

Hypergraph based Persistent Cohomology for molecular representations in drug design

Xiang Liu

Supervised by Asst. Prof. Kelin Xia(NTU)

Nankai University

2020-12-03

Outline

1 Background

2 Model construction

3 Application

1 Background

2 Model construction

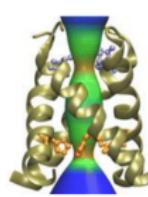
3 Application

Drug design and discovery



General flow

- 1) Disease identification (**physiology**)
- 2) Target hypothesis (**biochem./mole. biol.**)
- 3) Virtual screening: 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 test
- 7) Optimize drug's efficacy, pharmacokinetics, and pharmacodynamics properties (**quantitative systems pharmacology**)



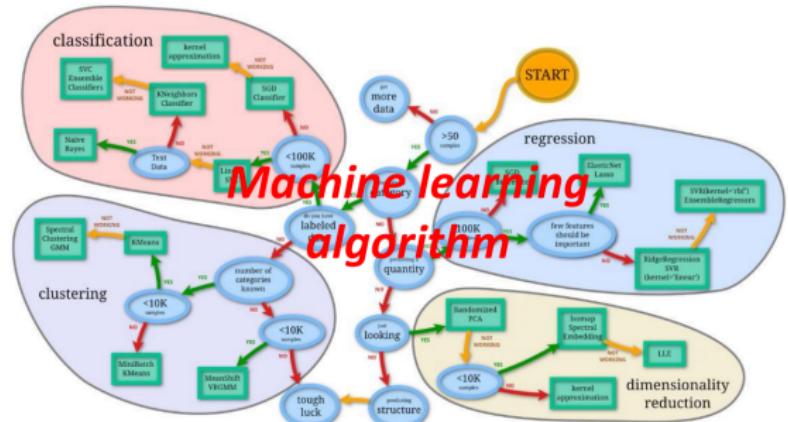


NATURE | NEWS

Gene data to hit milestone

With close to one million gene-expression data sets

How to deal with big data ?



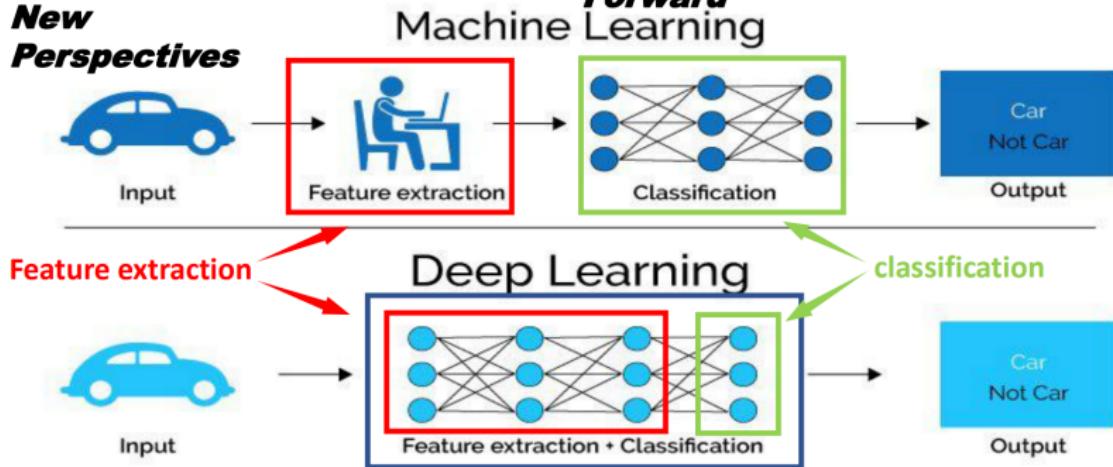
Feature extraction and feature learning

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

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

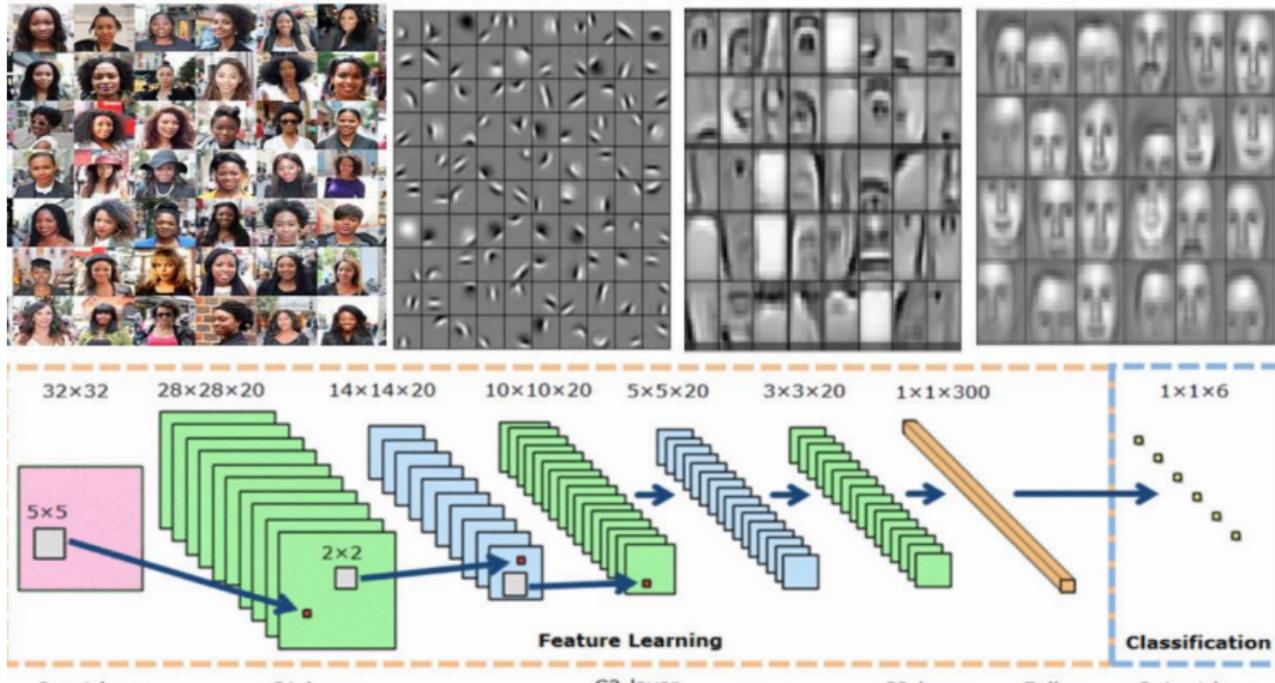
"The deep learning research aims at discovering learning algorithms that discover multiple levels of distributed representations..."

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



Deep learning

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



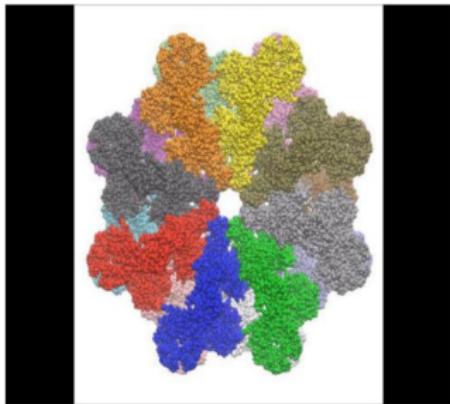
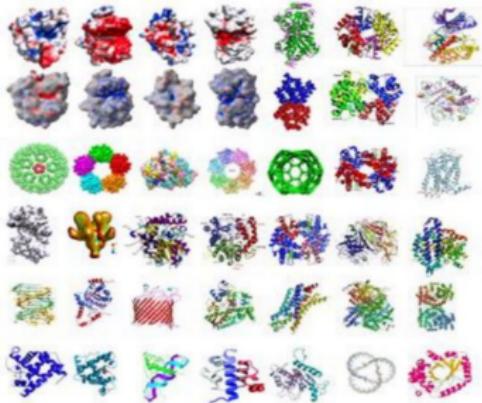
How to do deep learning for 3D biomolecular data?

Obstacles for deep learning of 3D biomolecules:

- Geometric dimensionality: R^{3N} , where $N \sim 5500$ for a protein.
- Machine learning dimensionality: $> 1024^3 m$, where m is the number of atom types in a protein.
- Molecules have different sizes --- non-scalable.
- Complexity: biochemistry & biophysics

Solution:

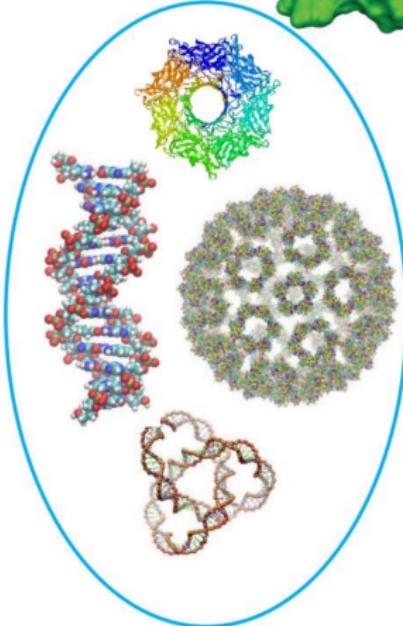
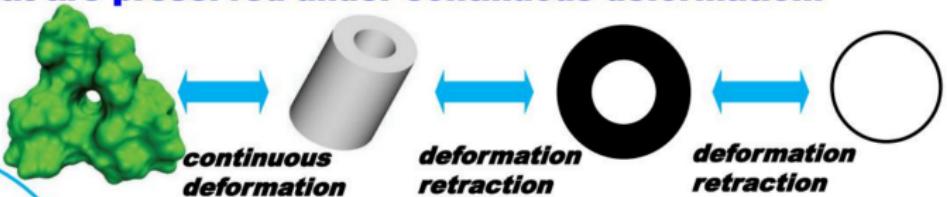
- Dimensionality reduction & unification (scalability)
- Topological simplification/geometric simplification/graph theory simplification



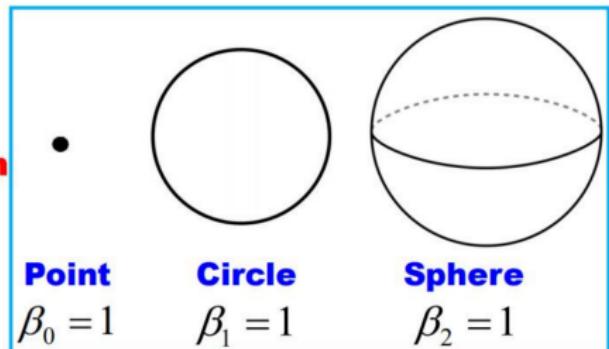
Topological invariant--Betti number

Properties that are preserved under continuous deformation!

Homotopy equivalent:



Topological simplification

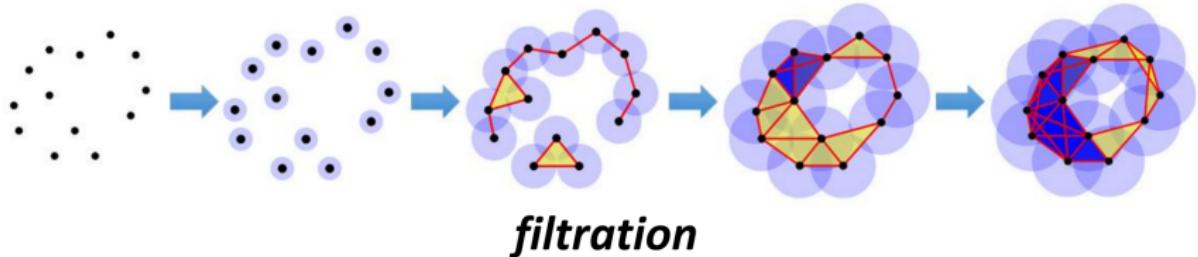


β_0 is the number of connected components

β_1 is the number of tunnels or circles

β_2 is the number of voids or cavities

Topological data analysis -- *Persistent homology*



Chain group: $C_n(K, \mathbb{Z}_p)$

Boundary operator:

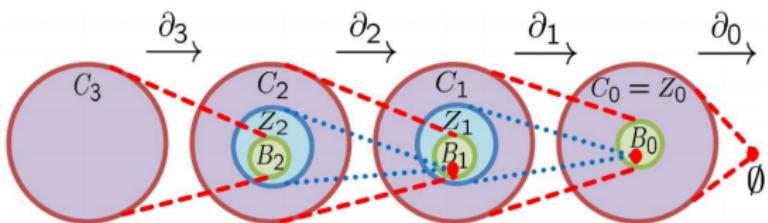
$$\partial_n(\sigma_n) = \sum_{i=0}^n (-1)^i \{v_0, v_1, \dots, \hat{v_i}, \dots, v_n\}$$

$$Z_n = \text{Ker} \partial_n$$

$$B_n = \text{Im} \partial_{n+1}$$

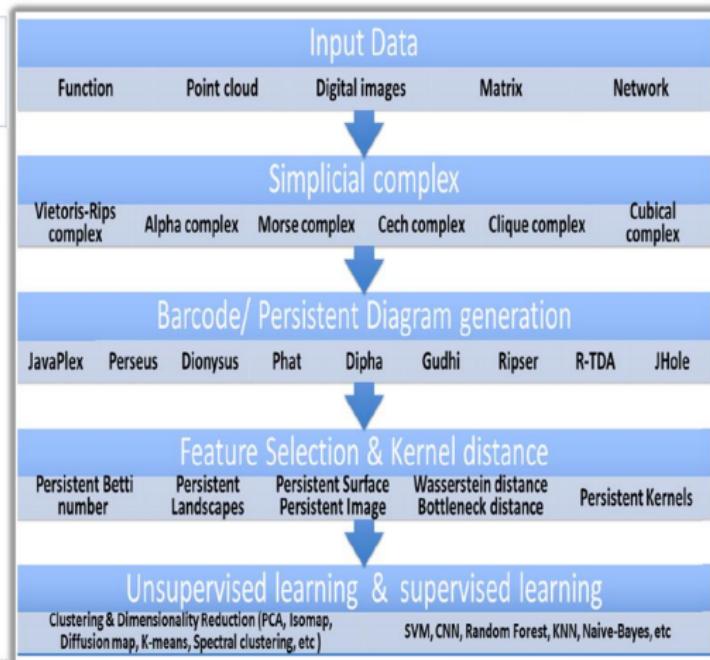
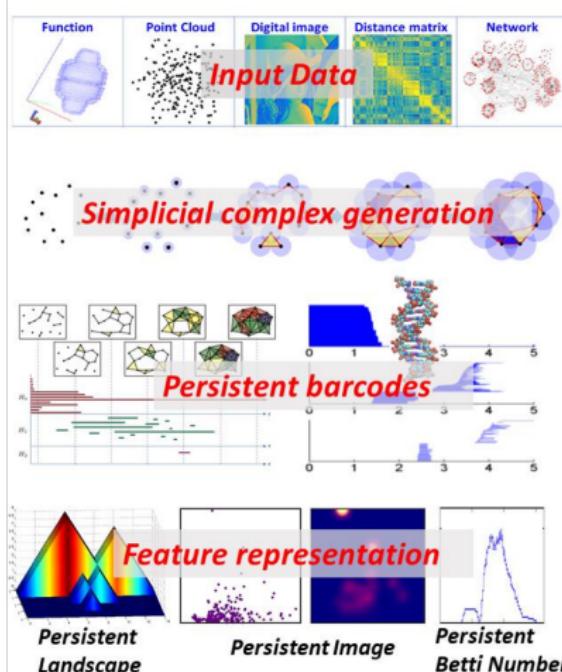
$$H_n = Z_n / B_n$$

$$\beta_n = \text{rank}(H_n)$$



TDA based machine learning models

(Pun, Xia and Lee, arXiv, 2019)

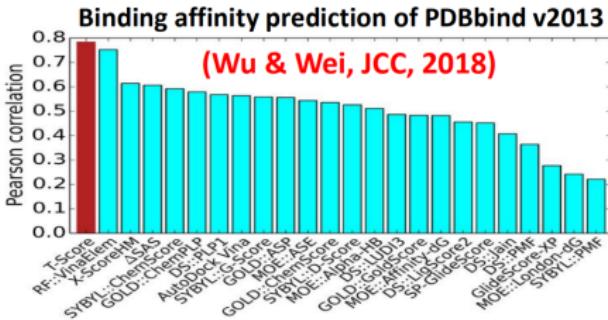
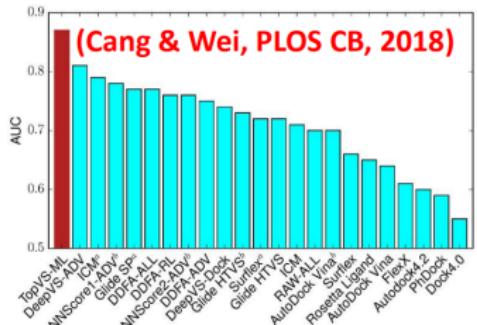


Recent progress in TDA based drug design

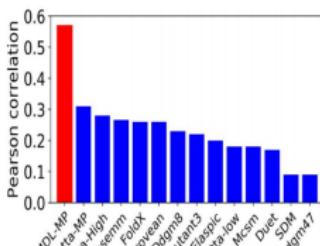
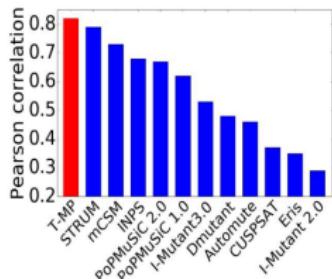
Guowei Wei
MSU, USA



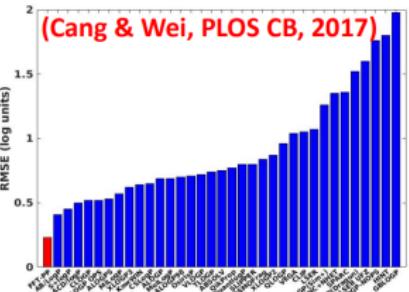
DUD database 128374 protein-ligand/decoy pairs



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



Prediction RMSD of logP(star set)



Recent progress in TDA based drug design



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 \(partials\)](#)

Affinity Rankings excluding Kds > 10 μ M

Cathepsin Stage 1

[Scoring \(partials\)](#)

[Free Energy Set](#)

[VEGFR2](#)

[Scoring \(partials\)](#)

[JAK2 SC3](#)

[Scoring](#)

[Free Energy Set](#)

Active / Inactive Classification

[VEGFR2](#)

[Scoring \(partials\)](#)

[JAK2 SC3](#)

[Scoring](#)

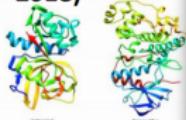
[Free Energy Set](#)

Affinity Rankings for Cocrystallized Ligands

Cathepsin Stage 1

[Scoring \(partials\)](#)

[Free Energy Set](#)



[Cathepsin Stage 2](#)

[Scoring \(partials\)](#)

[Free Energy Set](#)

[JAK2 SC2](#)

[Scoring \(partials\)](#)

[TIE2](#)

[Scoring \(partials\)](#)

[Free Energy Set 1](#)

[Free Energy Set 2](#)

[p38- \$\alpha\$](#)

[Scoring \(partials\)](#)

[ABL1](#)

[Scoring \(partials\)](#)

[p38- \$\alpha\$](#)

[Scoring \(partials\)](#)

[ABL1](#)

[Scoring \(partials\)](#)

We Team's performance at D3R Grand Challenge

D3R Grand Challenge 4 (2018-2019)

Pose Predictions

BACE Stage 1A

[Pose Predictions \(Partials\)](#)

BACE Stage 1B

[Pose Prediction \(Partials\)](#)

Affinity Predictions

Cathepsin Stage 1

[Combined Ligand and Structure Based Scoring](#)

Ligand Based Scoring (No participation)

Structure Based Scoring

[Free Energy Set](#)

[Scoring \(Partials\)](#)

[Scoring \(Partials\)](#)

[Scoring \(Partials\)](#)

[BACE Stage 1](#)

[Combined Ligand and Structure \(No participation\)](#)

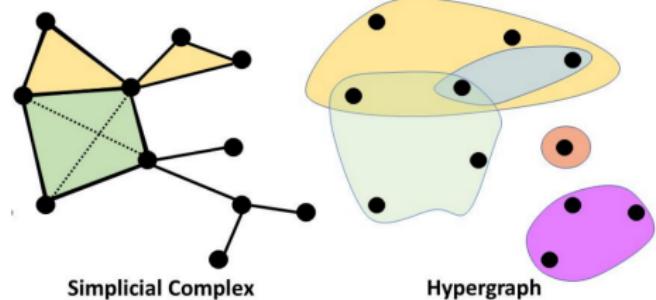
[Ligand Based Scoring \(Partials\) \(No participation\)](#)

[Structure Based Scoring \(Partials\) \(No participation\)](#)

[Free Energy Set \(No participation\)](#)

[Structure Based Scoring \(Partials\)](#)

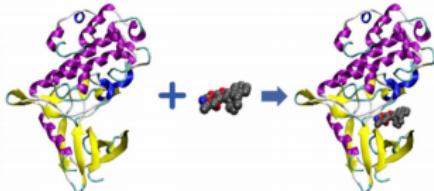




**hypergraph is a generalization
of simplicial complex,**

**how about doing persistence
by hypergraph?**

Recent progress in hypergraph



*Biomolecular interactions
characterized by hypergraph*

Jie Wu
Math, HEBNU



The embedded homology of
hypergraphs and applications
(Bressan, Li, Ren & Wu, 2016)

Operators on random
hypergraphs and random
simplicial complexes
(Ren, Wu & Wu, 2017)

A discrete Morse Theory for
hypergraphs (Ren, Wang, Wu
& Wu, 2018)

Hodge Decompositions for
Weighted Hypergraphs
(Ren, Wu & Wu, 2018)

Stability of persistent
homology for hypergraphs
(Ren & Wu, 2020)

① Background

② Model construction

③ Application

Biomolecular hypergraph representation

The key idea is to describe atom interaction relations as hyperedges.

Take the protein-ligand interactions as an example. We denote the set of protein atoms, the set of ligand atoms, the hypergraph and the vertex set of the hypergraph as $V_P = \{v_i; i = 1, 2, \dots, N_p\}$, $V_L = \{v_j; j = 1, 2, \dots, N_l\}$, \mathcal{H} and $V_{\mathcal{H}}$ respectively.

- ① For the vertex set, we take $V_{\mathcal{H}} = V_L \cup V_P$
- ② For the hyperedge set, we define an n-hyperedge in \mathcal{H} as

$$\sigma^n = \begin{cases} \{\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_n\}; \exists i, j, \mathbf{v}_i \in V_P, \mathbf{v}_j \in V_L, & n > 0 \\ \{\mathbf{v}_0\}; \mathbf{v}_0 \in V_{\mathcal{H}}, & n = 0. \end{cases} \quad (1)$$

For simplicity, we denote the hypergraph $(\mathcal{H}, V_{\mathcal{H}})$ as \mathcal{H} .

Example

Remark: for each n-hyperedge($n > 0$), there must be two vertices come from different molecules.

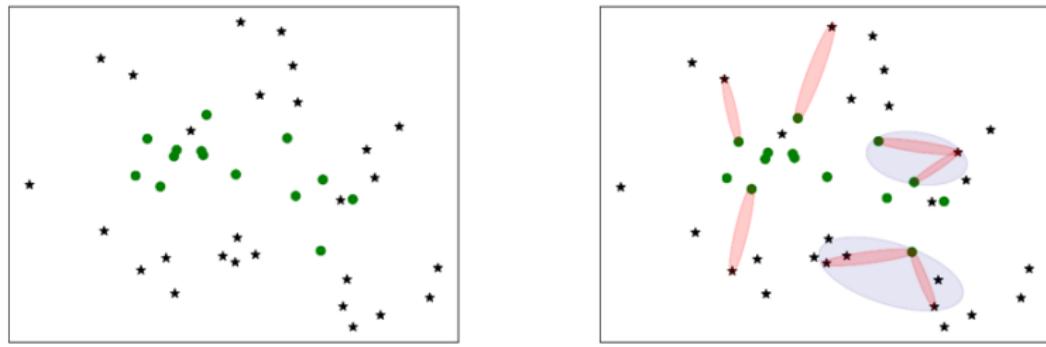


Figure: C-C pair of PDBID 3P2E with cutoff 5

Hypergraph filtration

The key point of all the persistent models is the filtration process. Here we assign nondecreasing filtration values for all the hyperedges, then the hypergraph together with these filtration values will naturally form a filtration process.

For the protein-ligand interactions, we use two steps to build the filtration.

- ① we define interactive distance between two atoms \mathbf{v}_i and \mathbf{v}_j as

$$d(\mathbf{v}_i, \mathbf{v}_j) = \begin{cases} \|\mathbf{v}_i - \mathbf{v}_j\|, & \text{if } \mathbf{v}_i \in \mathbf{V}_P, \mathbf{v}_j \in \mathbf{V}_L \text{ or } \mathbf{v}_i \in \mathbf{V}_L, \mathbf{v}_j \in \mathbf{V}_P \\ g(\mathbf{v}_i, \mathbf{v}_j), & \text{otherwise.} \end{cases} \quad (2)$$

- ② using the idea of Rips complex, the filtration value for a hyperedge $\sigma^n = \{\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_n\}$ is defined as,

$$f(\sigma^n) = \begin{cases} \max_{0 \leq i < j \leq n} d(\mathbf{v}_i, \mathbf{v}_j), & n > 0 \\ 0, & n = 0. \end{cases} \quad (3)$$

Remark: different $g(\mathbf{v}_i, \mathbf{v}_j)$ can result in different filtration processes. I will give more explanation later. Here, we use this formula

$$g(\mathbf{v}_i, \mathbf{v}_j) = \begin{cases} \max_{\mathbf{v}_k \in \mathbf{V}_P, \{\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k\} \in \mathcal{H}_2} \{d(\mathbf{v}_i, \mathbf{v}_k), d(\mathbf{v}_j, \mathbf{v}_k)\}, & \text{if } \mathbf{v}_i, \mathbf{v}_j \in \mathbf{V}_L \\ \max_{\mathbf{v}_k \in \mathbf{V}_L, \{\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k\} \in \mathcal{H}_2} \{d(\mathbf{v}_i, \mathbf{v}_k), d(\mathbf{v}_j, \mathbf{v}_k)\}, & \text{if } \mathbf{v}_i, \mathbf{v}_j \in \mathbf{V}_P \end{cases} \quad (4)$$

Example

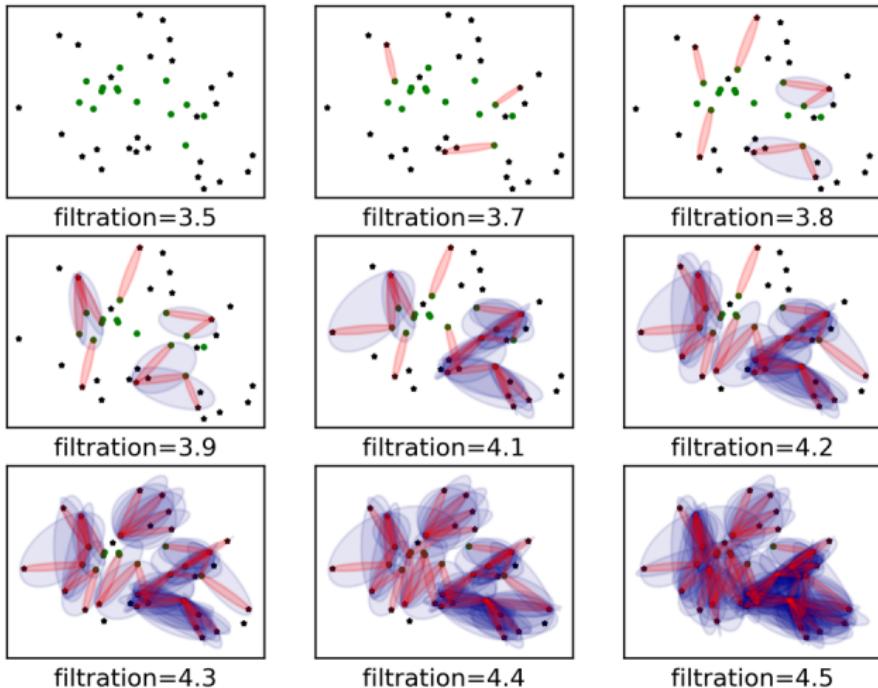


Figure: Hypergraph filtration for C-C pair of PDBID 3P2E

(Co)homology of the hypegraph

Given a hypergraph \mathcal{H} , its associated simplicial complex $K_{\mathcal{H}}$ is defined in 1991. In 2016, a more general definition — the embedded homology is proposed.

Here, we consider the embedded homology and corresponding persistent homology, for our hypergraph models.

Further, to incorporate the information that persistent homology cannot capture, we propose the hypergraph based persistent cohomology(HPC) and hypergraph based weighted persistent cohomology(HWPC).

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

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

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

Definition (supremum chain complex)

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

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

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

Proposition

Given a hypergraph \mathcal{H} , the infimum chain complex of \mathcal{H} with coefficient R is given by

$$Inf_n(\mathcal{H}, R) = R(\mathcal{H}_n) \cap \partial_n^{-1}(R(\mathcal{H}_{n-1}))$$

Proposition

Given a hypergraph \mathcal{H} , the supremum chain complex of \mathcal{H} with coefficient R is given by

$$Sup_n(\mathcal{H}, R) = R(\mathcal{H}_n) + \partial_{n+1}(R(\mathcal{H}_{n+1}))$$

Proposition

Given a hypergraph \mathcal{H} , the homology of the infimum chain complex 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

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

$$C_0 = Z\{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}\}$$

$$C_1 = Z\{\{0,1\}, \{2,3\}, \{2,4\}, \{3,4\}\}$$

$$C_2 = Z\{\{0,1,2\}\}$$

$$A_0 = Z\{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}\}$$

$$A_1 = Z\{\{0,1\}, \{0,2\}, \{1,2\}, \{2,3\}, \{2,4\}, \{3,4\}\}$$

$$A_2 = Z\{\{0,1,2\}\}$$

$$\rightarrow A_3 \xrightarrow{\partial_3} A_2 \xrightarrow{\partial_2} A_1 \xrightarrow{\partial_1} A_0$$

$$S_n = C_n + \partial_{n+1}(C_{n+1}), I_n = I_n \cap \partial_n^{-1}(C_{n-1})$$

$$H_0^s = \text{Ker}(\partial_0^s) / \text{Im}(\partial_1^s)$$

$$= S_0 / \text{Im}(\partial_1^s)$$

$$= Z\{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}\} / Z\{\{1\} - \{0\}, \{3\} - \{2\}, \{4\} - \{2\}, \{4\} - \{3\}\}$$

$$= I_0 / \text{Im}(\partial_1^i)$$

$$= \text{Ker}(\partial_0^i) / \text{Im}(\partial_1^i)$$

$$= H_0^i$$

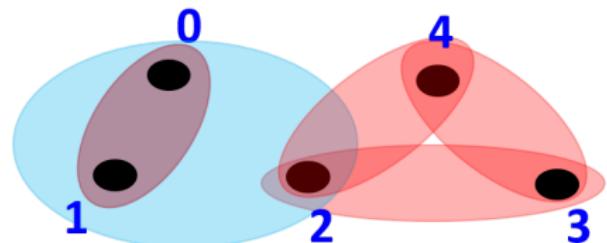
$$H_1^s = \text{Ker}(\partial_1^s) / \text{Im}(\partial_2^s)$$

$$= Z\{\{3,4\} - \{2,4\} + \{2,3\}, \partial\{0,1,2\}\} / Z\{\partial\{0,1,2\}\}$$

$$= Z\{\{3,4\} - \{2,4\} + \{2,3\}\}$$

$$= \text{Ker}(\partial_1^i) / \text{Im}(\partial_2^i)$$

$$= H_1^i$$



$$I_0 = Z\{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}\}$$

$$I_1 = Z\{\{0,1\}, \{2,3\}, \{2,4\}, \{3,4\}\}$$

$$I_2 = 0$$

$$S_0 = Z\{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}\}$$

$$S_1 = Z\{\{0,1\}, \{2,3\}, \{2,4\}, \{3,4\}, \partial\{0,1,2\}\}$$

$$S_2 = Z\{\{0,1,2\}\}$$

$$H_2^s = \text{Ker}(\partial_2^s) / \text{Im}(\partial_3^s)$$

$$= \text{Ker}(\partial_2^s)$$

$$= 0$$

$$= \text{Ker}(\partial_2^i) / \text{Im}(\partial_3^i)$$

$$= H_2^i$$

Embedded cohomology of hypergraph

Definition (Infimum cochain complex)

Given a hypergraph \mathcal{H} , we define the infimum cochain complex $\text{Inf}^*(\mathcal{H}, G)$ with coefficient G as

$$\text{Inf}^n(\mathcal{H}, G) = (\text{Inf}_n(\mathcal{H}, R))^* = \text{Hom}(\text{Inf}_n(\mathcal{H}, R), G),$$

which the dual of $\text{Inf}_n(\mathcal{H}, R)$.

Definition (Supremum cochain complex)

Given a hypergraph \mathcal{H} , we define the Supremum cochain complex $\text{Sup}^*(\mathcal{H}, G)$ with coefficient G as

$$\text{Sup}^n(\mathcal{H}, G) = (\text{Sup}_n(\mathcal{H}, R))^* = \text{Hom}(\text{Sup}_n(\mathcal{H}, R), G),$$

which the dual of $\text{Sup}_n(\mathcal{H}, R)$.

Proposition

Given a hypergraph \mathcal{H} , the cohomology of $\text{Sup}^*(\mathcal{H}, G)$ and the cohomology of $\text{Inf}^*(\mathcal{H}, G)$ are isomorphic.

Definition (Hypergraph embedded cohomology)

Given a hypergraph \mathcal{H} , we define the n -th embedded cohomology of \mathcal{H} with coefficient G as

$$H^n(\mathcal{H}, G) = H^n(\text{Sup}^*(\mathcal{H}, G)) = H^n(\text{Inf}^*(\mathcal{H}, G)).$$

From now on, the coefficient R and G are both Z_2 . So we have $\{\mathbf{v}_i, \mathbf{v}_j\} = -\{\mathbf{v}_i, \mathbf{v}_j\} = \{\mathbf{v}_j, \mathbf{v}_i\}$, which means that for any two n -hyperedges $\sigma_n, \sigma_{n'}$ ($0 \leq n \leq m$), if the vertices of σ_n and $\sigma_{n'}$ are same. Then, $\sigma_n = \sigma_{n'}$.

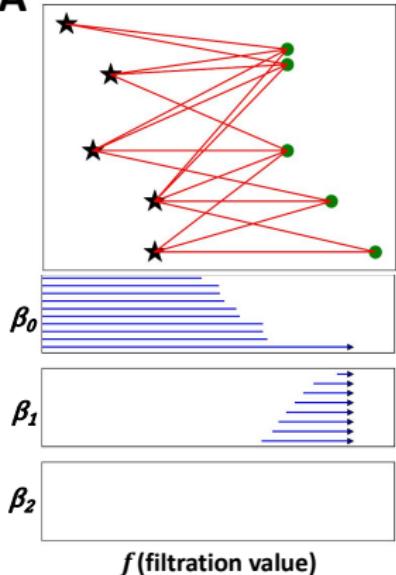
Theorem

For an m -dimension hypergraph \mathcal{H} , its embedded cohomology is the simplicial cohomology of $K_{\mathcal{H}}$, if it satisfies the following two conditions.

- ① For each vertex of \mathcal{H} , it is a 0-hyperedge.
- ② For each n -hyperedge $\sigma_n = \{\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_n\}$ ($1 \leq n \leq m$) of \mathcal{H} , it has an associated ("face") set $\{\sigma_{n-1}^0, \sigma_{n-1}^1, \dots, \sigma_{n-1}^n\}$ with σ_{n-1}^i ($0 \leq i \leq n$) = $\{\mathbf{v}_0, \dots, \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, \dots, \mathbf{v}_n\}$ generated from σ_n by removing the vertex \mathbf{v}_i . It only allows at most one ("face") element (among the $n + 1$ elements) from the set is not an $(n - 1)$ -hyperedge of \mathcal{H} .

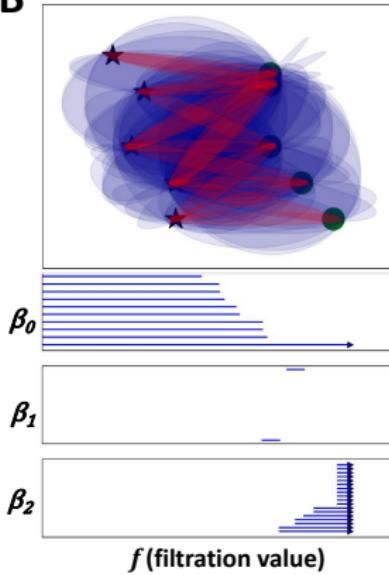
Hypergraph based persistent cohomology(HPC)

A



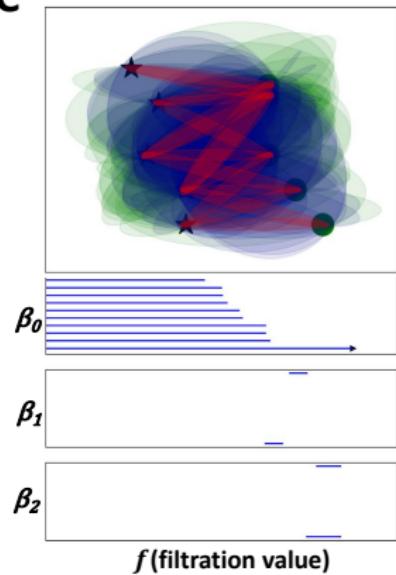
Bipartite graph

B



Hypergraph(up to 2)

C



Hypergraph(up to 3)

Figure: Bipartite-graph VS hypergraph

Weighted persistent cohomology has been proposed to incorporate more structure, physical, chemical and biological information into a unified representation, i.e., persistent cohomology enriched barcode[?]. Different from all previous models, here we consider a new weight scheme, derived from graph centrality, for persistent cohomology. More specifically, we define the weights for a 0-hyperedge $\sigma_0 = \{\mathbf{v}_i\}$ as,

$$w(\sigma_0) = \begin{cases} \sum_{\mathbf{v}_k \in V_L} e^{-\frac{\|\mathbf{v}_k - \mathbf{v}_i\|^2}{\eta^2}}, & \mathbf{v}_i \in V_P \\ \sum_{\mathbf{v}_k \in V_P} e^{-\frac{\|\mathbf{v}_k - \mathbf{v}_i\|^2}{\eta^2}}, & \mathbf{v}_i \in V_L \end{cases} \quad (5)$$

and an 1-hyperedge $\sigma_1 = \{\mathbf{v}_i, \mathbf{v}_j\}$ as,

$$w(\sigma_1) = e^{-\frac{\|\mathbf{v}_i - \mathbf{v}_j\|^2}{\eta^2}}. \quad (6)$$

The scale parameter η controls the influence range of the nodes. Smaller η values mean local interactions and larger η values mean global interactions.

Further, we define the weight for a 0-cohomology generator δ^0 as,

$$w(\delta^0) = \frac{\sum_{\sigma_0^i \in \mathcal{H}_0} \delta^0(\sigma_0^i) * w(\sigma_0^i)}{\sum_{\sigma_0^i \in \mathcal{H}_0} \delta^0(\sigma_0^i)}, \quad (7)$$

and an 1-cohomology generator δ^1 as,

$$w(\delta^1) = \frac{\sum_{\sigma_1^i \in \mathcal{H}_1} \delta^1(\sigma_1^i) * w(\sigma_1^i)}{\sum_{\sigma_1^i \in \mathcal{H}_1} \delta^1(\sigma_1^i)}. \quad (8)$$

Note that Z_2 is used in computation, and the term $\delta^1(\sigma_1^i)$ is either 0 or 1. For persistent cohomology enriched barcodes, each barcode represents a generator and is colored by the weight values defined above.

Computationally, our persistent cohomology enriched barcodes are calculated based on the associated simplicial complex. And the weights defined in Eqs. (5) and (6) are generalized to any 0-simplex and 1-simplex, respectively. Note that distance between two atoms from the same molecule is defined as Eq. (2). Similarly, weights for cohomology generators in Eqs. (7) and (8) are also extended to associated simplicial complex counterparts.

HPC/HWPC base machine learning

Persistent barcodes from HPC and HWPC can be discretized into feature vectors. We combine these feature vectors with gradient boosting tree to build our HPC/HWPC-ML model.

For the discretization, we consider the binning approach. That is, The total number of the barcodes (i.e., Betti numbers for HPC) or the sum of the weight values of enriched barcodes (for HWPC) within each bin, is used as molecular descriptors.

Protein-ligand binding affinity prediction

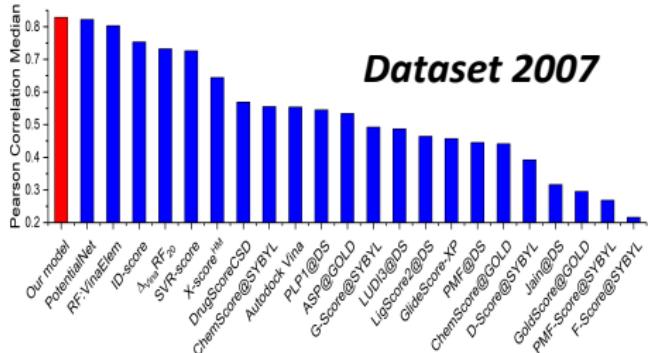
A drug design process covers various steps. Among all the steps, one of the key issue is to identify ligands (drug) that have higher binding affinity with the target biomolecules. During the past few decades, a variety of empirical, physics-based, knowledge-based, and machine-learning-based models are proposed. The databank PDBbind (www.pdbsbind.org.cn) is established to systematically evaluate and compare their performance for protein-ligand binding affinity prediction.

Parameter setting

We use our HPC/HWPC-ML to make the protein-ligand binding affinity prediction, the cutoff is 10.5Å, filtration region is [2.0Å,7.5Å] with bin size 0.1Å. In this way, the feature size are $3960=36\times55\times2\times1$, $7920=36\times55\times2\times2$ and $11880=36\times55\times2\times3$ for HPC, HWPC and combined-HWPC respectively. The parameter setting for GBT is as follows.

No. of Estimators	Learning rate	Max depth	Subsample
40000	0.001	9	0.7
Min_samples_split	Loss function	Max features	Repetitions
2	Least square	SQRT	10

Results



HPC/HWPC
+
GradientBoostingTree

	HPC	HWPC2.5	HWPC10	HWPC(2.5+10)
PDBbind2007	0.813(1.423)	0.823(1.418)	0.827(1.395)	0.829(1.403)
PDBbind2013	0.770(1.508)	0.780(1.498)	0.779(1.486)	0.784(1.483)
PDBbind2016	0.810(1.359)	0.825(1.322)	0.825(1.324)	0.831(1.307)

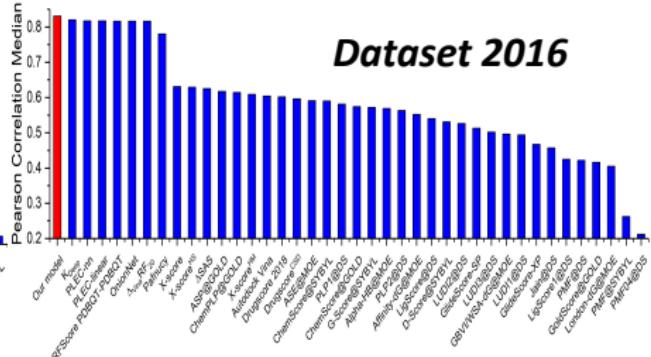
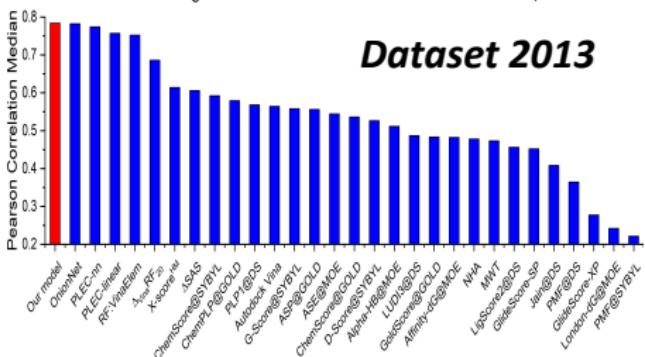


Figure: Comparison of our model with other traditional models

Thank You