# Persistent function based machine learning for drug design

## Xiang Liu
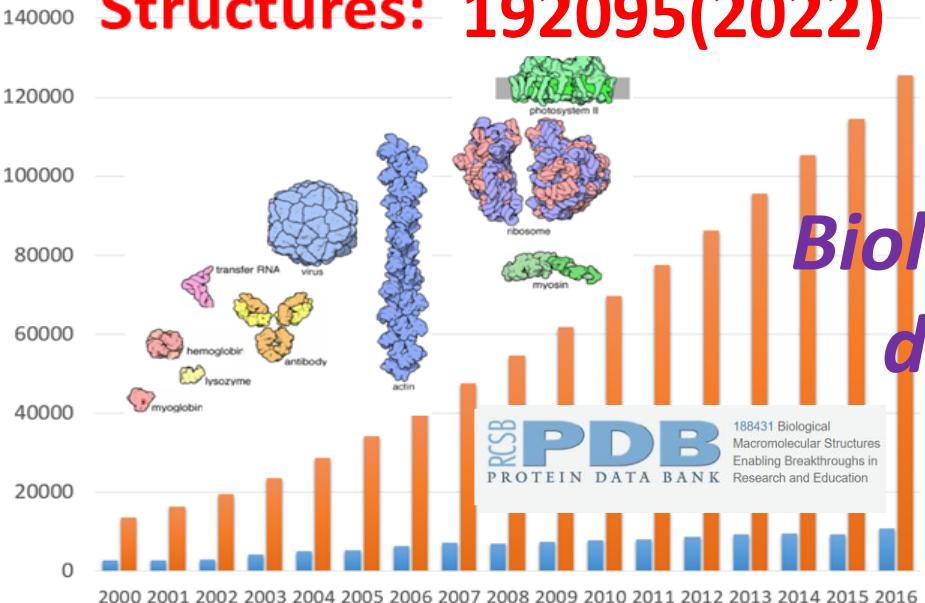
*Nankai University & BIMSA*

## September 19, 2022

# Drug Discovery Process (simplified)

## Clinical Trials

| Target Discovery | Lead Discovery | Lead Optimization | •Preclinical Development | Phase 1 | Phase 2 | Phase 3 | Launch |
|---|---|---|---|---|---|---|---|
| •Target identification<br>•Microarray profiling<br>•Target validation<br>•Assay development<br>•Biochemistry<br>•Clinical/Animal disease models | •High-throughput Screening (HTS)<br>•Fragment-based screening<br>•Focused libraries<br>•Screening collection | •Medicinal Chemistry<br>•Structure-based drug design<br>•Selectivity screens<br>•ADMET screens<br>•Cellular/Animal disease models<br>•Pharmacokinetics | •Toxicology<br>•In vivo safety pharmacology<br>•Formulation<br>•Dose prediction | PK tolerability | Efficacy | Safety & Efficacy | Indication Discovery & expansion |

| Discovery | Development | Use |
|---|---|---|

| Med. Chem. ML, | Clinical Candidates | Drugs |
|---|---|---|

>450,000 distinct compounds
~25,000 distinct lead series

~12,000 candidates

~1,200 drugs

**Time: > 10 years**       **Cost: > 2.6 billion$**       **high failure rate**

Structures: 192095(2022)

Gene: >220 Millions (2022)

Biological data

Computational power

Learning algorithm

# Artificial Intelligence

Enabling machines to think like humans

## Machine Learning

Training machines to get better at a task without explicit programming

### Deep Learning

Using multi-layered networks for machine learning
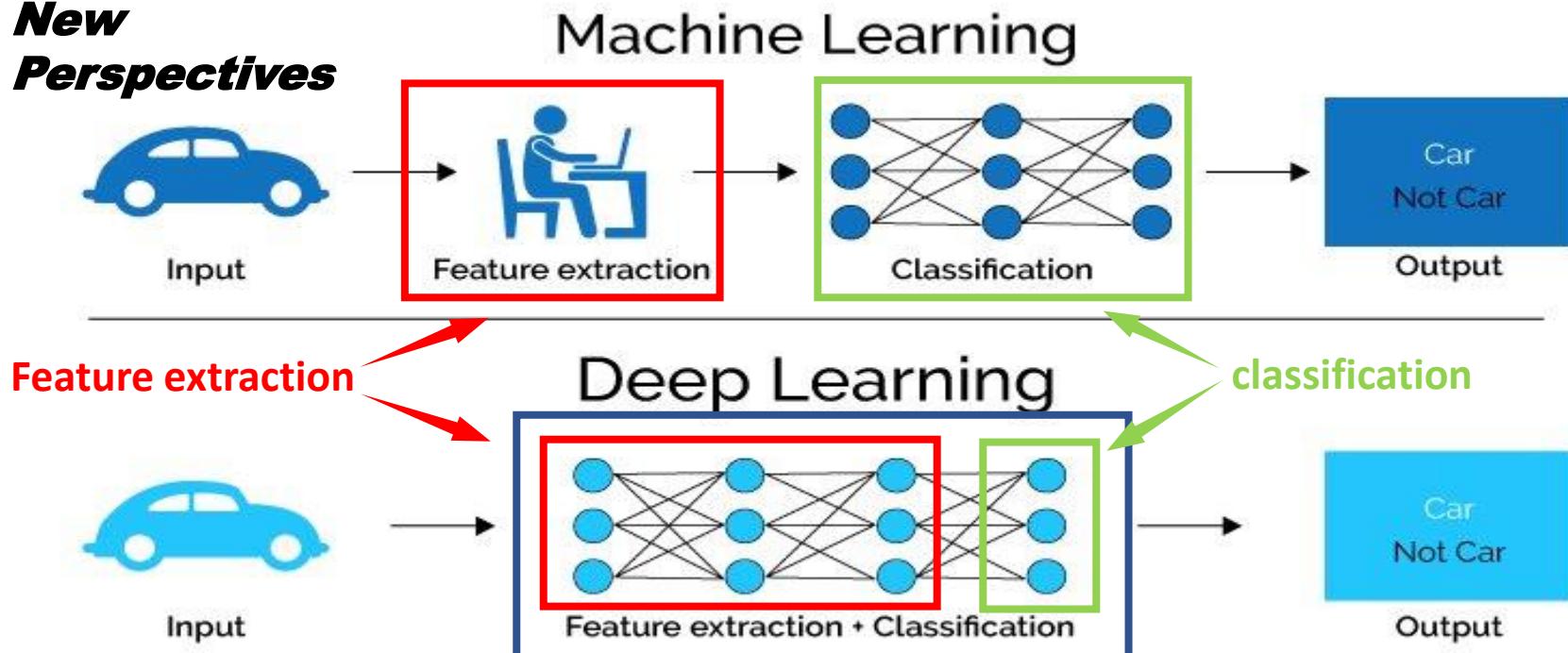
# Feature extraction and feature learning

"**The success of machine learning algorithms generally depends on data representation...**"
*Y. Bengio, etc, "Representation Learning: A Review and New Perspectives*
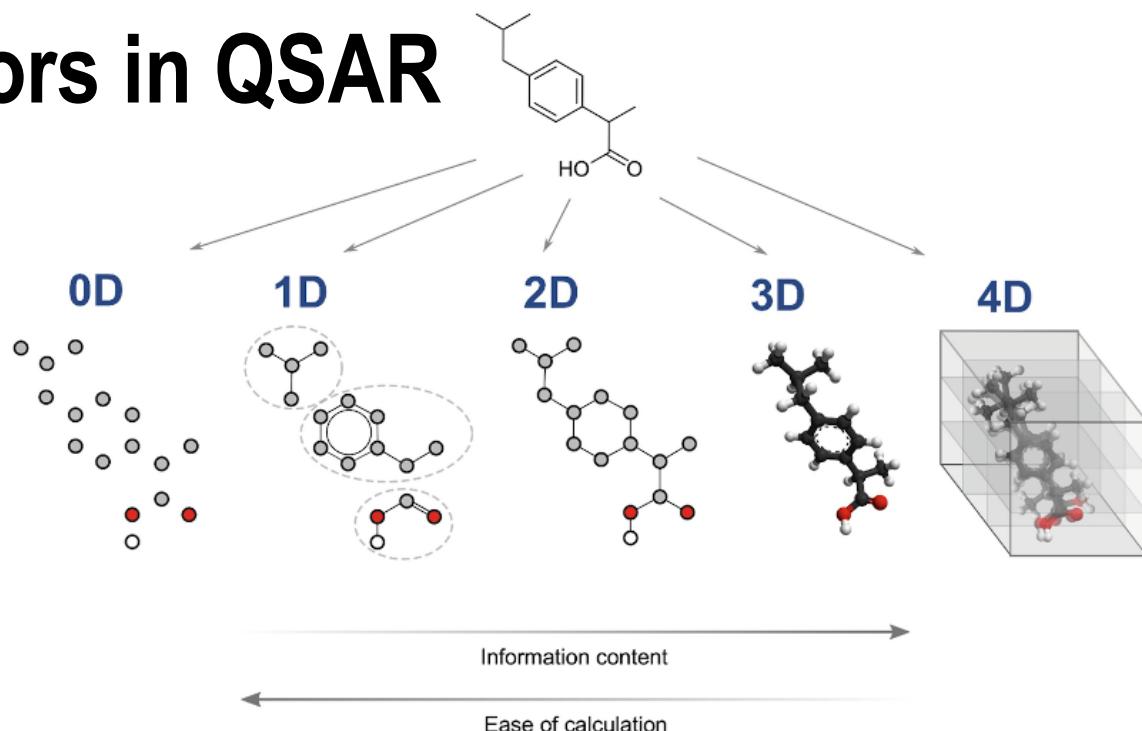
"**The deep learning research aims at discovering learning algorithms that discover multiple levels of distributed representations...**"
*Y. Bengio, "Deep Learning of Representations: Looking Forward*



**Machine Learning**

Input → Feature extraction → Classification → Output (Car / Not Car)

**Deep Learning**

Input → Feature extraction + Classification → Output (Car / Not Car)

Feature extraction

classification

# Molecular descriptors in QSAR models

**More than 5000 Molecular descriptors in Quantitative Structure Activity relationship (QSAR) models.**

Grisoni F, Ballabio D, Todeschini R, et al. Molecular descriptors for structure–activity applications: a hands-on approach[M]// Computational Toxicology. Humana Press, New York, NY, 2018: 3-53.



**0D    1D    2D    3D    4D**

Information content →

← Ease of calculation
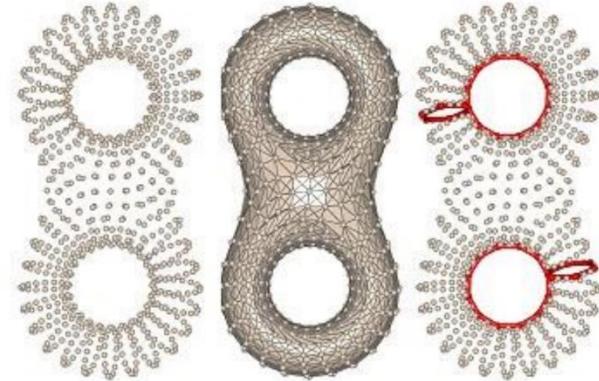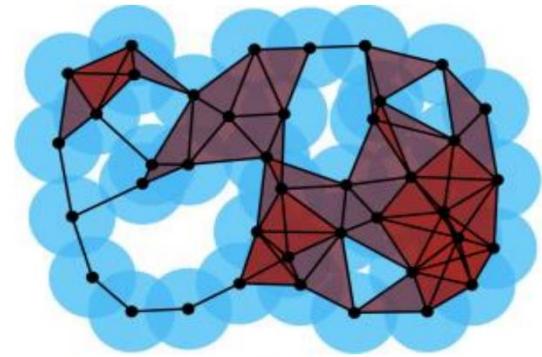
## Common chemical descriptors for QSAR/QSPR analysis

| Chemical descriptors | Based on | Examples |
|---|---|---|
| **Theoretical descriptors** | | |
| 0D | Molecular formula | Molecular weights, atom counts, bond counts |
| 1D | Chemical graph | Fragment counts, functional group counts |
| 2D | Structural topology | Weiner index, Balaban index, Randic index, BCUTS |
| 3D | Structural geometry | WHIM, autocorrelation, 3D-MORSE, GETAWAY |
| 4D | Chemical conformation | Volsurf, GRID, Raptor |
| | | |
| **Experimental descriptors** | | |
| Hydrophobic parameters | Hydrophobicity | Partition coefficents (logP), hydrohobic substituent constant ($\pi$) |
| Electronic parameters | Electronic properties | Acid dissociation constant, Hammett constant |
| Steric parameters | Steric properties | Taft steric constant, Charton's constant |

# Topological Data Analysis (TDA)



**Topological invariant;
Homology;
Homotopy;
Simplicial complex;
Morse theory;
Reeb graph;**

**......**
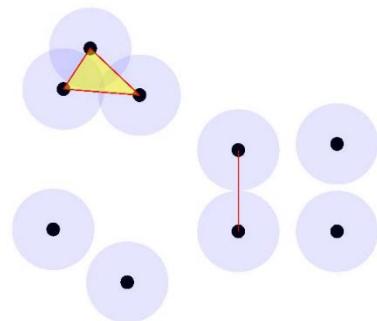
Computational Geometry;
Computational topology;
Algebraic topology
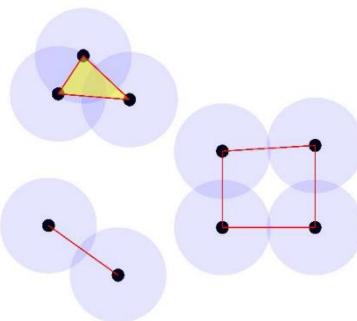
# Topological data analysis---persistent homology



**A**

$f_1=0.4$            $f_2=0.5$            $f_3=0.6$            $f_4=0.8$

$\beta_0:6$  $\beta_1:0$      $\beta_0:3$  $\beta_1:1$      $\beta_0:1$  $\beta_1:1$      $\beta_0:1$  $\beta_1:1$

**B**

$\beta_0=2$  $\beta_1=1$  $\beta_2=1$

**C**

$f_1$  $f_2$  $f_3$        $f_4$

$\beta_0$

$\beta_1$

*filtration*
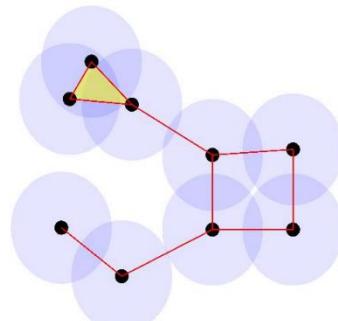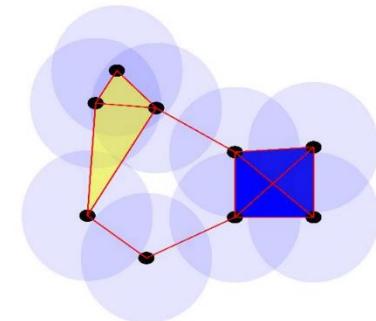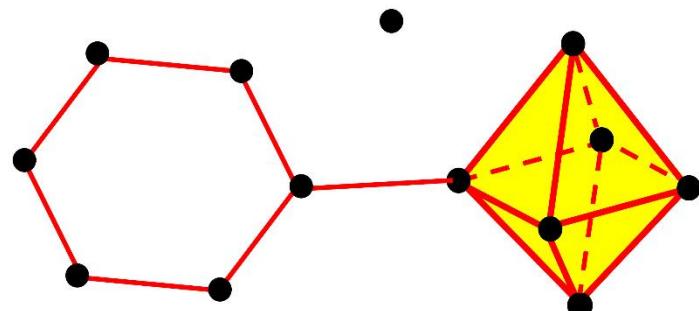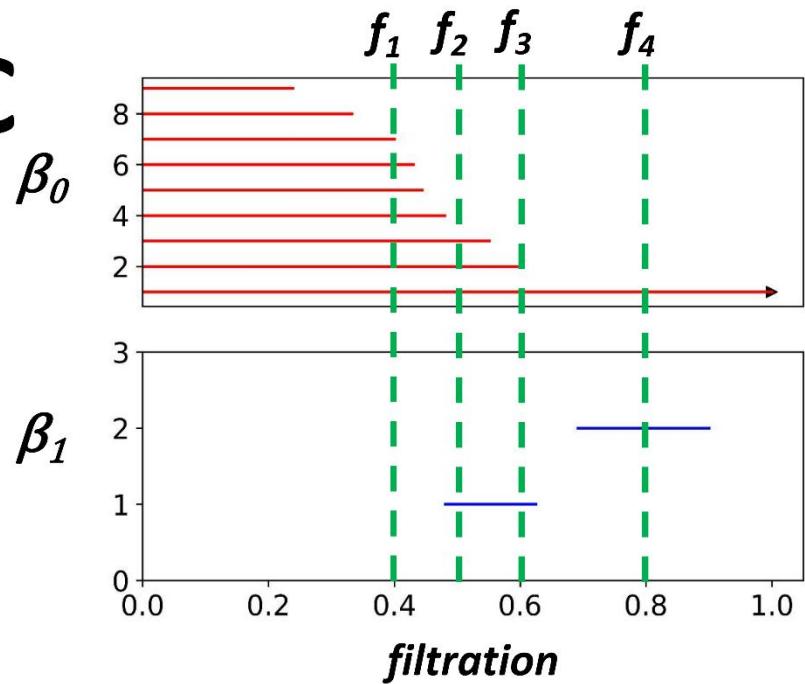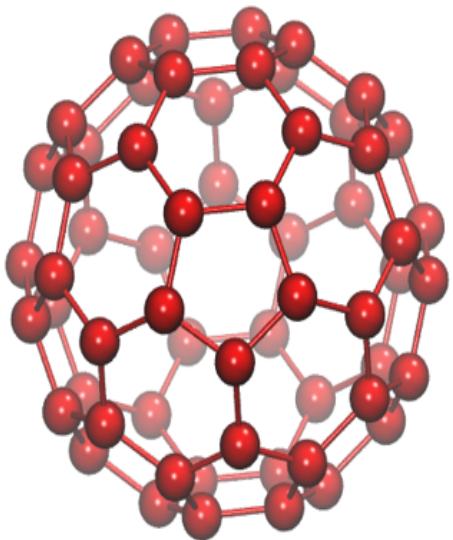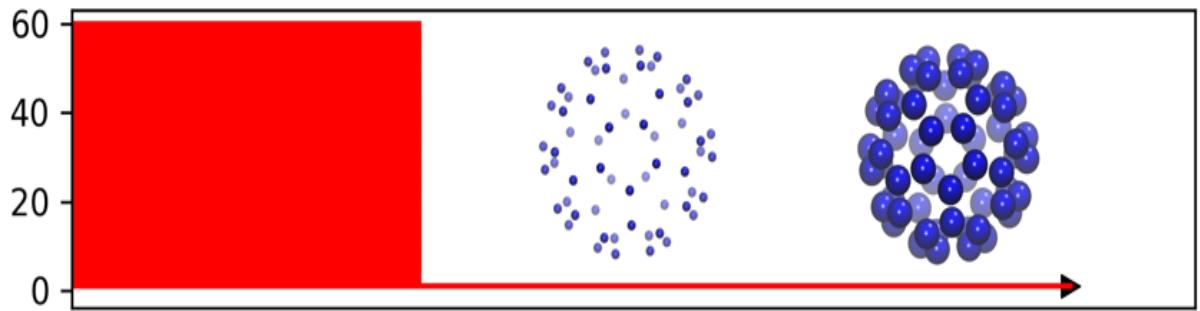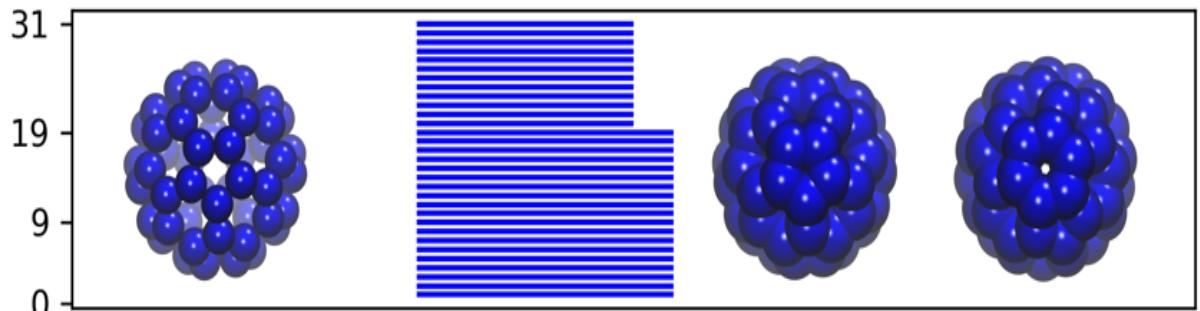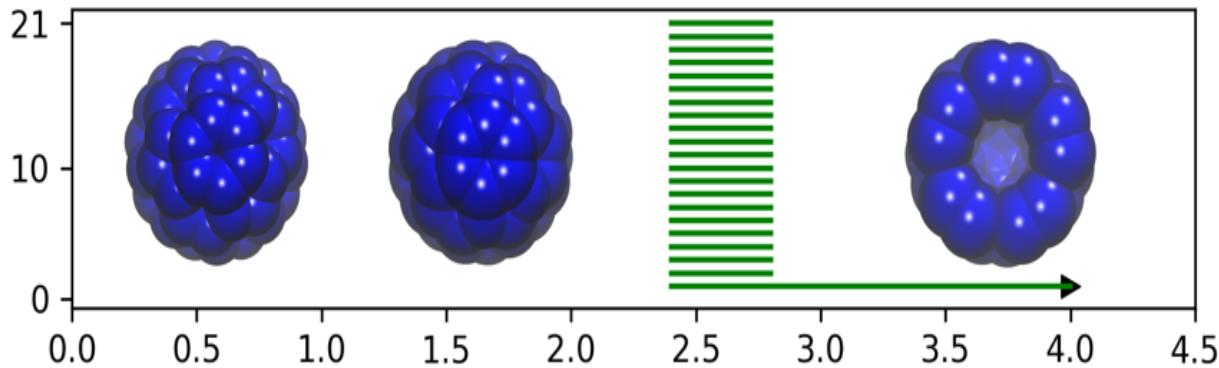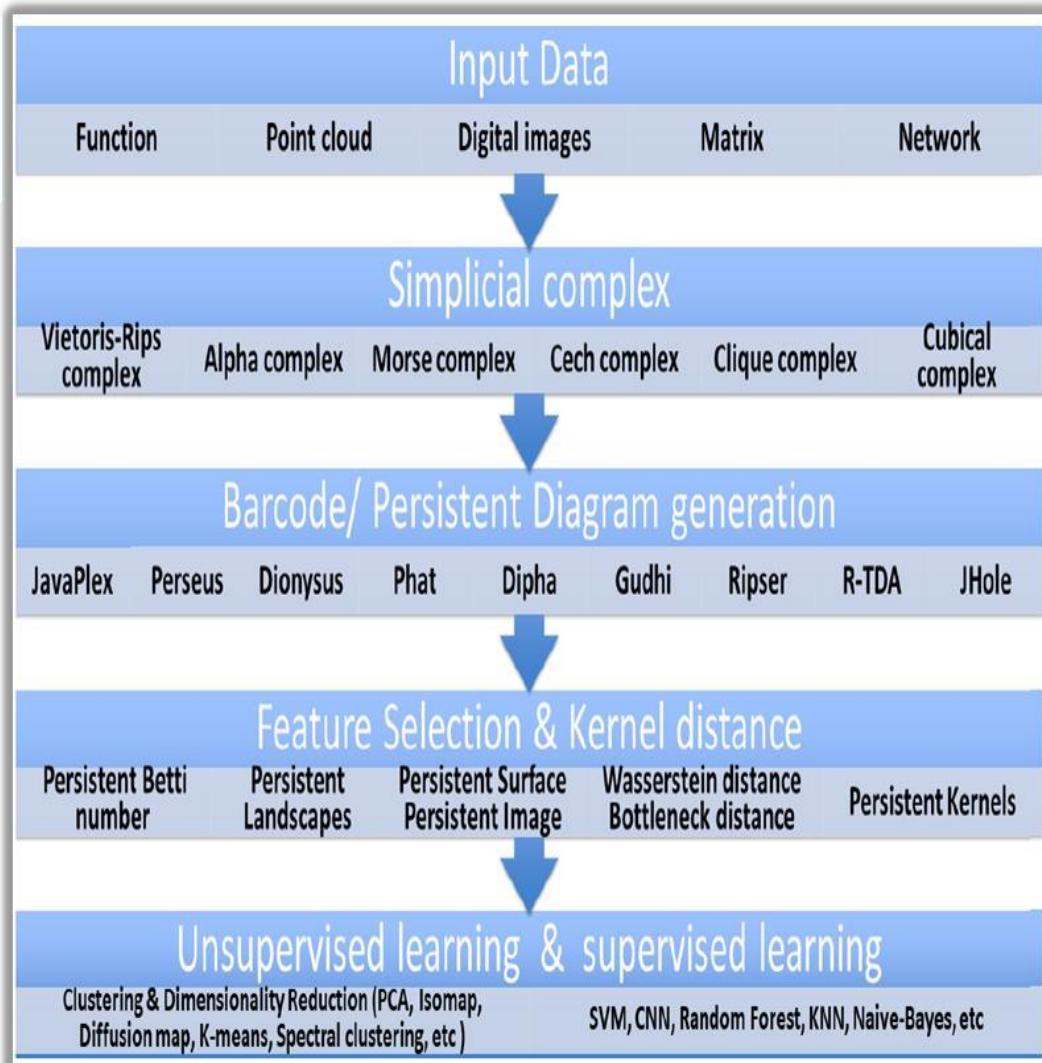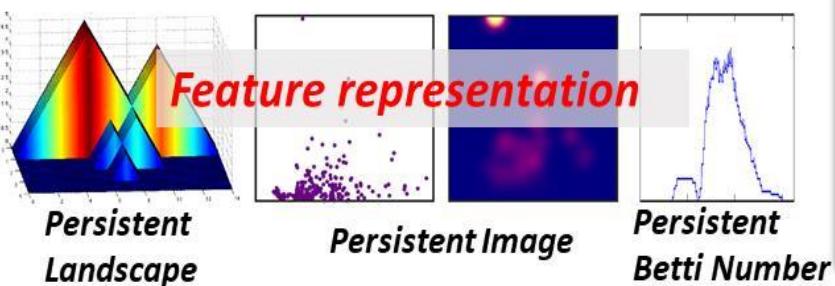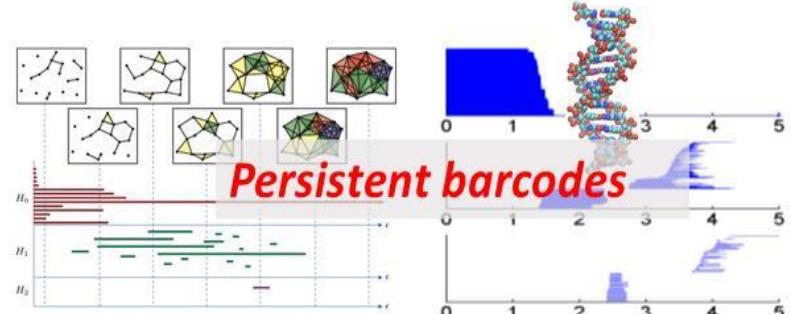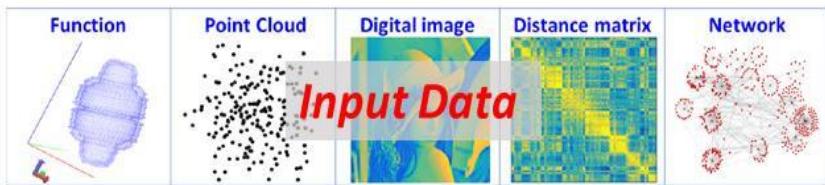
# Persistent Homology Analysis of Carbon-60

$C_{60}$

# TDA based machine learning models
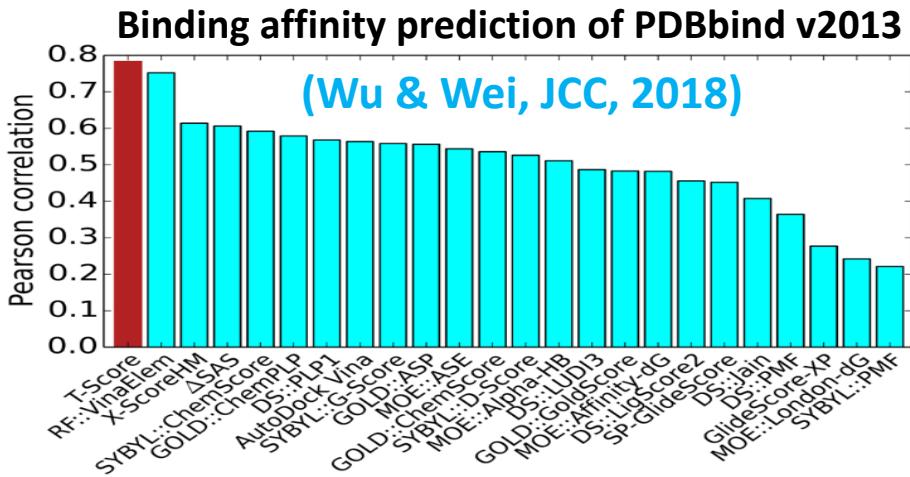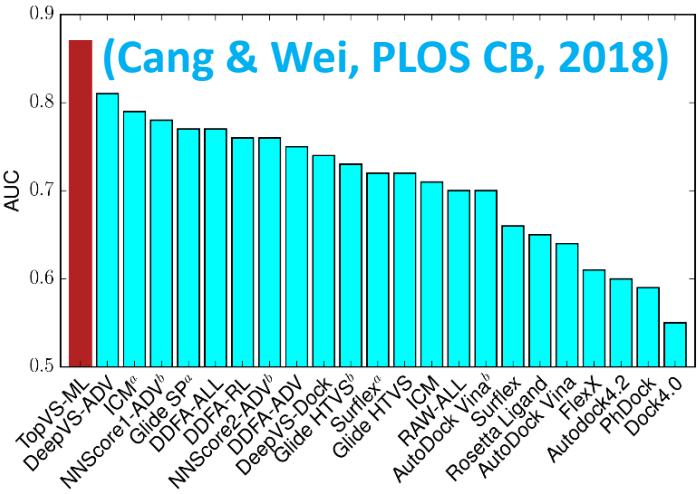
# Recent progress in TDA based drug design
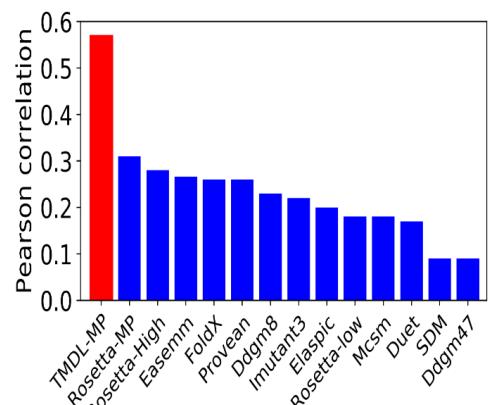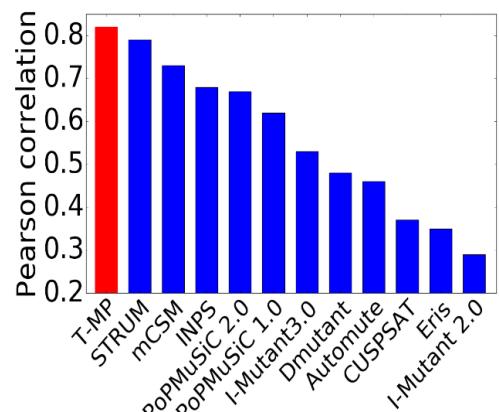
**Guowei Wei**
**MSU Foundation professor**
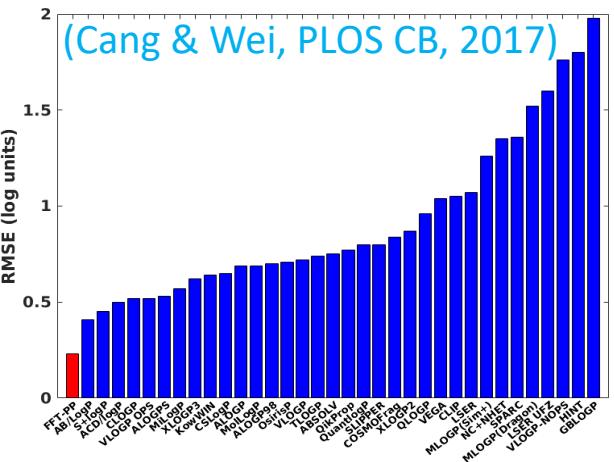
**DUD database** **128374 protein-ligand/decoy pairs**



(Cang & Wei, PLOS CB, 2018)



**Binding affinity prediction of PDBbind v2013**

(Wu & Wei, JCC, 2018)

**Prediction correlations for 2648 mutations on globular proteins**
(Cang & Wei, PLOS CS, 2017)



**Prediction RMSD of logP(star set )**

(Cang & Wei, PLOS CB, 2017)

# Recent progress in TDA based drug design

**Drug Design Data Resource (D3R) Grand Challenges**

**Grand Challenge 2：** win **14%**
**Grand Challenge 3:** win **38%** while the **second** winner had a rate of **19%**
**Grand Challenge 4:** win **50%**

*Wei Team's performance at D3R Grand Challenge*

## D3R Grand Challenge 2

**Stage 1**
Pose Predictions (partials)
Scoring (partials)
Free Energy Set 1 (partials)
Free Energy Set 2 (partials)
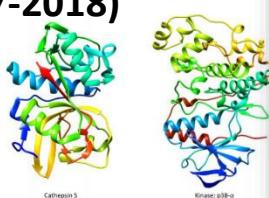
**Stage 2**
Scoring (partials)
Free Energy Set 1 (partials)
Free Energy Set 2 (partials)

## D3R Grand Challenge 3 (2017-2018)

**Pose Prediction**

**Cathepsin Stage 1A**
Pose Predictions (partials)

**Cathepsin Stage 1B**
Pose Prediction

**Affinity Rankings excluding Kds > 10 µM**

**Cathepsin Stage 1**
Scoring (partials)
Free Energy Set

**Cathepsin Stage 2**
Scoring (partials)
Free Energy Set

**VEGFR2**
Scoring (partials)

**JAK2 SC2**
Scoring (partials)

**p38-α**
Scoring

**JAK2 SC3**
Scoring
Free Energy Set

**TIE2**
Scoring
Free Energy Set 2

**ABL1**
Scoring (partials)

**Active / Inactive Classification**

**VEGFR2**
Scoring (partials)

**JAK2 SC2**
Scoring (partials)

**p38-α**
Scoring (partials)

**JAK2 SC3**
Scoring
Free Energy Set

**TIE2**
Scoring (partials)
Free Energy Set 1

**ABL1**
Scoring (partials)

**Affinity Rankings for Cocrystalized Ligands**

**Cathepsin Stage 1**
Scoring (partials)
Free Energy Set

**Cathepsin Stage 2**
Scoring (partials)
Free Energy Set

## D3R Grand Challenge 4 (2018-2019)

**Pose Predictions**

**BACE Stage 1A**  1/2  3/3
Pose Predictions (Partials)

**BACE Stage 1B**  2/2  1/2
Pose Prediction (Partials)
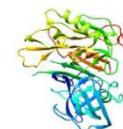
**Affinity Predictions**

**Cathepsin Stage 1**
Combined Ligand and Structure Based Scoring  2/5  2/3  2/4

Ligand Based Scoring (No participation)

Structure Based Scoring  2/4  3/3  3/3

Free Energy Set  1/7  2/5

**BACE Stage 1**
Combined Ligand and Structure (No participation)

Ligand Based Scoring (Partials) (No participation)

Structure Based Scoring (Partials) (No participation)

Free Energy Set (No participation)
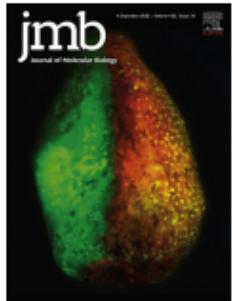
**BACE Stage 2**
Combined Ligand and Structure

Ligand Based Scoring (No participation)

Structure Based Scoring (Partials)

Free Energy Set  3/4  1/4

# TDA-based learning models in SARS-Cov-2

## Mutations Strengthened SARS-CoV-2 Infectivity

**Jiahui Chen**[1], **Rui Wang**[1], **Menglun Wang**[1] and **Guo-Wei Wei**[1,2,3]

1 - Department of Mathematics, Michigan State University, MI 48824, USA
2 - Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA
3 - Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA

Correspondence to Guo-Wei Wei: wei@math.msu.edu

**Wei's Team predicts key mutation sites in prevailing variants**

Mutations at 501 and 452 in prevailing SARS-Cov-2 variants

Alpha: N501Y
Beta: K417N, E484K, N501Y
Gamma: K417T, E484K, N501Y
Delta: L452R, T478K
Epsilon: L452R
Kappa: L452R, E484Q
Omicron: N501,…

**They discovered the mechanism of viral transmission and evolution:
more infectious**

## Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infectivity is a major concern in coronavirus disease 2019 (COVID-19) prevention and economic reopening. However, rigorous determination of SARS-CoV-2 infectivity is very difficult owing to its continuous evolution with over 10,000 single nucleotide polymorphisms (SNP) variants in many subtypes. We employ an algebraic topology-based machine learning model to quantitatively evaluate the binding free energy changes of SARS-CoV-2 spike glycoprotein (S protein) and host angiotensin-converting enzyme 2 receptor following mutations. We reveal that the SARS-CoV-2 virus becomes more infectious. Three out of six SARS-CoV-2 subtypes have become slightly more infectious, while the other three subtypes have significantly strengthened their infectivity. We also find that SARS-CoV-2 is slightly more infectious than SARS-CoV according to computed S protein-angiotensin-converting enzyme 2 binding free energy changes. Based on a systematic evaluation of all possible 3686 future mutations on the S protein receptor-binding domain, we show that most likely future mutations will make SARS-CoV-2 more infectious. Combining sequence alignment, probability analysis, and binding free energy calculation, we predict that a few residues on the receptor-binding motif, i.e., 452, 489, 500, 501, and 505, have high chances to mutate into significantly more infectious COVID-19 strains.

# TDA is based on the multiscale simplicial complex
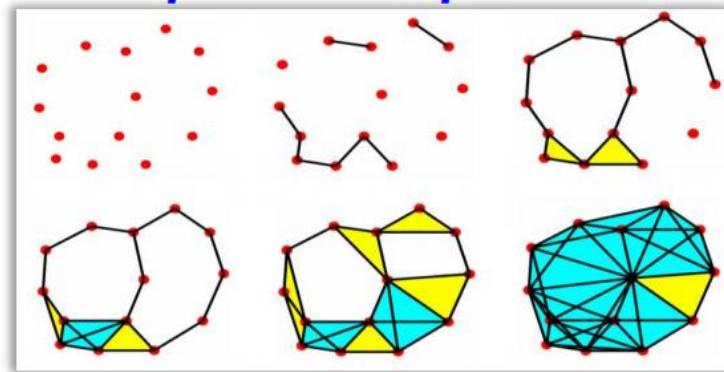
❑ **Graph**          ❑ **Simplicial complex**          ❑ **Multiscale simplicial complexes**



❖ **Graph models and measurements:**

*Graph Laplacian; Fiedler Eigenvalue; Fiedler eigenvector; Shortest path; Clique; Cluster coefficient; Closeness; Centrality; Betweenness; Modularity; Cheeger constant; Erdos number; Percolation…*

❖ **Simplicial complex models and measurements:**
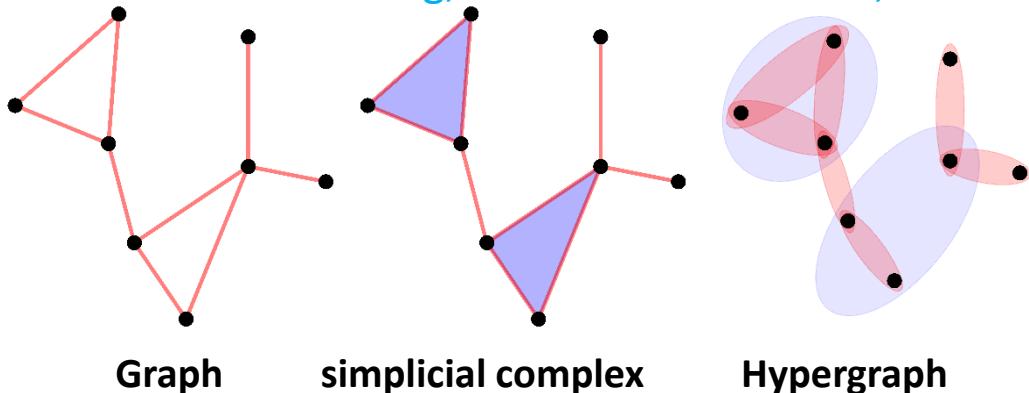
*Combinatorial Laplacian; Hodge theory; Betti number; Euler characteristics; Homology; Cohomology; Morse theory; Knot polynomials…*

❖ **Multiscale simplicial complex:**

*Persistent homology; Persistent cohomology…*

# Persistent Spectral (PerSpect)

Meng, Xia. Science Advance, 2021

## Spectral models

Spectral graph
Spectral simplicial complex
Spectral hypergraph

## Filtration

Nested sequence of graphs
Nested sequence of simplicial complex
Nested sequence of hypergraphs

## Spectral models + filtration

Persistent spectral graph
Persistent spectral simplicial complex
Persistent spectral hypergraph



Graph    simplicial complex    Hypergraph

*F. Chung, and S.T. Yau."A Harnack inequality for homogeneous graphs and subgraphs."Comm.Anal. Geom 2.4 (1994):627-640.*
*F. Chung,"Spectral graph theory".American Mathematical Societ, 1997*
*D. Spielman,"Spectral graph theory", Combinatorial scientific computing. No.18.Boca Raton, FL: CRC Press,2012.*
*D.Horak,and J.Jost,"Spectra of combinatorial Laplace operators on simplicial complexes".Advances in Mathematics,244,303-336, 2013*
*O. Parzanchevski, and R. Ron"Simplicial complexes: spectrum, homology and random walks."Random Structures &Algorithms 50.2 (2017): 225-261.*
*M. T.Schaub, et al."Random walks on simplicial complexes and the normalized Hodge 1-Laplacian." SIAM Review 62.2 (2020):353-391.*
*K. Q. Feng, "Spectra of hypergraphs and applications."Journal of number theory 60.1(1996):1-22.*
*J.Cooper, and D.Aaron."Spectra of uniform hypergraphs."Linear Algebra and its applications 436.9 (2012):3268-3292.*
*L.Q.Qi, and Z.Y.Luo."Tensor analysis: spectral theory and special tensors".Society for Industrial and Applied Mathematics, 2017*
*Wang R, Nguyen D D, Wei G W. Persistent spectral graph[J]. International journal for numerical methods in biomedical engineering, 2020, 36(9): e3376.*
*......*

# Combinatorial Laplacian

Let $K$ be a simplicial complex and $C_k(K)$ be a vector space over some field $\mathbb{F}$ whose basis is all $k$-simplices of $K$.

### Definition

The *dual* of $C_k(K)$, denoted by $C^k(K)$, is the set of all linear functionals on $C_k(K)$:

$$C^k(K) = \{\phi : C_k(K) \to \mathbb{F} : \phi \text{ is linear}\}.$$

**Note:** Both $C_k(K)$ and $C^k(K)$ have the same dimension = no. of $k$-simplices of $K$.

**❶** *Boundary map* $\partial_k : C_k(K) \to C_{k-1}(K)$

$$\partial_k([u_0, u_1, \ldots, u_k]) = \sum_{i=0}^{k}(-1)^i[u_0, \ldots, u_{i-1}, u_{i+1}, \ldots, u_k],$$

where $[u_0, u_1, \ldots, u_k]$ is a basis element of $C_k(K)$.

**❷** *Coboundary map* $\delta_k : C^k(K) \to C^{k+1}(K)$

$$\delta_k(\phi)(\sigma^{k+1}) = \sum_{i=0}^{k+1}(-1)^i\phi([u_0, \ldots, u_{i-1}, u_{i+1}, \ldots, u_{k+1}]),$$

where $\phi \in C^k(K)$ and $\sigma^{k+1} = [u_0, \ldots, u_{k+1}]$ is a basis element of $C_{k+1}(K)$.

## Combinatorial Laplacian

Another important map that is crucial in the formulation of Hodge Decomposition Theorem is the combinatorial Laplacian:

### Definition

The $k$-dimensional *combinatorial Laplacian* is the linear operator $\Delta_k : C^k(K) \to C^k(K)$ is defined as follows:

$$\Delta_k = \begin{cases} \delta_k^* \circ \delta_k + \delta_{k-1} \circ \delta_{k-1}^* & \text{if } k \geq 1, \\ \delta_k^* \circ \delta_k & \text{if } k = 0. \end{cases}$$

$\delta_k^* : C^{k+1}(K) \to C^k(K)$ is the *adjoint/transpose map* of $\delta_k$ where

$$\langle \delta_k(f), g \rangle = \langle f, \delta_k^*(g) \rangle$$

for every $f \in C^k(K)$, $g \in C^{k+1}(K)$ and a suitable inner product $\langle , \rangle$ for $C^k(K)$ and $C^{k+1}(K)$.

**Boundary matrix**
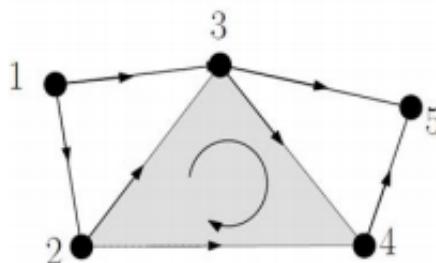
**face**

**same orientation**

$$B_k(i,j) = \begin{cases} 1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \text{ and } \sigma_i^{k-1} \sim \sigma_j^k \\ -1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \text{ and } \sigma_i^{k-1} \not\sim \sigma_j^k \\ 0, & \text{if } \sigma_i^{k-1} \not\subset \sigma_j^k. \end{cases}$$

**Combinatorial Laplacian (Hodge Laplacian)**

$$\mathbf{L}_k = \mathbf{B}_k^T \mathbf{B}_k + \mathbf{B}_{k+1} \mathbf{B}_{k+1}^T.$$

**upper degree (number of cofaces)**

**lower degree (number of faces)**

**lower adjacent**

**For simplicial complexes, if they are upper adjacent, they must be lower adjacent!**

$$L_k(i,j) = \begin{cases} d(\sigma_i^k) + k + 1, & \text{if } i = j \\ 1, & \text{if } i \neq j, \sigma_i^k \not\frown \sigma_j^k, \sigma_i^k \smile \sigma_j^k \text{ and } \sigma_i^k \sim \sigma_j^k \\ -1, & \text{if } i \neq j, \sigma_i^k \not\frown \sigma_j^k, \sigma_i^k \smile \sigma_j^k \text{ and } \sigma_i^k \not\sim \sigma_j^k \\ 0, & \text{if } i \neq j, \sigma_i^k \frown \sigma_j^k \text{ or } \sigma_i^k \not\smile \sigma_j^k. \end{cases}$$

**k>0**

**upper adjacent**



$$\mathbf{B}_1 = \begin{array}{c|ccccccc|c} & [12] & [13] & [23] & [24] & [34] & [35] & [45] & \\ \hline & -1 & -1 & 0 & 0 & 0 & 0 & 0 & [1] \\ & 1 & 0 & -1 & -1 & 0 & 0 & 0 & [2] \\ & 0 & 1 & 1 & 0 & -1 & -1 & 0 & [3] \\ & 0 & 0 & 0 & 1 & 1 & 0 & -1 & [4] \\ & 0 & 0 & 0 & 0 & 0 & 1 & 1 & [5] \end{array}$$

$$\mathbf{B}_2 = \begin{array}{c|c} & [2,3,4] \\ \hline [1,2] & 0 \\ [1,3] & 0 \\ [2,3] & 1 \\ [2,4] & -1 \\ [3,4] & 1 \\ [3,5] & 0 \\ [4,5] & 0 \end{array}$$

The Laplacians are computed as

$$\mathcal{L}_0 = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{pmatrix}, \quad \mathcal{L}_1 = \begin{array}{ccccccc} [12] & [13] & [23] & [24] & [34] & [35] & [45] \\ \begin{pmatrix} 2 & 1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & -1 & -1 & 0 \\ -1 & 1 & 3 & 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 3 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 3 & 1 & -1 \\ 0 & -1 & -1 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & -1 & -1 & 1 & 2 \end{pmatrix} \end{array}, \quad \mathcal{L}_2 = 3.$$
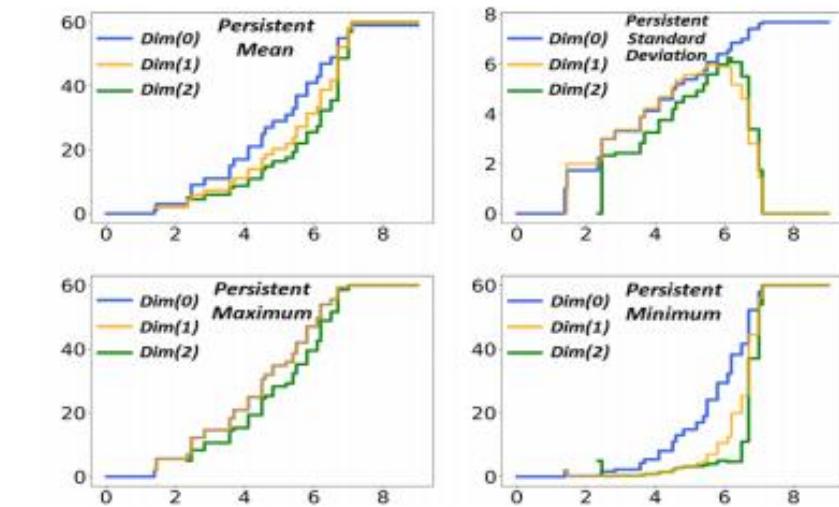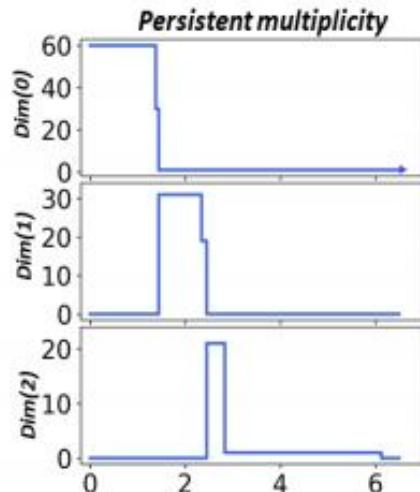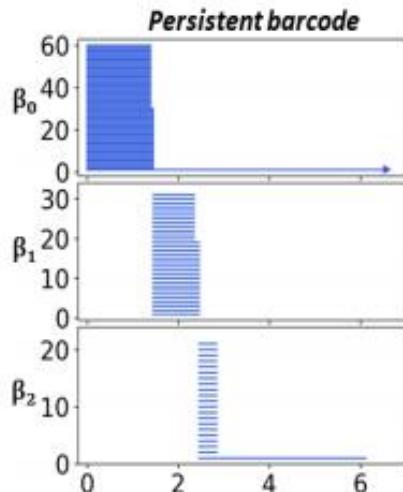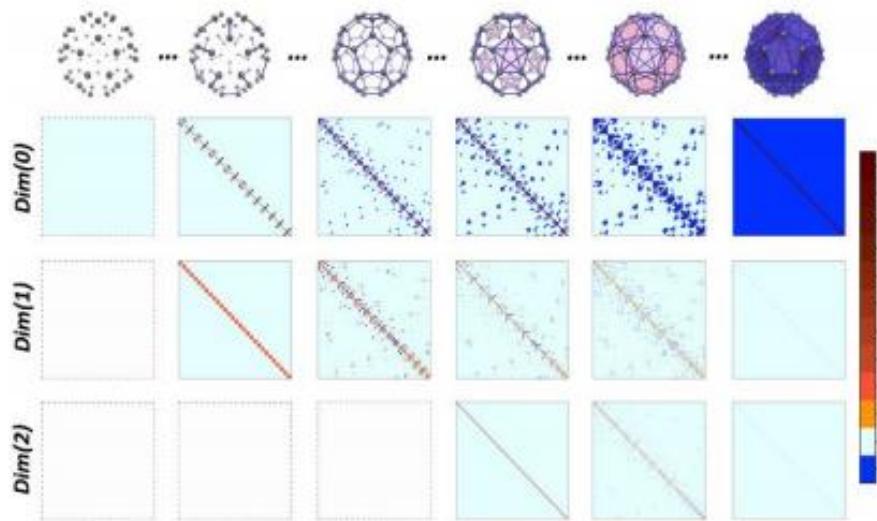
# Persistent spectral simplicial complex

Meng, Xia. Science Advance, 2021

## Combinatorial Laplacian (Hodge Laplacian)
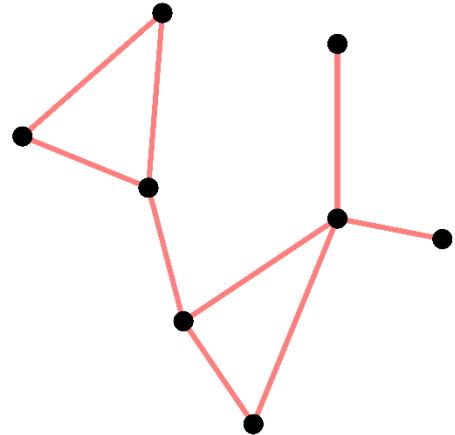
$$\mathbf{L}_k = \mathbf{B}_k^T \mathbf{B}_k + \mathbf{B}_{k+1}\mathbf{B}_{k+1}^T.$$
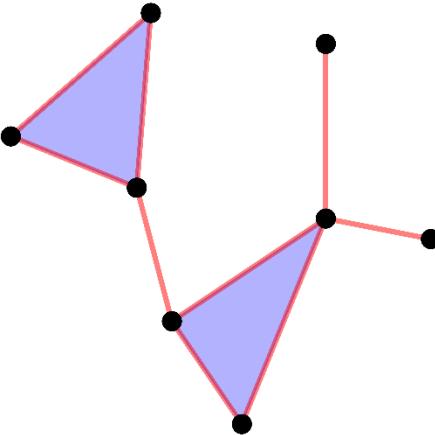




**Multiplicity of zero eigenvalues (Persistent multiplicity) from PerSpect simplicial complex is equivalent to persistent Betti number.**

**PerSpect variables change with filtration parameter and incorporate in them related geometric information.**
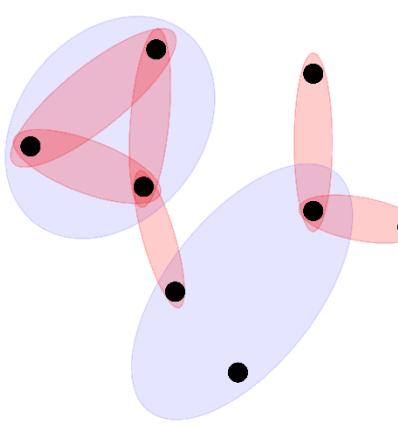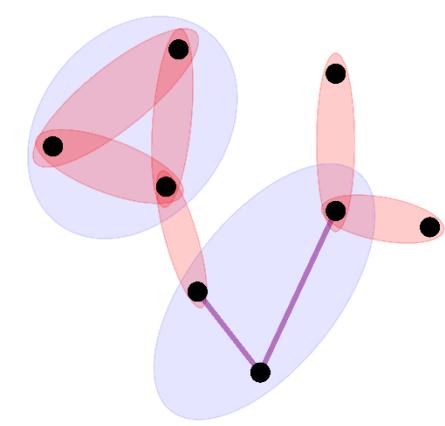
**Graph**     **Simplicial complex**     **Hypergraph**     **Super-hypergraph**

Jie Wu,
BIMSA

# Hypergraph based data representation
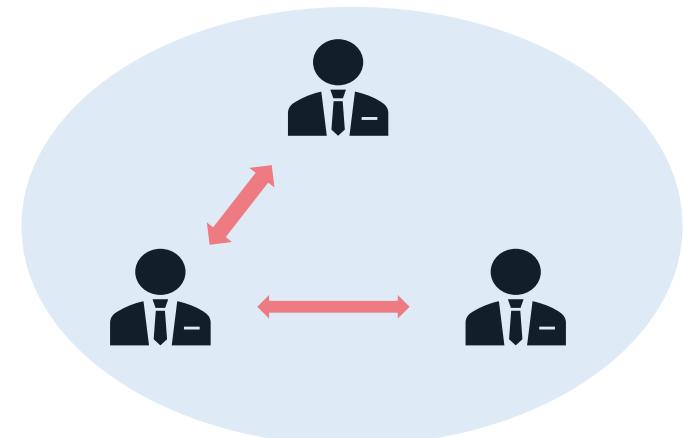
Grbic J, Wu J, Xia K, Wei GW. Aspects of topological approaches for data science[J]. Foundations of Data Science, 2022.

Bressan, Li, Ren, Wu. The embedded homology of hypergraphs and applications , 2016

Ren, Shiquan, et al. "Computing the Homology of Hypergraphs." *arXiv preprint arXiv:1705.00151* (2017).

Ren, Shiquan, Chengyuan Wu, and Jie Wu. "Operators on random hypergraphs and random simplicial complexes." *arXiv preprint arXiv:1712.02045* (2017).

Ren, Shiquan, and Jie Wu. "Stability of persistent homology for hypergraphs." *arXiv preprint arXiv:2002.02237* (2020).

Ren, Shiquan, et al. "A Discrete Morse Theory for Hypergraphs." *arXiv preprint arXiv:1804.07132* (2018).

# Embedded homology of hypergraph

**Definition (infimum chain complex)**

Given a hypergraph $\mathcal{H}$, the infimum chain complex of $\mathcal{H}$ with coefficient $R$ is defined as

$$Inf_n(\mathcal{H}, R) = \sum \{C_n \mid C_\star \text{ is a subchain complex of } R((K_\mathcal{H})_\star) \text{ and } C_n \subset R(\mathcal{H}_n)\}$$

which is the largest subchain complex of the chain complex of $K_\mathcal{H}$ that is contained in the graded modules $R(\mathcal{H}_\star)$
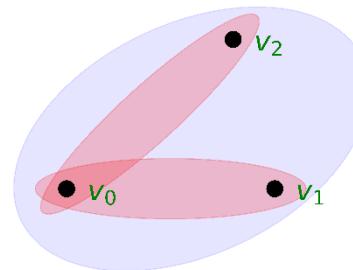
**Definition (supremum chain complex)**

Given a hypergraph $\mathcal{H}$, the supremum chain complex of $\mathcal{H}$ with coefficient $R$ is defined as
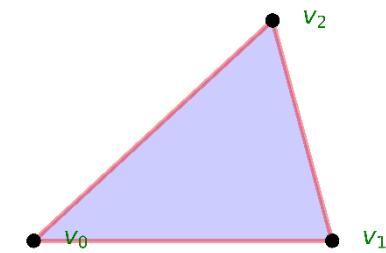
$$Sup_n(\mathcal{H}, R) = \bigcap \{C_n \mid C_\star \text{ is a subchain complex of } R((K_\mathcal{H})_\star) \text{ and } R(\mathcal{H}_n) \subset C_n\}$$

which is the smallest subchain complex of the chain complex of $K_\mathcal{H}$ that contains $R(\mathcal{H}_\star)$ as a graded modules.



**Hypergraph H**

**Associated simplicial complex $K_H$**

## Proposition

Given a hypergraph $\mathcal{H}$, the homology of the infimum chain complex of and supremum chain complex of $\mathcal{H}$ with coefficient $R$ are isomorphic.

## Definition (Hypergraph embedded homology)

Given a hypergraph $\mathcal{H}$, the $n$-th embedded homology of $\mathcal{H}$ with coefficient $R$ is defined as

$$H_n(\mathcal{H}, R) = H_n(Sup_\star(\mathcal{H}, R)) = H_n(Inf_\star(\mathcal{H}, R))$$

$$C_0 = Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\}$$
$$C_1 = Z\{\{0,1\},\{2,3\},\{2,4\},\{3,4\}\}$$
$$C_2 = Z\{\{0,1,2\}\}$$
$$A_0 = Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\}$$
$$A_1 = Z\{\{0,1\},\{0,2\},\{1,2\},\{2,3\},\{2,4\},\{3,4\}\}$$
$$A_2 = Z\{\{0,1,2\}\}$$
$$\to A_3 \xrightarrow{\partial_3} A_2 \xrightarrow{\partial_2} A_1 \xrightarrow{\partial_1} A_0$$
$$S_n = C_n + \partial_{n+1}(C_{n+1}), I_n = C_n \cap \partial_n^{-1}(C_{n-1})$$
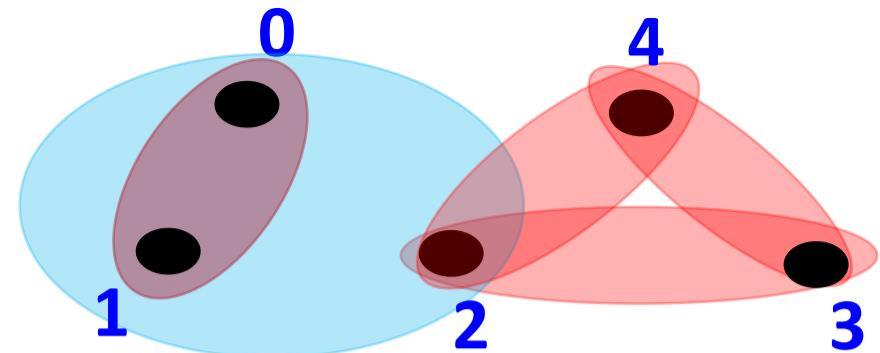


$$I_0 = Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\}$$
$$I_1 = Z\{\{0,1\},\{2,3\},\{2,4\},\{3,4\}\}$$
$$I_2 = 0$$
$$S_0 = Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\}$$
$$S_1 = Z\{\{0,1\},\{2,3\},\{2,4\},\{3,4\},\partial\{0,1,2\}\}$$
$$S_2 = Z\{\{0,1,2\}\}$$

$$H_0^s = Ker(\partial_0^s)/\operatorname{Im}(\partial_1^s)$$
$$= S_0 / \operatorname{Im}(\partial_1^s)$$
$$= Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\}/Z\{\{1\}-\{0\},\{3\}-\{2\},\{4\}-\{2\},\{4\}-\{3\}\}$$
$$= I_0 / \operatorname{Im}(\partial_1^i)$$
$$= Ker(\partial_0^i)/\operatorname{Im}(\partial_1^i)$$
$$= H_0^i$$

$$H_1^s = Ker(\partial_1^s)/\operatorname{Im}(\partial_2^s)$$
$$= Z\{\{3,4\}-\{2,4\}+\{2,3\},\partial\{0,1,2\}\}/Z\{\partial\{0,1,2\}\}$$
$$= Z\{\{3,4\}-\{2,4\}+\{2,3\}\}$$
$$= Ker(\partial_1^i)/\operatorname{Im}(\partial_2^i)$$
$$= H_1^i$$

$$H_2^s = Ker(\partial_2^s)/\operatorname{Im}(\partial_3^s)$$
$$= Ker(\partial_2^s)$$
$$= 0$$
$$= Ker(\partial_2^i)/\operatorname{Im}(\partial_3^i)$$
$$= H_2^i$$

Liu, Wang, Wu, Xia, BIB, 2021
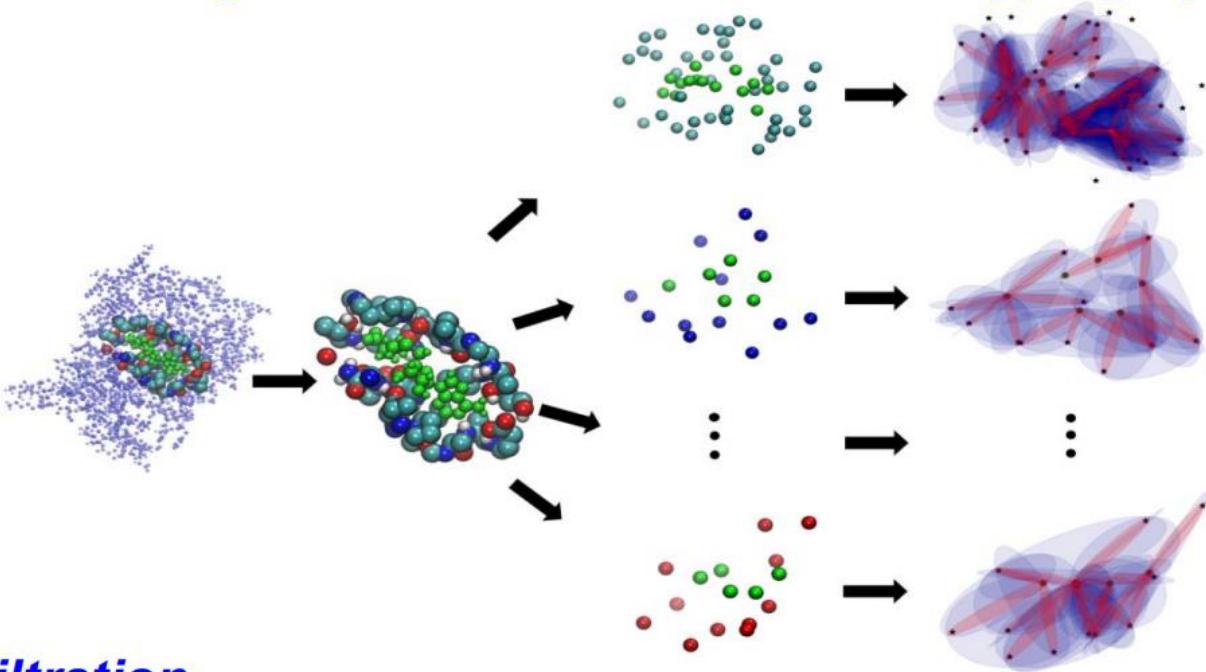
# Hypergraph-based models

**Protein-ligand interaction modeled as hypergraph**

**Hypergraph-based filtration**

filtration=3.5
filtration=3.7
filtration=3.8
filtration=3.9
filtration=4.1
filtration=4.2
filtration=4.3
filtration=4.4
filtration=4.5
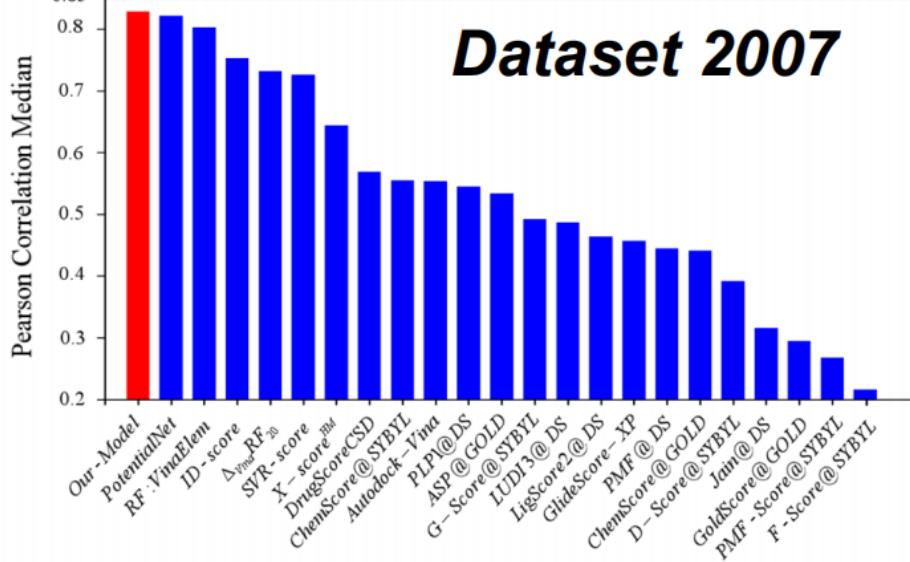
**Bipartite graph VS Hypergraph**

(a) (b) (c)

0 Dim 1 Dim 2 Dim

(d) (e) (f)
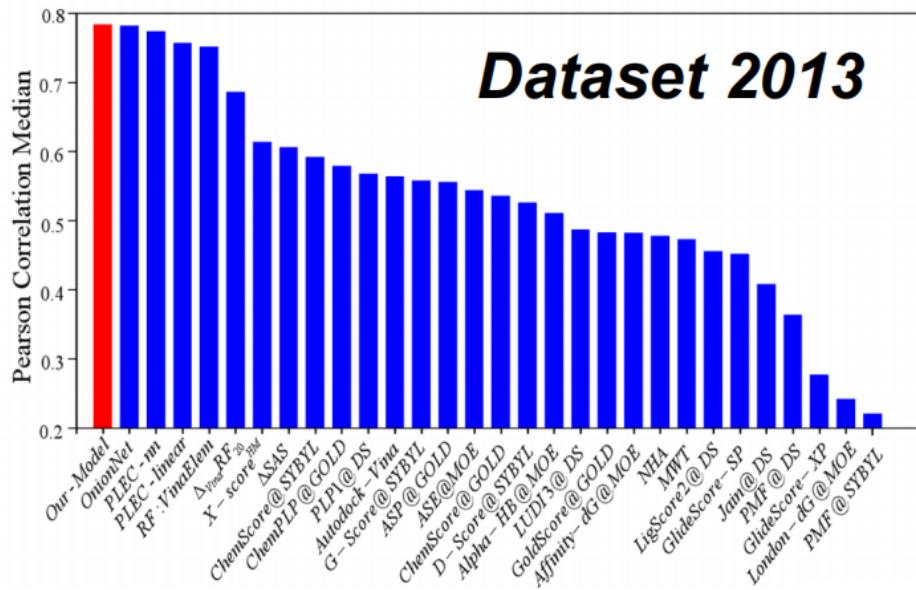
**Benchmark testing with PDBbind datasets**
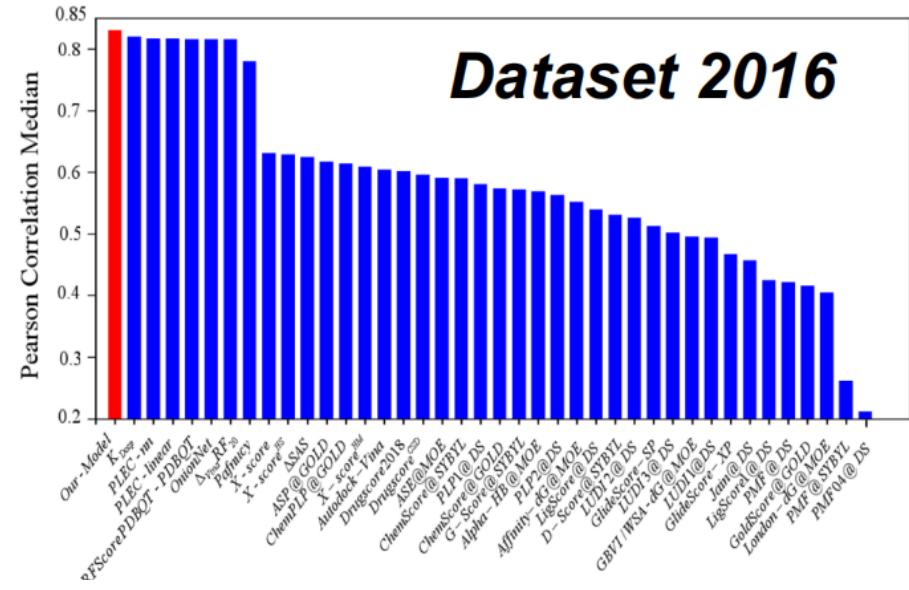
*Model setting:*
homology vectors
+
Gradientboostingtree

# Dowker Complex based molecular representation
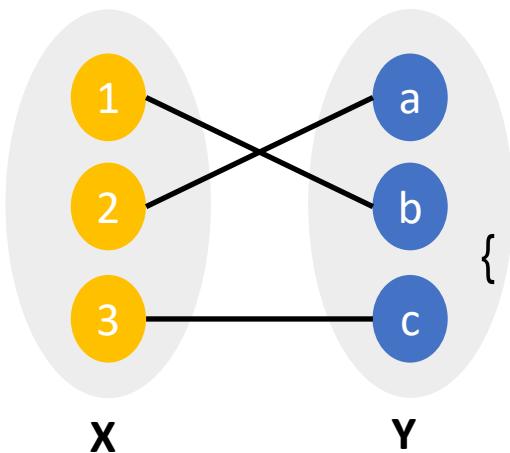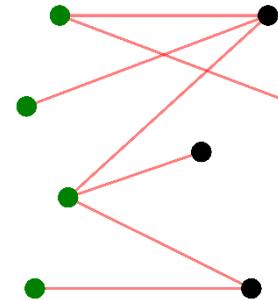
**Relation:**
{ (1,b), (2,a), (3,c) }

X          Y

**Bipartite graph**        **Dowker complex**



DNA          Bipartite graph          Dowker complex          DC (chain A)          DC (chain B)

Persistence barcode (chain A)

Persistence barcode (chain B)

**Dowker complex for interactions of two chains of a DNA**
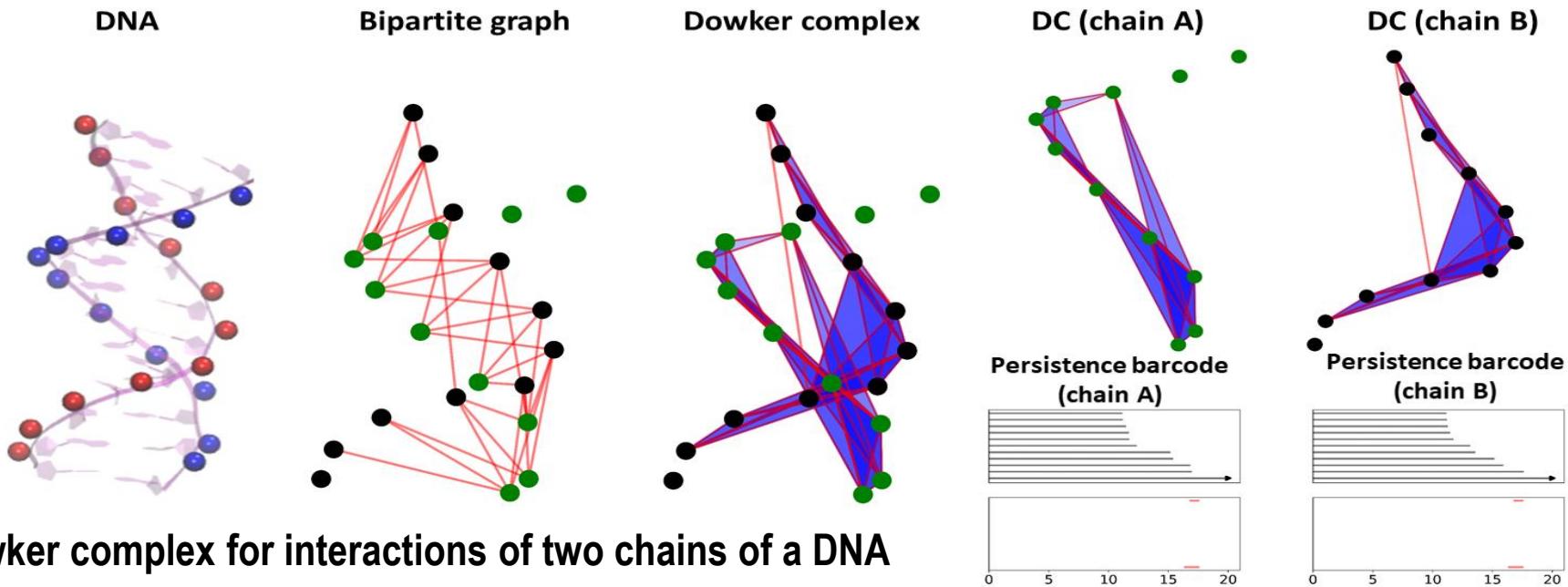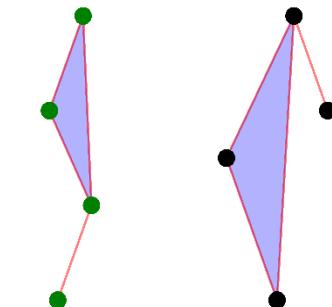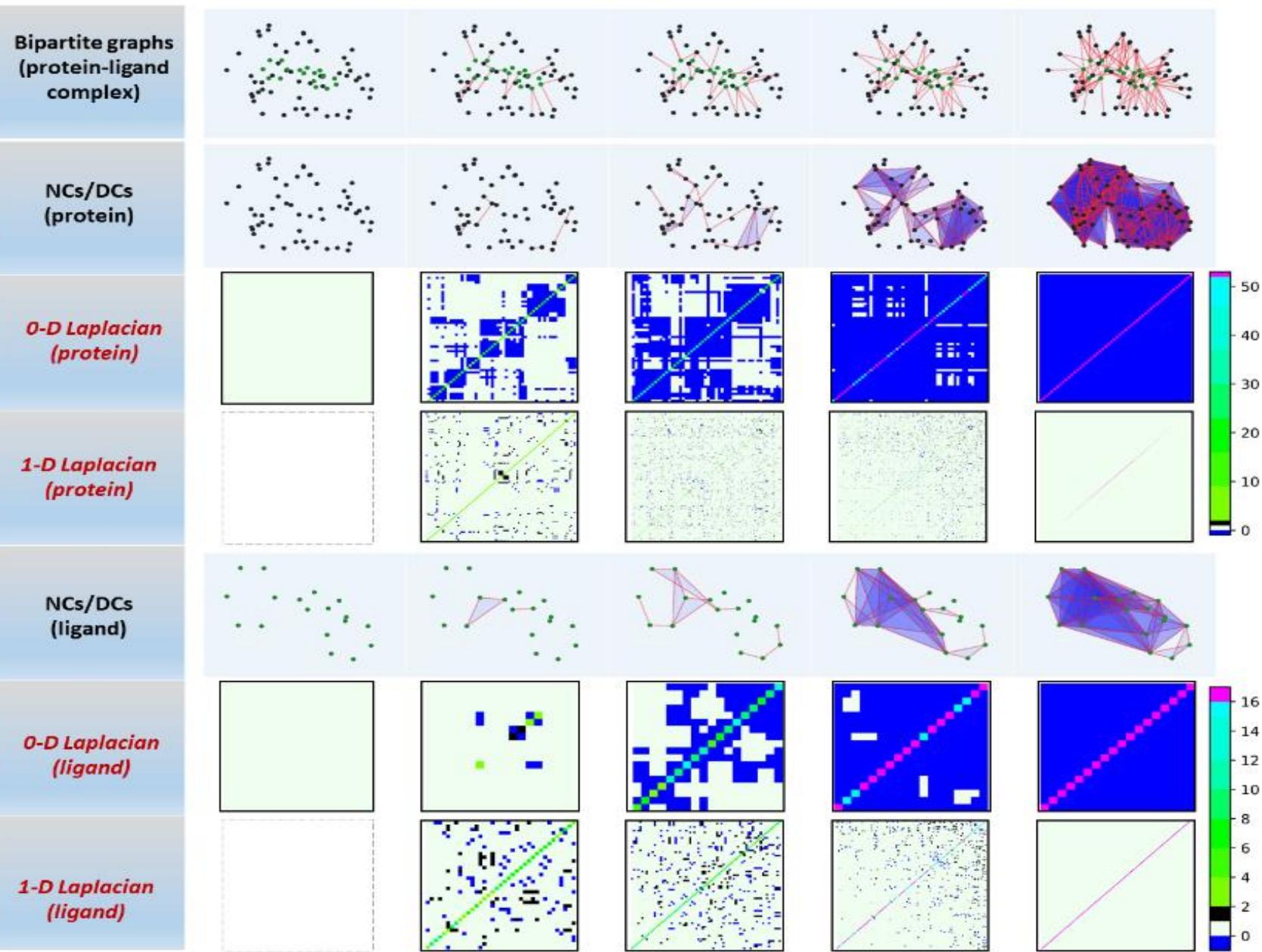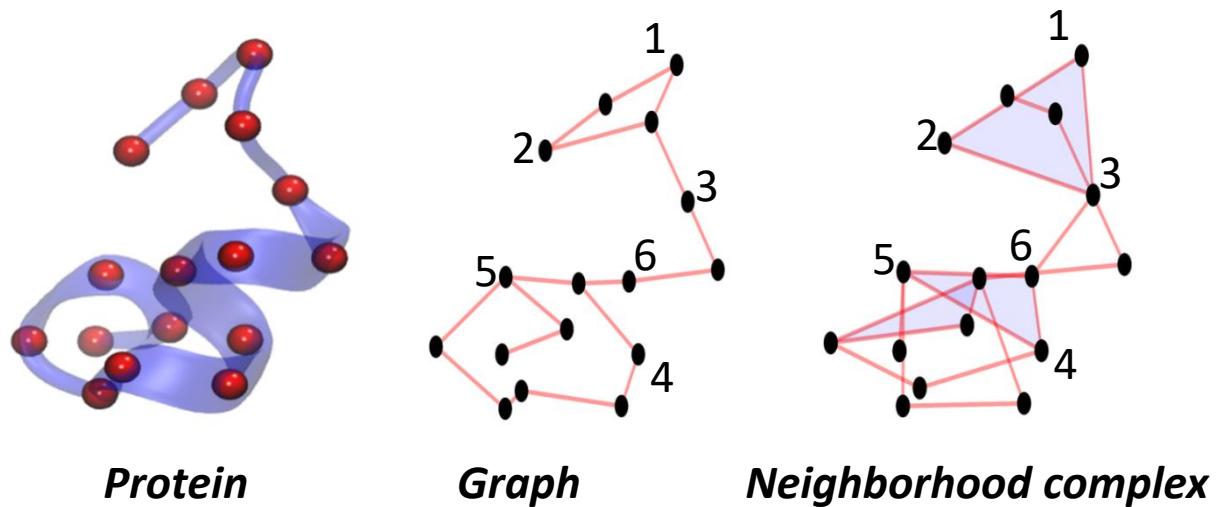
# Dowker complex based persistent Laplacian
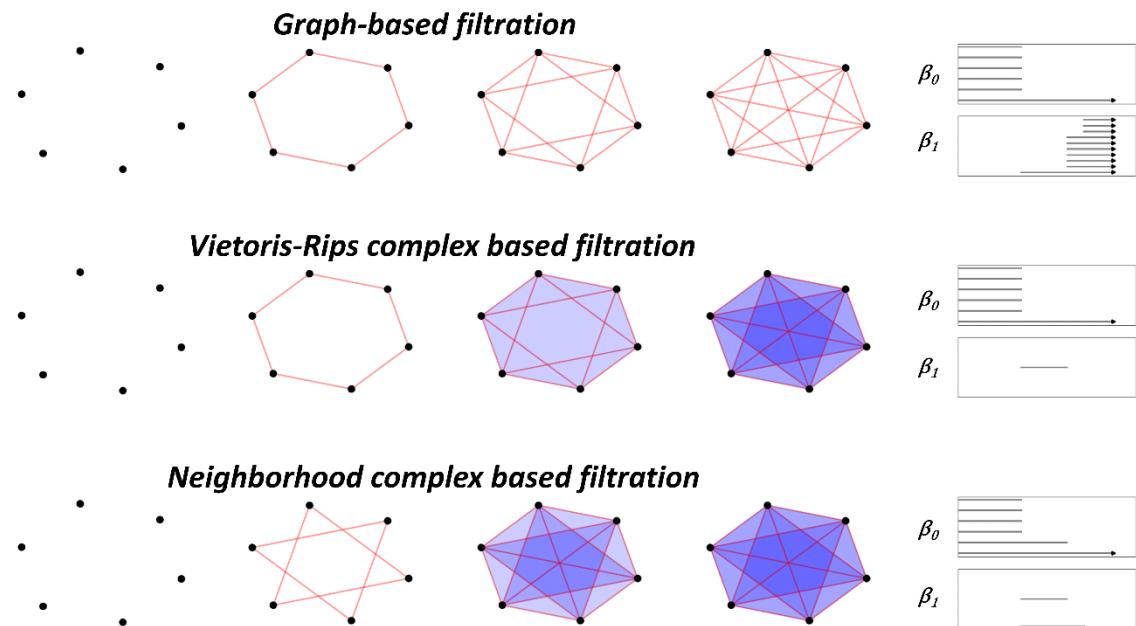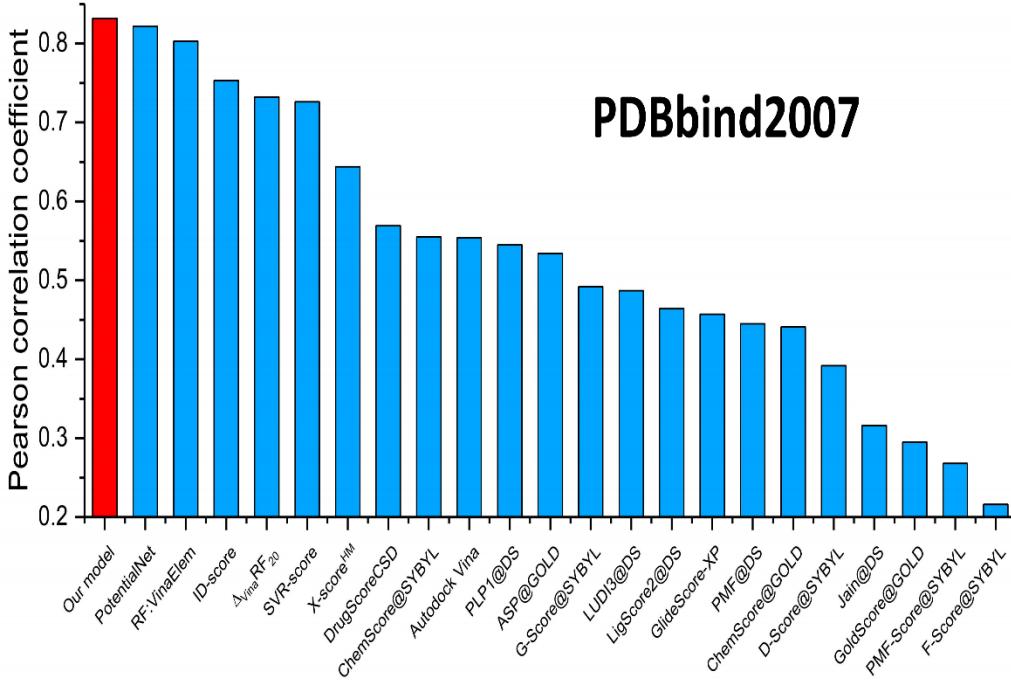
# Neighborhood complex based molecular representation



**László Lovász**

The neighborhood complexes of graphs were introduced by Lov´asz in his proof of the Kneser conjecture. He showed that a certain topological property of N(G) gives a lower bound for the chromatic number of G

**Protein**

**Graph**

**Neighborhood complex**

*Graph-based filtration*

$\beta_0$

$\beta_1$

*Vietoris-Rips complex based filtration*

$\beta_0$

$\beta_1$

*Neighborhood complex based filtration*

$\beta_0$

$\beta_1$
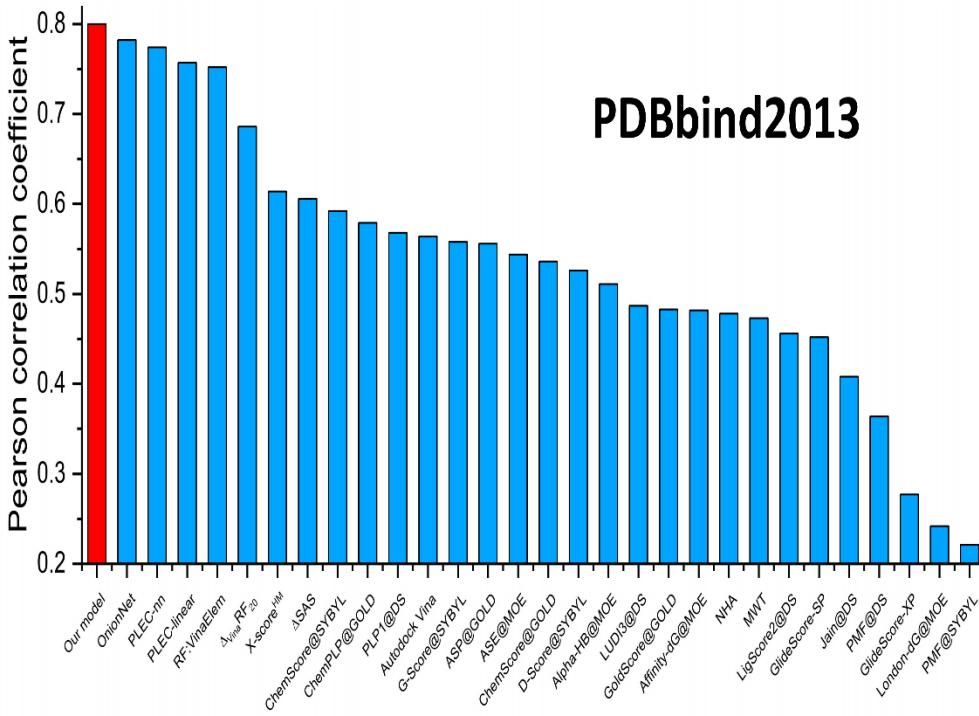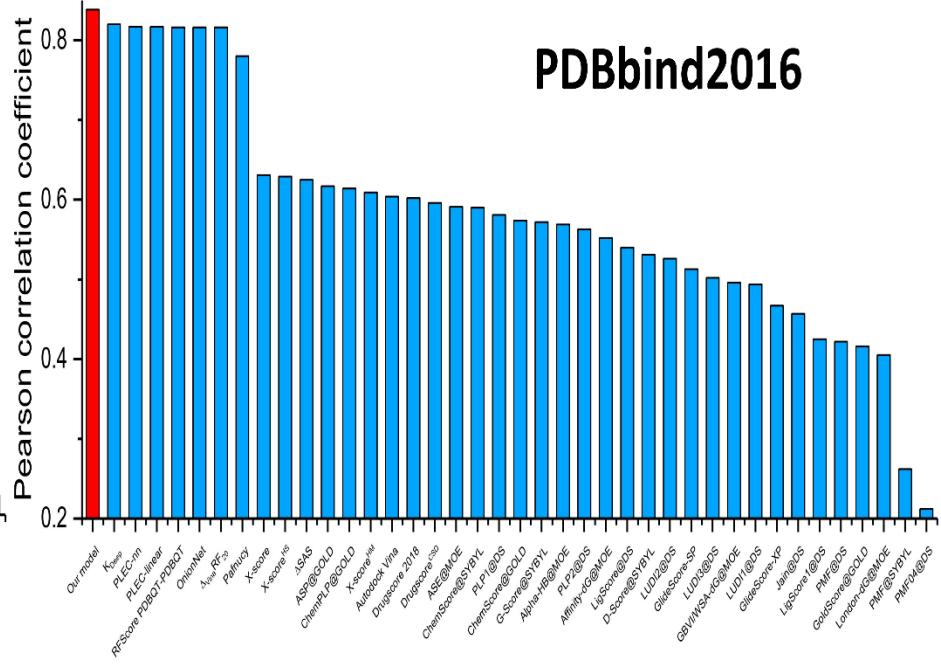
Liu, Xia. TDA4MedicalData 2021

*Neighborhood complex*
*Based*
*Molecular descriptors*
*+*
*Gradient boosting tree*

# Hom complex Hom(G1,G2)

$K_2$     $K_3$     $Hom(K_2, K_3)$

$x_1$   $x_2$   $a$   $c$   $b$

$(b, a)$

$(\{b, c\}, a)$    $(b, \{a, c\})$

$(c, a)$    $(b, c)$

$(c, \{a, b\})$    $(\{a, b\}, c)$

$(c, b)$    $(a, c)$

$(\{a, c\}, b)$    $(a, \{b, c\})$

$(a, b)$
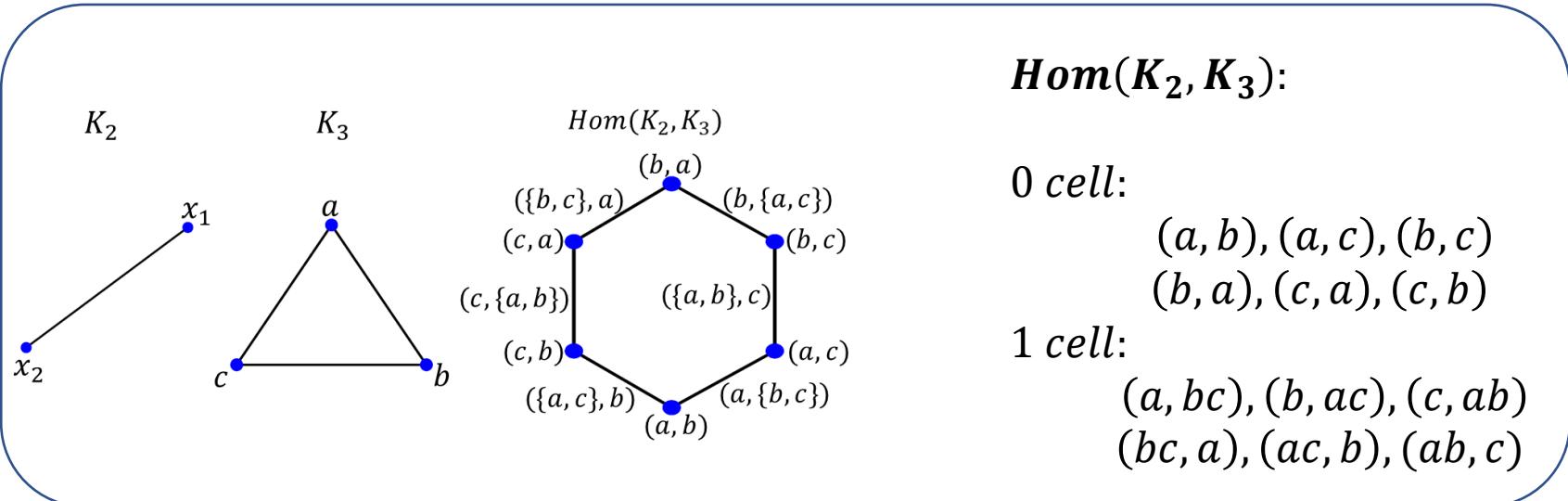
$Hom(K_2, K_3)$:
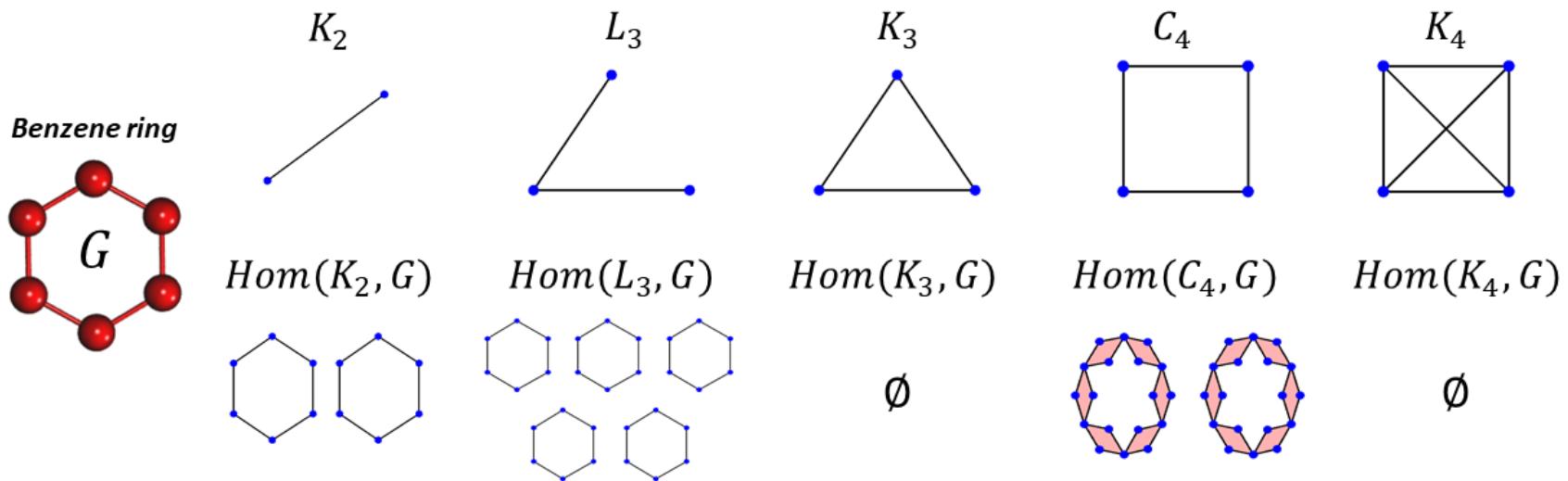
$0\ cell$:

$(a, b), (a, c), (b, c)$
$(b, a), (c, a), (c, b)$
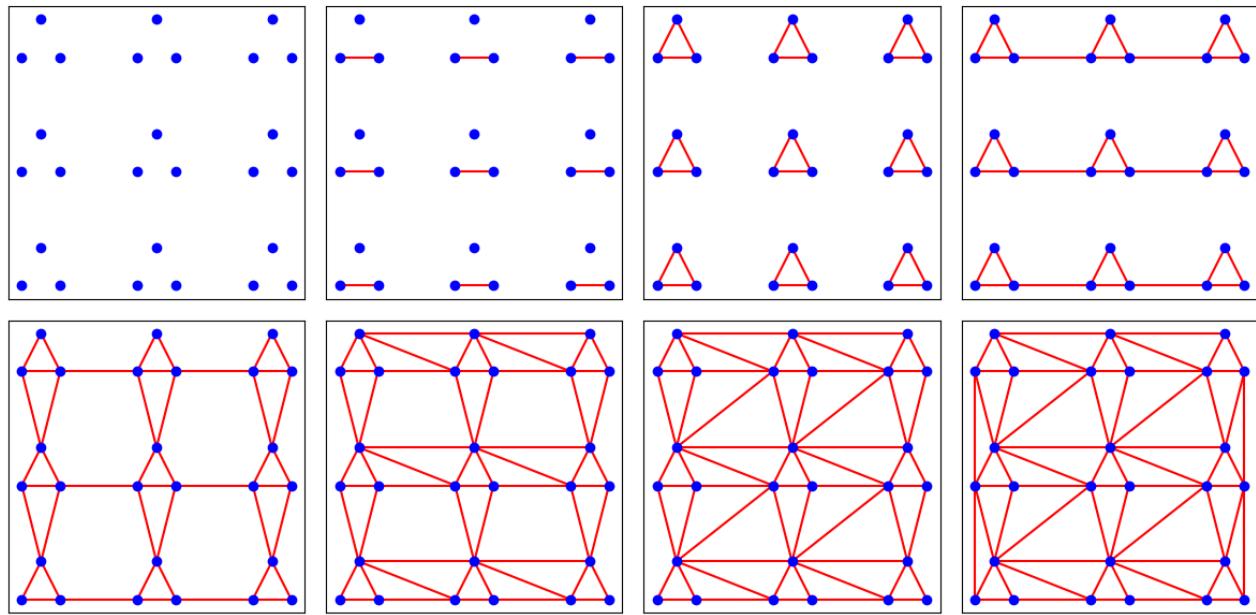
$1\ cell$:

$(a, bc), (b, ac), (c, ab)$
$(bc, a), (ac, b), (ab, c)$
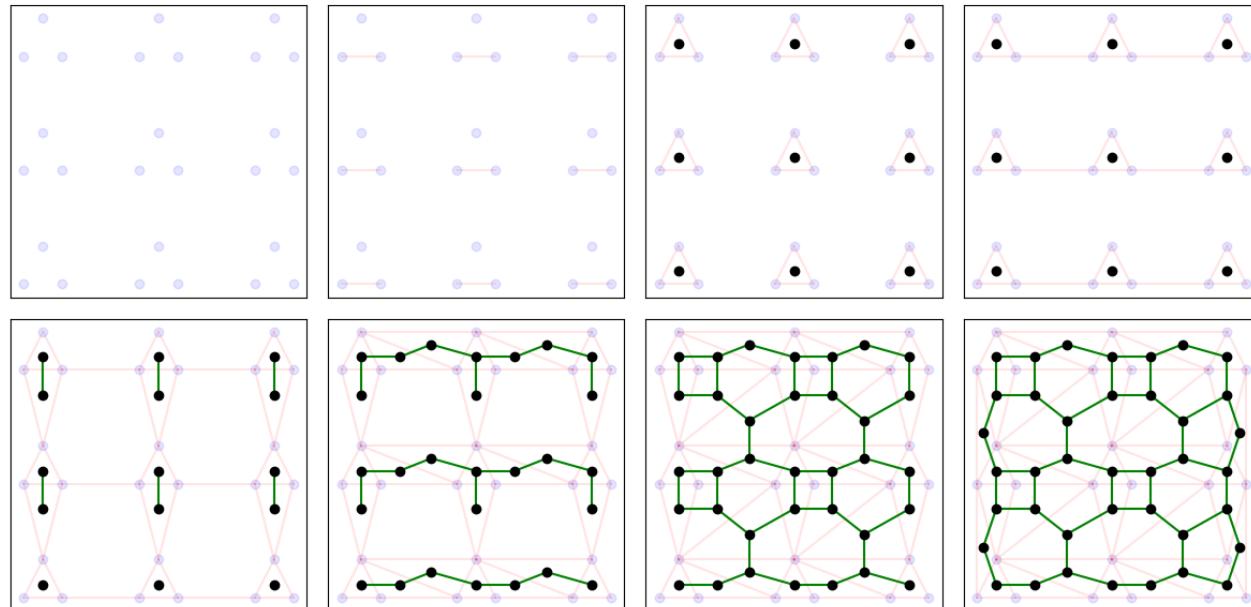
## *Hom-complex representation of Benzene ring from different graphs*



Benzene ring   $K_2$   $L_3$   $K_3$   $C_4$   $K_4$

$G$

$Hom(K_2, G)$   $Hom(L_3, G)$   $Hom(K_3, G)$   $Hom(C_4, G)$   $Hom(K_4, G)$

$\emptyset$     $\emptyset$

# Graph G filtration



# Hom(K₃, G) filtration



## Persistence barcode
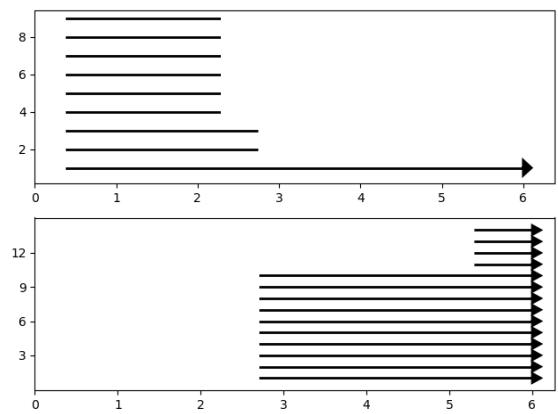
Table 1. Comparison of the performance between our model and other models on SKEMPI S1131 dataset.

| Method | PCC |
| --- | --- |
| **Hom-ML(2)** | **0.857** |
| TopNetTree | 0.850 |
| **Hom-ML(1)** | **0.792** |
| BindProfX | 0.738 |
| Profile-score+FoldX | 0.738 |
| Profile-score | 0.675 |
| SAAMBE | 0.624 |
| FoldX | 0.457 |
| BeAtMuSic | 0.272 |
| Dcomplex | 0.056 |

Table 2. Comparison of the performance between our model and other models on AB-Bind S645.

| Method | PCC |
| --- | --- |
| TopNetTree | 0.65(0.68) |
| **Hom-ML(2)** | **0.58(0.70)** |
| **Hom-ML(1)** | **0.58(0.68)** |
| mCSM-AB | 0.53(0.56) |
| Discovery Studio | 0.45 |
| mCSM-PPI | 0.35 |
| FoldX | 0.34 |
| STATIUM | 0.32 |
| DFIRE | 0.31 |
| bAsA | 0.22 |
| dDFIRE | 0.19 |
| Rosetta | 0.16 |

# Thank You!