REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
05-01-2023	Final Report	13-Jul-2018 - 8-Oct-2022

4. TITLE AND SUBTITLE
Final Report: A topological heat map for data analysis (TopHeat)

5a. CONTRACT NUMBER
W911NF-18-1-0307

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Florida - Gainesville
219 Grinter Hall
PO Box 115500
Gainesville, FL 32611 -5500

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
72995-MI.15

12. DISTRIBUTION AVAILIBILITY STATEMENT
2 Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Peter Bubenik
UU	UU	UU	UU		19b. TELEPHONE NUMBER 352-294-2342

RPPR

as of 02-Feb-2023

Agency Code:

Proposal Number: **Agreement Number:**

Organization:
Address: , ,
Country:
DUNS Number: EIN:
Report Date: Date Received:
for Period Beginning and Ending
Title:
Begin Performance Period: **End Performance Period:**
Report Term: -
Submitted By: Email:
Phone:

Distribution Statement: -

STEM Degrees: **STEM Participants:**

Major Goals:

Accomplishments:

Training Opportunities:

Results Dissemination:

Plans Next Period:

Honors and Awards:

Protocol Activity Status:

Technology Transfer:

I certify that the information in the report is complete and accurate:
Signature:
Signature Date:

Abstract

Topological data analysis provides summaries for the "shape" of data. Some of these summaries, such as the PI's persistence landscape, are feature maps and kernels, and can be easily combined with standard methods of statistics and machine learning. However, these topological summaries can be difficult for non-experts to interpret. In this project, we produced a new summary, that may be visualized as a heat map on the underlying data. We developed theory for this heat map, showing that it is stable under perturbations of the input. Furthermore, we showed how to combine this summary with statistics and machine learning and applied it to synthetic data and real data.

1 Objectives

The main goal of this project was to develop a new visualization tool for a topological summary that "lives on the data" and its underlying theory. An additional goal was to incorporate this tool into current approaches to statistics and machine learning with topological data analysis (TDA).

Objective 1: A topological heat map

The persistence diagram is one of the main reasons for the success of topological data analysis. It provides a succinct and stable visualization of how the shape of the data changes with a parameter of interest. However, it is also an impediment to the wider use of TDA as it is an abstract summary whose interpretation requires significant understanding of both the data and this topological tool. For example, the points in the persistence diagram furthest away from the diagonal correspond to the most topologically significant features of the data. But what are these features in the data? A stable visualization of persistent homology on top of the starting data would be of great interest to practitioners. For example, persistent homology might detect an anomaly in an image in comparison to reference images. Our new visualization would help locate this anomaly in the image.

We proposed to construct an alternative topological summary to the persistence diagram that would provide a heat map on the data. Whereas the persistence diagram only pairs critical values, we will also pair critical points. Unfortunately, unlike the pairing of critical values, the pairing of critical points is unstable: small perturbations of the data can lead to arbitrarily large changes in the pairings.

To deal with this instability, we proposed randomly perturb the data a number of times and takes an average. The result of these computations could be visualized as a heat map.

Objective 2: Underlying theory

Crucial to validating this topological heat map is developing a mathematical framework in which this signal is well defined and one can prove that it cannot be corrupted by noise in the data (i.e. the heat map is stable). The main goal of this task was the construction of a persistence measure. This requires the use of some sophisticated ideas from functional analysis that are new to TDA.

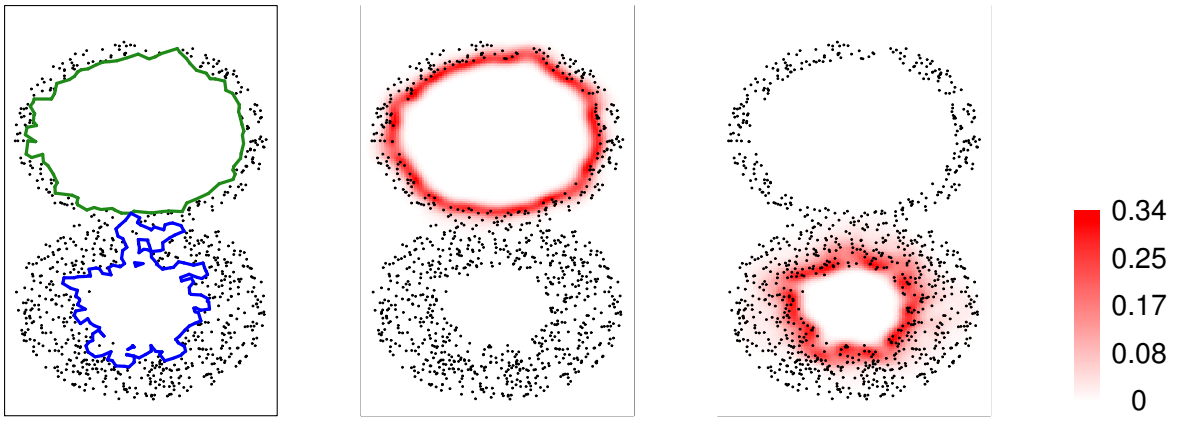


Figure 1: Left: we sample points from a figure eight, where the top annulus is thinner than the bottom one. Persistent homology in degree one sees two long bars corresponding to the two holes in the figure eight. The representative cycles for these persistent homology classes are indicated in red and blue respectively. Middle: the topological heat map for the most persistent topological feature. Right: the topological heat map for the second most persistent topological feature.

Objective 3: Combining topological heat maps with statistics and machine learning

The completion of Objectives 1 and 2 would enable new approaches to combining TDA with statistics and machine learning. Two such possible interactions will be considered. In the first, we will use the topological heat map to visualize the location of statistically significant topological features and classification vectors. In the second, we will consider our topological heat map as a feature map and perform statistics and machine learning directly on it.

Objectives for each grant year

Year 1: Objectives 1 and 2. Year 2: Objectives 1 and 2. Year 3: Objective 3.

2 Findings

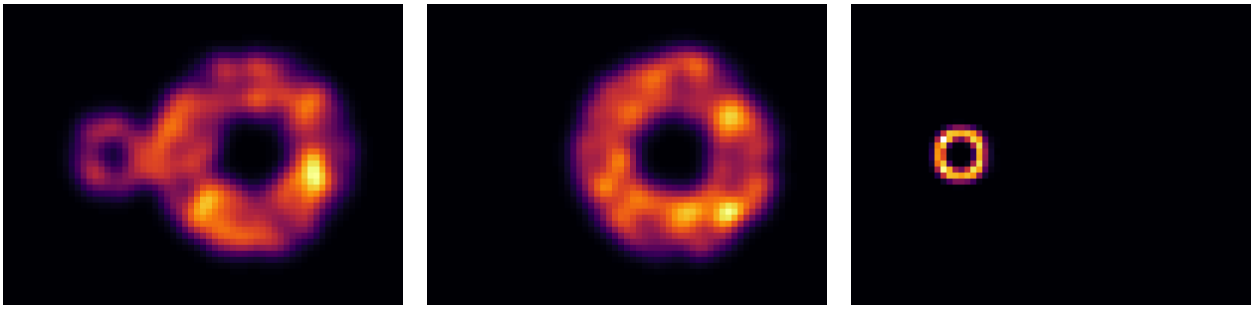
2.1 A topological heat map [1]

We developed a stable topological heat map that identifies the parts of data that are responsible for a significant topological feature. See Figures 1, 2, and 3.

2.2 Underlying theory

2.2.1 Stabilization [1]

A key ingredient for our topological heat maps is the following a simple procedure for stabilizing unstable persistent homology computations: perturb the input by adding, for example, Gaussian noise, and redo the computation; repeat and average. See Algorithm 1. By Theorem 1, the law



(a) Sample from double annulus (b) Sample from single annulus (c) Learned heat map

Figure 2: Left and middle: kernel density estimators for points sampled from two different regions. Right: the learned heat map for the difference between the two regions. That is, samples from two distributions are used to produce an empirical approximation of the expected learned heat map.

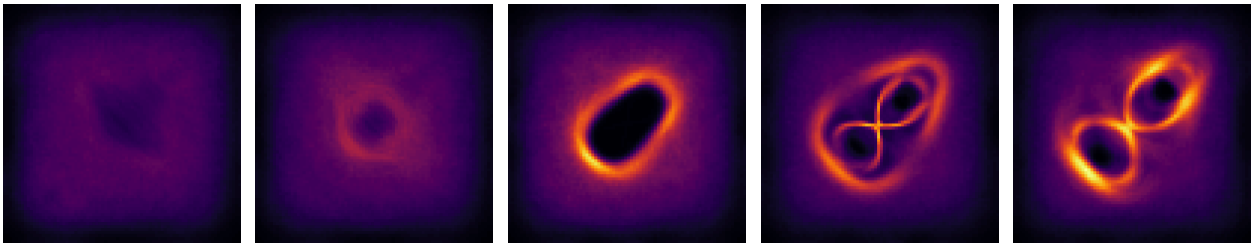


Figure 3: We consider a one-parameter family of discrete dynamical systems called the linked twist map. The five images show the topological heat maps for the values $r = 3.9, 4, 4.1, 4.2, 4.3$, respectively.

of large numbers, and our Theorem 2, the result converges to the desired stable value. As a result our topological heat maps are stable to perturbations of the input.

Algorithm 1 Stabilizing unstable persistence computations

Require: $h : \mathbb{R}^n \to \mathbb{R}$, $a \in \mathbb{R}^n$
Ensure: $M \in \mathbb{N}$, $\sigma > 0$
  for $i \leftarrow 1, M$ **do**
    for $j \leftarrow 1, n$ **do**
      Sample ϵ_j from $N(0, \sigma^2)$
    end for
    $y_i \leftarrow h(a + \epsilon)$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$
  end for
  return the average value of $y_1, \ldots, y_M$

Theorem 1 *Let $\epsilon_1, \ldots, \epsilon_M$ be drawn independently from a kernel K. Then*

$$\frac{1}{M} \sum_{i=1}^{M} h(a - \epsilon_i) \to (h * K)(a).$$

Theorem 2 *If h is locally essentially bounded then for the triangular and Epanechnikov kernels, $h * K$ is locally Lipschitz. If h is essentially bounded then for the Gaussian kernel, $h * K$ is Lipschitz.*

2.2.2 The persistence landscape kernel [3]

We proved the persistence landscape kernel is characteristic for generic empirical measures. That is, the mapping from such a measure to its expected persistence landscape is injective. Equivalently, not only can we recover a persistence diagram from a persistence landscape, we can recover a finite set of persistence diagrams from their average persistence landscape.

2.2.3 The graded persistence diagram [2]

We introduced a refinement of the persistence diagram, the graded persistence diagram. It is the Möbius inversion of the graded rank function, which is obtained from the rank function using the unary numeral system. Both persistence diagrams and graded persistence diagrams are integer-valued functions on the Cartesian plane. Whereas the persistence diagram takes non-negative values, the graded persistence diagram takes values of 0, 1, or -1. We proved that the sum of the graded persistence diagrams is the persistence diagram. Furthermore we showed that the positive and negative points in the kth graded persistence diagram correspond to the local maxima and minima, respectively, of the kth persistence landscape. Finally, we proved a stability theorem for graded persistence diagrams.

2.2.4 Embedding into Hilbert space [12]

Since persistence diagrams do not admit an inner product structure, a map into a Hilbert space is needed in order to use kernel methods. It is natural to ask if such maps necessarily distort the metric on persistence diagrams. We proved that persistence diagrams with the bottleneck distance do not even admit a coarse embedding into a Hilbert space. As part of our proof, we showed that any separable, bounded metric space isometrically embeds into the space of persistence diagrams with the bottleneck distance. As corollaries, we obtained the generalized roundness, negative type, and asymptotic dimension of this space.

2.2.5 Homological algebra for TDA [9]

Underlying TDA are algebraic objects called persistence modules. We applied homological algebra to study these persistence modules. We classified the projective, injective and flat interval modules and computed their Hom and Ext functors.

2.2.6 Universal constructions for TDA [5]

We proved that the standard tools of TDA, the persistence diagrams and the Wasserstein distances, are obtained from universal constructions.

2.2.7 Algebraic distances for TDA [11]

It is crucial to applications to measure distances between topological summaries. We developed new algebraic methods for measuring distances between persistence modules.

2.2.8 Embedding into Banach spaces [6]

We showed that persistence diagrams can be isometrically embedded into a Banach space of signed measures.

2.2.9 Geometric topology and TDA [4]

We used cobordism theory, generalized Morse theory, and Cerf theory to construct new tools for applying TDA to one-parameter families for functions.

2.2.10 A common framework for discrete and continuous TDA [10]

We developed TDA in the setting of filtered Čech closure spaces, providing a uniform treatment for the persistent homology of filtered topological spaces, metric spaces, weighted graphs, and weighted directed graphs.

2.2.11 Topological and metric properties for TDA [7]

We considered persistence diagrams as formal sum on a metric space and studied their topological and metric properties.

2.3 Combining topological heat maps with statistics and machine learning

2.3.1 Topological heat map for images [13]

We developed TDAExplore, a machine learning image analysis pipeline based on topological data analysis. It can classify different types of cellular perturbations after training with only 20–30 high-resolution images and performs robustly on images from multiple subjects and microscopy modes. Using only images and whole-image labels for training, TDAExplore provides quantitative, spatial information, characterizing which image regions contribute to classification. See Figures 4 and 5.

2.3.2 Curvature [8]

In topological data analysis, it is often said that the long intervals in the barcode represent the 'topological signal' and the short intervals represent 'noise'. We give evidence to dispute this thesis, showing that the short intervals encode geometric information. Specifically, we prove that persistent homology detects the curvature of disks from which points have been sampled. See Figures 6 and 7.

2.3.3 Topological Decomposition of Videos [14]

We applied topological data analysis in a novel way to a video to produce synthetic periodic videos that represented the characteristic motions contained in the video. The video in this case was of *C. elegans*, an extensively studied model organism in biology. The locomotion and behavior of this worm (nematode) is quasi-periodic. Applying persistent homology to a high-dimensional time-delay embedding of this video, we obtained cycle representatives of the most important

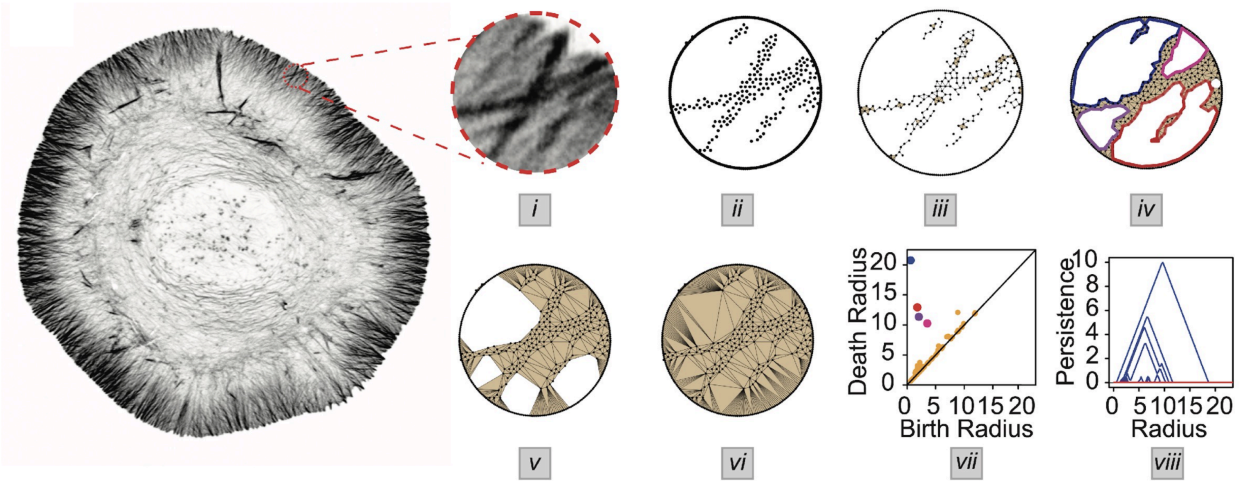


Figure 4: Extracting topological summaries from biological images. From a super-resolution microscopy image of actin in a cell (left), smaller patches are sampled (i). For each patch, a subsample of points is chosen, together with evenly spaced points along the boundary. From these, an increasing sequence of simplicial complexes is produced (iii-vi). From this filtered simplicial complex, the persistence diagram (vii) and persistence landscape (viii) are computed. Note that the four most persistent dots in the persistence diagram (vii) are colored and have corresponding colored representative cycles (iv).

topological features. These cycles were used to produce synthetic videos that visualized the most important characteristic behaviors of the worm contained in the video. See Figures 8, 9, and 10.

2.3.4 Topology of deep neural networks [15]

We used topological data analysis to study how data transforms as it passes through successive layers of a deep neural network. We computed the persistent homology of the activation data for each layer of the network and summarized this information using persistence landscapes. The resulting feature map provides both an informative visualization of the network and a kernel for statistical analysis and machine learning. A statistical test showed that it correlates with classification accuracy. We observed that the topological complexity often increases with training and that the topological complexity does not decrease with each layer. See Figure 11.

References

[1] Paul Bendich, Peter Bubenik, and Alexander Wagner. Stabilizing the unstable output of persistent homology computations. *J. Appl. Comput. Topol.*, 4(2):309–338, 2020.

[2] Leo Betthauser, Peter Bubenik, and Parker B. Edwards. Graded persistence diagrams and persistence landscapes. *Discrete Comput. Geom.*, 67(1):203–230, 2022.

[3] Peter Bubenik. The persistence landscape and some of its properties. In *Topological data analysis—the Abel Symposium 2018*, volume 15 of *Abel Symp.*, pages 97–117. Springer, Cham, [2020] ©2020.

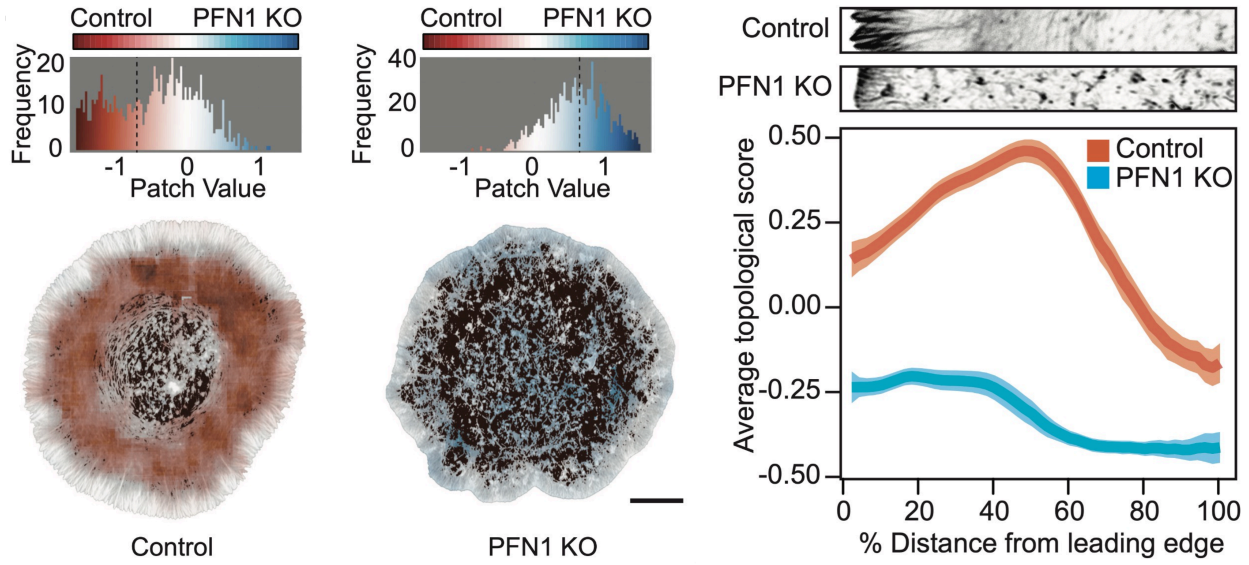


Figure 5: TDA highlights biological structures. (Left) Using persistence landscapes to summarize the shape of patches in training data, as in Figure 1, support vector regression is used to train a model to score patches labeled as controls at -1 (red) and to score patches labeled as mutants at 1 (blue). This model is applied to the persistence landscapes of patches for two test cells: a control cell (left) and a mutant cell (right). A histogram of the patch scores is given above the images of the cells. The scores of the patches containing a pixel are averaged to assign a value to the pixel which is used to color the image of the cell. Scale bar is 10 μm. (Right) We give the patch scores as a function of the distance to the leading edge (i.e. outer boundary) of the cell.

[4] Peter Bubenik and Michael J. Catanzaro. Multiparameter persistent homology via generalized morse theory. *Fields Institute Communications*, accepted, 2022.

[5] Peter Bubenik and Alex Elchesen. Universality of persistence diagrams and the bottleneck and Wasserstein distances. *Computational Geometry*, 105-106:101882, 2022.

[6] Peter Bubenik and Alex Elchesen. Virtual persistence diagrams, signed measures, Wasserstein distances, and banach spaces. *Journal of Applied and Computational Topology*, 6(4):429–474, 2022.

[7] Peter Bubenik and Iryna Hartsock. Topological and metric properties of spaces of generalized persistence diagrams. 05 2022.

[8] Peter Bubenik, Michael Hull, Dhruv Patel, and Benjamin Whittle. Persistent homology detects curvature. *Inverse Problems*, 36(2):025008, 23, 2020.

[9] Peter Bubenik and Nikola Milićević. Homological algebra for persistence modules. *Foundations of Computational Mathematics*, 21(5):1233–1278, 2021.

[10] Peter Bubenik and Nikola Milićević. Homotopy, homology, and persistent homology using closure spaces and filtered closure spaces. 04 2021.

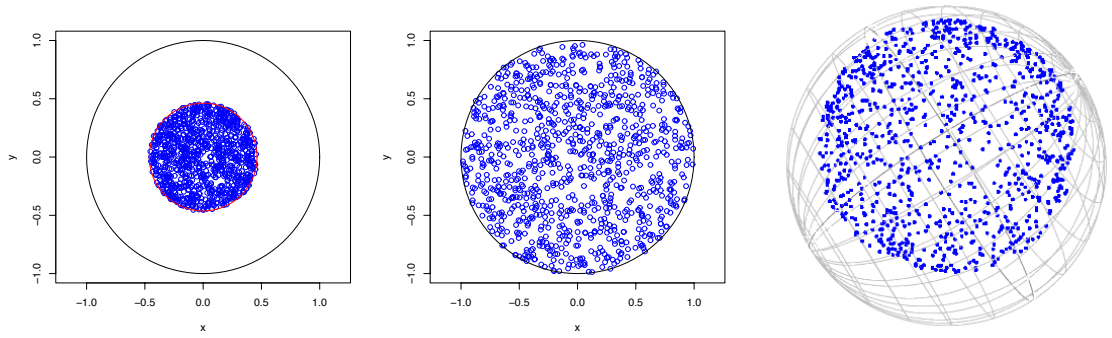


Figure 6: Plots of 1000 points sampled independently, with each point sampled uniformly with respect to area for the unit disk on the Poincaré disk model of the hyperbolic plane (left), the Euclidean plane (center), and a sphere of radius 1 (right).

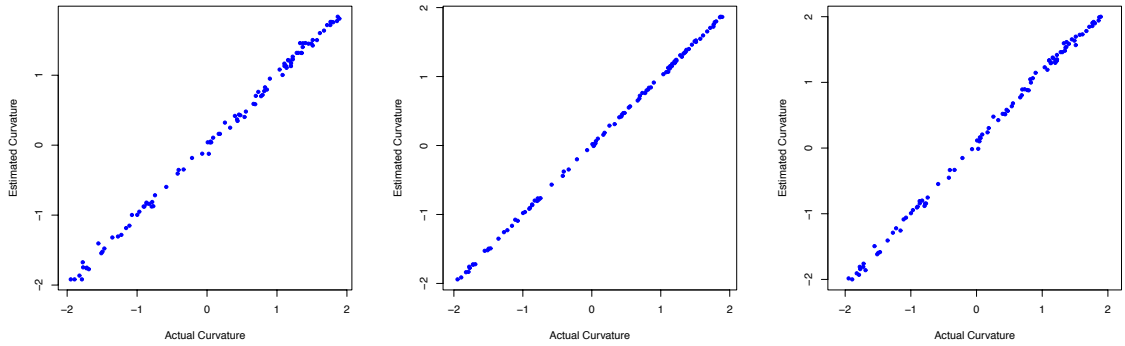


Figure 7: Plots showing actual curvature and estimated curvature using H_0 and H_1 from distance data, for nearest neighbors (left), support vector regression (center), and the first principal component (right).

[11] Peter Bubenik, Jonathan Scott, and Donald Stanley. Exact weights, path metrics, and algebraic Wasserstein distances. *Journal of Applied and Computational Topology*, 2022.

[12] Peter Bubenik and Alexander Wagner. Embeddings of persistence diagrams into Hilbert spaces. *J. Appl. Comput. Topol.*, 4(3):339–351, 2020.

[13] Parker Edwards, Kristen Skruber, Nikola Milićević, James B. Heidings, Tracy-Ann Read, Peter Bubenik, and Eric A. Vitriol. Tdaexplore: Quantitative analysis of fluorescence microscopy images through topology-based machine learning. *Patterns*, page 100367, 2021.

[14] Ashleigh Thomas, Kathleen Bates, Alex Elchesen, Iryna Hartsock, Hang Lu, and Peter Bubenik. Topological data analysis of c. elegans locomotion and behavior. *Frontiers in Artificial Intelligence*, 4:69, 2021.

[15] Matthew Wheeler, Jose Bouza, and Peter Bubenik. Activation landscapes as a topological summary of neural network performance. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3865–3870, 2021.

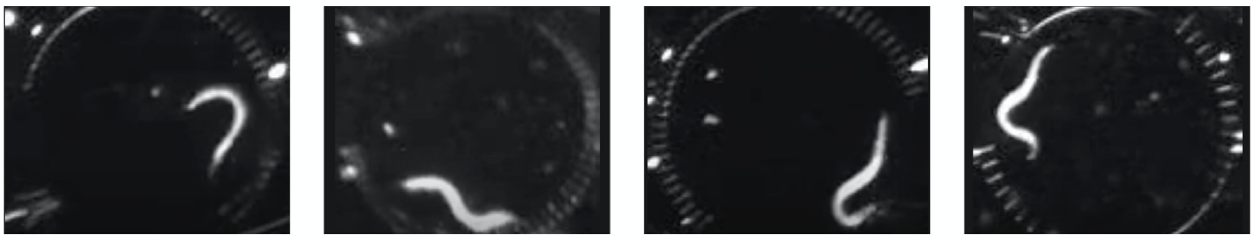


Figure 8: Sample frames from a video capturing the behavior and locomotion of *C. elegans*.

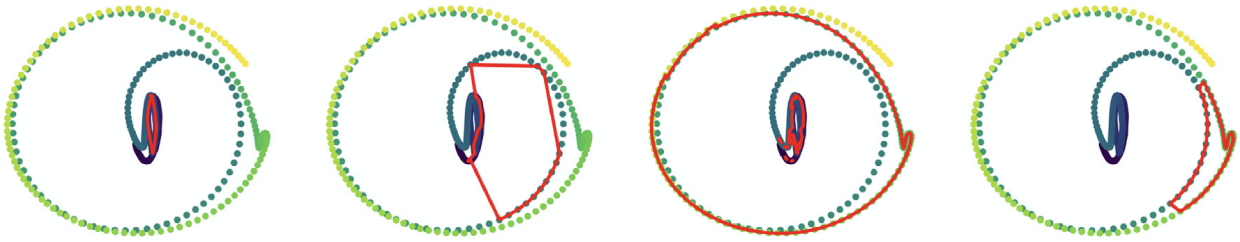


Figure 9: The cycle representatives (red) of most significant persistent homology features in the time delay embedding of the video (green to yellow dots seen in a 2D PCA projection).

Figure 10: Sample frames from the four synthetic periodic videos visualizing the characteristic behaviors in the video. From left to right: forward motion; transition from forward to backward motion; backward motion; and pause.

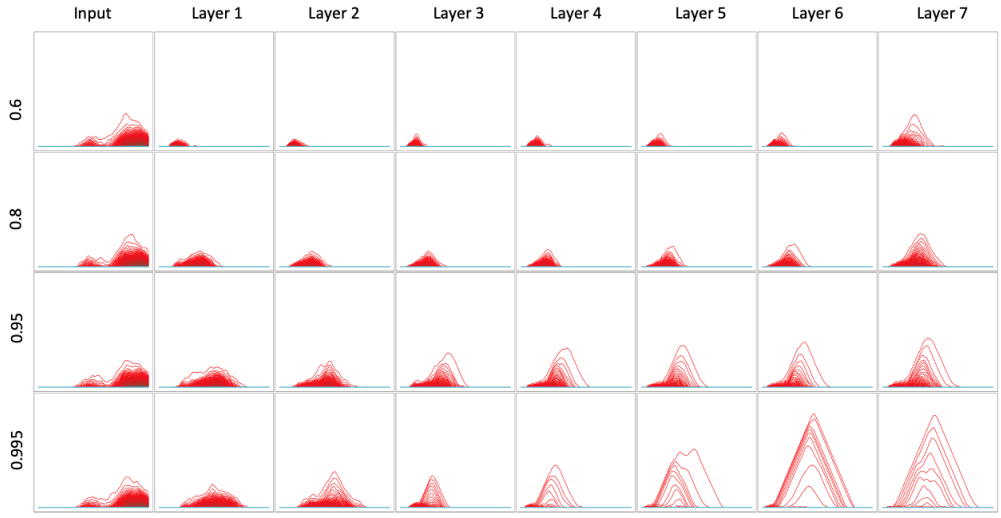


Figure 11: Activation landscapes. For the MNIST database of handwritten digits, we present the activation landscapes for local homology in degree one for each layer and various training thresholds averaged across 100 network initializations. Each row represents a training threshold and each column represents a layer of the network. The batch of input data consisting of points from all classes.