Topological modelling and analysis for biomolecular data: 3 TDA-based machine learning

Kelin Xia

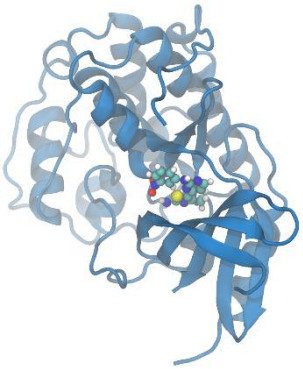
School of Physical and Mathematical Sciences, Nanyang Technological University

应用拓扑短期课程与报告会, 12月13-16日, 2020

Fund: NTU-JSPS(2019), MOE-Tier 1(2018,2019), MOE-Tier 2(2018,2021), Alibaba-NTU(2020), Merlion(2020)

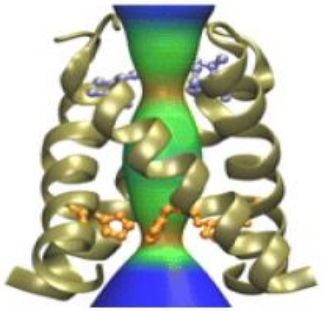
Drug design and discovery

1) Disease identification (physiology)
2) Target hypothesis (biochem./mole. biol.)
3) Virtual screening: drug pose, binding affinity, solubility, partition coefficient, toxicity, and side-effects (biophysics/bioinformatics)
4) Drug structural optimization in the target binding site (biochemistry/biophysics/synthetic chem.)
5) Preclinical *in vitro* and *in vivo* test
6) Clinical trials
7) Optimize drug's efficacy, pharmacokinetics, and pharmacodynamics properties (quantitative systems pharmacology)



Influenza -- flu virus

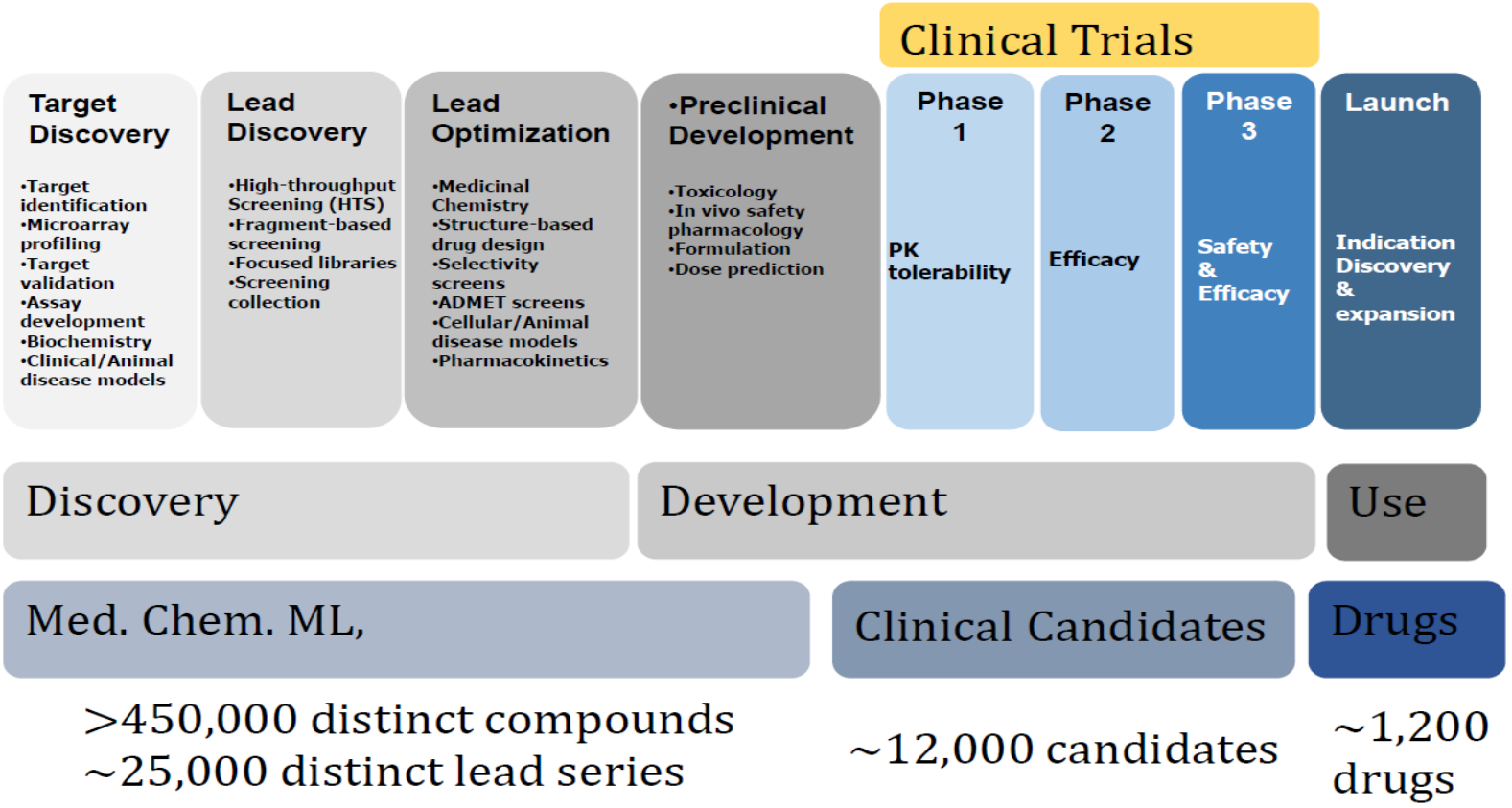
M2 channel



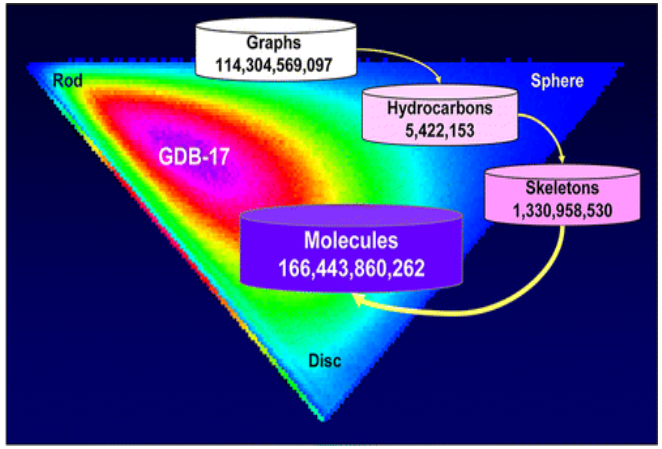
Amantadine

M2-A complex

Drug Discovery Process (simplified)

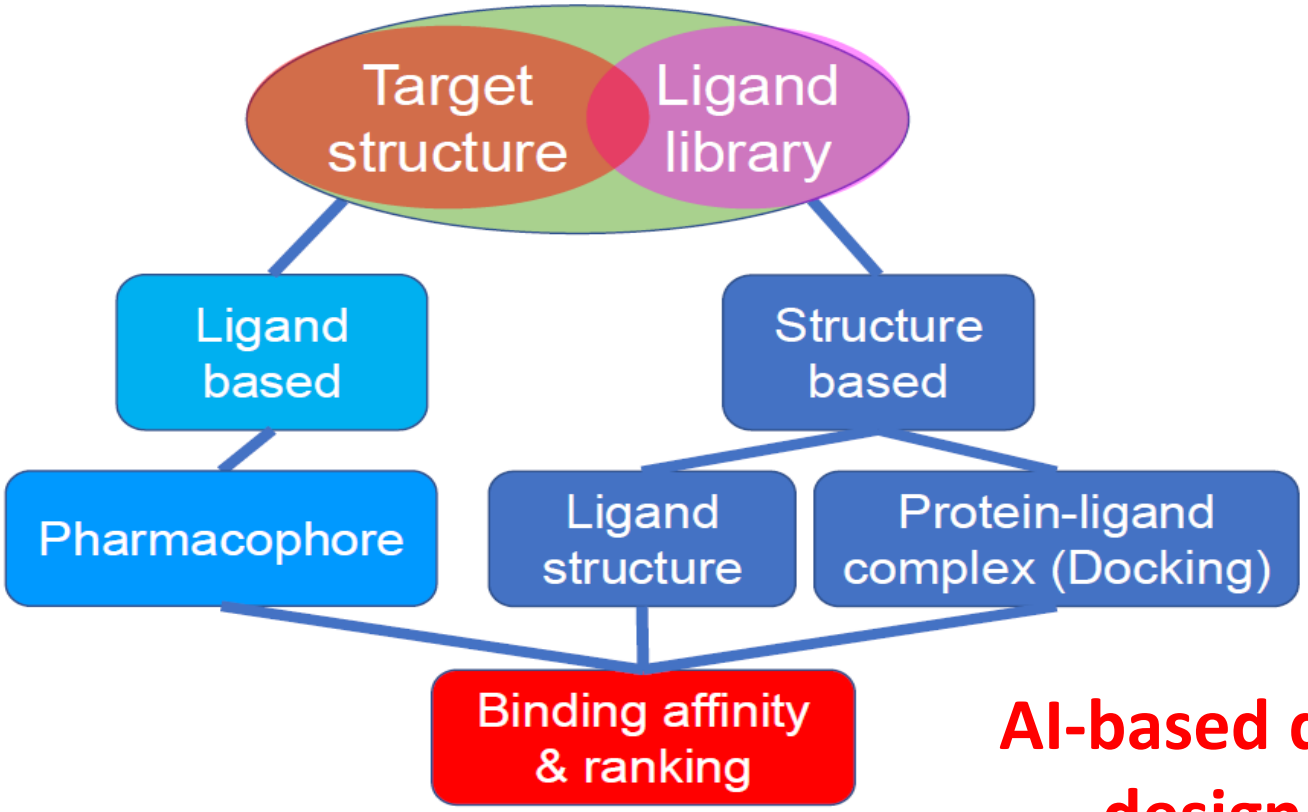


"...166.4 billion molecules of up to 17 atoms of C, N, O, S, and halogens forming the chemical universe database GDB-17"



Ruddigkeit, Lars, et al. *"Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17."* Journal of chemical information and modeling 52(11): 2864-2875, 2012.

Virtual Screening



AI-based drug design!

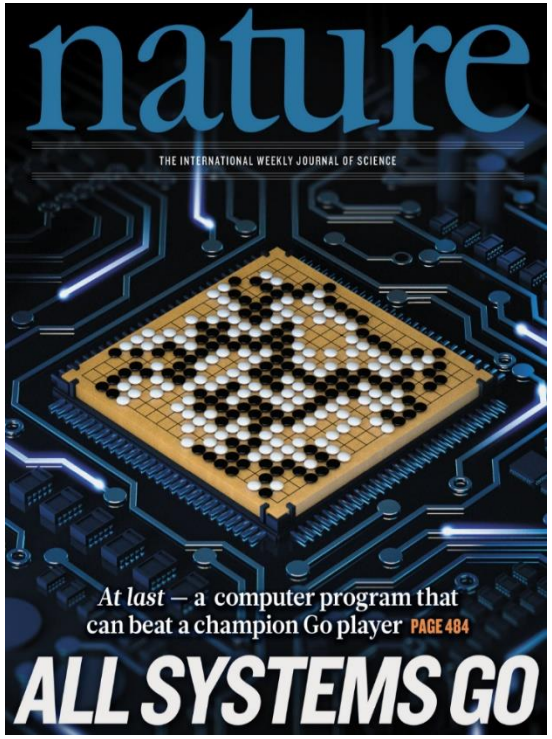
Force field | Empirical | Knowledge-based | Machine learning

P. J. Ballester and J. B. Mitchell, "A machine learning approach to predicting protein– ligand binding affinity with applications to molecular docking," Bioinformatics, 26 (9): 1169–1175, 2010

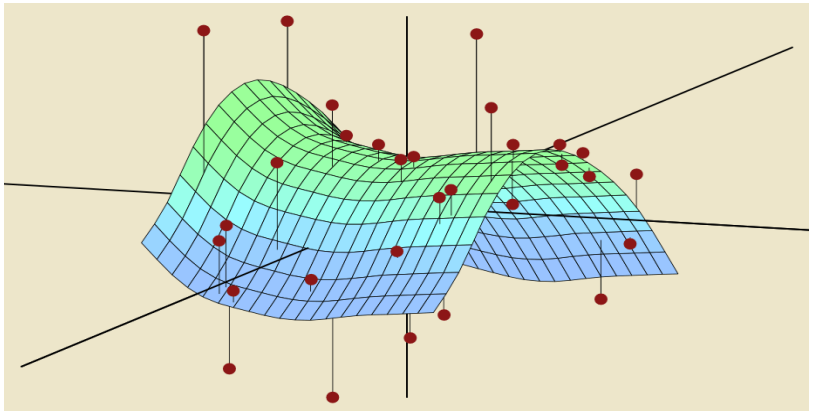
Q. U. Ain, A. Aleksandrova, F. D. Roessler, and P. J. Ballester, "Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening," Wiley Interdisciplinary Reviews: Computational Molecular Science, 5(6): 405– 424, 2015.

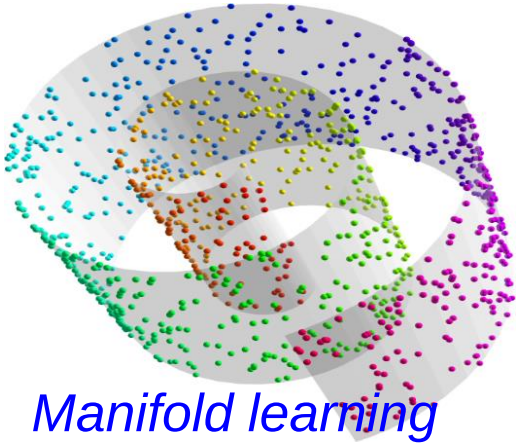
Y. C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, "Machine learning in chemoinformatics and drug discovery," Drug discovery today, 23 (8), 1538–1546, 2018.

Deep learning



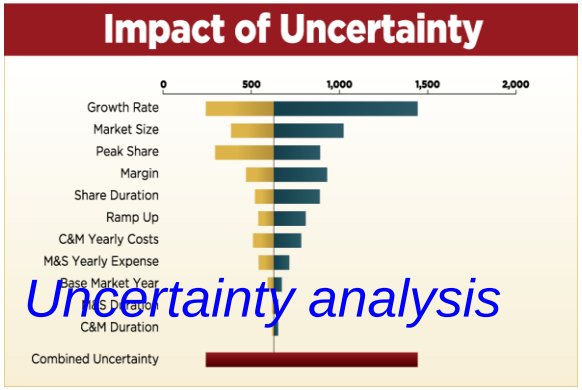
Statistic learning





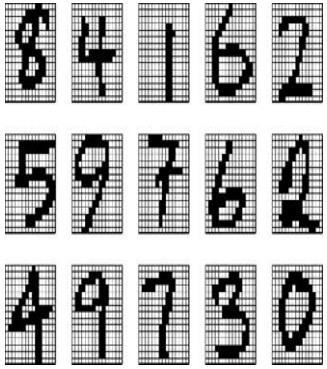
Manifold learning

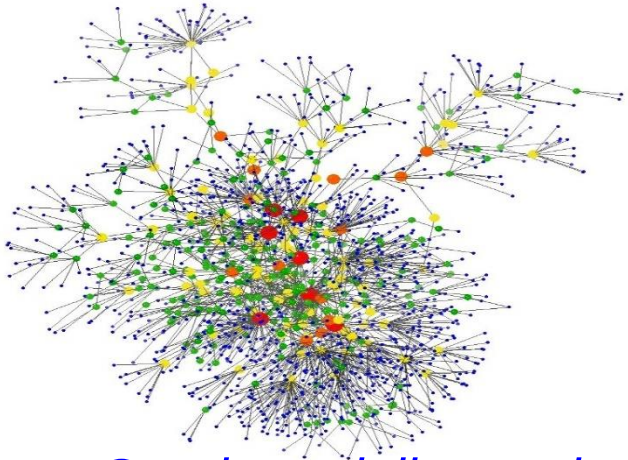
How to deal with big data?



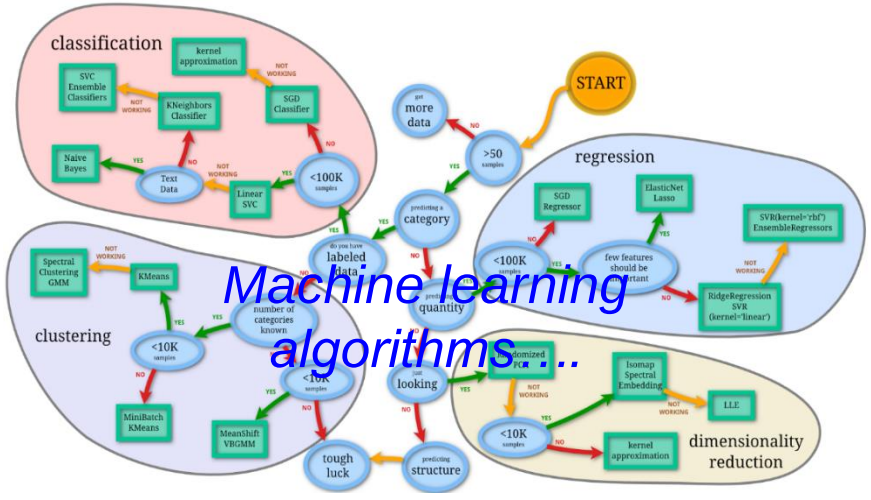
Uncertainty analysis

Pattern recognition





Graph modeling and analysis



Machine learning algorithms ...

Representation and Feature Learning

"The success of machine learning algorithms generally depends on data representation..."

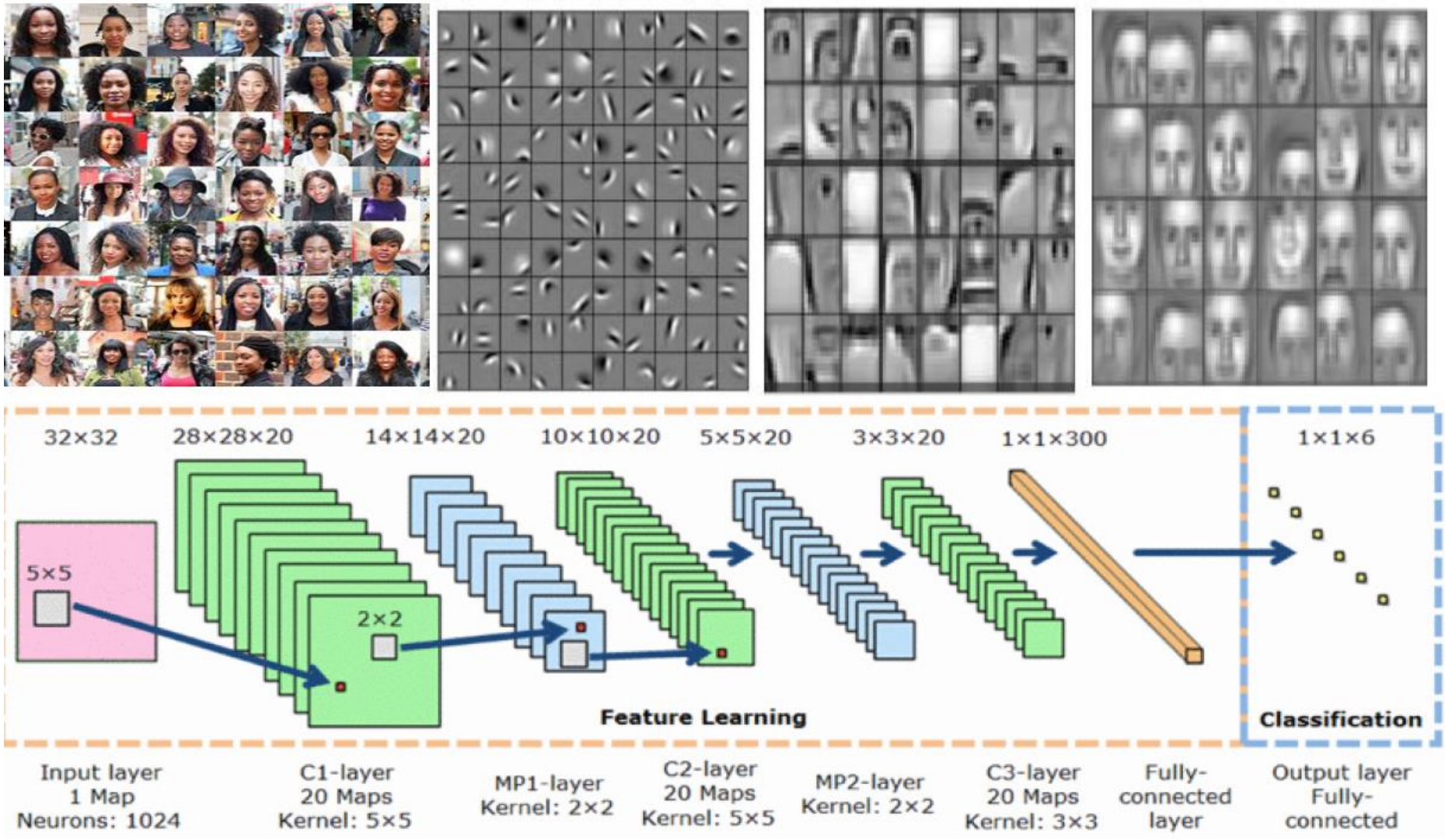
Y. Bengio, etc,"Representation Learning: A Review and New Perspectives"

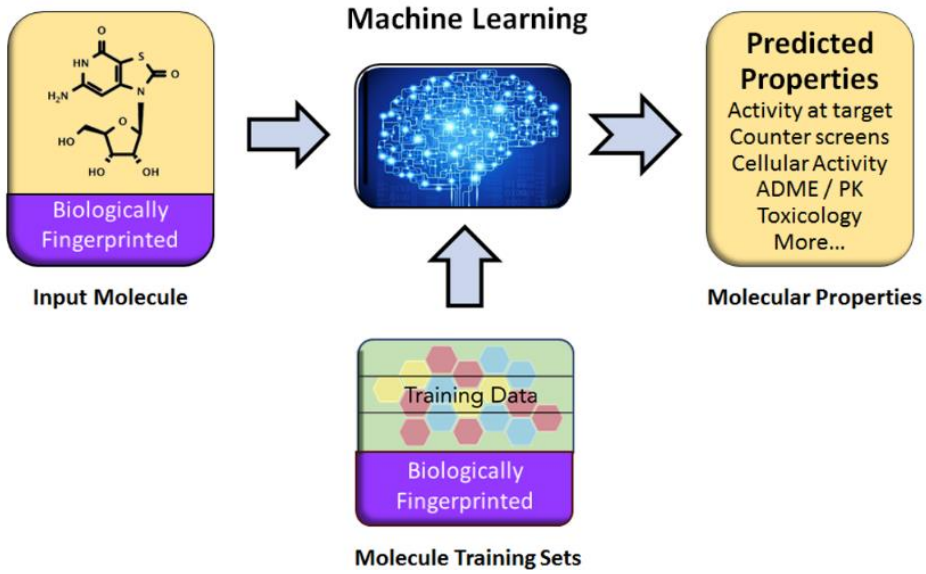
"The deep learning research aims at discovering learning algorithms that discover multiple levels of distributed representations..."

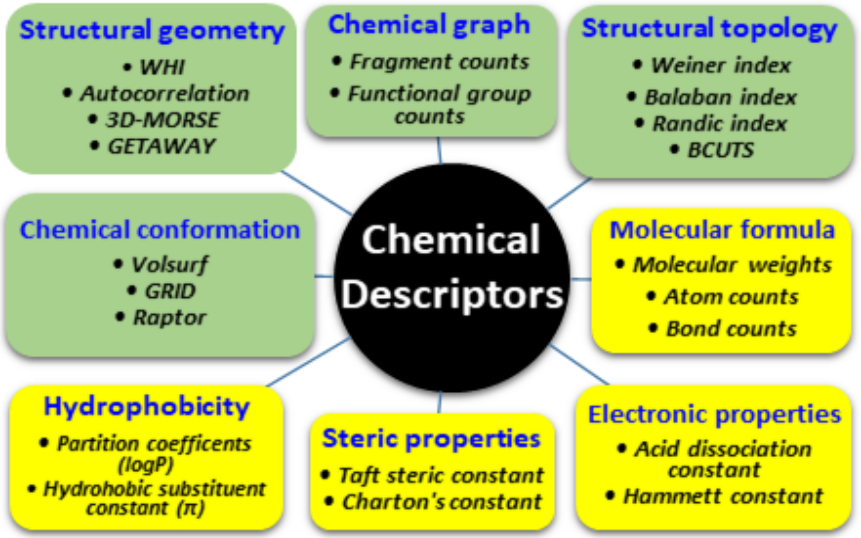
Y. Bengio, "Deep Learning of Representations: Looking Forward"

Feature learning is key to data analysis!

Fukushima (1980) – Neo-Cognitron; LeCun (1998) – Convolutional Neural Networks (CNN);…







Molecular (chemical) descriptors (>5000) directly determine the performance of learning models!

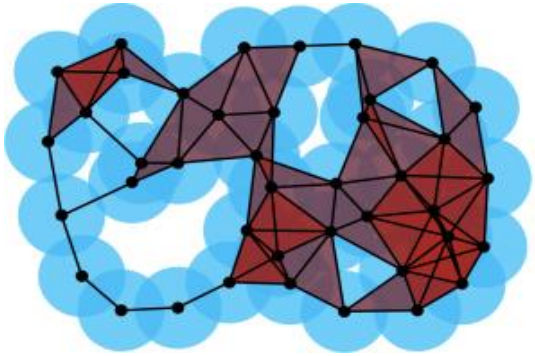
Common chemical descriptors for QSAR/QSPR analysis

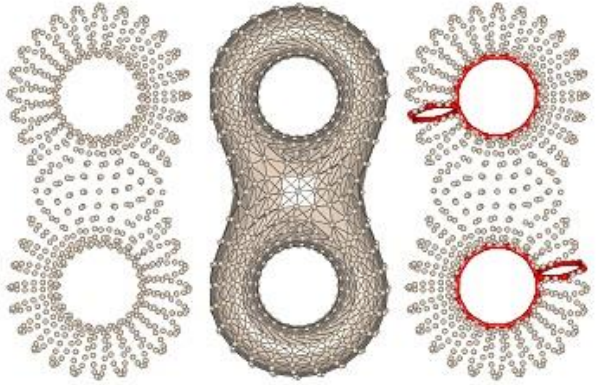
Chemical descriptors	Based on	Examples
Theoretical descriptors		
0D	Molecular formula	Molecular weights, atom counts, bond counts
1D	Chemical graph	Fragment counts, functional group counts
2D	Structural topology	Weiner index, Balaban index, Randic index, BCUTS
3D	Structural geometry	WHIM, autocorrelation,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

Fingerprint	Description	Number of features	Package
FP2	A path-based fingerprint which indexes small molecule fragments based on linear segments of up to 7 atoms[20]	256	Openbabel[20]
Daylight	A path-based fingerprint consisting 2048 bits and encoding all connectivity pathways in a given length through a molecule[21]	2048	RDKit[23]
MACCS	A substructure keys-based fingerprint with 166 structural keys based on SMARTS patterns[19]	166	
Estate1	A topological fingerprint based on electro-topological State Indices, which encodes the intrinsic electronic state of the atom as perturbed by the electronic influence of all other atoms in the molecule within the context of the topological character of the molecule. Estate 1 represents the number of times each atom type is hit[22]	79	
Estate2	Similar to estate 1, however it contains the sum of the EState indices for atoms of each type[22]	79	
ECFP4	The de facto standard circular fingerprint based on the Morgan algorithm,[42] which uses an iterative process to assign numeric identifiers to each atom[15]	2048	
Pharm2D	Each bit corresponds to a particular combination of features and interactions needed for a molecule to be active against a given target[23]	990	
ERG	A Pharmacophore fingerprint, which is an extended reduced graph approach using pharmacophore-type node descriptions to encode the relevant molecular properties[24]	315	

Topological Data Analysis (TDA)

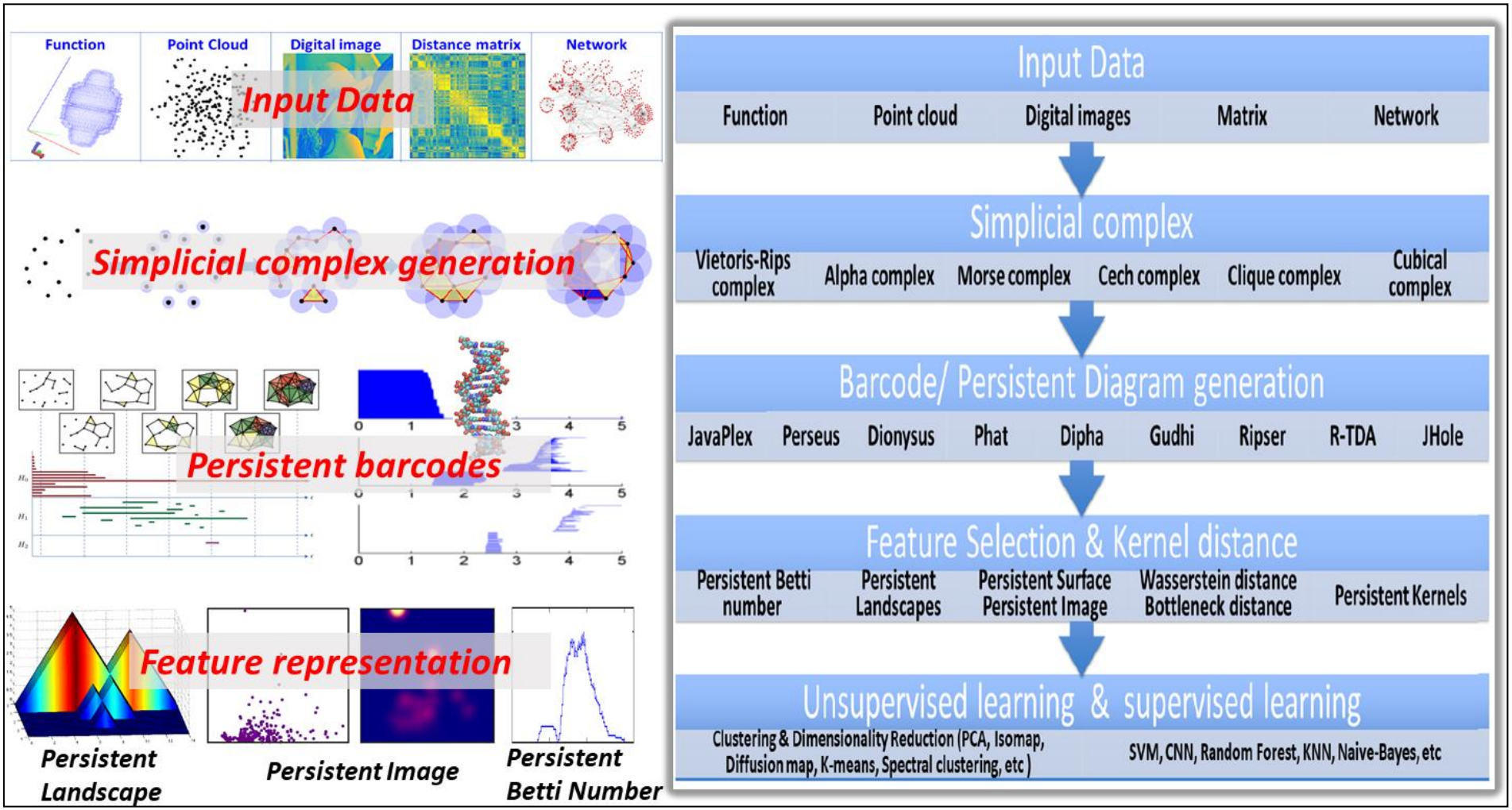
Topological invariant;
Homology;
Homotopy;
Simplicial complex;
Morse theory;
Reeb graph;
.....





Computational Geometry;
Computational topology;
Algebraic topology

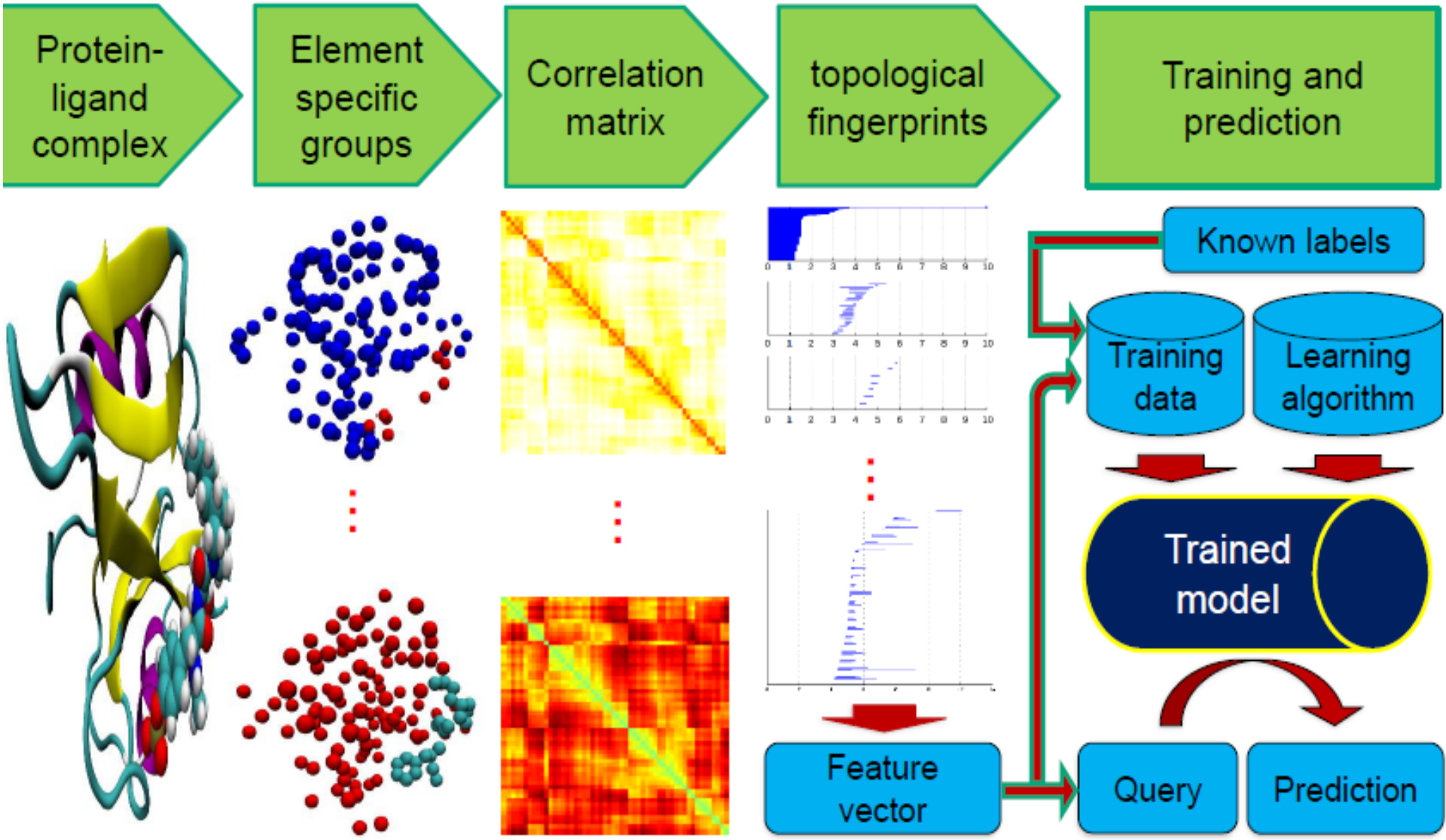
TDA based machine learning models



(Pun, Xia and Lee, arXiv, 2018)

Topology based learning architecture

(Cang & Wei, IJNMBE, 2017)



Recent progress in TDA based drug design

Collaborator
Guowei Wei
MSU, USA

MORE AT SIAM | siam news

HOME | HAPPENING NOW | GET INVOLVED | RESEARCH

Get Involved | September 01, 2016

Mathematical Molecular Bioscience and Biophysics

A Recurring Theme at the SIAM Conference on the Life Sciences

MORE AT SIAM | siam news

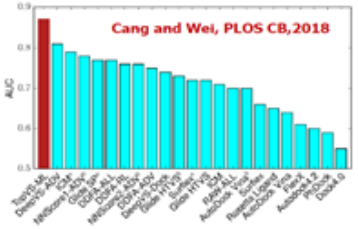
HOME | HAPPENING NOW | GET INVOLVED | RESEARCH

Research | December 01, 2017

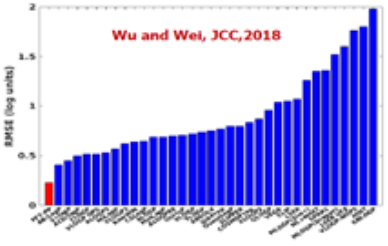
Persistent Homology Analysis of Biomolecular Data

Topological learning based predictions

Classification of ligands & decoys
DUD database 128,374 protein-ligand/decoy pairs



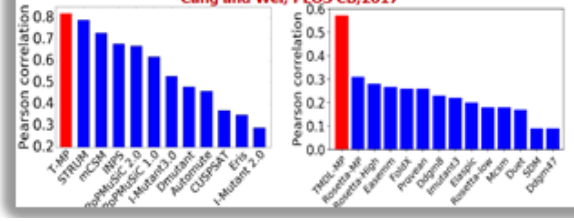
Prediction RMSD of LogP (Star set)



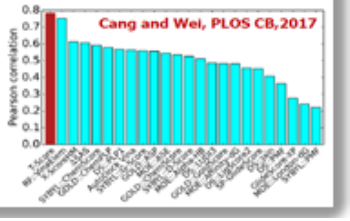
Prediction correlations for 2648 mutations on globular proteins

Prediction correlations for 223 mutations on membrane proteins

Cang and Wei, PLOS CB,2017



Binding affinity prediction of PDBBind v2013 core set of 195 complexes



D3R Grand Challenge 2 (2016-2017)

Given: Farnesoid X receptor (FXR) and 102 ligands
Tasks: Dock 102 ligands to FXR, and predict their poses, binding free energies and energy ranking

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)

(Nguyen et al, JCAMD, 2018)

D3R Grand Challenge 3 (2017-2018)

(Nguyen et al, JCAMD, 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

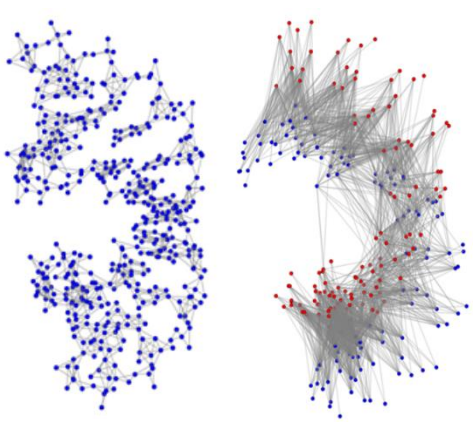
D3R Grand Challenge 4 (2018-2019)

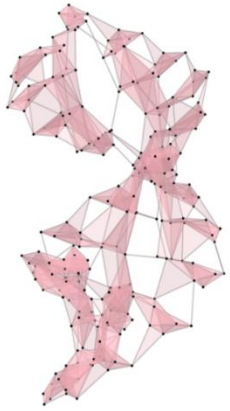
Pose Predictions
BACE Stage 1A
Pose Predictions (Partials) 2/3 2/3
BACE Stage 1B
Pose Prediction (Partials) 2/2 1/2

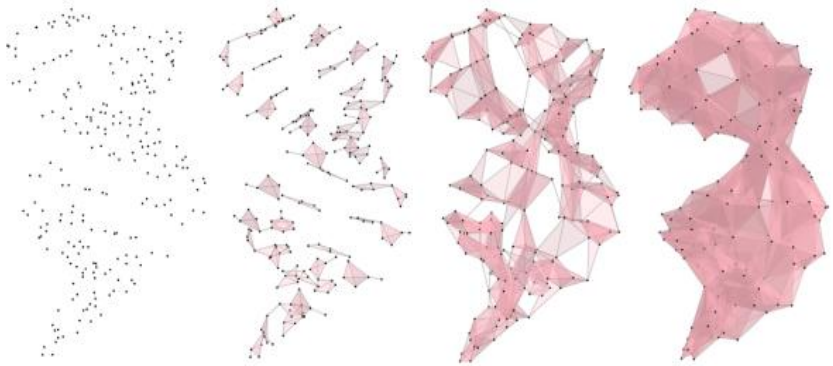
Affinity Predictions
Cathepsin Stage 1
Combined Ligand and Structure Based Scoring 2/5 2/3 2/4
Ligand Based Scoring (No participation)
Structure Based Scoring 2/4 1/3 3/3
Free Energy Set 1/7 1/7 2/5

Mathematical representations for molecules

- ❑ *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; Erdős number; Percolation…

Advanced mathematical representations and intrinsic invariants

Simplicial complex models and measurements:

Combinatorial Laplacian; Hodge theory; Betti number; Euler characteristics; Homology; Cohomology; Morse theory; Knot polynomials…

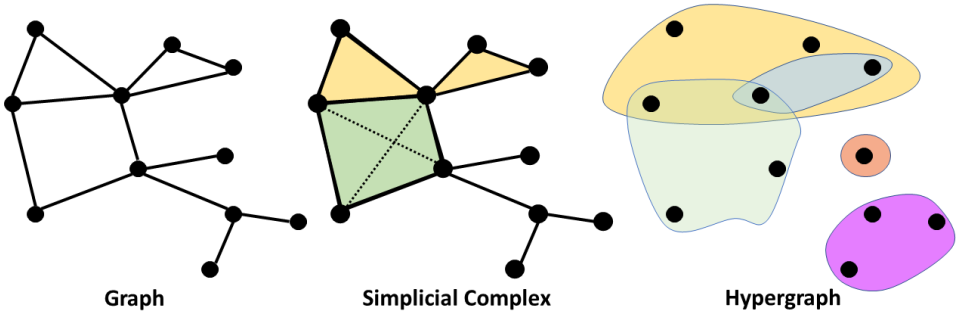
Multiscale simplicial complex:

Persistent homology; Persistent cohomology…

Persistent Spectral (PerSpect)

Spectral models

- ❑ *Spectral graph*
- ❑ *Spectral simplicial complex*
- ❑ *Spectral hypergraph*



Filtration

- ❑ *Nested sequence of Graphs*
- ❑ *Nested sequence of Simplicial Complexes*
- ❑ *Nested sequence of Hypergraphs*

Spectral models + Filtration

- ❖ **Persistent spectral (PerSpect) graph**
- ❖ **Persistent spectral (PerSpect) simplicial complexes**
- ❖ **Persistent spectral (PerSpect) 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 Society, 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

.....

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.

1. *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 $C_k(K)$.

2. *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 δ_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

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

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)

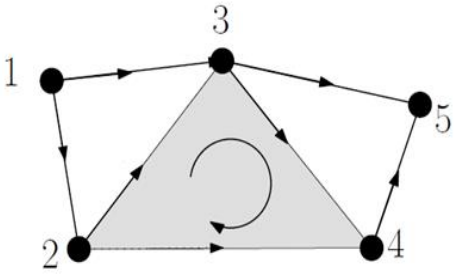
lower adjacent

$$L_k(i,j) = \begin{cases} d(\sigma_i^k) + k + 1, & \text{if } i = j \\ 1, & \text{if } i \neq j, \sigma_i^k \nprec \sigma_j^k, \sigma_i^k \succ \sigma_j^k \text{ and } \sigma_i^k \sim \sigma_j^k \\ -1, & \text{if } i \neq j, \sigma_i^k \nprec \sigma_j^k, \sigma_i^k \succ \sigma_j^k \text{ and } \sigma_i^k \nsim \sigma_j^k \\ 0, & \text{if } i \neq j, \sigma_i^k \succ \sigma_j^k \text{ or } \sigma_i^k \nprec \sigma_j^k. \end{cases}$$

k>0

upper adjacent

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



$\mathbf{B_1} =$

[12]	[13]	[23]	[24]	[34]	[35]	[45]	
−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]

$\mathbf{B_2} =$

[2,3,4]	
0	[1,2]
0	[1,3]
1	[2,3]
−1	[2,4]
1	[3,4]
0	[3,5]
0	[4,5]

The Laplacians are computed as

$$\mathcal{L}_0 = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{pmatrix},$$

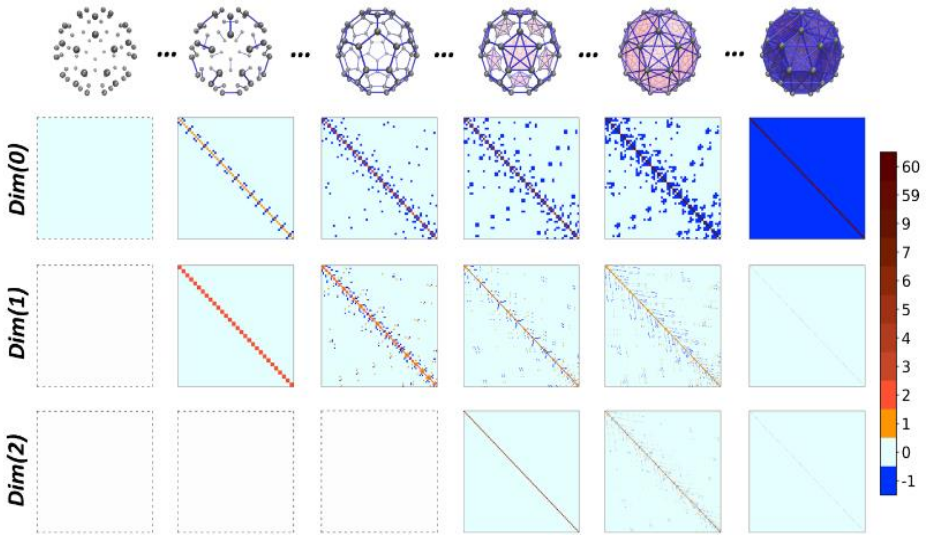
$$\mathcal{L}_1 = \begin{array}{c} \begin{matrix} [12] & [13] & [23] & [24] & [34] & [35] & [45] \end{matrix} \\ \begin{pmatrix} 2 & 1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & -1 & -1 & 0 \\ -1 & 1 & 3 & 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 3 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 3 & 1 & -1 \\ 0 & -1 & -1 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & -1 & -1 & 1 & 2 \end{pmatrix} \end{array},$$

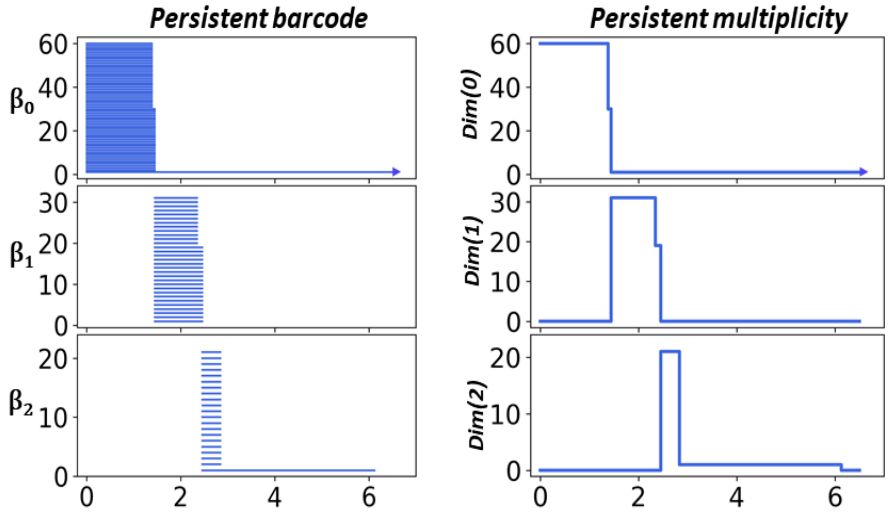
$$\mathcal{L}_2 = 3.$$

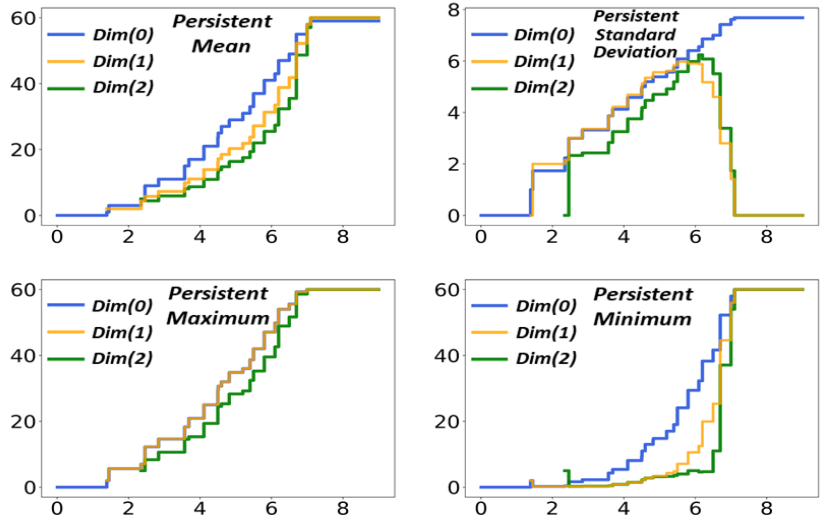
Persistent spectral simplicial complex

Combinatorial Laplacian (Hodge Laplacian)

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





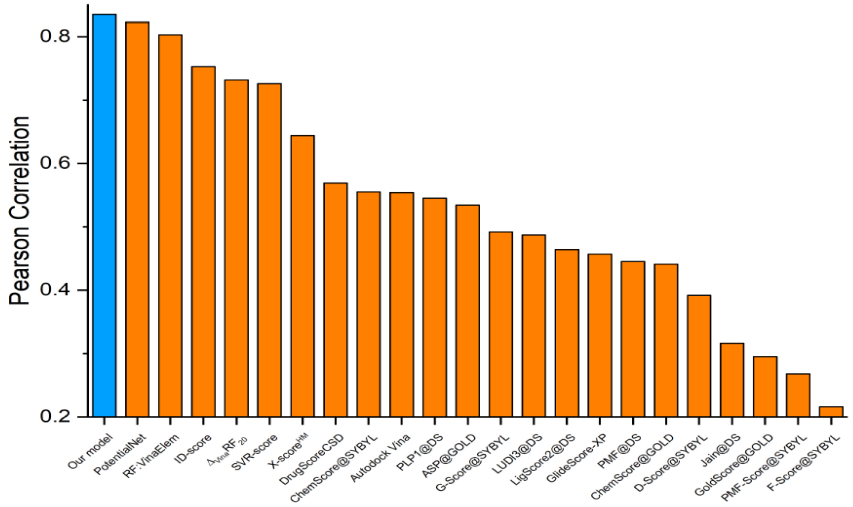


Multiplicity of zero eigenvalues (Persistent multiplicity) from PerSpect simplicial complex is equivalent to persistent Betti number.

PerSpect variables change with filtration parameter and incorporate in them related geometric information.

Benchmark testing with PDBbind datasets

Ours: 0.836 *Dataset 2007*



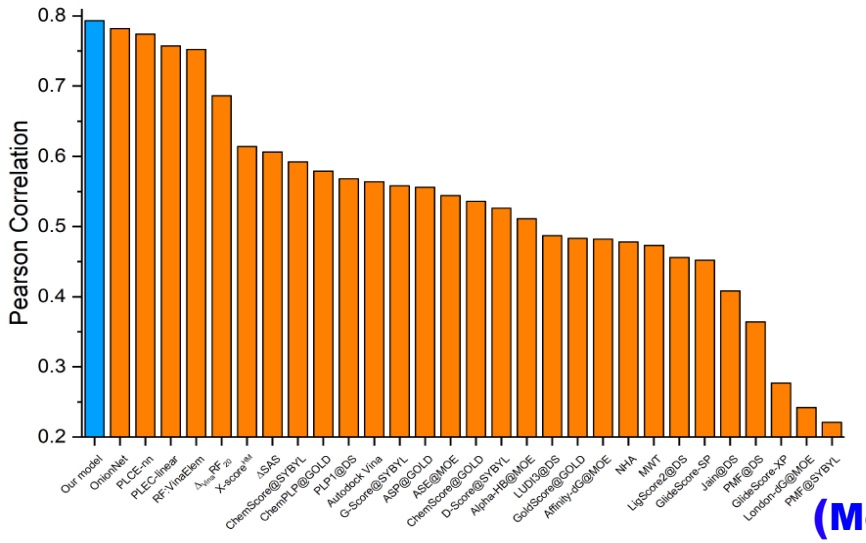
Model setting:

Spectral vectors

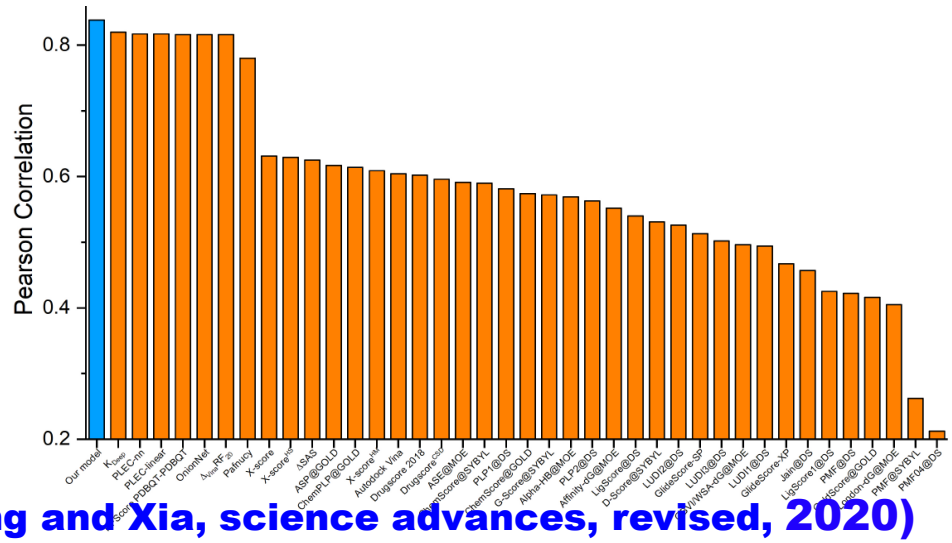
+

Random forest

Ours: 0.793 *Dataset 2013*

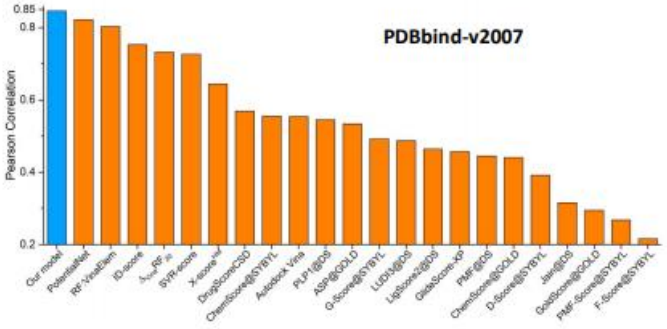


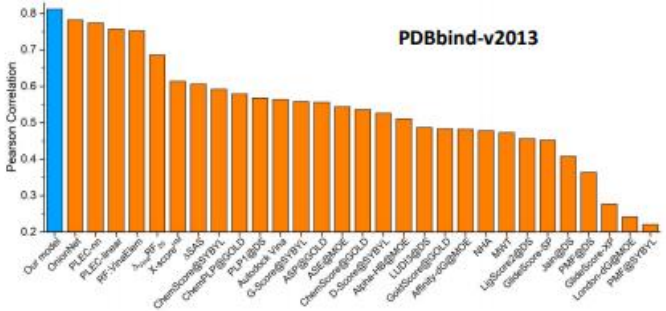
Ours: 0.840 *Dataset 2016*

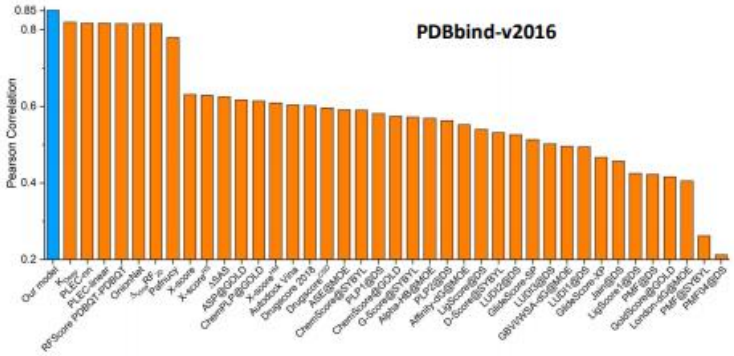


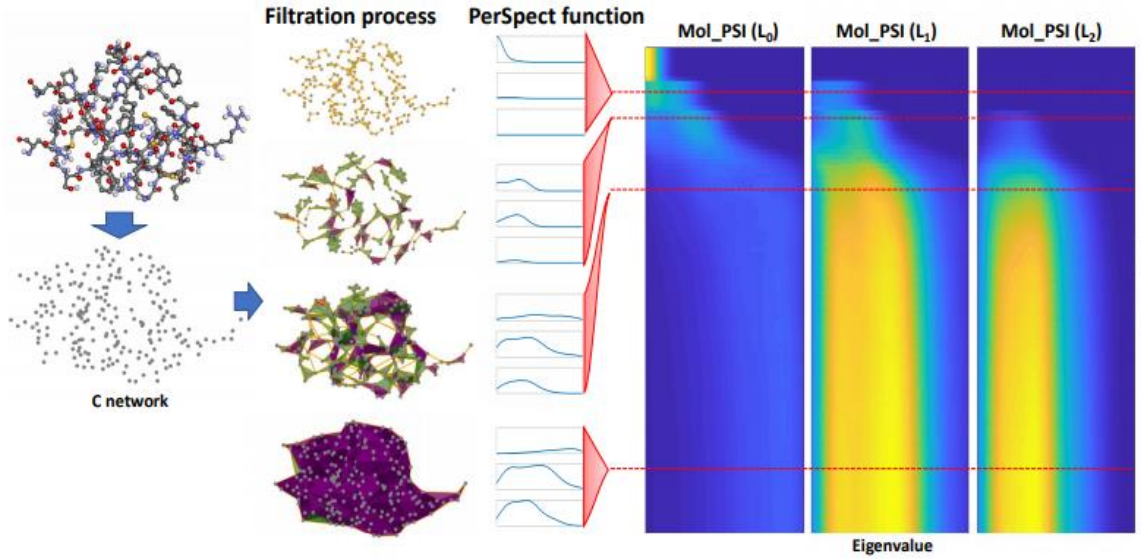
(Meng and Xia, science advances, revised, 2020)

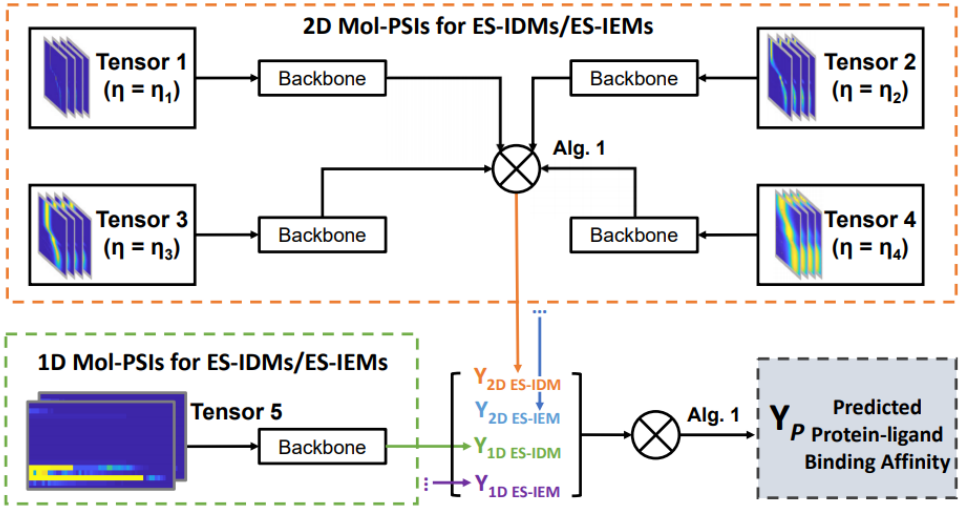
Molecular persistent spectral image (Mol-PSI) based CNNs





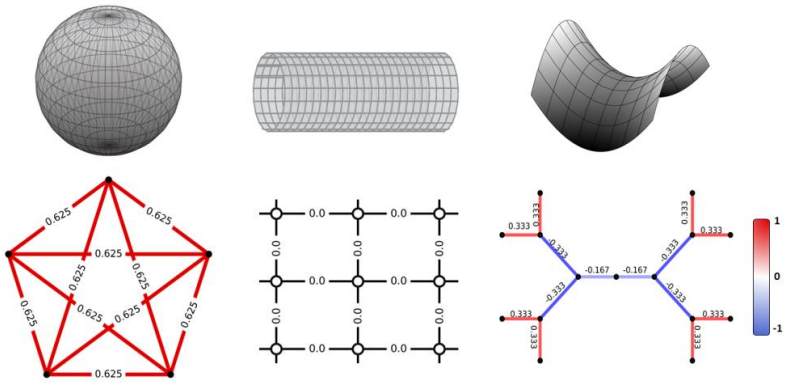




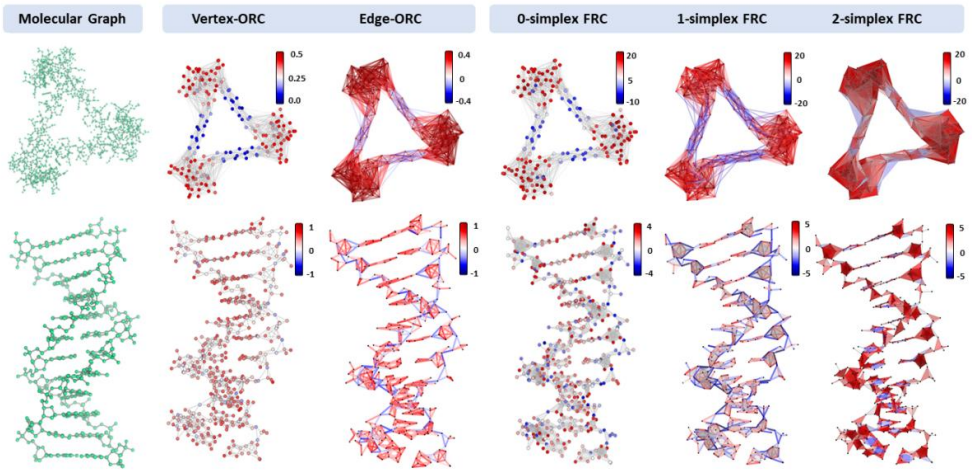


Joint work with Alibaba group (Peiran Jiang, Zhenyu Meng, Ying Chi, Lei Zhang, Xian-Sheng Hua, Kelin Xia, submitted, 2020)

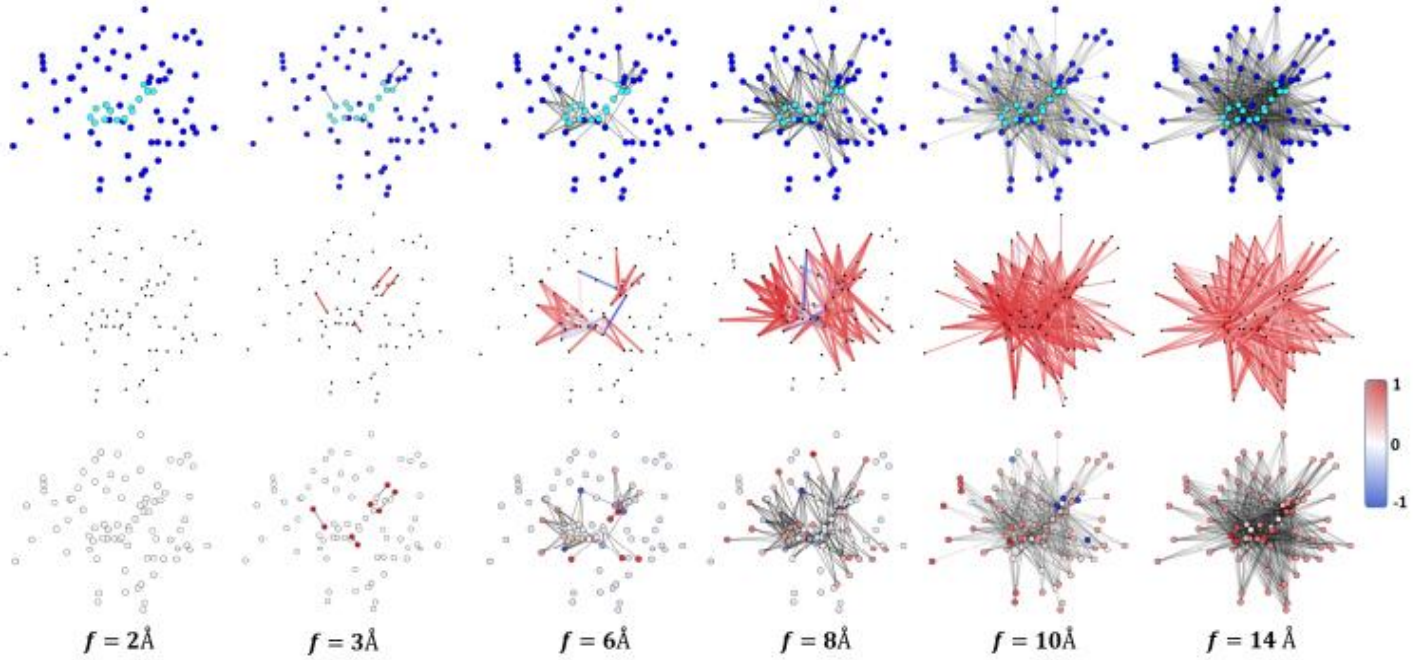
ORCs for three graphs



Ollivier Ricci curvature (ORC) VS Forman Ricci curvature (FRC)



Persistent Ricci curvature



Olliver Ricci curvature (ORC)

Vertex-x related probability measure:

constant between 0 and 1

$$m_x(x_i) = \begin{cases} \alpha & \text{if } x_i = x. \\ (1-\alpha)/k_x & \text{if } x_i \in \Gamma_x. \\ 0 & \text{otherwise.} \end{cases}$$

set of all the neighboring vertices of x

L^1 Wasserstein distance:

distance between two vertices

$$W_1(m_x, m_y) = \inf_{\xi} \sum_{x_i \in V} \sum_{y_j \in V} d(x_i, y_j)\xi(x_i, y_j)$$

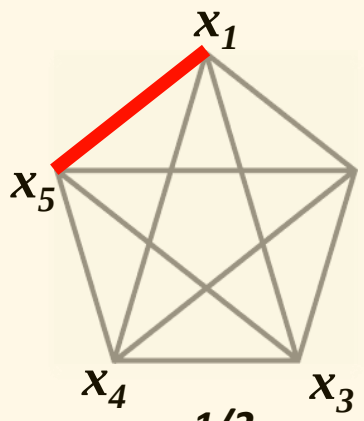
transportation distance

$$\sum_{y_j \in V} \xi(x, y_j) = m_x$$
$$\sum_{x_i \in V} \xi(x_i, y) = m_y$$

Ollivier Ricci curvature:

$$c(x, y) := 1 - \frac{W_1(m_x, m_y)}{d(x, y)}$$

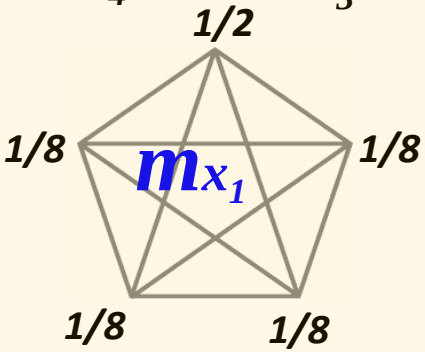
An example

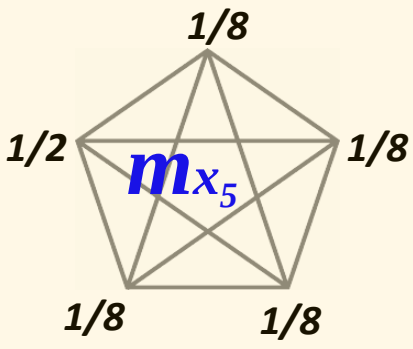


set $\alpha = 1/2$, check edge (x_1, x_5)

$\Gamma x_1 = \{x_2, x_3, x_4, x_5\}$

$\Gamma x_5 = \{x_2, x_3, x_4, x_1\}$





$$W_1(m_{x_1}, m_{x_5}) = (\tfrac{1}{2} - \tfrac{1}{8}) \times d(x_1, x_5)$$
$$c(x_1, x_5) = 1 - \frac{W_1(m_{x_1}, m_{x_5})}{d(x_1, x_5)}$$
$$= 1 - \frac{(\frac{1}{2} - \frac{1}{8}) \times d(x_1, x_5)}{d(x_1, x_5)}$$
$$= \tfrac{5}{8}$$

F. R. Chung and S.-T. Yau, *"Logarithmic harnack inequalities,"* Mathematical Research Letters, 3(6),793–812, 1996.

Y. Ollivier, *"Ricci curvature of metric spaces,"* Comptes Rendus Mathematique, 345 (11), 643–646, 2007.

J. Lott and C. Villani, *"Ricci curvature for metric-measure spaces via optimal transport,"* Annals of Mathematics, 903–991, 2009

Y. Lin, L. Lu, and S.-T. Yau, *"Ricci curvature of graphs,"* Tohoku Mathematical Journal, Second Series, 63(4),605–627, 2011.

Y. Lin and S.-T. Yau, "Ricci curvature and eigenvalue estimate on locally finite graphs," Mathematical research letters, 17(2), 343–356, 2010.

C.-C. Ni, Y.-Y. Lin, J. Gao, X. D. Gu, and E. Saucan, *"Ricci curvature of the internet topology,"* in 2015 IEEE Conference on Computer Communications (INFOCOM), 2758– 2766, IEEE, 2015.

W. Zeng, D. Samaras, and X. D. Gu. *"Ricci flow for 3D shape analysis."* IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (4) 662-677, 2010.

.....

Forman Ricci curvature (FRC)

k-simplex

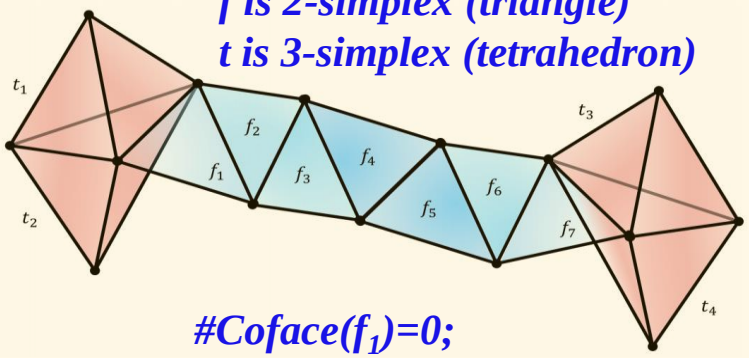
$$\mathcal{F}_k^{\sharp}(\sigma^k) = \sharp\{\sigma^k \subset \sigma^{(k+1)} >\} + \sharp\{\sigma^{(k-1)} \subset \sigma^k\} - \sharp\{\text{parallel neighbours of } \sigma^k\}$$

upper degree (number of "cofaces")

lower degree (number of "faces")

either upper adjacent or lower adjacent, but not both

f is 2-simplex (triangle)
t is 3-simplex (tetrahedron)



#Coface(f_1)=0;
#face(f_1)=3;
#Neighbour (f_1)=4.

$$\mathcal{F}_2^{\sharp}(f_1) = \mathcal{F}_2^{\sharp}(f_7) = -1.$$
$$\mathcal{F}_2^{\sharp}(f_i) = 1 \text{ for } i = 2, 3, \cdots, 6.$$
$$\mathcal{F}_3^{\sharp}(t_i) = 3 \text{ for } i = 1, 2, 3, 4.$$

ORC VS FRC

(1) ORC is defined on graph; FRC is defined on cell complex.

(2) ORC is for network probabilistic properties; FRC is for combinatorial properties.

(3) ORC is highly correlated with FRC:

Positive ORCs or FRCs are commonly found in densely-packed clusters or "communities", while negative ORCs or FRCs usually represent bridges or links between two clusters.

R. Forman, *"Bochner's method for cell complexes and combinatorial Ricci curvature,"* Discrete and Computational Geometry, 29 (3), 323–374, 2003.

Desbrun, Mathieu, Eva Kanso, and Yiying Tong. *"Discrete differential forms for computational modeling."* Discrete differential geometry. Birkhäuser Basel, 287-324, 2008.

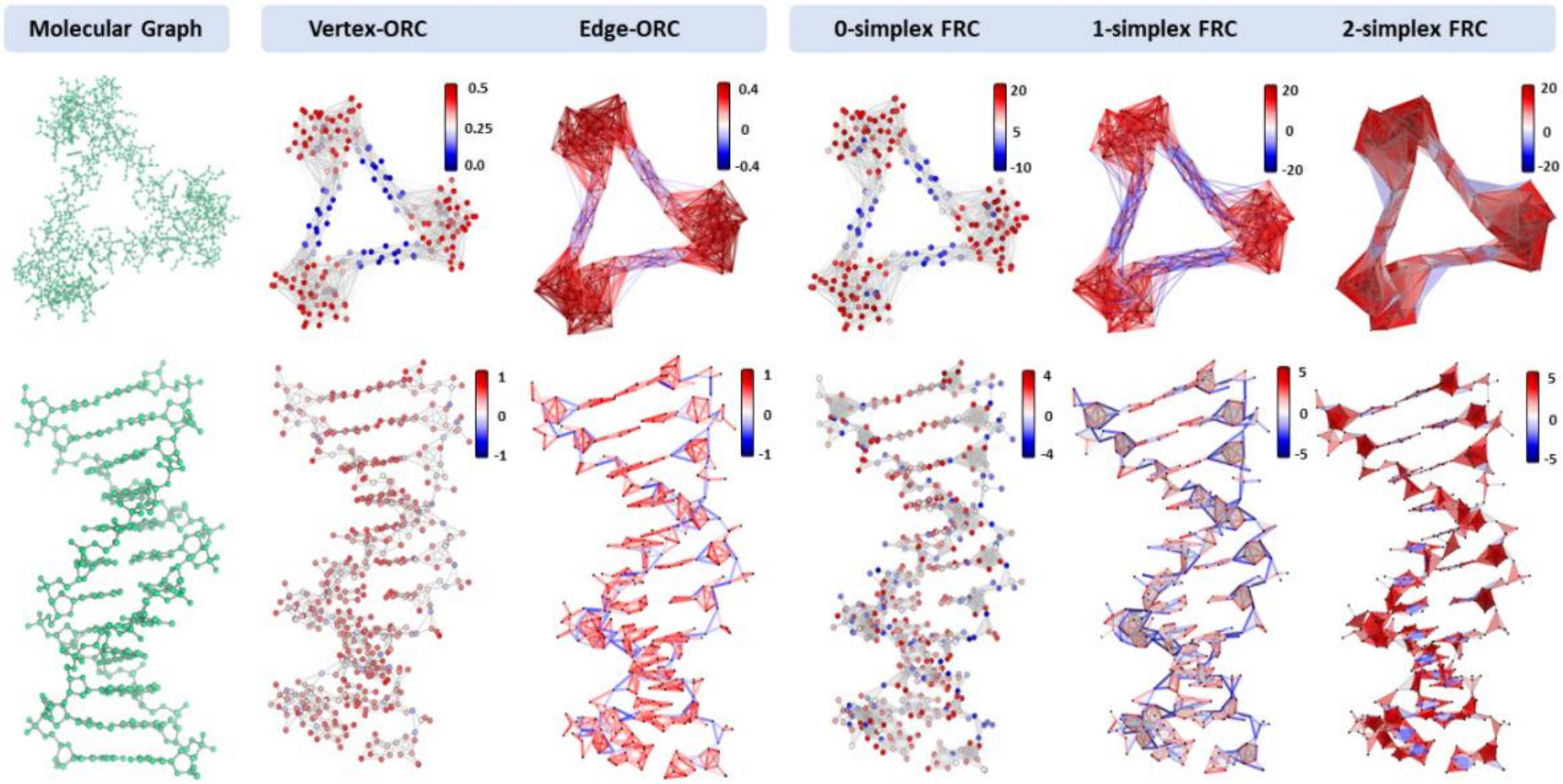
Jost, Jürgen, and Jèurgen Jost. *"Riemannian geometry and geometric analysis"*. Vol. 42005. Berlin: Springer, 2008.

R. Sreejith, K. Mohanraj, J. Jost, E. Saucan, and A. Samal, *"Forman curvature for complex networks,"* Journal of Statistical Mechanics: Theory and Experiment, 2016(6), 063206, 2016.

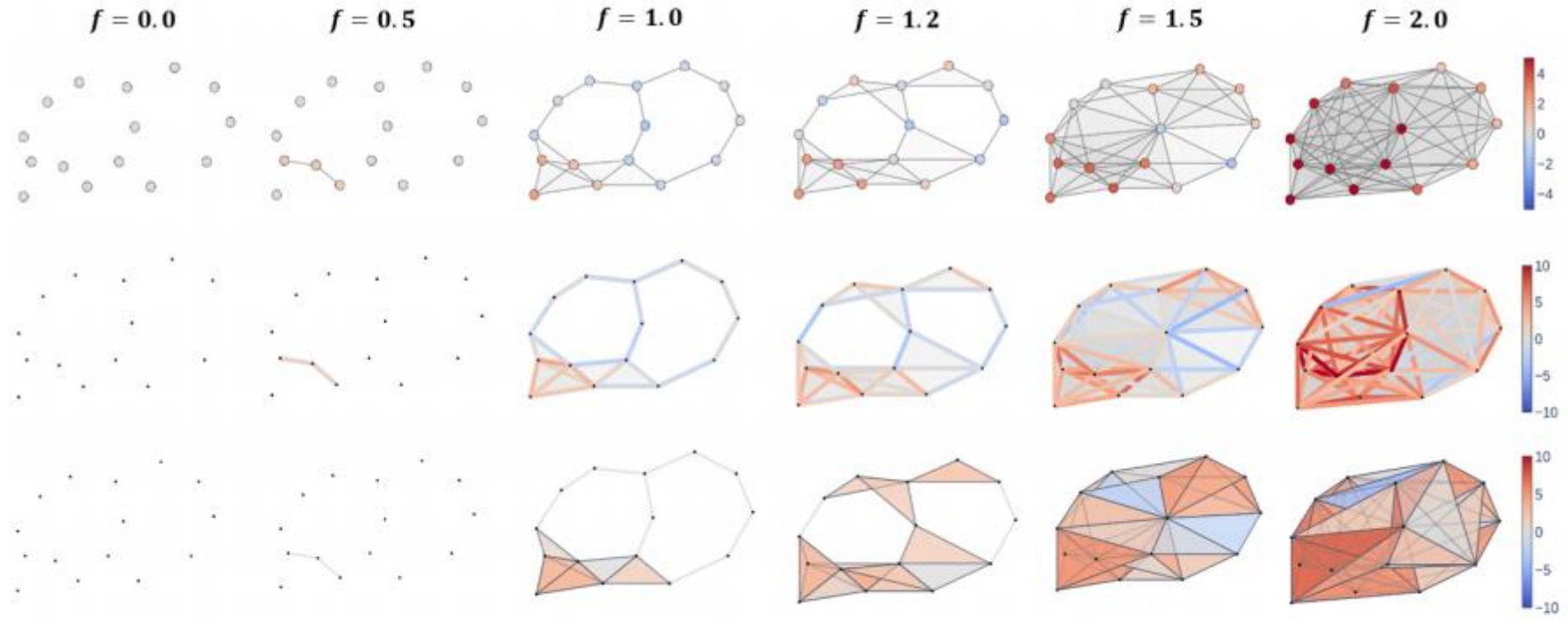
A. Samal, R. Sreejith, J. Gu, S. Liu, E. Saucan, and J. Jost. *"Comparative analysis of two discretizations of Ricci curvature for complex networks".* Scientific reports, 8(1):1–16, 2018.

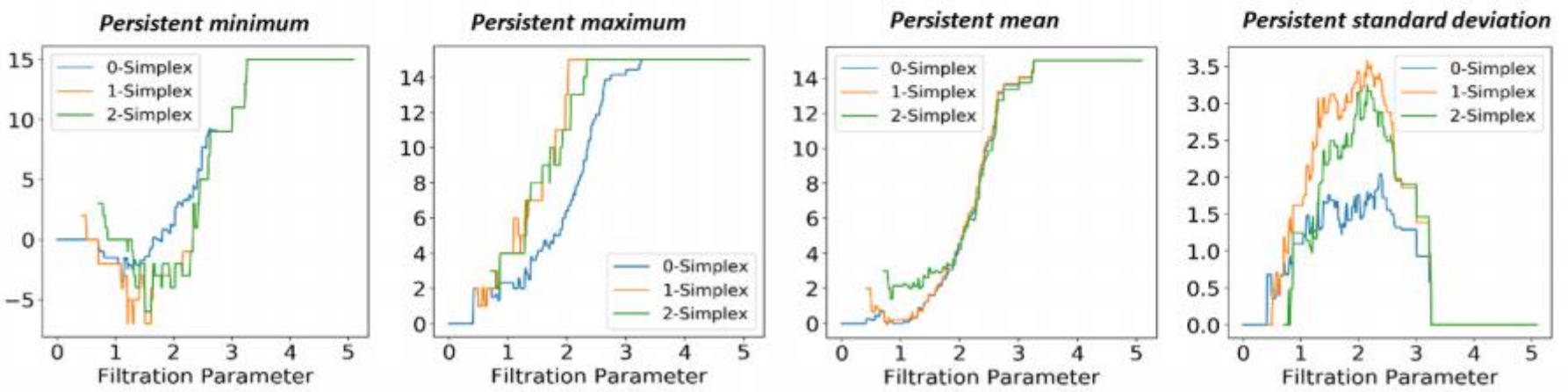
……

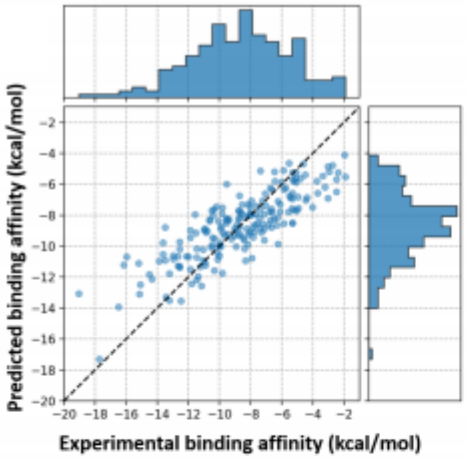
ORC VS FRC

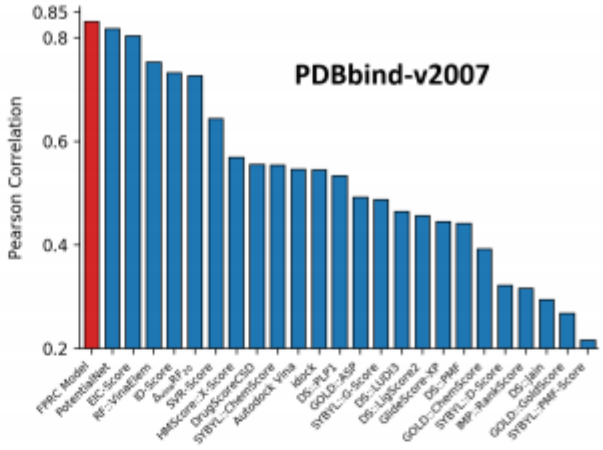

Molecular Graph
Vertex-ORC
Edge-ORC
0-simplex FRC
1-simplex FRC
2-simplex FRC
0.5
0.25
0.0
0.4
0
-0.4
20
5
-10
20
0
-20
20
0
-20
1
0
-1
1
0
-1
4
0
-4
5
0
-5
5
0
-5

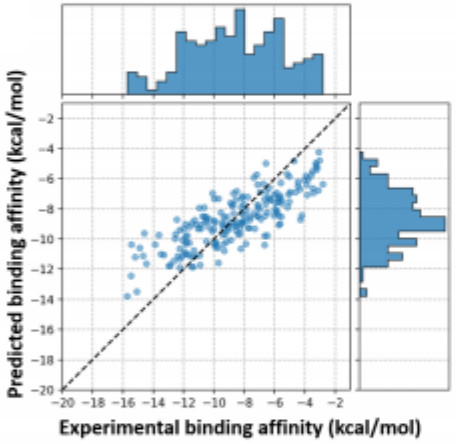
(Forman) Persistent Ricci curvature

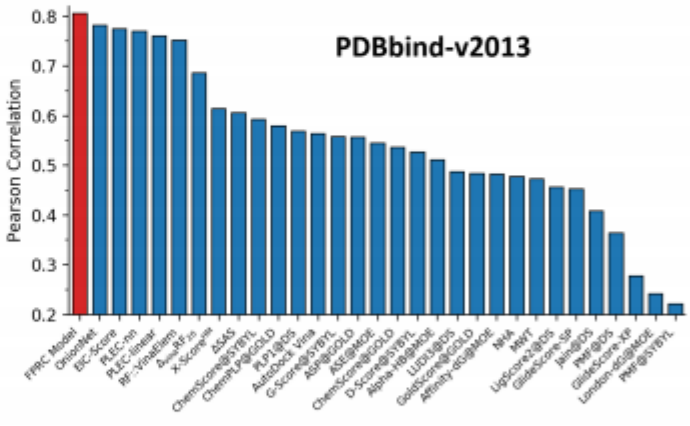


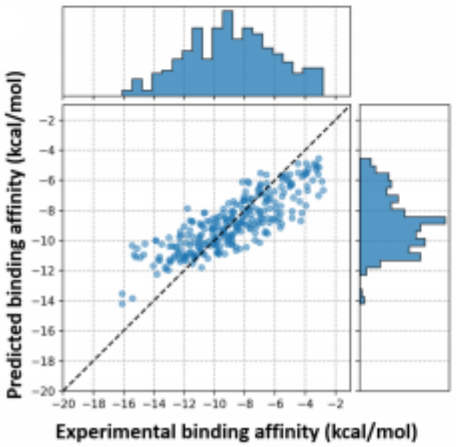


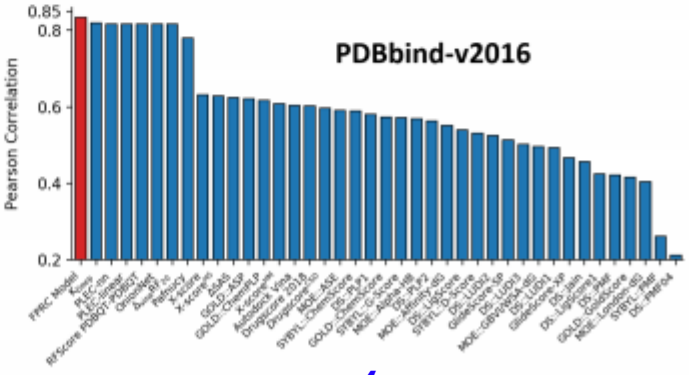












Benchmark testing with PDBbind datasets

Model setting:

Persistent attributes

+

Random forest / gradient boosting tree

(Wee, Xia, submitted, 2020)

Persistent spectral models

Combinatorial Bochner-Weitzenböck decomposition

Hodge decomposition

Persistent Ricci curvature

Gauss-Bonnet

Persistent homology

Combinatorial Bochner-Weitzenböck decomposition

$$\mathrm{L}_k = \Delta_k + \mathrm{Ricc}_k^{\mathcal{F}}$$

Combinatorial Bochner Laplacian

$$\Delta_k(i,j) = \begin{cases} \sum_{m \neq i} |\mathrm{L}_k(i,m)|, & i = j \\ \mathrm{L}_k(i,j), & i \neq j \end{cases}$$

Forman Ricci matrix

$$\mathrm{Ricc}_k^{\mathcal{F}}(i,j) = \begin{cases} \mathcal{F}_k^{\sharp}(\sigma_i^k), & i = j \\ 0, & i \neq j \end{cases}$$

Forman Ricci curvature

$$\mathcal{F}_k^{\sharp}(\sigma_i^k) = \underbrace{d(\sigma_i^k) + k + 1}_{\mathrm{L}_k(i,i)} - \underbrace{\sum_{m \neq i} |(B_{k+1}B_{k+1}^T)(i,m) + (B_k^T B_k)(i,m)|}_{\Delta_k(i,i)}$$

Hodge decomposition

harmonic flow *(topological invariant)*

$$C^1(K) = \mathrm{Im}(\delta_0) \oplus \ker(\Delta_1) \oplus \mathrm{Im}(\delta_1^*)$$

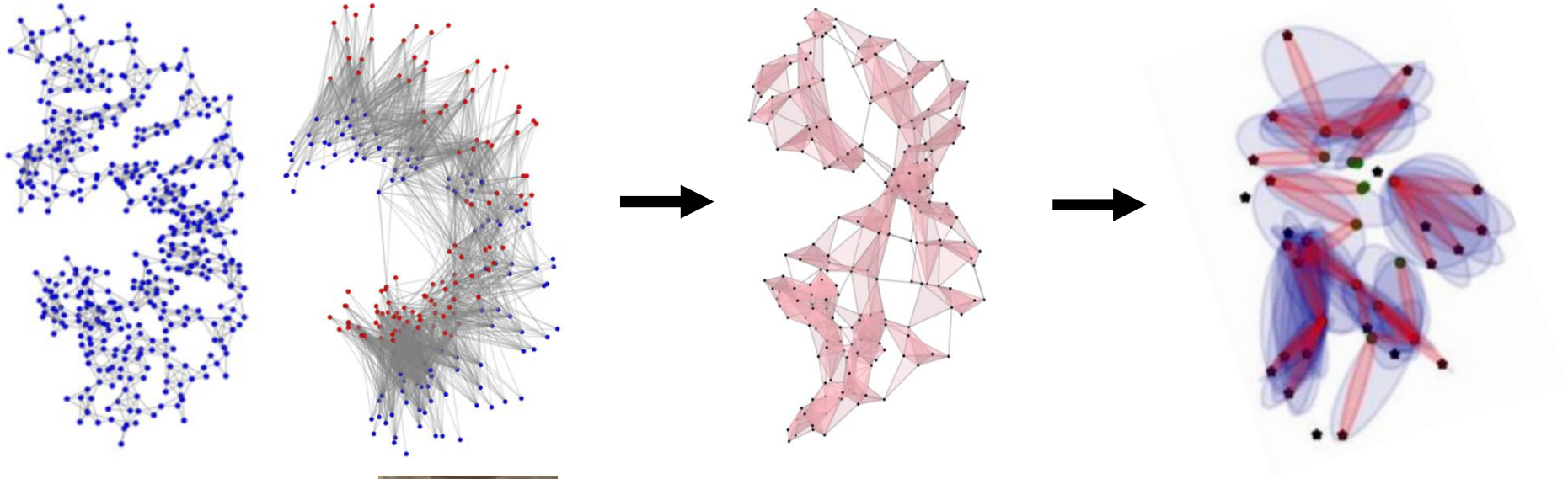
gradient flow

curl flow

Graph

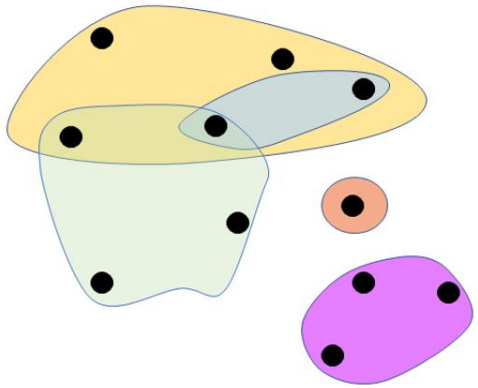
Simplicial complex

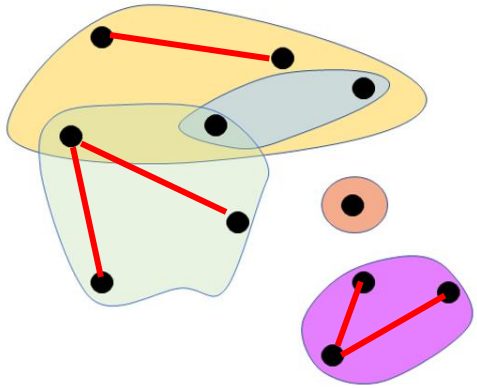
Hypergraph



Collaborator
Jie Wu
Math, HEBNU

Hypergraph VS super hypergraph





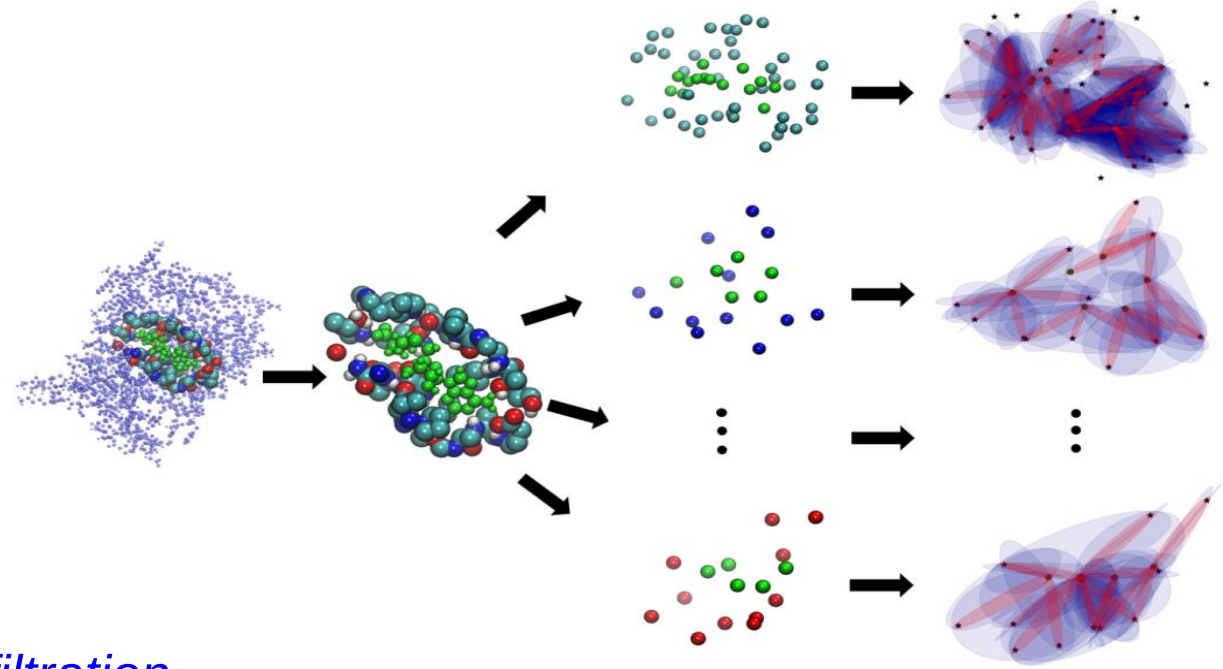
hyperedge is a set of vertices

super hyperedge is a subgraph

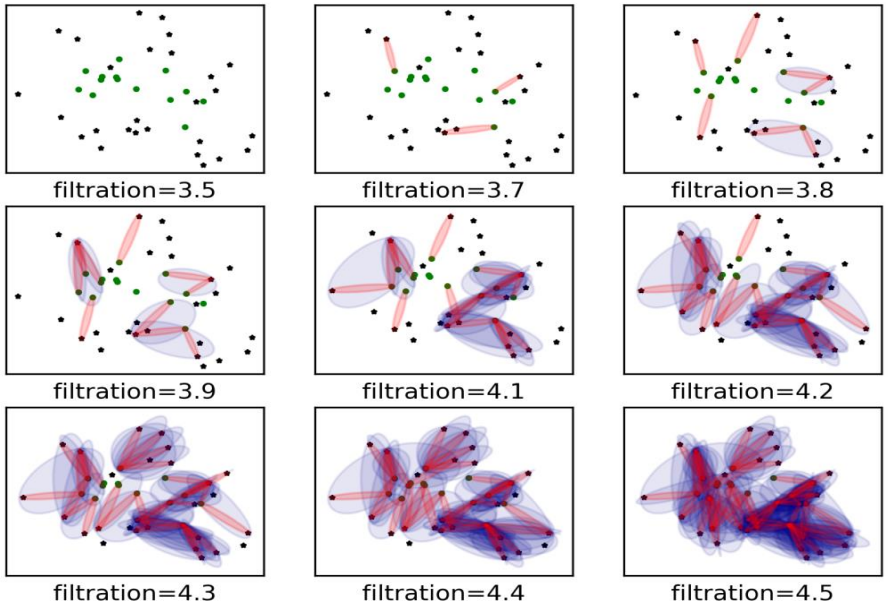
Ren, Shiquan, et al. "Computing the Homology of Hypergraphs." *arXiv preprint arXiv:1705.00151* (2017).
Ren, Shiquan, Chengyuan Wu, and Jie Wu. "Operators on random hypergraphs and random simplicial complexes." *arXiv preprint arXiv:1712.02045* (2017).
Ren, Shiquan, and Jie Wu. "Stability of persistent 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).

Hypergraph-based models

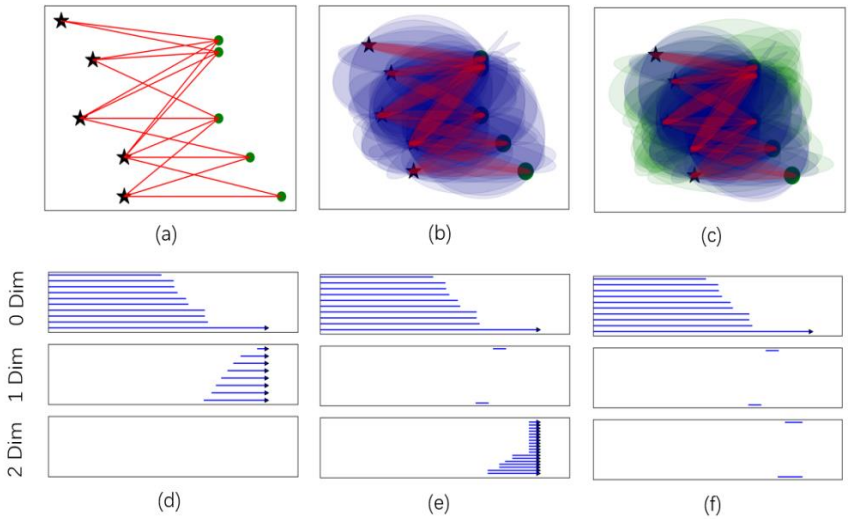
Protein-ligand interaction modeled as hypergraph



Hypergraph-based filtration



Bipartite graph VS Hypergraph



Hypergraph for protein-ligand interactions

Hypergraph: $(V_{\mathcal{H}}, \mathcal{H})$

set of vertices

set of hyperedges

$V_{\mathcal{H}} = V_P \cup V_L$

set of protein atoms

set of ligand atoms

Hyperedge:

At least one atom from protein and another atom from ligand

n-hyperedge

$$\sigma_n = \begin{cases} \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_n\}; \mathbf{v}_k \in V_{\mathcal{H}} (0 \leqslant k \leqslant n), \exists i, j \in [0, n], \mathbf{v}_i \in V_P, \mathbf{v}_j \in V_L, & n > 0 \\ \{\mathbf{v}_0\}; \mathbf{v}_0 \in V_{\mathcal{H}}, & n = 0. \end{cases}$$

Filtration value for hyperedge:

$$f(\sigma_n) = \begin{cases} \max_{0 \leq i < j \leq n} d(\mathbf{v}_i, \mathbf{v}_j), & n > 0 \\ 0, & n = 0. \end{cases}$$

Interactive distance:

$$d(\mathbf{v}_i, \mathbf{v}_j) = \begin{cases} \|\mathbf{v}_i - \mathbf{v}_j\|, & \text{if } \mathbf{v}_i \in \mathbf{V}_P, \mathbf{v}_j \in \mathbf{V}_L \text{ or } \mathbf{v}_i \in \mathbf{V}_L, \mathbf{v}_j \in \mathbf{V}_P \\ g(\mathbf{v}_i, \mathbf{v}_j), & \text{otherwise.} \end{cases}$$

Distance between two atoms from the same molecule is defined as the maximal distance of the two atoms to the other molecule

$$g(\mathbf{v}_i, \mathbf{v}_j) = \begin{cases} \max_{\mathbf{v}_k \in V_P, \{\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k\} \in \mathcal{H}} \{\|\mathbf{v}_i, \mathbf{v}_k\|, \|\mathbf{v}_j, \mathbf{v}_k\|\} + d_0, & \text{if } \mathbf{v}_i, \mathbf{v}_j \in \mathbf{V}_L \\ \max_{\mathbf{v}_k \in V_L, \{\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k\} \in \mathcal{H}} \{\|\mathbf{v}_i, \mathbf{v}_k\|, \|\mathbf{v}_j, \mathbf{v}_k\|\} + d_0, & \text{if } \mathbf{v}_i, \mathbf{v}_j \in \mathbf{V}_P \end{cases}$$

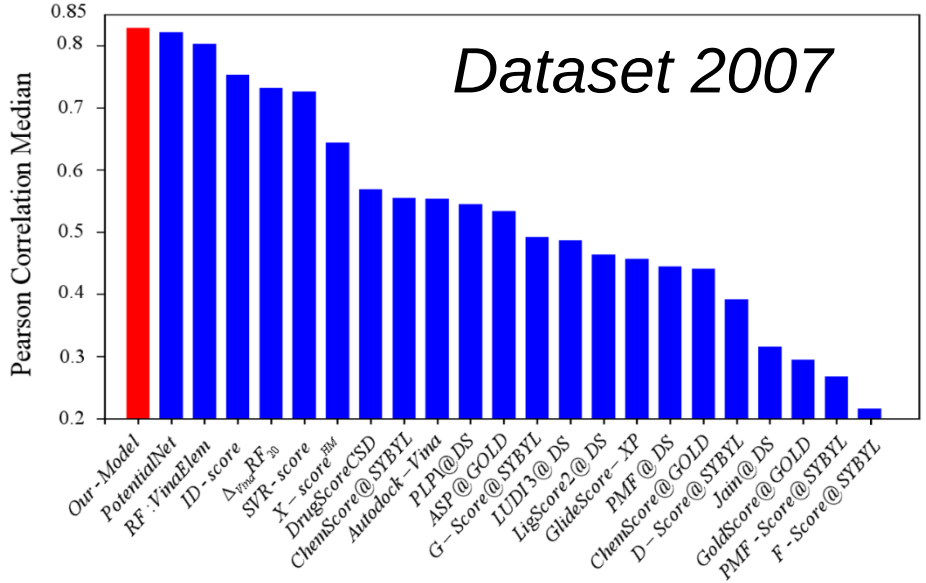
Benchmark testing with PDBbind datasets

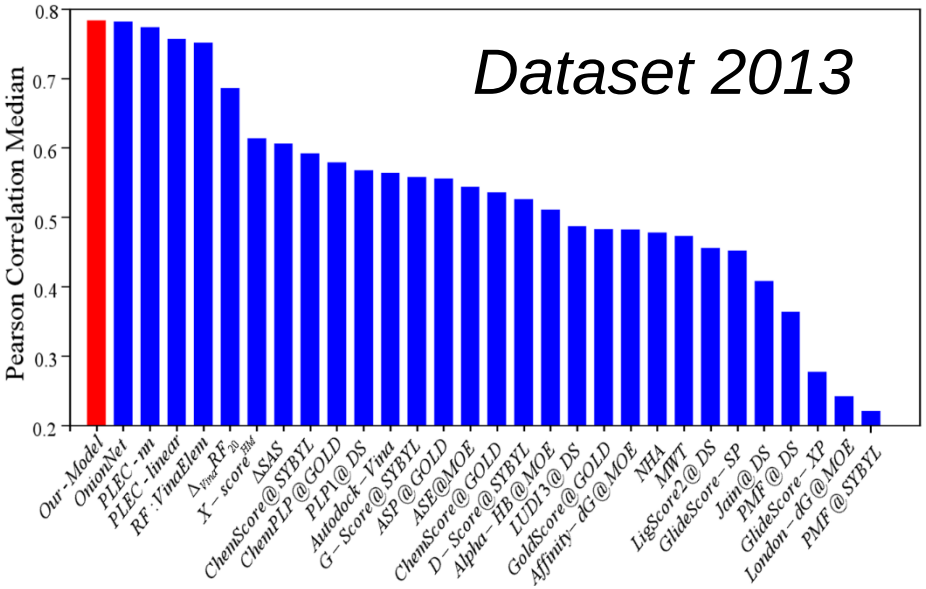
Model setting:

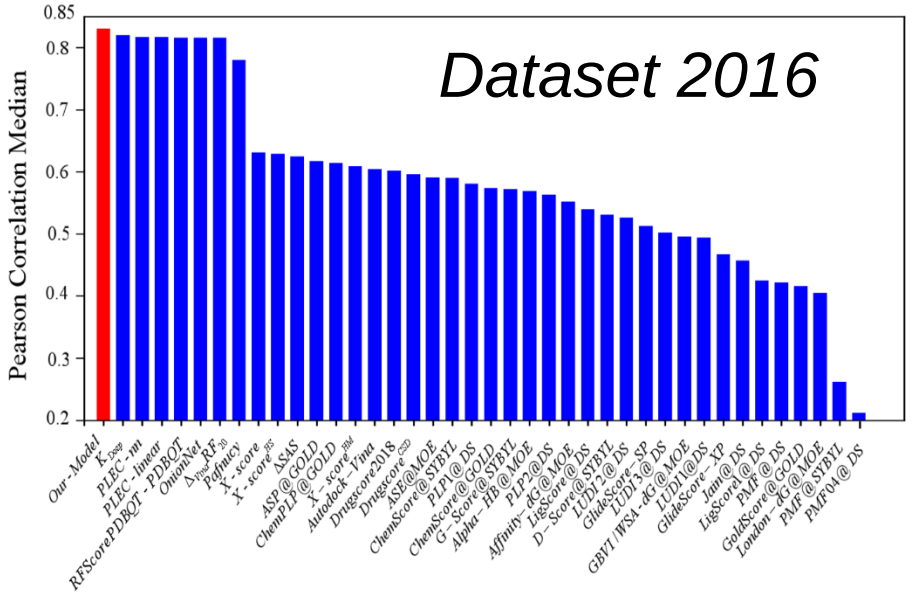
homology vectors

+

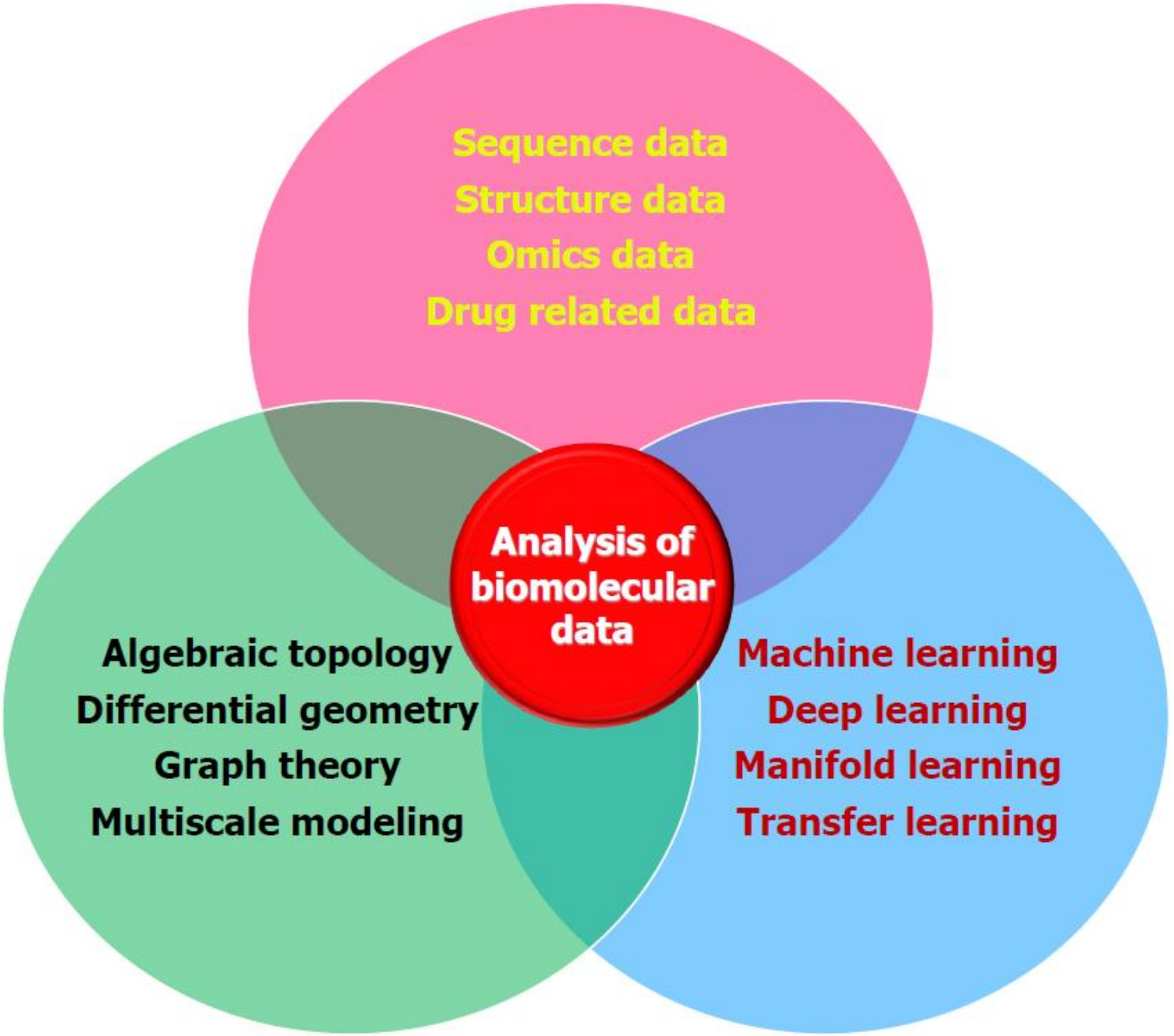
Random forest







(Liu, Wu, Wang, Xia, briefings in bioinformatics, accepted, 2020)


Sequence data
Structure data
Omics data
Drug related data
Analysis of biomolecular data
Algebraic topology
Differential geometry
Graph theory
Multiscale modeling
Machine learning
Deep learning
Manifold learning
Transfer learning

Group members

Bin Liu *Xiang Liu* *Zhenyu Meng*

Joon Wei KOH *JunJie WEE* *Vijai Anand*

Two postdocs (2021)

Geometric/topological modelling, Machine learning, Mathematical biology, Bioinformatics…

阿里巴巴达摩院AI医疗团队实习生 (机器学习+拓扑/几何)

Grant support

NTU-JSPS (2019-2022)
Alibaba-NTU (2020-2021)
Merlion (2020-2022)
MOE-Tier 1 (2018-2021, 2019-2022)
MOE-Tier 2 (2018-2021, 2021-2024)

Email : xiakelin@ntu.edu.sg

Conference Program

The Third Conference on Computational and Mathematical Bioinformatics and Biophysics

Tsinghua Sanya International Mathematics Forum

December 20 – 24, 2020, (Beijing Time)
December 19 – 23, 2020, (US Central Time)

Zoom Meeting ID: 981 2055 6545

Password: CMBB

Zoom Link:
https://uasystem.zoom.us/j/98120556545?pwd=OXJhZENlTDhpaUsrNDQ5a2NvQVNudz09

Organizing Committee:
Stephen Shing-Toung Yau, Tsinghua University, China
Guowei Wei, Michigan State University, USA
Changchuan Yin, University of Illinois at Chicago, USA
Shan Zhao, University of Alabama, USA