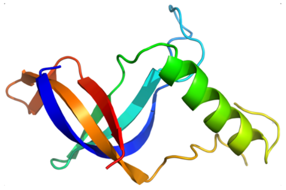
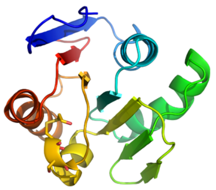
EMALCA 2021 Peru – Análisis Topológico de Datos

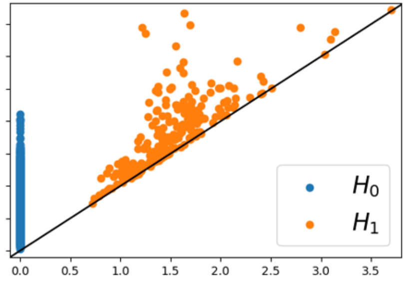
Clase 4: Classification with Persistence Diagrams

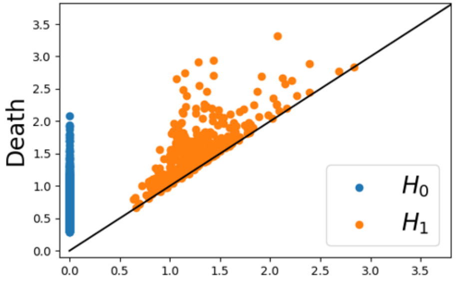
* Clustering ✓

* Series de tiempo ✓

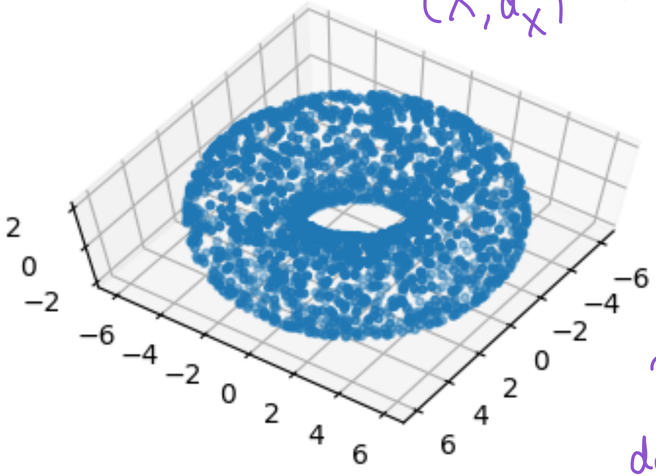


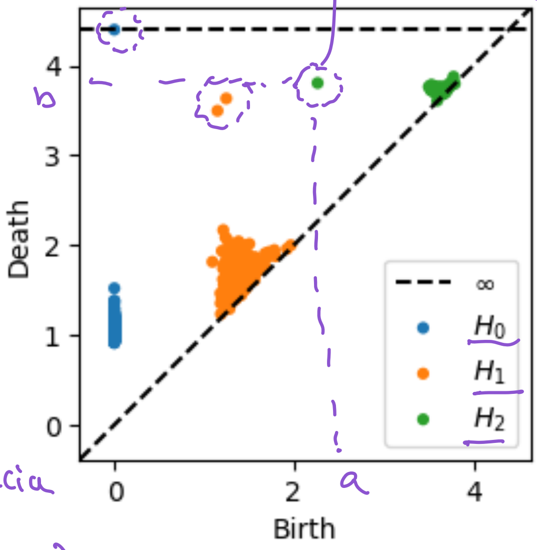




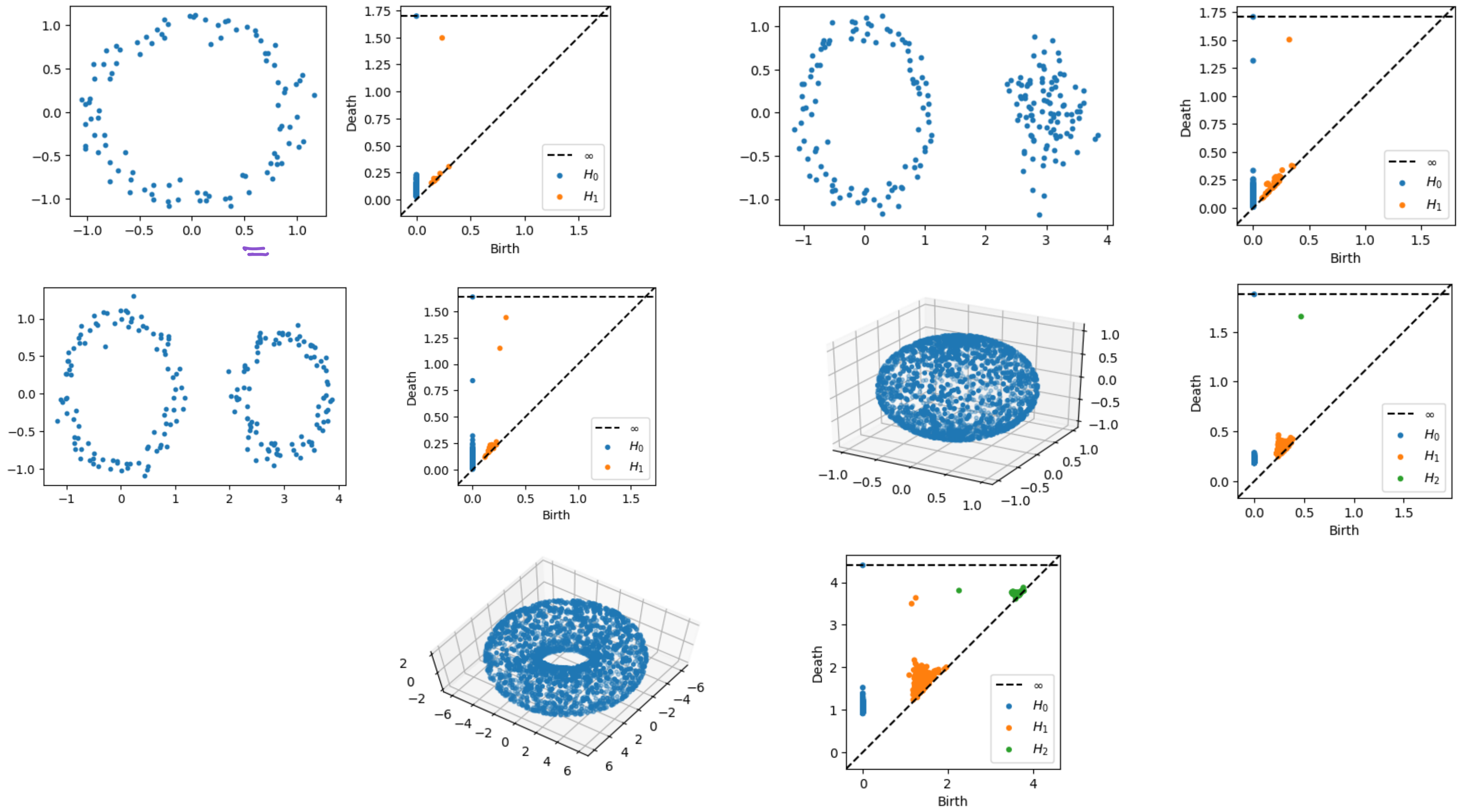


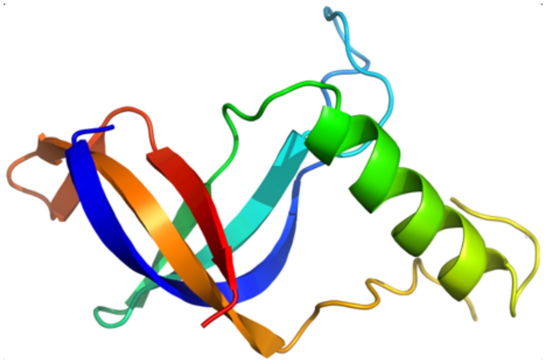
$(X, d_X) \rightarrow H_n(R_\varepsilon(X); \mathbb{Z}_p)$ ← elementos base en $a \leq \varepsilon \leq b$





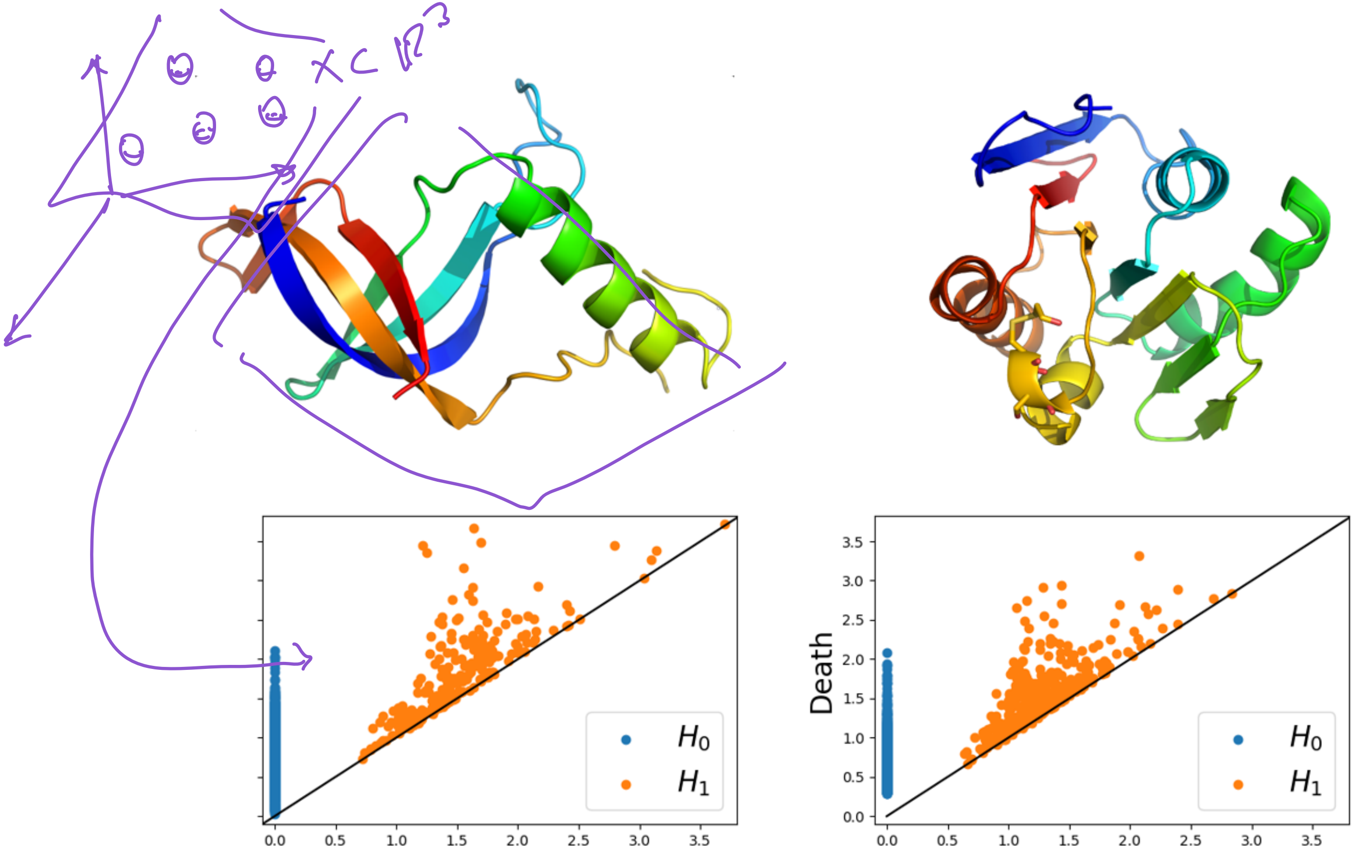
Diagramas de persistencia $dgm_n(\cdot)$


Birth
Death
∞
H0
H1
H2

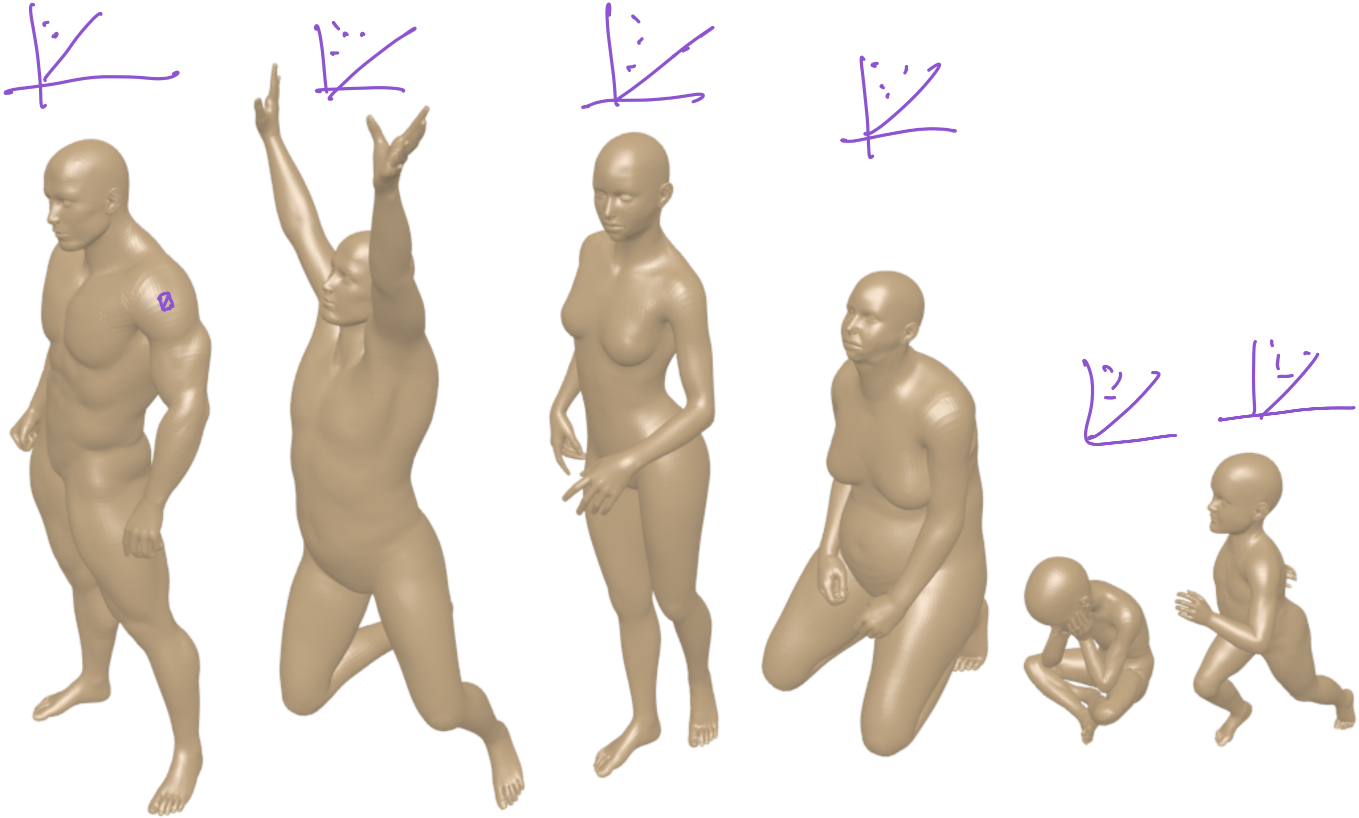


Protein Classification Benchmark Collection
(PCB00019)

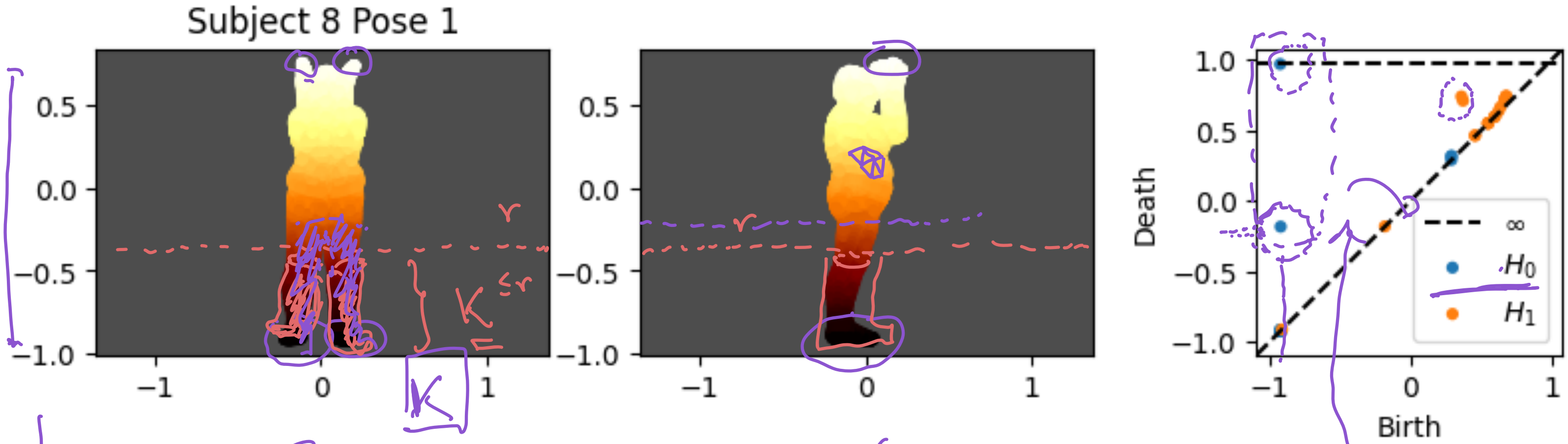
A protein classification benchmark collection for machine learning, P. Sonego et. al., Nucleic acids research, 2007.



X ⊂ ℝ³
Death
H_0
H_1
H_0
H_1
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5



Shape retrieval of non-rigid 3d human models. D. Pickup et. al., 3DOR '14,2014



paso 1: Fijar $f : K^{(0)} \to \mathbb{R}$ (color)

paso 2: Construir $K^{\leq r} = \{ \sigma \in K \mid \max_{x \in \sigma} f(x) \leq r \}$

paso 3: Calcular $H_n(K^{\leq r}; \mathbb{F})$ variando r.

A Stable Multi-Scale Kernel for Topological Machine Learning

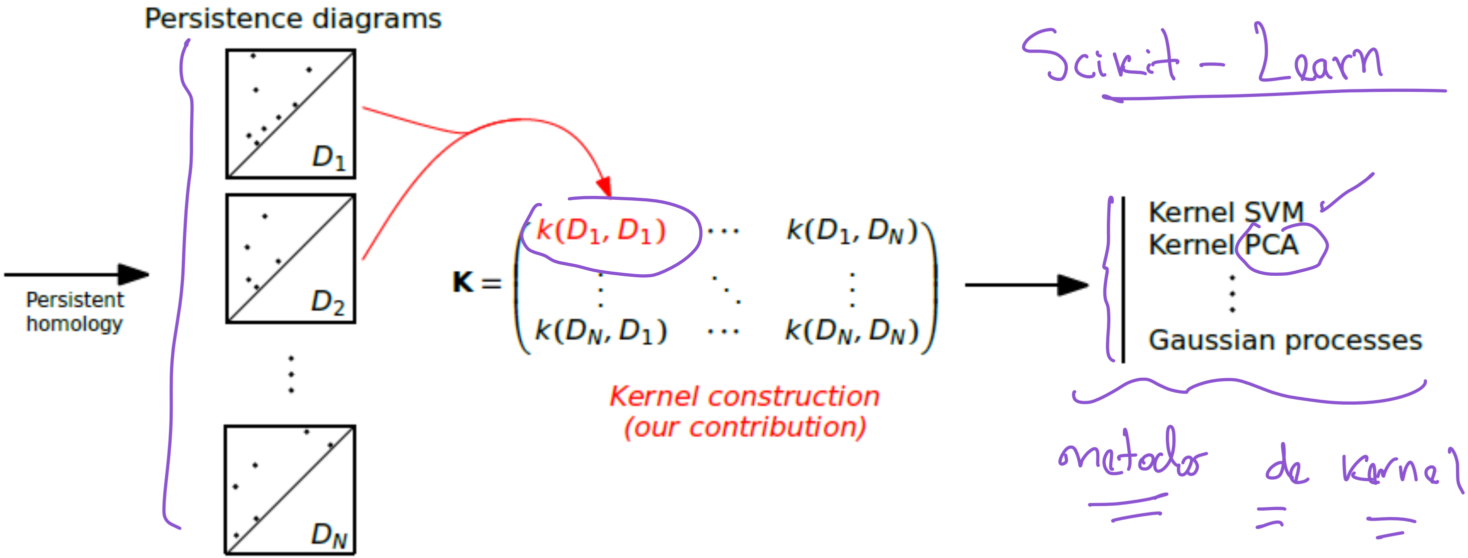
Jan Reininghaus, Stefan Huber
IST Austria

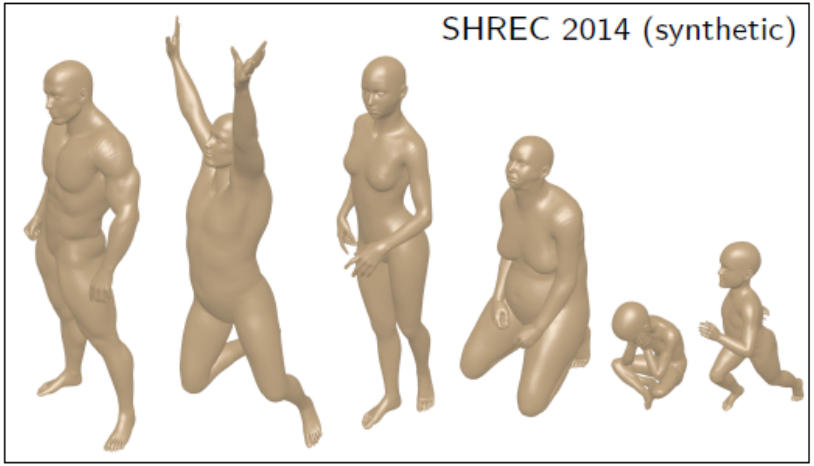
Ulrich Bauer
IST Austria, TU München

Roland Kwitt
University of Salzburg, Austria

$$K_\sigma(\mathsf{dgm}, \mathsf{dgm}') = \sum_{\substack{(a,b)\in\mathsf{dgm} \\ (x,y)\in\mathsf{dgm}'}} \exp\left(\frac{-\|(a-x, b-y)\|^2}{8\sigma}\right) - \exp\left(\frac{-\|(a-y, b-x)\|^2}{8\sigma}\right)$$

$$K_\sigma(\mathsf{dgm}, \mathsf{dgm}') = \sum_{\substack{(a,b)\in \mathsf{dgm}\\ (x,y)\in \mathsf{dgm}'}} \exp\left(\frac{-\|(a-x, b-y)\|^2}{8\sigma}\right) - \exp\left(\frac{-\|(a-y, b-x)\|^2}{8\sigma}\right)$$





HKS t_i	k^L	k_σ	Δ
t_1	68.0 ± 3.2	94.7 ± 5.1	+26.7
t_2	$\mathbf{88.3} \pm 3.3$	$\mathbf{99.3} \pm 0.9$	+11.0
t_3	61.7 ± 3.1	96.3 ± 2.2	+34.7
t_4	81.0 ± 6.5	97.3 ± 1.9	+16.3
t_5	84.7 ± 1.8	96.3 ± 2.5	+11.7
t_6	70.0 ± 7.0	93.7 ± 3.2	+23.7
t_7	73.0 ± 9.5	88.0 ± 4.5	+15.0
t_8	81.0 ± 3.8	88.3 ± 6.0	+7.3
t_9	67.3 ± 7.4	88.0 ± 5.8	+20.7
t_{10}	55.3 ± 3.6	91.0 ± 4.0	+35.7

The Annals of Applied Statistics
2016, Vol. 10, No. 1, 94–117
DOI: 10.1214/15-AOAS874

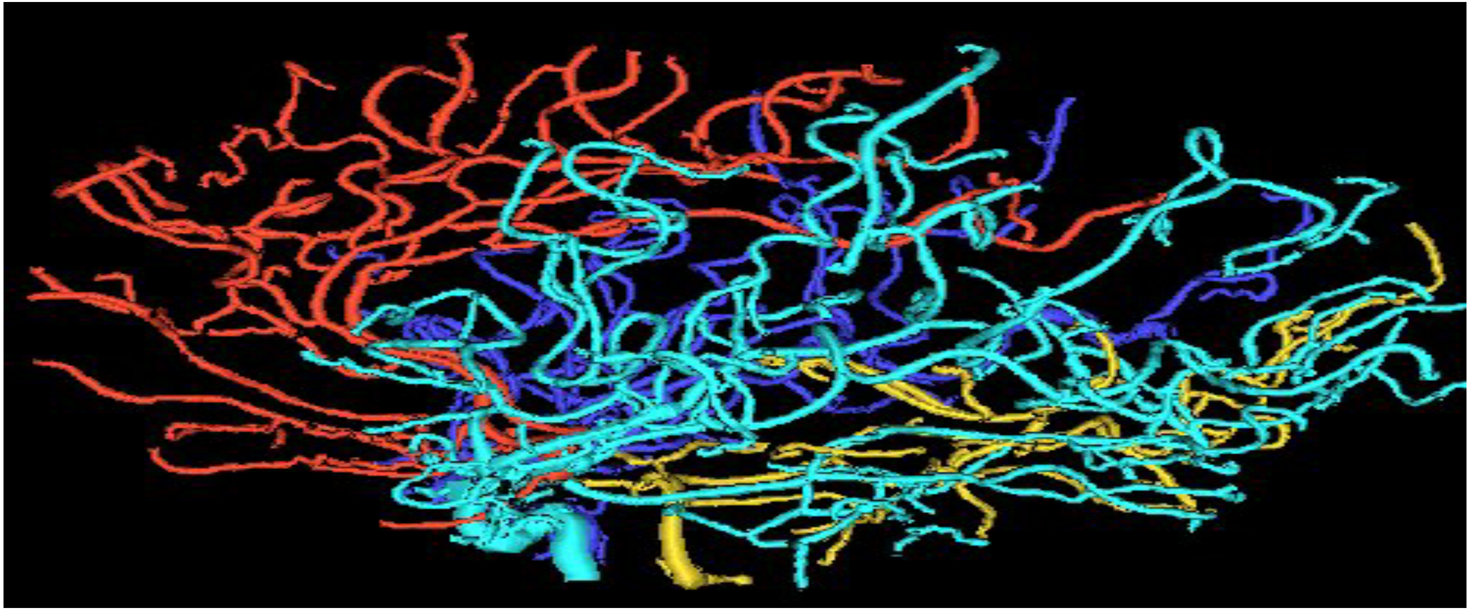
PERSISTENT HOMOLOGY ANALYSIS OF BRAIN ARTERY TREES[1]

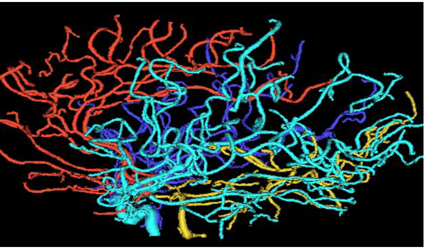
BY PAUL BENDICH*, J. S. MARRON†, EZRA MILLER*, ALEX PIELOCH* AND SEAN SKWERER‡

*Duke University**, *University of North Carolina*† *and Yale School of Public Health*‡

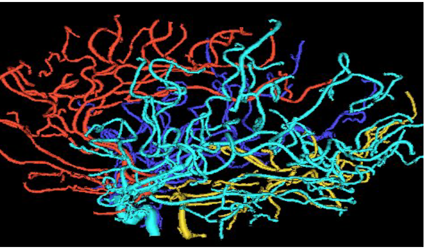
Brain Vasculature
Age (and Gender)

Bendich
Marron
Miller
Pieloch
Skwerer

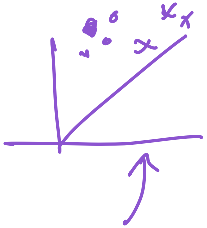




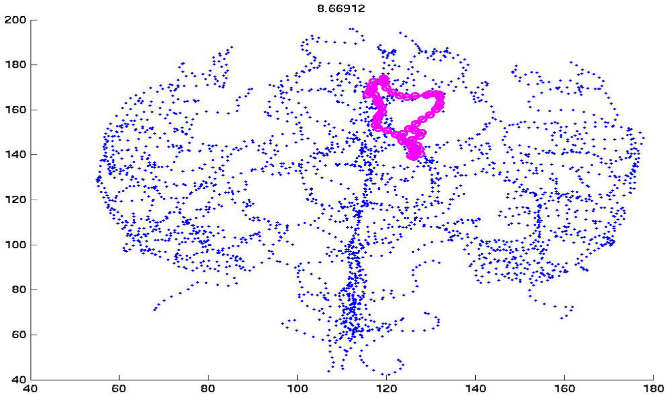
Persistence Features

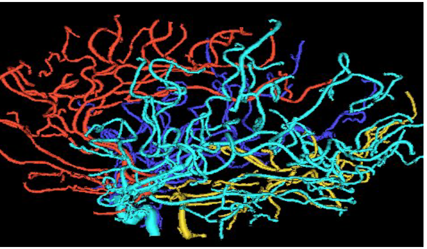


Persistence Features



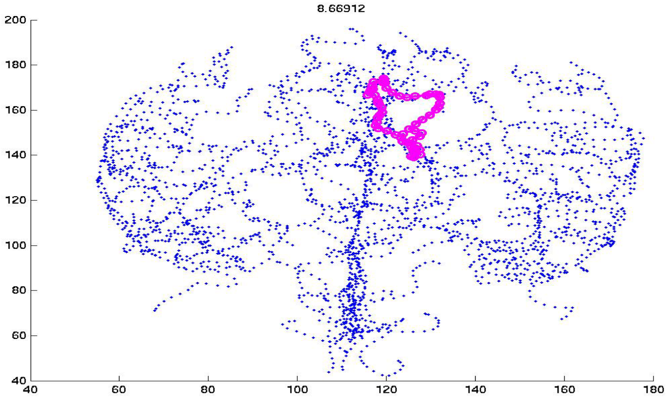
$H_n(R_\varepsilon(X);\mathbb{F})$

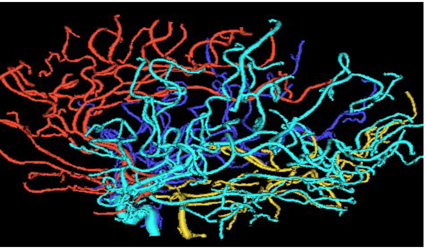




Persistence Features

0D persistence by height
1D persistence by growth

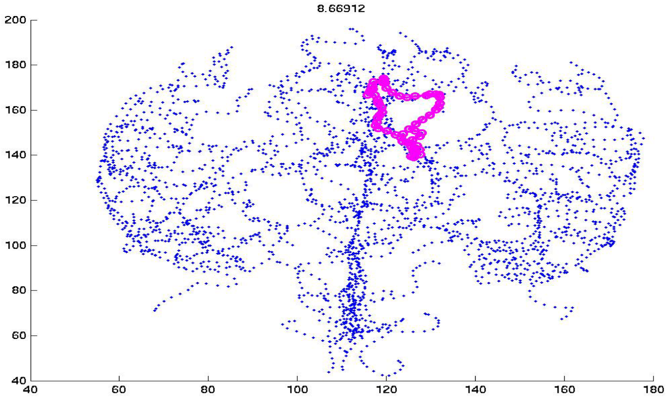


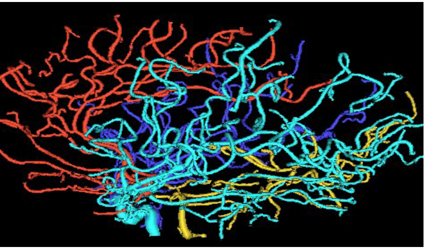


Persistence Features

Example of a 1-cycle
"Medium Size"

- 0D persistence by height
- 1D persistence by growth

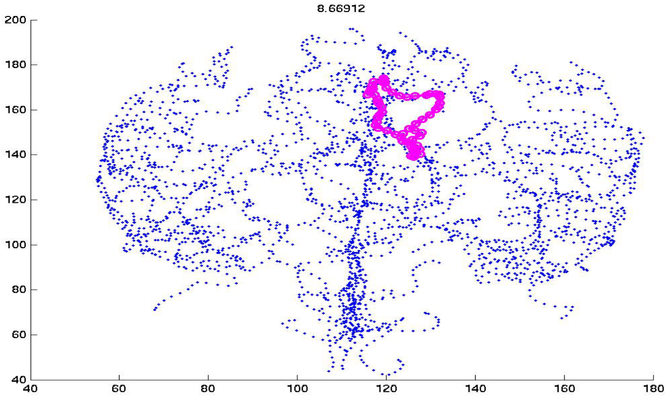




Persistence Features

Example of a 1-cycle
“Medium Size”

0D persistence by height
1D persistence by growth



Feature Extraction

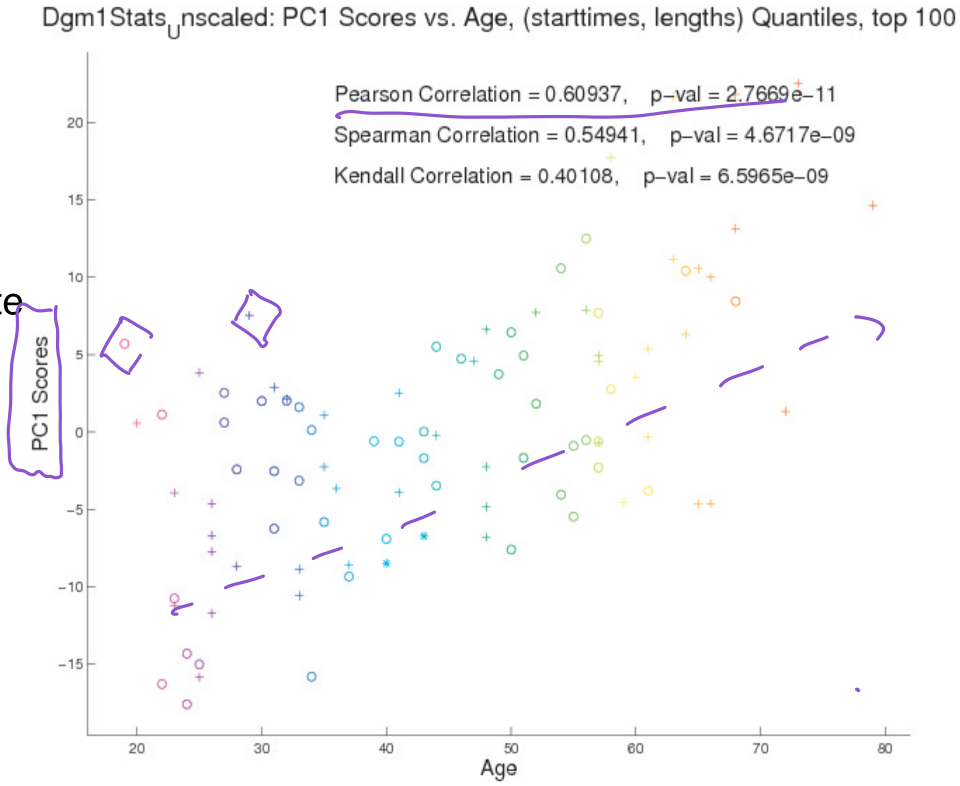
Brain ? top 100 persistences
ranked by size

Analysis

Brains -> ran PCA on these features and selected first eigenvector. Projected onto that to get a coordinate called PC1

Correlate this with age and it does really well – better than all other known methods.

Colors are age



PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions

Zixuan Cang[1], Guo-Wei Wei[1,2,3]*

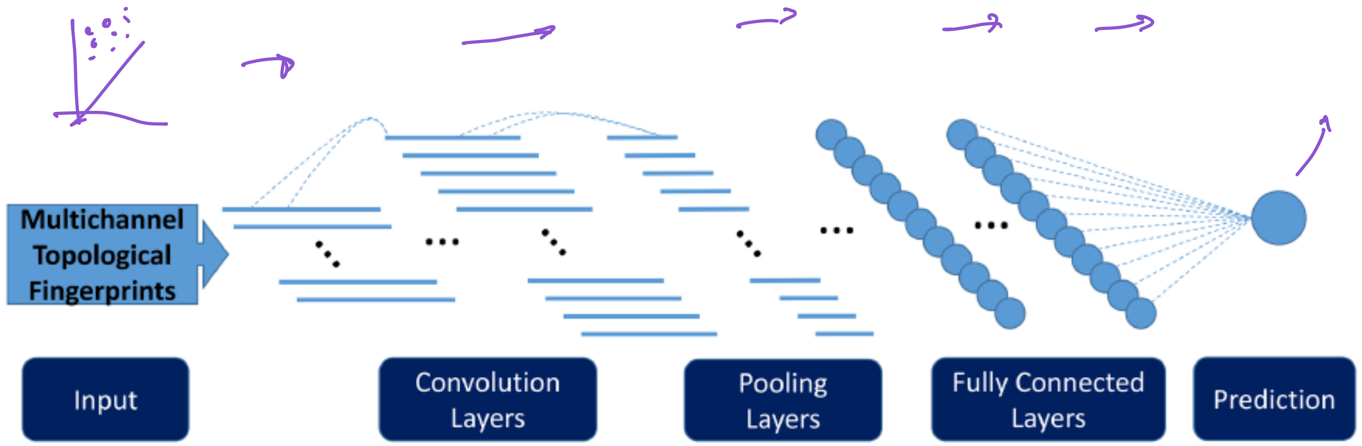


Fig 5. An illustration of the 1D convolutional neural network. The network consists of repeated convolution layers and pooling layers followed by several fully connected layers.

https://doi.org/10.1371/journal.pcbi.1005690.g005

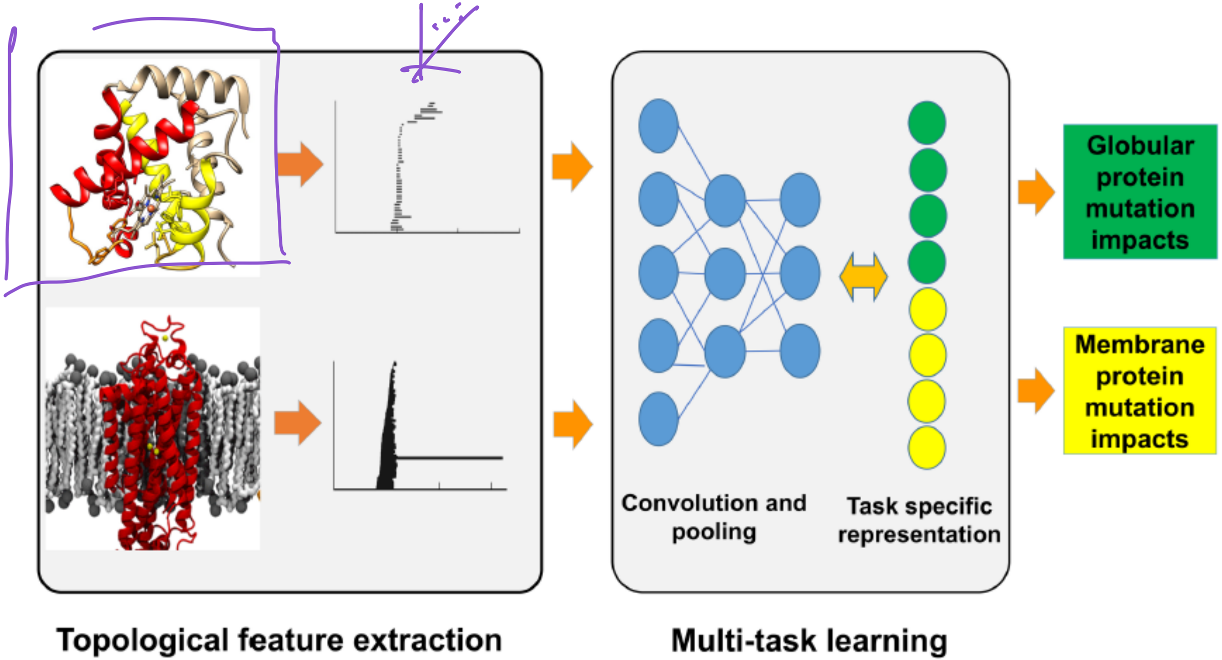


Fig 7. Workflow of the multi-task topological deep learning model. The multi-task multichannel topological convolutional neural network model shares and transforms topological information for the simultaneous training and prediction of globular protein and membrane protein mutation impacts on protein stability.

https://doi.org/10.1371/journal.pcbi.1005690.g007

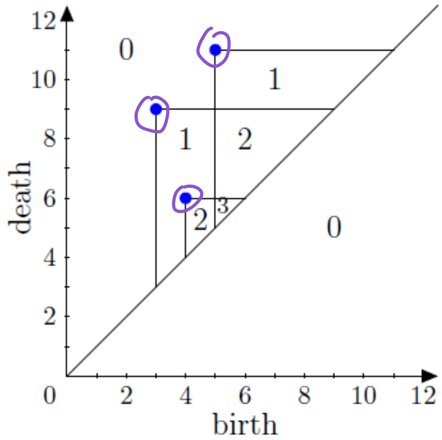
Table 1. Performance comparisons of TNet-BP and other methods.

Method	R_P	RMSE
TNet-BP	0.826[a]	1.37
RF::VinaElem	0.803	1.42
RF:Vina	0.739	1.61
Cyscore	0.660	1.79
X-Score::HMScore	0.644	1.83
MLR::Vina	0.622	1.87
HYDE2.0::HbondsHydrophobic	0.620	1.89
DrugScore	0.569	1.96
SYBYL::ChemScore	0.555	1.98
AutoDock Vina	0.554	1.99
DS::PLP1	0.545	2.00
GOLD::ASP	0.534	2.02
SYBYL::G-Score	0.492	2.08
DS::LUDI3	0.487	2.09
DS:LigScore2	0.464	2.12
GlideScore-XP	0.457	2.14

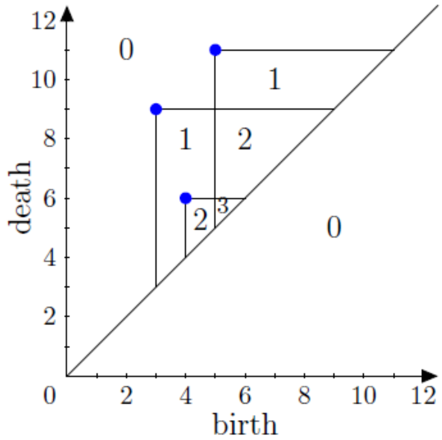
Journal of Machine Learning Research 16 (2015) 77-102 Submitted 7/14; Published 1/15

Statistical Topological Data Analysis using Persistence Landscapes

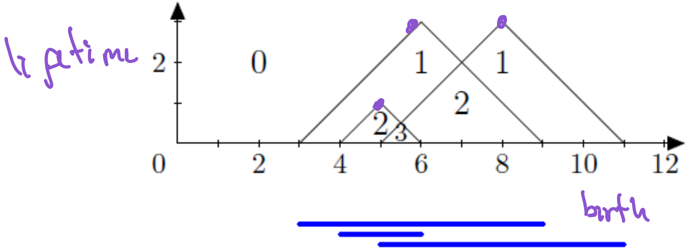
Peter Bubenik PETER.BUBENIK@GMAIL.COM
Department of Mathematics
Cleveland State University
Cleveland, OH 44115-2214, USA



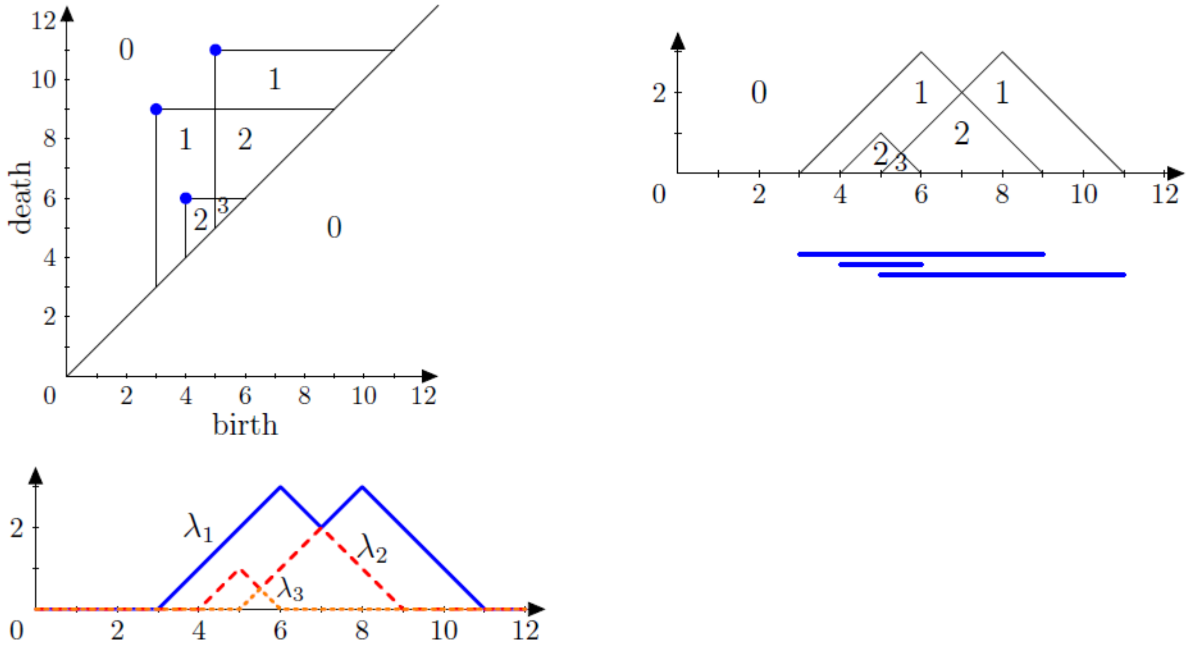
12
10
8
6
4
2
0
0
2
4
6
8
10
12
death
birth
0
1
1
2
2
3
0

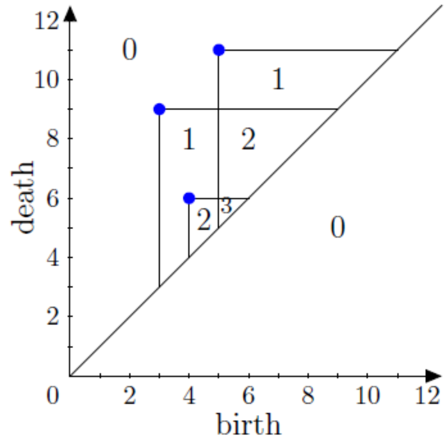


death
birth

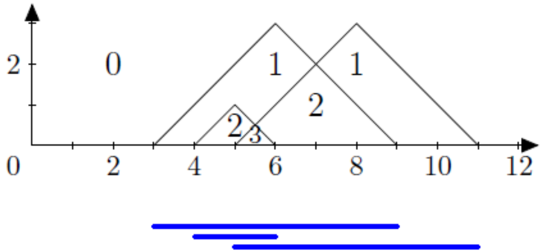


birth

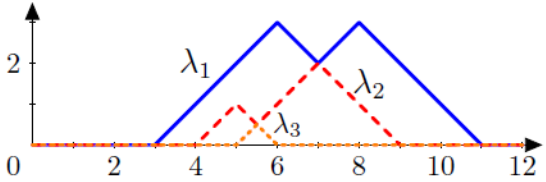

death
birth
0
1
2
3
λ1
λ2
λ3



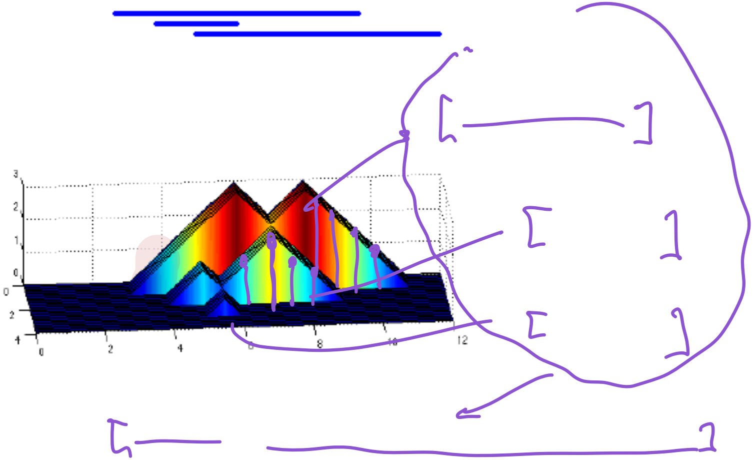
death
birth
0 2 4 6 8 10 12
0
1
1 2
2 3
0



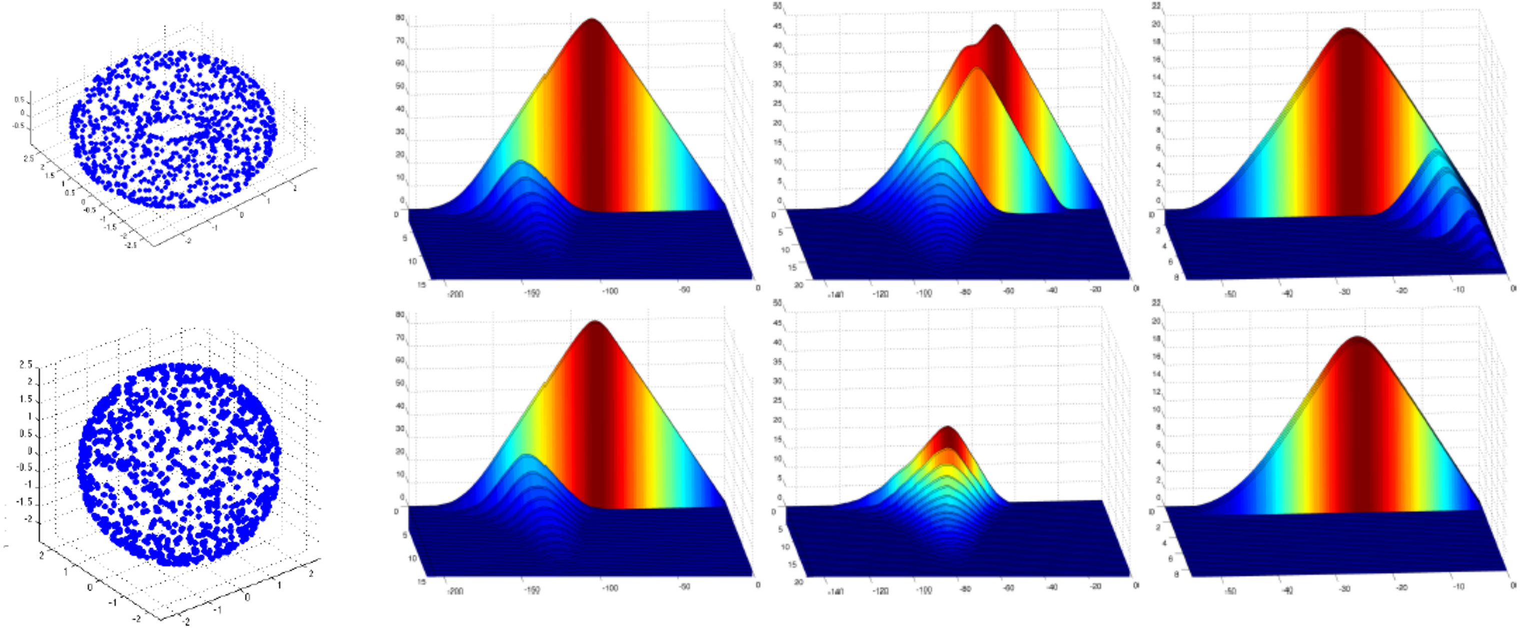
0 2 4 6 8 10 12
2
0
1 1
2
2 3



0 2 4 6 8 10 12
2
λ_1
λ_2
λ_3



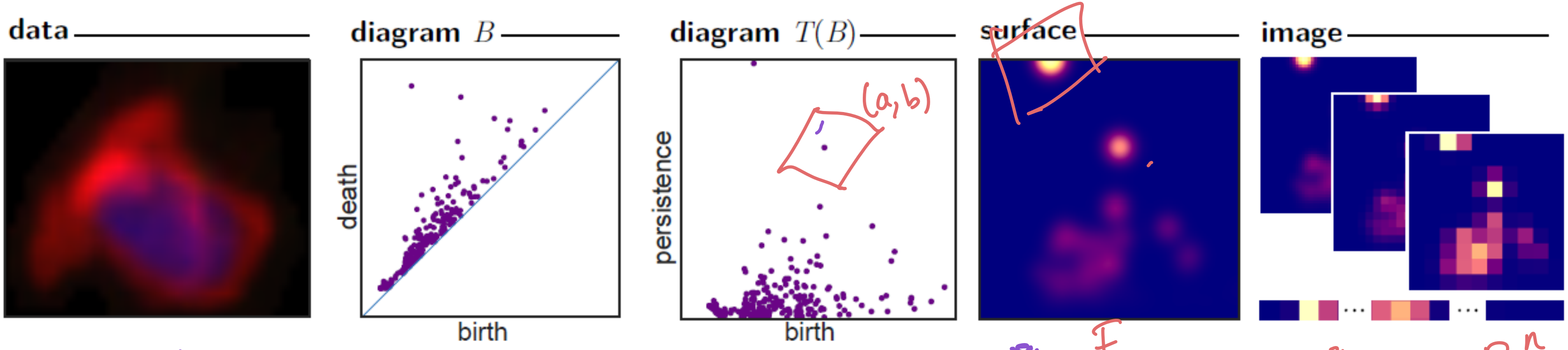
dgm_0 dgm_1 dgm_2



Journal of Machine Learning Research 18 (2017) 1-35 Submitted 7/16; Published 2/17

Persistence Images: A Stable Vector Representation of Persistent Homology

Henry Adams ADAMS@MATH.COLOSTATE.EDU
Tegan Emerson EMERSON@MATH.COLOSTATE.EDU
Michael Kirby KIRBY@MATH.COLOSTATE.EDU
Rachel Neville NEVILLE@MATH.COLOSTATE.EDU
Chris Peterson PETERSON@MATH.COLOSTATE.EDU
Patrick Shipman SHIPMAN@MATH.COLOSTATE.EDU



F_α $\in \mathbb{R}^n$

(x, d_x)

K, $f: K^{(0)} \to \mathbb{R}$

fija $\alpha > 0$ (kernel bandwith Ancho de Banda)

$$F_\alpha(x,y) = \frac{1}{\#(dgm)\cdot\alpha} \sum_{(a,b)\in dgm} b \cdot e^{-\|(x,y)-(a,b)\|^2/\alpha}$$