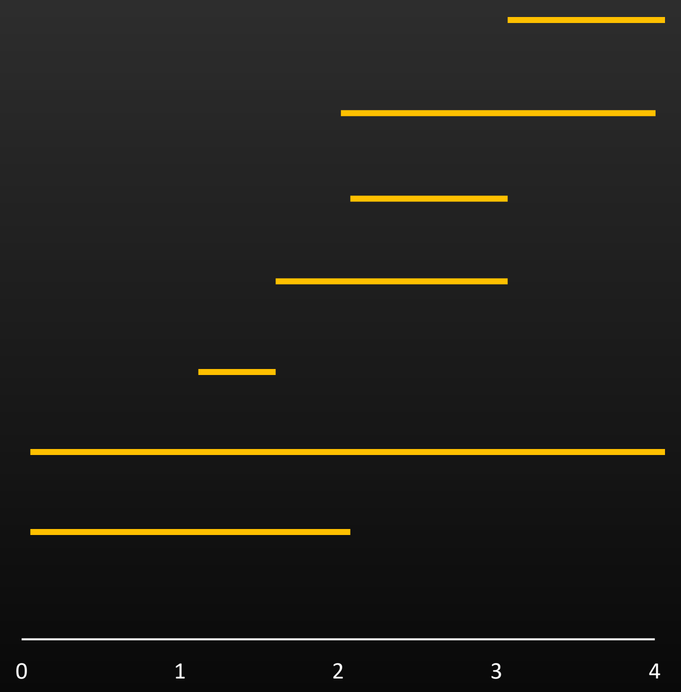
L. Munch, 2019.

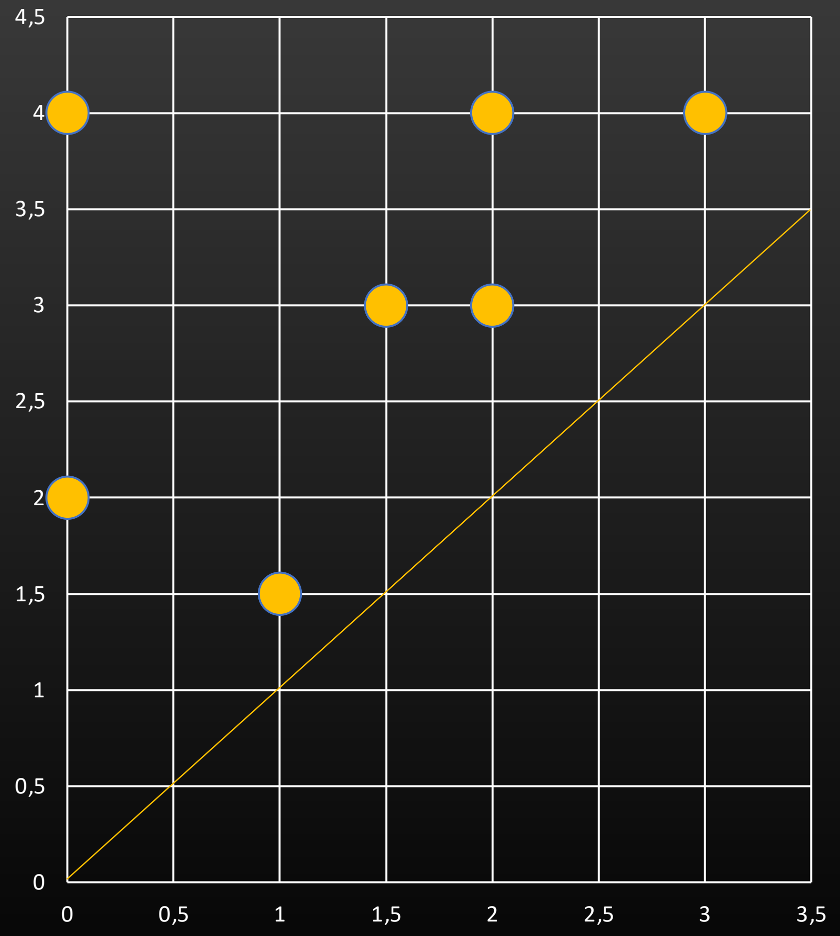
Step 3: Nested complexes to barcode



Otter et al., arXiv, 2016.

Barcodes vs persistence diagrams (PD)





Stability

- The set of barcodes/persistence diagrams can be equipped with a variety of earthmover-type distances: the Wasserstein distances of L_p-type and the bottleneck distance of L_∞ -type.

- Most reasonable known instantiations of the TDA pipeline are Lipschitz continuous with respect to Hausdorff distance on point clouds and bottleneck distance on persistence diagrams.

Practicalities

- There are extensive libraries of software, mostly open source, for TDA computations (e.g., GUDHI, Ripser, Flagser, Giotto-TDA,…).
- There exist “inverse analysis” tools for interpreting results of TDA computations (e.g., work of Hiraoka et al.).

From one to many parameters

- In real data, there are often several parameters along which it would be natural to filter (e.g., some notion of density or time).
- Generalization from one to many parameters poses serious problems, for reasons of both theory and implementation: in general, there is no analogue of barcodes or persistence diagrams.
- Common approaches for two parameters
  - Restrict to lines in the plane determined by the two parameters: fibered bar code.
  - Focus on decompositions into blocks (instead of bars) when possible.

Static TDA input to ML

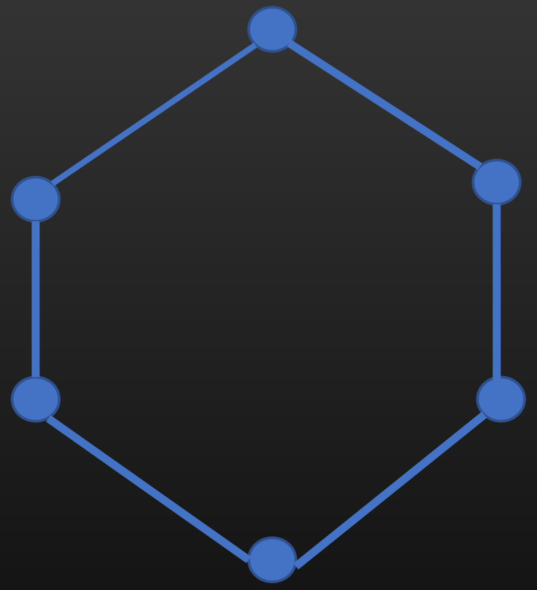
Strategies for vectorization/featurization

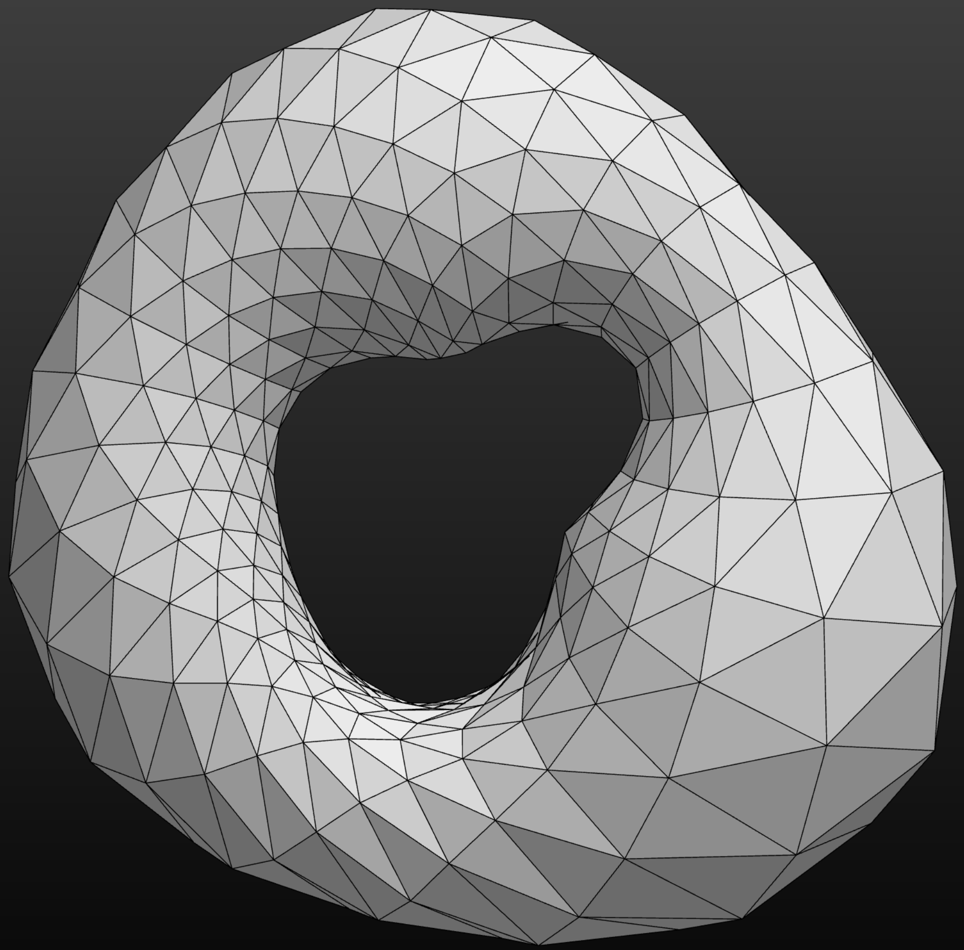
- Problem: Cannot compute statistics in the space of barcodes or the space of persistence diagrams.
- Solution:
  - Define a Lipschitz-continuous mapping from the space of barcodes/persistence diagrams to a vector space $\mathcal{V}$ equipped with an inner product.
  - Compute statistics in $\mathcal{V}$!
  - Two main types:
    - Embeddings into finite-dimensional Euclidean spaces

      Few trainable parameters
    - Kernel methods: defining generalized scalar product on PD, i.e., see PD as elements of a Hilbert space

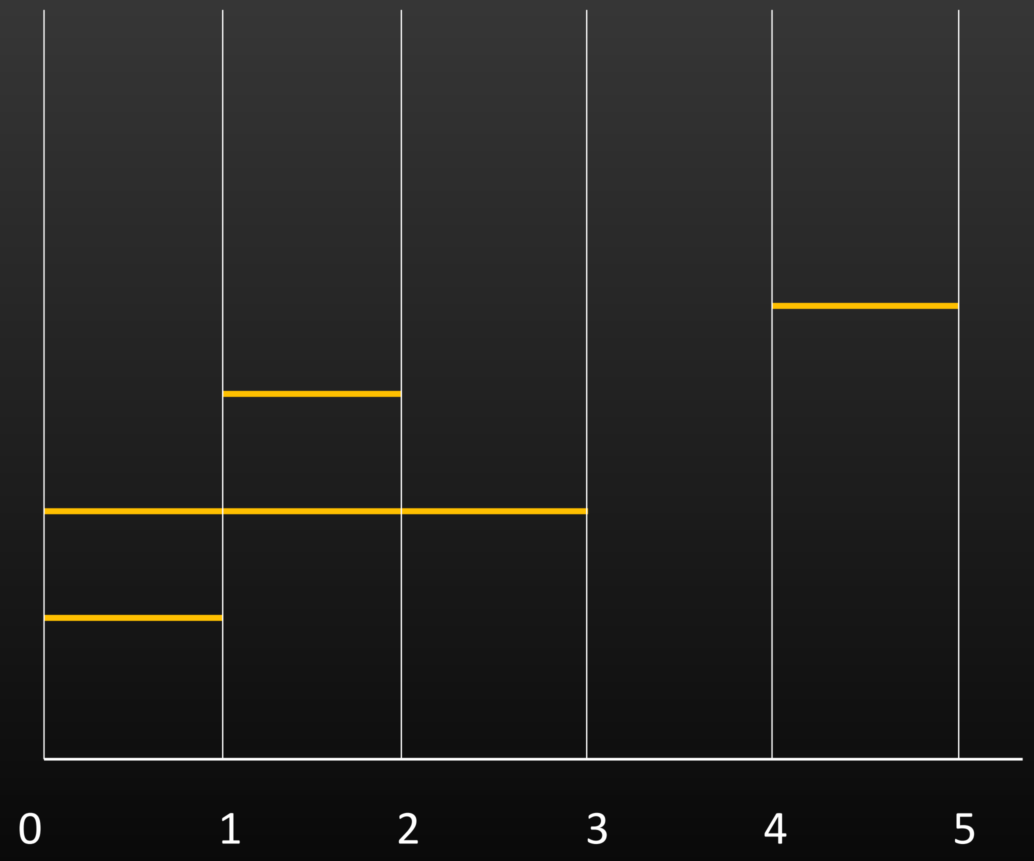
      Expensive

Cavities

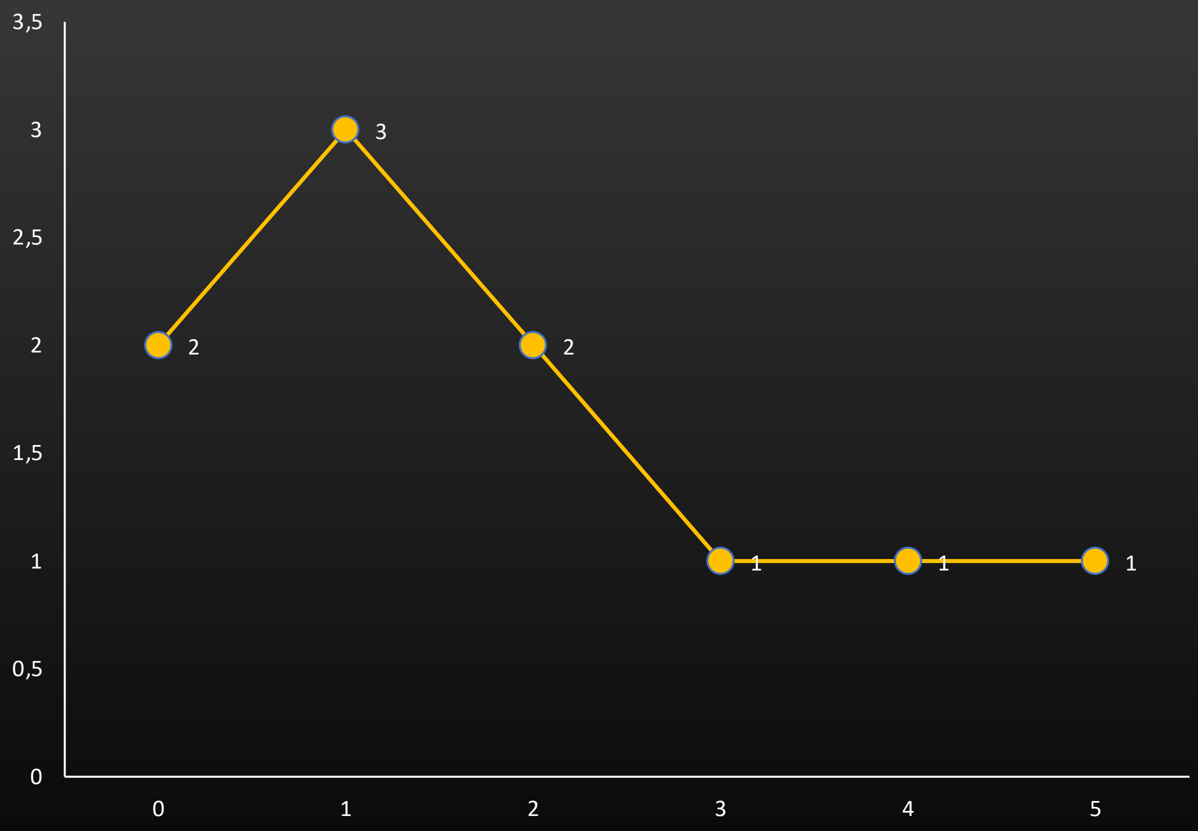




Betti curves

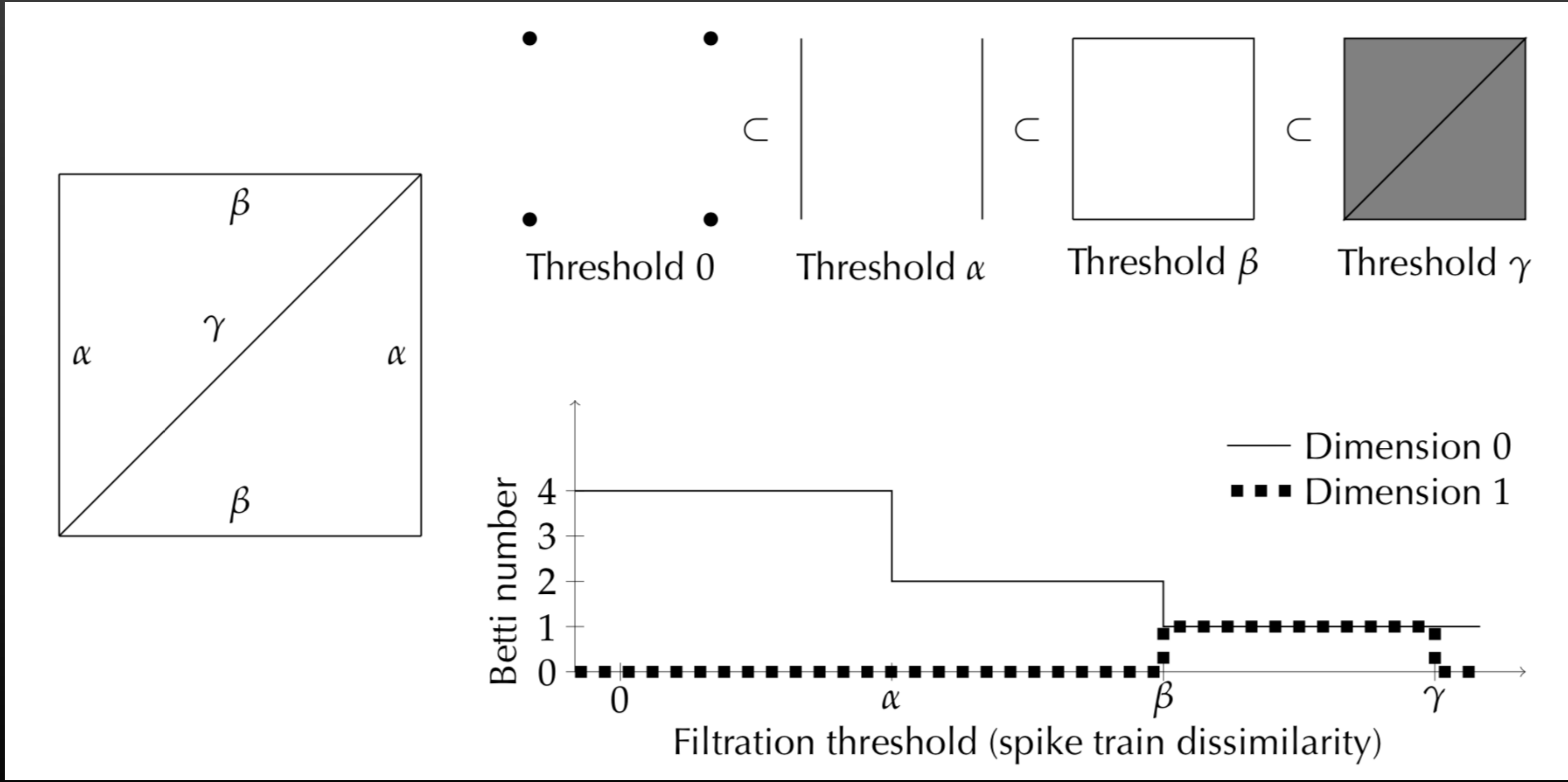


Bar code for cavities of dimension k



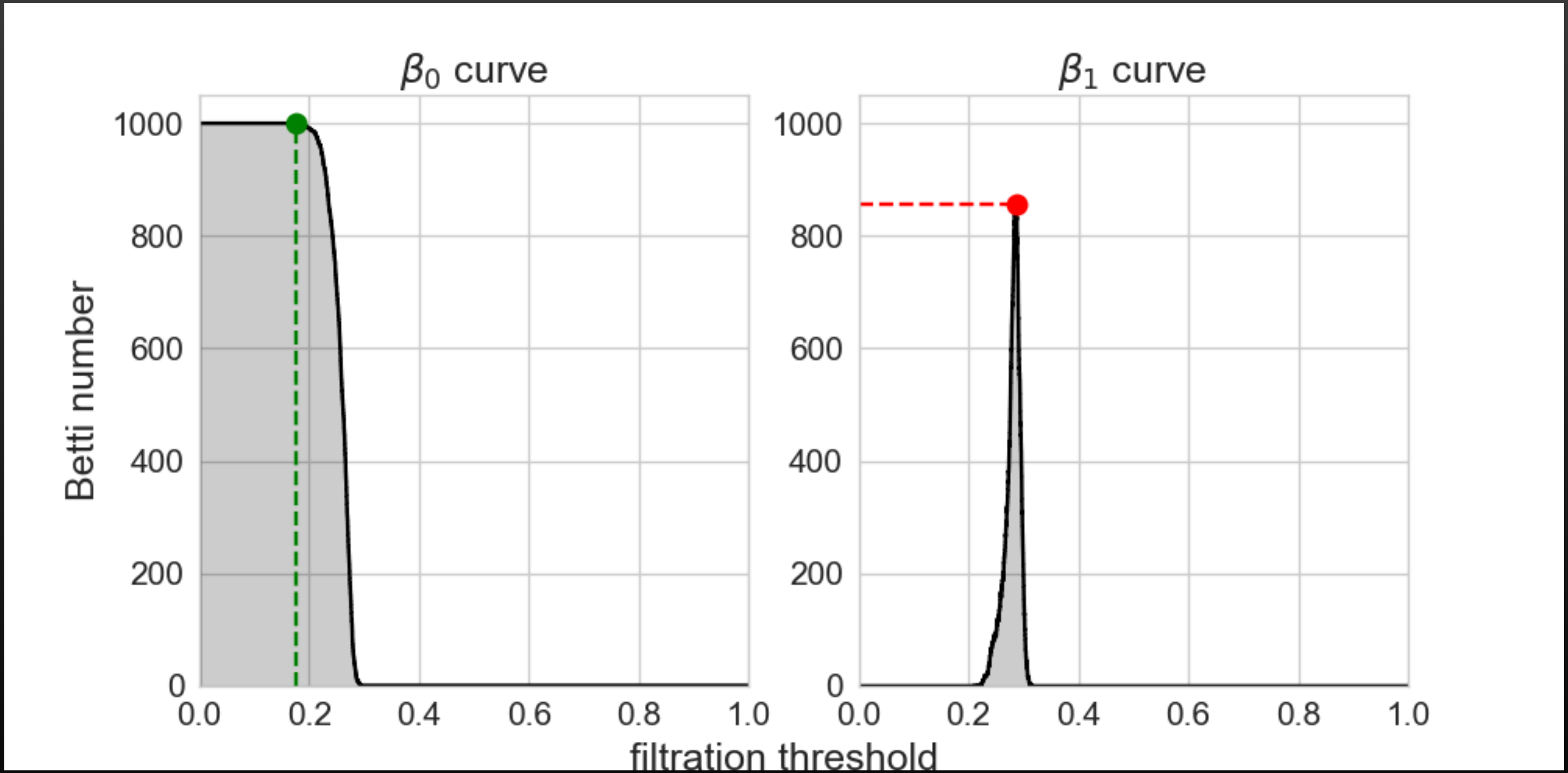
Betti_k curve

Nested complex to Betti curve



Bardin, et al., Network Neuroscience, 2019.

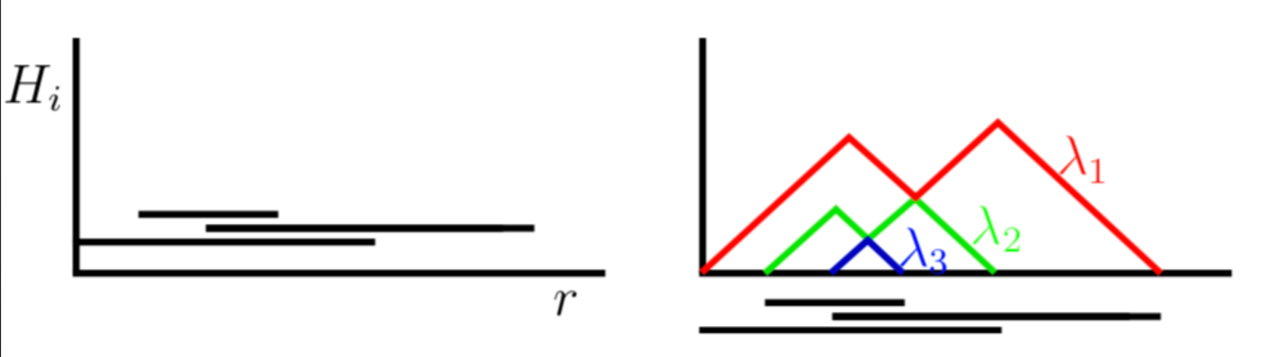
Extracting numerical features



Bardin, et al., Network Neuroscience, 2019.

Persistence landscapes

- Barcodes also give rise to persistence landscapes.



$$\lambda = \{\lambda_k : \mathbb{R} \to \mathbb{R} \cup \{\infty\} \mid k \in \mathbb{N}\}$$

- The L2-landscape distance between barcodes B and B' with associated landscapes $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$:

$$\Lambda(B, B') = \|\lambda - \lambda'\|_2 = \sum_{k=1}^{\infty} \left(\int |\lambda_k(t) - \lambda'_k(t)|^2 dt \right)^{\frac{1}{2}}$$

Bubenik, JMLR 2015
Dlotko & Bubenik, J Symbolic Comp 2017

Persistence curves

Name	Notation	$\psi(b,d,t)$	T
Betti	$\boldsymbol{\beta(D)}$	1	sum
Midlife	$\mathbf{ml}(D)$	$(b+d)/2$	sum
Life	$\boldsymbol{\ell}(D)$	$d-b$	sum
Multiplicative Life	$\mathbf{mul}(D)$	d/b	sum
Life Entropy [2]	$\mathbf{le}(D)$	$-\frac{d-b}{\sum(d-b)}\log\frac{d-b}{\sum(d-b)}$	sum
Midlife Entropy	$\mathbf{mle}(D)$	$-\frac{d+b}{\sum(d+b)}\log\frac{d+b}{\sum(d+b)}$	sum
Mult. Life Entropy	$\mathbf{mule}(D)$	$-\frac{d/b}{\sum(d/b)}\log\frac{d/b}{\sum(d/b)}$	sum
k-th Landscape [5]	$\boldsymbol{\lambda_k}(D)$	$\min\{t-b,d-t\}$	$\max_k$

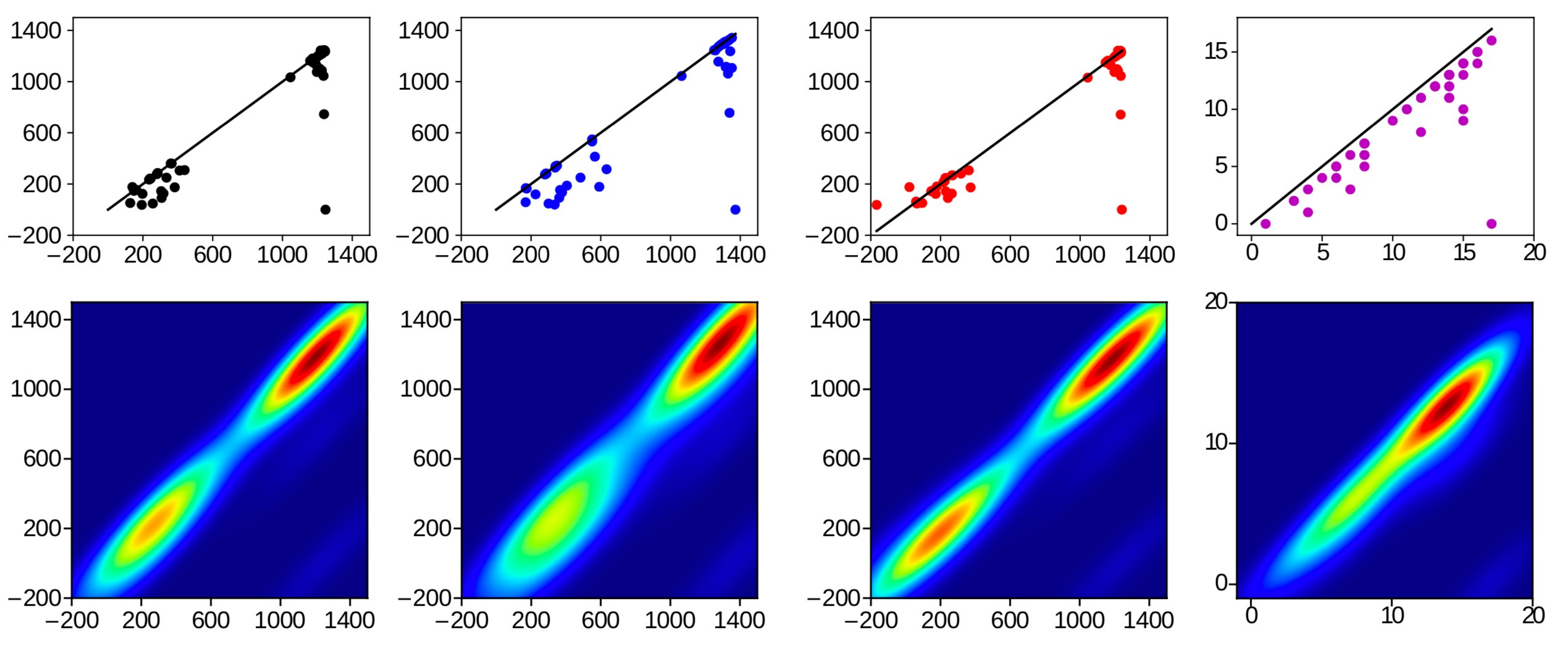
Simultaneous generalization of Betti curves and persistence landscapes. Robust to input noise, efficient to compute, interpretable, and allowing weighting of relative importance of different regions in the PD.

For each t, compute T({ ψ(b,d,t) | b ≤ t, d > t}).

Chung and Lawson, arXiv, 2019

Persistence images

- Smooth the PD: replace each point by a Gaussian kernel, then sum
- Discretize



(Image from Kanari, et al., Neuroinformatics, 2018.)

Adams et al., JMLR 2017

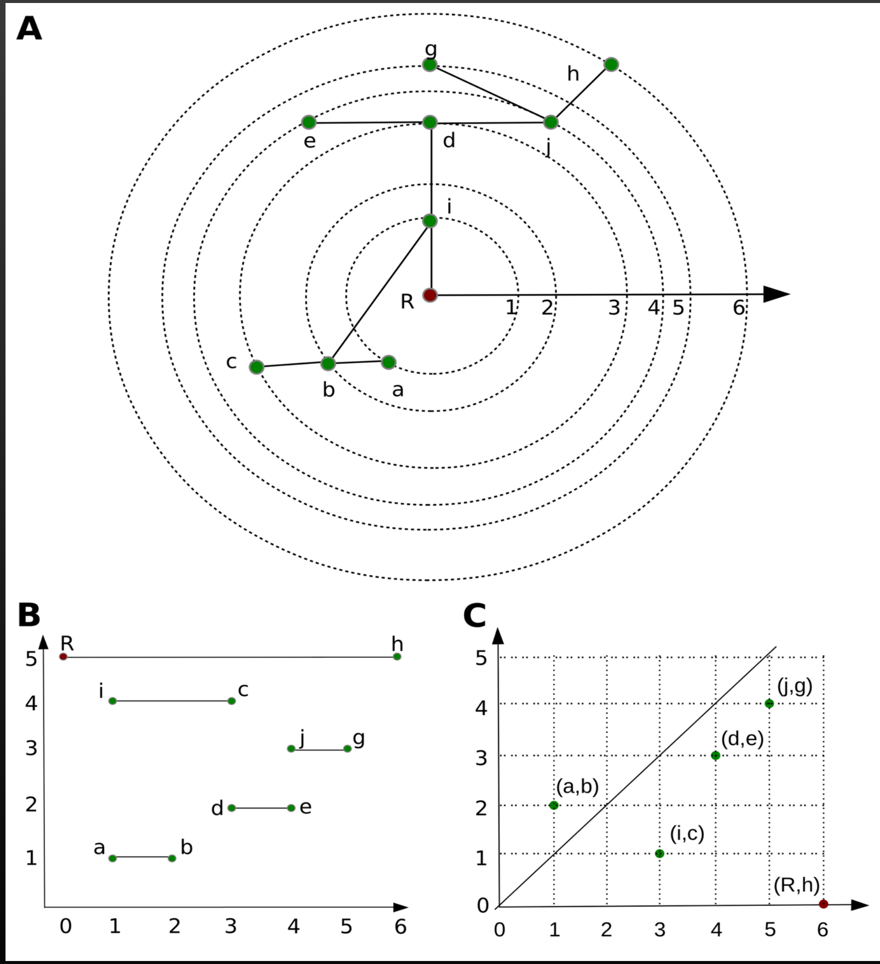
ML methods applied to vectorized TDA

- Decision tree
- Random forest
- Support Vector Machine
- CNN
- GNN

Also possible to integrate a TDA layer into an ML model!

Applications

Classification of neuron morphologies

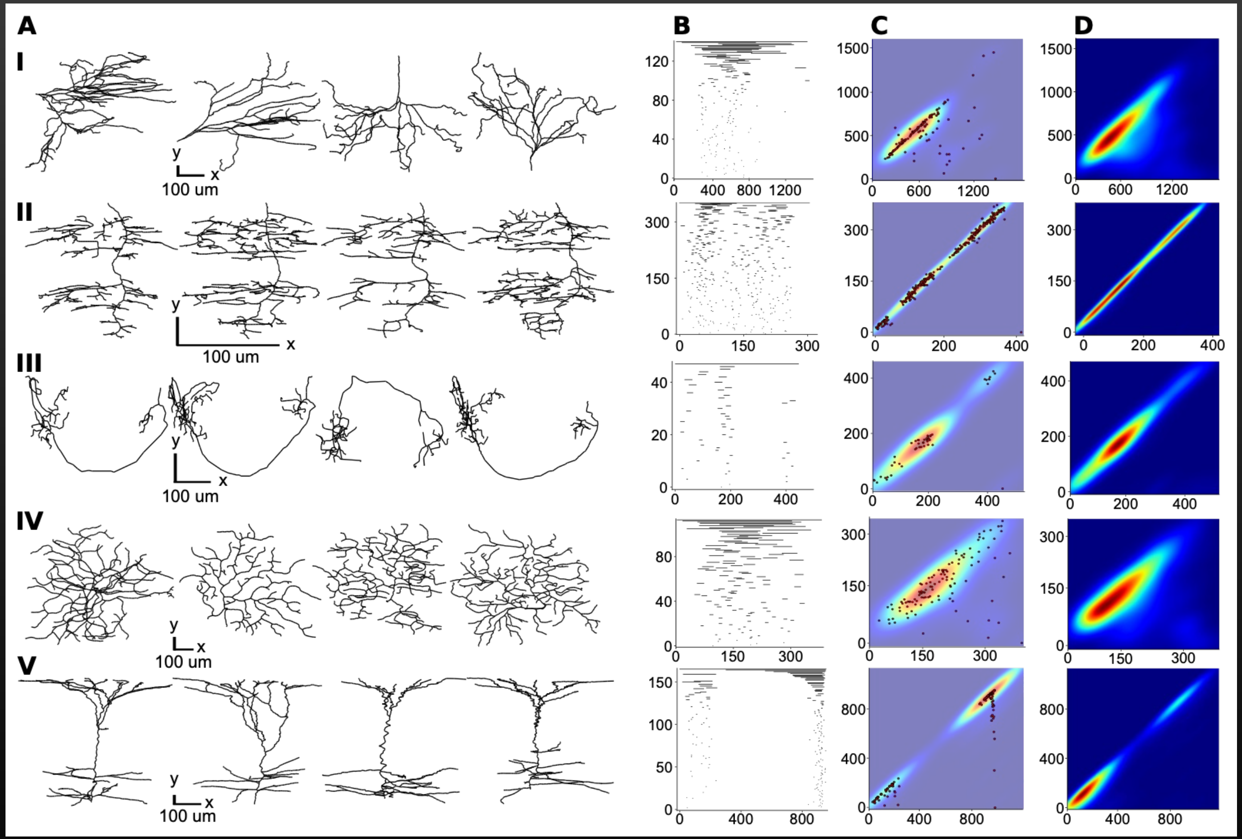


Idea: Starting at the leaves and descending recursively to the root, decompose the tree into branches, while respecting the Elder Rule, i.e., at any bifurcation, the elder (longer) branch survives and the younger branch is broken off.

Integrate the topology of the tree and the geometry of its embedding in space into a surprisingly powerful global descriptor.

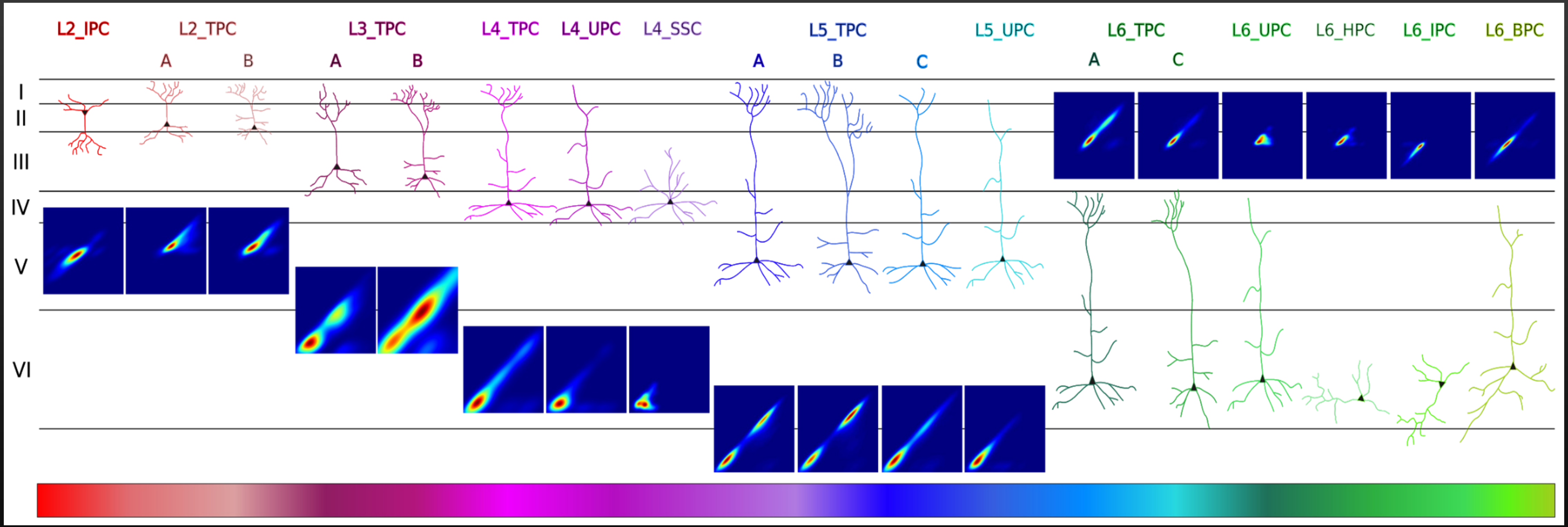
Kanari et al., Neuroinformatics 2017
Kanari et al., Cerebral Cortex 2019

Classification of neuron morphologies



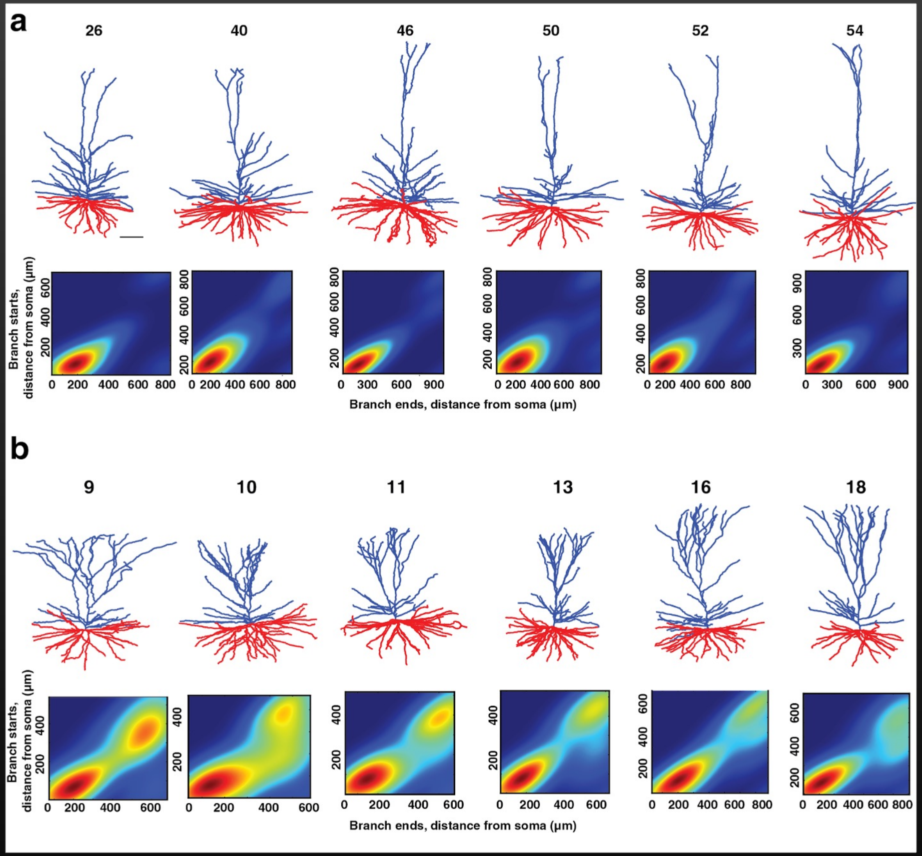
Kanari et al., Neuroinformatics 2017
Kanari et al., Cerebral Cortex 2019

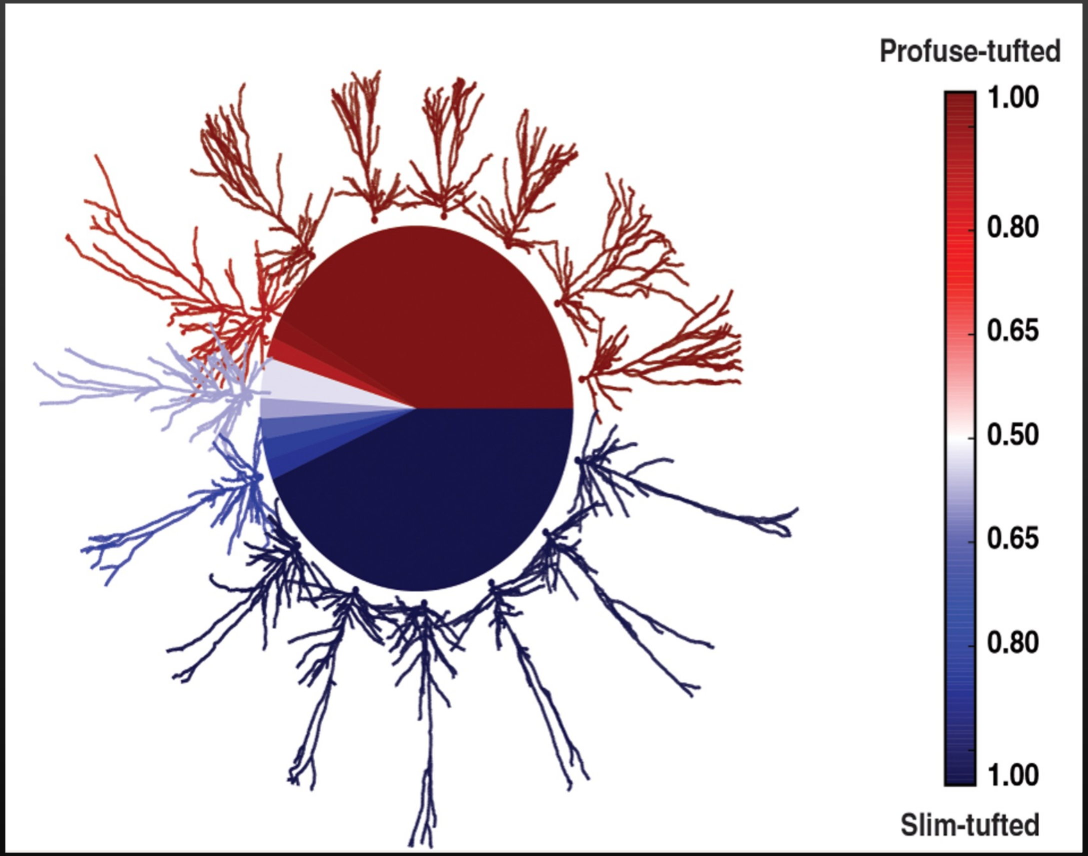
Classification of neuron morphologies



Kanari et al., Neuroinformatics 2017
Kanari et al., Cerebral Cortex 2019

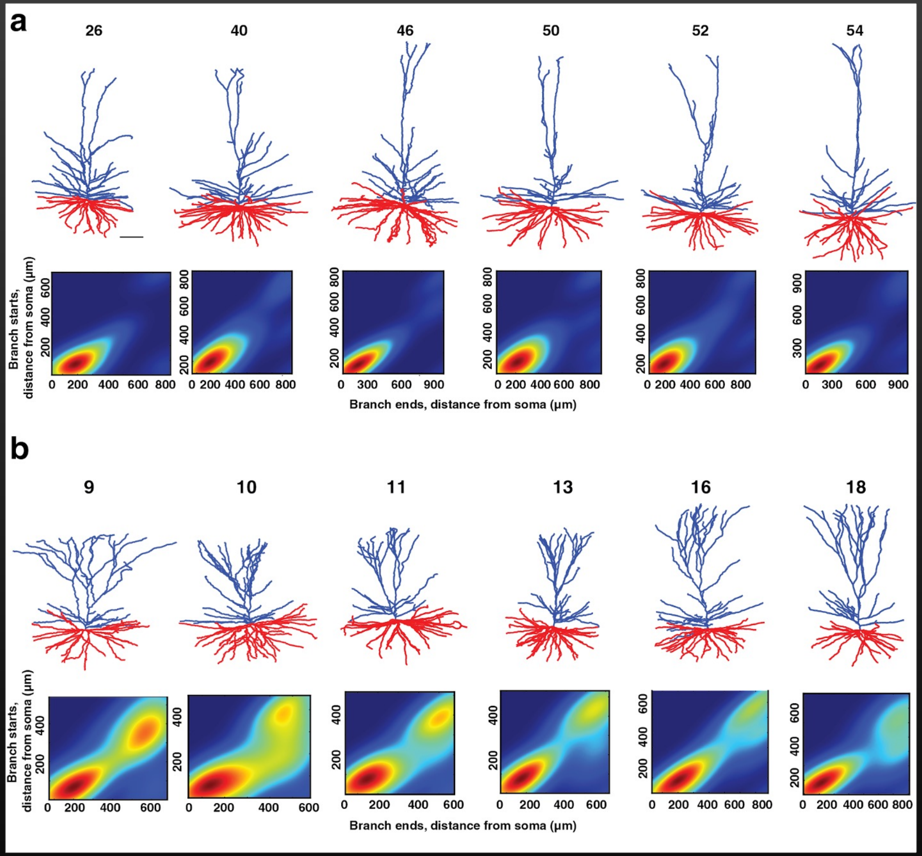
Classification of neuron morphologies

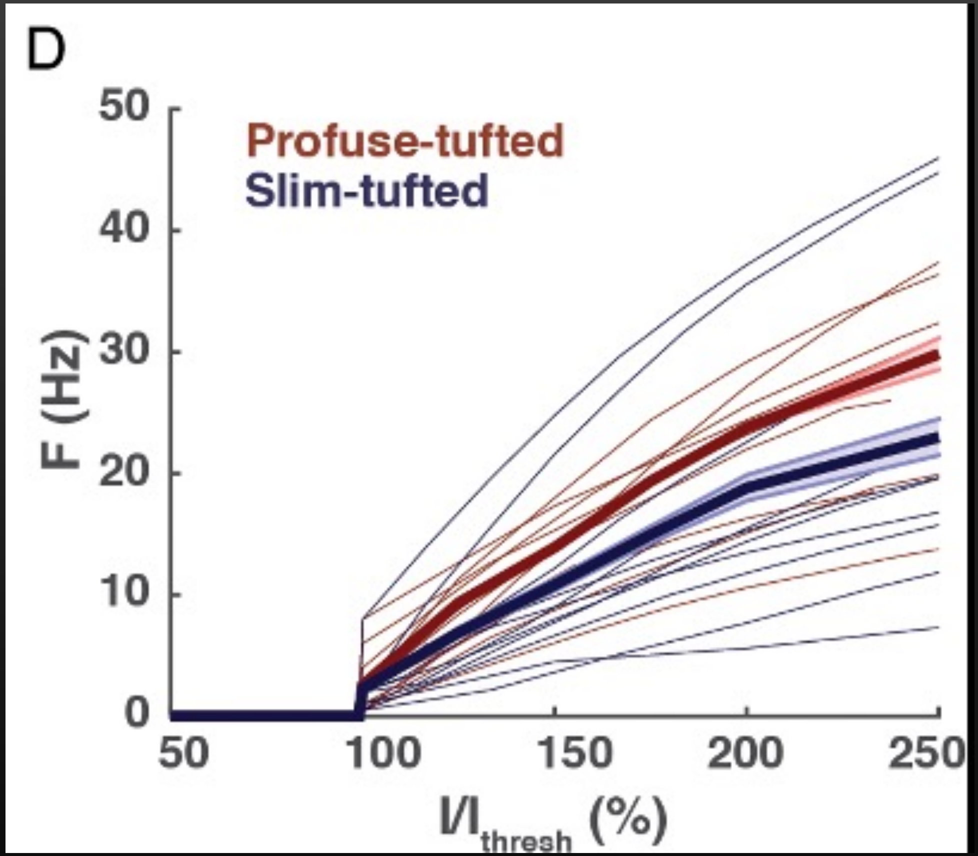




Kanari et al., Cerebral Cortex 2017

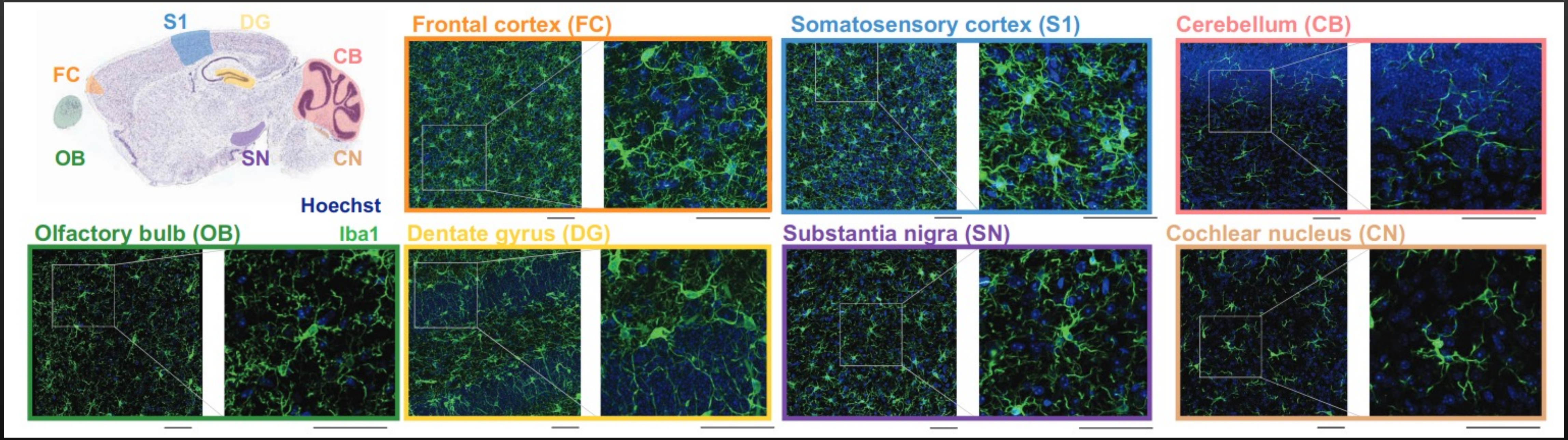
Classification of neuron morphologies





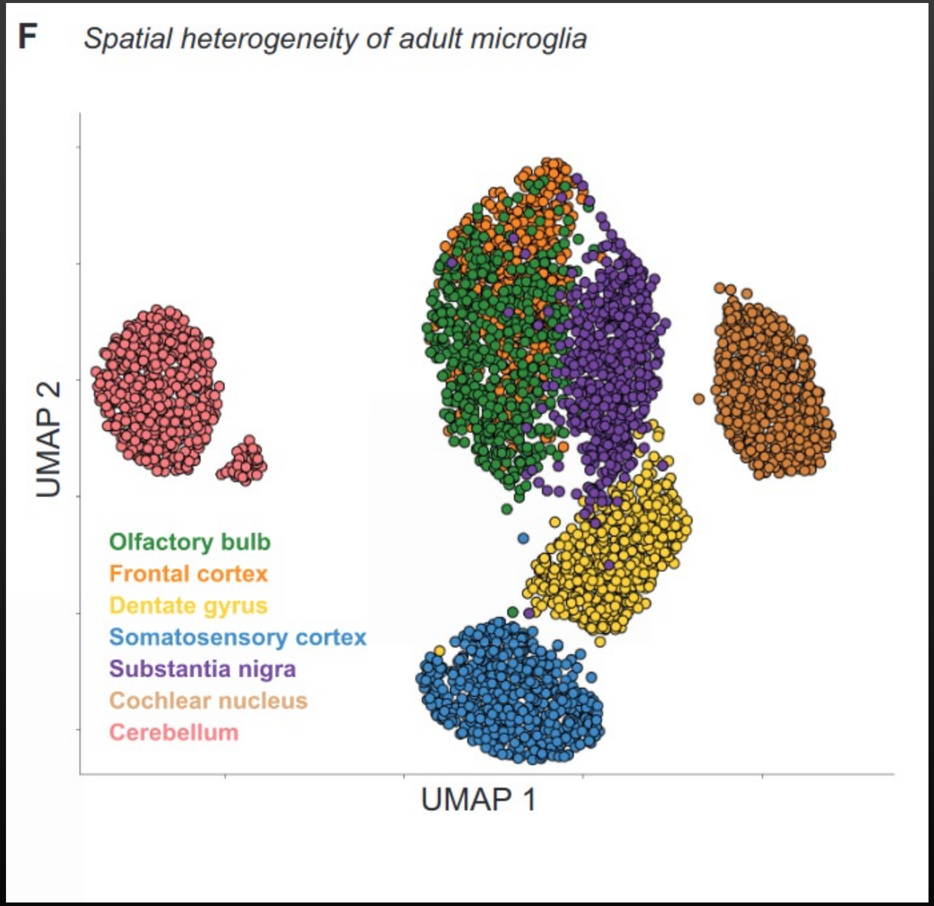
Kanari et al., Cerebral Cortex 2017

Classification of microglia



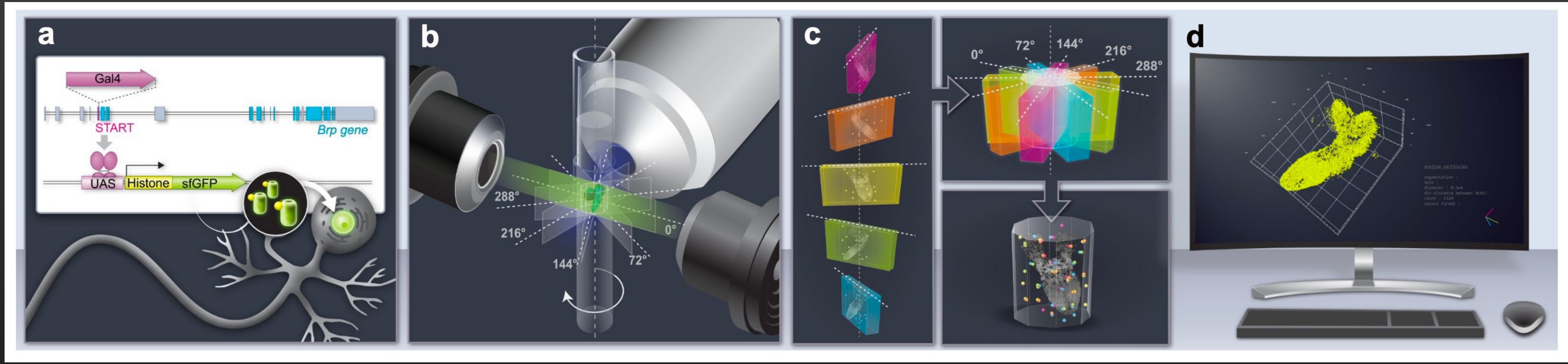
Colombo, et al., Nature Neuroscience, 2022.

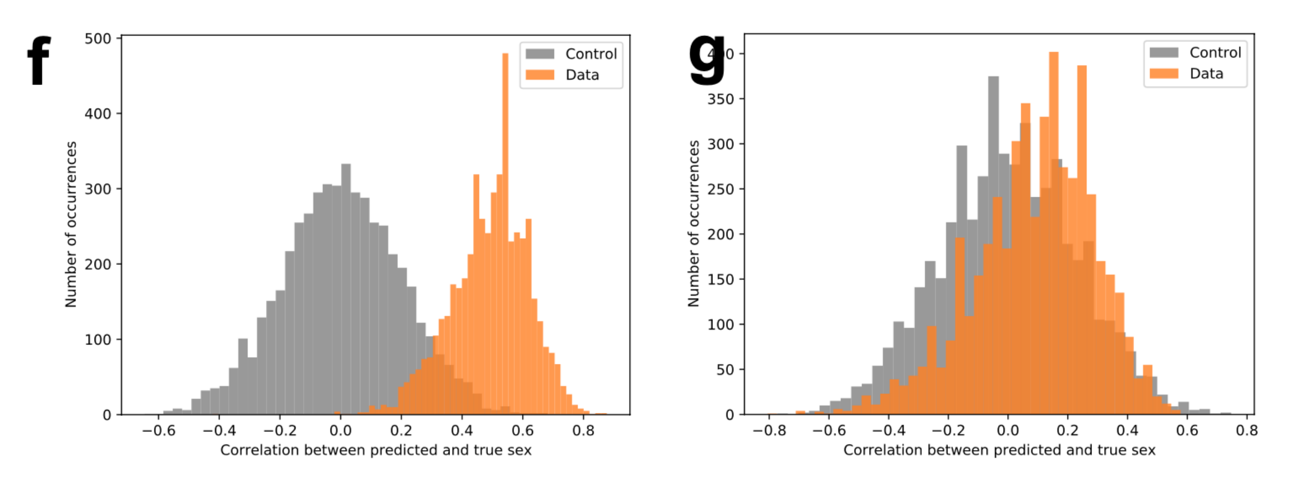
Classification of microglia



Colombo, et al., Nature Neuroscience, 2022.

Sexual dichotomy in larval fruitflies

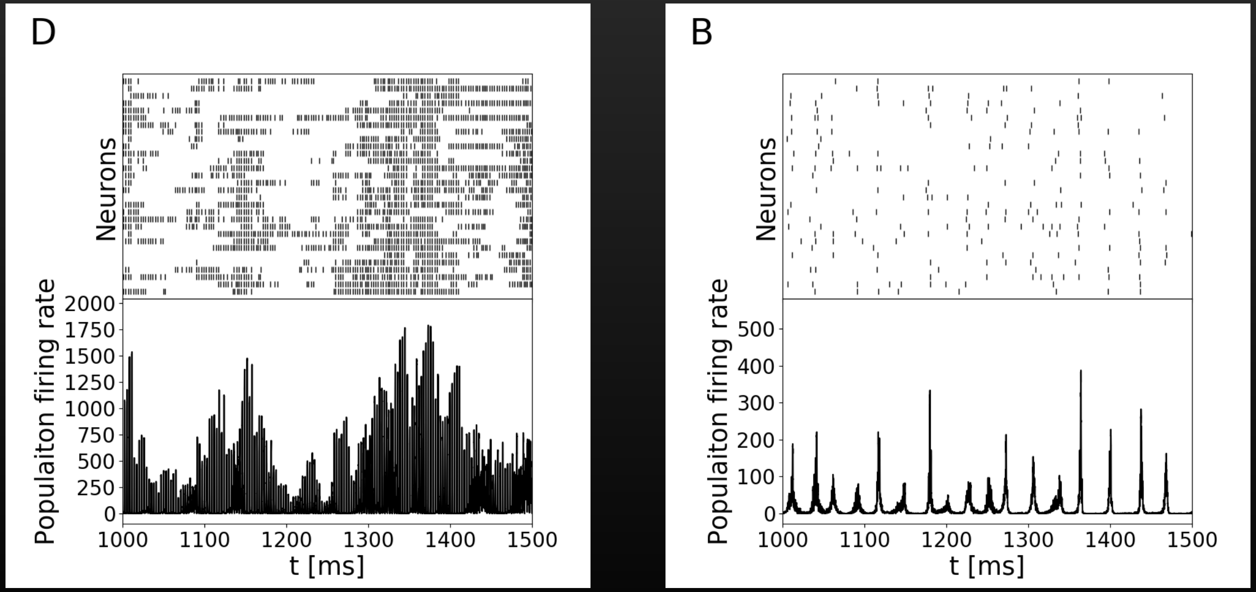




Jiao et al, eLife, 2022

Classification of neural dynamics

- For a range of activity parameters, associate to an active Brunel network a weighted graph, to which we apply tools of persistent homology.
- Extract simple topological features of each dynamic regime.
- Use these to train a (highly accurate!) classifier.



Bardin, et al., Network Neuroscience, 2019.

Classification of neural dynamics

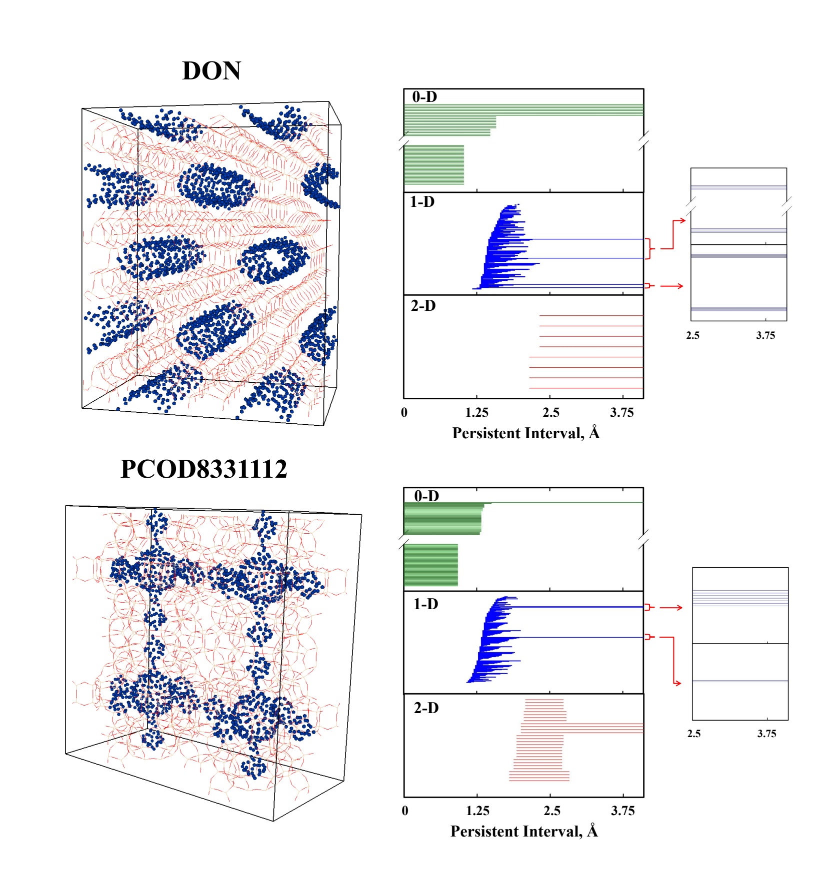
Automated classification of network dynamics

- For a range of activity parameters, associate to an active Brunel network a weighted graph, to which we apply tools of persistent homology.
- Extract simple topological features of each dynamic regime.
- Use these to train a (highly accurate!) classifier.

	Testing set			
Training set	**Ver. 1**	**Ver. 2**	**Ver. 3**	**All ver.**
Version 1	100% (28)	86.67% (180)	91.18% (170)	89.68% (378)
Version 2	97.69% (130)	100% (24)	93.33% (240)	95.18% (394)
Version 3	99.23% (130)	99.17% (240)	100% (24)	99.23% (394)
All versions	100% (28)	100% (24)	100% (24)	100% (76)

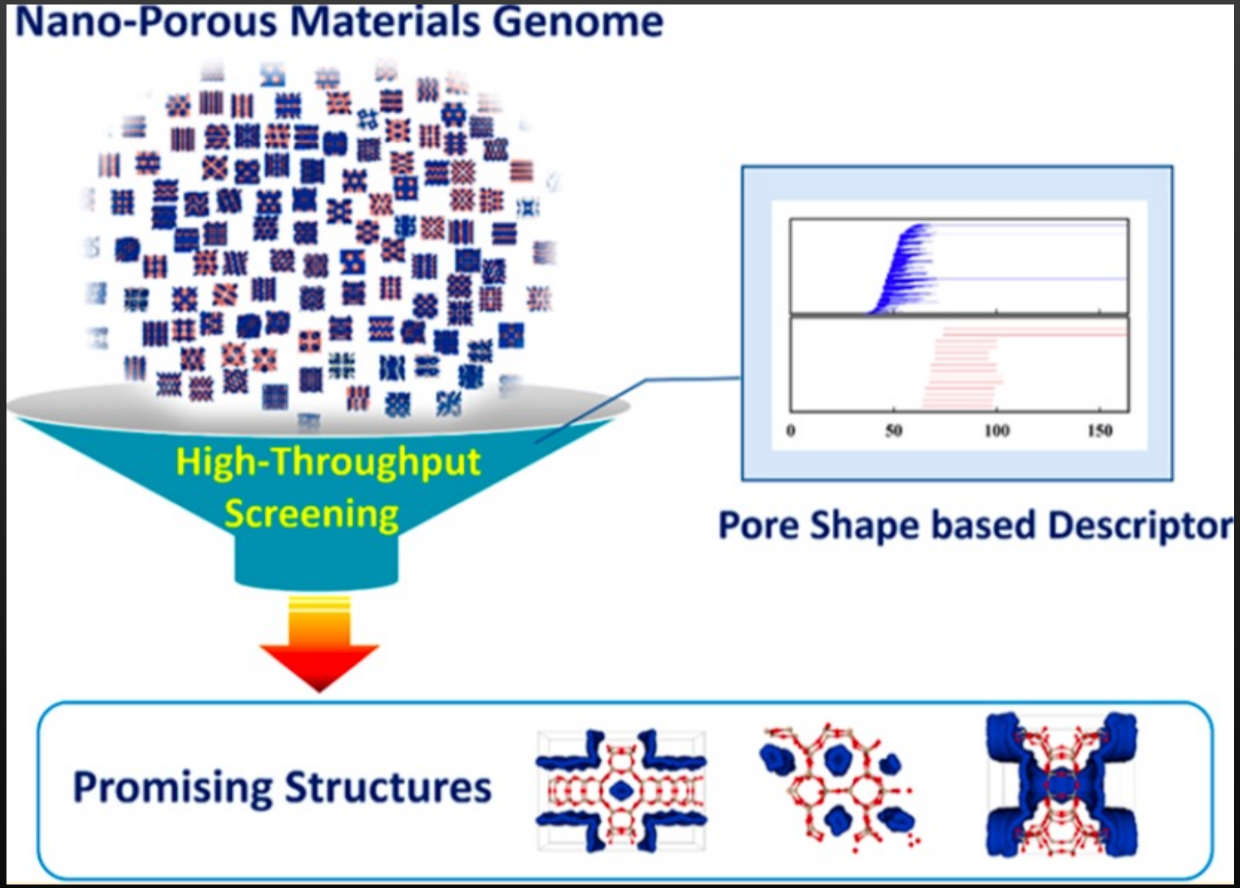
Bardin, et al., Network Neuroscience, 2019.

Classification of nanoporous crystalline materials



Lee et al., Nature Communications 2017
Lee et al., J Chem Thy Comput 2018

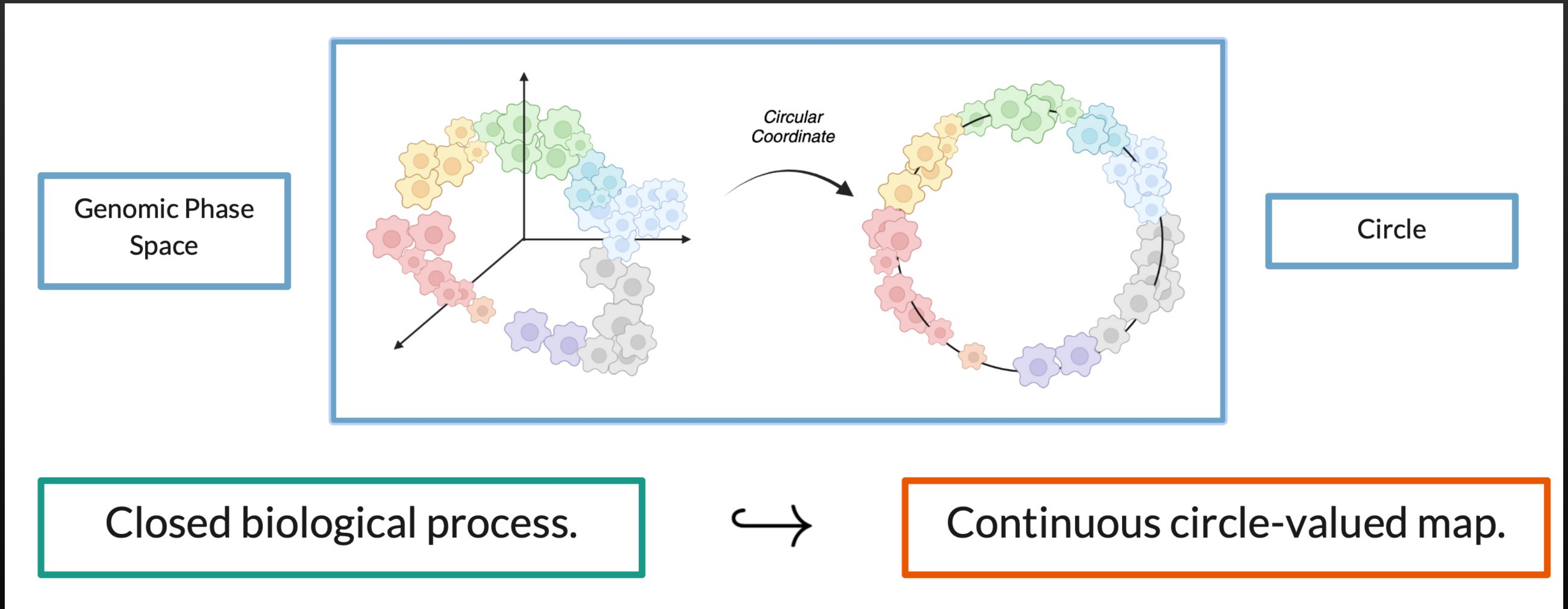
Classification of nanoporous crystalline materials



Lee et al., Nature Communications 2017
Lee et al., J Chem Thy Comput 2018

Detection of gene cascades in single-cell data

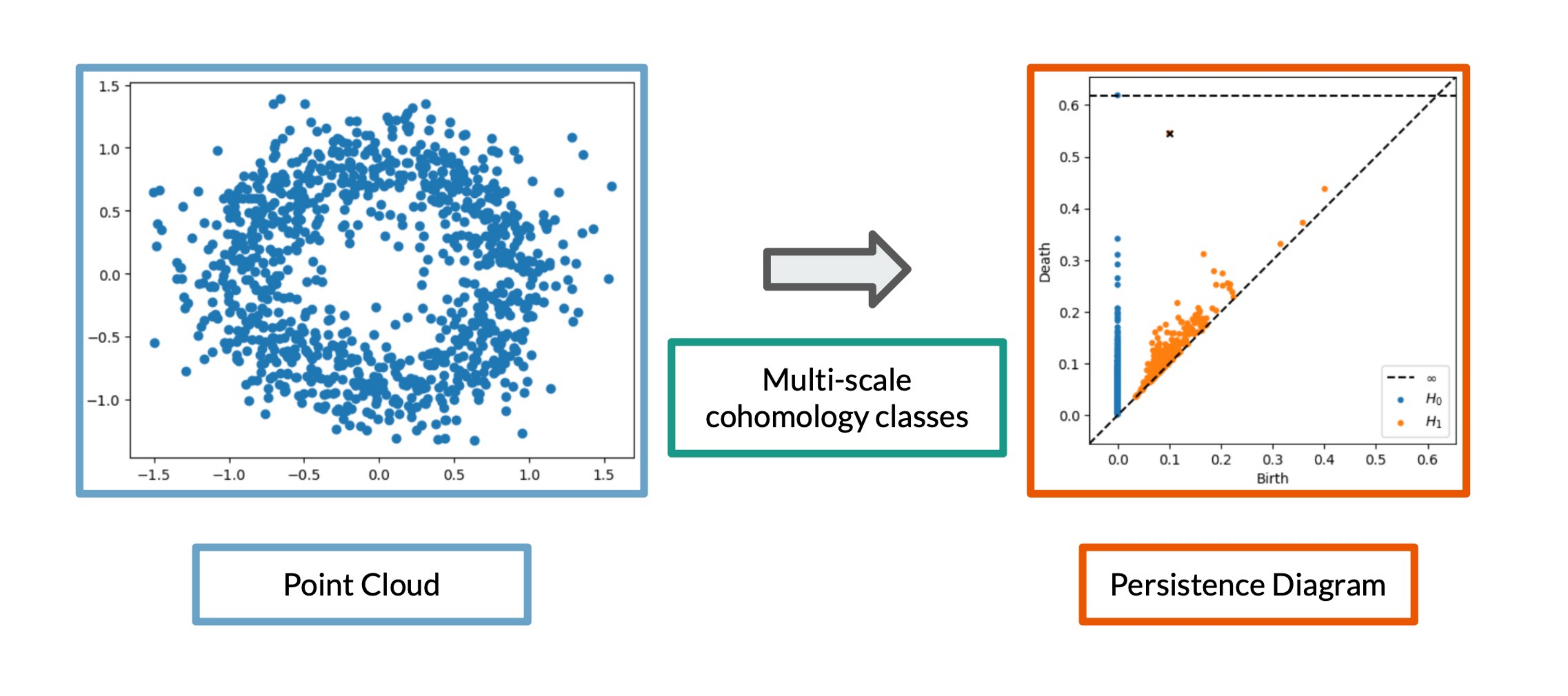
- Idea: Generate biological hypotheses about closed processes in single-cell RNA seq data using topology and geometry



Maggs, Nguyen, Youssef, in progress

Detection of gene cascades in single-cell data

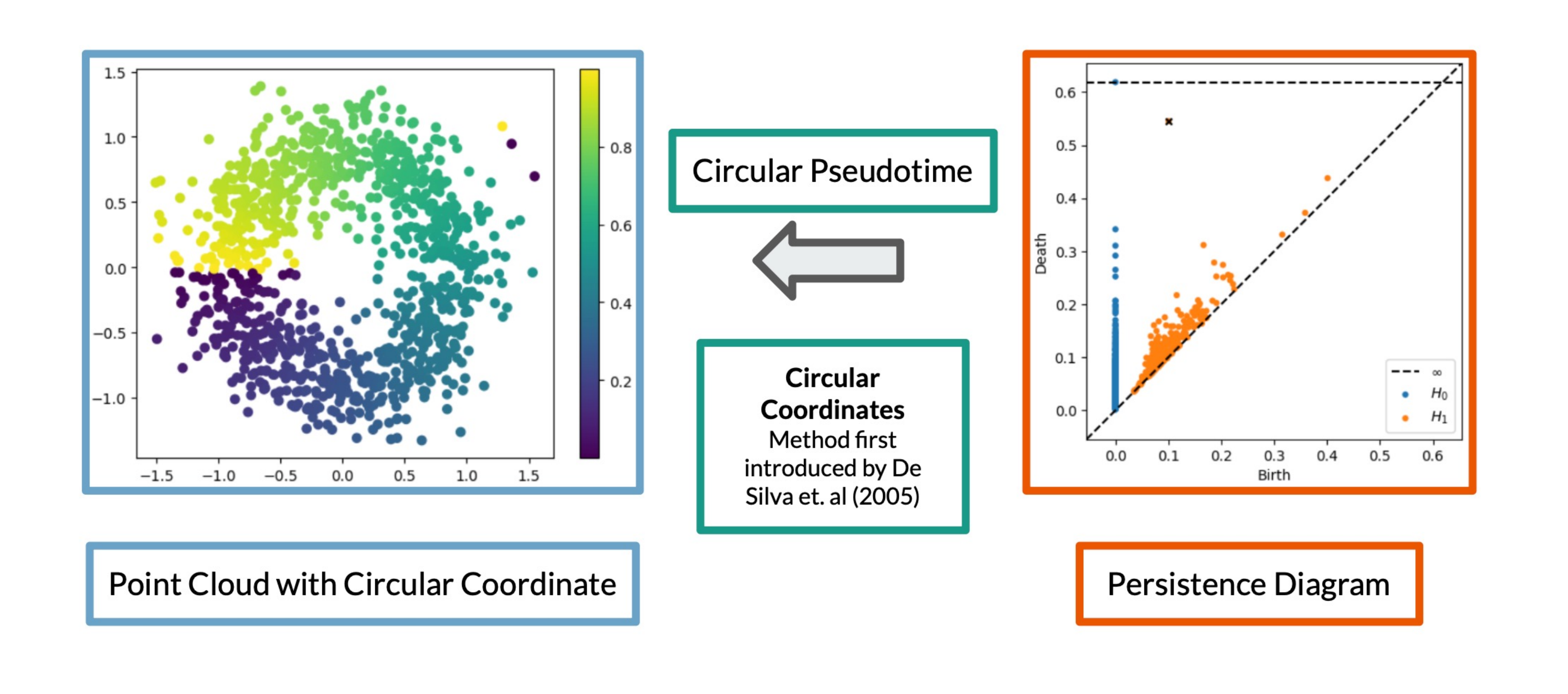
- Idea: Generate biological hypotheses about closed processes in single-cell RNA seq data using topology and geometry



Maggs, Nguyen, Youssef, in progress

Detection of gene cascades in single-cell data

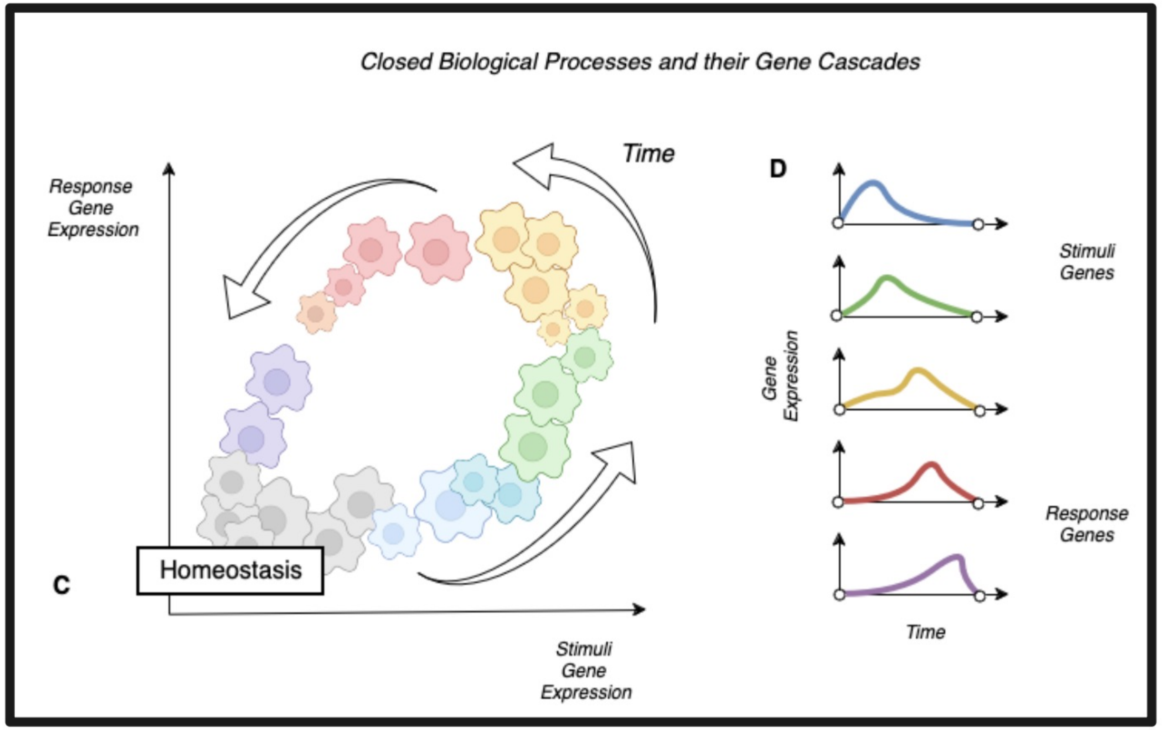
- Idea: Generate biological hypotheses about closed processes in single-cell RNA seq data using topology and geometry



Maggs, Nguyen, Youssef, in progress

Detection of gene cascades in single-cell data

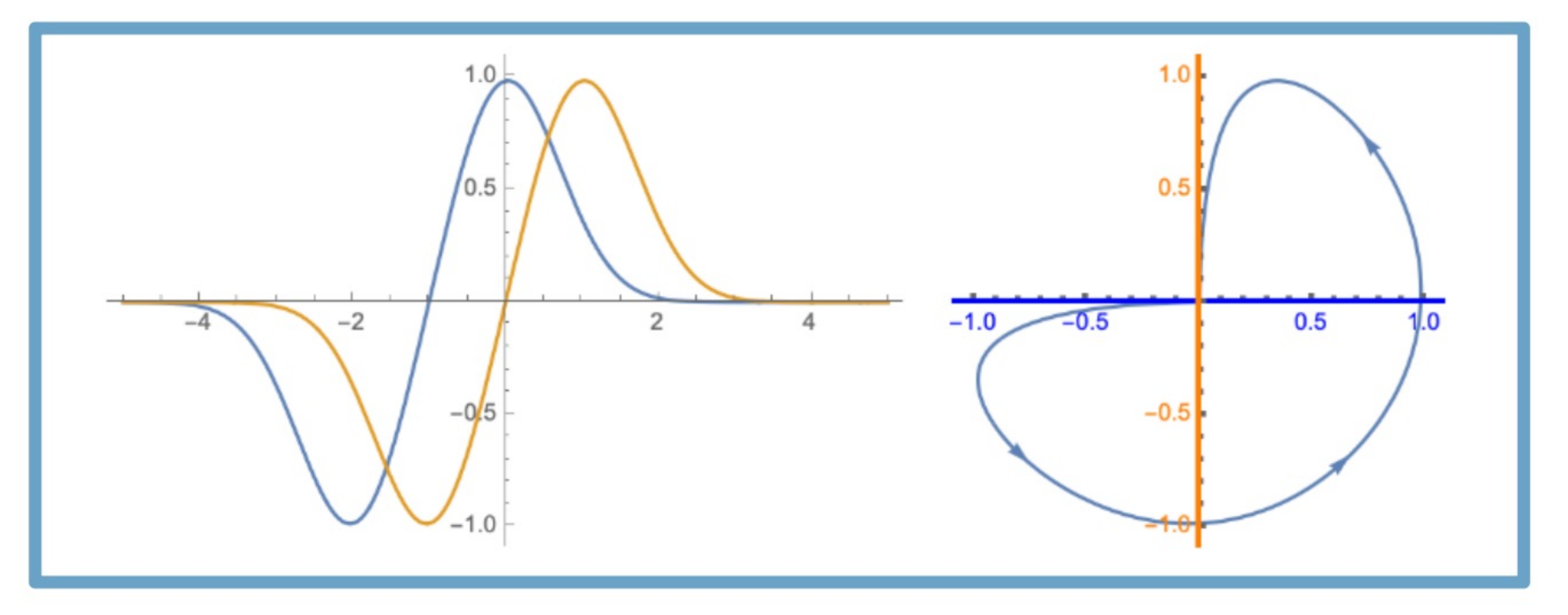
Closed biological process



Gene expression cascade

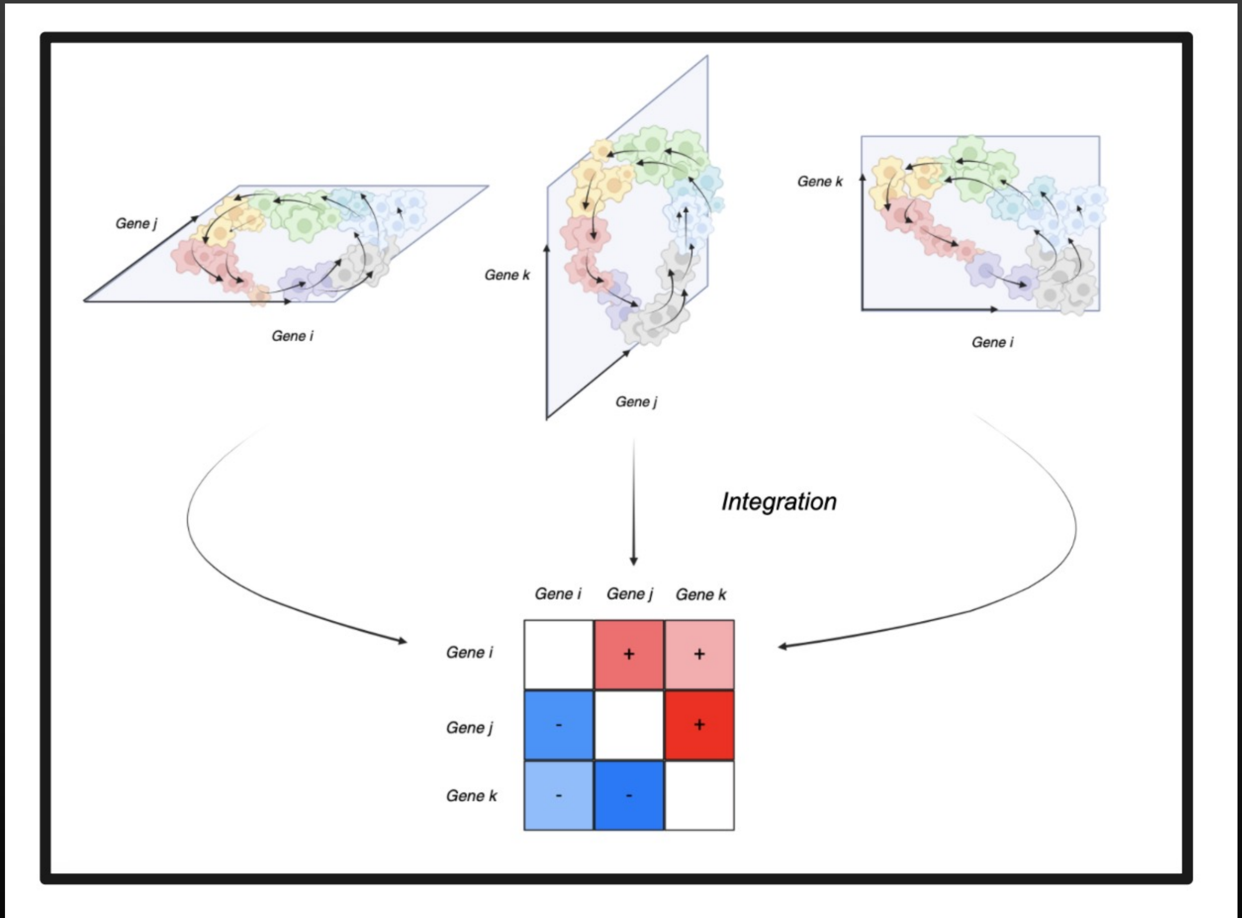
Maggs, Nguyen, Youssef, in progress

Detection of gene cascades in single-cell data



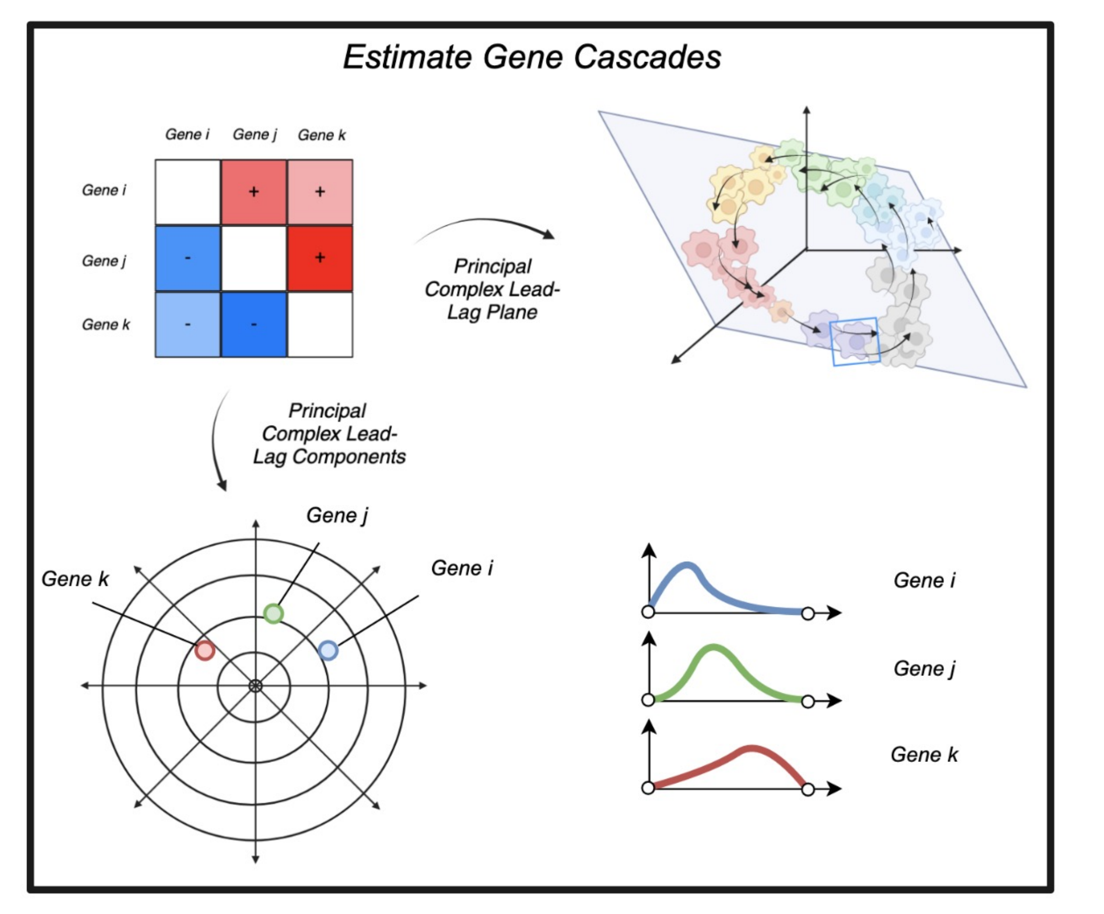
Maggs, Nguyen, Youssef, in progress

Detection of gene cascades in single-cell data



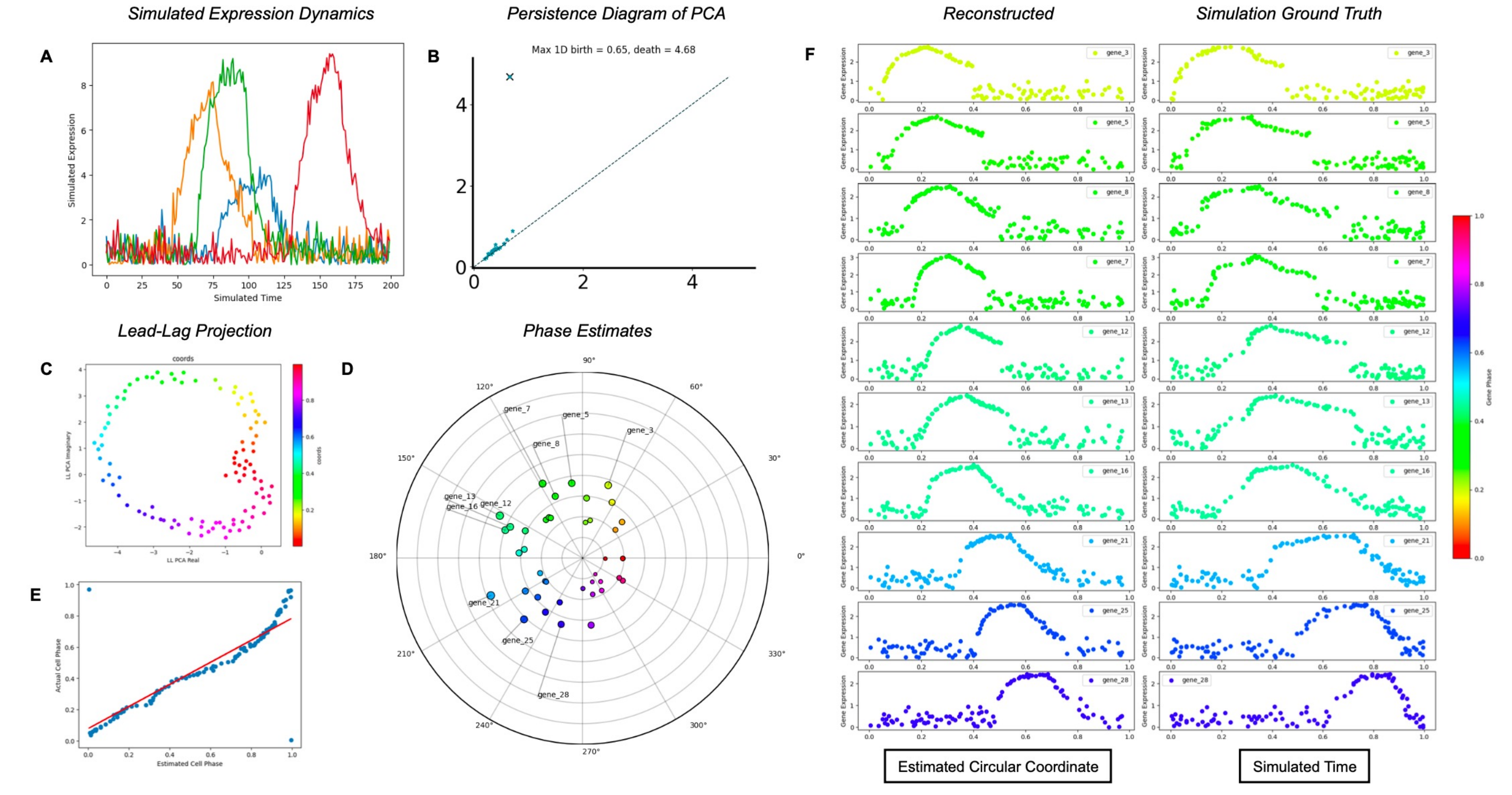
Maggs, Nguyen, Youssef, in progress

Detection of gene cascades in single-cell data



Maggs, Nguyen, Youssef, in progress

Detection of gene cascades in single-cell data



Maggs, Nguyen, Youssef, in progress

Thank you to our funding agencies

- SNSF
- Innosuisse
- Blue Brain Project

Merci !

Members of the Laboratoire de Topologie et Neurosciences