# Topological Data Analysis

Peter Bubenik

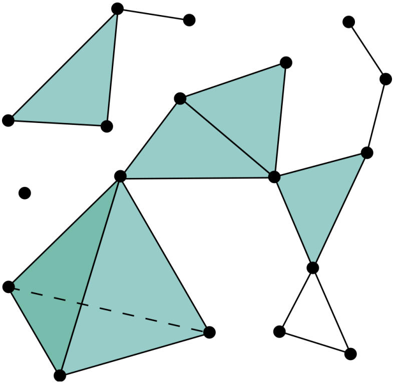
University of Florida
Department of Mathematics,
peter.bubenik@ufl.edu
http://people.clas.ufl.edu/peterbubenik/

British Applied Mathematics Colloquium
University of Oxford
April 6, 2016

# Topological Data Analysis

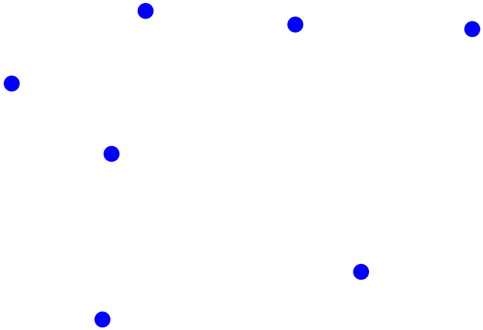
## Idea

Use topology to summarize and learn from the “shape” of data.



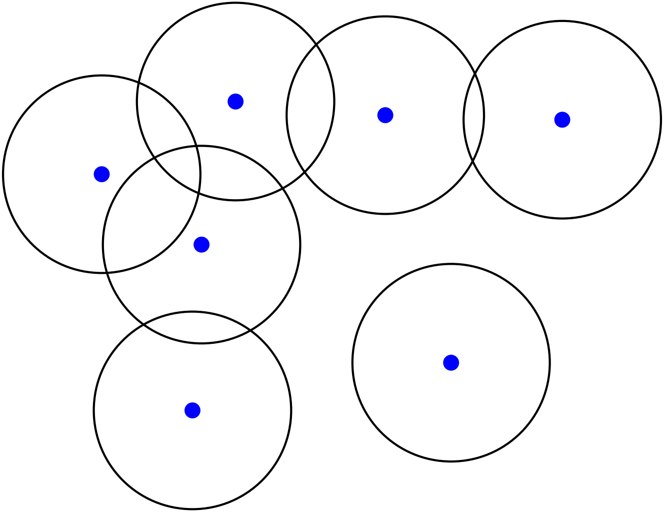
# Simplicial complexes from point data

The Čech construction



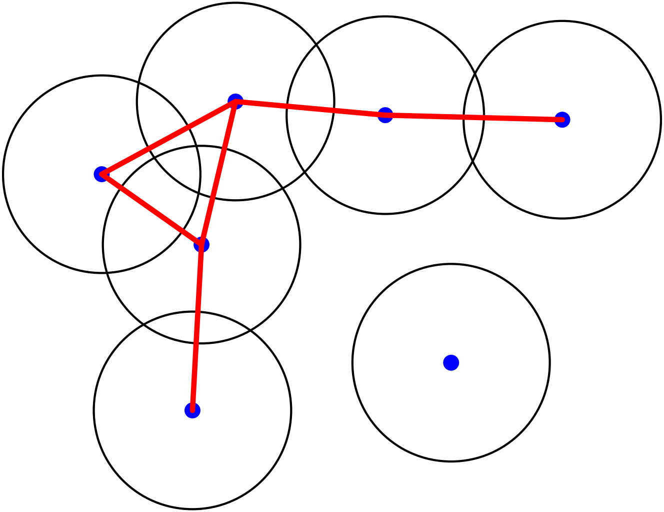
# Simplicial complexes from point data

The Čech construction



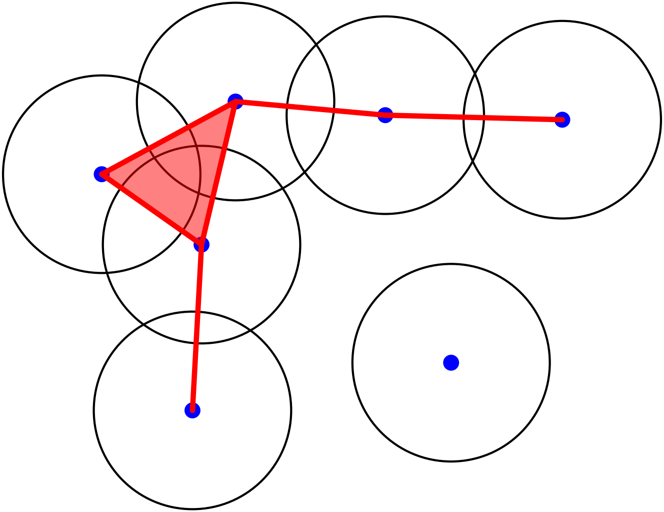
# Simplicial complexes from point data

The Čech construction



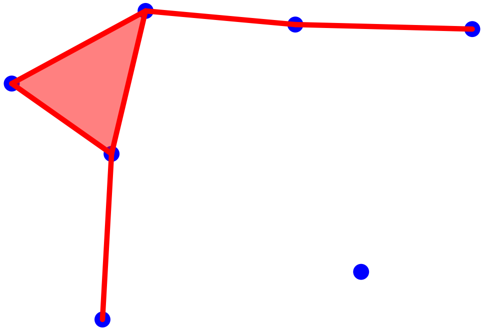
# Simplicial complexes from point data

The Čech construction



# Simplicial complexes from point data

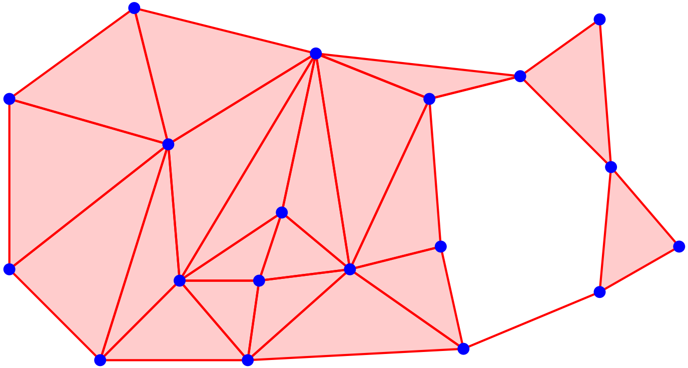
The Čech construction



# Homology of simplicial complexes

## Definition

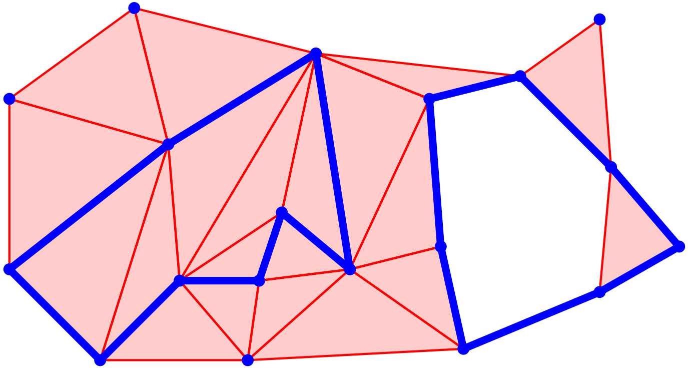
Homology in degree $k$ is given by $k$-cycles modulo the $k$-boundaries.



# Homology of simplicial complexes

## Definition

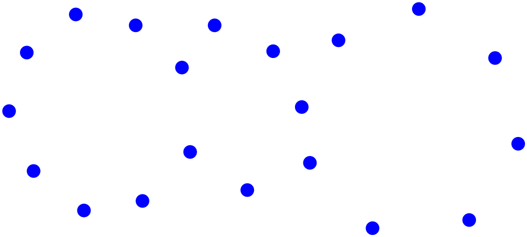
Homology in degree $k$ is given by $k$-cycles modulo the $k$-boundaries.



# Persistence

## Main idea

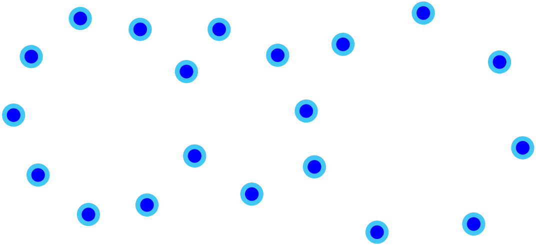
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

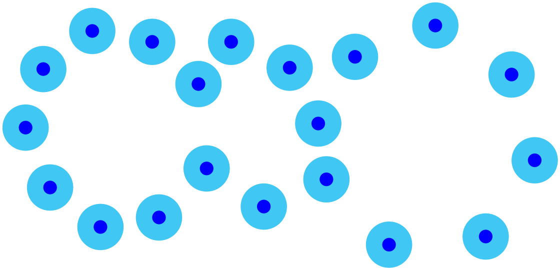
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

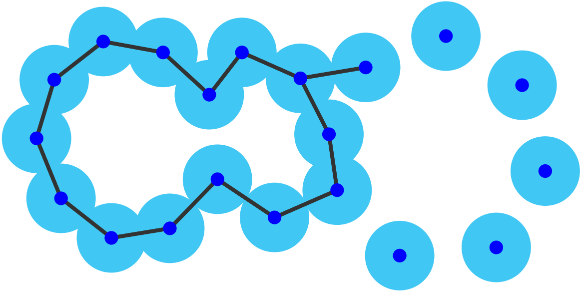
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

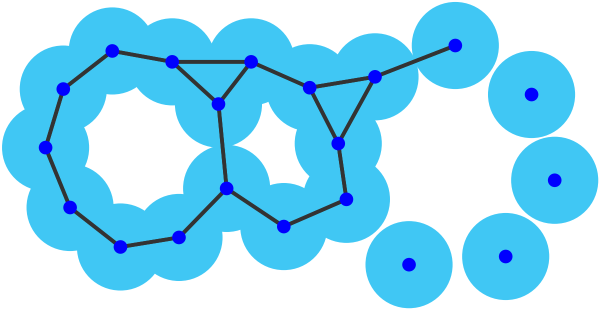
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

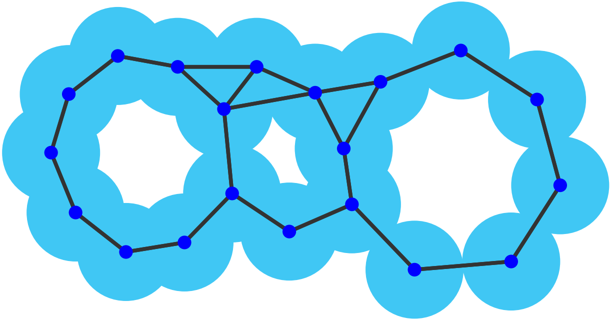
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

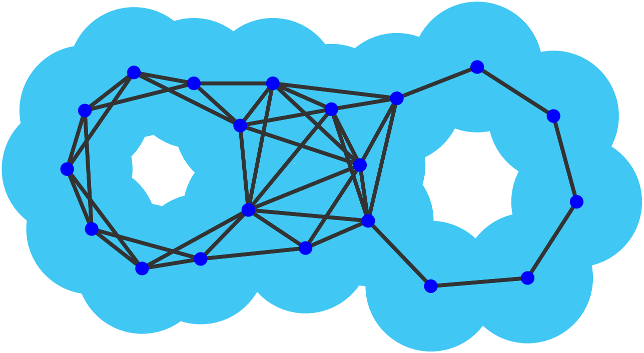
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

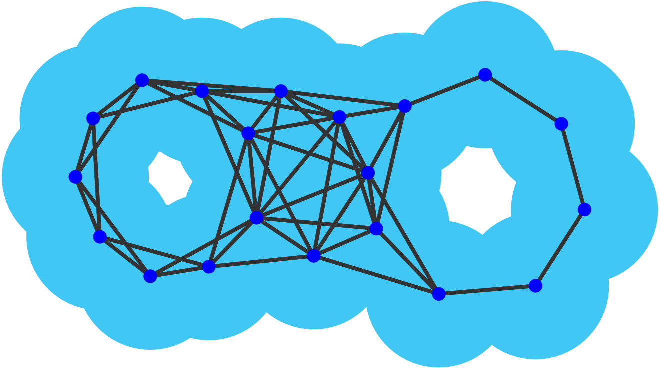
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

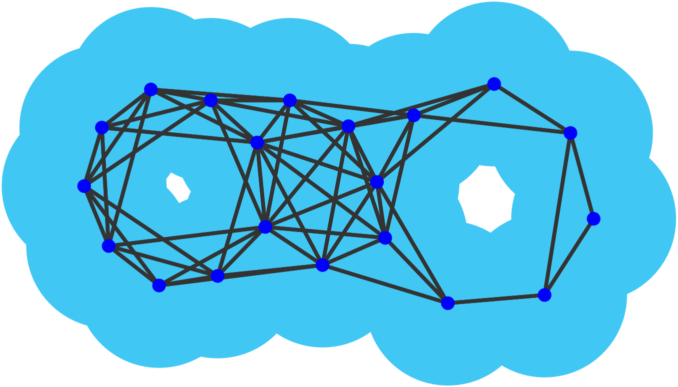
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

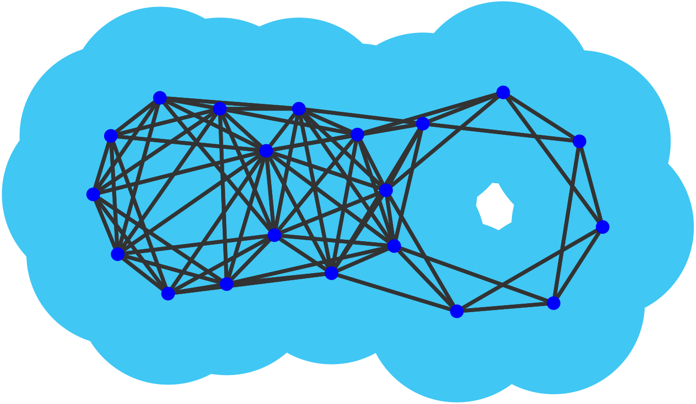
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

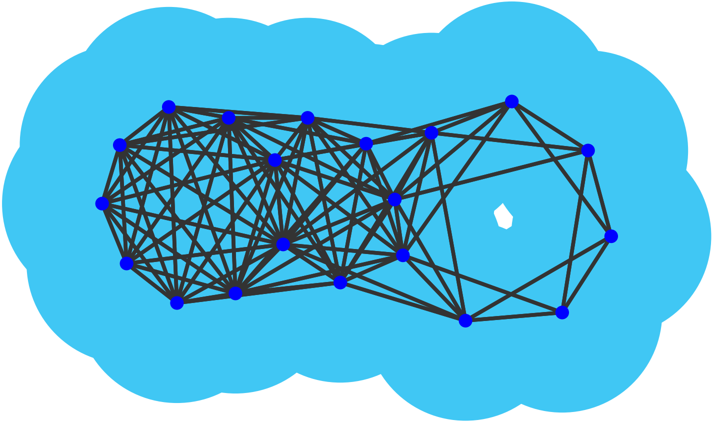
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

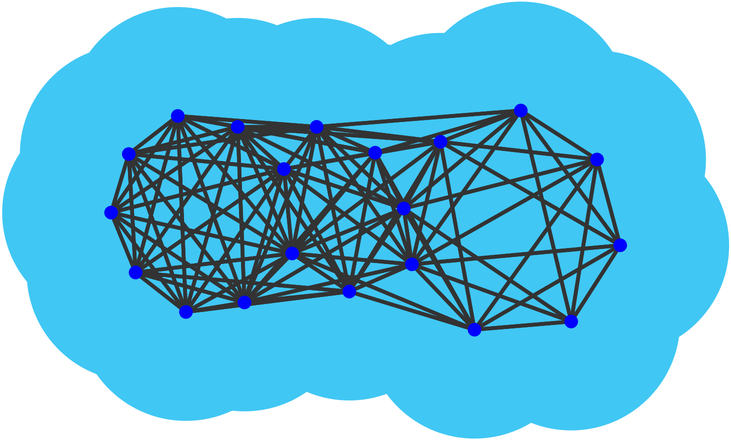
Vary a parameter and keep track of when features appear and disappear.



# Persistence

## Main idea

Vary a parameter and keep track of when features appear and disappear.



# Mathematical encoding

We have an increasing sequence of simplicial complexes

$$X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots \subseteq X_m$$

called a filtered simplicial complex.

Apply homology.

We get a sequence of vector spaces and linear maps

$$V_0 \to V_1 \to V_2 \to \cdots \to V_m$$

called a persistence module.

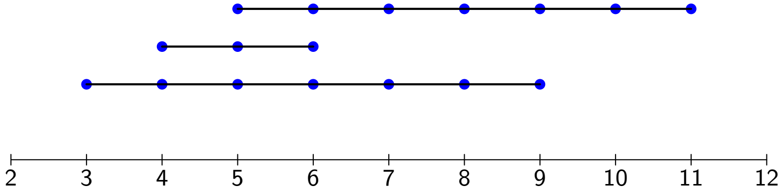
# Persistence module to Barcode

$$V_0 \to V_1 \to V_2 \to V_3 \to V_4 \to V_5 \to V_6 \to V_7 \to \cdots \to V_m$$

**Fundamental Theorem of Persistent Homology**

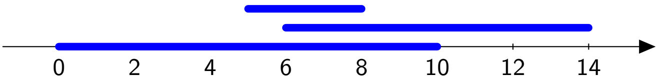
There exists a choice of bases for the vector spaces $V_i$ such that each map is determined by a bipartite matching of basis vectors.

Get a barcode:

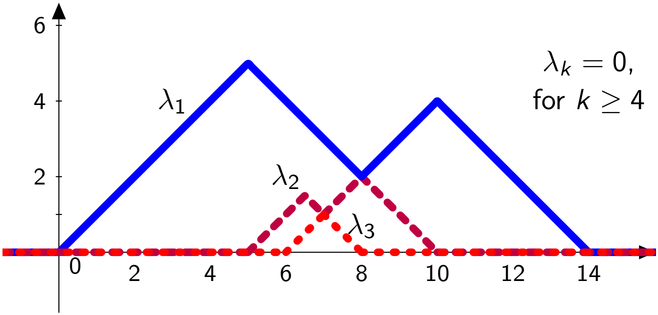


# Barcode to Persistence Landscape

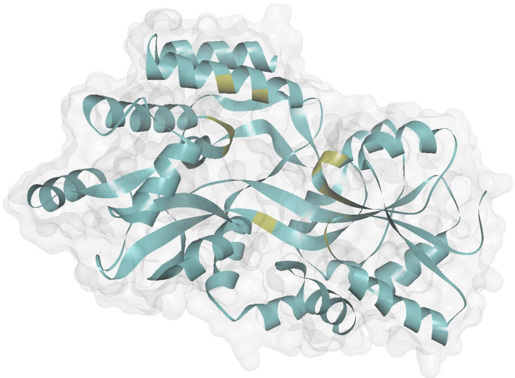
Barcode:




Persistence Landscape:

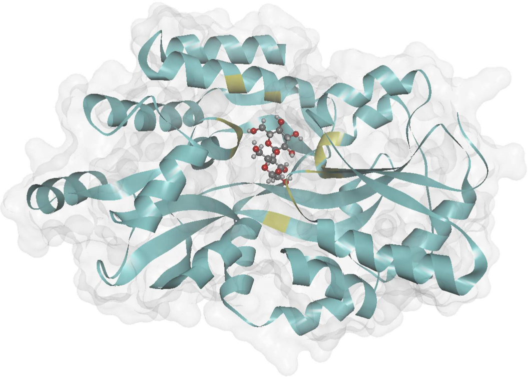


# Maltose Binding Protein, two conformations



V. Kovacev-Nikolic, P. Bubenik, D. Nikolic, and G. Heo. Using persistent homology and dynamical distances to analyze protein binding. Statistical Applications in Genetics and Molecular Biology, **15** (2016) no. 1, 19–38.

# Maltose Binding Protein, two conformations



V. Kovacev-Nikolic, P. Bubenik, D. Nikolic, and G. Heo. Using persistent homology and dynamical distances to analyze protein binding. Statistical Applications in Genetics and Molecular Biology, **15** (2016) no. 1, 19–38.

# Maltose Binding Protein Data

## The Data

Fourteen MBP structures from the Protein Data Bank.

- 7 closed conformations
- 7 open conformations

X-ray crystallography: coordinates of atoms.

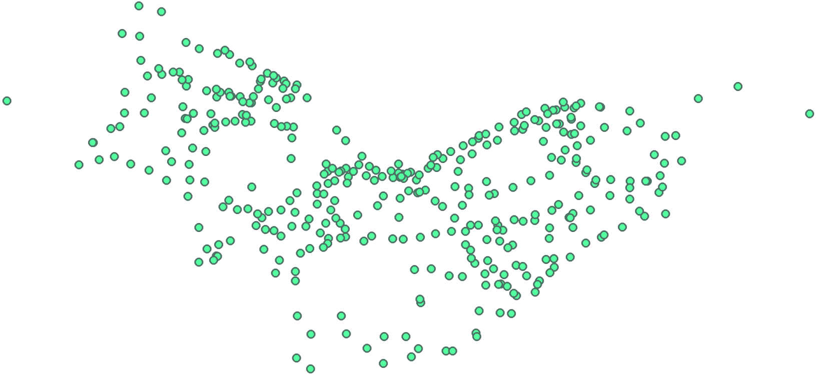
Represent each amino acid residue by its $C\alpha$ atom.

Have 14 sets of 370 points in $\mathbb{R}^3$.

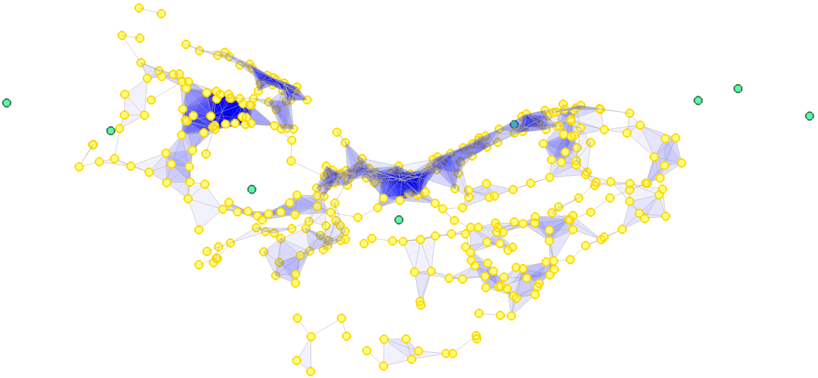
## The Goal

Can we use topological data analysis to distinguish the open and closed conformations?

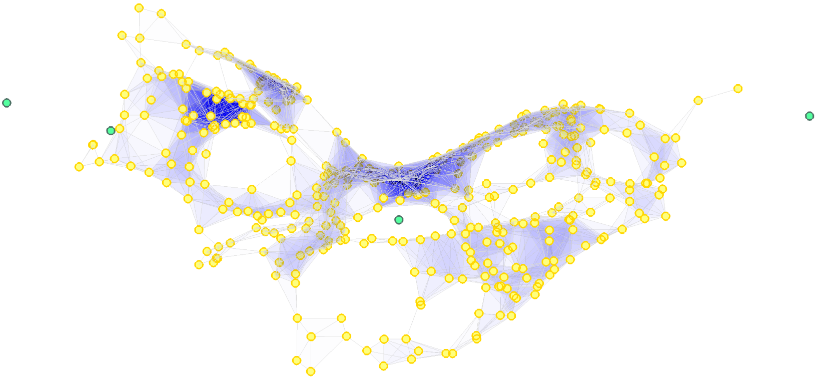
# Filtered simplicial complex



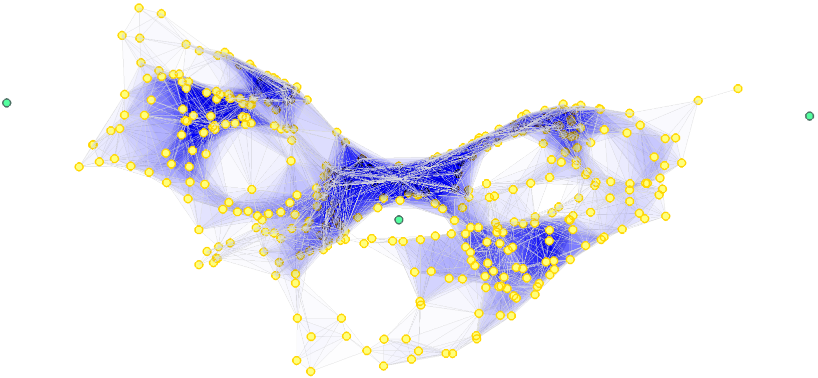
# Filtered simplicial complex



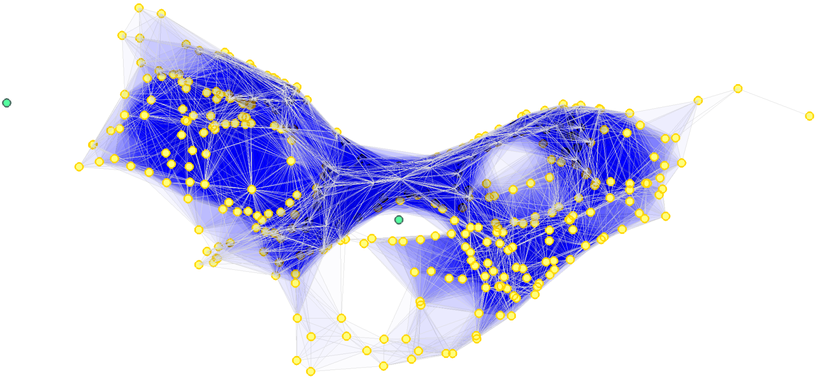
# Filtered simplicial complex



# Filtered simplicial complex

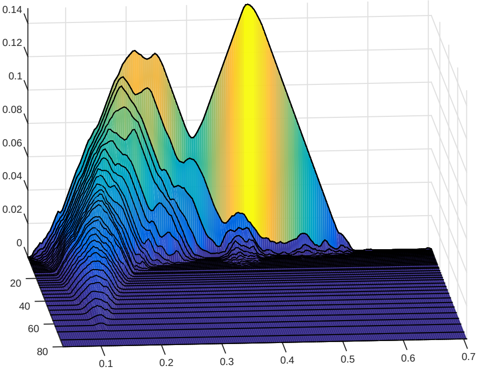


# Filtered simplicial complex

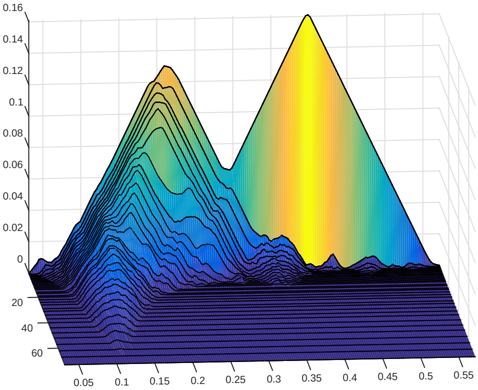


# Average persistence landscapes

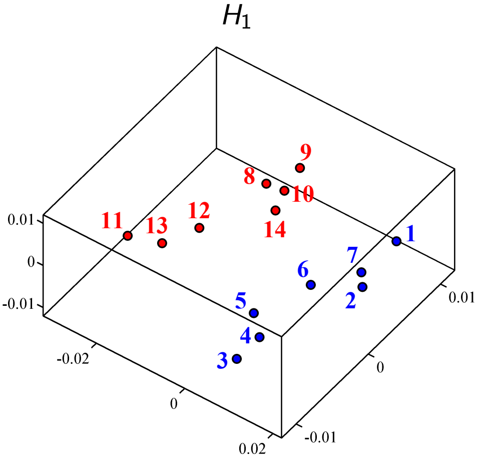
$H_1$ closed




$H_1$ open

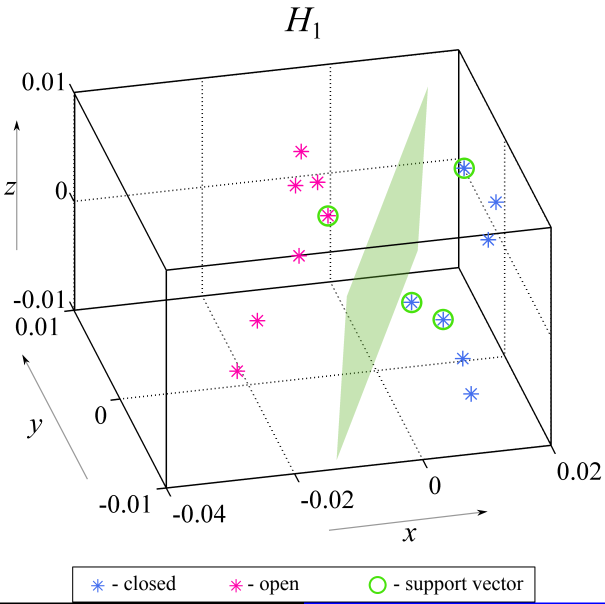


# Clustering of protein conformations




Projection of the $L^2$ distance matrix to $\mathbb{R}^3$ using Isomap.

# Classification of protein conformations



# Software

Persistent Homology software:

- JavaPlex
- PHAT, DIPHA
- Perseus
- Dionysus
- CHOMP
- GUDHI

Persistence Landscape software:

- The Persistence Landscape Toolbox
- the R package TDA

# Topological Data Analysis Summary