Applications

Dr. Elena Xinyi Wang

November 11, 2026

Persistence Landscape

Journal of Machine Learning Research 16 (2015) 77-102 Submitted 7/14; Published 1/15

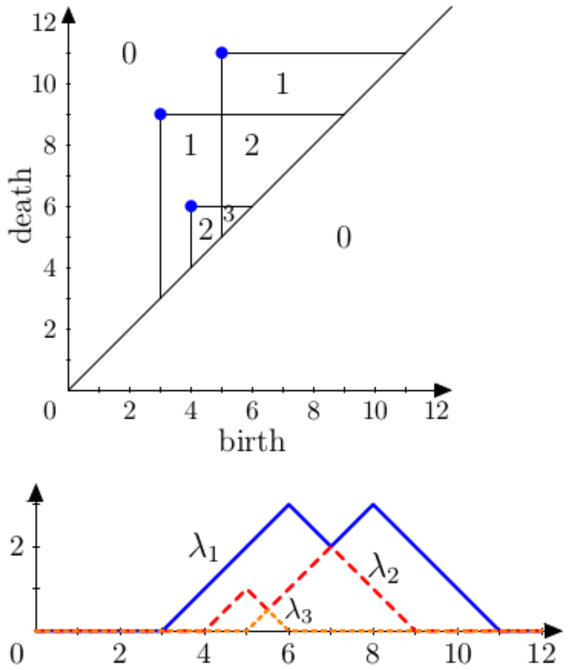
Statistical Topological Data Analysis using Persistence Landscapes

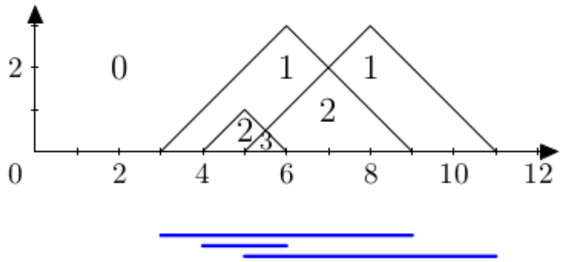
Peter Bubenik PETER.BUBENIK@GMAIL.COM
Department of Mathematics
Cleveland State University
Cleveland, OH 44115-2214, USA

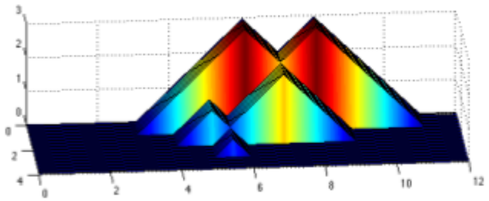
Editor: David Dunson

Abstract

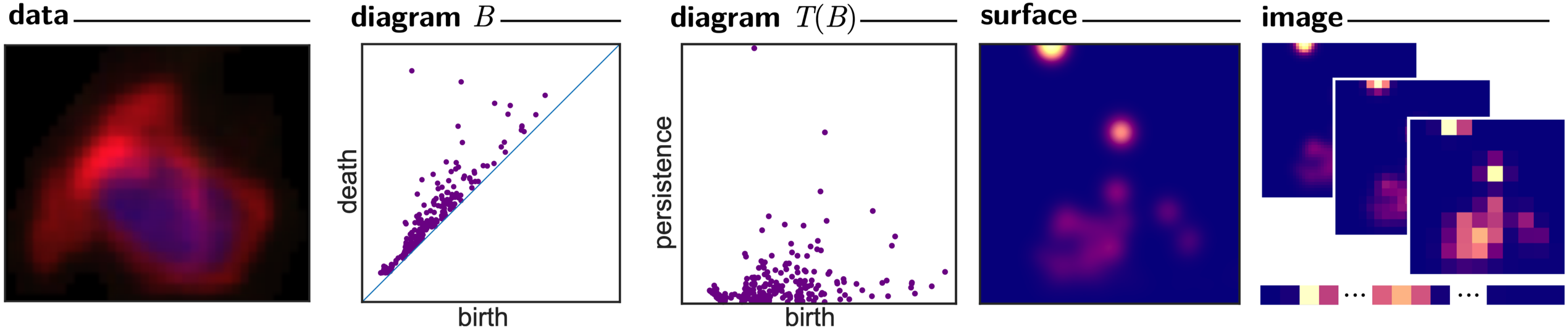
We define a new topological summary for data that we call the persistence landscape. Since this summary lies in a vector space, it is easy to combine with tools from statistics and machine learning, in contrast to the standard topological summaries. Viewed as a random variable with values in a Banach space, this summary obeys a strong law of large numbers and a central limit theorem. We show how a number of standard statistical tests can be used for statistical inference using this summary. We also prove that this summary is







Persistence Image



Porus Material

Topological Descriptors Help Predict Guest Adsorption in Nanoporous Materials

Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

Aditi S. Krishnapriyan,* Maciej Haranczyk, and Dmitriy Morozov

Cite This: *J. Phys. Chem. C* 2020, 124, 9360–9368

Read Online

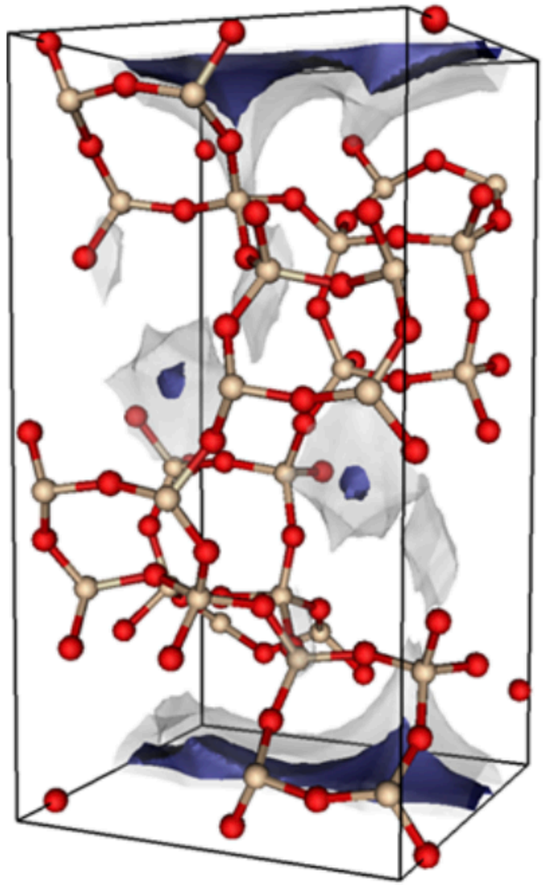
ACCESS | Metrics & More | Article Recommendations



ABSTRACT: Machine learning has emerged as an attractive alternative to experiments and simulations for predicting material properties. Usually, such an approach relies on specific domain knowledge for feature design: each learning target requires careful selection of features that an expert recognizes as important for the specific task. The major drawback of this approach is that computation of only a few structural features has been implemented so far, and it is difficult to tell a priori which features are important for a particular application. The latter problem has been empirically observed for predictors of guest uptake in nanoporous materials: local and global porosity features become dominant descriptors at low and high pressures, respectively. We investigate a feature representation of materials using tools from topological data analysis. Specifically, we use persistent homology to describe the geometry of nanoporous materials at various scales. We combine our topological descriptor with traditional structural features and investigate the relative importance of each to the prediction tasks. We demonstrate an application of this feature representation by predicting methane adsorption in zeolites, for pressures in the range 1−200 bar. Our results not only show a considerable improvement compared to the baseline, but they also highlight that topological features capture information complementary to the structural features. This is especially important for the adsorption at low pressure, a task particularly difficult for the traditional features. Furthermore, by investigation of the importance of individual topological features in the adsorption model, we are able to pinpoint the location of the pores that correlate best to adsorption at different pressure, contributing to our atom-level understanding of structure−property relationships.

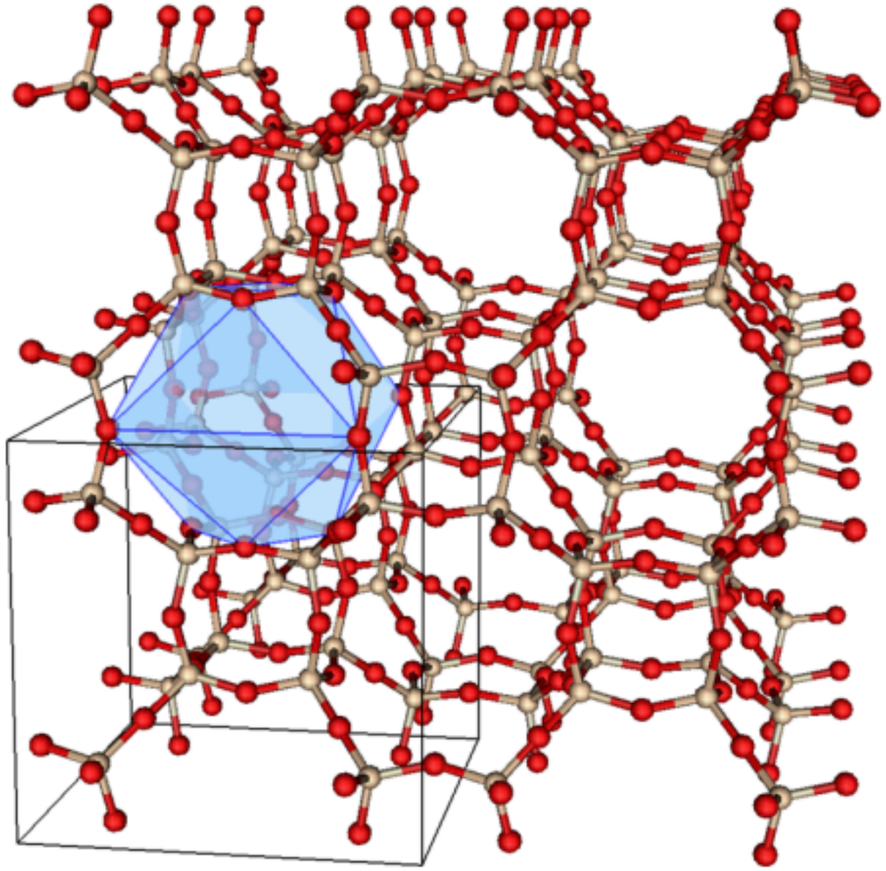
Problem setup

- Supervised regression: given a zeolite's crystal structure, how much methane it absorbs
  - Microscopic sponge that contains passages and pockets
  - Gas molecules can collect inside these spaces
- Input: Atomic coordinates of the periodic crystal
- Target: Methane adsorption loading at pressure p

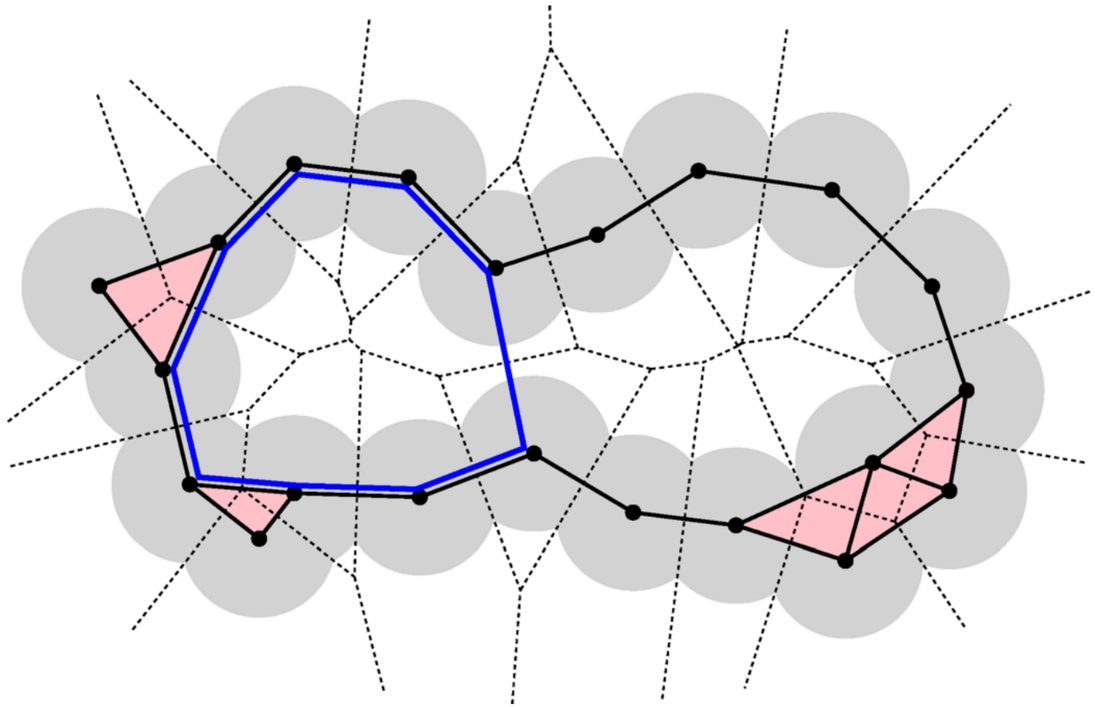


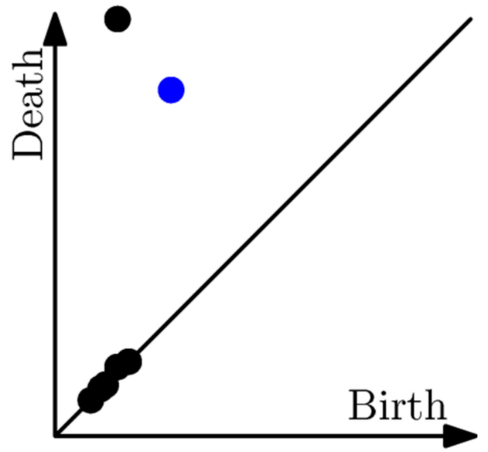
Measurements

- How much empty space is available?
- How much internal surface is available?
- How wide are the passages and pockets?
- **Could these measurements miss important details?**



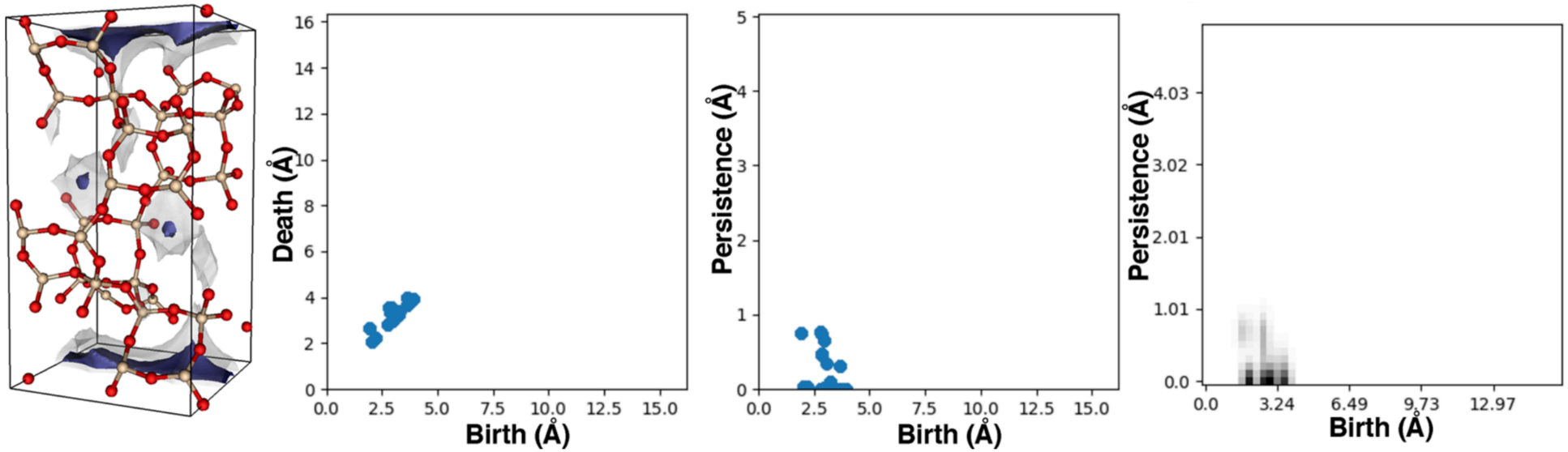
Alpha Complex



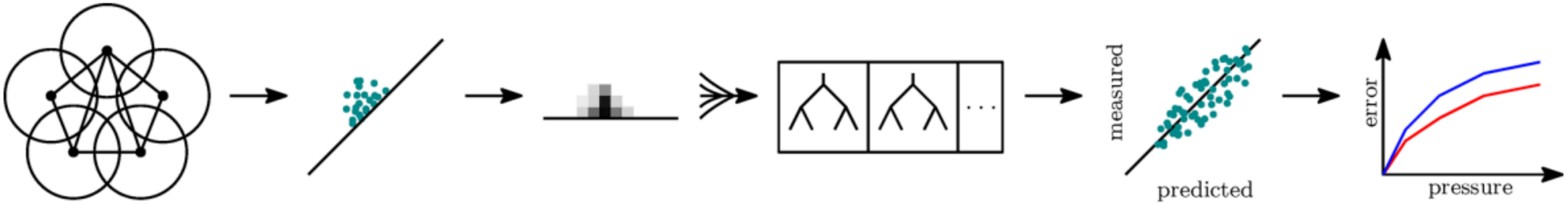


Persistence Images

- Grow neighborhoods around the atoms
- Track persistence via Alpha complexes

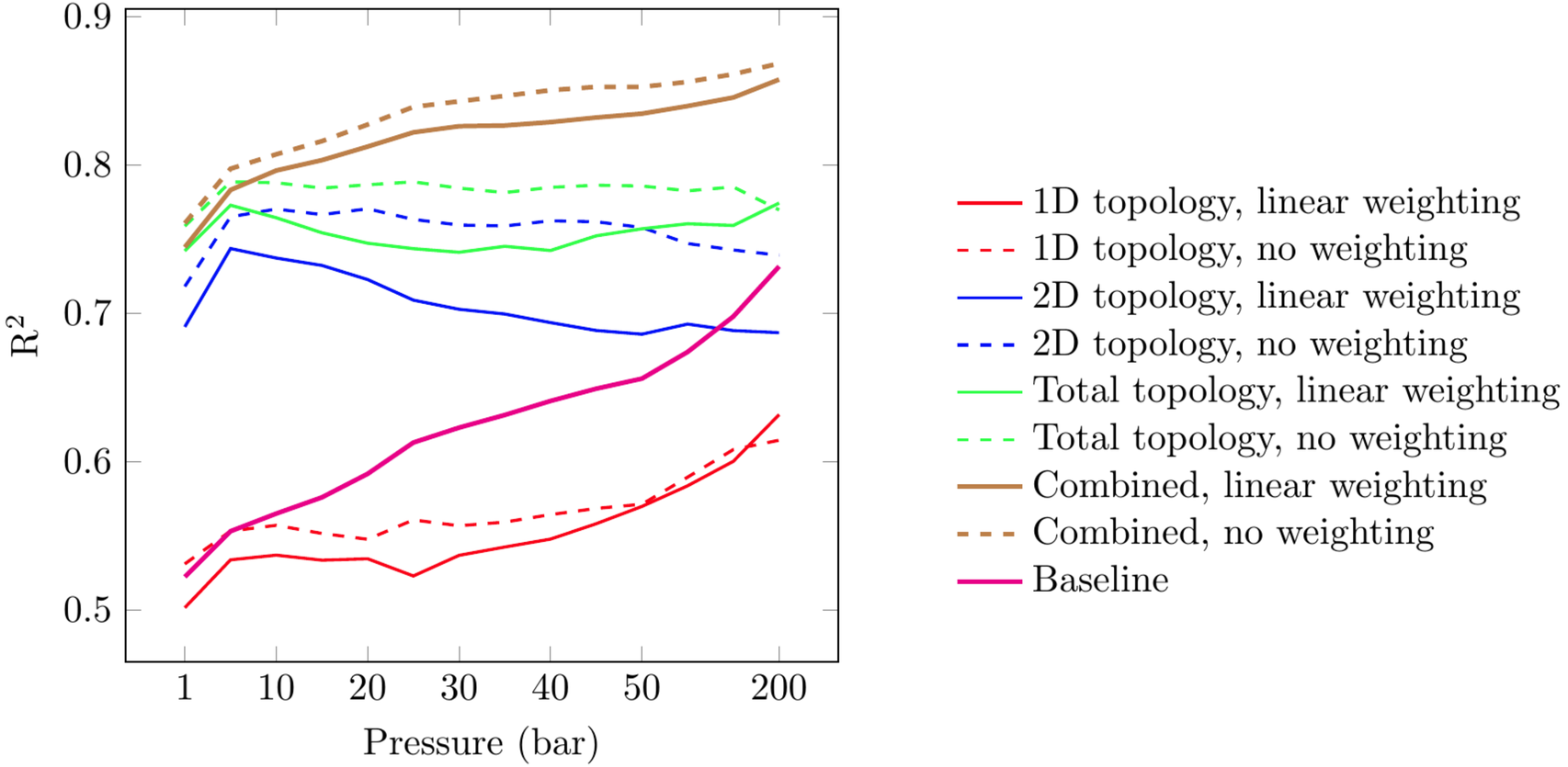


The pipeline



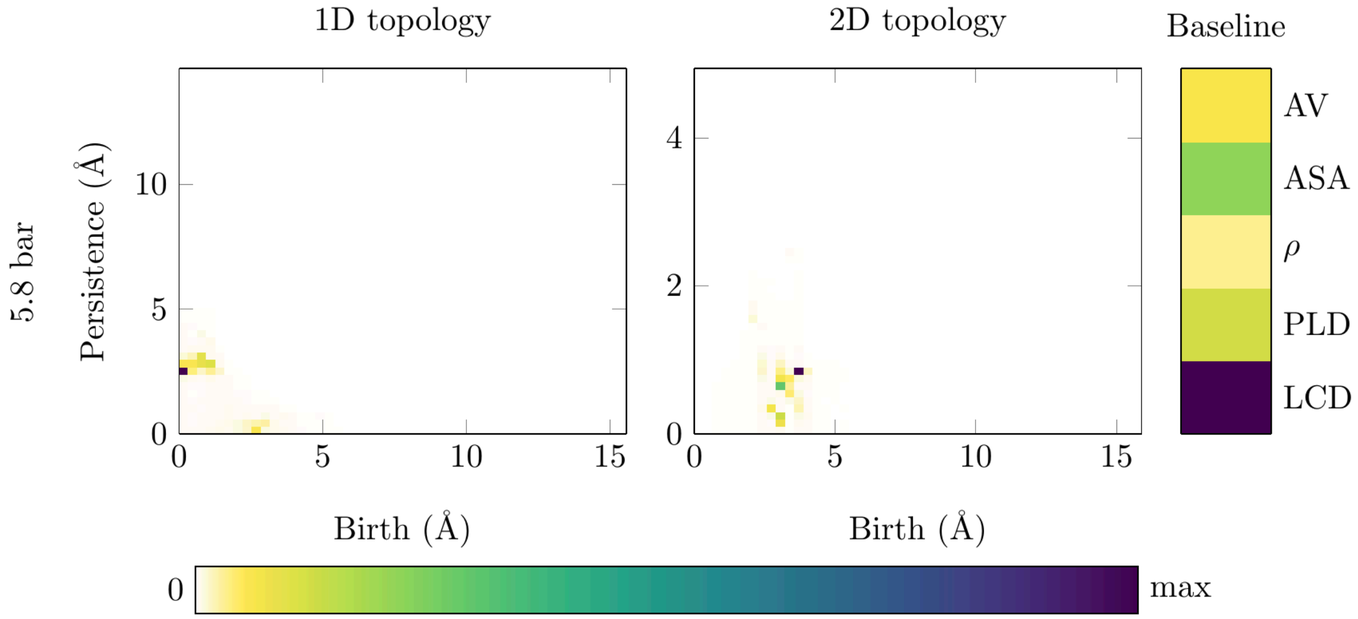
- Conventional measurements: accessible volume, surface area, density, pore-limiting diameter, largest-cavity diameter
- H_1 and H_2 persistence image
- Combine features
- Optional weighting (emphasize high-persistence regions)

Results


0.9
0.8
0.7
0.6
0.5
R^2
1
10
20
30
40
50
200
Pressure (bar)
1D topology, linear weighting
1D topology, no weighting
2D topology, linear weighting
2D topology, no weighting
Total topology, linear weighting
Total topology, no weighting
Combined, linear weighting
Combined, no weighting
Baseline

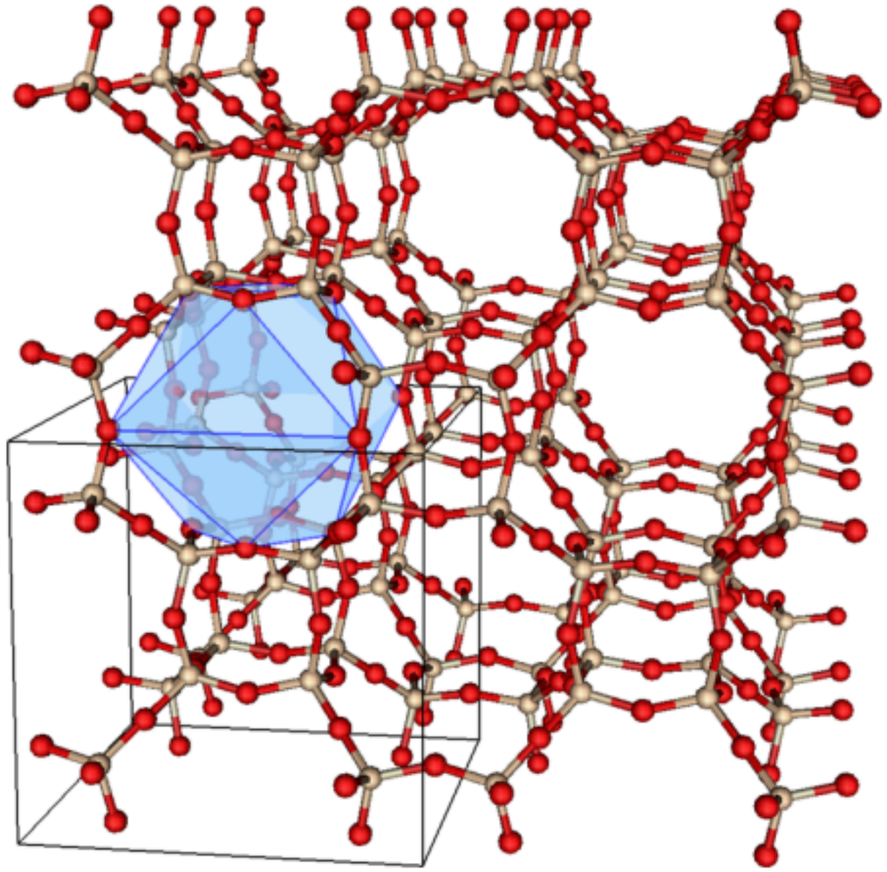
Which features does the model use?

- Left: importance of pixels in the H_1 persistence image
- Middle: importance of pixels in the H_2 persistence image
- Right: importance of the five conventional measurements (LCD: largest cavity diameter)
- Can we trace an important feature back to an actual cavity in the material?



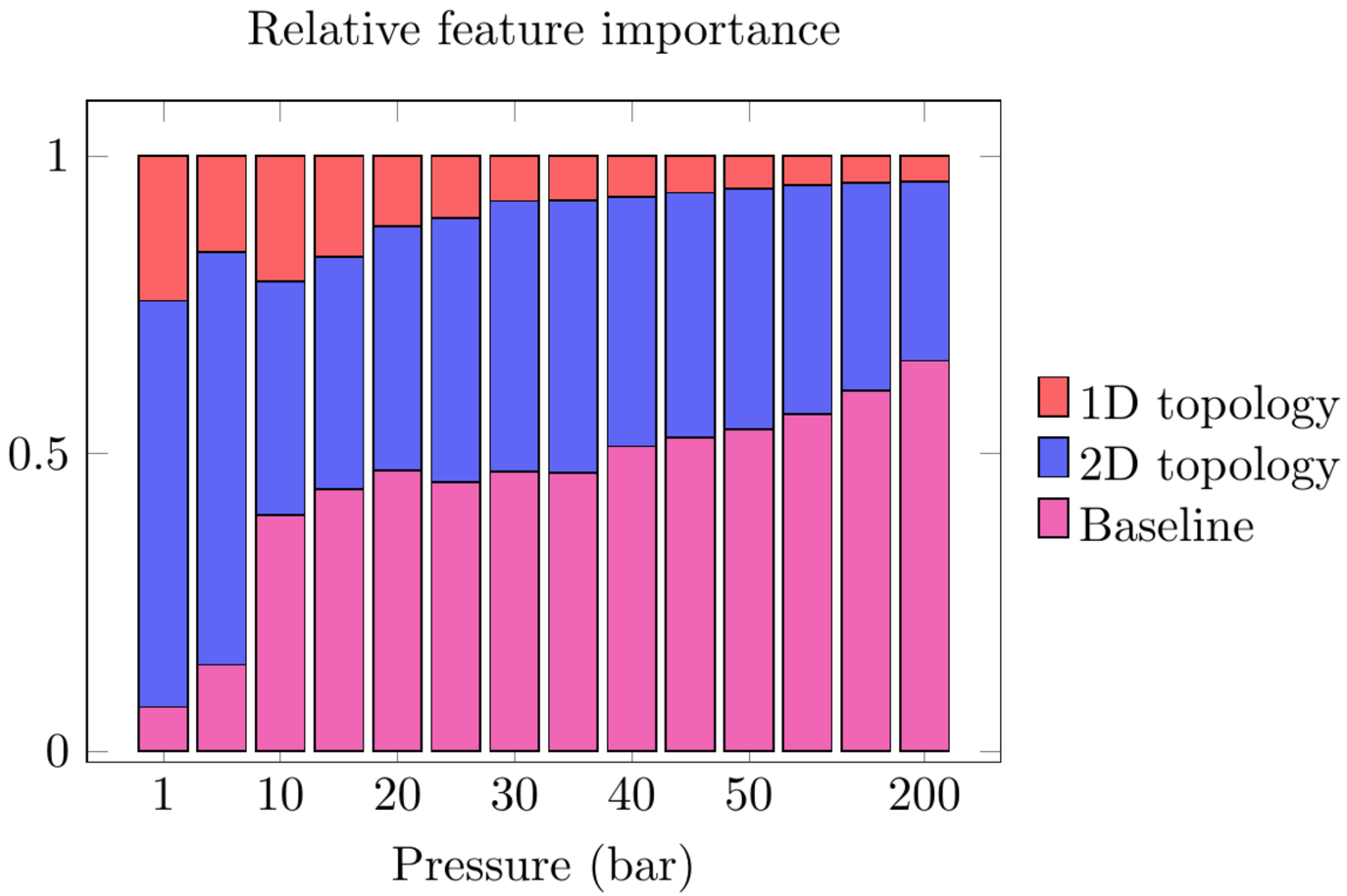
Predictive Feature to Cavity

- The atomic framework: part of the zeolite called ACO, which has high methane adsorption
- The blue-highlighted surface: a representative H_2 cycle surrounding a cavity
- What about this cavity's geometry might help the material hold methane?
- Does the useful structural information stay the same when we change the gas pressure?



What does pressure change?

- Feature groups contribute differently across pressures.
- Topological and conventional descriptions play different roles.



Return to materials

- Can structure help predict gas uptake?
- Does topology add information beyond familiar measurements?
- What other hypotheses can we test next?

Protein Binding

Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolić and Giseon Heo*

Using persistent homology and dynamical distances to analyze protein binding

DOI 10.1515/sagmb-2015-0057

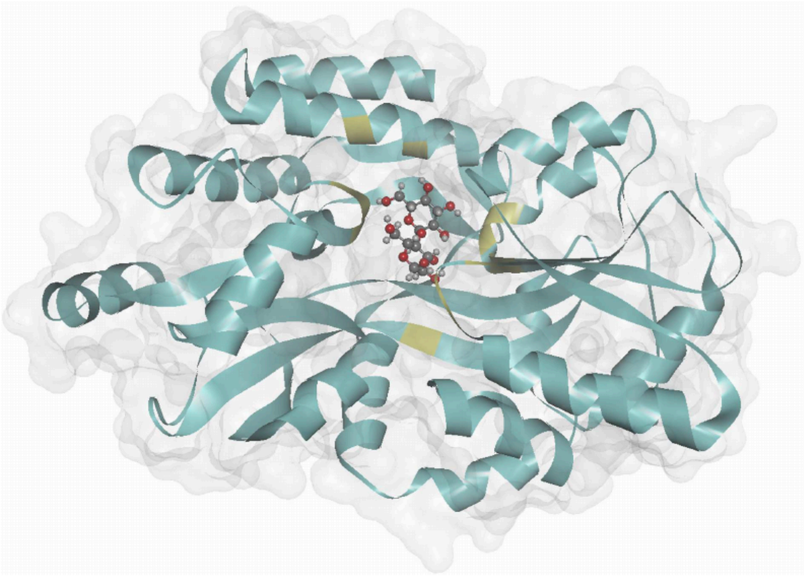
Abstract: Persistent homology captures the evolution of topological features of a model as a parameter changes. The most commonly used summary statistics of persistent homology are the barcode and the persistence diagram. Another summary statistic, the persistence landscape, was recently introduced by Bubenik. It is a functional summary, so it is easy to calculate sample means and variances, and it is straightforward to construct various test statistics. Implementing a permutation test we detect conformational changes between closed and open forms of the maltose-binding protein, a large biomolecule consisting of 370 amino acid residues. Furthermore, persistence landscapes can be applied to machine learning methods. A hyperplane from a support vector machine shows the clear separation between the closed and open proteins conformations. Moreover, because our approach captures dynamical properties of the protein our results may help in identifying residues susceptible to ligand binding; we show that the majority of active site residues and allosteric pathway residues are located in the vicinity of the most persistent loop in the corresponding filtered Vietoris-Rips complex. This finding was not observed in the classical anisotropic network model.

Keywords: dynamical distance; persistence landscape; persistent homology; simplicial complex; support vector machine.

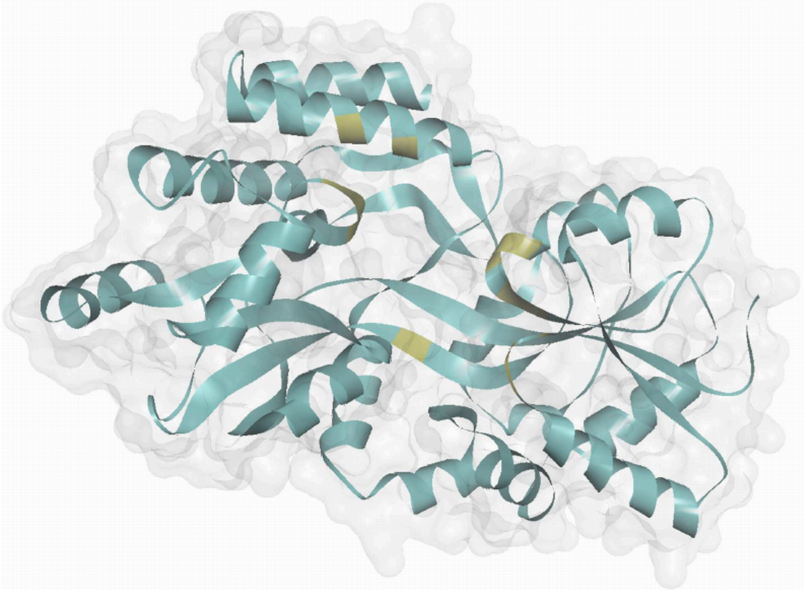
Proteins move as they work

- Proteins are flexible structures
- Maltose-binding protein adopts open and closed forms
- How can we describe this change quantitatively?

Closed + sugar (1MPD)



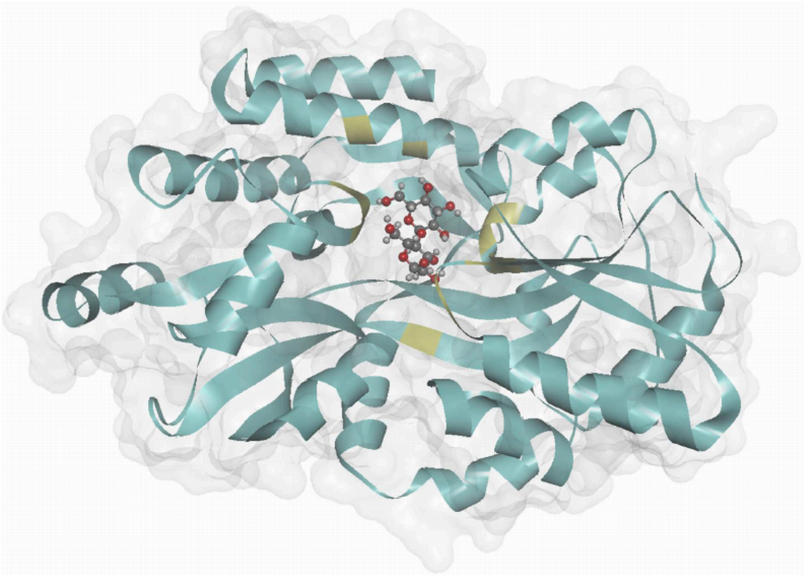
Open, no sugar (1OMP)



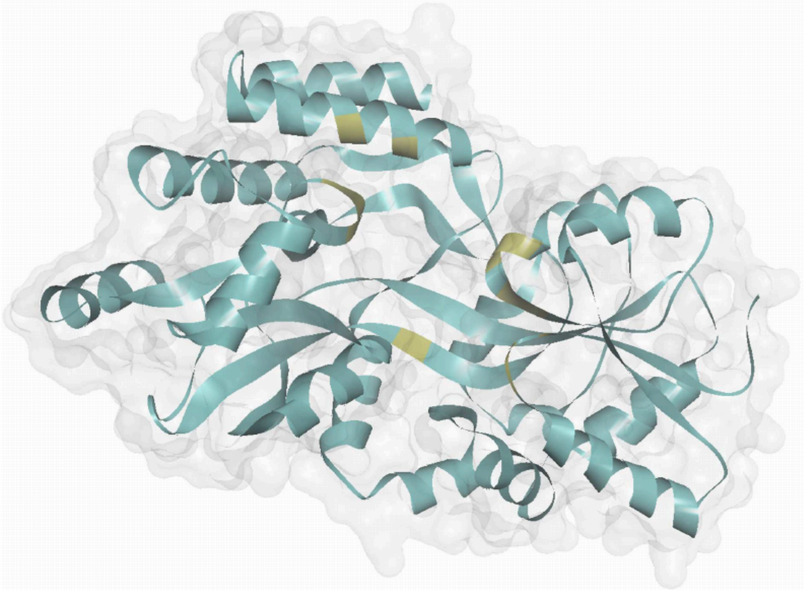
What would we like to know

- Do the two forms organize motion differently?
- Can we detect a consistent difference?
- Could we recognize a new structure's group?

Closed + sugar (1MPD)



Open, no sugar (1OMP)



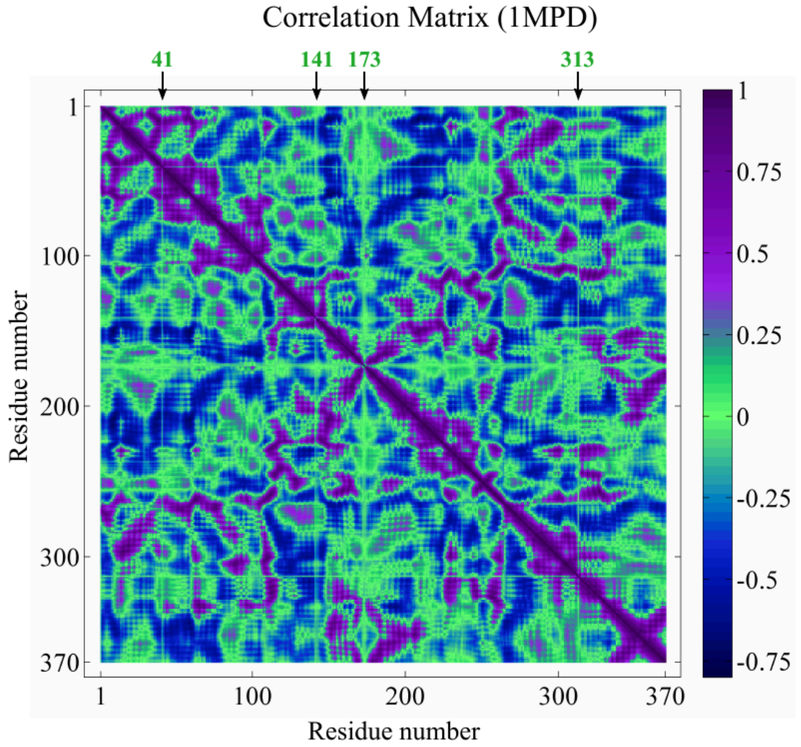
The data

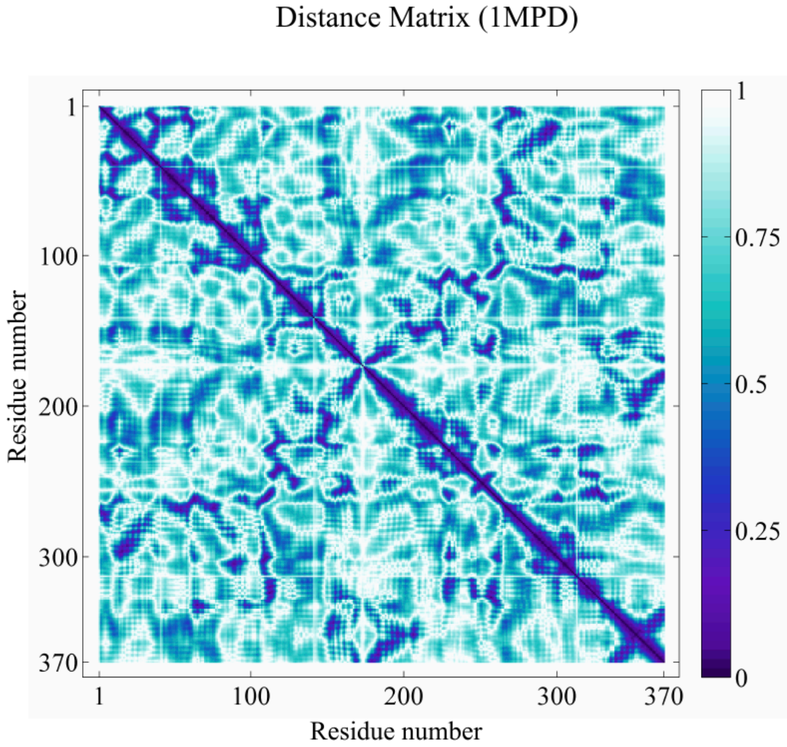
- 7 open structures
- 7 closed structures
- One structure gives one topological summary

No.	PDB code	Ligand name	Protein structure
1	1ANF	maltose	closed-*holo*
2	1FQC	maltotriotol	closed-*holo*
3	1FQD	maltotetraitol	closed-*holo*
4	1MPD	maltose	closed-*holo*
5	3HPI	sucrose	closed-*holo*
6	3MBP	maltotriose	closed-*holo*
7	4MBP	maltotetraose	closed-*holo*
8	1EZ9	maltotetraitol	open-*holo*
9	1FQA	maltotetraitol	open-*holo*
10	1FQB	maltotetraitol	open-*holo*
11	1JW4	-	open-*apo*
12	1JW5	maltose	open-*holo*
13	1LLS	-	open-*apo*
14	1OMP	-	open-*apo*

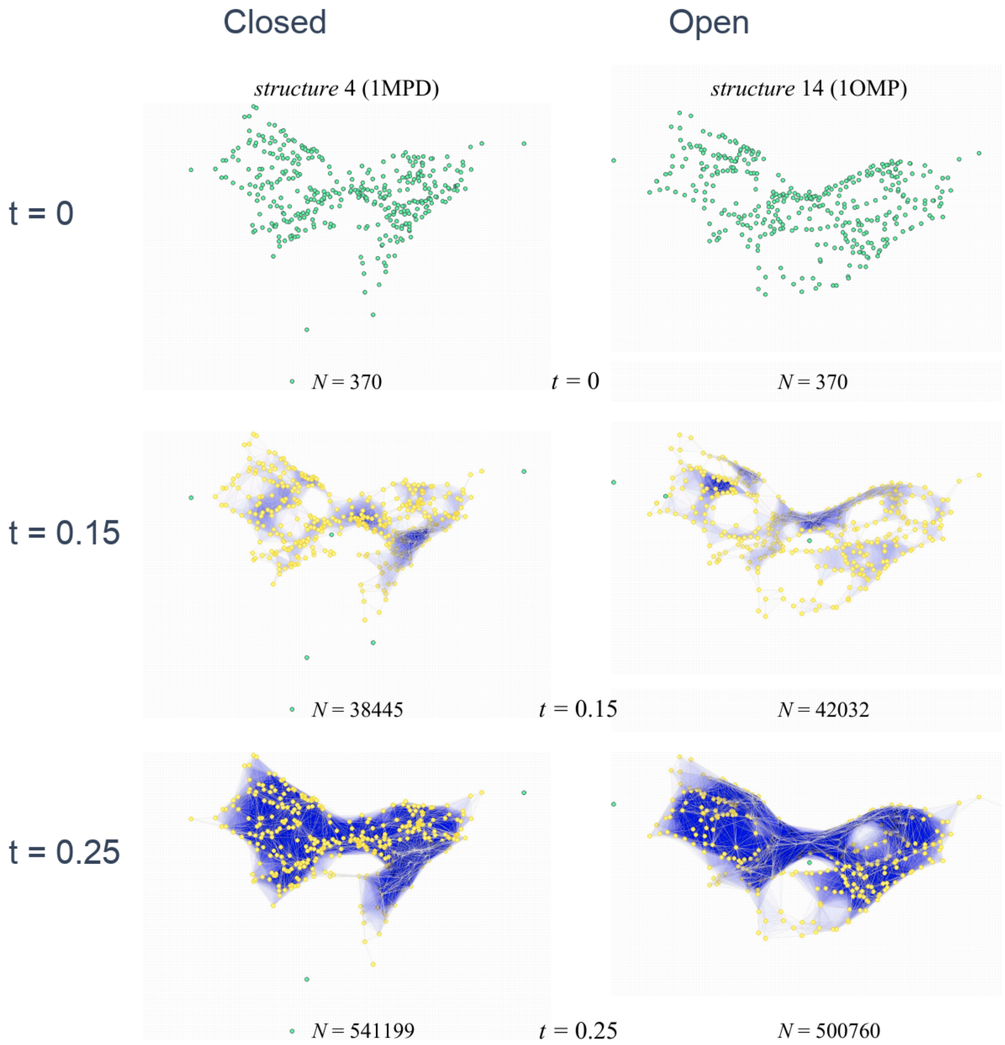
Notion of Closeness based on Motion

- Model how residues move together (370 residues)
- Convert correlations into dissimilarities
- Strongly correlated or anticorrelated residues are close

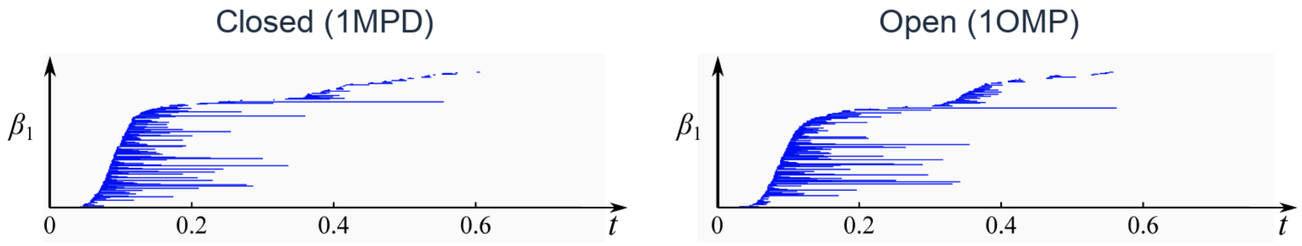




Coordinated motion to persistence

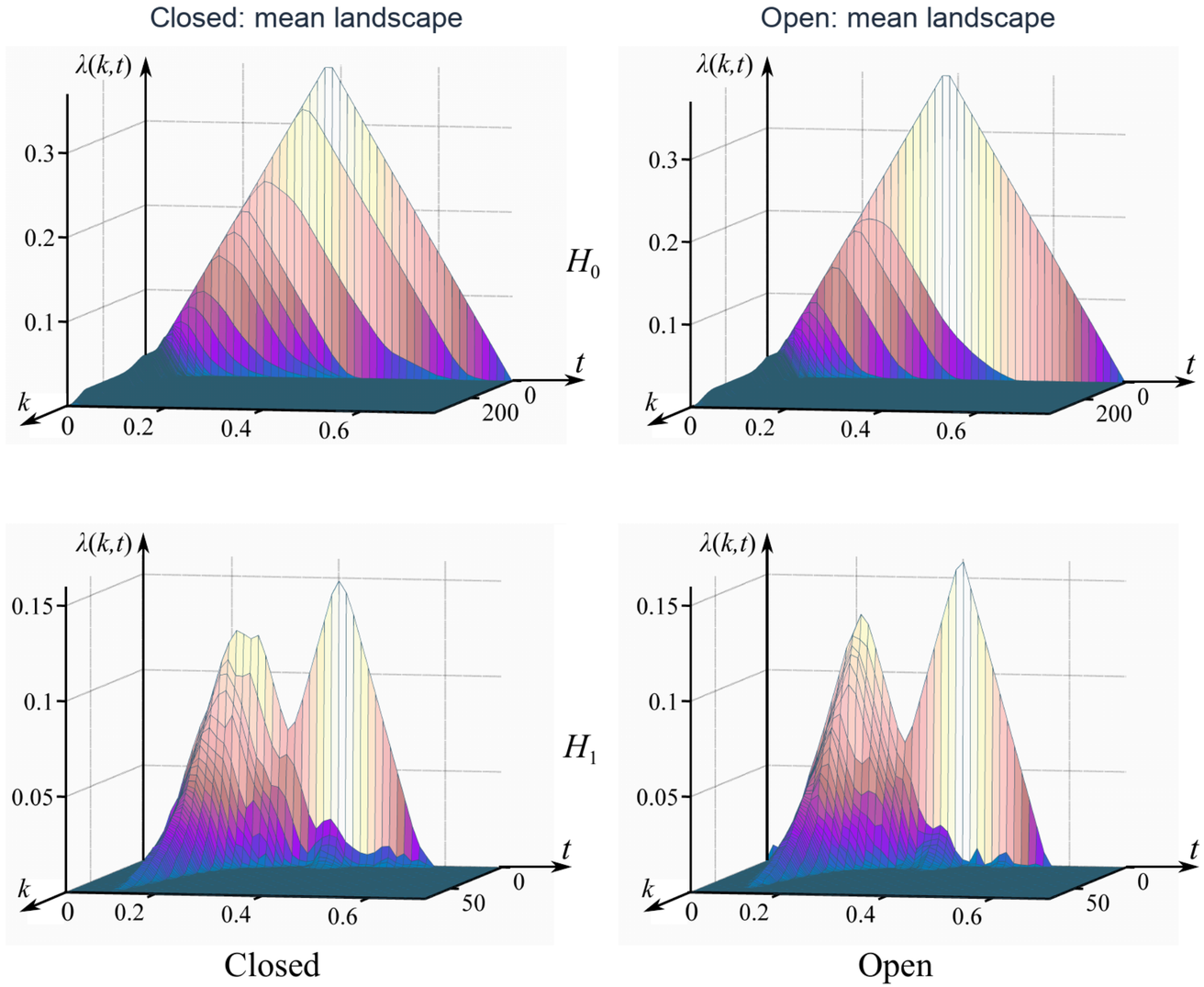


- Motion dissimilarities determine the filtration
- Compute the 0- and 1-dimensional barcodes
- Each structure produces a barcode
- Then compute the landscape



Average Landscape

- Individual barcodes might look similar
- Averaging makes group-level patterns available for comparison
- Notice the differences in small peaks in H_1
- Is this a consistent group difference? Or could it arise from variation among these fourteen structures?

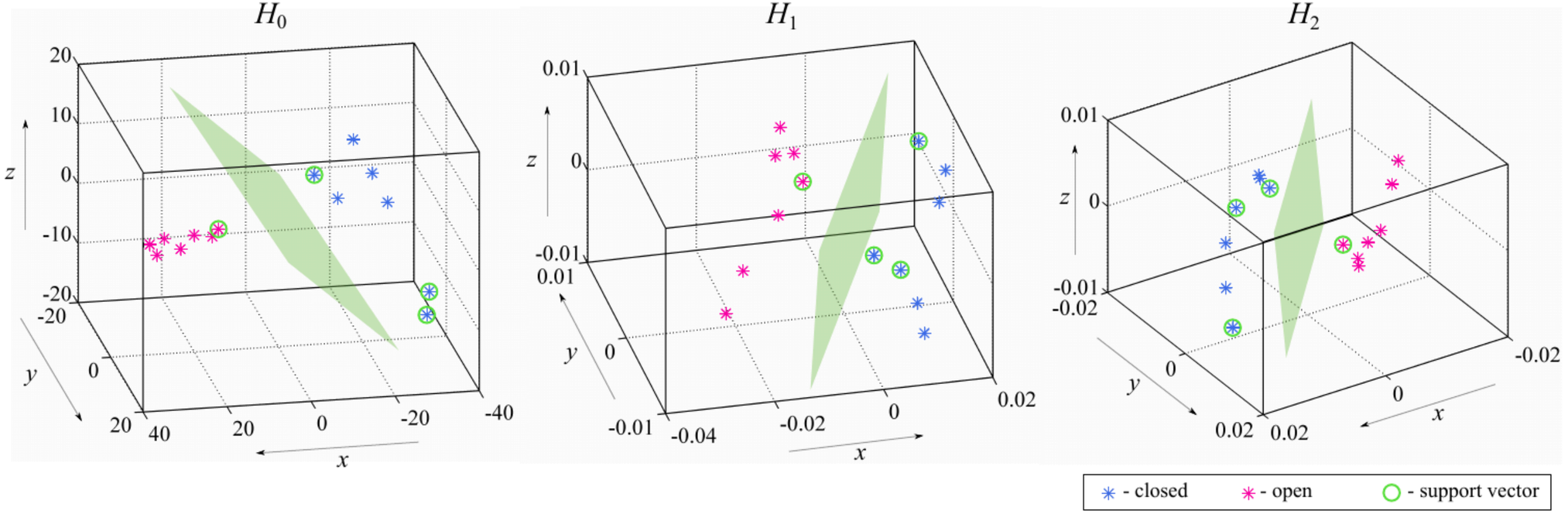


Test a specific difference

- Summarize each landscape by its total area
- Compare these values between groups
- Assess the difference using label permutations
- Result: open and closed structures differ in both H_0 and H_1

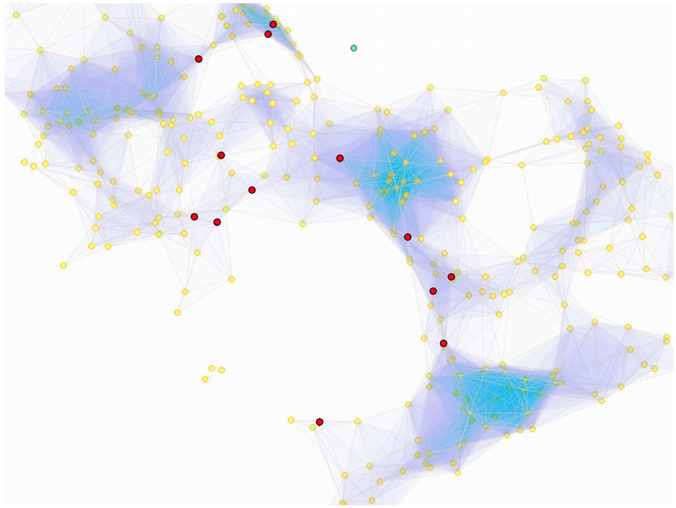
Linear Classification

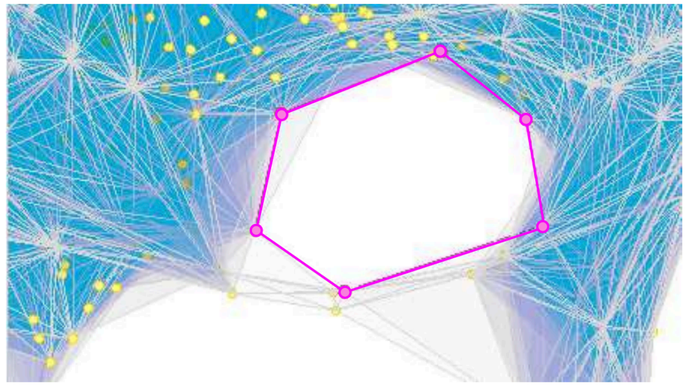
- 3 principal components on persistence landscape
- Fit an SVM
- Limitation: all 14 structures were used for training; no cross-validation or held-out evaluation



What does persistence tell us?

- Red dots: residues that participate in sugar binding.
  - For structure 1MPD, 10 of the 13 lie near the longest-lived loop
  - The other three lie near the second-longest-lived loop
- Polygon: representative cycle for the most persistent H_1 feature; identifies specific residues forming a representative of that feature
- Persistent features may help prioritize residues for further investigation





Neuroscience

Clique topology reveals intrinsic geometric structure in neural correlations

Chad Giusti[a,b], Eva Pastalkova[c], Carina Curto[b,d,1], and Vladimir Itskov[b,d,1,2]

[a]Warren Center for Network and Data Science, Departments of Bioengineering and Mathematics, University of Pennsylvania, Philadelphia, PA 19104; [b]Department of Mathematics, University of Nebraska, Lincoln, NE 68588; [c]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147; and [d]Department of Mathematics, The Pennsylvania State University, University Park, PA 16802

Edited by William Bialek, Princeton University, Princeton, NJ, and approved September 23, 2015 (received for review April 28, 2015)

Detecting meaningful structure in neural activity and connectivity data is challenging in the presence of hidden nonlinearities, where traditional eigenvalue-based methods may be misleading. We introduce a novel approach to matrix analysis, called clique topology, that extracts features of the data invariant under nonlinear monotone transformations. These features can be used to detect both random and geometric structure, and depend only on the relative ordering of matrix entries. We then analyzed the activity of pyramidal neurons in rat hippocampus, recorded while the animal was exploring a 2D environment, and confirmed that our method is able to detect geometric organization using only the intrinsic pattern of neural correlations. Remarkably, we found similar results during nonspatial behaviors such as wheel running and rapid eye movement (REM) sleep. This suggests that the geometric structure of correlations is shaped by the underlying hippocampal circuits and is not merely a consequence of position coding. We propose that clique topology is a powerful new tool for matrix analysis in biological settings, where the relationship of observed quantities to more meaningful variables is often nonlinear and unknown.

structure of neural correlation | neural coding | Betti curves | clique topology | topological data analysis

should be invariant under matrix transformations of the following form:

$$C_{ij} = f(A_{ij}), \qquad [1]$$

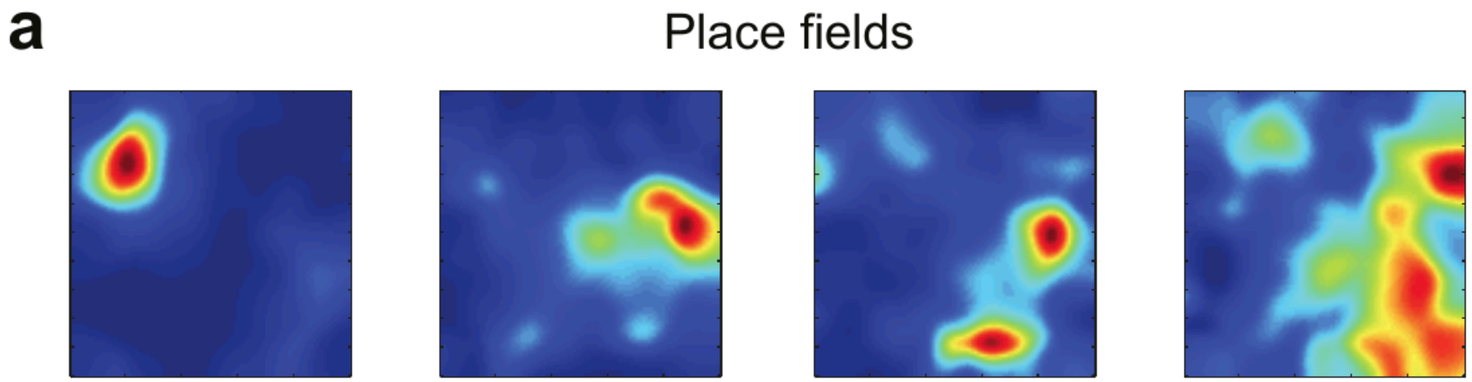
where f is a nonlinear monotonic function (Fig. 1*A*). In the case of hippocampal place cells, f captures the manner in which pairwise correlations C_{ij} decrease with distance between place field centers (3). In less studied contexts, the represented stimuli—and the function f—may be completely unknown.

Unfortunately, eigenvalues are not invariant under transformations of the form (Eq. **1**) (Fig. 1*B* and *SI Appendix*, Fig. S1). Although large random matrices have a reliable eigenvalue spectrum [e.g., Wigner's semicircle law (4)], it is possible that a random matrix with independent and identically distributed (i.i.d.) entries could be mistaken as structured, purely as an artifact of a monotonic nonlinearity (Fig. 1*B*).* The results of eigenvalue-based analyses can thus be difficult to interpret, and potentially misleading.

Here, we introduce a new tool to reliably detect signatures of structure and randomness that are invariant under nonlinear monotone transformations of the form (Eq. **1**). Using pairwise correlations of hippocampal place cells recorded during both spatial and nonspatial behaviors, we demonstrate that our method is

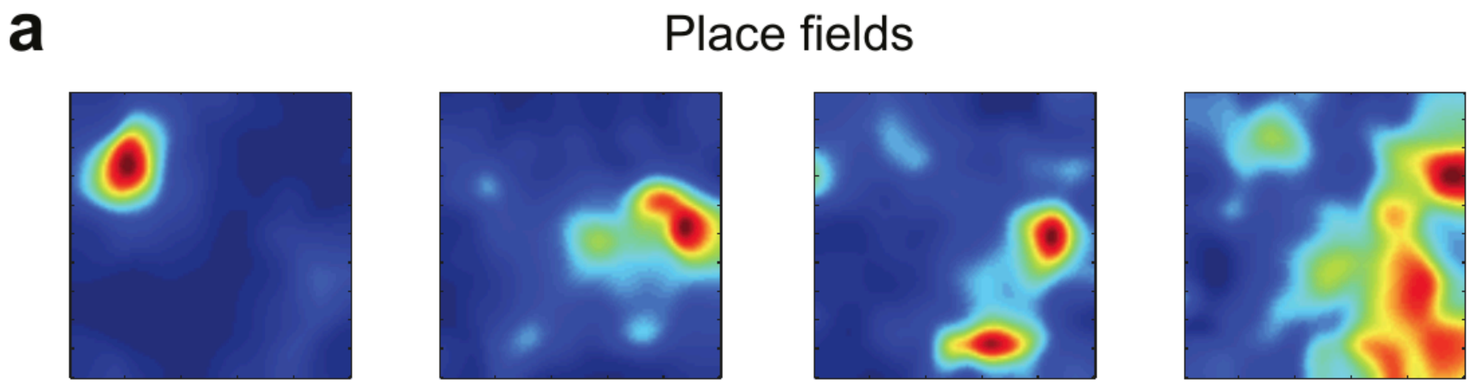
Place fields

- Some neurons prefer particular locations
- These preferred locations are called place fields
- Can activity alone reveal organization?



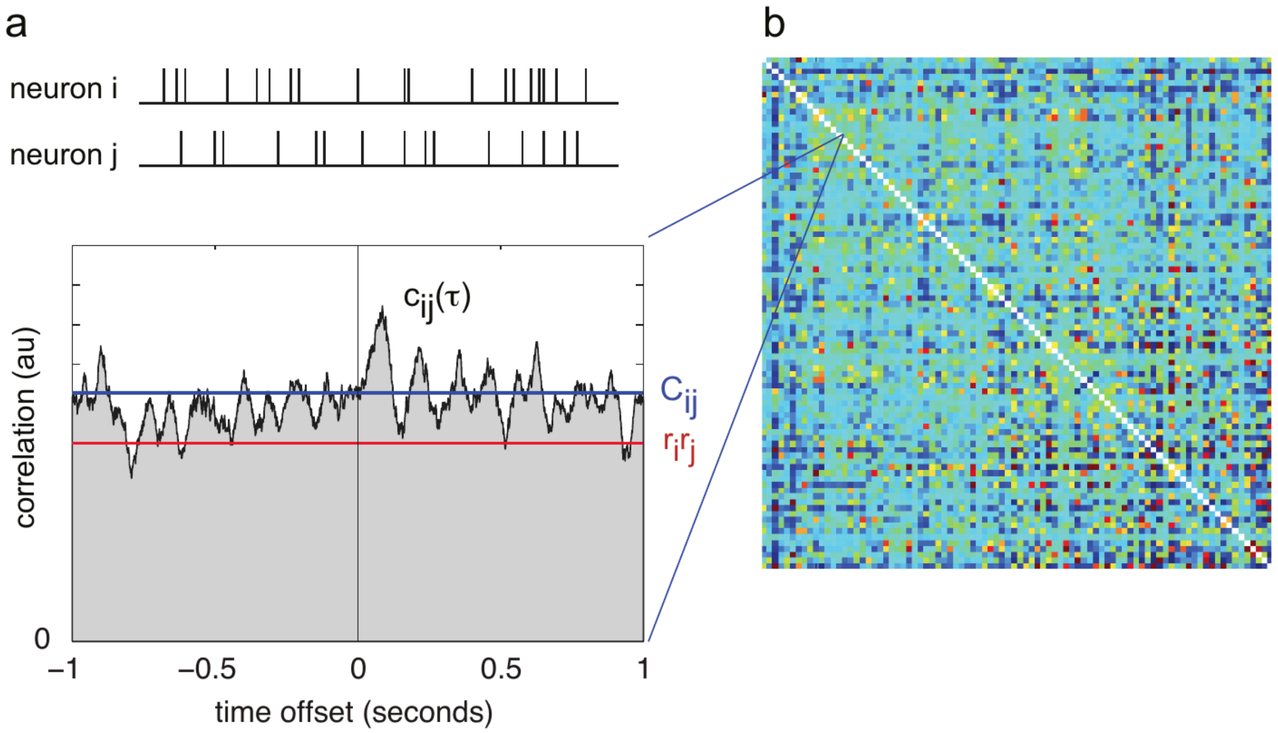
Our questions

- Is the activity pattern unstructured?
- Does it resemble geometric organization?



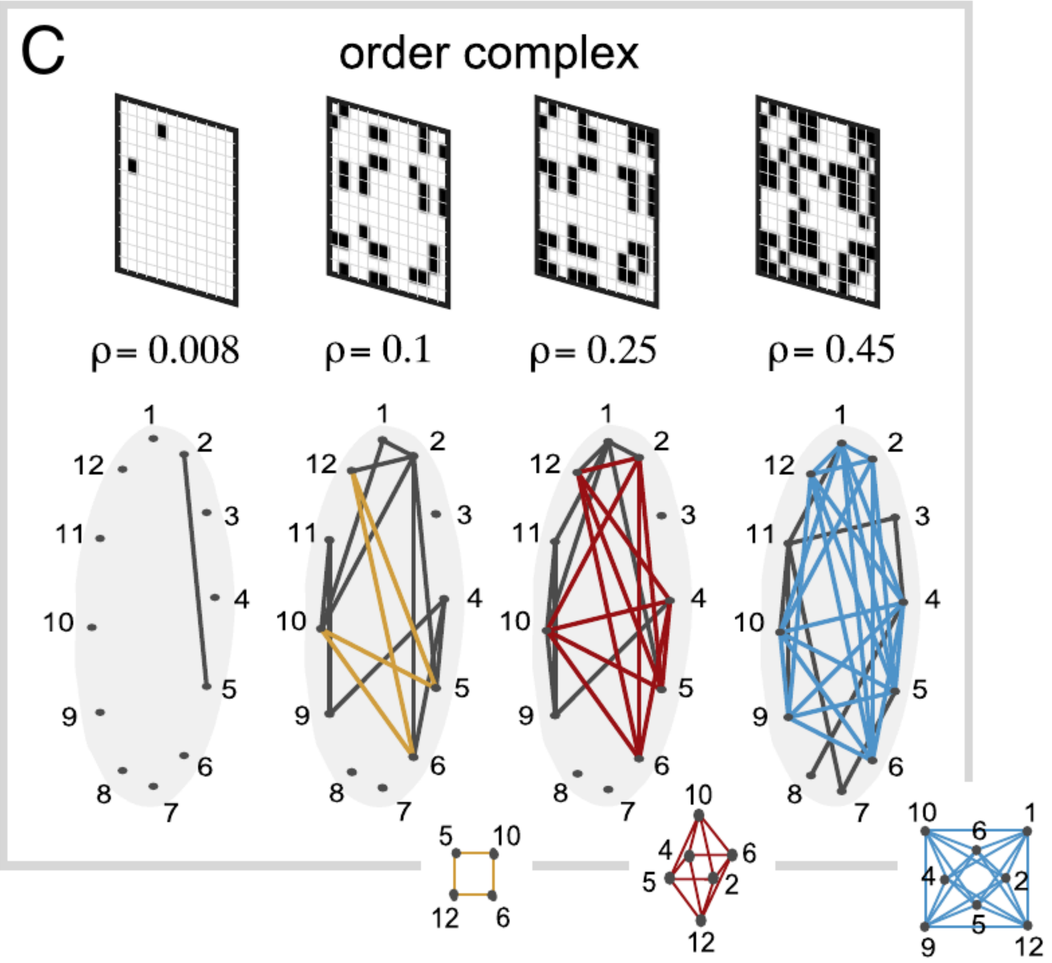
The data

- 88 neurons
- Correlation summarize relationships in their activity
- One recording produces one correlation matrix



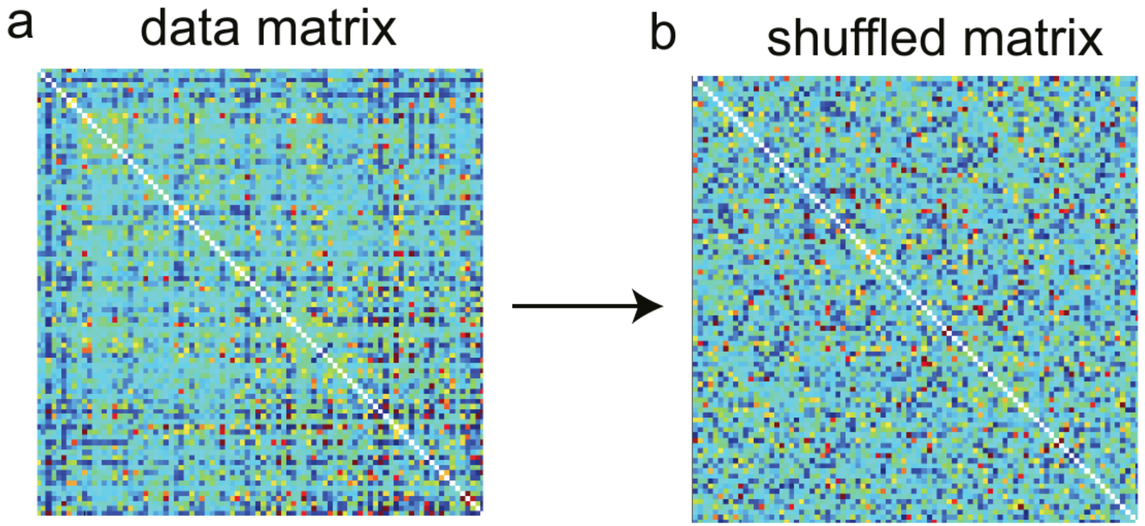
Order Complex

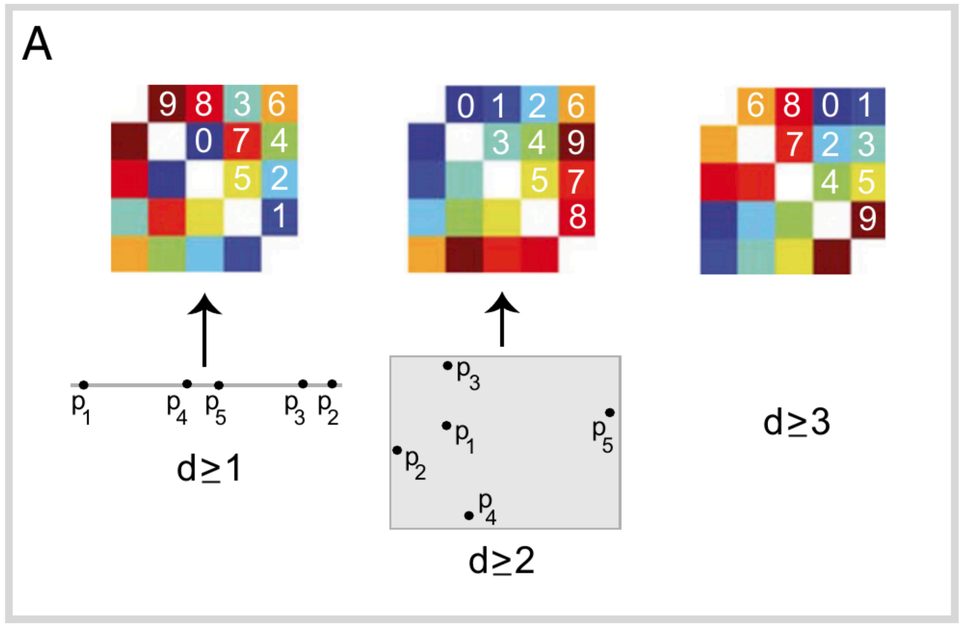
- Each neuron is a vertex
- Add connections from strongest to weakest
- Build the corresponding clique complexes
- Track Betti curves



Two Reference Models

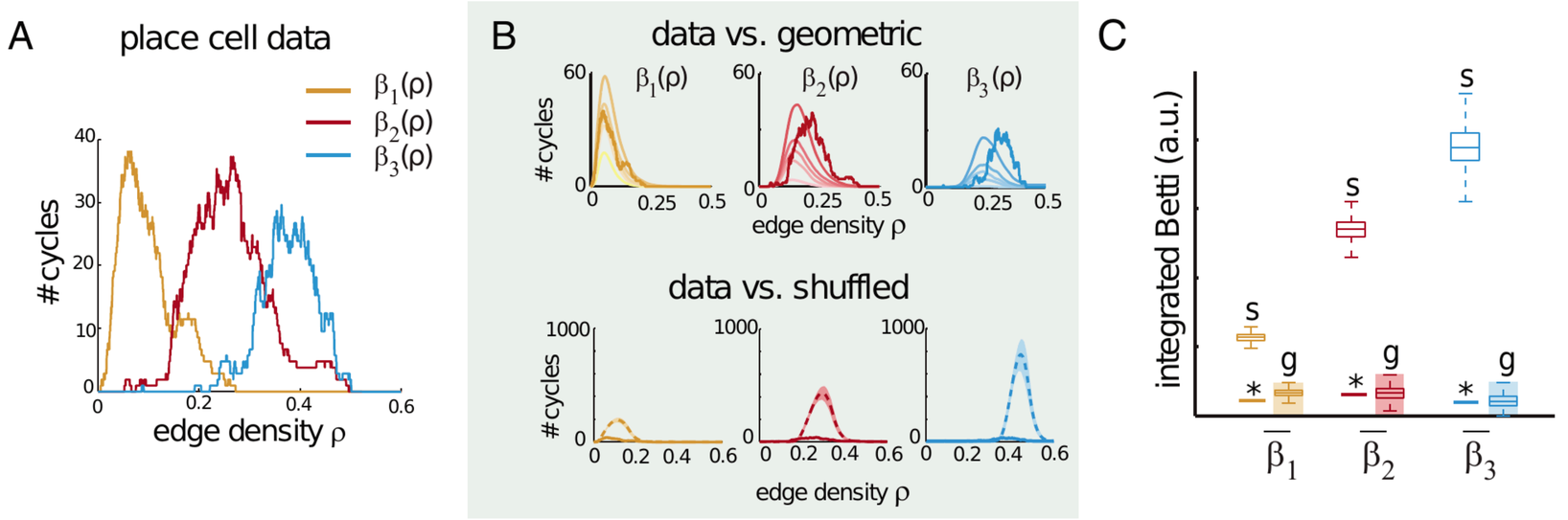
- Shuffled: disrupt correlation organization
- Geometric: derive relationships from distances





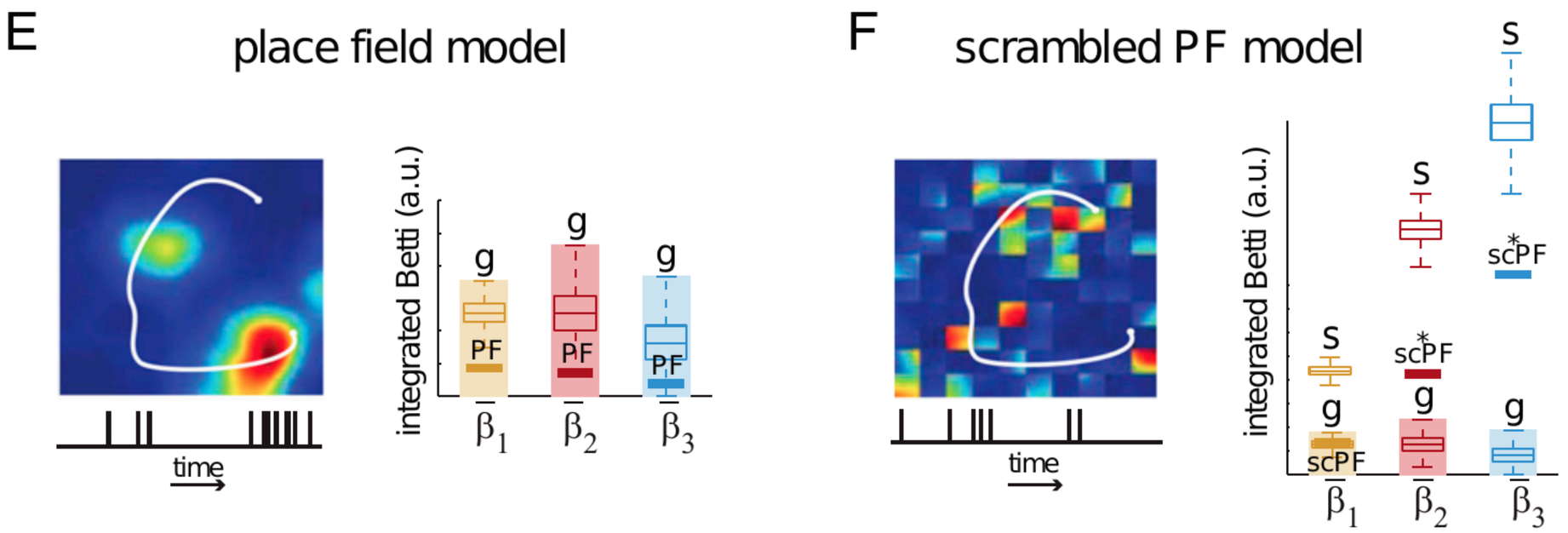
Which reference resembles the recording?

- Observed summaries differ strongly from shuffled controls
- Compatible with geometric controls



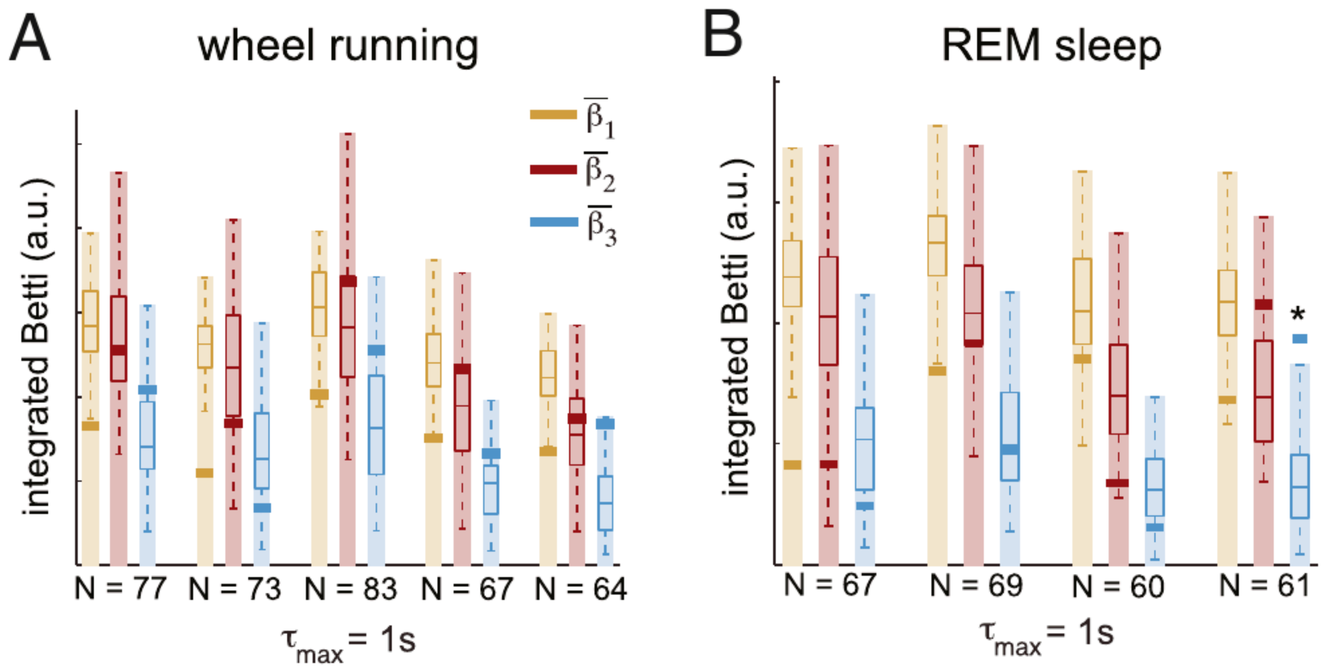
Scramble the Place Fields

- Challenge the role of spatial organization
- H_1 alone could miss the difference
- The second and third summaries reveal it



Beyond Navigation

- Examine wheel running and REM sleep
- Similar geometric signatures appear in these settings



Clarification of “Geometry”

- The observations support a geometric explanation.
- They do not identify a unique geometry or precise dimension.
- Additional controls can test competing explanations.

Finance Data

Topological data analysis of financial time series: Landscapes of crashes

Marian Gidea [a,b,*], Yuri Katz [c]

[a] School of Civil Engineering and Architecture, Xiamen University of Technology, Fujian, China
[b] Department of Mathematical Sciences, Yeshiva University, New York, NY 10016, USA
[c] S&P Global Market Intelligence, 55 Water Str., New York, NY 10040, USA

HIGHLIGHTS

- We introduce a new method, based on topological data analysis (TDA), to analyze financial time series, and detect possible early signs prior to financial crashes.
- We analyze the time-series of daily log-returns of four major US stock market indices: S&P 500, DJIA, NASDAQ, and Russell 2000.
- We use persistence homology to detect and quantify topological patterns that appear in the multidimensional time series.
- We find that, in the vicinity of financial meltdowns, the L^p-norms of persistence landscapes exhibit strong growth prior to the primary peak, which ascends during a crash.
- Our method is very general and can be applied to any asset-types and mixtures of time series.

ARTICLE INFO

Article history:
Received 7 April 2017
Available online 9 October 2017

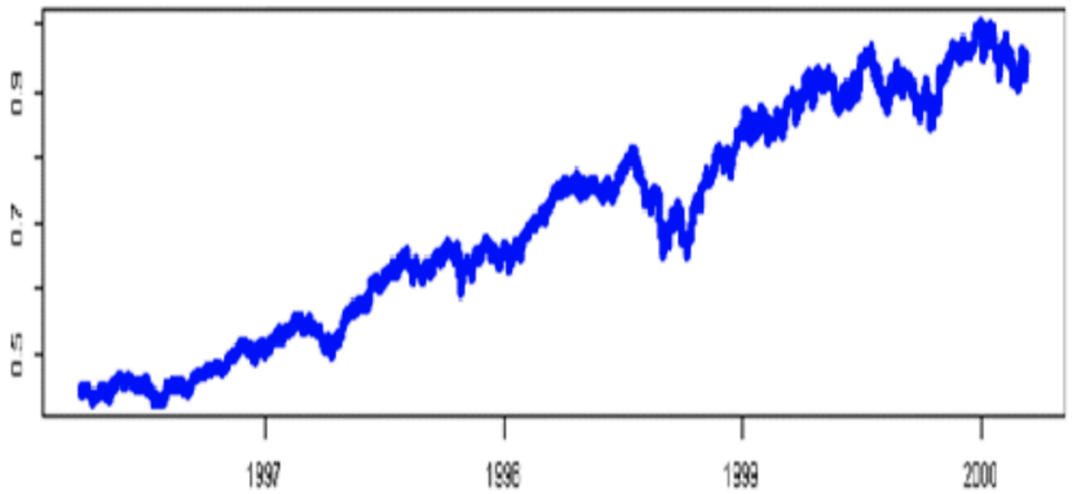
Keywords:
Topological data analysis
Financial time-series
Early warning signals

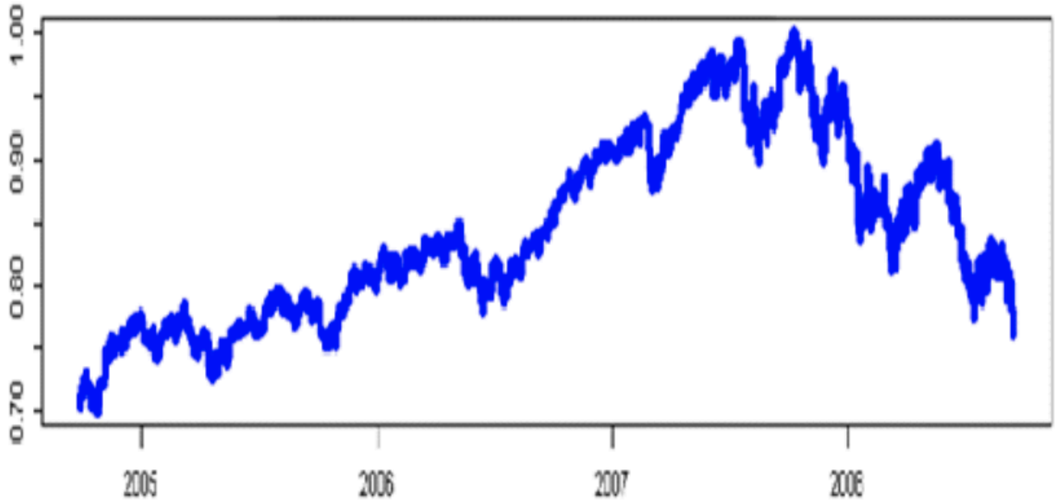
ABSTRACT

We explore the evolution of daily returns of four major US stock market indices during the technology crash of 2000, and the financial crisis of 2007–2009. Our methodology is based on topological data analysis (TDA). We use persistence homology to detect and quantify topological patterns that appear in multidimensional time series. Using a sliding window, we extract time-dependent point cloud data sets, to which we associate a topological space. We detect transient loops that appear in this space, and we measure their persistence. This is encoded in real-valued functions referred to as a 'persistence landscapes'. We quantify the temporal changes in persistence landscapes via their L^p-norms. We test this procedure on multidimensional time series generated by various non-linear and non-equilibrium models. We find that, in the vicinity of financial meltdowns, the L^p-norms exhibit strong growth prior to the primary peak, which ascends during a crash. Remarkably, the average spectral density at low frequencies of the time series of L^p-norms of the persistence landscapes demonstrates a strong rising trend for 250 trading days prior to either dotcom crash on 03/10/2000, or to the Lehman bankruptcy on 09/15/2008. Our study suggests that TDA provides a new type of econometric analysis, which complements the standard statistical measures. The method can be used to detect early warning signals of imminent market crashes. We believe that this approach can be used beyond the analysis of financial time series presented here.

© 2017 Elsevier B.V. All rights reserved.

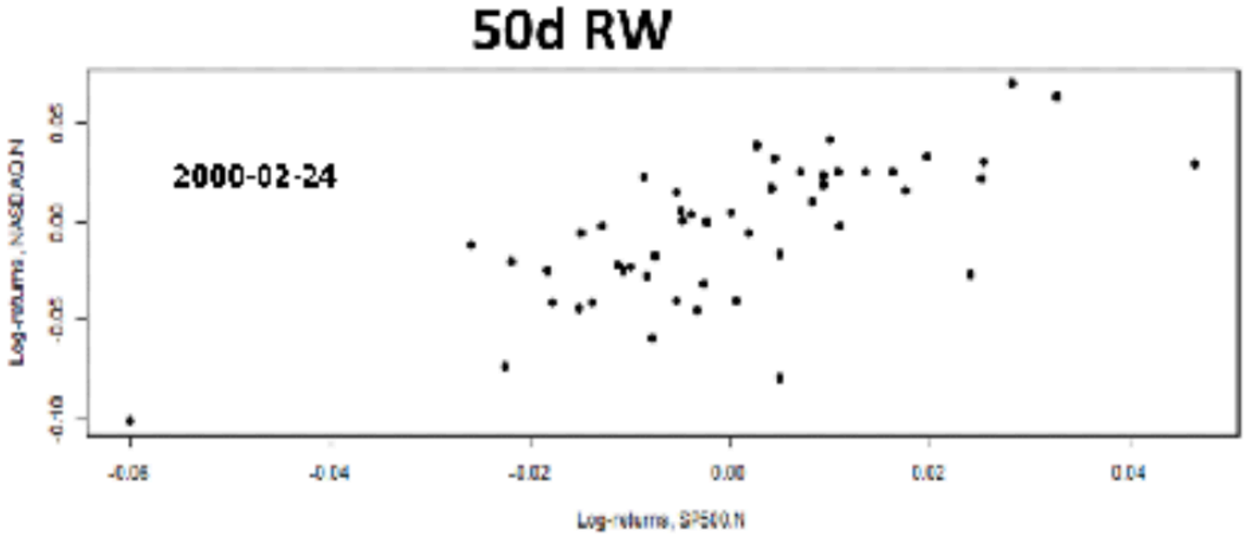
Market Disruption

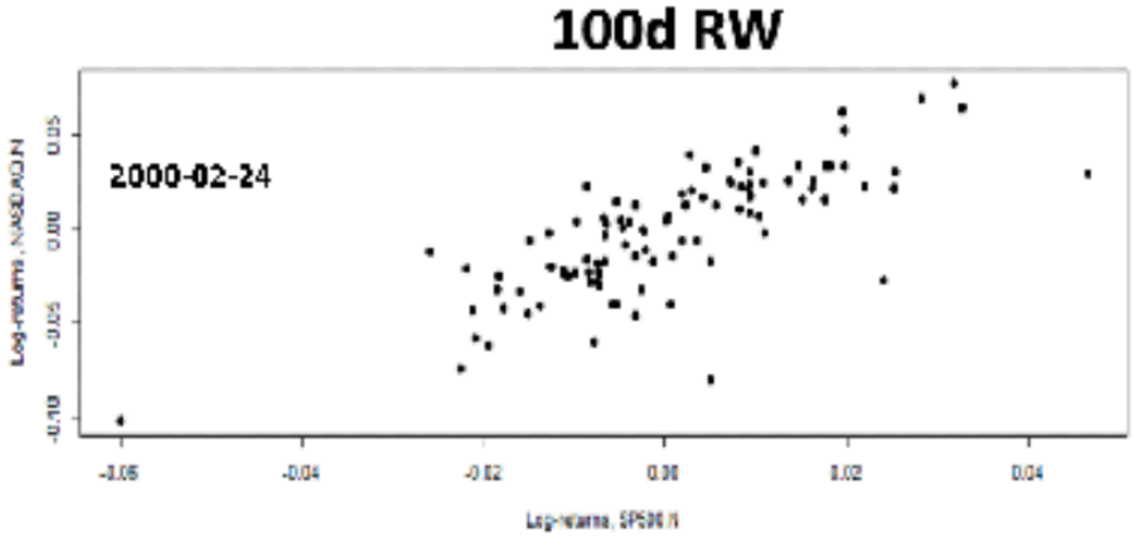




Point Cloud

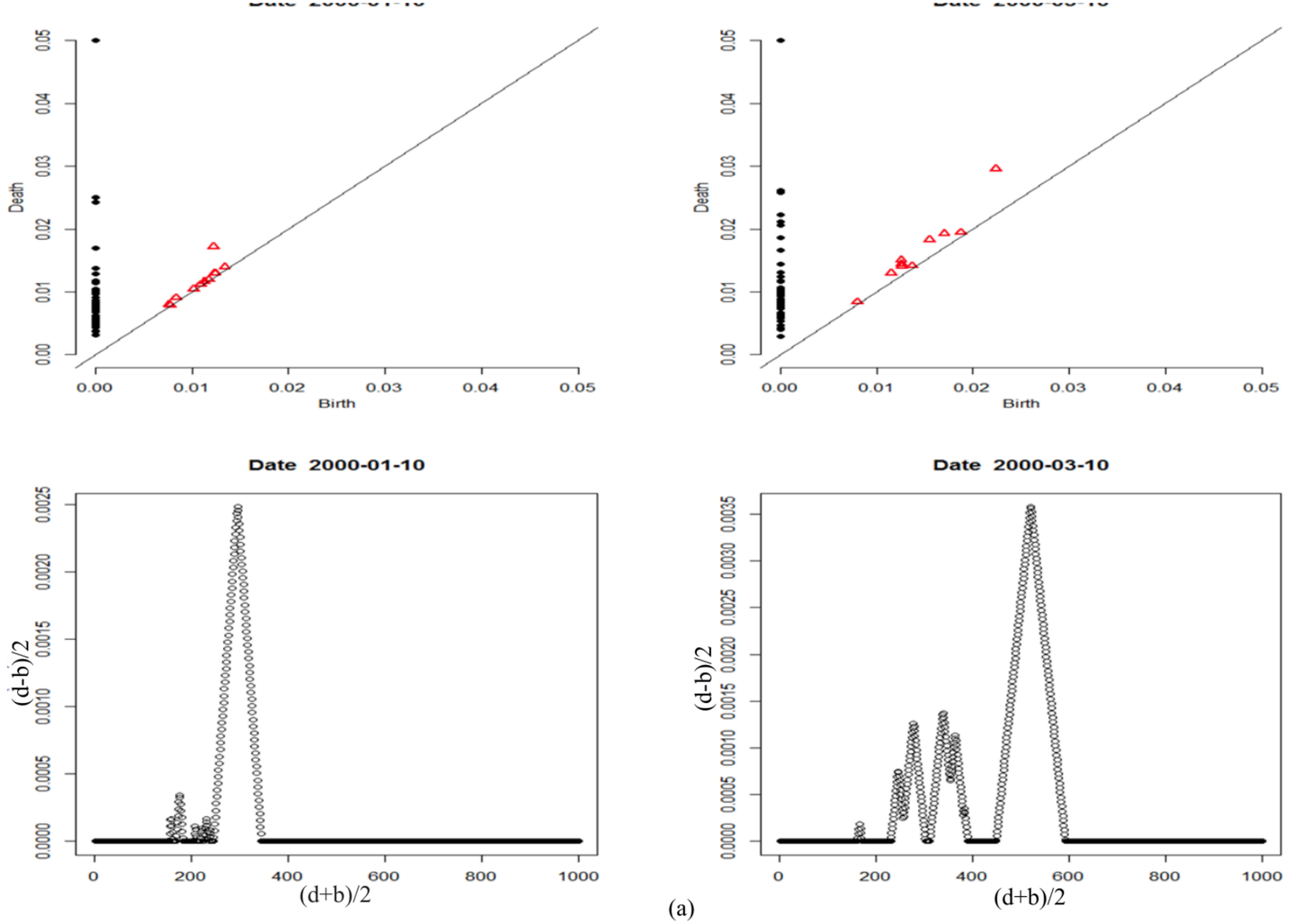
- One point contains a day's return for four indices (S&P 500, Dow Jones, NASDAQ, Russell 2000)
- A window collects consecutive days
- Window length controls how much history is included





What changes in the representation?

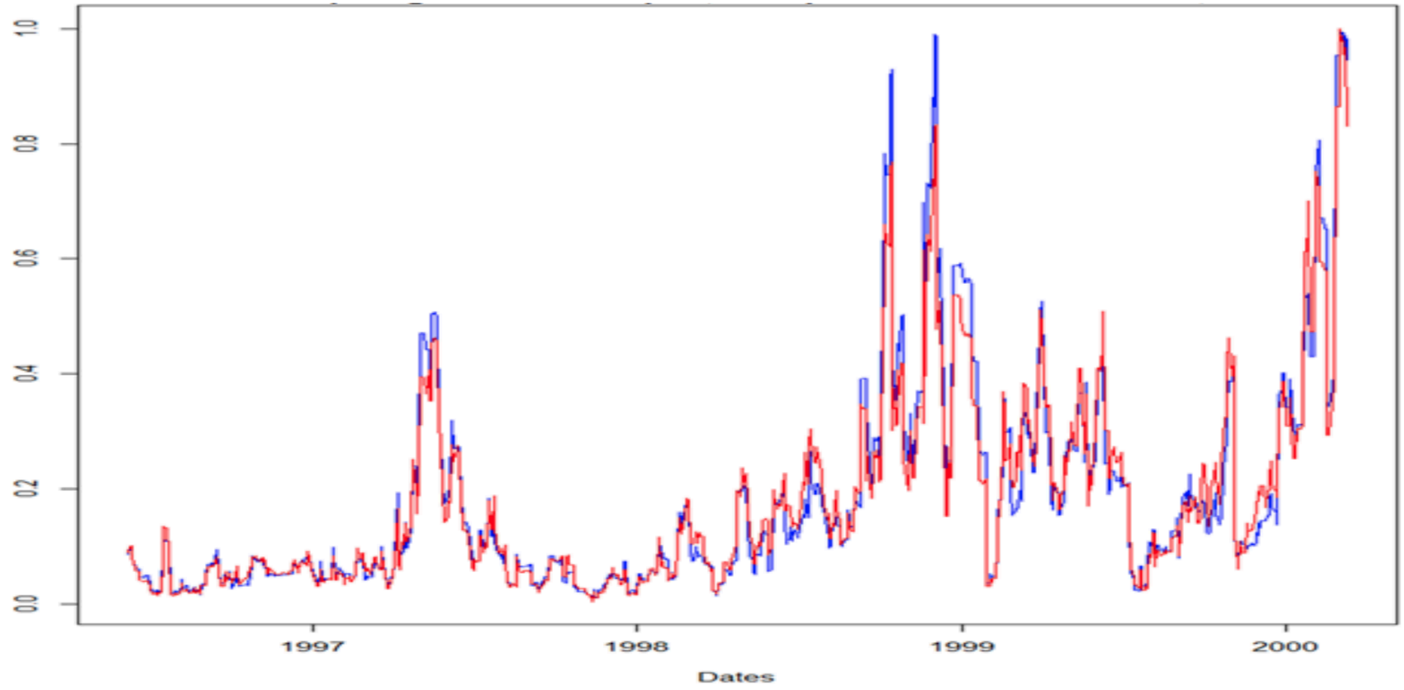
- Compute persistence for each window
- Compare diagrams and landscapes across dates
- Identify which changes survive in the summary



(a)

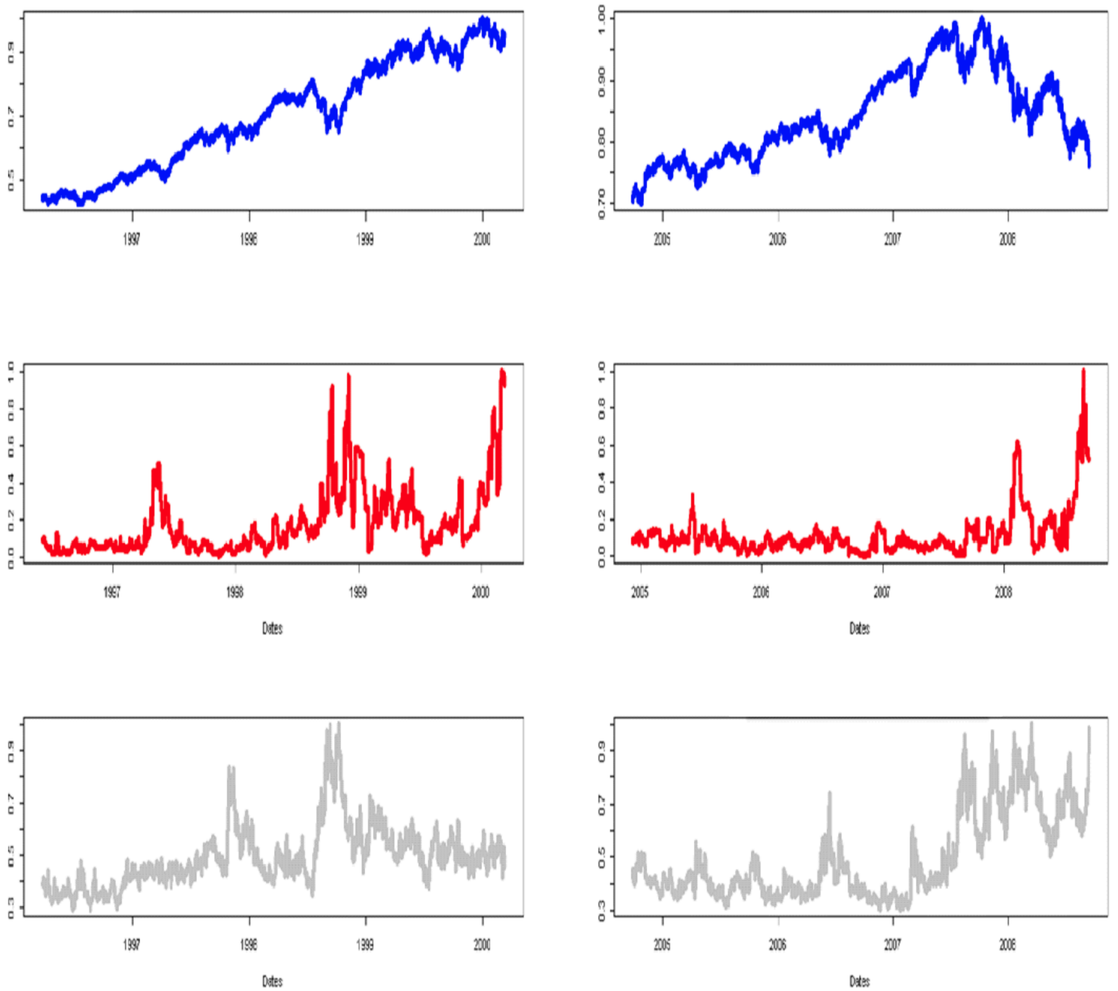
Track one summary through time

- Compute a landscape norm for each window
- Blue: L_1 norm (total area under all landscape layers)
- Red: L_2 norm (taller peaks have more influence)
- What changes does the signal respond to?



Compare with market behavior

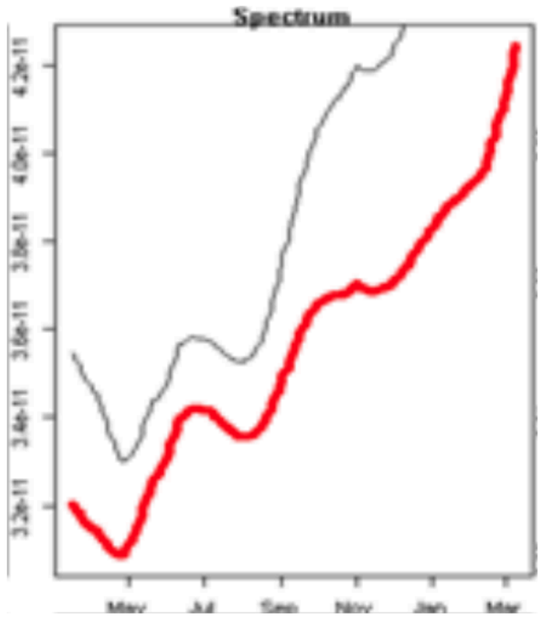
- Top: S&P 500
- Middle: L_1 landscape norm
- Bottom: VIX (a measure of expected market volatility) (my husband: “Like how scared ppl are. It indicates how uncertain and scared people are about the market. So when market crashes happen, VIX goes up a lot)



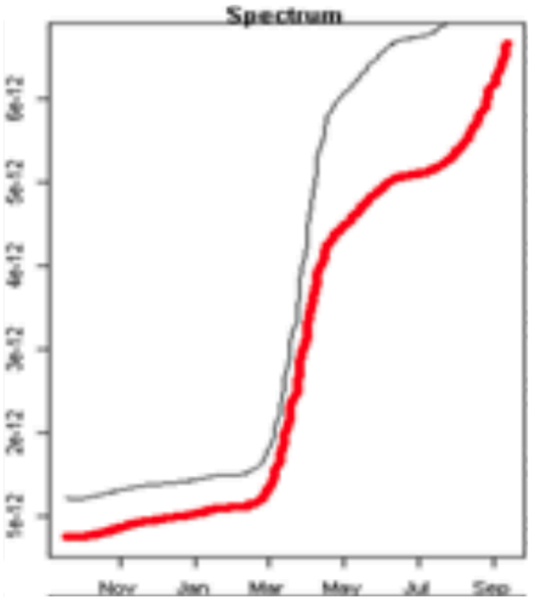
Does behavior change before the event?

- Examine statistical indicators of the signal
  - Grey: low-frequency spectral power of L_1
  - Red: low-frequency spectral power of L_2
- Look for a sustained change before each event
- Does the pattern generalize?

2000 episode

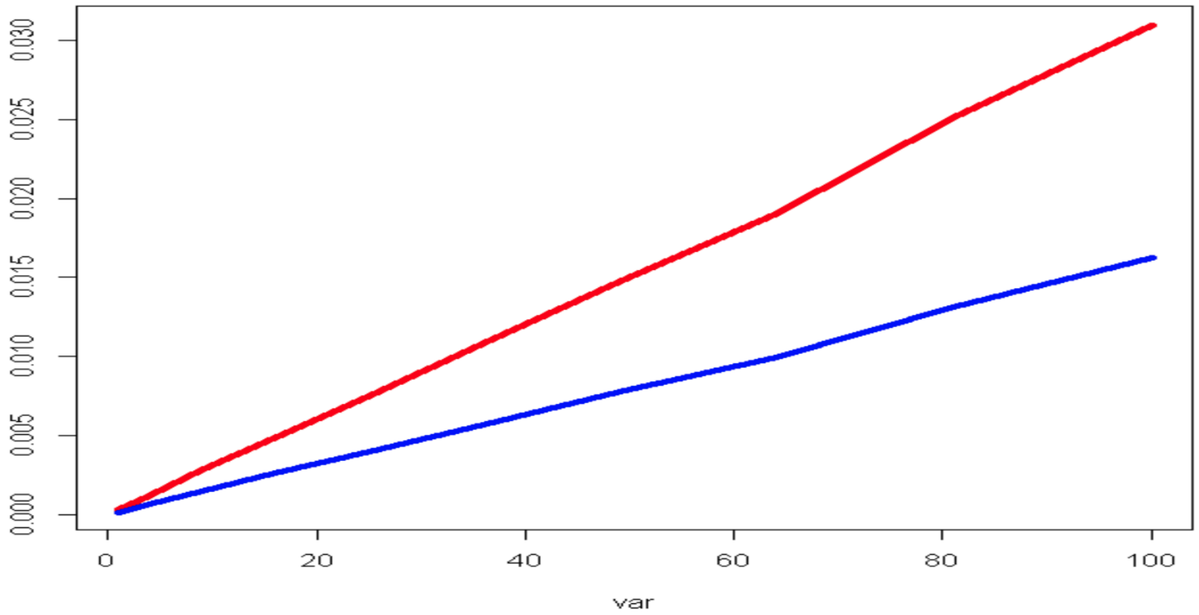


2008 episode



What could make the topological signal increase?

- Generate random data and increase its variance (simulated Gaussian noise)
- Horizontal axis: noise variance
- Vertical axis: average landscape norm
- Red: L_1
- Blue: L_2



Using Persistence in Applications

Match the evidence to the question

- Prediction: relevant baselines and unseen data
- Group differences: suitable samples and tests
- Explanation: informative controls
- Warning: prospective evaluation and false alarms

- **Material**: predict outcomes
- **Proteins**: compare groups
- **Neuroscience**: compare explanations
- **Finance**: evaluate a warning

Directional Transform — Intro

Recall from last week

International Press *publishers of scholarly mathematical and scientific journals and books*

Home

Journals

Journal Content Online

Books

Information & Ordering

Company Contacts

Join Our Mailing Lists

CONTENTS ONLINE

HHA Home Page | HHA Content Home | All HHA Volumes
This Volume | This Issue

Homology, Homotopy and Applications
Volume 18 (2016)
Number 1

The ring of algebraic functions on persistence bar codes

Pages: 381 – 402

DOI: http://dx.doi.org/10.4310/HHA.2016.v18.n1.a21

Authors

Aaron Adcock (Facebook, Inc., New York, N.Y., U.S.A.)

Erik Carlsson (Center of Mathematical Sciences and Applications, Harvard University, Cambridge, Massachusetts, U.S.A.)

Gunnar Carlsson (Department of Mathematics, Stanford University, Stanford, California, U.S.A.)

Abstract

Persistent homology is a rapidly developing field in the study of numerous kinds of data sets. It

Features — Digits Example

For each digit:

Diagrams

- 4 directions
- 0- and 1-dimensional diagrams for each direction
- = 8 diagrams each

Features:

- Each diagram has 4 features
- = 32 features total

Features:

$y_{\max}$ = max death for all diagrams

Diagrams $X = \{(x_i, y_i)\}$

- $\sum x_i(y_i - x_i)$
- $\sum (y_{max} - y_i)(y_i - x_i)$
- $\sum x_i^2(y_i - x_i)^4$
- $\sum (y_{max} - y_i)^2(y_i - x_i)^4$

What if we take more directions?

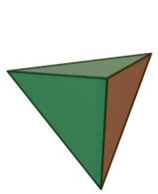
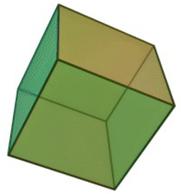
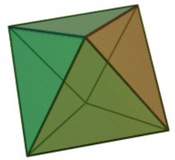
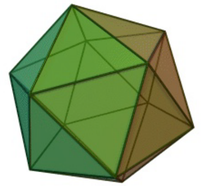
The Euler Characteristic

$$\chi(K) = \beta_0(K) - \beta_1(K)$$

	$\beta_0(K)$	$\beta_1(K)$	$\chi(K)$
C, E, F,...	1	0	1
A, D, O,...	1	1	0
B	1	2	-1
i, j	2	0	2

The Euler Characteristic

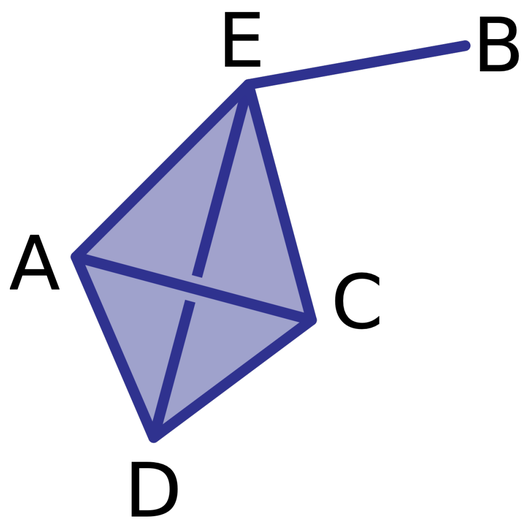
$$\chi(K) = V - E + F$$

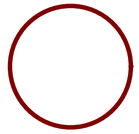
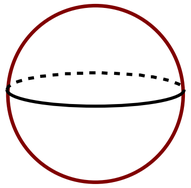
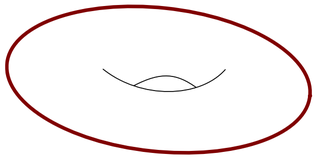
	Tetrahedron	Cube	Octahedron	Dodecahedron
V	4	8	6	20
$-E$	-6	-12	-12	-30
$+F$	4	6	8	12
χ	2	2	2	2

Generalized Euler Characteristic

For any simplicial complex K where n_p is the number of p-dimensional simplices, the Euler characteristic is $\chi(K) = \sum_p (-1)^p n_p$



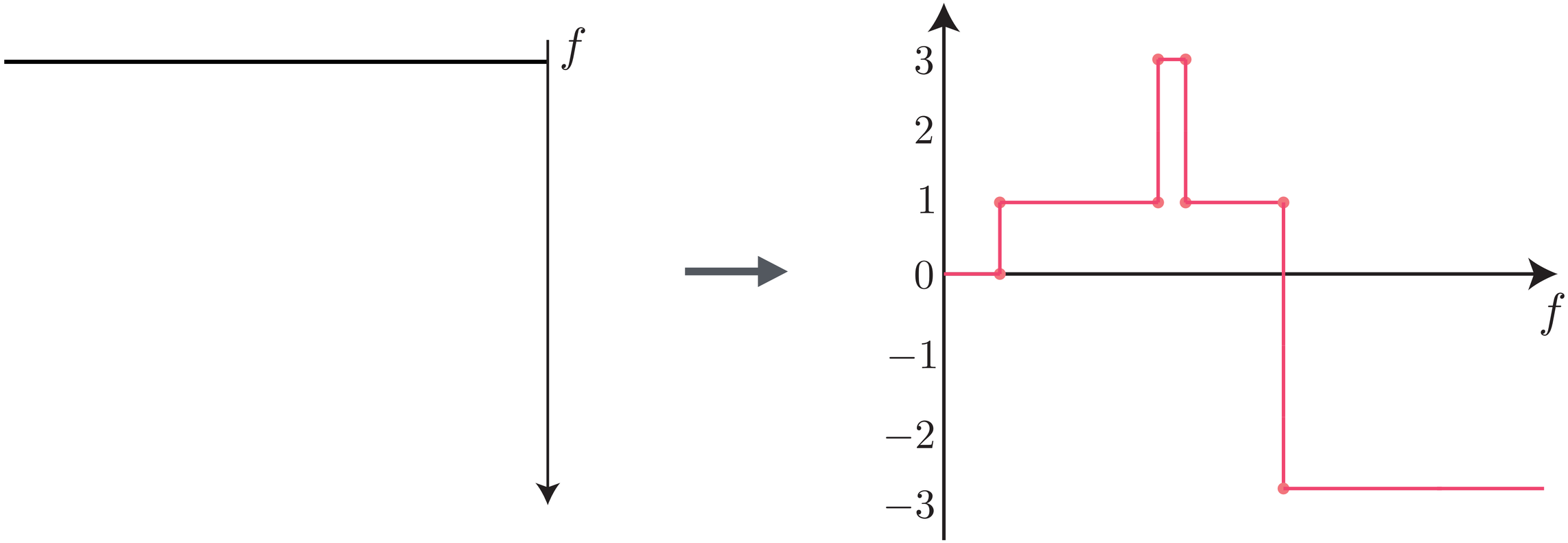
Euler Characteristic of Spaces

K	•			
$\chi(K)$	1	0	2	0

- If the objects are the **same**, the ECs are the **same**
- If the ECs are **different**, the objects are **different**

The Euler Characteristic is a topological signature

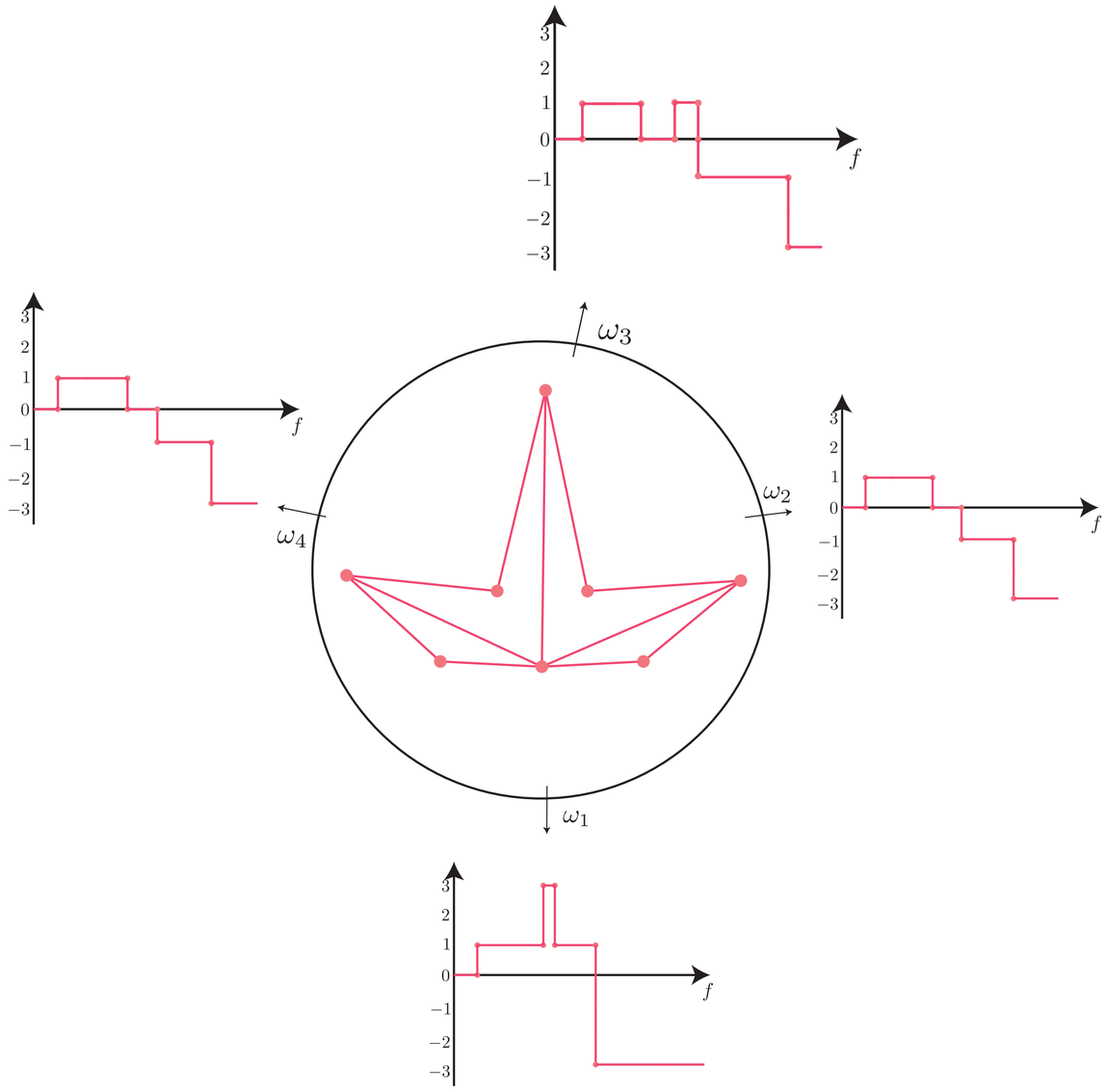
The Euler Characteristic as a Function



Euler Characteristic Curve: $\mathbb{R} \to \mathbb{Z}$

$$a \mapsto \chi(K_a)$$

Euler Characteristic Transform (ECT)



Data: $\boldsymbol{D} \subseteq \mathbb{R}^2$ and

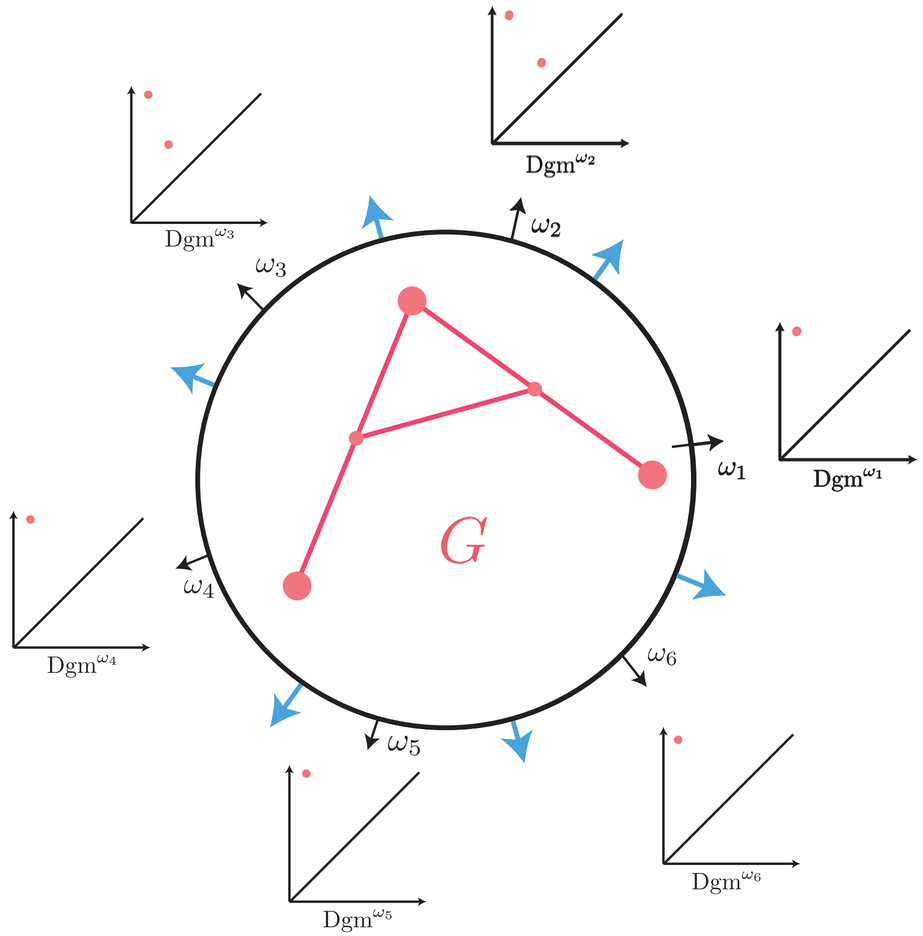
Direction $\boldsymbol{\omega} \in \mathbb{S}^1$

$$\mathrm{ECT}(\boldsymbol{D}) : \mathbb{S}^1 \to \mathrm{Func}(\mathbb{R}, \mathbb{Z})$$

$$\omega \mapsto \{a \mapsto \chi(f_\omega(-\infty, a])\}$$

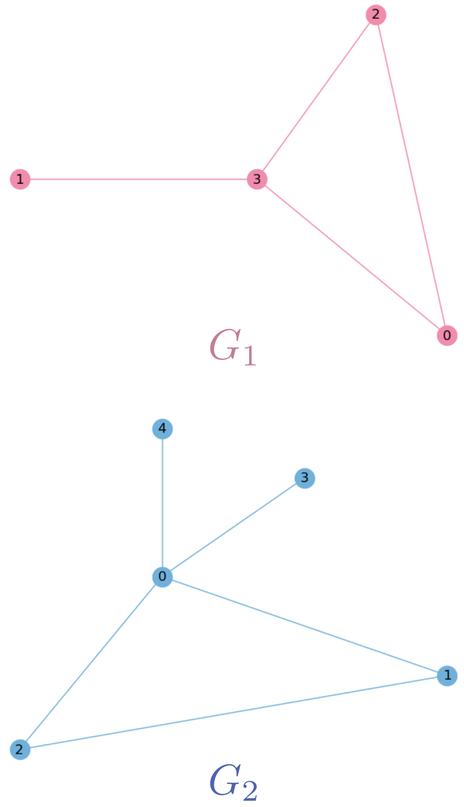
Persistent Homology Transform

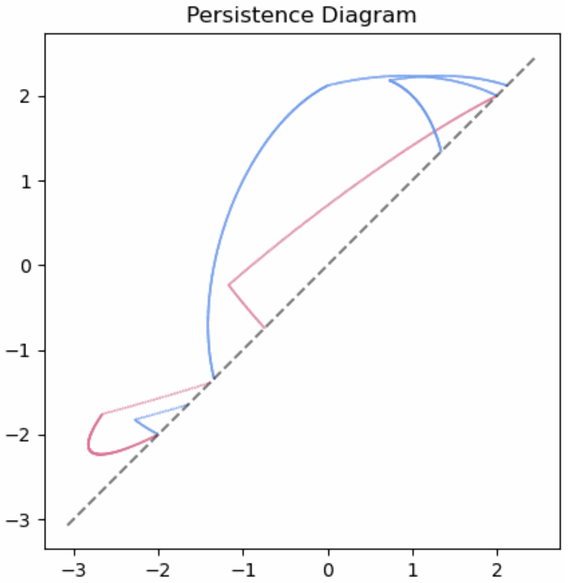
If we compute the persistence diagrams from all directions of $\mathbb{S}^1$, we obtain a family of diagrams (vineyards), called the **Persistent Homology Transform (PHT).**



- Formally, the PHT of $K \subseteq \mathbb{R}^d$ is
  $\mathrm{PHT}(K) : \mathbb{S}^{d-1} \to \mathrm{Dgm}^d$
  $\omega \to (\mathrm{Dgm}_0^{\omega}(K), \mathrm{Dgm}_1^{\omega}(K), \ldots, \mathrm{Dgm}_{d-1}^{\omega}(K))$
- The PHT is **injective**, i.e., if $\mathrm{PHT}(A) = \mathrm{PHT}(B)$, then $A = B$
- The points in PHT are **continuous**

Persistent Homology Transform





Mathematical Guarantees

Theorem (Turner, Mukherjee, Boyer, 2014)

The ECT and PHT is **injective** !

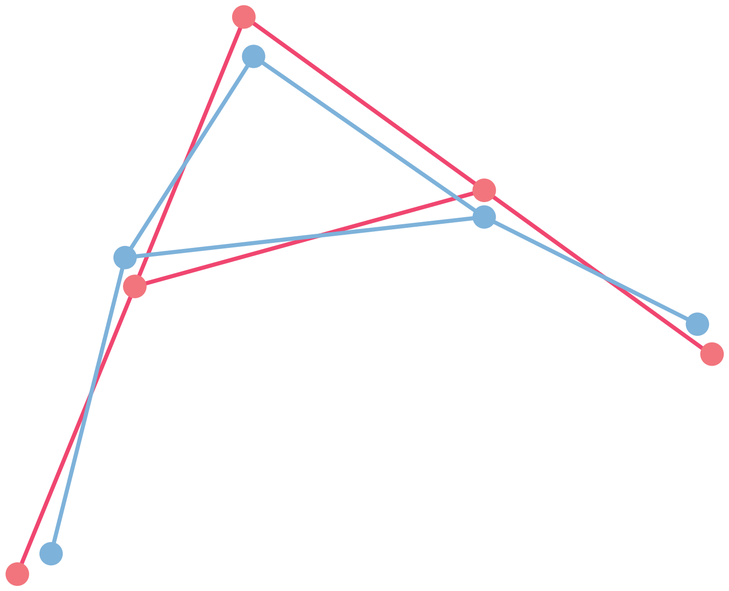
If $\mathrm{ECT}(A) = \mathrm{ECT}(B)$, then $A = B$

If $\mathrm{PHT}(A) = \mathrm{PHT}(B)$, then $A = B$

ECT Stability

Work by EXW, Jasmine George, Oscar Lledo Osborn, Messiah Ridgley II, Liz Munch

On the Stability of the Euler Characteristic Transform for a Perturbed Embedding, submitted to *Involve*



K_1 K_2

Objects

$$\rightarrow \quad d\left(\mathrm{ECT}(K_1), \mathrm{ECT}(K_2)\right) \leq C\|K_1 - K_2\|_2$$

ECT Stability

Homework Presentation

Consider the V-shaped graph K with vertices $A = (-1,1),\ B = (0,0),\ C = (1,1)$ with edges AB and BC. For a direction v, use the height function $h_v(x) = x \cdot v$. A vertex enters at its height, and an edge enters when both endpoints have entered.

- Scan the graph upward, using $v = (0,1)$. Sketch the filtration, compute its Euler characteristic curve, and draw its H_0 persistence barcode.

- Repeat for the downward direction $v = (0, -1)$.