Ollivier persistent Ricci curvature (OPRC) based molecular representations for drug design

Wee JunJie

Supervisor: Asst. Prof. Xia Kelin

Nanyang Technological University

WEEJ0019@e.ntu.edu.sg

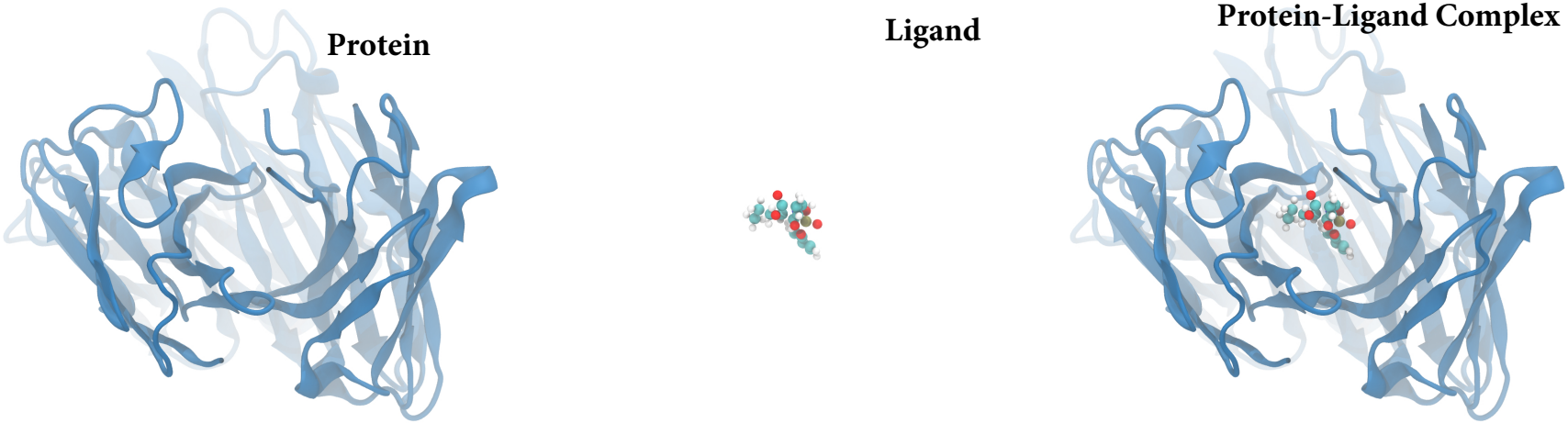
19-Nov-2020

Outline

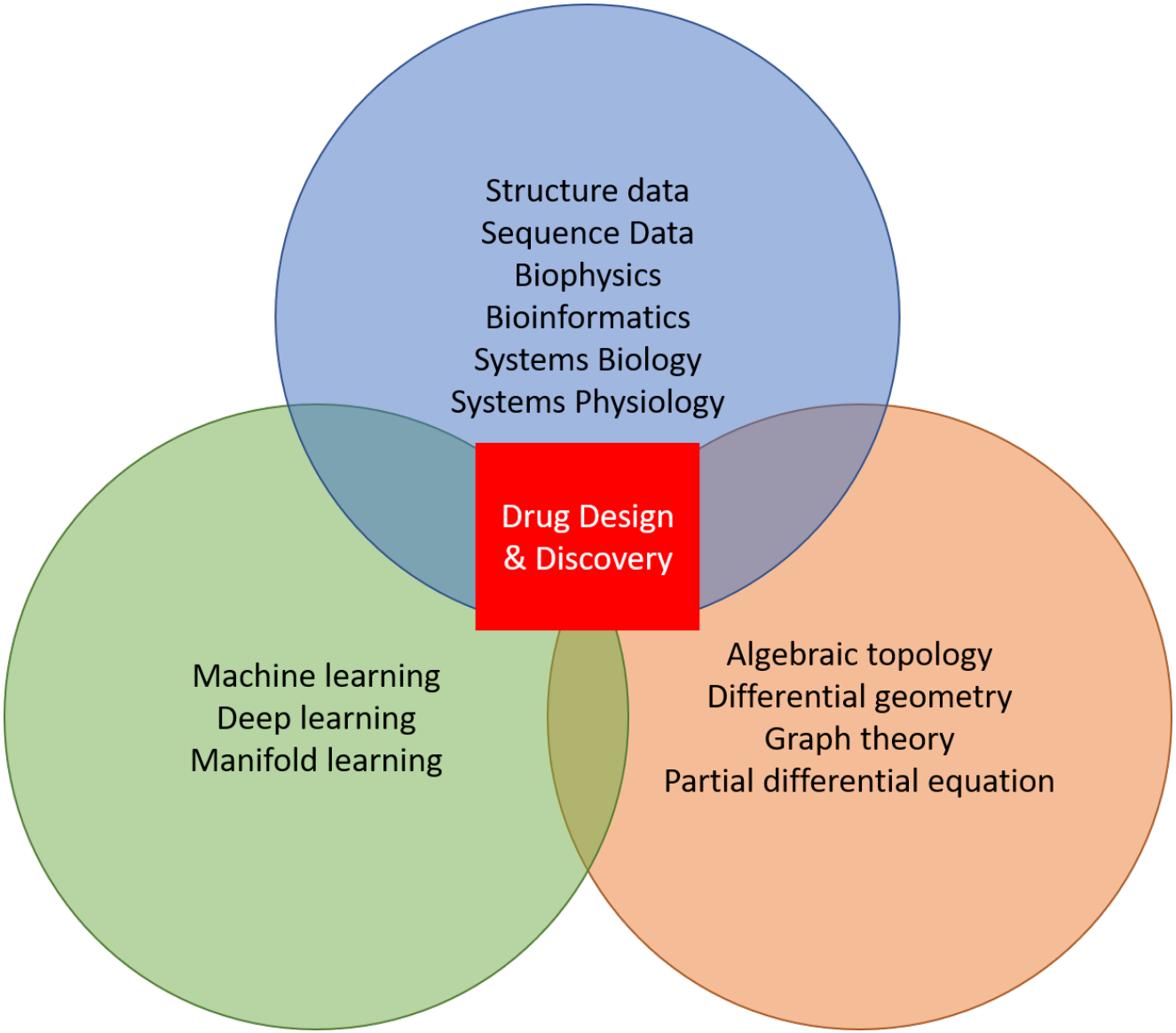
Outline

Why Study Protein-Ligand Interactions?

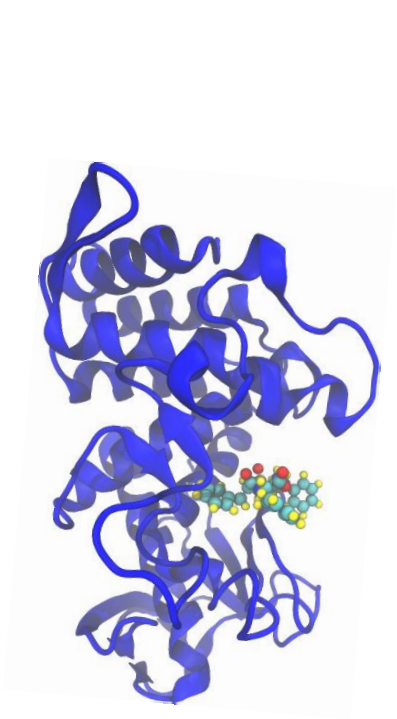
- Currently, it takes over 10 years and about USD2.6 billion to bring an average drug into the market.
- Protein-Ligand Interactions are core mechanisms for actions in drug design.
- One of the essential information needed in drug design is the binding affinity which measures the strength of binding interaction between protein and its ligand partner.

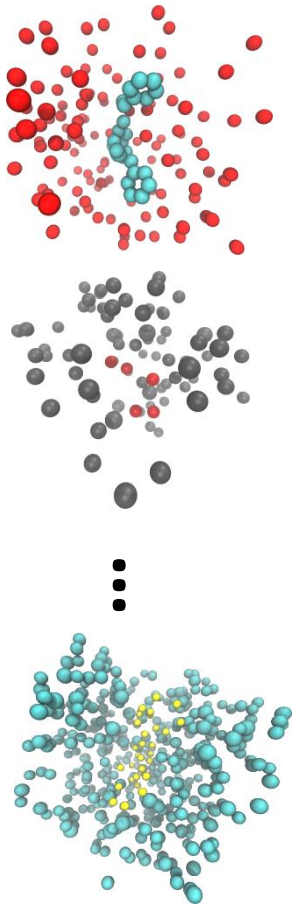


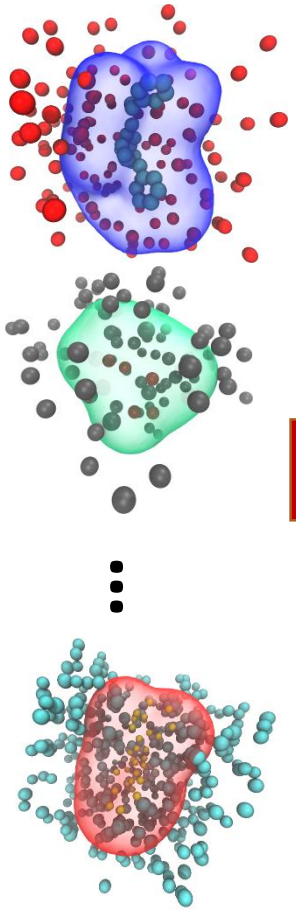
Drug Design & Discovery



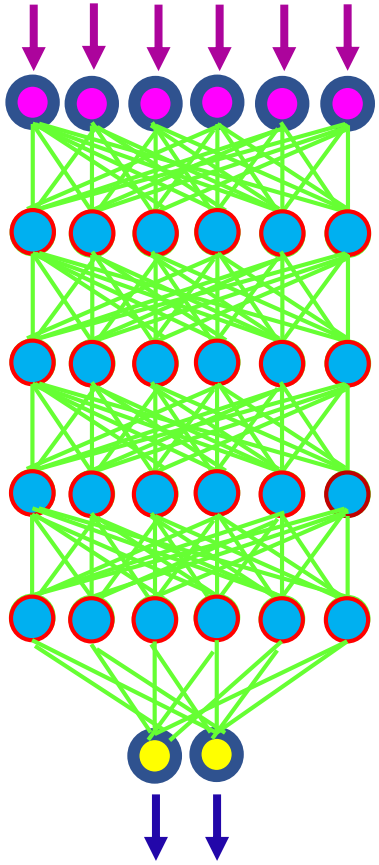
Mathematical deep learning







Algebraic topology
or
Differential geometry
or
Graph theory
or
Differential equation
or
.....



Protein-ligand complex

Element specific groups

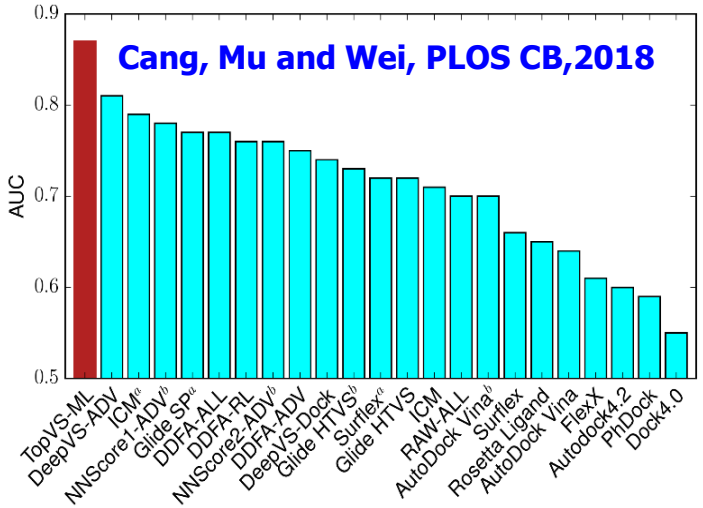
Element interactive manifolds

Various Mathematical features

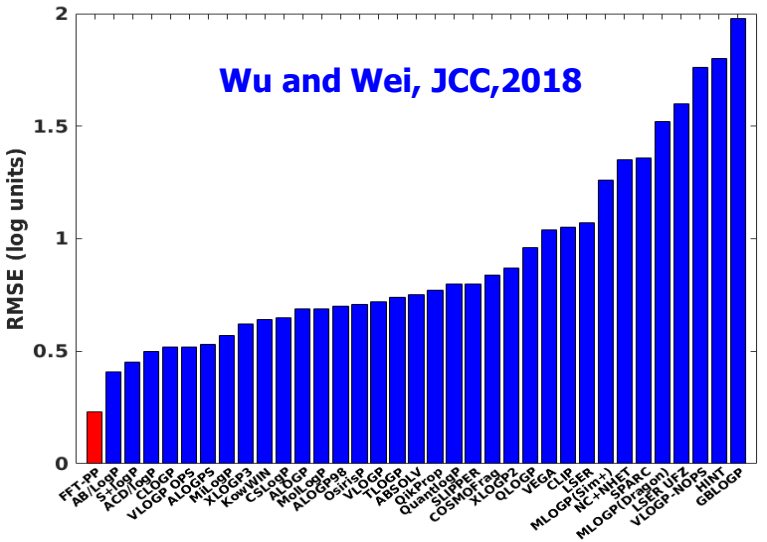
Machine learning prediction

Topological learning based predictions

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



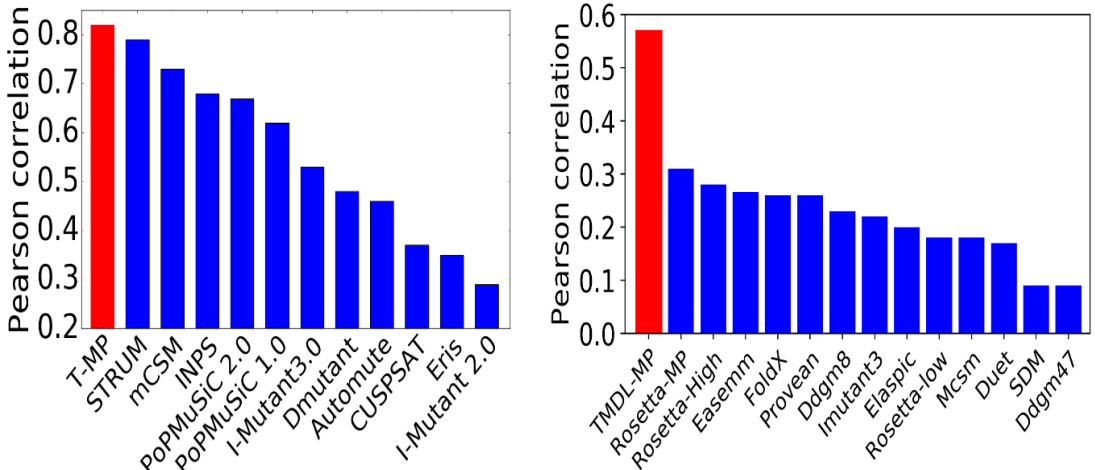
Prediction RMSD of LogP (Star set)



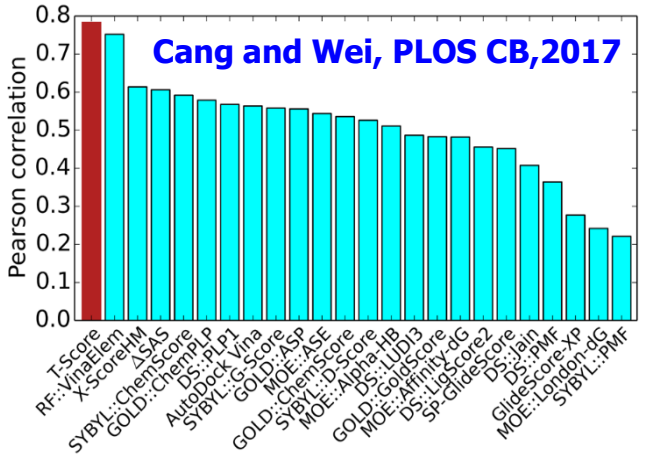
Predicting mutations on 2648 globular proteins

Predicting mutations on 223 membrane proteins

(Cang and Wei, Bioinformatics, 2017)

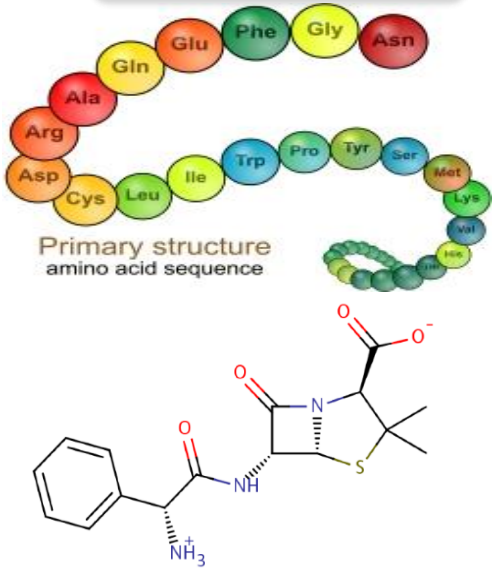


Binding affinity prediction of PDBBind v2013 core set of 195 protein-ligand complexes

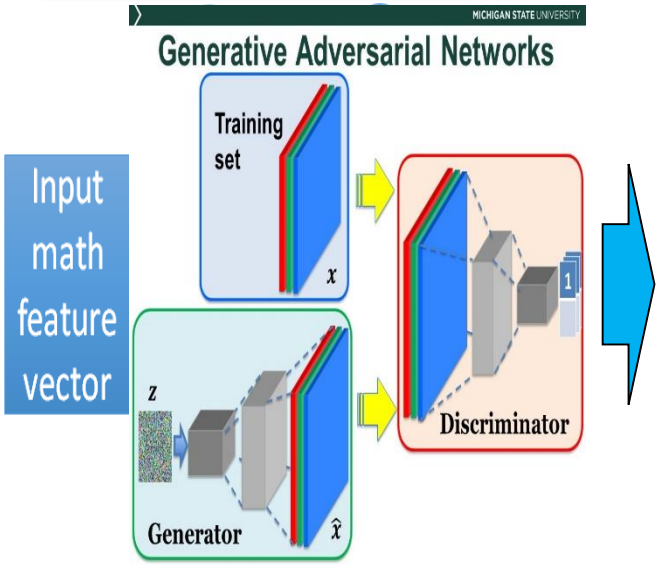


Drug Design Data Resource (D3R) Grand Challenge

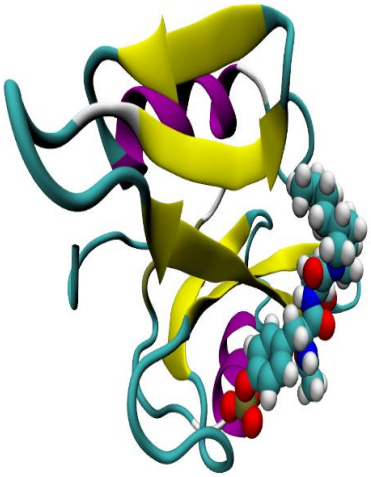
Given data



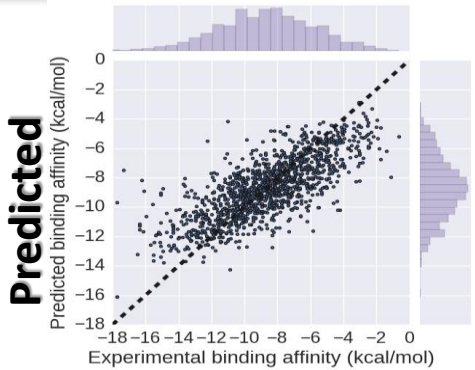
Math based GAN



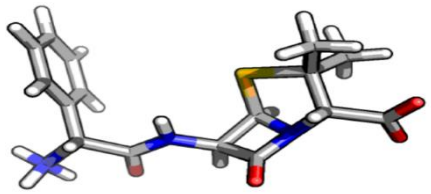
Predicted complex



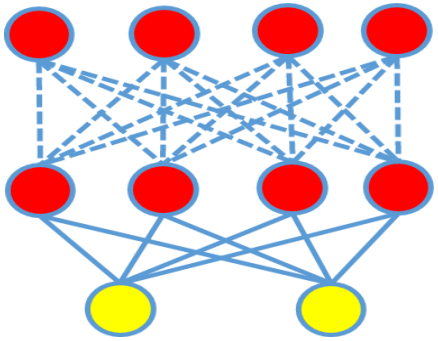
Final predictions to be compared with experiments



Experimental



Drug pose



(Nguyen et al, JCAMD, 2018)

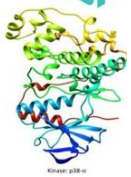
D3R Grand Challenge 3 (2017-2018)

(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


Zixuan Cang


Dr D Nguyen

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

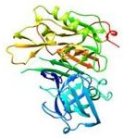

Dr. Kaifu Gao


Dr. D Nguyen

Ligand Based Scoring (No participation)

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

Free Energy Set 1/7 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

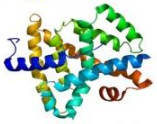
D3R Grand Challenge 2 (2016-2017)

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)

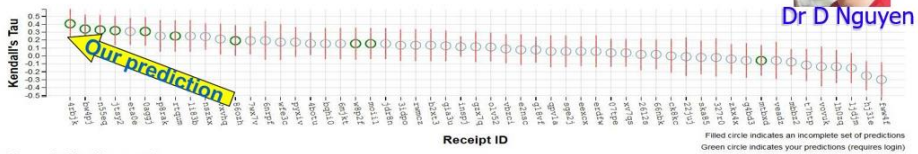



Dr D Nguyen

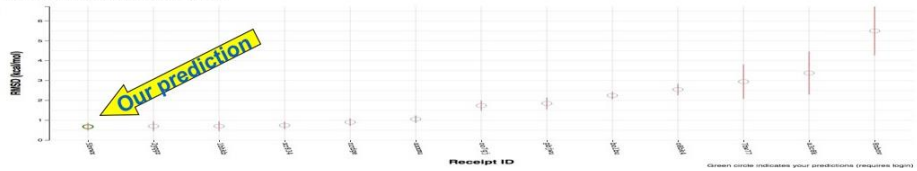
Grand Challenge 2

(Nguyen et al, JCAMD, 2018)

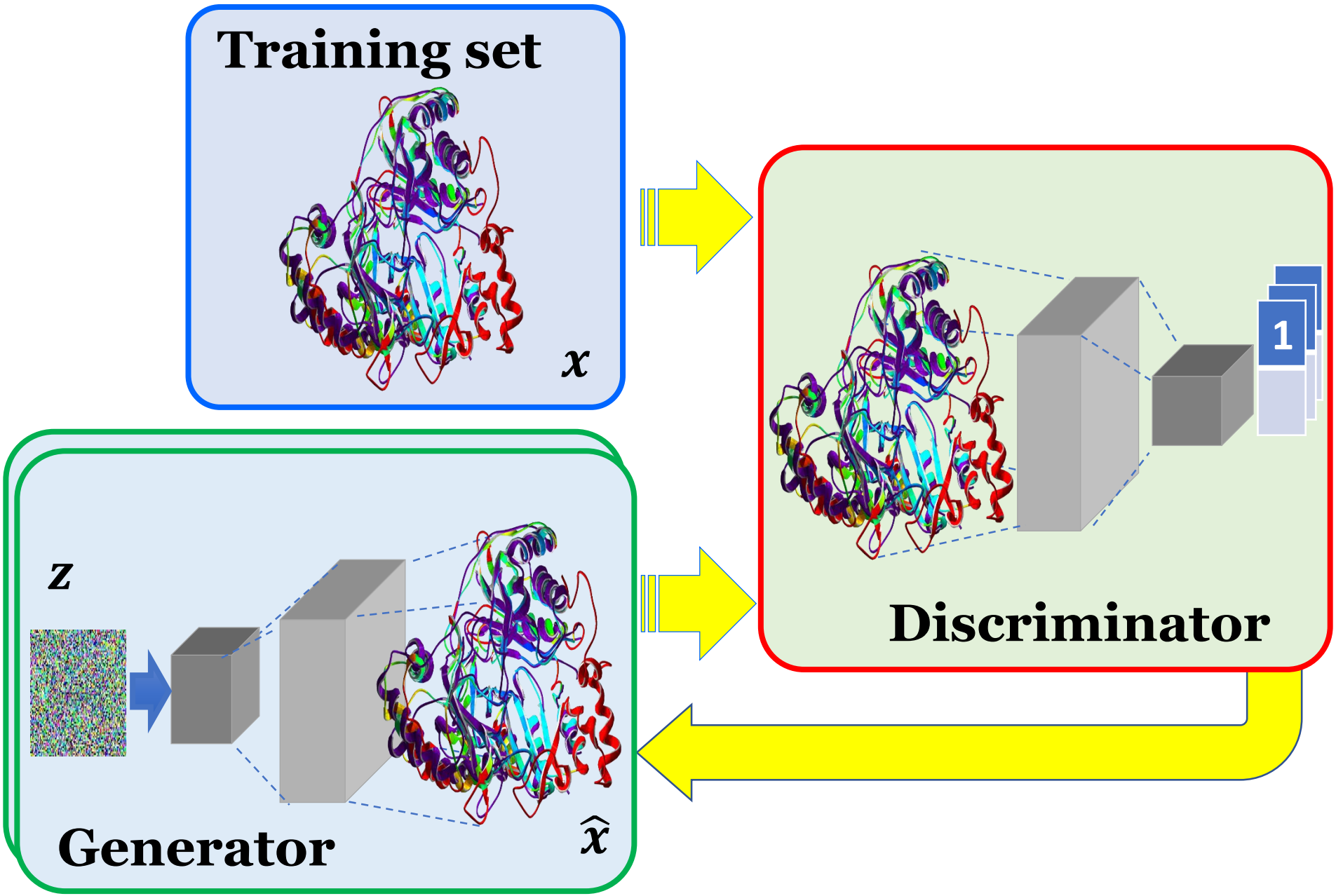
Free Energy Set 1 (Stage 2) - Kendall's Tau



Grand Challenge 2
Free Energy Set 1 (Stage 1) - RMSD



Generative Adversarial Networks for Drug Design



Ollivier-Ricci Curvature

- The topological fingerprints generated were able to be applied to different areas in biomolecular data analysis.
- Main research problem is in protein-ligand interactions, where we use the fingerprints in machine learning models to predict its binding affinities.
- The main aim in this research is thus to generate new fingerprints from other mathematical concepts to apply in biology. E.g. Ollivier-Ricci Curvature.

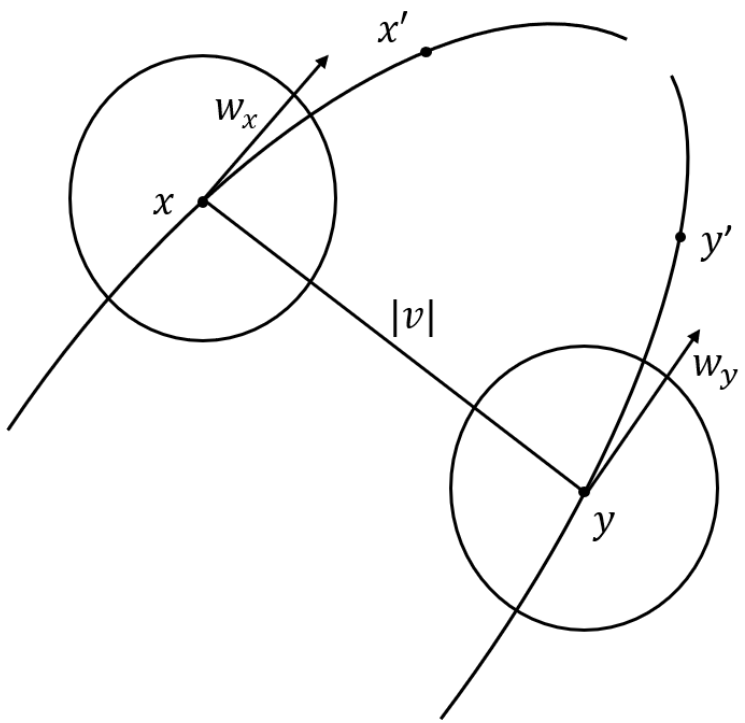
Ollivier-Ricci Curvature

- Previous results in ORC
  - Theoretical graph theory. (see e.g. [2, 6, 8, 9, 10])
  - Applications complex networks and internet topologies. [11]
  - Differentiating cancer genetic networks [20]
  - Analyzing fragilities in financial markets [21]
  - Community detections in large network models [12].
- Ricci Curvature was an important concept in Perelman's proof of Poincaré conjecture which made developments in Ricci Flow and remained as the only solved Millennium problem till today.
- Advanced mathematical graph based ML models also became popular in generating features for protein-ligand interactions [18, 32].
- Thus far, advanced mathematical graph based ML models based on Ollivier-Ricci Curvature have not been implemented for biomolecular data analysis.

Outline

Ricci Curvature

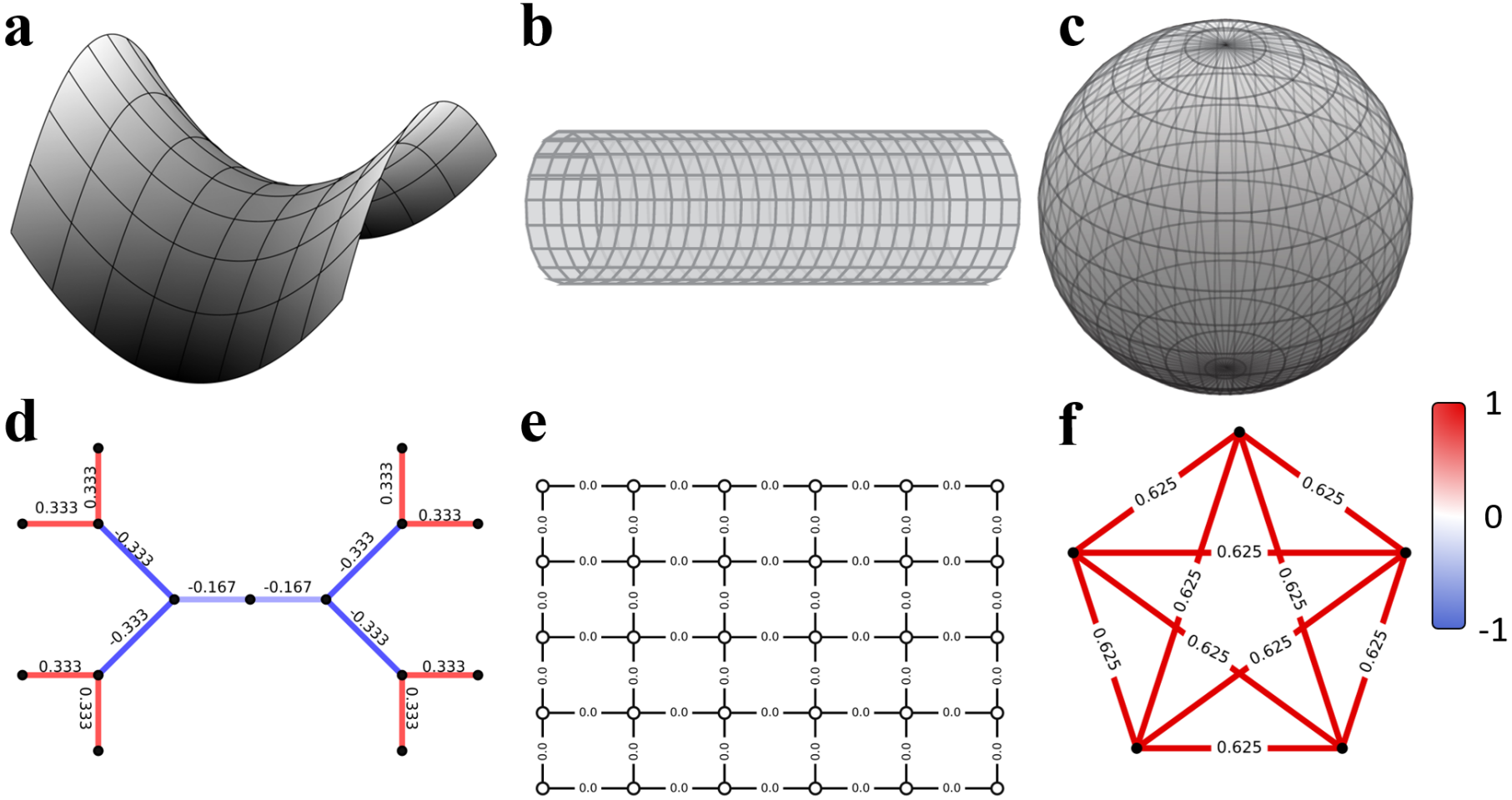
In sectional curvature, we consider a tangent vector w_x at x and another tangent vector w_y at y. If $|x'y'| < |xy| = |v|$, then the sectional curvature is positive.



Ricci Curvature of x in an n-dimensional Riemannian manifold is the $(n-1)$ times of mean sectional curvature at x over all possible directions w.

Discrete Ricci Curvature

In Discrete Ricci Curvature, we associate the geometric intuitions of negative, zero and positive curvatures with simple graphs such as trees, infinitely sized grid graph and the complete graph respectively.



Basic Definitions

In Ollivier-Ricci Curvature, the common probability distribution m_x which has an additional parameter α (see [10]), for a graph G is defined by the following definition. A typical value of α taken is 0.5.

Definition 2.1

For a vertex $x \in V$ with degree k, let $N_x = \{x_1, x_2, \cdots, x_k\}$ denote the set of neighbors of x. For any $\alpha \in [0, 1]$, the probability measure m_x is defined as

$$m_x(x_i) := \begin{cases} \alpha & \text{if } x_i = x. \\ (1-\alpha)/k & \text{if } x_i \in N_x. \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Basic Definitions

In Ollivier-Ricci Curvature, Ollivier Yann [19] incorporated the optimal transportation theory which involves the Wasserstein distance.

Definition 2.2 (Wasserstein Distance or Earth Mover Distance)

Let $\alpha \in [0,1]$ and the vertices in graph G be the metric space with two probability measures m_x and m_y. A transportation plan from m_x to m_y is a measure $\xi \in \prod(m_x, m_y)$ that is mass-preserving. The L^1 Wasserstein distance between m_x and m_y denoted by $W_1(m_x, m_y)$, is the minimum average travelling distance that can be achieved by any transportation plan:

$$W_1(m_x, m_y) = \inf_{\xi} \sum_{x_i \in V} \sum_{y_j \in V} d(x_i, y_j)\xi(x_i, y_j). \tag{2.2}$$

Definition 2.3 (Ollivier-Ricci Curvature with Idleness)

Let $\alpha \in [0,1]$ and the vertices in graph G be the metric space. Let m_x, m_y be two probability measures with respect to x and y in V respectively. For any two distinct points x, $y \in V$, the Ollivier-Ricci Curvature along the edge between x and y is defined as

$$c(x,y) := 1 - \frac{W_1(m_x, m_y)}{d(x,y)}, \tag{2.3}$$

where $W_1(m_x, m_y)$ is the Wasserstein distance between two probability measures m_x and m_y and $d(x,y)$ is the distance between x and y.

Note that the Definition 2.3 is defined on edges. The Ollivier-Ricci-Curvature of a vertex is then defined as the average of curvature of its adjacent edges.

Basic Definitions

In [11], the Ollivier-Ricci Curvature is computed via the optimizing Wasserstein Distance using Linear Programming (LP). First, let $\rho(x_i, y_j)$ be the proportion of "mass" transported from x_i to y_j, which is in an $m \times n$ matrix representing variables of $W_1(m_x, m_y)$. In this LP, the total mass is preserved while the objective function is to minimize the total transportation distance. The LP can written as follows:

$$\min \sum_{y_j \in V} \sum_{x_i \in V} d(x_i, y_j)\rho(x_i, y_j)m_x(x_i) \tag{2.4}$$

$$\text{such that} \sum_{y_j \in V} \rho(x_i, y_j) = 1, \quad 0 \leq \rho(x_i, y_j) \leq 1, \tag{2.5}$$

$$\sum_{x_i \in V} \rho(x_i, y_j)m_x(x_i) = m_y(y_j). \tag{2.6}$$

Theoretical Computations of ORCs

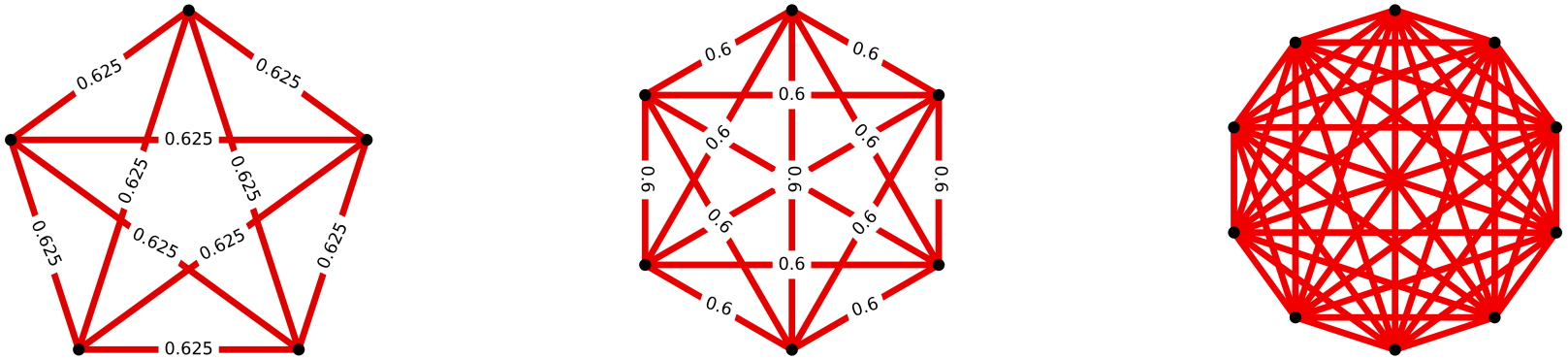
The discoveries in Ricci Curvatures has led to many results in theoretical graph theory (see [6, 10]) and many applications in complex networks and internet topologies also (see [11]). In order to understand the Ollivier-Ricci Curvature, we take a look at some simple graphs.

Example 2.4 (Complete Graph K_n)

We consider the complete graph K_n, $n \geq 3$. For any two vertices x and y in $V(K_n)$, the Ollivier-Ricci Curvature of xy is

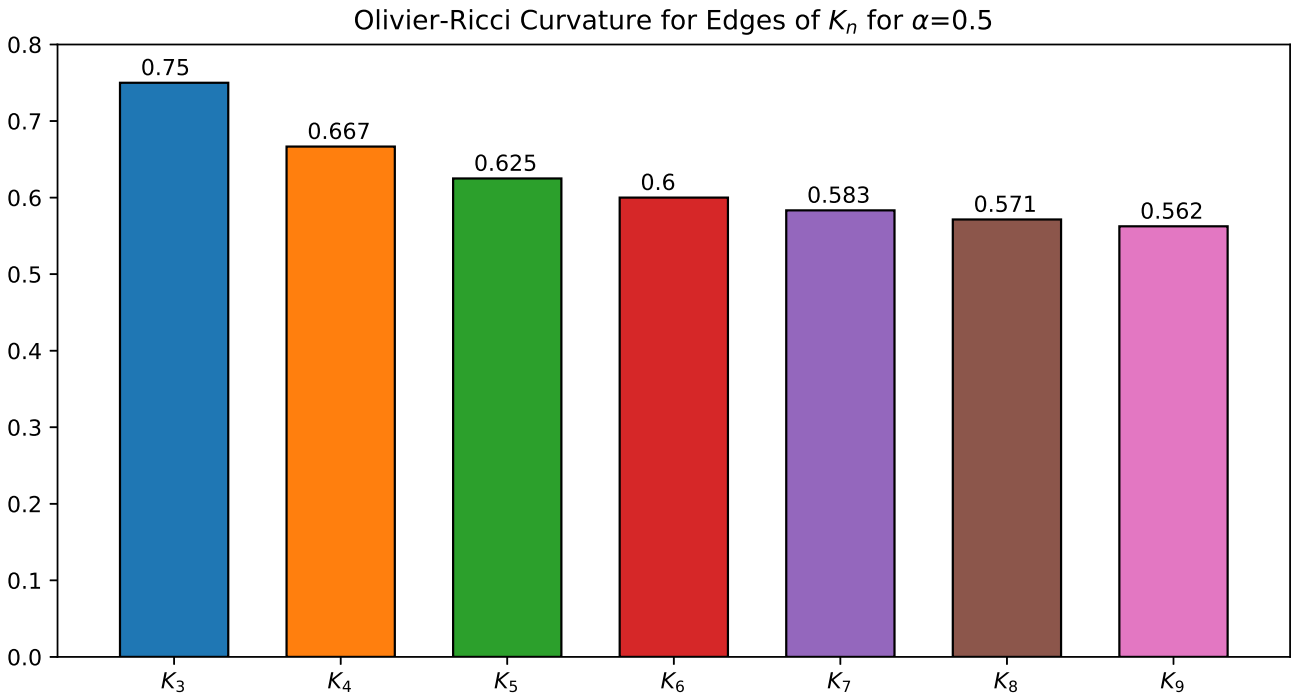
$$(1-\alpha)\frac{n}{n-1}. \tag{2.7}$$

Now, for $\alpha = 0.5$, the following are ORCs for some K_n.



(a) (b) (c)

Figure 4: Positive Olivier-Ricci Curvature of Complete Graphs



Theoretical Computations of ORCs

Example 2.5 (Trees)

We consider graphs without any cycles. i.e. Trees T. For any two vertices x and y in $V(\mathrm{T})$, the Ollivier-Ricci Curvature of xy is

$$\frac{1}{d_x} + \frac{1}{d_y} - 1, \tag{2.8}$$

where d_x, d_y are the degrees of vertices x and y respectively.

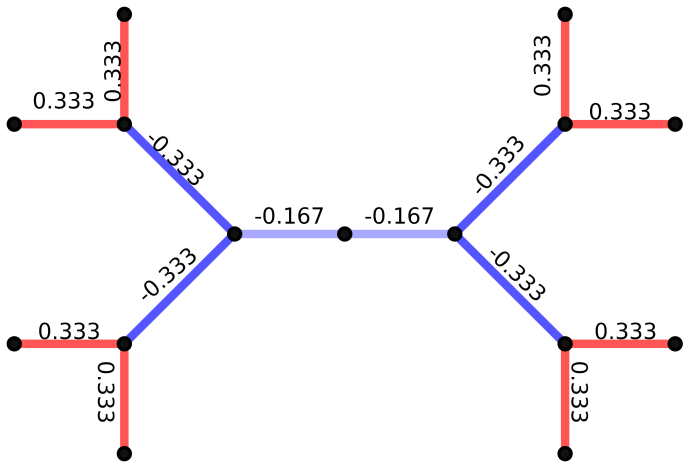


Figure 6: Negative Olivier-Ricci Curvature on Bridges

Example 2.6

An infinitely sized $m \times n$ grid graph has all edges with zero curvature. This is due to the fact that the cost of moving m_x to m_y is equal to $d(x, y)$.

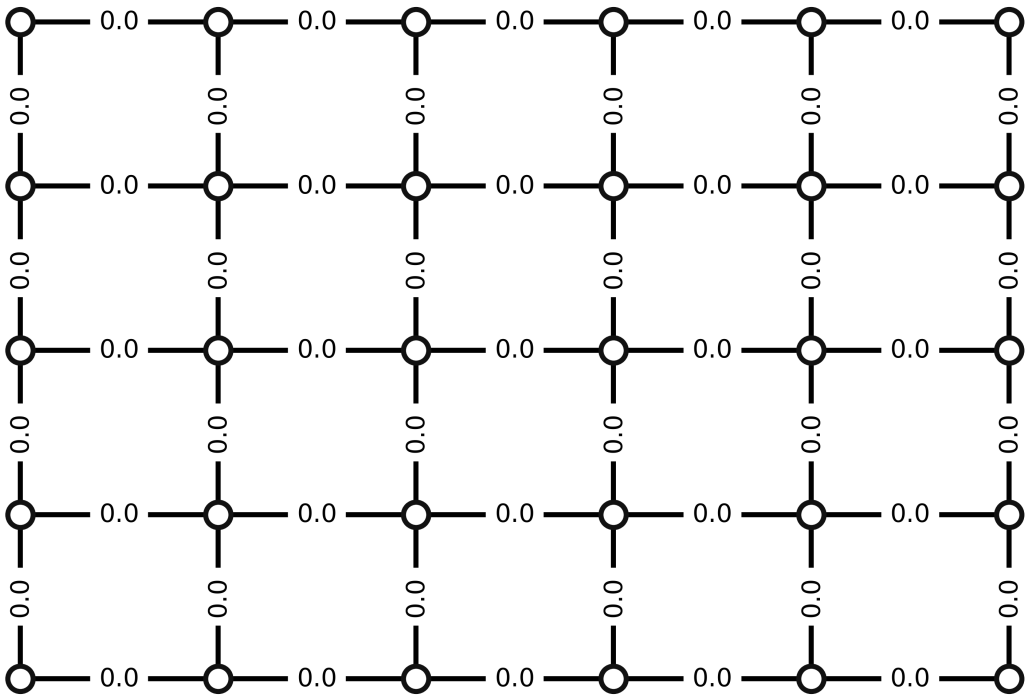
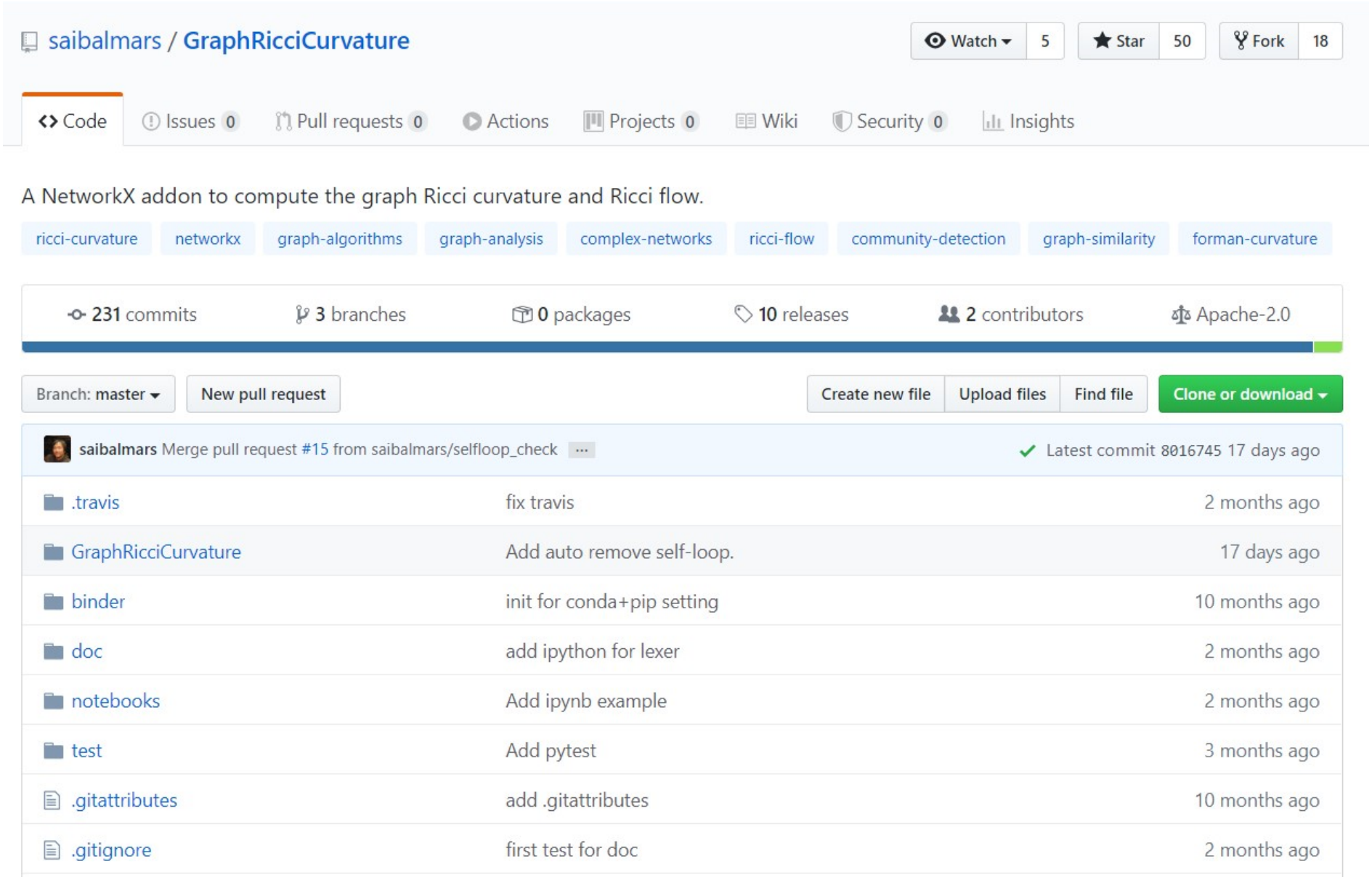


Figure 7: Zero Curvature

Ollivier-Ricci Curvature in Implementation



Ollivier-Ricci Curvature Based Molecular Representation

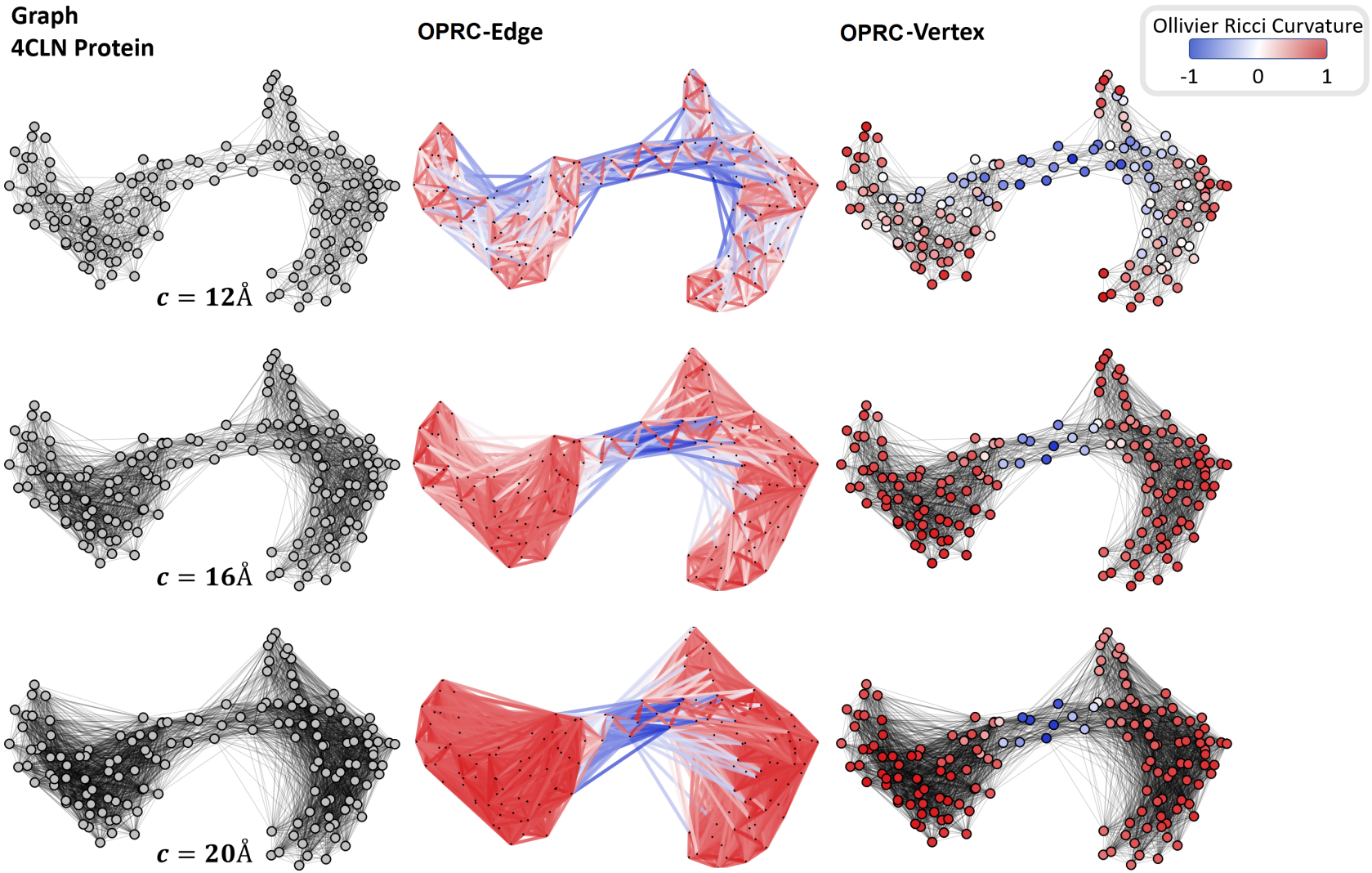


Figure 9: Illustration of C_α atoms of Calmodulin Protein (PDBID: 4CLN).

Ollivier-Ricci Curvature Based Molecular Representation

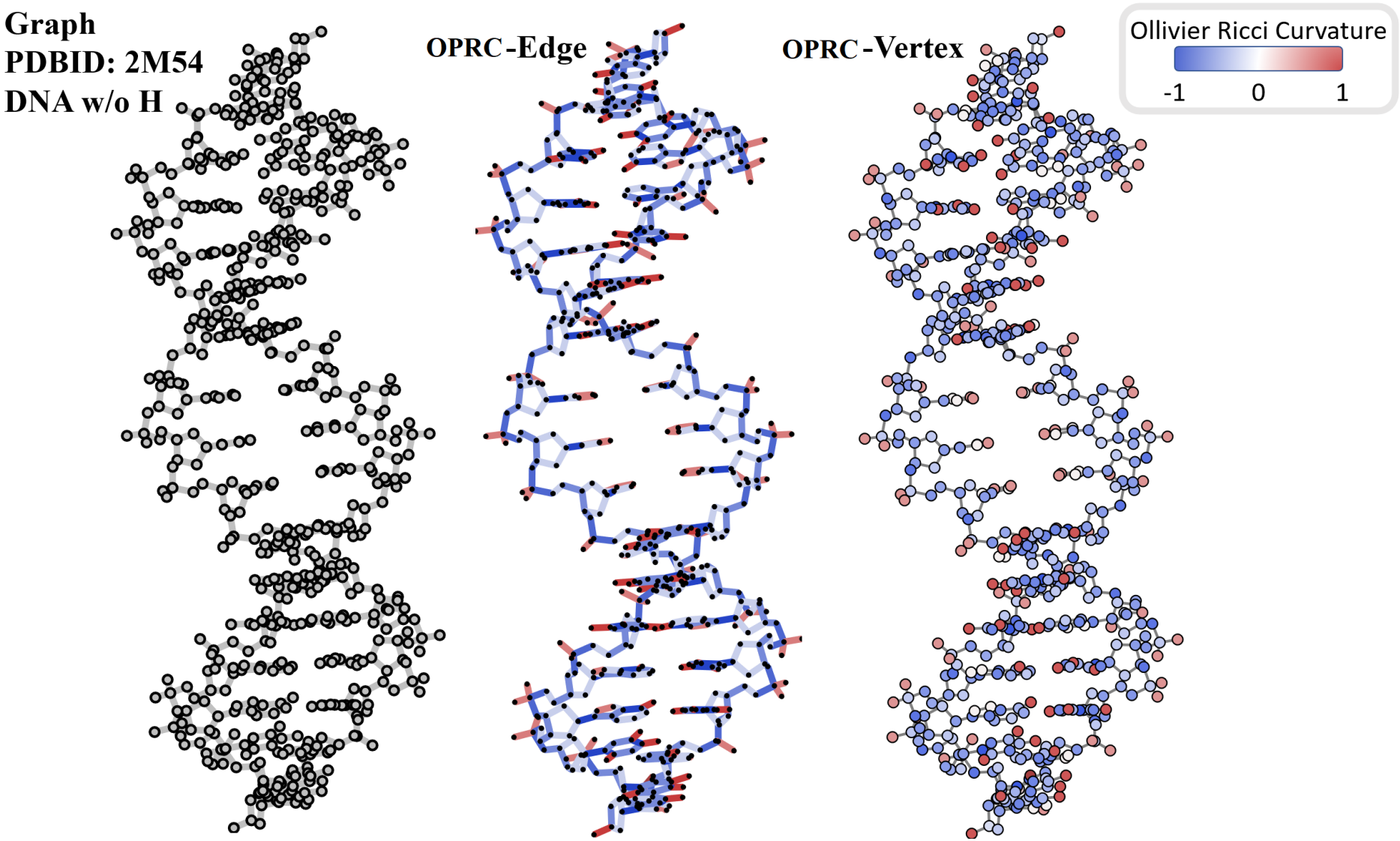


Figure 10

Ollivier-Ricci Curvature Based Molecular Representation

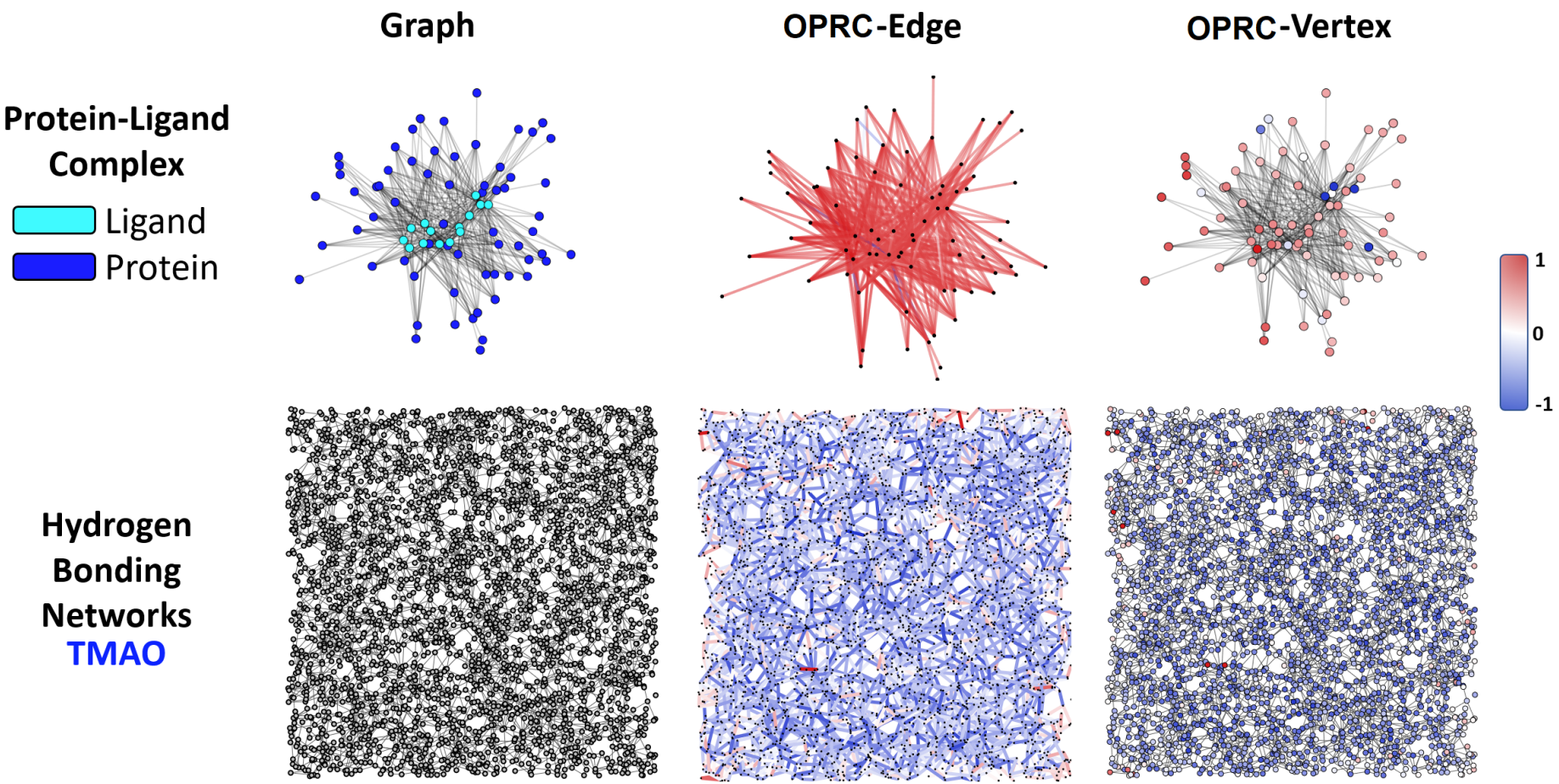


Figure 11

Outline

Application in Hydrogen-Bonding Networks

One of the applications of Persistent Homology is differentiating osmolytes, TMAO and urea, in hydrogen bonding networks. Persistent ORC can perform a similar application as Persistent Homology (see [28, 33]). Cutoff distance taken here is 4Å for both osmolytes in all frames and all ion concentrations.

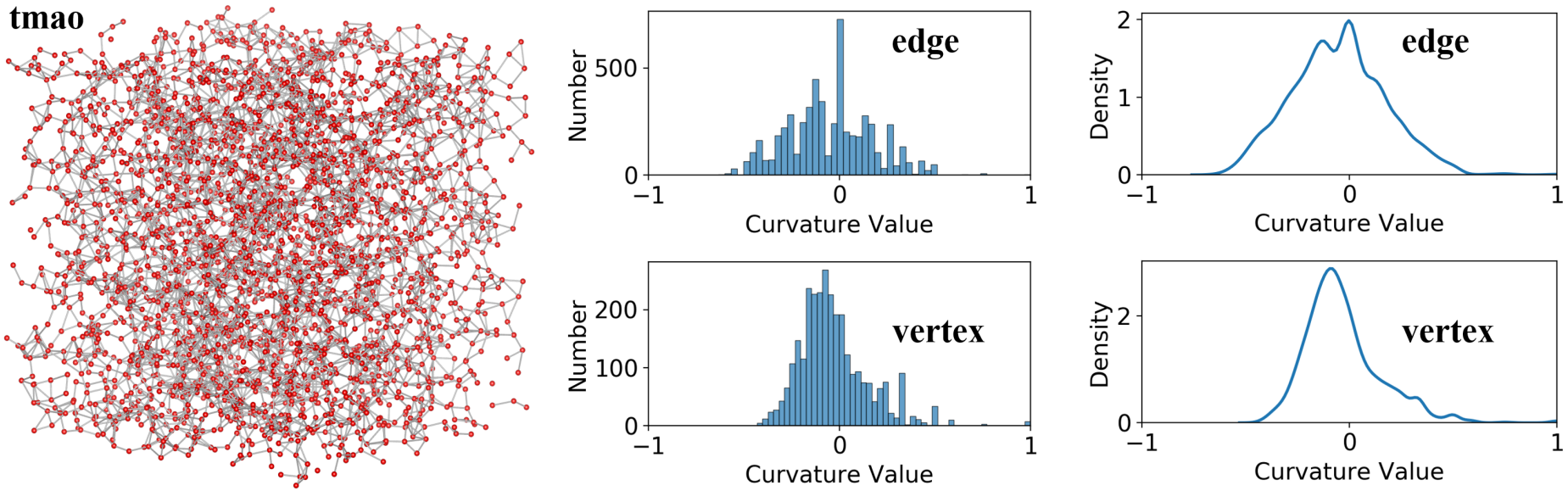


Figure 12: The Ollivier-Ricci curvatures for edges and vertices of the hydrogen-bonding networks for TMAO.

Application in Hydrogen-Bonding Networks

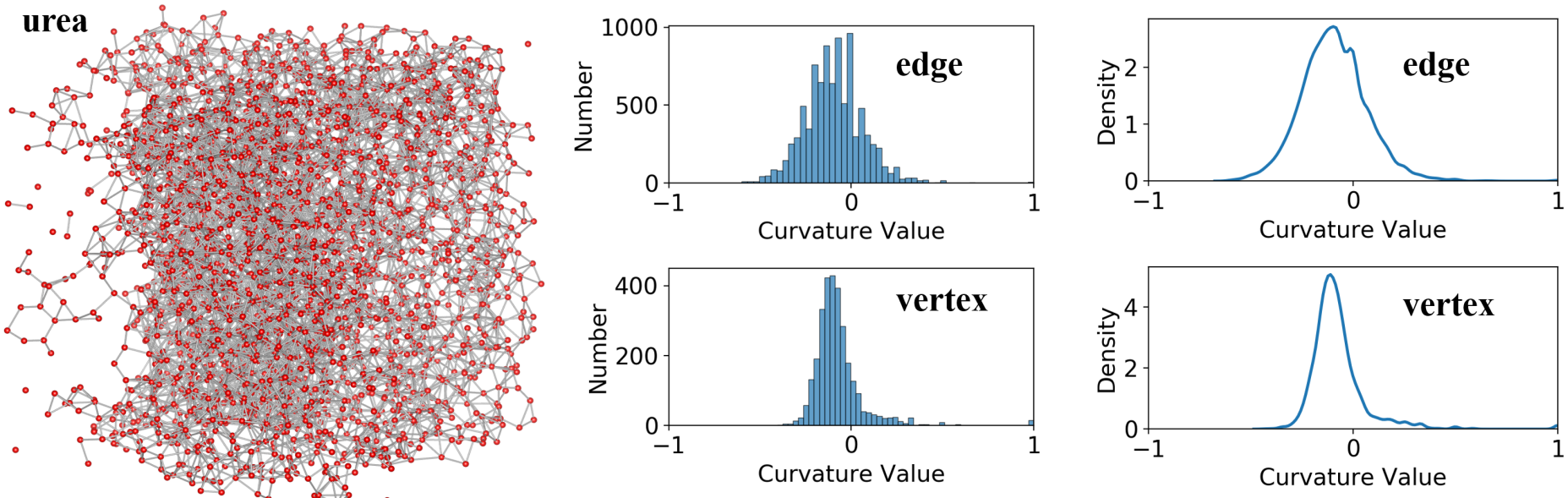


Figure 13: The Ollivier-Ricci curvatures for edges and vertices of the hydrogen-bonding networks for urea.

Application in Hydrogen-Bonding Networks

We generate the 101 frames of Molecular Dynamics (MD) simulations for each TMAO and urea in ion concentrations of 1M, 2M, ..., 8M.

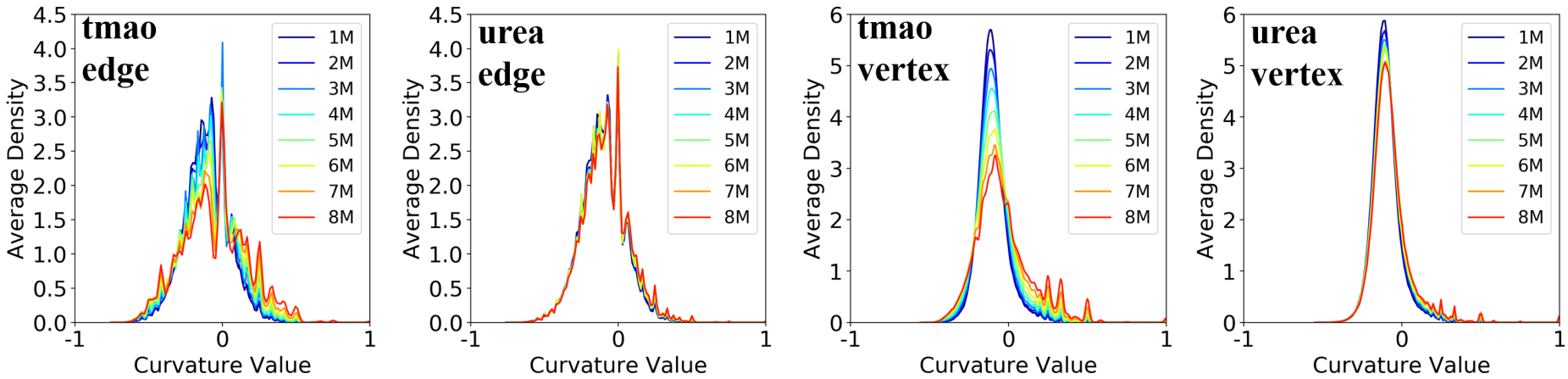


Figure 14: Illustration of average density distributions of edge and vertex ORCs of TMAO and urea.

OPRC Machine Learning Models for Drug Design

We consider element-specific (ES) groups for our Proteins and Ligands. We also consider the Interactive Distance Matrix (IDM).

Definition 3.1

Let α be a fixed atom type in $X \in$ Protein and β be a fixed atom type in $Y \in$ Ligand. We denote R_α and R_β as the atom coordinate sets for α and β respectively. Then for every i^{th} atom of α and j^{th} atom of β,

$$d_{\text{IDM}}(i,j) = \begin{cases} \|\mathrm{r}_i - \mathrm{r}_j\|, & \mathrm{r}_i \in \mathrm{R}_\alpha, \mathrm{r}_j \in \mathrm{R}_\beta \\ \infty, & \text{otherwise.} \end{cases}$$

where $\|\mathrm{r}_i - \mathrm{r}_j\|$ is the Euclidean distance between i^{th} atom of α and j^{th} atom of β.

For ES-IDM, we have the following ES groups.

- Proteins: {C, N, O, S}.
- Ligands: {C, N, O, S, P, F, Cl, Br, I}.

Persistent ORC in Protein-Ligand Binding

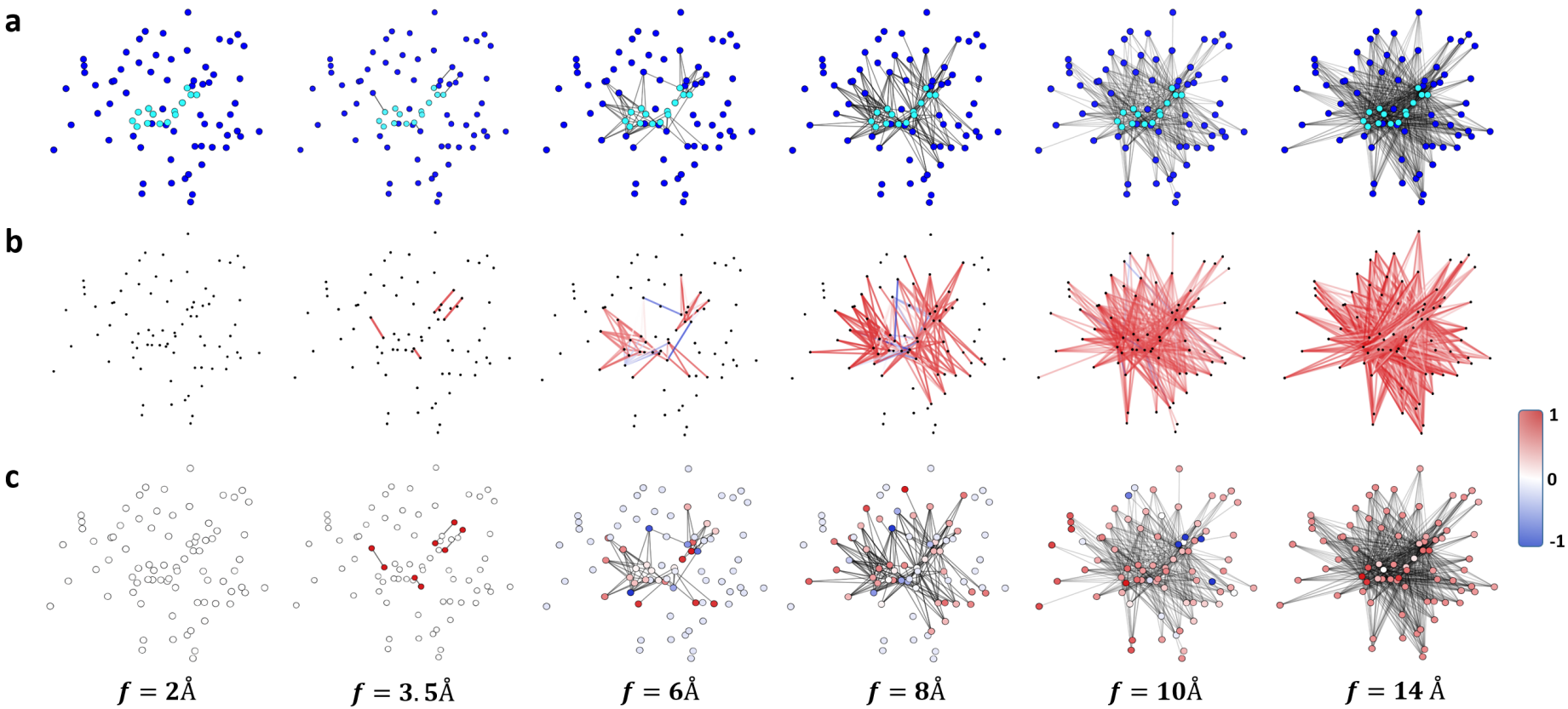


Figure 15: Filtration Process for PDBID: 1PXO O-C complex. The bipartite network changes and becomes increasingly connected between the carbon (ligand) and oxygen (protein) atoms. Eventually, the network stops changing when it becomes a complete bipartite network.

OPRC Model

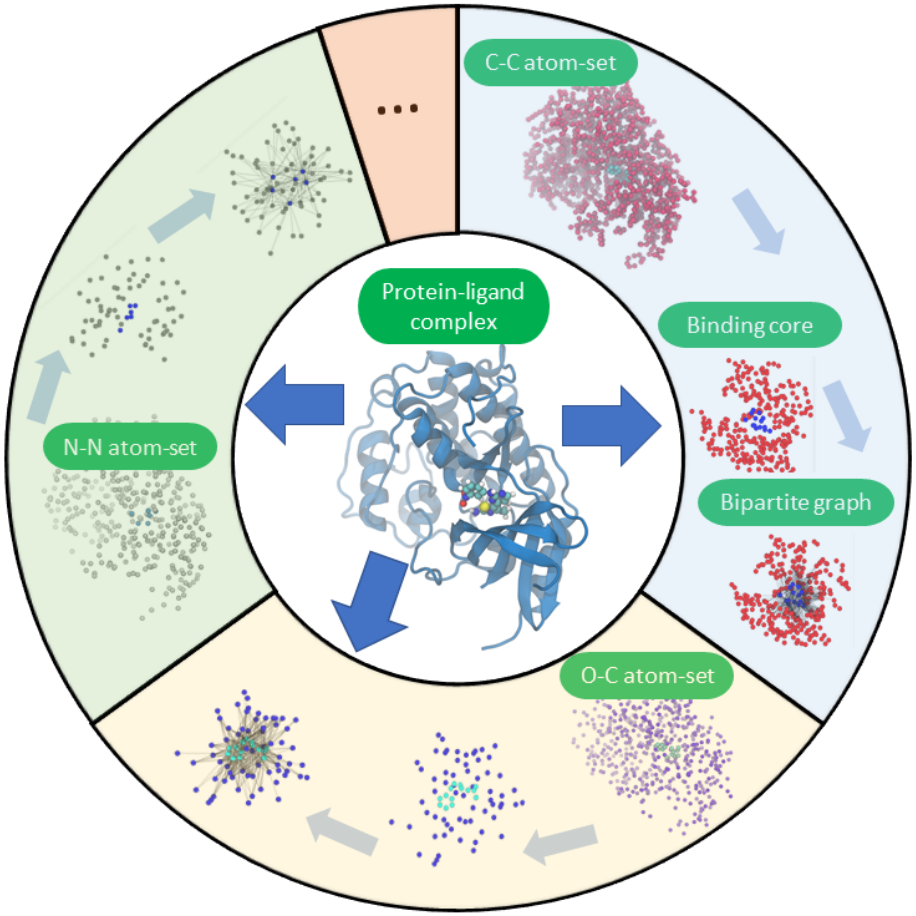


Figure 16: OPRC Model for protein-ligand binding. PDBID:1PXO.

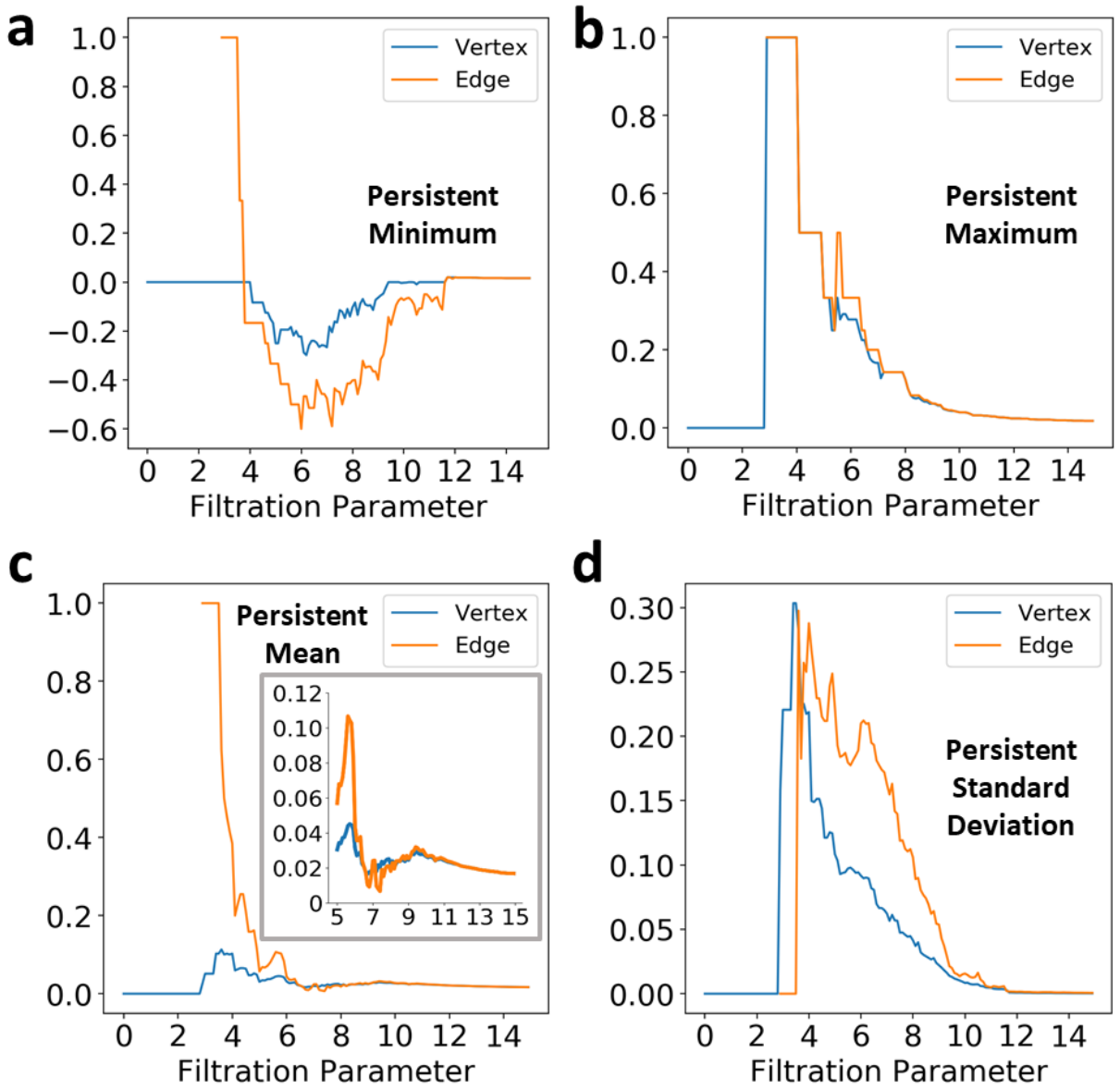
The set of values obtained is then converted into a set of features for machine learning.

OPRC Model Feature Generation

Let a distribution of ORC values be $(c_1, c_2, \cdots, c_N)$ where N is either the number of vertices or the number of edges. We use the following 10 statistical descriptors to convert our vertex and edge ORCs.

- Min $c_{\min}$
- Max $c_{\max}$
- Mean $\bar{c}$
- Standard Deviation σ_c
- Positive Sum $C_+ = \sum_{i=1}^{N} c_i \mathbb{1}_{\{c_i>0\}}$
- Absolute Deviation (AD) $C_{\text{AD}} = \sum_{i=1}^{N} |c_i - \bar{c}|$
- Total Sum of Squares $C^2_{\text{total}} = \sum_{i=1}^{N} c_i^2$
- Sum of Squares of Positive Terms $C^2_+ = \sum_{i=1}^{N} c_i^2 \mathbb{1}_{\{c_i>0\}}$
- Third Moment of AD $C^3_{\text{AD}} = \sum_{i=1}^{N} |c_i - \bar{c}|^3$
- Quasi-Wiener Index $C^* = \log\left(N \sum_{i=1}^{N^+} \frac{1}{c_i} + 1\right)$

Example of Persistent Attributes: 1PXO O-C Complex



OPRC Model Feature Generation

- Protein heavy atoms that have a $d_{\text{IDM}} \leq 10$Å with the ligand are considered.
- With 36 atom-atom type combinations, we filtrate each combination from 0Å to 15Å with gridsize 0.1.
- For each of 36×150 distribution of vertex and edge ORCs, we apply 10 statistical descriptors to convert our vertex and edge features (explained in the next slide).
- OPRC ES-IDM Model would have feature size $36 \times 150 \times 20$ for each protein-ligand complex.

Better Representation of Ligand Features

- Based on the comparisons of various atom combinations in ligand features in [3], we consider generating Persistent ORC features **solely** on ligand data to obtain further improvement for our model.
- For each atom combination in $\{\{C\}, \{C, N\}, \{C, O\}, \{C, N, O\}, \{C, N, O, F, P, Cl, Br, I\}\}$, we consider the ligand structures as graph networks and we compute the Persistent ORC for edges and vertices in ligand network.
- The 5 sets of features for each ligand is then combined to form a ligand multiscale feature **Lig**.
- The filtration is still taken from 0Å to 15Å with gridsize 0.1. By applying the same statistical descriptors, **Lig** has a feature size of 15000.
- We train and test the models of ES-IDM and ES-IDM + **Lig** features.

Databases and Parameter Settings

Version	Refined set	Training set	Core set (Test set)
v2007	1300	1105	195
v2013	2959	2764	195
v2016	4057	3772	285

Table 1: Breakdown of PDBbind v2007, v2013 and v2016 Databases.

No. of Estimators	Max Depth	Min. Sample Split	Learning Rate
40000	7	2	0.001

Loss Function	Max Features	Subsample Size	Repetition
Least Squares	Square Root	0.7	10 times

Table 2: GBT Parameters.

OPRC-GBT Model Results

	ES-IDM	**ES-IDM + Lig**
PDBbind v2007	0.820 (1.935)	0.821 (1.926)
PDBbind v2013	0.781 (2.035)	0.789 (2.010)
PDBbind v2016	0.835 (1.748)	0.838 (1.736)

Table 3: Results of OPRC-GBT Models for PDBbind v2007, v2013 and v2016 databases.

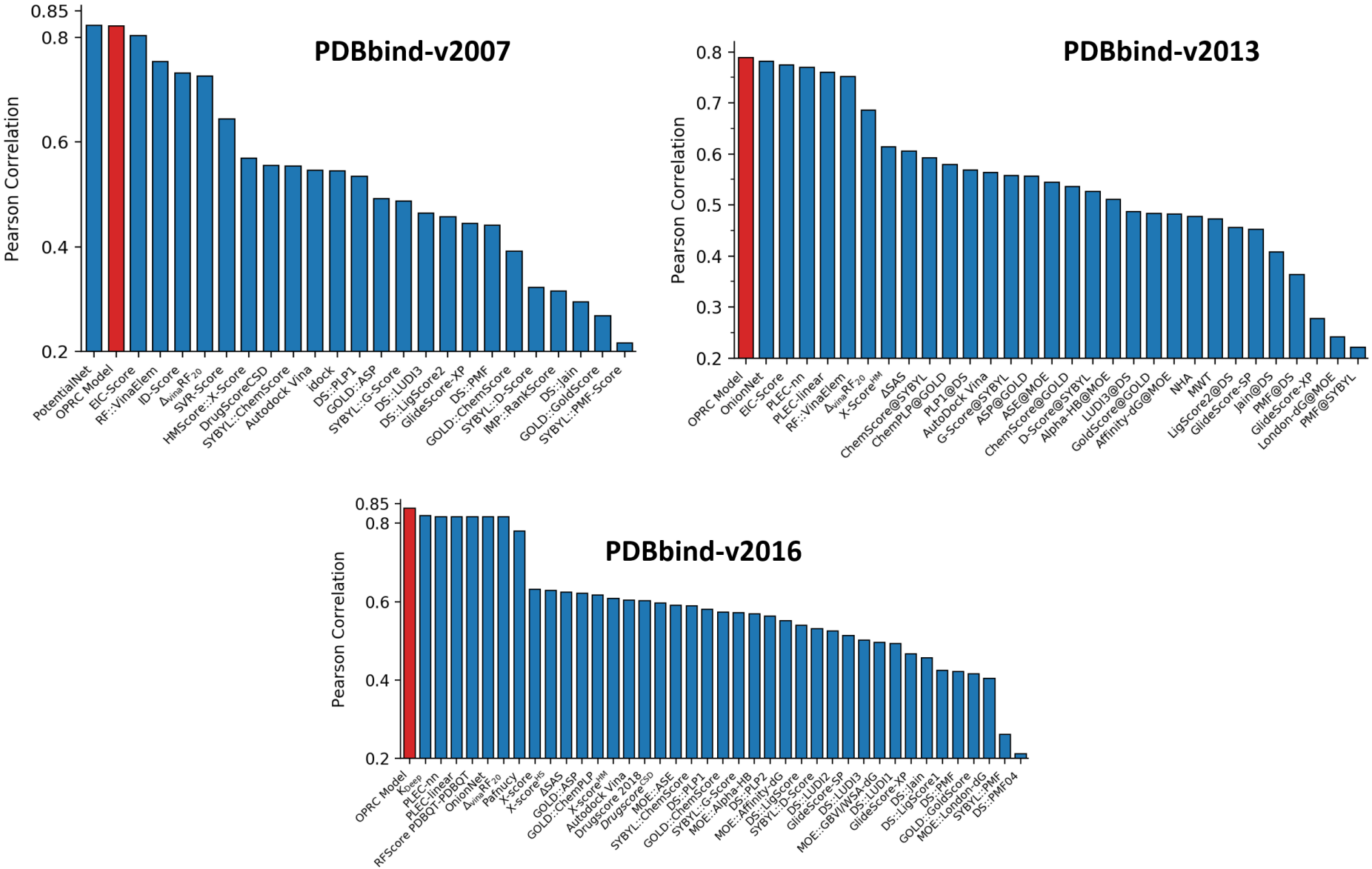


Figure 18: The comparison of our OPRC-GBT model with all traditional-molecular-descriptor based machine learning models.

Thank You

References I

[1] E. Appleboim, Y. Hyams, S. Krakovski, C. Sagiv & E. Saucan. 2013. The scale-curvature connection and its application to texture segmentation. Theory and Applications of Mathematics & Computer Science, 3(1), 38-55.

[2] B. B. Bhattacharya and S. Mukherjee, 2015. "Exact and asymptotic results on coarse Ricci curvature of graphs." Discrete Mathematics, 338(1), 23-42.

[3] Z. X. Cang, L. Mu, and G. W. Wei. 2018. "Representability of Algebraic Topology for Biomolecules in Machine Learning Based Scoring and Virtual Screening." PLoS Computational Biology, 14(1), e1005929.

[4] Z. X. Cang and G. W. Wei. 2017. "TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions." PLoS Computational Biology, 13(7), e1005690.

[5] Z. X. Cang and G. W. Wei. 2018. "Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction." International Journal for Numerical Methods in Biomedical Engineering, 34(2), e2914.

References II

[6] H. J. Cho and S. H. Paeng. 2013. "Ollivier's Ricci curvature and the coloring of graphs." European Journal of Combinatorics, 34(5), 916-922.

[7] Z. Gaieb, C. D. Parks, M. Chiu, H. Yang, C. Shao, W. P. Walters and T. Mirzadegan. 2019. "D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings." Journal of Computer-Aided Molecular Design, 33(1), 1-18.

[8] J. Jost and S. Liu, 2014. Ollivier's Ricci curvature, local clustering and curvature-dimension inequalities on graphs. Discrete & Computational Geometry, 51(2), 300-322.

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

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

[11] C. C. Ni, Y. Y. Lin, J. Gao, X. D. Gu and E. Saucan. 2015. "Ricci curvature of the Internet topology." In 2015 IEEE Conference on Computer Communications (INFOCOM) (pp. 2758-2766). IEEE.

References III

[12] C. C. Ni, Y. Y. Lin, F. Luo and J. Gao. 2019. "Community detection on networks with ricci flow." Scientific reports, 9(1), 1-12.

[13] D. D. Nguyen, Z. X. Cang, K. Wu, M. Wang, Y. Cao and G. W. Wei. 2019. "Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges." Journal of Computer-Aided Molecular Design, 33(1), 71-82.

[14] D. D. Nguyen, Z. X. Cang, and G. W. Wei. "A Review of Mathematical Representations of Biomolecular Data." Physical Chemistry Chemical Physics, 2020.

[15] D. D. Nguyen, K. Gao, M. Wang and G. W. Wei. 2020. "Mathdl: Mathematical deep learning for d3r grand challenge 4." Journal of Computer-Aided Molecular Design, 34(2), 131-147.

[16] D. D. Nguyen, K. L. Xia and G. W. Wei. 2016. "Generalized flexibility-rigidity index. The Journal of chemical physics." 144(23), 234106.

References IV

[17] D. D. Nguyen, T. Xiao, M. Wang and G. W. Wei. 2017. "Rigidity Strengthening: A Mechanism For Protein–ligand Binding." Journal of Chemical Information and Modeling, 57(7), 1715-1721.

[18] D. D. Nguyen and G. W. Wei. 2019. "AGL-Score: Algebraic Graph Learning Score for Protein–Ligand Binding Scoring, Ranking, Docking, and Screening." Journal of Chemical Information and Modeling, 59(7), 3291-3304.

[19] Y. Ollivier. 2009. "Ricci curvature of Markov chains on metric spaces." Journal of Functional Analysis, 256(3), 810-864.

[20] R. S. Sandhu, T. T. Georgiou, E. Reznik, L. Zhu, I. Kolesov, Y. Senbabaoglu and A. Tannenbaum. 2015. "Graph curvature for differentiating cancer networks." Scientific reports, 5(1), 1-13.

[21] R. S. Sandhu, T. T. Georgiou and A. R. Tannenbaum. 2016. "Ricci curvature: An economic indicator for market fragility and systemic risk." Science advances, 2(5), e1501495.

[22] E. Saucan, A. Samal & J. Jost. 2020. A Simple Differential Geometry for Complex Networks. arXiv preprint arXiv:2004.11112.

References V

[23] B. Wang, C. Wang, K. Wu and G. W. Wei. 2018. "Breaking the polar-nonpolar division in solvation free energy prediction." Journal of Computational Chemistry, 39(4), 217-233.

[24] B. Wang, Z. Zhao and G. W. Wei. 2016. "Automatic parametrization of non-polar implicit solvent models for the blind prediction of solvation free energies." The Journal of Chemical Physics, 145(12), 124110.

[25] G. W. Wei. 2017. "Mathematics at the Eve of a Historic Transition in Biology." Computational and Mathematical Biophysics, 5(1).

[26] G. W. Wei. 2017. "Persistent Homology Analysis of Biomolecular Data." Journal of Computational Physics. 305:276-299.

[27] K. Wu and G. W. Wei. 2018. "Quantitative toxicity prediction using topology based multitask deep neural networks." Journal of Chemical Information and Modeling, 58(2), 520-531.

References VI

[28] K. L. Xia, D. V. Anand, S. Shikar and Y. Mu. 2019. "Persistent homology analysis of osmolyte molecular aggregation and their hydrogen-bonding networks." Physical Chemistry Chemical Physics, 21(37), 21038-21048.

[29] K. L. Xia, X. Feng, Y. Y. Tong and G. W. Wei. 2015. "Persistent homology for the quantitative prediction of fullerene stability." Journal of Computational Chemistry, 36, 408-422.

[30] K. L. Xia. 2018. "Multiscale virtual particle based elastic network model (MVP-ENM) for biomolecular normal mode analysis." Physical Chemistry Chemical Physics, 20(1), 658-669.

[31] K. L. Xia and G. W. Wei. 2015. "Multiresolution topological simplification." Journal of Computational Biology, 22(9), 1-5.

[32] M. Zhenyu, K. L. Xia. 2020. "Persistent spectral based machine learning (PerSpect ML) for drug design." arXiv preprint arXiv:2002.00582.

[33] M. Zhenyu, D. V. Anand, Y. Lu, J. Wu, & K. L. Xia. 2020. Weighted persistent homology for biomolecular data analysis. Scientific reports, 10(1), 1-15.