Topological Machine Learning and Its Applications

Baris Coskunuzer

UT Dallas
Mathematics Department

December 2023

Topological Data Analysis

Aim:

To study the shape of the data, and obtain a unique fingerprint of its topological features.

Topological Data Analysis

Aim:

To study the shape of the data, and obtain a unique fingerprint of its topological features.

TDA methods are highly effective on various forms of data.

Topological Data Analysis

Aim:

To study the shape of the data, and obtain a unique fingerprint of its topological features.

TDA methods are highly effective on various forms of data.

- Point Clouds in High Dimensions

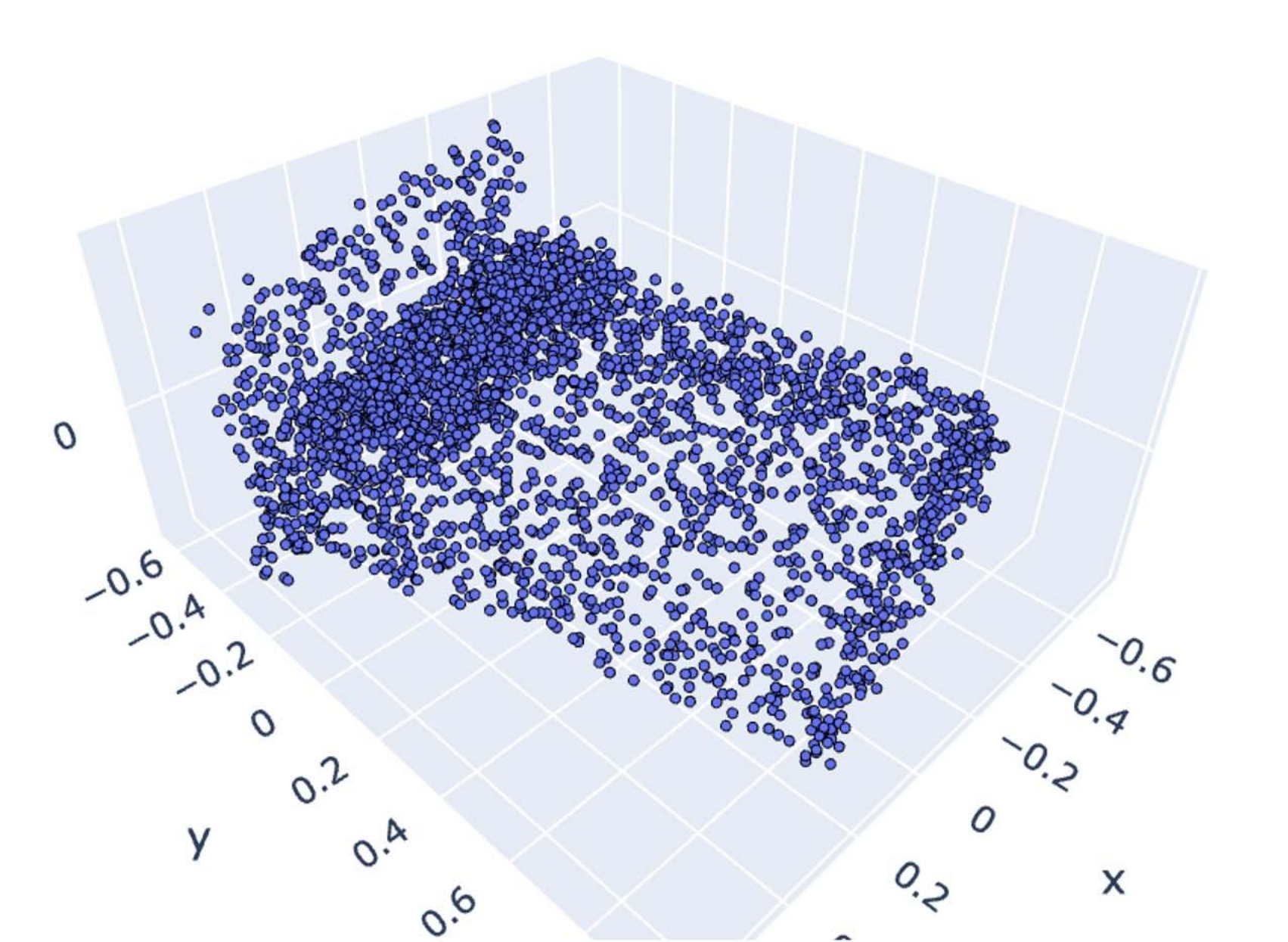

0
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
y
−0.6
−0.4
−0.2
0
0.2
x

Image credit H. Adams

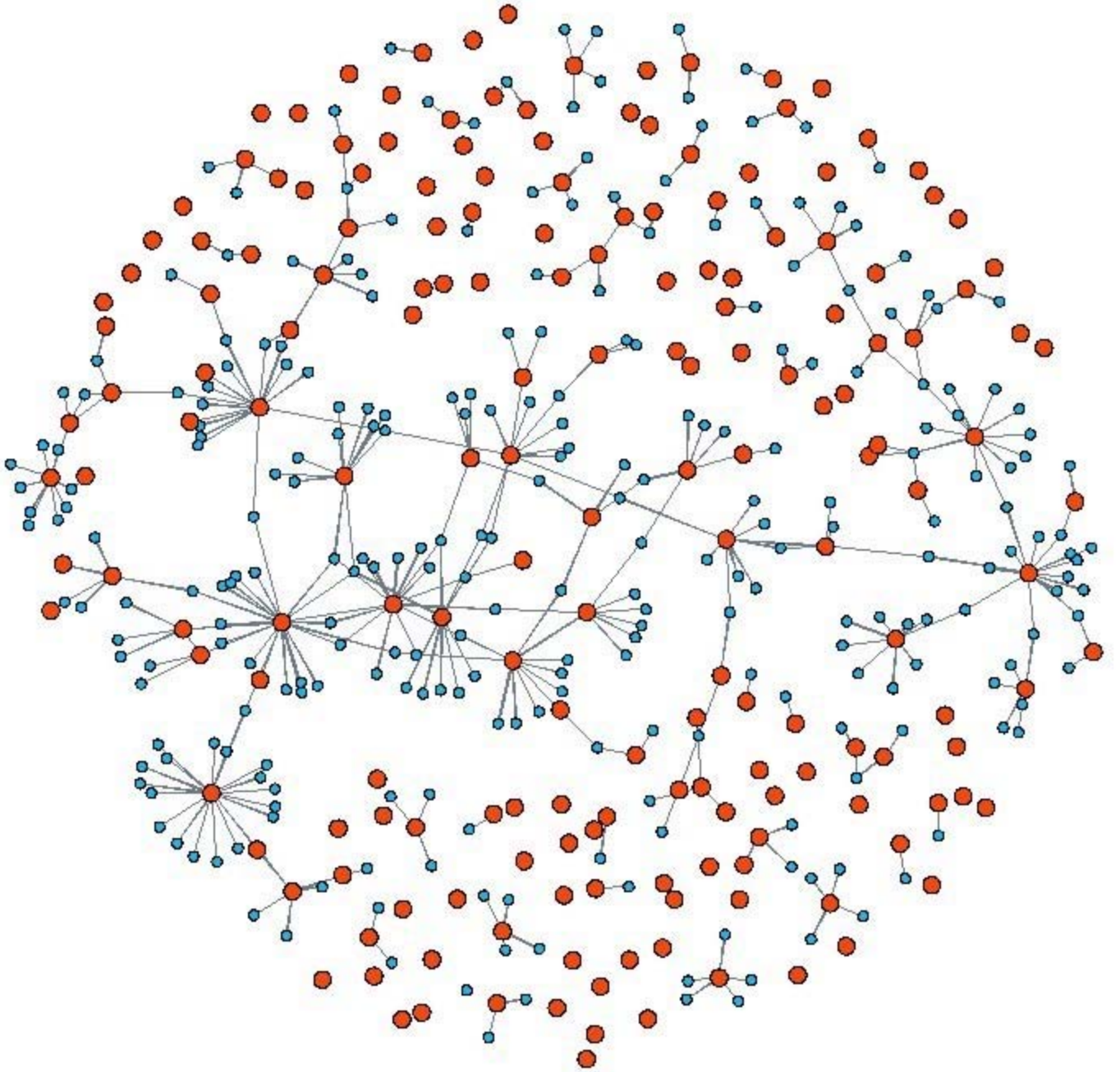
Topological Data Analysis

Aim:

To study the shape of the data, and obtain a unique fingerprint of its topological features.

TDA methods are highly effective on various forms of data.

- Point Clouds in High Dimensions
- Graphs and Networks



Topological Data Analysis

Aim:

To study the shape of the data, and obtain a unique fingerprint of its topological features.

TDA methods are highly effective on various forms of data.

- Point Clouds in High Dimensions
- Graphs and Networks
- Images

airplane

automobile

bird

cat

deer

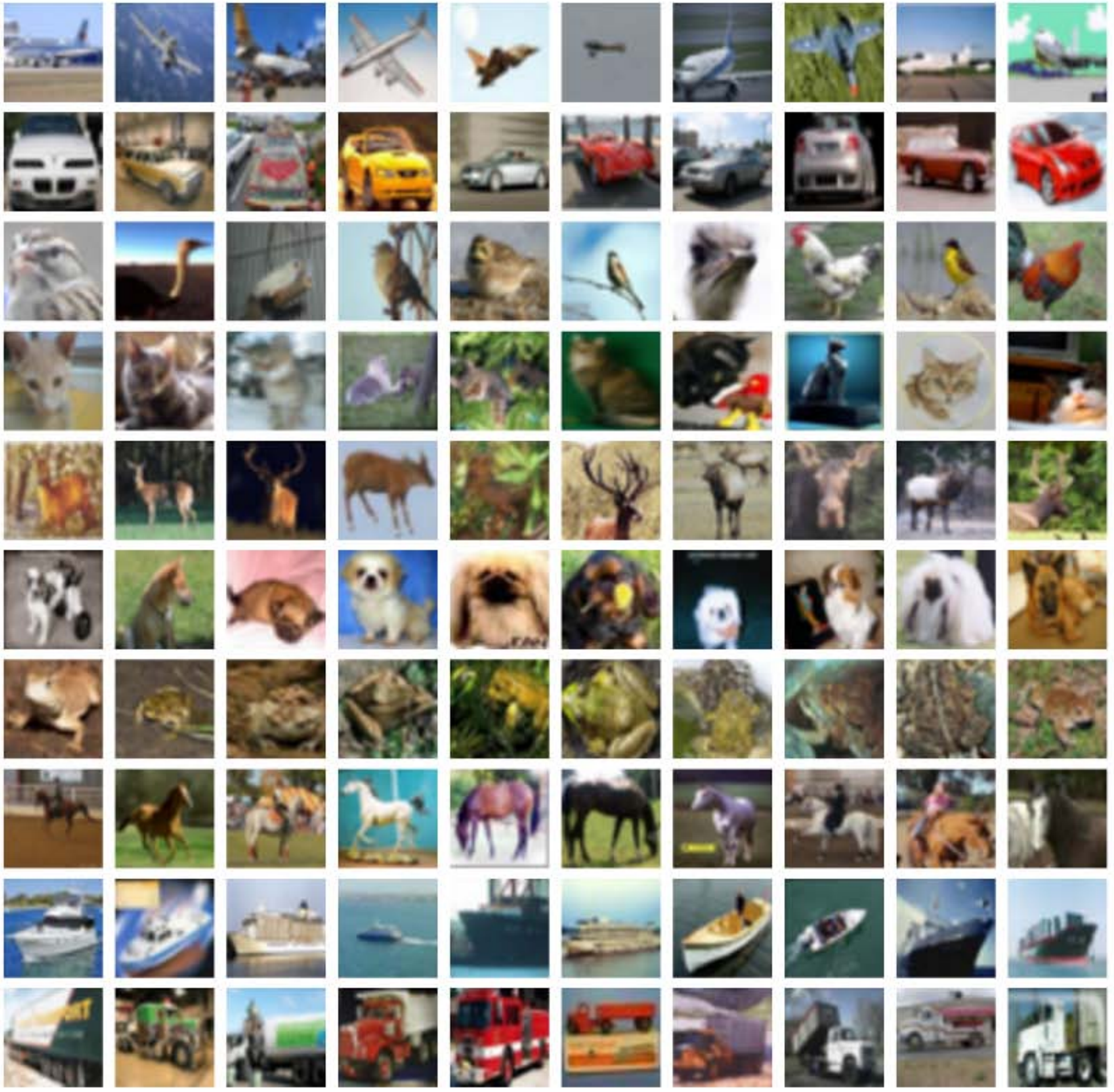
dog

frog

horse

ship

truck



Topological Data Analysis

Aim:

To study the shape of the data, and obtain a unique fingerprint of its topological features.

TDA methods are highly effective on various forms of data.

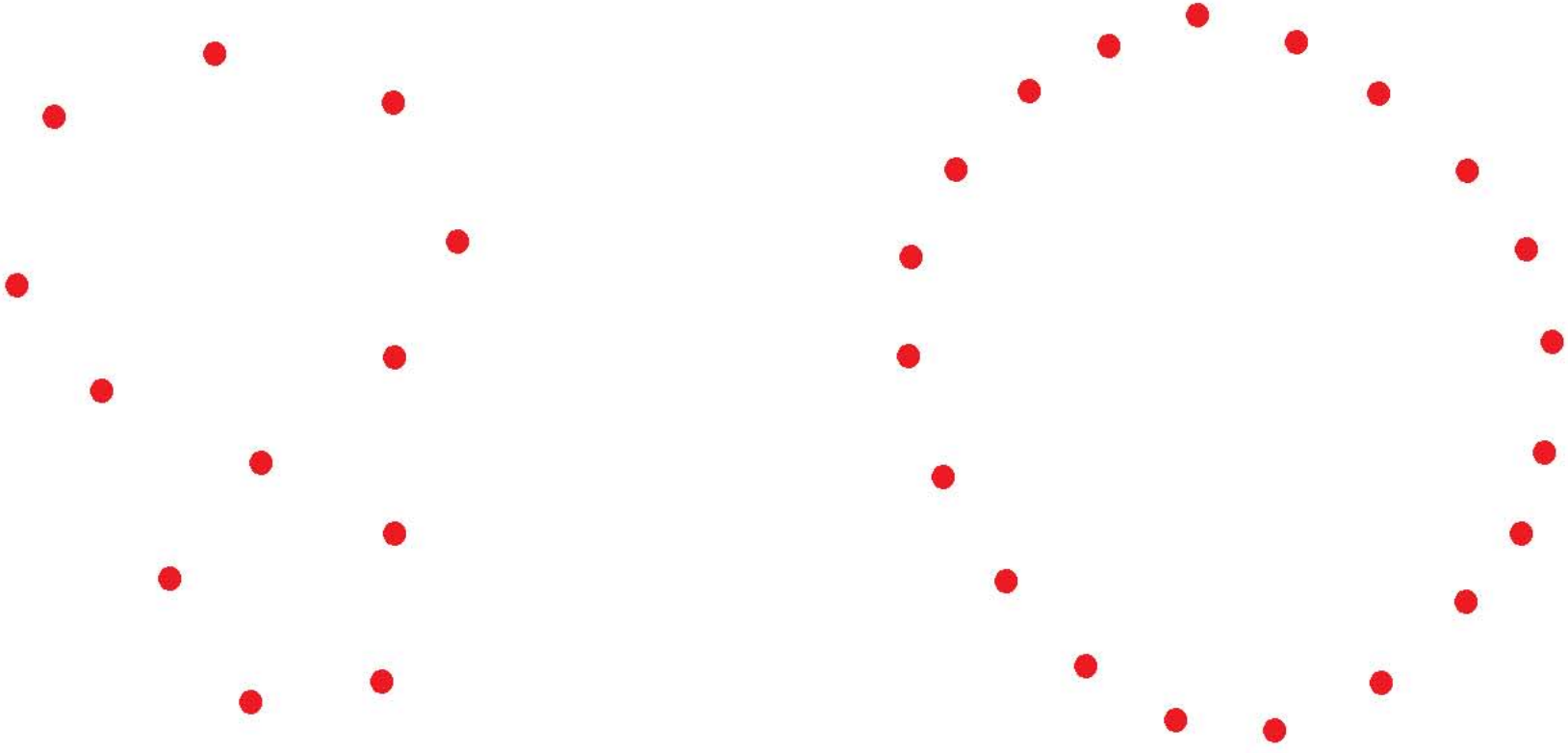
- Point Clouds in High Dimensions
- Graphs and Networks
- Images

Applications in different fields:

Bioinformatics, computational biology, finance, image recognition, material science, network analysis, combining with deep learning.

Shapes hidden in the Data

- Want to find topological patterns hidden in the data.



Shapes hidden in the Data

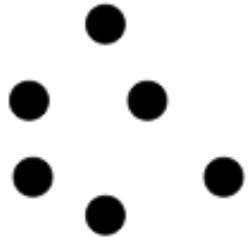
- Want to find topological patterns hidden in the data.
- Topological Features can be detected by using *Homology*.

Shapes hidden in the Data

- Want to find topological patterns hidden in the data.
- Topological Features can be detected by using *Homology*.
  - 0-dimensional features: *Components*
  - 1-dimensional features: *Holes*
  - 2-dimensional features: *Cavities*

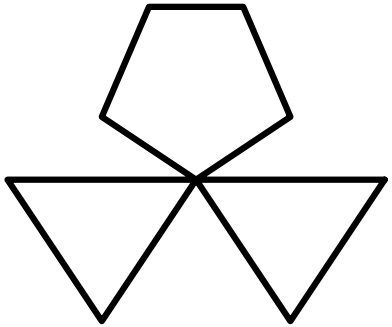
Homology

- i-dimensional homology H_i "counts the number of i-dimensional holes"
- i-dimensional homology H_i actually has the structure of a vector space!



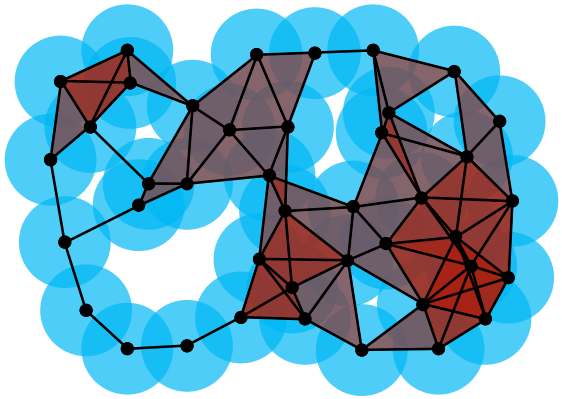
0-dimensional homology H_0: rank 6

1-dimensional homology H_1: rank 0



0-dimensional homology H_0: rank 1

1-dimensional homology H_1: rank 3

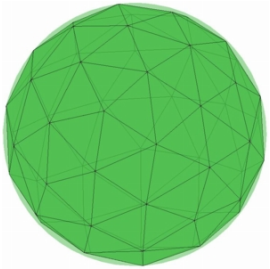


0-dimensional homology H_0: rank 1

1-dimensional homology H_1: rank 6

Homology

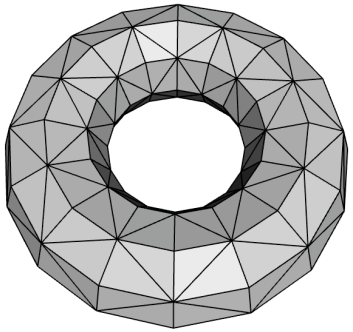
- *i*-dimensional homology "counts the number of *i*-dimensional holes"
- *i*-dimensional homology actually has the structure of a vector space!



0-dimensional homology H_0: rank 1

1-dimensional homology H_1: rank 0

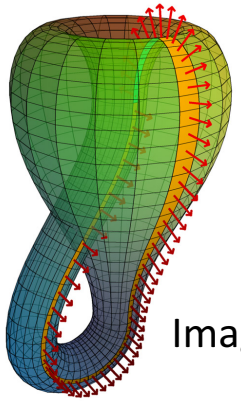
2-dimensional homology H_2: rank 1



0-dimensional homology H_0: rank 1

1-dimensional homology H_1: rank 2

2-dimensional homology H_2: rank 1



Be careful! (Same as torus over $\mathbb{Z}/2\mathbb{Z}$)

Image credit: https://plus.maths.org/content/imaging-maths-inside-klein-bottle

Shapes hidden in the Data

- Want to find topological patterns hidden in the data.
- Topological Features can be detected by using *Homology*.
  - 0-dimensional features: *Components*
  - 1-dimensional features: *Holes*
  - 2-dimensional features: *Cavities*
- Similar shape / topological patterns $\Rightarrow$ Same Class

Shapes hidden in the Data

- Want to find topological patterns hidden in the data.
- Topological Features can be detected by using *Homology*.
  - 0-dimensional features: *Components*
  - 1-dimensional features: *Holes*
  - 2-dimensional features: *Cavities*
- Similar shape / topological patterns $\Rightarrow$ Same Class
- How to obtain **a formal summary** of these topological features?

Persistent Homology

- *Persistent Homology* is one of the main methods of TDA.

Persistent homology

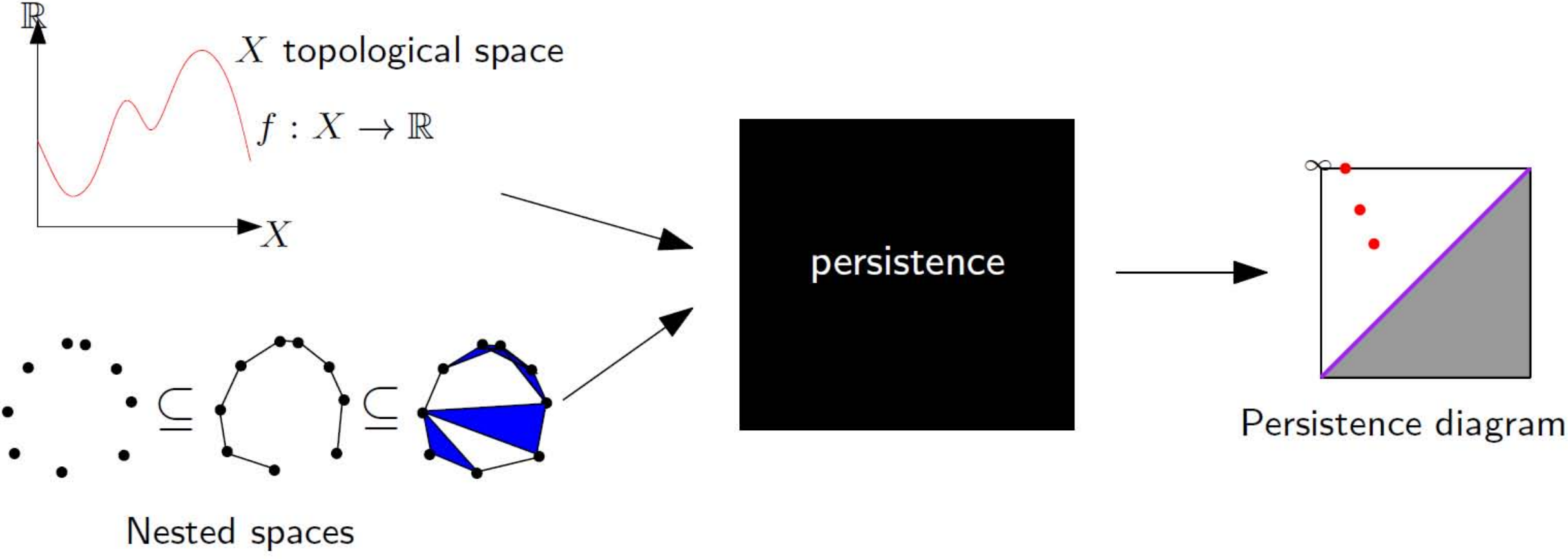


Image credit F. Chazal - B. Michel

Persistent Homology

- *Persistent Homology* is one of the main methods of TDA.
- It coarsely detects connected components, loops and cavities hidden in the shape along with their ”sizes”.

Persistent Homology

- *Persistent Homology* is one of the main methods of TDA.
- It coarsely detects connected components, loops and cavities hidden in the shape along with their ”sizes”.
- It is used in ML as a very powerful feature extraction method to capture shape patterns in the data.

Persistent Homology

- *Persistent Homology* is one of the main methods of TDA.
- It coarsely detects connected components, loops and cavities hidden in the shape along with their ”sizes”.
- It is used in ML as a very powerful feature extraction method to capture shape patterns in the data.
- PH is a 3-step process.

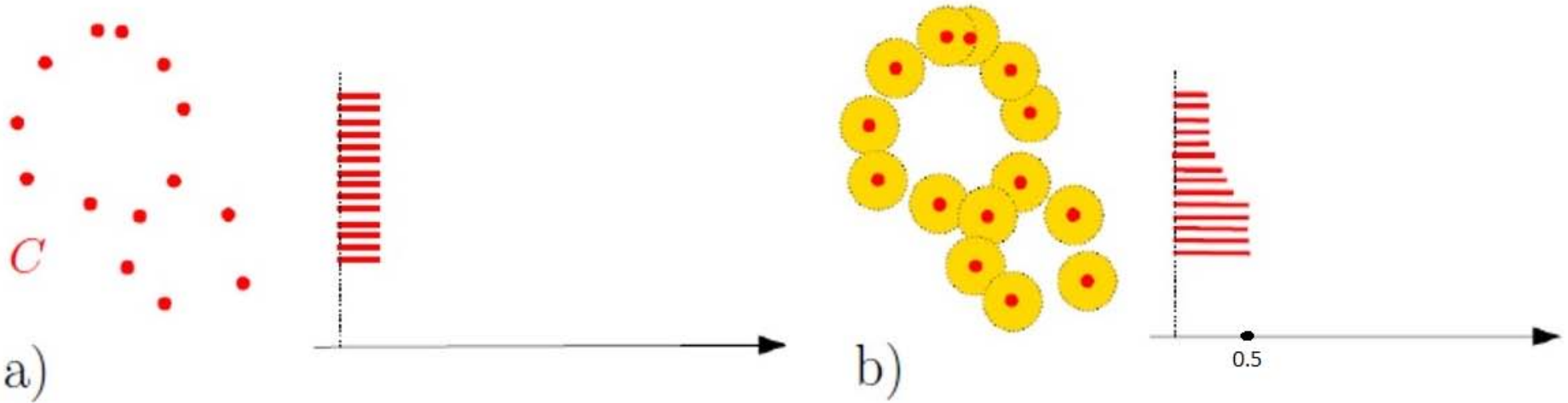
Persistent Homology

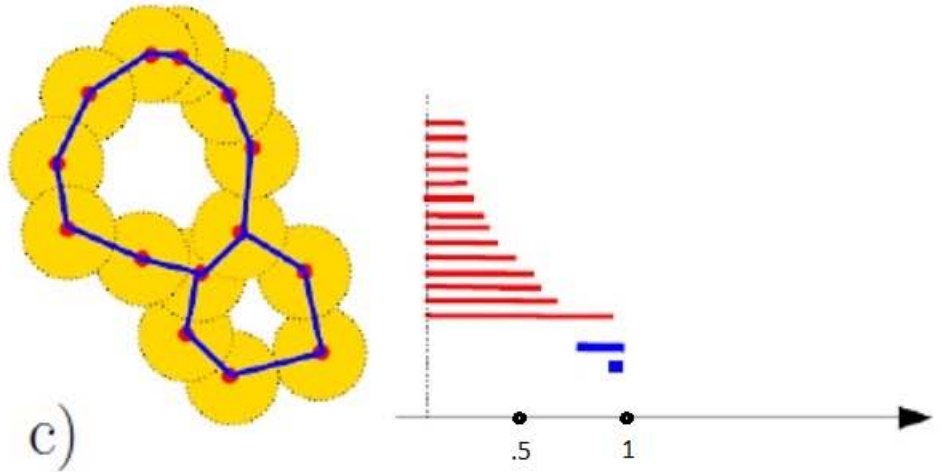
- *Persistent Homology* is one of the main methods of TDA.
- It coarsely detects connected components, loops and cavities hidden in the shape along with their ”sizes”.
- It is used in ML as a very powerful feature extraction method to capture shape patterns in the data.
- PH is a 3-step process.
  - **Step 1 - Constructing Filtration:** To obtain a sequence of simplicial complexes $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \cdots \subset \mathcal{X}_N$.

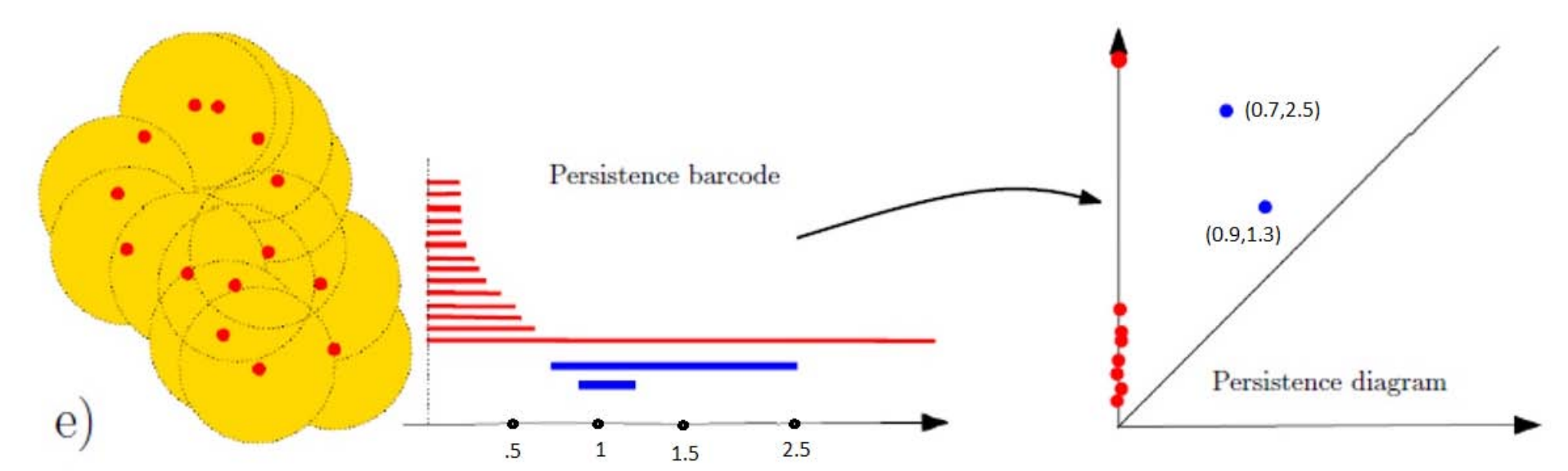
Persistent Homology

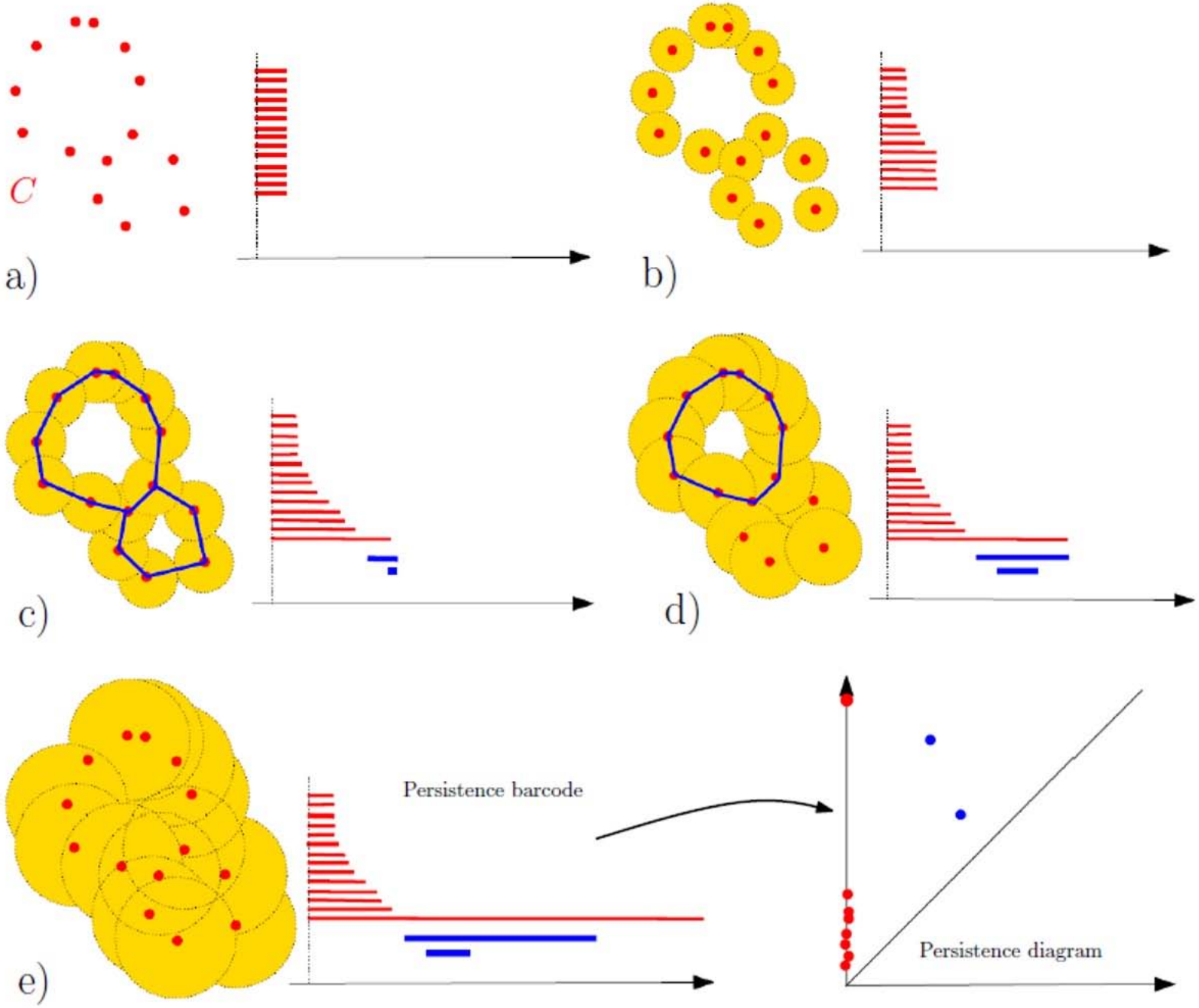
- *Persistent Homology* is one of the main methods of TDA.
- It coarsely detects connected components, loops and cavities hidden in the shape along with their ”sizes”.
- It is used in ML as a very powerful feature extraction method to capture shape patterns in the data.
- PH is a 3-step process.
  - **Step 1 - Constructing Filtration:** To obtain a sequence of simplicial complexes $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \cdots \subset \mathcal{X}_N$.
  - **Step 2 - Persistence Diagram:** Record the topological changes in the sequence $\{\mathcal{X}_i\}$.

Construction of Persistence Diagram



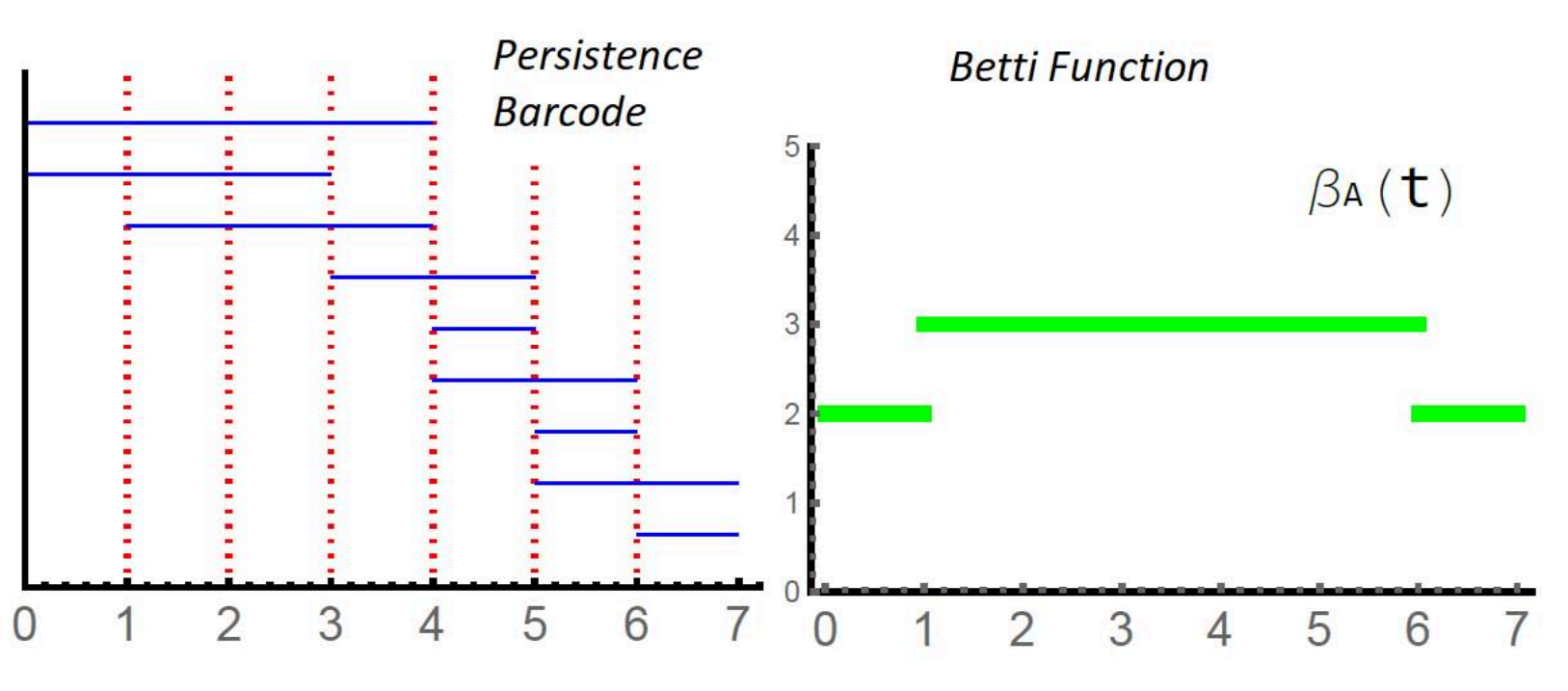

c)
.5
1
d)
.5
1
1.5

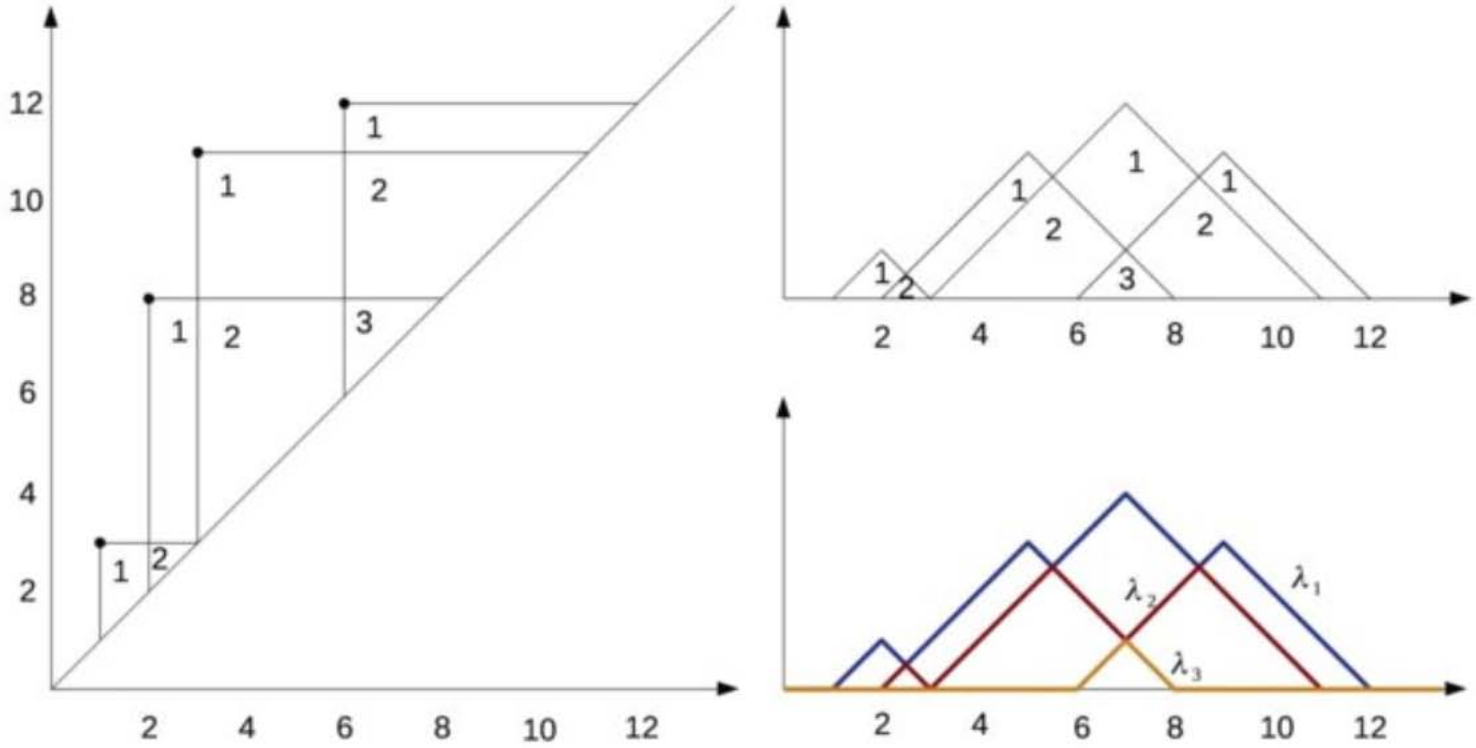

e)
Persistence barcode
.5
1
1.5
2.5
(0.7,2.5)
(0.9,1.3)
Persistence diagram


C
a)
b)
c)
d)
e)
Persistence barcode
Persistence diagram

Persistent Homology

- *Persistent Homology* is one of the main methods of TDA.
- It coarsely detects connected components, loops and cavities hidden in the shape along with their ”sizes”.
- It is used in ML as a very powerful feature extraction method to capture shape patterns in the data.
- PH is a 3-step process.
  - **Step 1 - Constructing Filtration:** To obtain a sequence of simplicial complexes $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \cdots \subset \mathcal{X}_N$.
  - **Step 2 - Persistence Diagram:** Record the topological changes in the sequence $\{\mathcal{X}_i\}$.
  - **Step 3 - Vectorization:** Convert PDs into vectors. Topological Feature Vectors


Persistence
Barcode
0
1
2
3
4
5
6
7
Betti Function
$\beta_A(t)$
0
1
2
3
4
5
6
7



Persistence Landscape

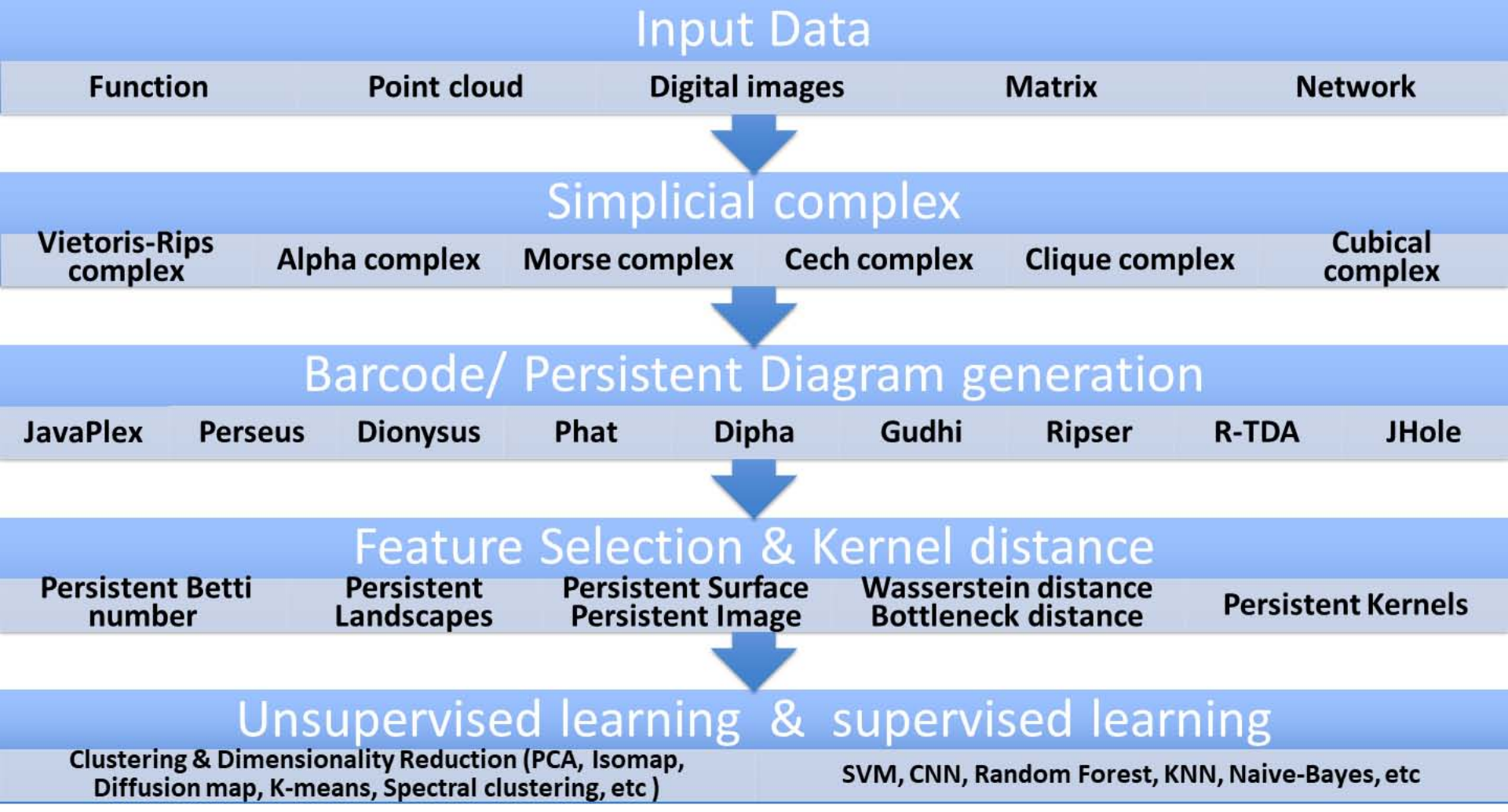


Image credit - Pun, Xia, Lee. TDA Survey (2018)

Applications

- **Point Cloud Setting:**

  Shape Recognition

Applications

- **Point Cloud Setting:**

  Shape Recognition

- **Graph Setting:**

  Computer-Aided Drug Design

Applications

- **Point Cloud Setting:**

  Shape Recognition

- **Graph Setting:**

  Computer-Aided Drug Design

- **Image Setting:**

  Cancer Detection from Histopathological Images

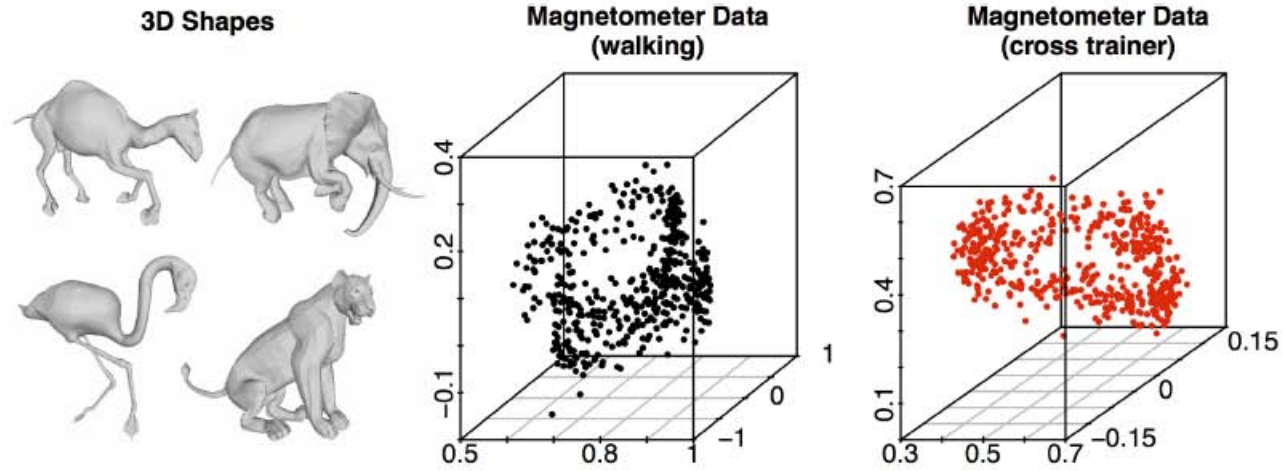


Figure 4. Left: Four 3D shapes. Middle and Left: 500 random points from the magnetometer data of the second experiment.

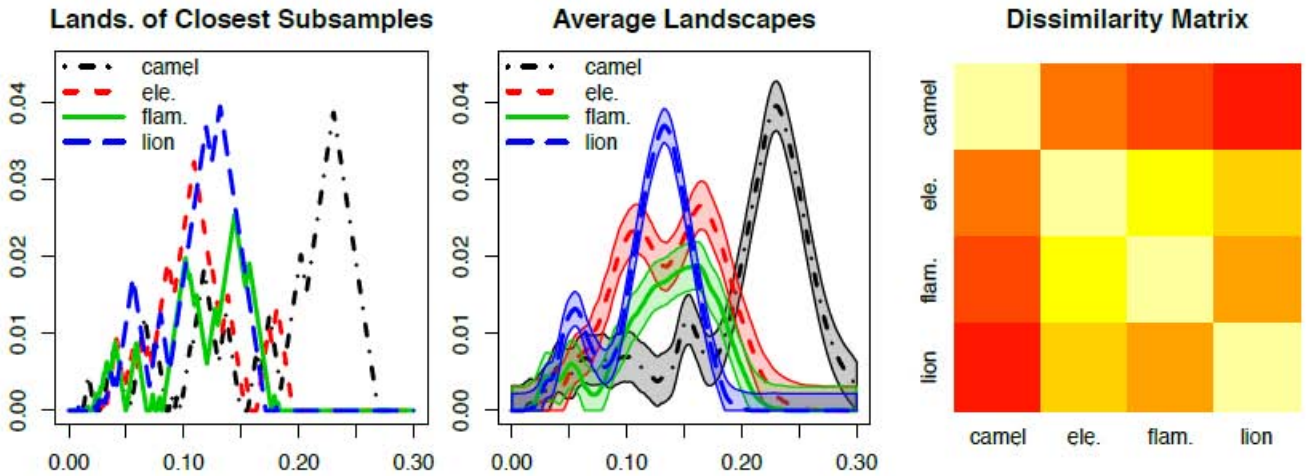


Figure 5. Subsampling methods applied to 3D shapes. For $n = 100$ subsamples of size $m = 300$, for each shape, we constructed the landscapes of the closest subsample (left), the average landscape with 95% confidence band (middle) and the dissimilarity matrix of the pairwise ℓ_∞ distance between average landscapes.

Histopathological Cancer Detection with Topological Machine Learning

Ankur Yadav[1], Faisal Ahmed[2], Ovidiu Daescu[1], Reyhan Gedik[3], and Baris Coskunuzer[2]

[1] UT Dallas, CS Dept.
[2] UT Dallas, Math Dept.
[3] Harvard - MGH, Pathology Dept.

IEEE - BIBM, Istanbul, December 2023

Motivation

- Examining tissue samples is the primary way to detect and grade cancer. However, this process requires **experienced pathologists** and tends to be **time-consuming.**

Motivation

- Examining tissue samples is the primary way to detect and grade cancer. However, this process requires **experienced pathologists** and tends to be **time-consuming.**

- ML methods offer **clinical decision support systems** that enhance accuracy, reproducibility, and speed in medical processes.

Motivation

- Examining tissue samples is the primary way to detect and grade cancer. However, this process requires **experienced pathologists** and tends to be **time-consuming.**

- ML methods offer **clinical decision support systems** that enhance accuracy, reproducibility, and speed in medical processes.

- DL methods exhibit significant potential. However, their utilization in clinical-stage implementation faces challenges due to extensive **preprocessing** periods, the substantial size of **training datasets**, the need for **high-performance computing** infrastructures, and the challenge of **interpretability** in decision-making.

Motivation

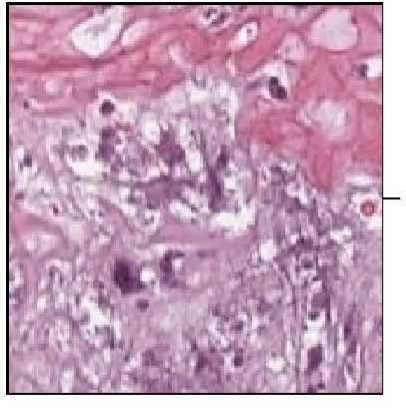
- Examining tissue samples is the primary way to detect and grade cancer. However, this process requires **experienced pathologists** and tends to be **time-consuming.**
- ML methods offer **clinical decision support systems** that enhance accuracy, reproducibility, and speed in medical processes.
- DL methods exhibit significant potential. However, their utilization in clinical-stage implementation faces challenges due to extensive **preprocessing** periods, the substantial size of **training datasets**, the need for **high-performance computing** infrastructures, and the challenge of **interpretability** in decision-making.
- In this project, to address these needs, we develop a **fast, and high-performing** topological ML method for this task.

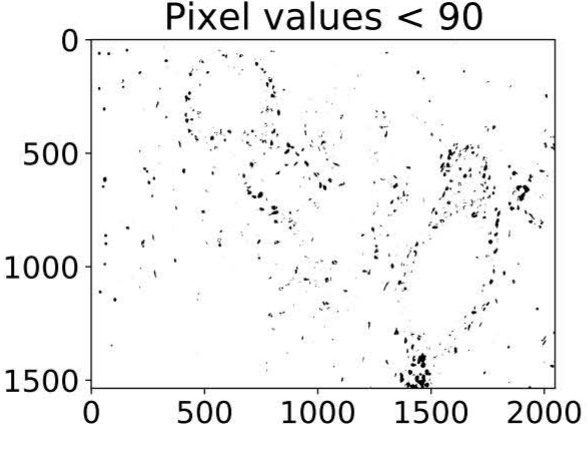
Motivation

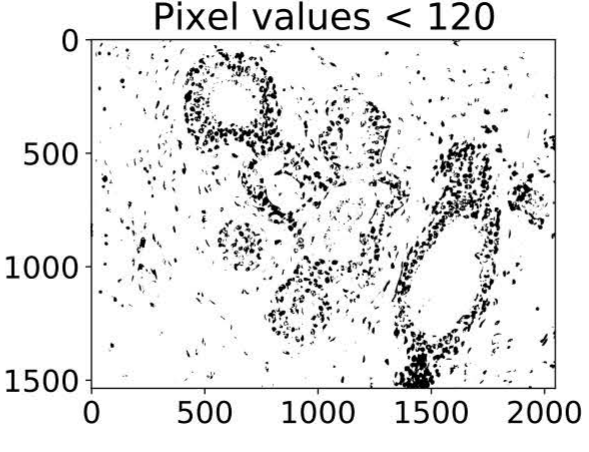
- Examining tissue samples is the primary way to detect and grade cancer. However, this process requires **experienced pathologists** and tends to be **time-consuming.**
- ML methods offer **clinical decision support systems** that enhance accuracy, reproducibility, and speed in medical processes.
- DL methods exhibit significant potential. However, their utilization in clinical-stage implementation faces challenges due to extensive **preprocessing** periods, the substantial size of **training datasets**, the need for **high-performance computing** infrastructures, and the challenge of **interpretability** in decision-making.
- In this project, to address these needs, we develop a **fast, and high-performing** topological ML method for this task.
- Further, our **topological features** hold the potential to significantly boost the performance of upcoming ML models within this domain.

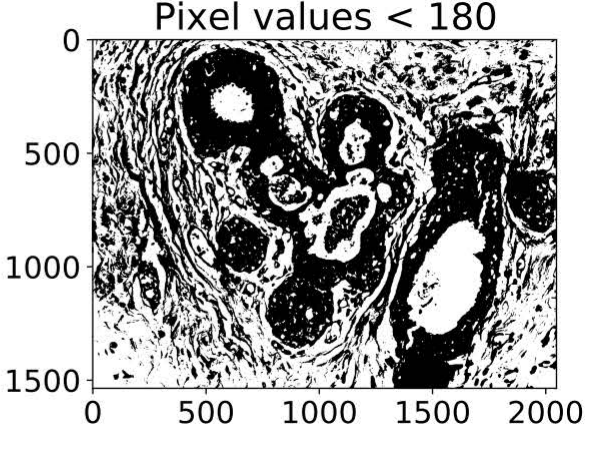
Cubical Persistence

- For Image Data, we use Cubical Persistence.
- Constructing filtration out of histopathological image.




Pixel values < 90
0
500
1000
1500
0
500
1000
1500
2000


Pixel values < 120
0
500
1000
1500
0
500
1000
1500
2000

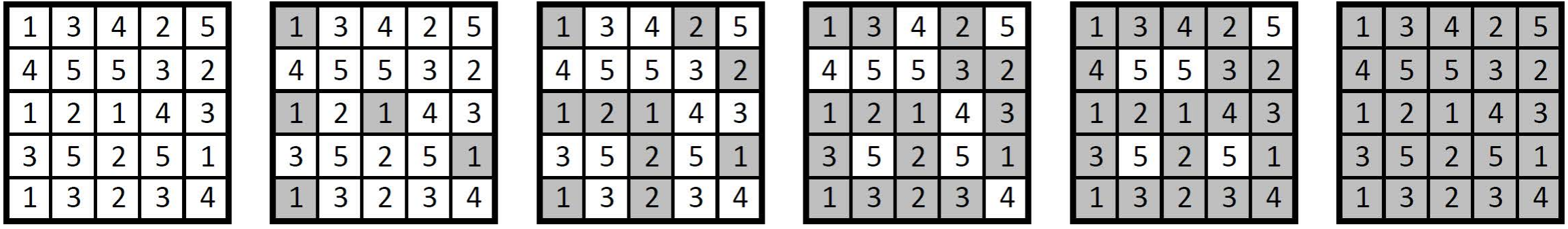

Pixel values < 180
0
500
1000
1500
0
500
1000
1500
2000

Cubical Persistence

- For Image Data, we use Cubical Persistence.
- Constructing filtration out of histopathological image.
- Persistence Diagram: Keep track of components and loops in binary images

Cubical Persistence

- For Image Data, we use Cubical Persistence.
- Constructing filtration out of histopathological image.
- Persistence Diagram: Keep track of components and loops in binary images
- Vectorization: Betti function.

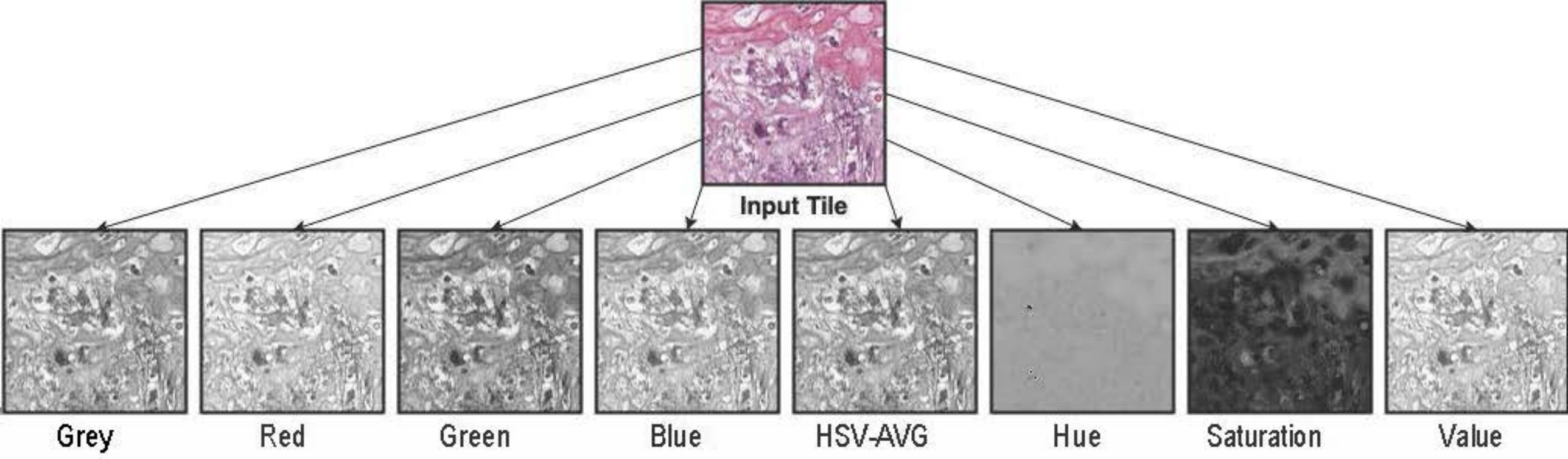


B_0 = [5 5 2 1 1]

B_1 = [0 0 2 3 0]

Topological Features of Histopathological Images

- **Topological Features:** For each histopathological image, for each of the 8 color channels, we obtain 100-dimensional Betti-0 and 100-dimensional Betti-1 vectors. This gives us a 1600-dimensional topological feature vector for each histopathological image.


Input Tile
Grey
Red
Green
Blue
HSV-AVG
Hue
Saturation
Value

Topological Features of Histopathological Images

- **Topological Features:** For each histopathological image, for each of the 8 color channels, we obtain 100-dimensional Betti-0 and 100-dimensional Betti-1 vectors. This gives us a 1600-dimensional topological feature vector for each histopathological image.

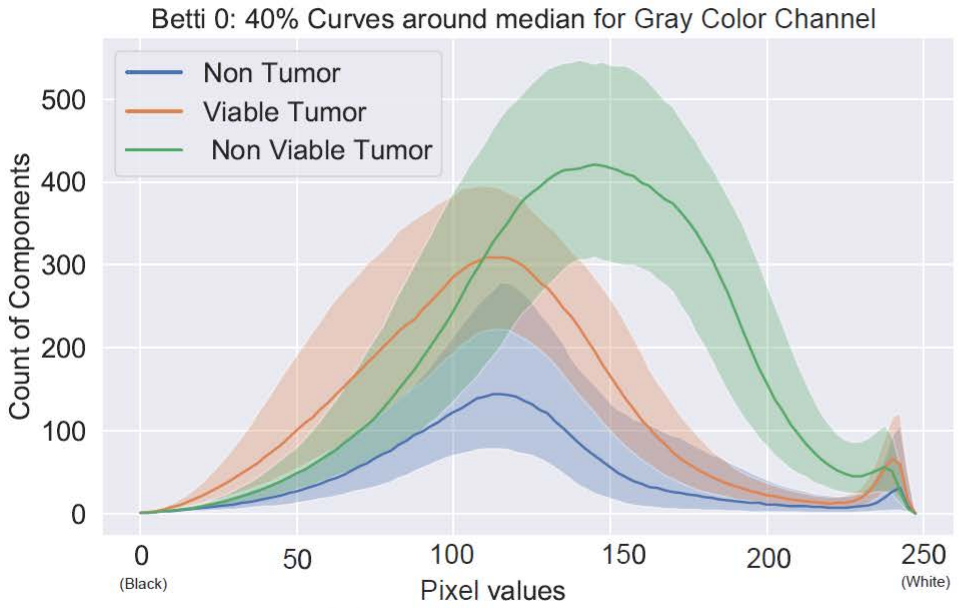
- **Topological Feature Vectors** distinguish normal and abnormal classes histopathological images for following cancer types.

Topological Features of Histopathological Images

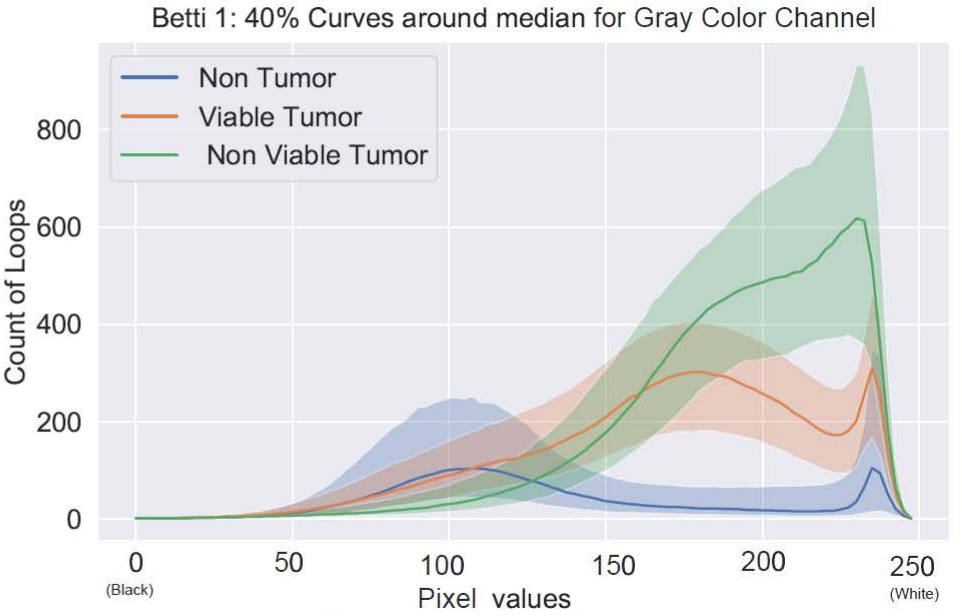
- **Topological Features:** For each histopathological image, for each of the 8 color channels, we obtain 100-dimensional Betti-0 and 100-dimensional Betti-1 vectors. This gives us a 1600-dimensional topological feature vector for each histopathological image.

- **Topological Feature Vectors** distinguish normal and abnormal classes histopathological images for following cancer types.
  - Bone Cancer

Topological Features for Bone Cancer



Betti-0 vectors for Gray color



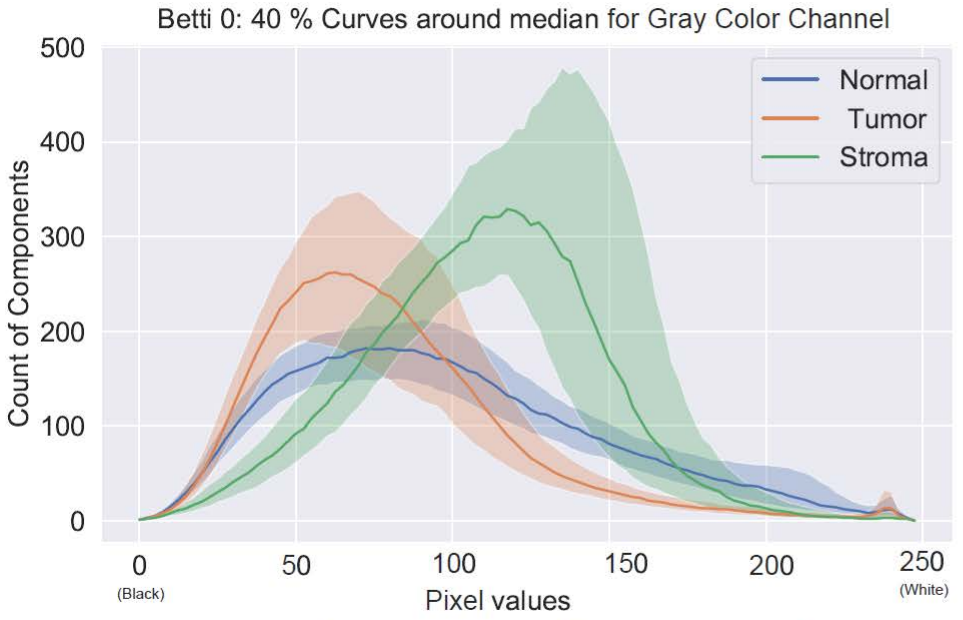
Betti-1 vectors for Gray color

Topological Features of Histopathological Images

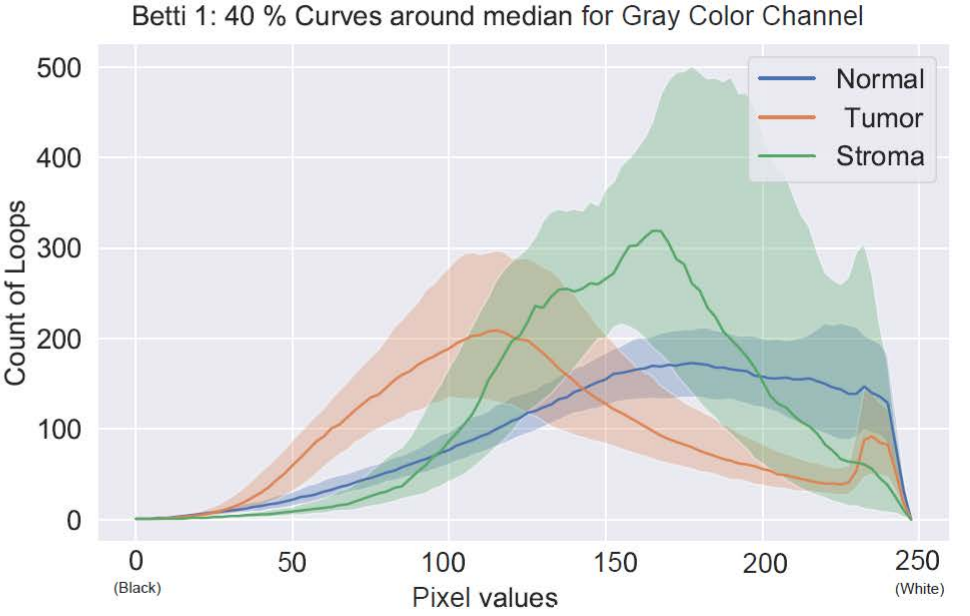
- **Topological Features:** For each histopathological image, for each of the 8 color channels, we obtain 100-dimensional Betti-0 and 100-dimensional Betti-1 vectors. This gives us a 1600-dimensional topological feature vector for each histopathological image.

- **Topological Feature Vectors** distinguish normal and abnormal classes histopathological images for following cancer types.
  - Bone Cancer
  - Colon Cancer

Topological Features for Colon Cancer



Betti-0 vectors for Gray color



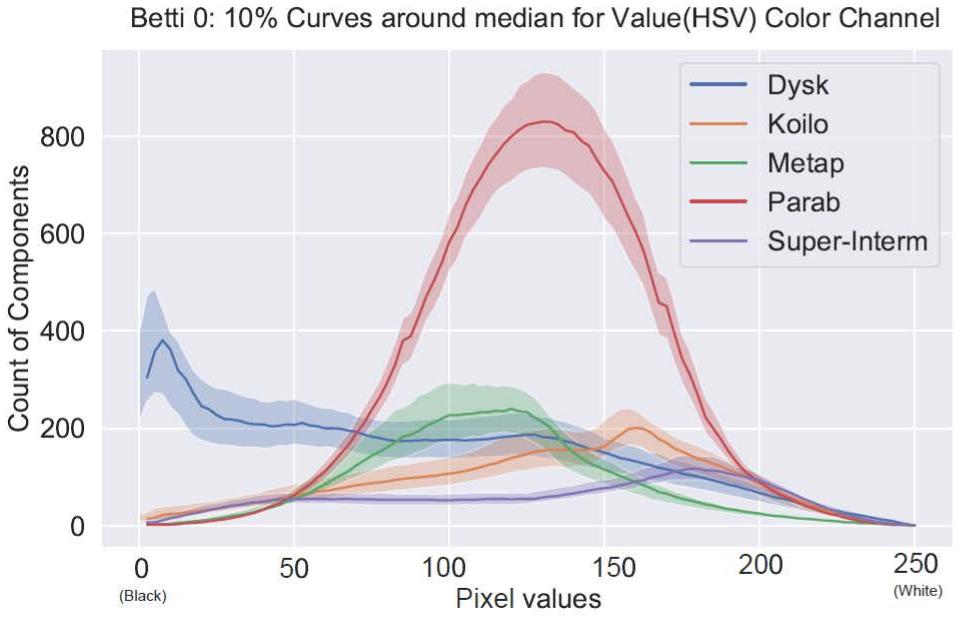
Betti-1 vectors for Gray color

Topological Features of Histopathological Images

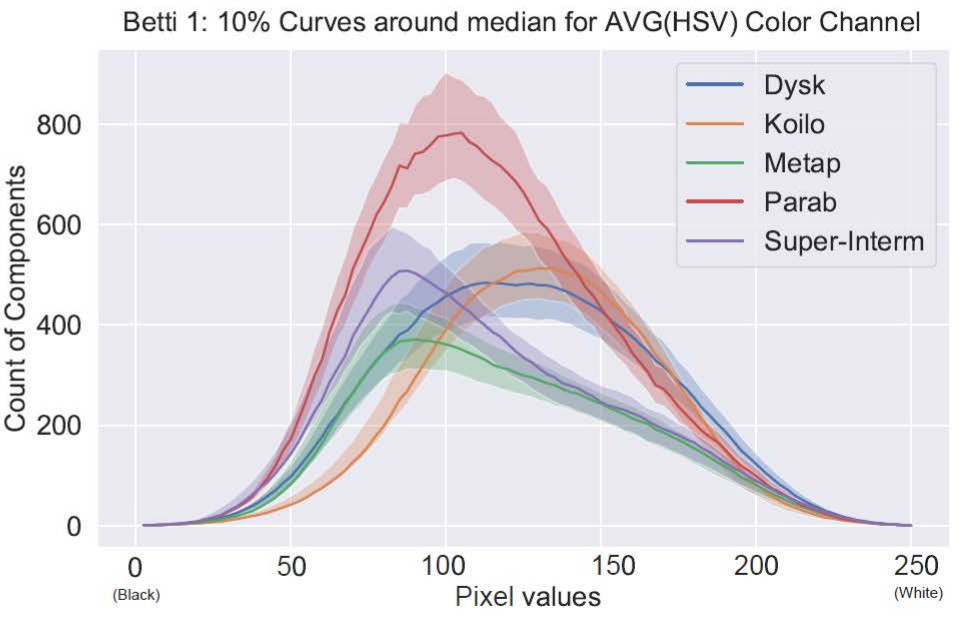
- **Topological Features:** For each histopathological image, for each of the 8 color channels, we obtain 100-dimensional Betti-0 and 100-dimensional Betti-1 vectors. This gives us a 1600-dimensional topological feature vector for each histopathological image.

- **Topological Feature Vectors** distinguish normal and abnormal classes histopathological images for following cancer types.
  - Bone Cancer
  - Colon Cancer
  - Cervical Cancer

Topological Features for Cervical Cancer



Betti-0 vectors for Value Color



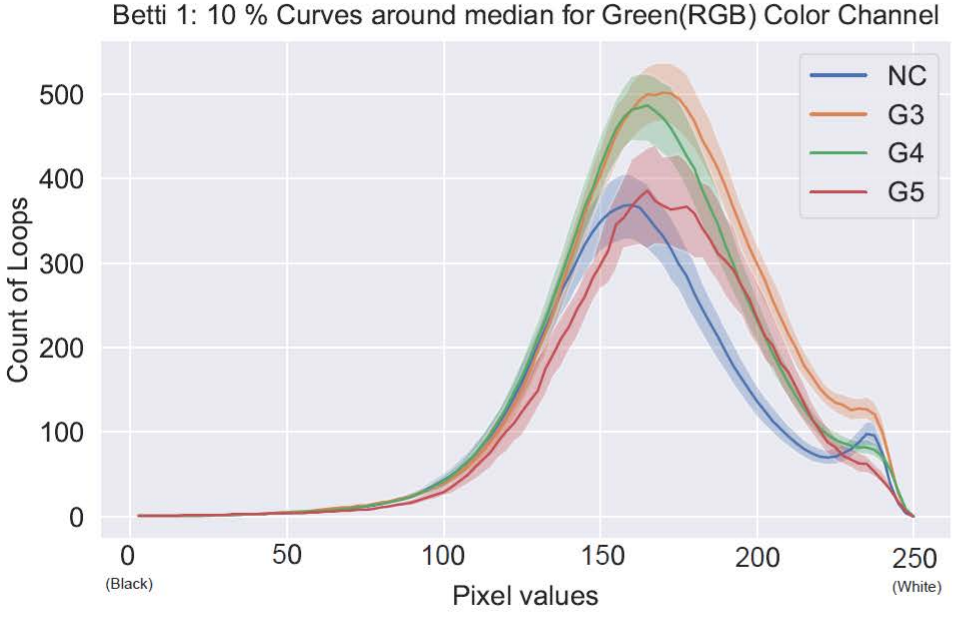
Betti-1 vectors for avg(HSV) color

Topological Features of Histopathological Images

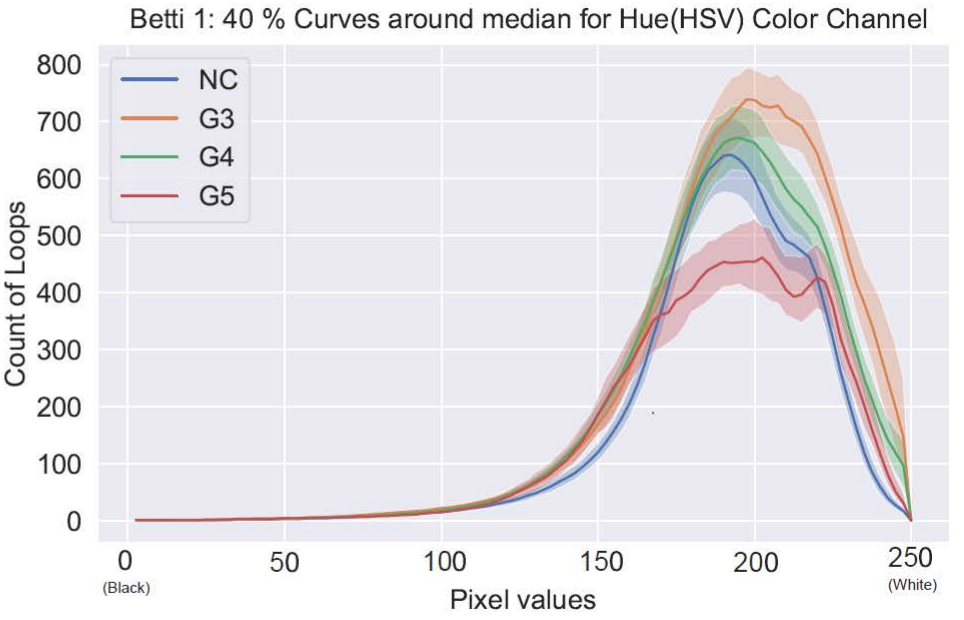
- **Topological Features:** For each histopathological image, for each of the 8 color channels, we obtain 100-dimensional Betti-0 and 100-dimensional Betti-1 vectors. This gives us a 1600-dimensional topological feature vector for each histopathological image.

- **Topological Feature Vectors** distinguish normal and abnormal classes histopathological images for following cancer types.
  - Bone Cancer
  - Colon Cancer
  - Cervical Cancer
  - Prostate Cancer

Topological Features for Prostate Cancer



Betti-1 vectors for Green color



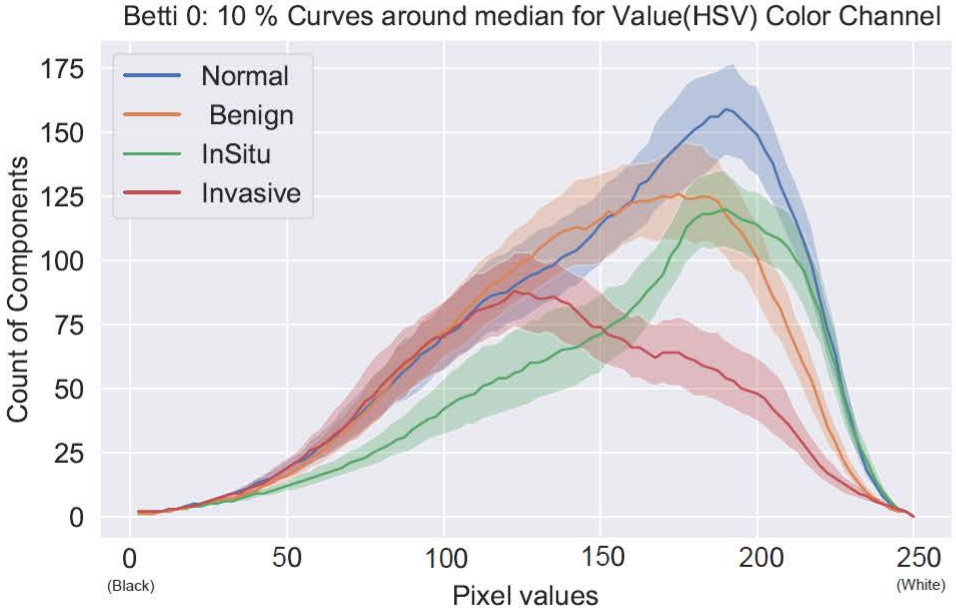
Betti-1 vectors for Hue color

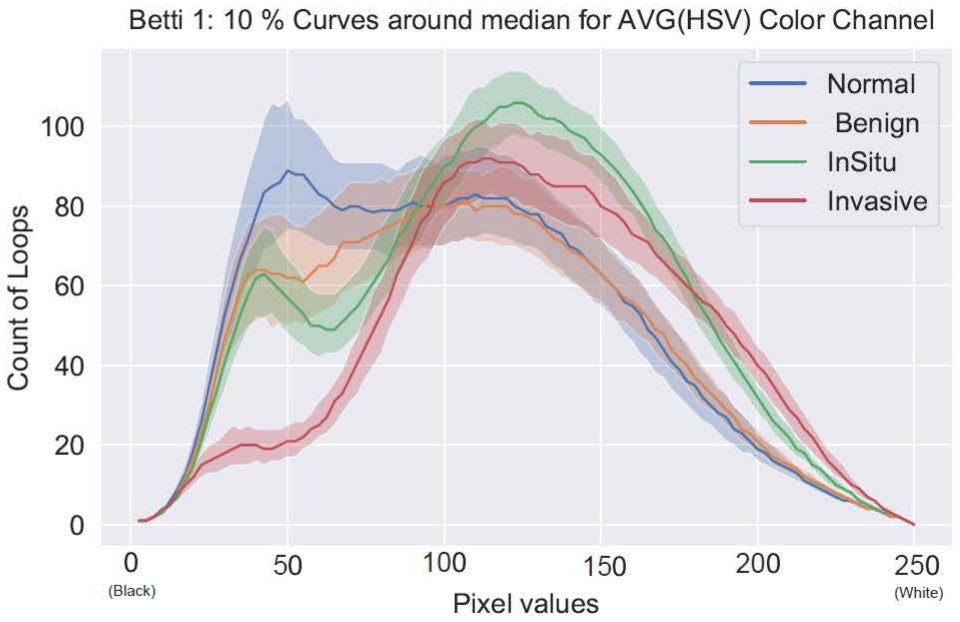
Topological Features of Histopathological Images

- **Topological Features:** For each histopathological image, for each of the 8 color channels, we obtain 100-dimensional Betti-0 and 100-dimensional Betti-1 vectors. This gives us a 1600-dimensional topological feature vector for each histopathological image.

- **Topological Feature Vectors** distinguish normal and abnormal classes histopathological images for following cancer types.
  - Bone Cancer
  - Colon Cancer
  - Cervical Cancer
  - Prostate Cancer
  - Breast Cancer

Topological Features for Breast Cancer





Betti-0 vectors for Value Color

Betti-1 vectors for avg(HSV) color

Topological Machine Learning Model

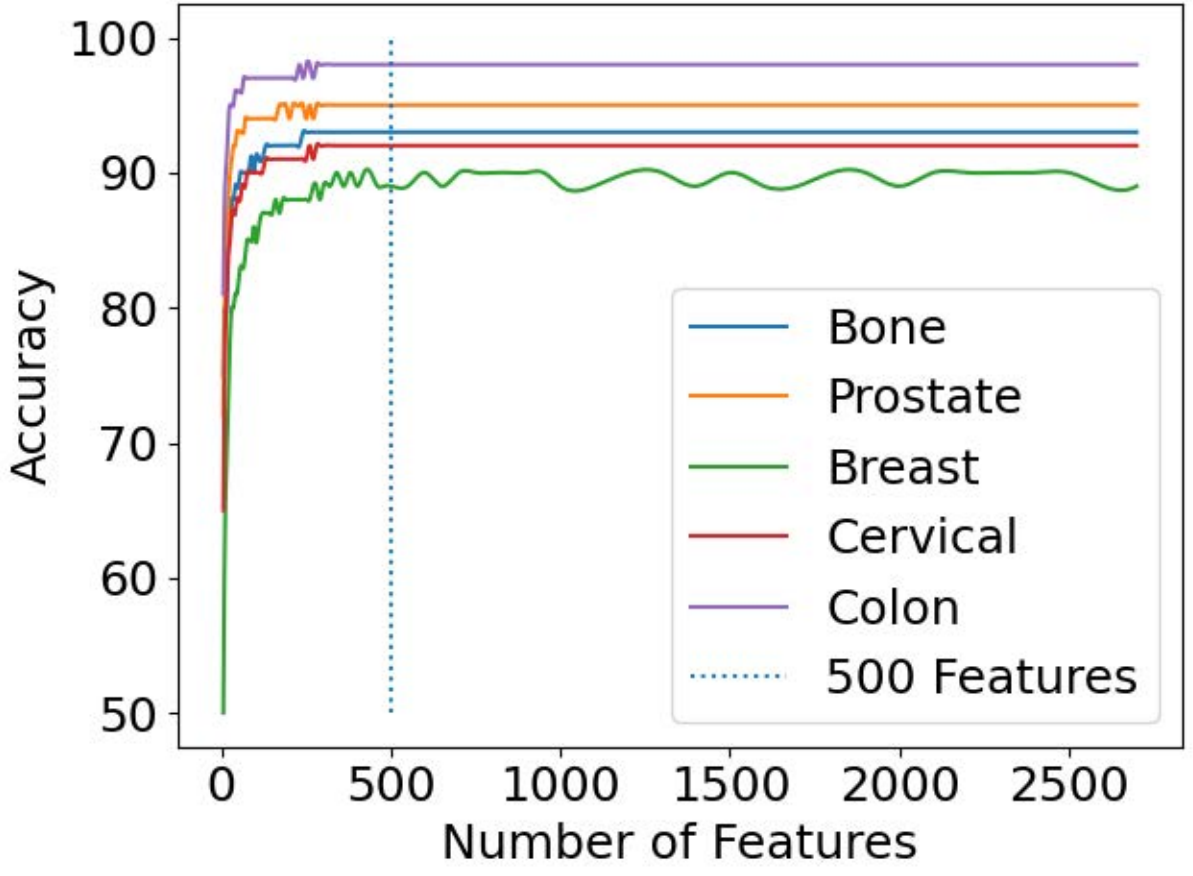
- For each histopathological image, we extract:

Topological Machine Learning Model

- For each histopathological image, we extract:
  - 1600-dimensional topological features
  - 800-dimensional Local Binary Pattern features
  - 400-dimensional Gabor Filter features

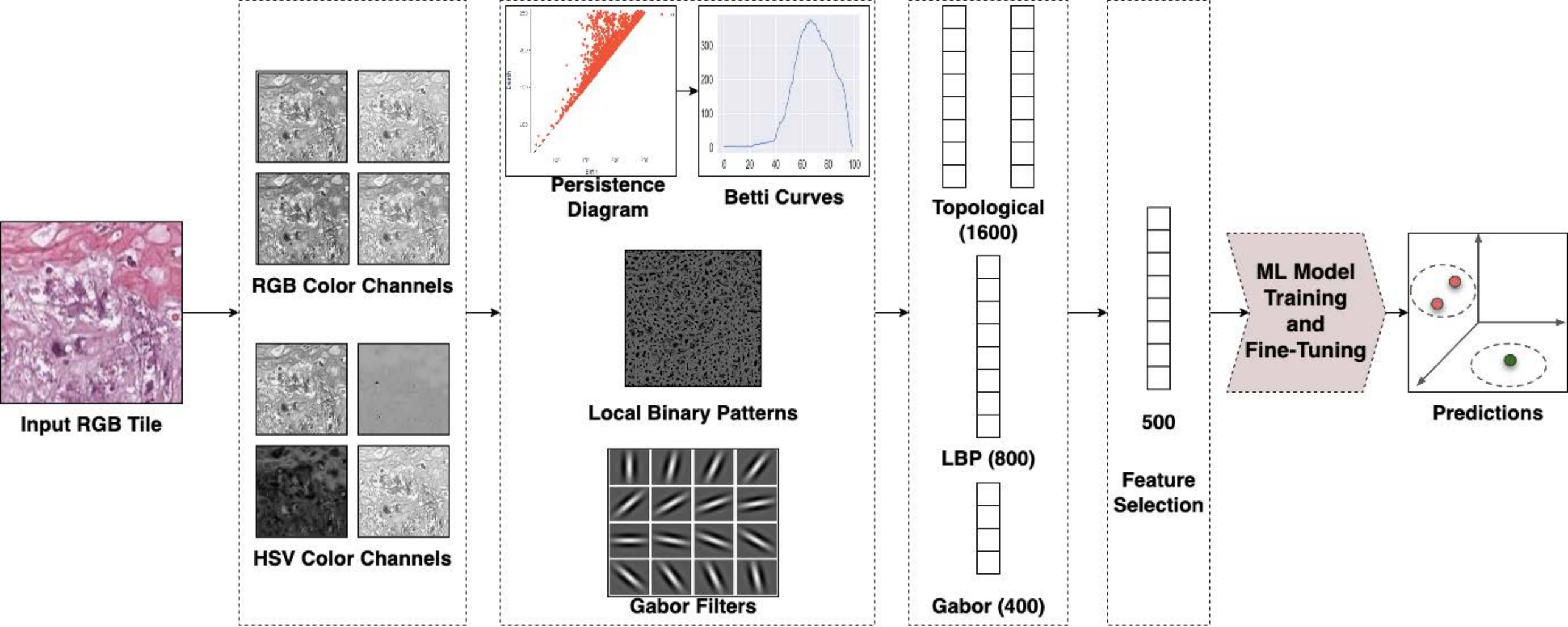
Topological Machine Learning Model

- For each histopathological image, we extract:
  - 1600-dimensional topological features
  - 800-dimensional Local Binary Pattern features
  - 400-dimensional Gabor Filter features
- We then use feature selection algorithm to choose the most important 500 features out of 2800 features.


100
90
80
70
60
50
Accuracy
0
500
1000
1500
2000
2500
Number of Features
Bone
Prostate
Breast
Cervical
Colon
500 Features

Topological Machine Learning Model

- For each histopathological image, we extract:
  - 1600-dimensional topological features
  - 800-dimensional Local Binary Pattern features
  - 400-dimensional Gabor Filter features
- We then use feature selection algorithm to choose the most important 500 features out of 2800 features.
- Finally, we feed these features into our ML classifier.


Input RGB Tile
RGB Color Channels
HSV Color Channels
Persistence Diagram
Betti Curves
Local Binary Patterns
Gabor Filters
Topological (1600)
LBP (800)
Gabor (400)
500
Feature Selection
ML Model Training and Fine-Tuning
Predictions

Experiments

- Datasets

TABLE VII: **Dataset Details.** Information about classes, base images and generated tiles for the experiments.

Dataset	# Classes	Images	Tile Size	Train Tiles	Test Tiles	Total Tiles
ICIAR2018 (Breast)	4	400	256x256	15078	3769	18848
SipakMed (Cervical)	5	966	256x256	51324	12831	64155
Sicapv2 (Prostate)	4	9959	256x256	27466	6866	34333
CRC100K (Colon)	3	33526	256x256	26820	6705	33526
UT-OSteo. (Bone)	3	1144	256x256	15587	6721	22308

Experiments

- Datasets
- Results for 5 Cancer types

TABLE I: Performance of our model in 5 cancer types

Cancer	***Dataset***	***# Image***	***# Class***	***Acc***	***AUC***
Breast	ICIAR2018	400	4	91.6	0.98
Prostate	Sicapv2	9959	4	95.2	0.99
Colon	CRC100K	33526	3	98.7	0.99
Bone	UT-Osteo.	1144	3	94.2	0.99
Cervical	SipakMed	966	5	94.2	0.99

Comparison Table for Breast Cancer

Method	Train:Test	# Class	Accuracy
Kwok [33]	75:25	4	79.00
Nawaz [39]	80:20	4	81.25
Rakhlin [26]	80:20	4	87.20
Vang [62]	75:25	4	87.50
DCNN [31]	75:25	4	**92.50**
Our Model	80:20	4	91.64

Comparison Table for Bone Cancer

Method	Train:Test	# Class	Accuracy
Mishra-CNN [8]	70:30	3	93.30
Mishra-SVM [8]	70:30	3	89.90
VGG19 [6]	70:30	3	93.91
Our Model	70:30	3	**94.20**

Comparison Table for Cervical Cancer

Method	Train:Test	# Class	Accuracy
Hayranto [29]	5 fold CV	5	87.32
ResNet [61]	5 fold CV	5	94.86
Fuzzy [36]	5 fold CV	5	**95.43**
Our Model	5 fold CV	5	94.21

Comparison Table for Prostate Cancer

Method	Train:Test	# Class	Accuracy
Arvaniti [9]	80:20	4	58.61
Gerytch [21]	80:20	4	51.36
FSConv+GMP [53]	80:20	4	83.50
Res-CAE [58]	80:20	4	85.00
Our Model	80:20	4	**95.20**

Cancer	*Classes*	*Sensitivity*	*Specificity*	*Precision*	*Recall*	F1-Score	AUC
Breast	Normal	0.8994	0.9652	0.8994	0.8935	0.8965	
	Benign	0.8911	0.9702	0.8911	0.9122	0.9015	
	InSitu	0.9195	0.9683	0.9195	0.9034	0.9114	
	Invasive	0.9560	0.9850	0.9560	0.9557	0.9558	
	Average	0.9166	0.9722	0.9166	0.9164	0.9164	0.9894
Cervical	im_Dyskeratotic	0.9272	0.9795	0.9272	0.9358	0.9315	
	im_Koil.	0.9280	0.9737	0.9280	0.9165	0.9222	
	im_Meta.	0.9517	0.9786	0.9517	0.9428	0.9472	
	im_Para.	0.9810	0.9983	0.9810	0.9880	0.9845	
	im_Sup.-Int.	0.9402	0.9941	0.9402	0.9582	0.9491	
	Average	0.9421	0.9819	0.9421	0.9421	0.9421	0.9960
Prostate	NC	0.9740	0.9975	0.9740	0.9964	0.9850	
	G3	0.9281	0.9791	0.9281	0.9012	0.9145	
	G4	0.9465	0.9731	0.9465	0.9512	0.9488	
	G5	0.9640	0.9913	0.9640	0.8845	0.9226	
	Average	0.9554	0.9852	0.9554	0.9553	0.9551	0.9916
Colon	Normal	0.9804	0.9936	0.9804	0.9820	0.9812	
	Stroma	0.9928	0.9958	0.9928	0.9906	0.9917	
	Tumor	0.9868	0.9906	0.9868	0.9874	0.9871	
	Average	0.9870	0.9930	0.9870	0.9870	0.9870	0.9995
Bone	VT	0.9498	0.9701	0.9498	0.9432	0.9465	
	NVT	0.9340	0.9771	0.9340	0.9366	0.9353	
	NT	0.9405	0.9646	0.9405	0.9445	0.9425	
	Average	0.9420	0.9698	0.9420	0.9420	0.9420	0.9553

Experiments

- Datasets
- Results for 5 Cancer types
- Ablation Study

TABLE VI: **Ablation Study.** Accuracy results of XGBoost model on different features.

Model	Bone	Breast	Prostate	Cervical	Colon
TDA (Grayscale)	85.2	69.1	92.1	54.9	95.7
TDA (All colors)	91.2	83.0	94.6	85.8	97.9
LBP	92.1	84.7	94.4	89.6	97.3
Gabor	88.1	70.7	92.7	84.9	94.9
Gabor+LBP	**93.7**	87.5	95.0	92.1	98.0
TDA+LBP	93.4	88.1	95.2	91.1	98.4
TDA+LBP+Gabor	93.1	**89.8**	**95.4**	**92.9**	**98.5**

	Bone		Breast		Prostate		Cervical		Colon	
Model	**RF**	**XG**	**RF**	**XG**	**RF**	**XG**	**RF**	**XG**	**RF**	**XG**
Gray (200)	84.0	84.6	69.6	68.8	92.3	91.1	53.7	54.6	94.5	95.9
G-RGB (800)	86.8	89.4	77.6	80.0	92.8	93.2	73.9	77.5	96.4	97.5
HSV AVG (200)	82.0	82.3	60.9	60.8	88.3	84.5	59.9	60.7	87.2	88.0
HSV AVG+HSV(800)	88.0	90.1	76.3	78.7	92.4	93.0	82.2	82.7	95.9	97.5
All Betti Features (1600)	87.9	90.8	79.4	82.8	92.8	93.9	82.2	85.5	96.5	97.8
All Features (2800)	90.7	93.7	85.4	88.4	93.8	94.7	90.5	92.6	97.1	98.5
Feature Selection	-	94.2	-	91.6	-	95.2	-	91.4	-	98.4

Final Remarks

- Our experiments show topological feature vectors are quite effective in cancer detection from histopathological images.

Final Remarks

- Our experiments show topological feature vectors are quite effective in cancer detection from histopathological images.
- Without using any deep learning, topological feature vectors give competitive results with SOTA models in 5 different cancer types.

Final Remarks

- Our experiments show topological feature vectors are quite effective in cancer detection from histopathological images.
- Without using any deep learning, topological feature vectors give competitive results with SOTA models in 5 different cancer types.
- We are positive that these novel topological features will substantially enhance the performance of any upcoming ML models in the domain.
- Moving forward, we aim to combine our topological feature vectors with the latest CNN models to obtain robust and effective computer-aided clinical decision support systems in histopathology.

Final Remarks

- Our experiments show topological feature vectors are quite effective in cancer detection from histopathological images.
- Without using any deep learning, topological feature vectors give competitive results with SOTA models in 5 different cancer types.
- We are positive that these novel topological features will substantially enhance the performance of any upcoming ML models in the domain.
- Moving forward, we aim to combine our topological feature vectors with the latest CNN models to obtain robust and effective computer-aided clinical decision support systems in histopathology.
- Thank you for your attention!

DRUG DISCOVERY with TOPOLOGICAL DATA ANALYSIS

Baris Coskunuzer

University of Texas at Dallas

with Andac Demir, Ignacio S. Dominguez, Yuzhou Chen, Yulia Gel, Bulent Kiziltan

Supported by NSF and Simons Foundation Collaboration with Novartis Inc.

NeurIPS 2022 https://arxiv.org/abs/2211.03808

Drug Discovery Process

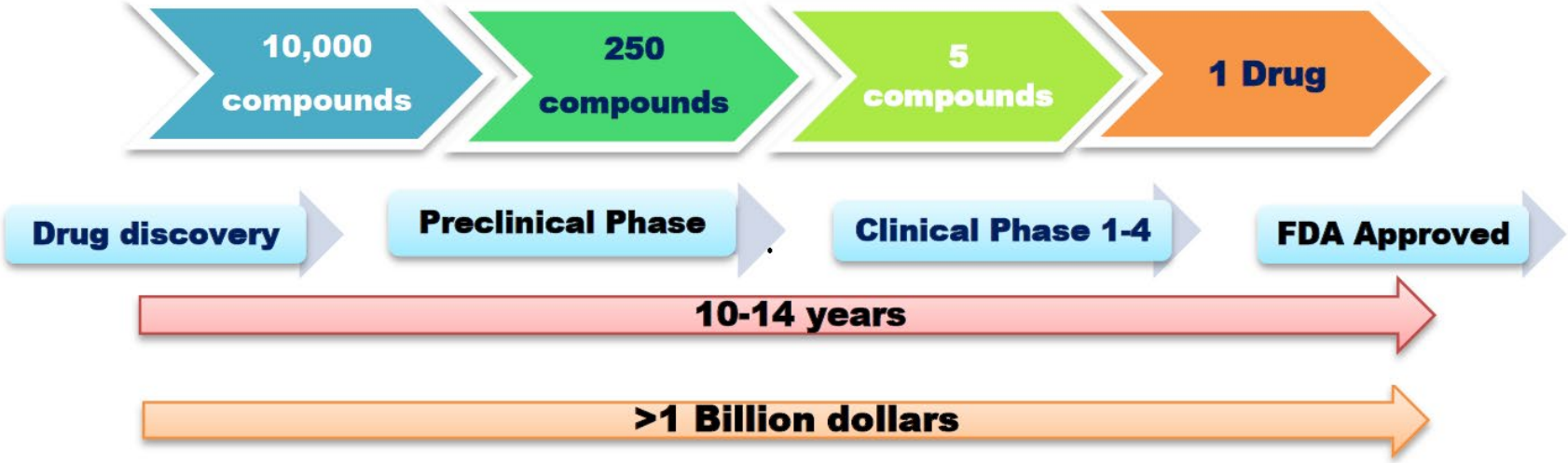


Figure 1: Traditional process of drug discovery and development.

Image Credit: Surabhi et al, Computer Aided Drug Design: An Overview, JDDT 2018

VIRTUAL SCREENING METHODS FOR HIT IDENTIFICATION

Structure-based methods

1. Molecular docking approach models the interaction between a small molecule and a drug target at the atomic level.

2. It requires:
   - **Knowledge of the binding site** before docking process.
   - Prediction of the **ligand conformation** as well as its position and orientation in binding site.

LOCK & KEY APPROACH

Ligand-based methods

1. We know a set of **active ligands** that can inhibit a drug target.

2. There is little or no structural information available for those drug targets.

3. Drug candidates are compared against a library of dozens/hundreds of active ligands and thousands of decoys (inactive ligands).

SIMILARITY BASED APPROACH

(Our method)

Structure (Docking) Based Virtual Screening

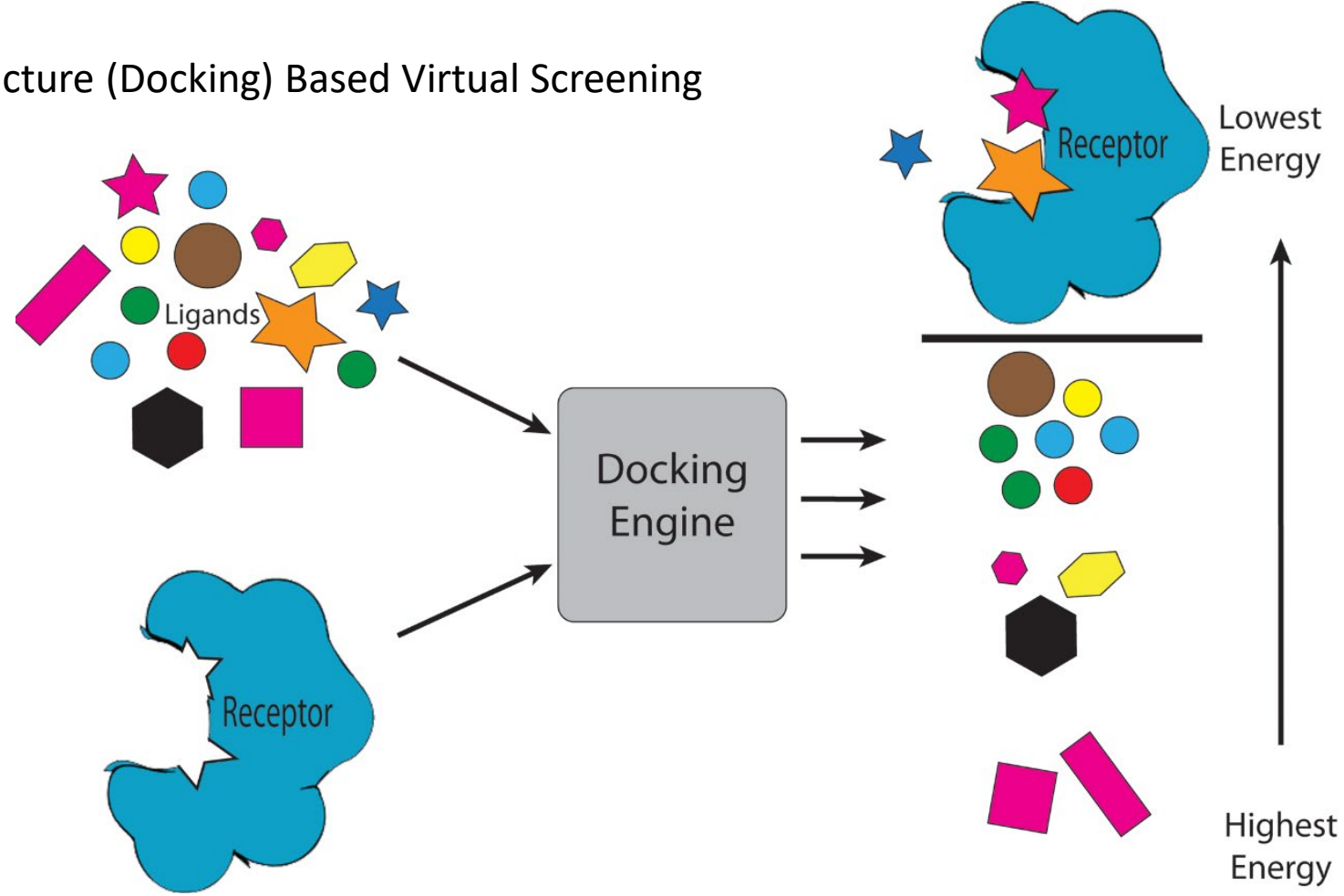


Image Credit: Jacob et al, Accessible HTVS Mol. Docking Software, PLOS 2012

RELATED WORK

- Most Ligand Based Models use an approach called **Fingerprinting**
- **Main idea is to convert/summarize the structure information into a suitable form to be used with ML algorithms**

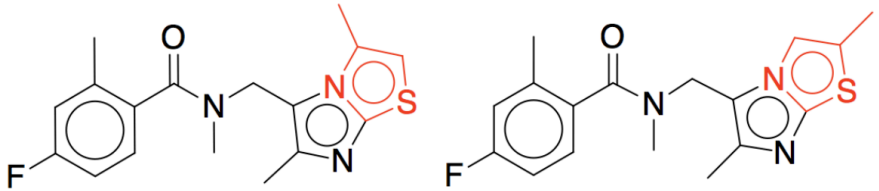
Two Examples:

1. Morgan Fingerprints: ECFP (Extended Connectivity Fingerprints)

- By using an hashing function, encode 0, 1, 2 neighborhoods of each atom as hash values. Similar to Weisfeiler-Lehman algorithm.
- By using 1024 dimensional bit-vector, summarize obtained hash values. Nonreversible.

RELATED WORK

2. SMILES METHOD



Cc1cn2c(CN(C)C(=O)c3ccc(F)cc3C)c(C)nc2s1
Cc1cc(F)ccc1C(=O)N(C)Cc1c(C)nc2scc(C)n12

Figure 1. Two almost identical molecules with markedly different canonical SMILES in RDKit. The edit distance between two strings is 22 (50.5% of the whole sequence).

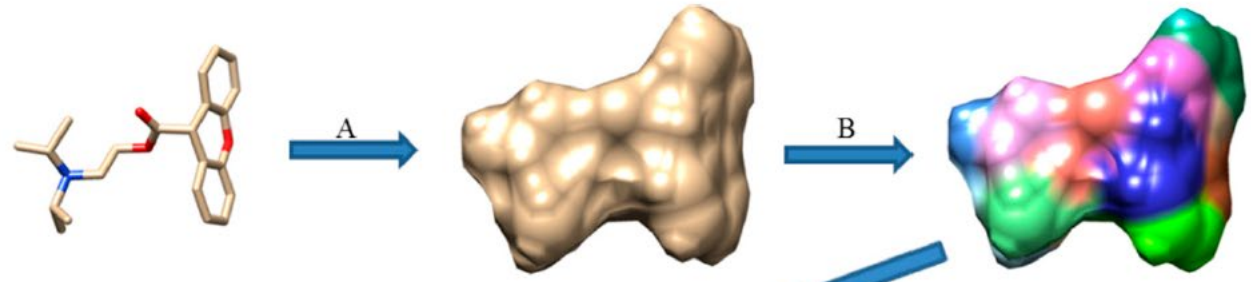
- Another popular technique **SMILES** formulated the compound fingerprinting task as string generation problem.
- SMILES strings are reversible i.e., they can be translated into graphs.

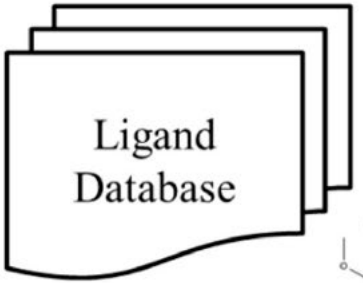
- However. SMILES has 2 limitations:

1. Two molecules with similar chemical structures may be encoded into markedly different SMILES strings.

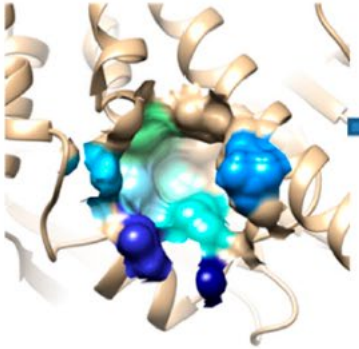
2. Essential chemical properties such as molecule validity are easier to express on graphs rather than linear SMILES representations.

RELATED WORK

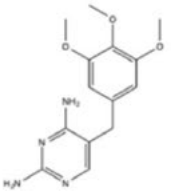


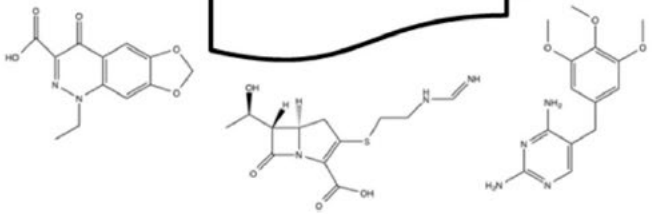


C



Molecule	Score
1	1.7
2	0.8
3	1.3





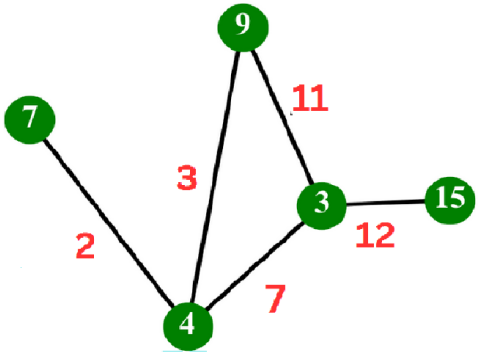
BACKGROUND - PERSISTENT HOMOLOGY

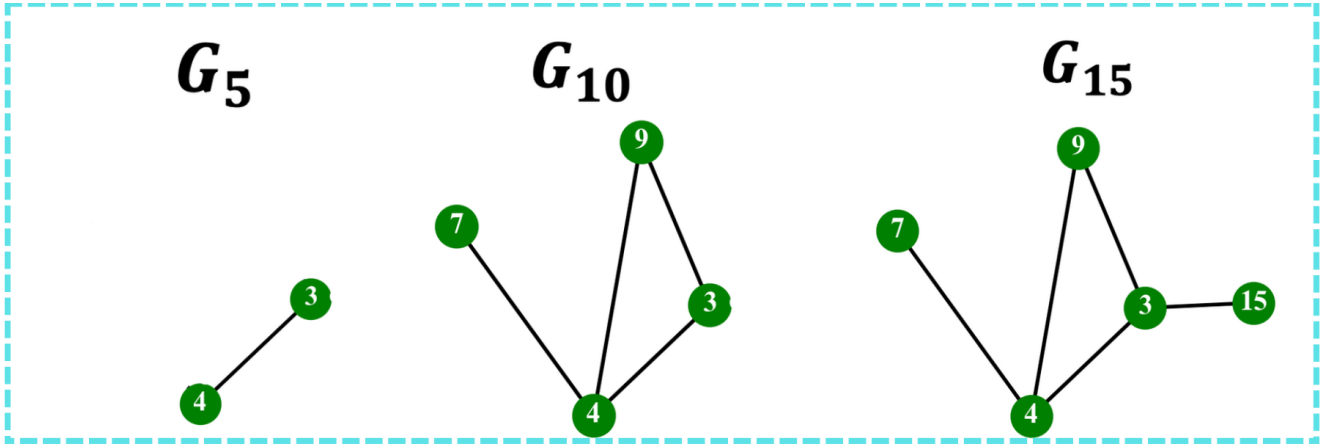
- A novel feature extraction method by using topology.
- Captures and summarizes the shape patterns developed in the data.
- 3-step Process:

1. **Filtration:** For a given dataset, induce a meaningful sequence of topological spaces.
2. **Persistence Diagram:** Record the evolution of topological features in this sequence.
3. **Vectorization:** Convert PDs into suitable functions/vectors for ML methods.

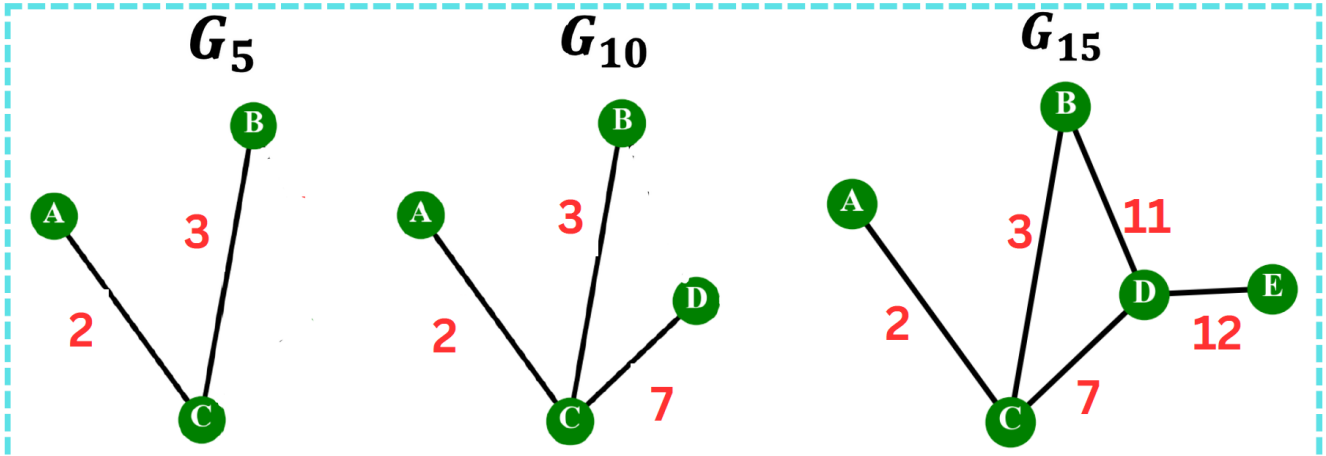
FILTRATION FOR GRAPHS

node filtration





edge filtration



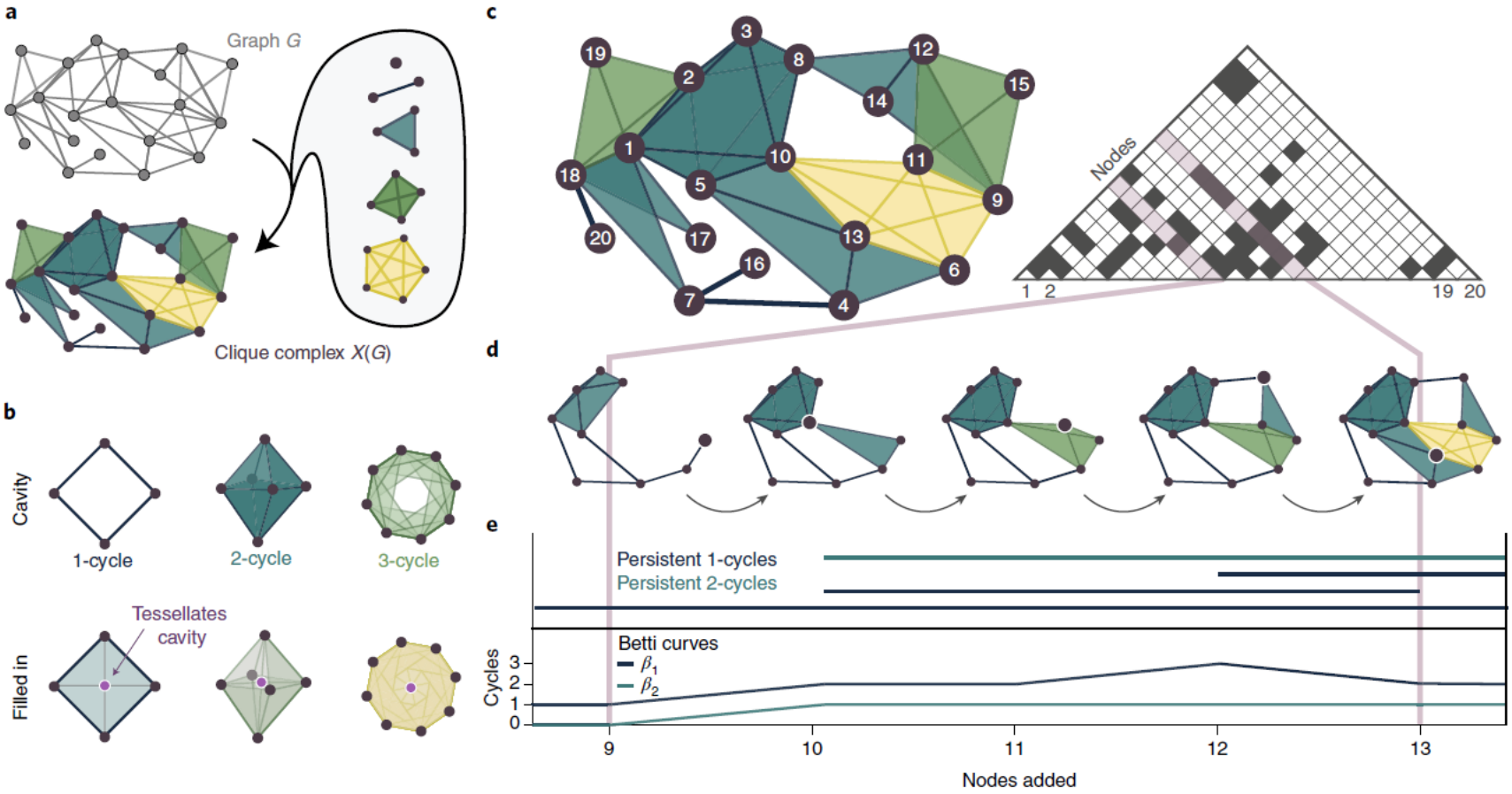
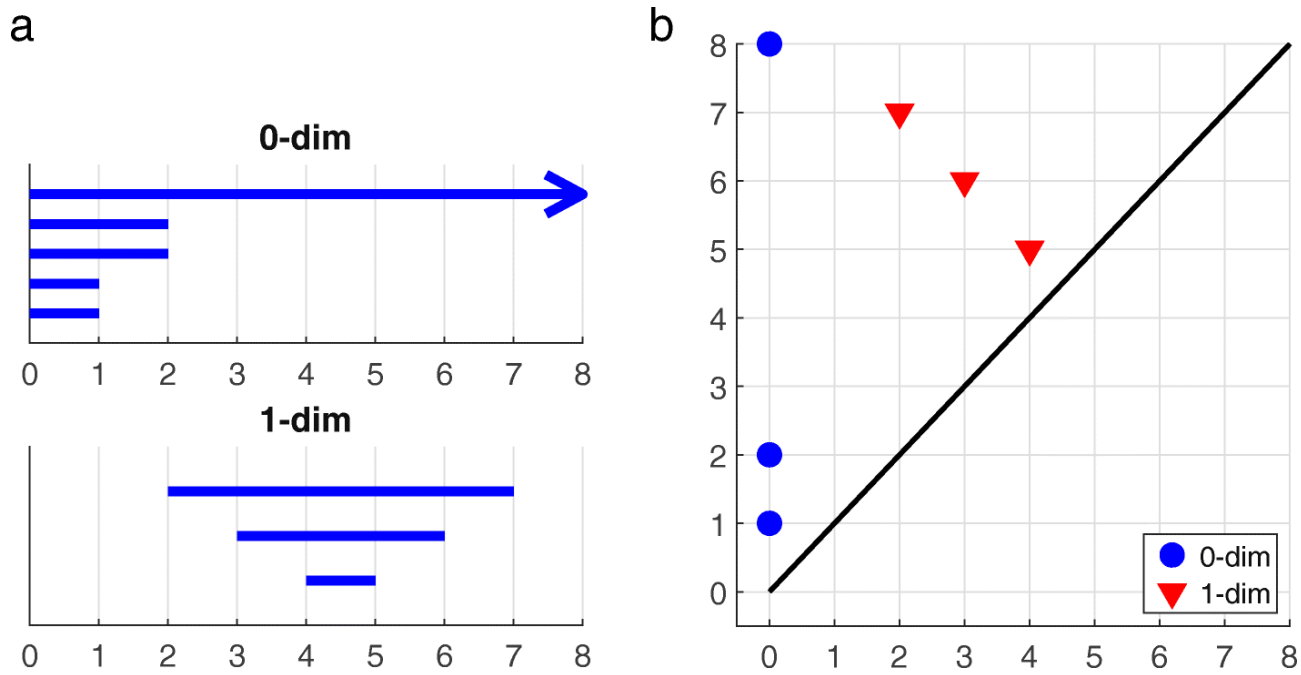


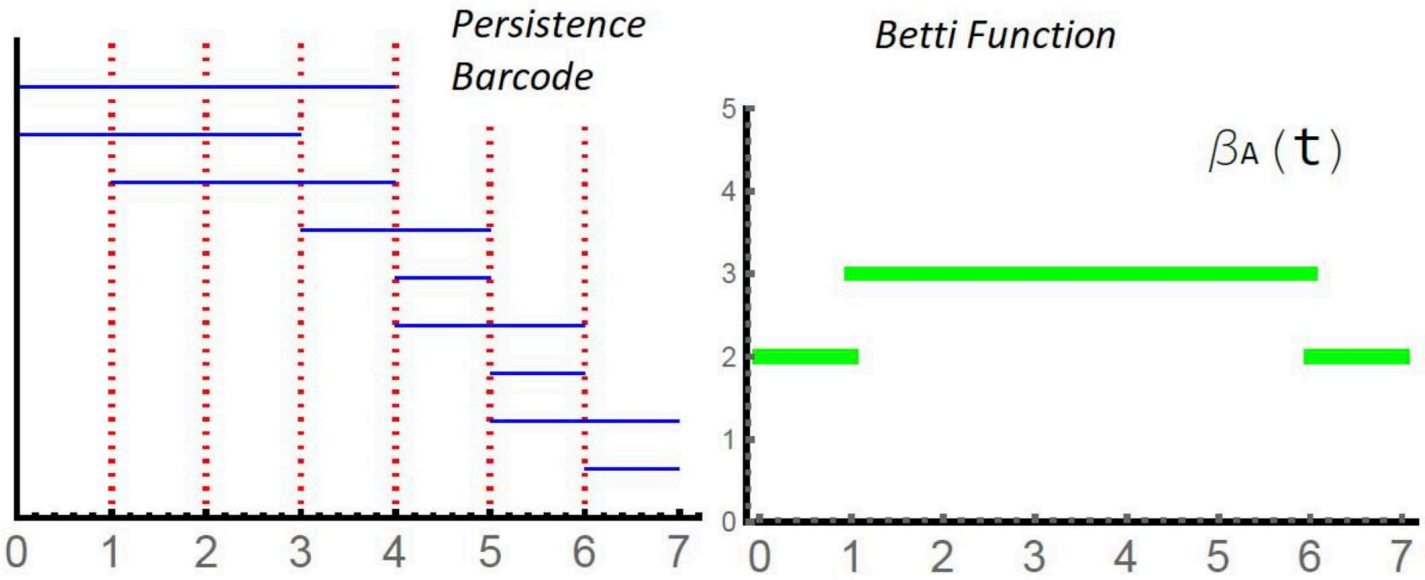
Image Credit: Sizemore et al, Knowledge gaps in the early growth of semantic feature networks, Nature Human Behavior, 2018

Persistence barcode

Persistence Diagram



Vectorization of Persistence Diagrams (Barcode)



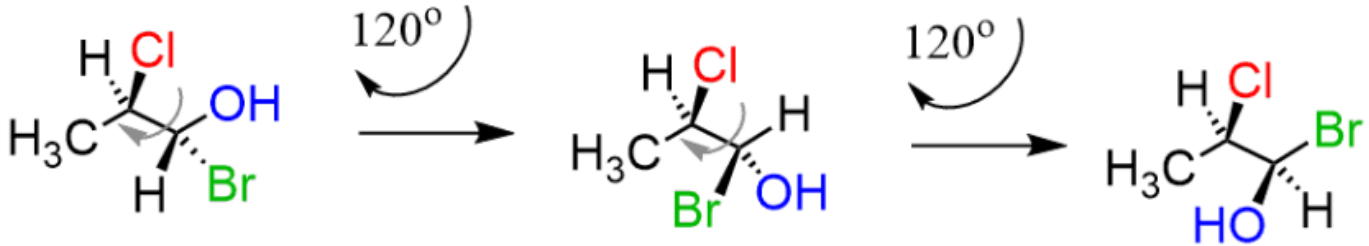
MULTIPARAMETER PERSISTENCE

- While single parameter persistence tracks the topological changes in 1-parameter sequence, multiparameter persistence aims to do it for 2 or more parameters.
- Serious mathematical problems make multiparameter approach infeasible in general.
- There are remedial methods to bypass these serious problems.
- We propose and apply one of these approaches in this work.

2D VS. 3D MODELS

Conformation Problem

Conformation also referred to as conformers or conformational isomers, are different arrangements of atoms that occur as a **result of rotation about single bonds**. For example, in the following molecule, we can have a different arrangement of atoms by rotating around the middle σ bond:



Different conformers = conformational isomers

Image credit: https://www.chemistrysteps.com

DRUG DISCOVERY WITH TDA

- We define a new fingerprinting method for compounds by using persistent homology.
- **Topological Fingerprints of Compounds:**

1. Realize a compound as a **graph** (Atoms→ Nodes, Bonds→Edges)
2. Use important chemical quantities as **node and edge functions**.
3. Obtain a **bifiltration** decomposing the compound into substructures (subgraphs G_{ij})
4. **Extract a vector** (mxn matrix) from bifiltration (multipersistence module) summarizing topology of G_{ij}

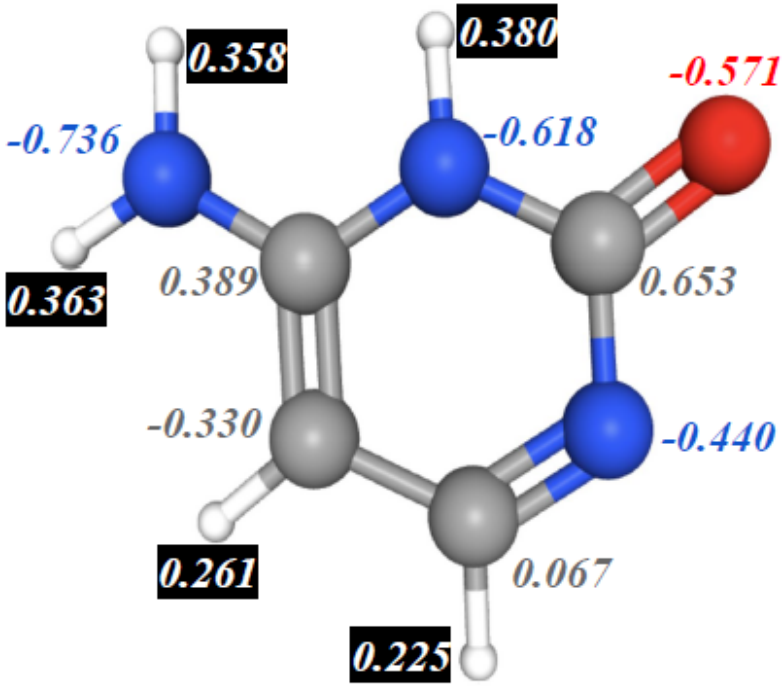


Figure 2: **Cytosine**. Atom types are coded by their color: White=Hydrogen, Gray=Carbon, Blue=Nitrogen, and Red=Oxygen. The decimal numbers next to atoms represent their partial charges.

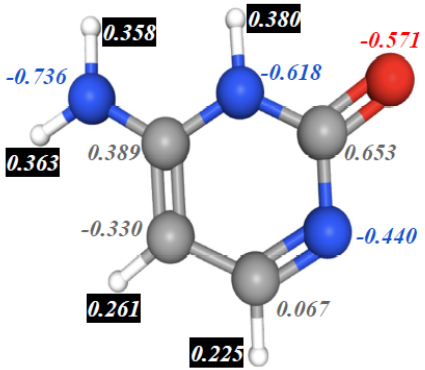
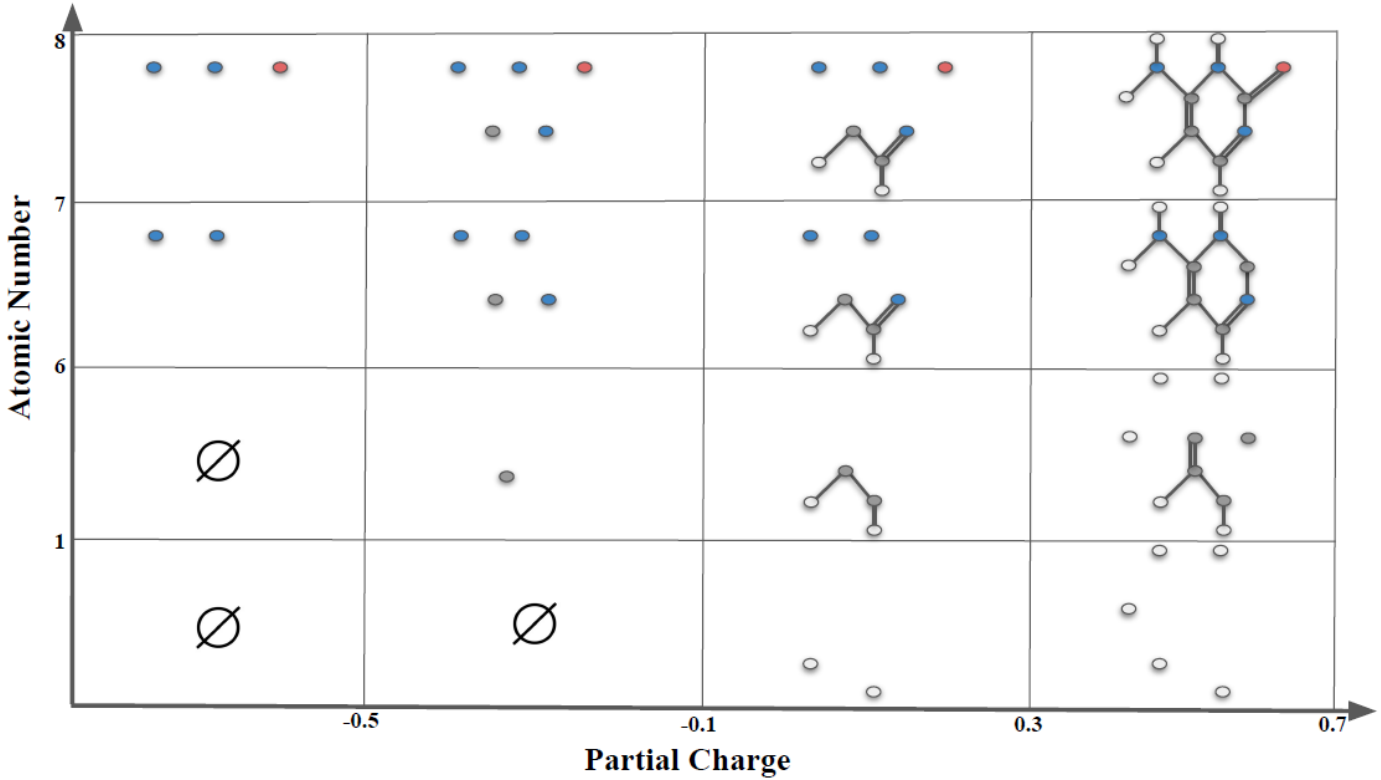
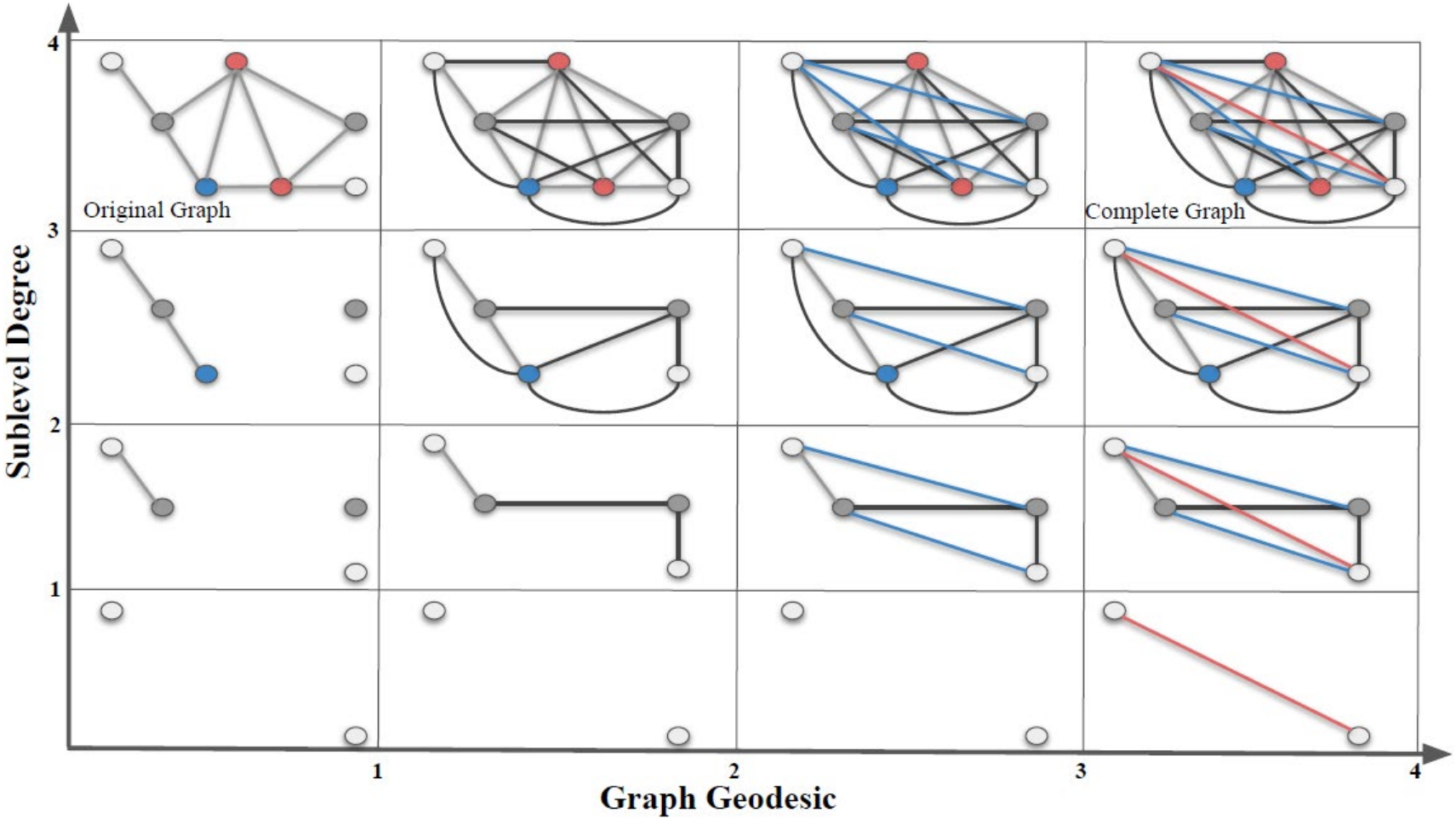


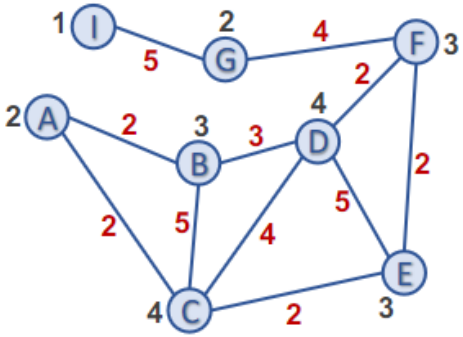
Figure 2: **Cytosine**. Atom types are coded by their color: White=Hydrogen, Gray=Carbon, Blue=Nitrogen, and Red=Oxygen. The decimal numbers next to atoms represent their partial charges.

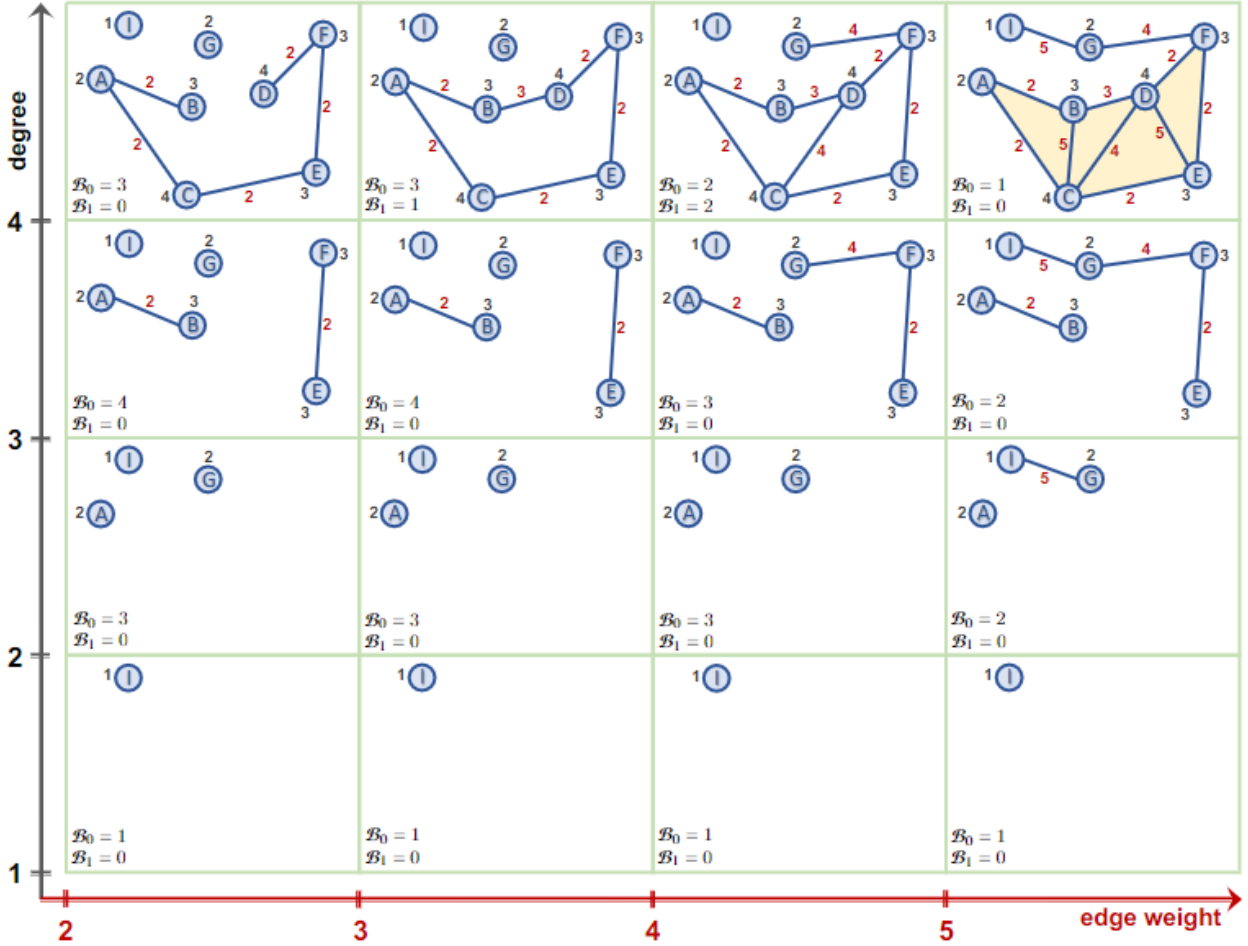


Atomic Number + Partial Charge Sublevel Bifiltration



Degree + Vietoris-Rips Filtration





Degree + Edge Weight Filtration

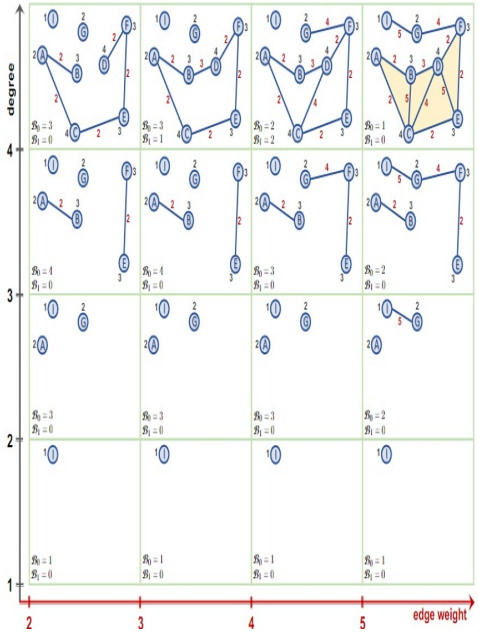
TOPOLOGICAL COMPOUND FINGERPRINTING

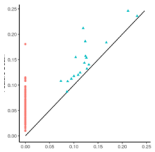
- mxn bifiltration
- horizontal slicing

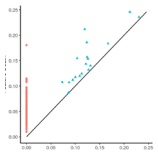
- m single persistence diagram

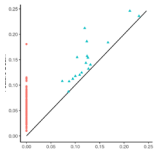
- m vectorization (Betti, Silhouette)

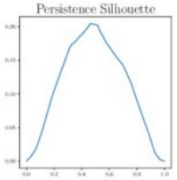
- m x n matrix

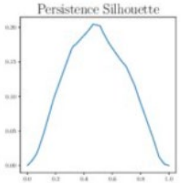


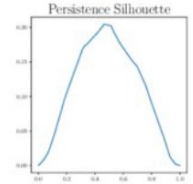












$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}$$

EXPERIMENTS - DATASETS

Table 3: Summary statistics of the Cleves-Jain dataset.

Target	# Training Samples	# Test Samples
a	3	6
b	3	22
c	2	13
d	3	6
e	2	5
f	2	4
g	2	5
h	2	5
i	2	5
j	3	14
k	3	14
l	3	10
m	3	9
n	2	10
o	3	30
p	3	23
q	3	11
r	2	14
s	3	15
t	2	5
u	3	9
v	3	7
Decoy	0	850

Table 4: Summary statistics of the DUD-E Diverse dataset.

Target	Description	# Active	# Decoy
AMPC	beta-lactamase	62	2902
CXCR4	C-X-C chemokine receptor type 4	122	3414
KIF11	kinesin-like protein 1	197	6912
CP3A4	cytochrome P450 3A4	363	11940
GCR	glucocorticoid receptor	563	15185
AKT1	serine/threonine-protein kinase Akt-1	423	16576
HIVRT	HIV type 1 reverse transcriptase	639	19134
HIVPR	HIV type 1 protease	1395	36278

PERFORMANCE METRIC

- **Enrichment Factor** (EF) is the most common performance evaluation metric for Virtual Screening methods.

- Let N be the total number of ligands in the dataset, A_{φ} be the number of true positives (i.e., correctly predicted active ligands) in the first α% of all ligands and $N_{actives}$ be the number of active ligands in the whole dataset. Then,

$$EF_{\alpha\%} = \frac{A_{\varphi}/N_{total.\alpha\%}}{N_{actives}/N_{total}} = \frac{A_{\varphi}}{N_{actives} \cdot \alpha\%}$$

- Notice that the **maximum score** for $EF_{\alpha\%}$ is $\frac{100}{\alpha}$, i.e., 100 for $EF_{1\%}$ and 20 for $EF_{5\%}$.

- **Example:** N=1000, $N_{actives}$= 40 α=5 A_{φ} = 18 => $EF_{5\%}$ = 9 max $EF_{5\%}$= 20.

RESULTS

Table 1: Comparison of EF 2%, 5%, 10% and AUC values between ToDD and other virtual screening methods on the Cleves-Jain dataset.

Model	EF 2% (max. 50)	EF 5% (max. 20)	EF 10% (max. 10)	AUC
USR [7]	10.0	6.2	4.1	0.76
GZD [83]	13.4	8.0	5.3	0.81
PS [42]	10.7	6.6	4.9	0.78
ROCS [36]	20.1	10.7	6.2	0.83
USR + GZD [75]	13.7	7.7	4.7	0.81
USR + PS [75]	13.1	7.9	5.0	0.80
USR + ROCS [75]	17.1	9.1	5.4	0.83
GZD + PS [75]	16.0	9.1	5.9	0.82
PH_VS [48]	18.6	NA	NA	NA
GZD + ROCS [75]	20.3	10.8	5.3	0.83
PS + ROCS [75]	20.5	10.7	6.4	0.83
ToDD-RF	35.2±2.3	15.6±1.0	8.1±0.4	**0.94**±0.02
ToDD-ViT	**39.6**±1.4	**18.6**±0.4	**9.9**±0.1	0.90±0.01
Relative gains	92.9%	83.7%	54.1%	13.3%

Relative gains are relative to the next best performing model.
Mean and standard deviation of EF scores evaluated by 5-fold cross-validation.

RESULTS

Table 2: Comparison of EF 1% (max. 100) between ToDD and other virtual screening methods on 8 targets of the DUD-E Diverse subset.

Model	AMPC	CXCR4	KIF11	CP3A4	GCR	AKT1	HIVRT	HIVPR	Avg.
Findsite [90]	0.0	0.0	0.9	21.7	34.2	39.0	1.2	34.7	16.5
Fragsite [91]	4.2	42.5	0.0	32.9	29.1	47.1	2.4	48.7	25.9
Gnina [78]	2.1	15.0	38.0	1.2	39.0	4.1	11.0	28.0	17.3
GOLD-EATL [87]	25.8	20.0	33.5	17.9	34.6	29.2	28.7	23.4	26.6
Glide-EATL [87]	35.5	20.8	30.5	15.1	24.0	31.6	29.0	22.0	26.1
CompM [87]	32.3	25.0	35.5	33.6	37.1	44.2	30.2	25.0	32.9
CompScore [66]	39.6	51.6	51.3	14.0	27.1	37.6	21.8	18.2	32.7
CNN [68]	2.1	5.0	11.2	28.7	12.8	84.6	12.2	9.9	20.8
DenseFS [44]	14.6	5.0	4.3	44.3	20.9	89.4	12.8	8.4	25.0
SIEVE-Score [88]	30.7	61.1	53.4	6.7	33.3	42.1	39.8	38.3	38.2
DeepScore [85]	28.1	56.8	54.3	37.1	40.9	59.0	43.8	62.8	47.9
RF-Score-VSv3 [88]	32.3	60.9	4.5	25.9	32.5	41.9	39.8	65.7	37.9
ToDD-RF	42.9±4.5	**92.3**±3.2	**75.0**±5.0	**67.6**±3.4	**78.9**±4.0	**90.7**±1.3	**64.1**±2.3	**92.1**±1.5	**73.7**
ToDD-ConvNeXt	**46.2**±3.6	84.6±2.8	72.5±3.6	28.8±2.8	46.0±2.0	81.2±2.5	37.5±3.6	74.6±1.0	58.9
Relative gains	16.7%	51.1%	38.1%	52.6%	92.9%	1.5%	46.3%	40.2%	53.9%

Relative gains are relative to the next best performing model.
Mean and standard deviation of EF scores evaluated by 5-fold cross-validation.

RESULTS SUMMARY

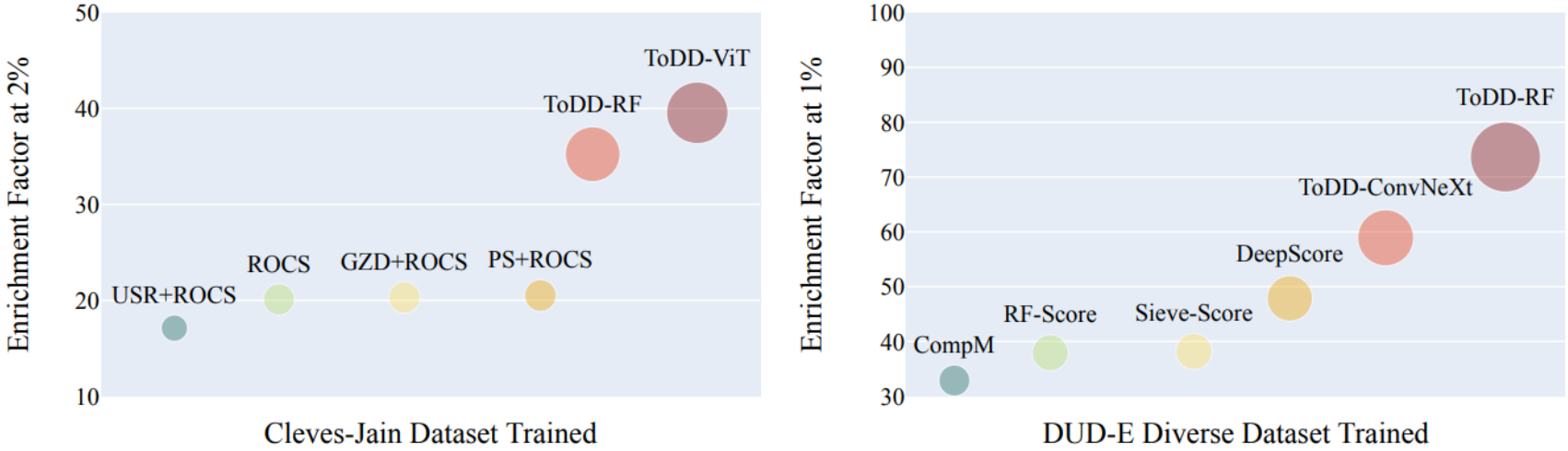


Figure 1: **Comparison of virtual screening performance.** Each bubble's diameter is proportional to its EF score. ToDD offers significant gain regardless of the choice of classification model such as random forests (RF), vision transformer (ViT) or a modernized ResNet architecture ConvNeXt. The standard performance metric $EF_{\alpha\%}$ is defined as $\frac{100}{\alpha}$, and therefore the maximum attainable value is 50 for $EF_{2\%}$, and 100 for $EF_{1\%}$.

COMPARISON OF FILTRATIONS

Table 5: EF 2% values and ROC-AUC scores across different modalities on Cleves-Jain dataset using **ToDD-RF**.

Target	Atomic Mass	Partial Charge	Bond Type	Atomic Mass & Partial Charge	All Modalities
a	33.3	33.3	33.3	33.3	41.7
b	25.0	29.5	31.8	27.3	25.0
c	19.2	7.7	15.4	26.9	34.6
d	33.3	33.3	41.7	50.0	50.0
e	30.0	30.0	30.0	40.0	40.0
f	25.0	50.0	37.5	50.0	37.5
g	30.0	30.0	30.0	30.0	40.0
h	40.0	50.0	30.0	50.0	50.0
i	40.0	40.0	30.0	40.0	40.0
j	17.9	39.3	35.7	28.6	28.6
k	21.4	21.4	17.9	35.7	32.1
l	15.0	15.0	15.0	30.0	25.0
m	44.4	50.0	33.3	50.0	38.9
n	15.0	25.0	10.0	25.0	10.0
o	21.7	20.0	25.0	23.3	23.3
p	10.9	8.7	13.0	17.4	26.1
q	45.5	27.3	22.7	40.9	40.9
r	42.9	42.9	42.9	39.3	32.1
s	26.7	16.7	20.0	20.0	30.0
t	30.0	50.0	50.0	50.0	50.0
u	33.3	38.9	27.8	38.9	50.0
v	21.4	28.6	28.6	28.6	28.6
Mean	28.3	31.3	28.3	35.2	35.2
ROC-AUC	0.92	0.90	0.88	0.94	0.93

KEY TAKEAWAYS

1. **Novelty**: We developed a new compound fingerprinting method by facilitating multiparameter persistence approach.

2. **Performance**: We develop and benchmark ML approaches; and outperform the SOTA by a wide and statistically significant margin: **93% gain** for Cleves-Jain and **54% gain** for DUD-E Diverse dataset.

3. **Small data sets**: effective **few-shot classification** (only 2-3 active ligands per drug target for training)

4. **ML integration**: features suited for SoTA Neural Networks, as well as traditional ML methods

5. **Computational efficiency**: Full training + analysis on a laptop **~7 minutes** (for a library of 1100 compounds, distributed across the 8 cores of an Intel Core i7 CPU (100GB RAM))

POTENTIAL FUTURE WORK

1. Testing the performance of ToDD on **ultra-large Virtual Screening datasets** with millions of compounds such as MUV, DUD-E and custom datasets of Novartis.

2. Trying other MP vectorization methods, e.g., MP Image - vineyards (Carriere-Blumberg), MP Landscapes (Vipond).

3. Using transfer learning to adapt state-of-the-art convolutional and transformer based computer vision models to extract complex chemical properties of compounds, specifically for **few-shot learning problems.**

4. There are other subdomains in chemistry that ToDD can be benchmarked and tested such as: **molecular property prediction,** e.g. solubility, polarization, binding affinity.