Topology based deep learning for drug discovery

Guowei Wei

Departments of Mathematics

Michigan State University

http://www.math.msu.edu/~wei

The 3rd Annual Meeting of SIAM Central States Section

September 29 — October 1, 2017

Colorado State University

Grant support:

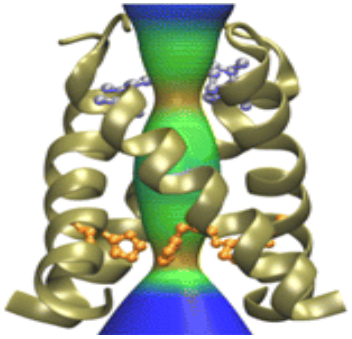
NSF, NIH, MSU and BMS

Drug design and discovery

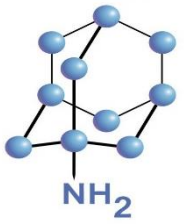
1) Disease identification
2) Target hypothesis
3) Virtual screening
4) Drug structural optimization in the target binding site
5) Preclinical *in vitro* and *in vivo* test
6) Clinical test
7) Optimize drug's efficacy, toxicity, pharmacokinetics, and pharmacodynamics properties (quantitative systems pharmacology)

Influenza -- flu virus

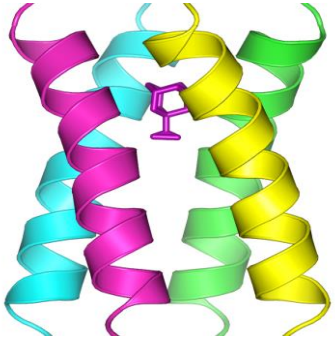
M2 channel



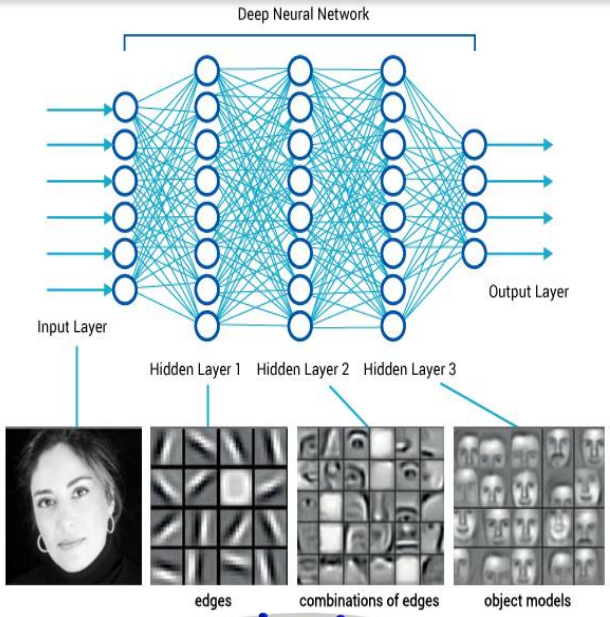
Amantadine

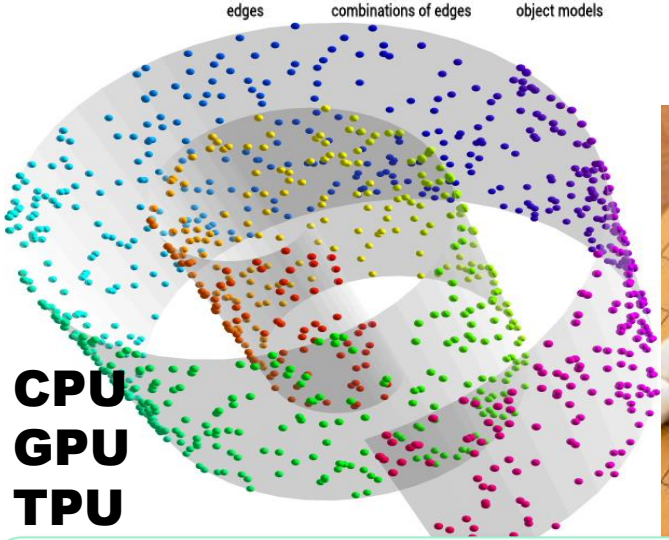


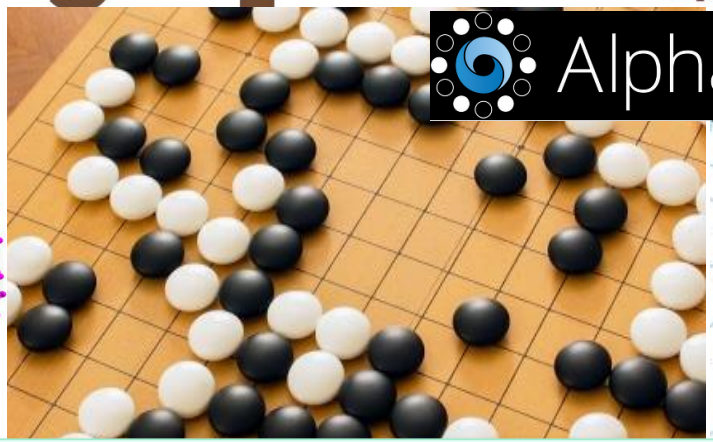
M2-A complex

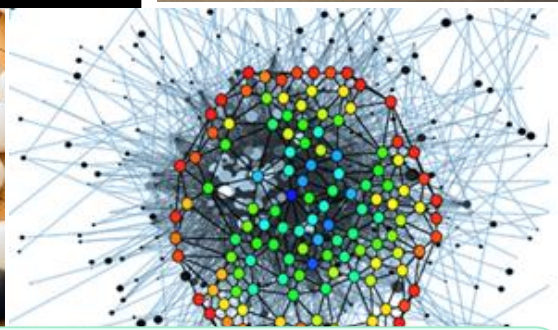


Welcome to big-data era

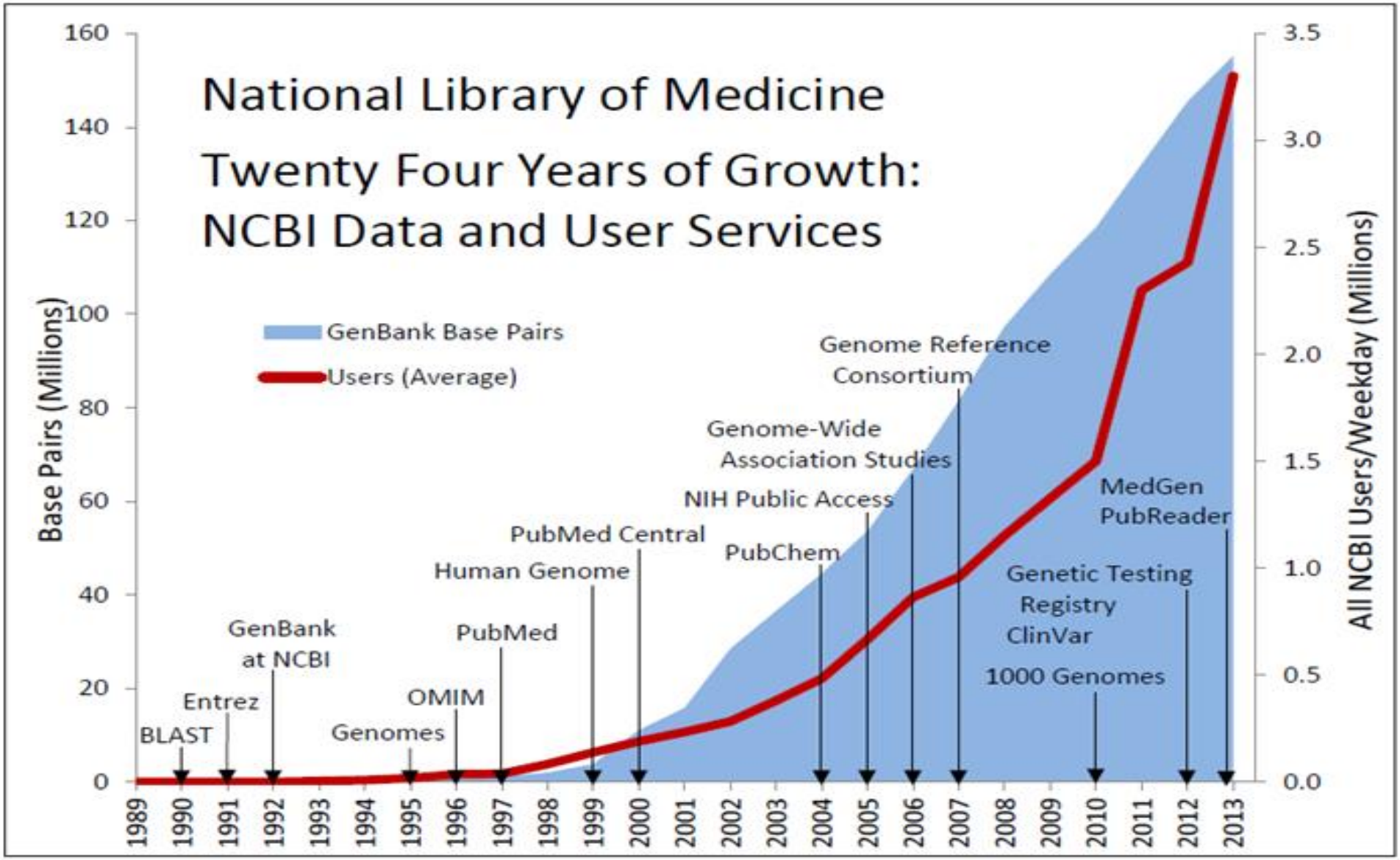






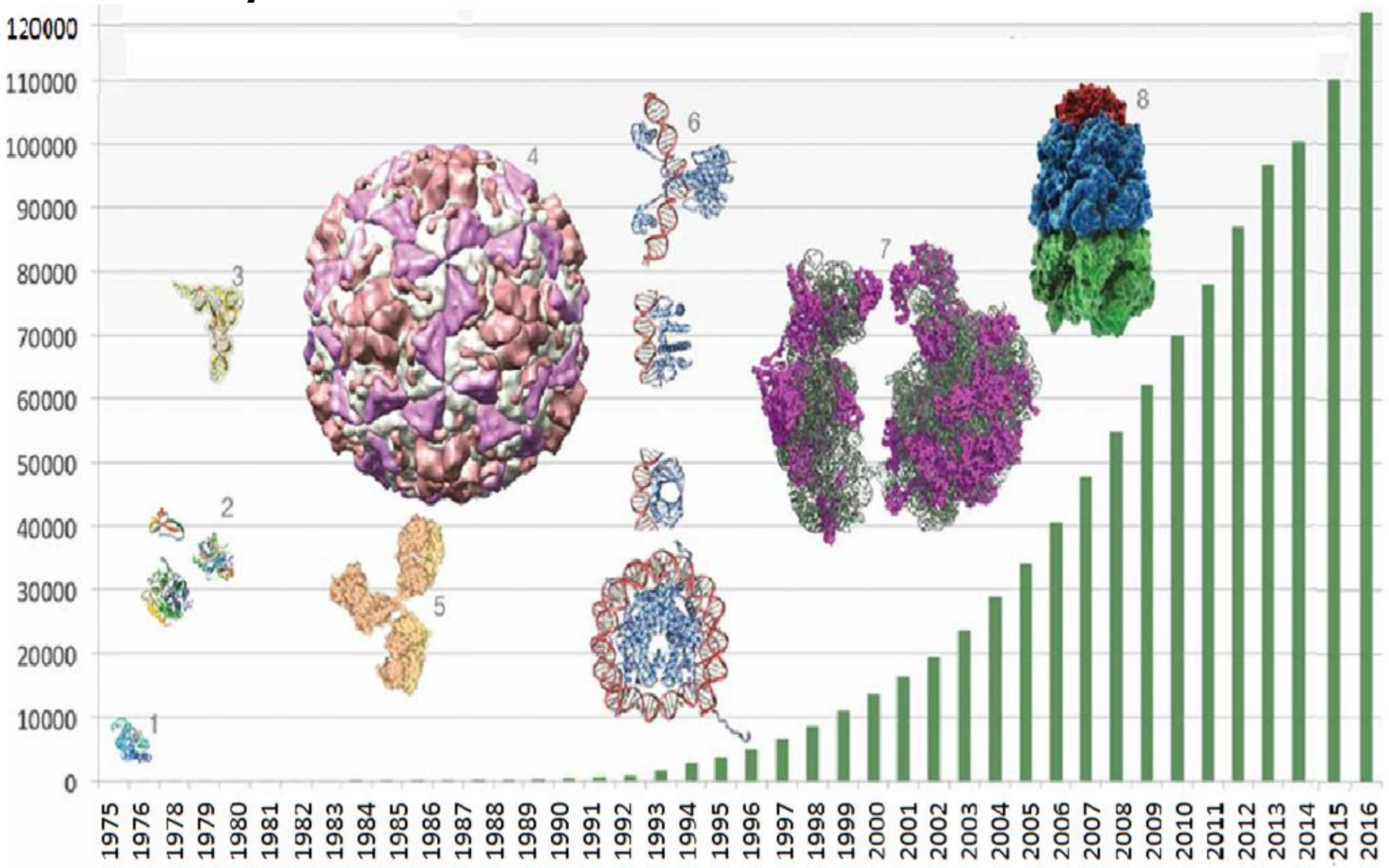


Half of all jobs will be done by robots in the near future



		GenBank		Whole Genome Shotgun	
Release	**Date**	**Bases**	**Sequences**	**Bases**	**Sequences**
219	Apr 2017	231824951552	200877884	2035032639807	451840147

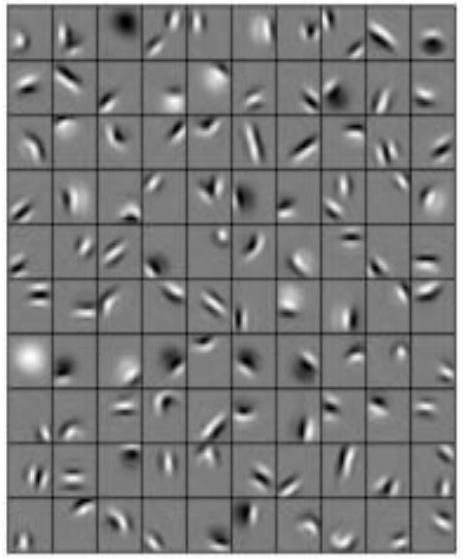
Yearly Growth of Total Structures in the Protein Data Bank

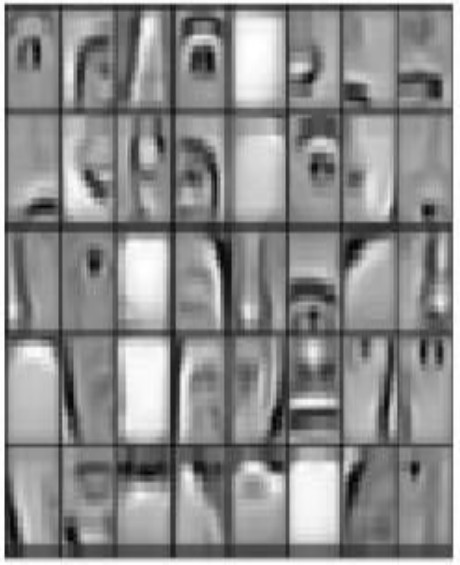


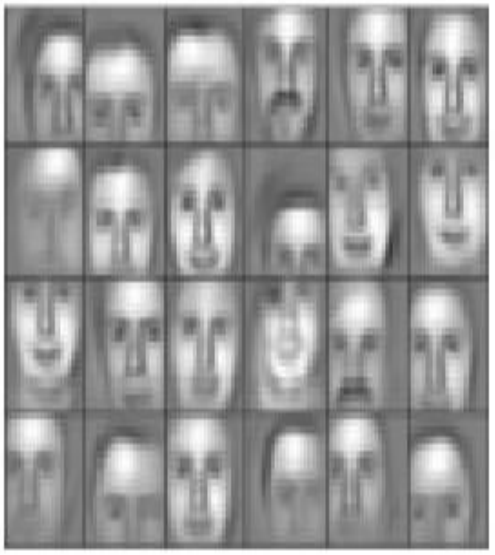
Biological sciences are undergoing a historic transition: From qualitative, phenomenological, and descriptive to quantitative, analytical and predictive, as quantum physics did a century ago

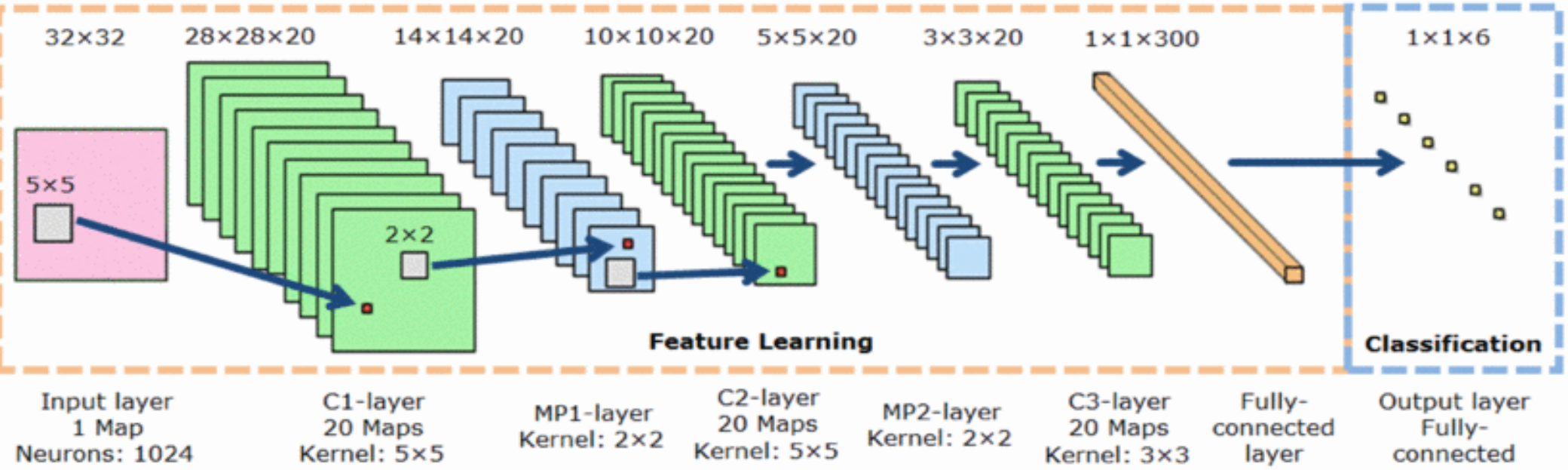
Deep learning

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









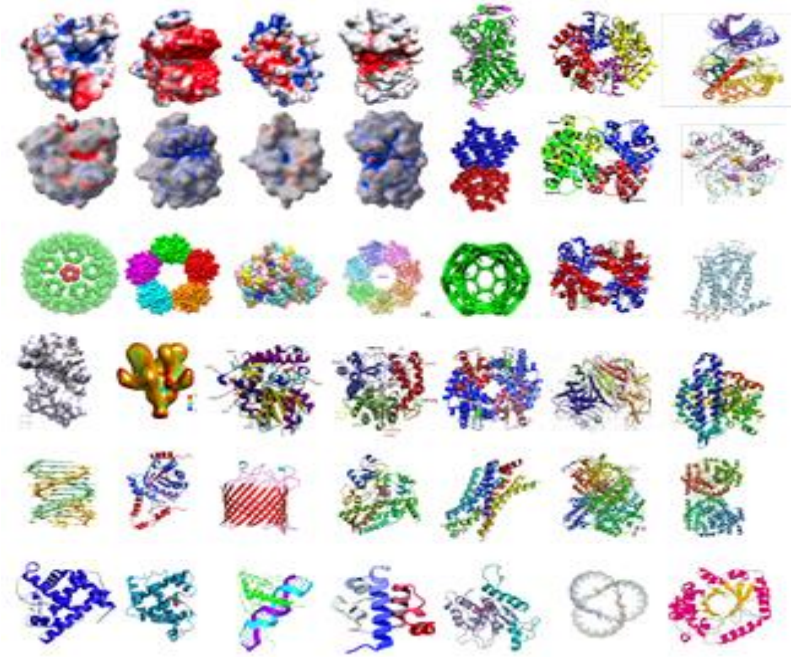
How to do deep learning for 3D biomolecular data?

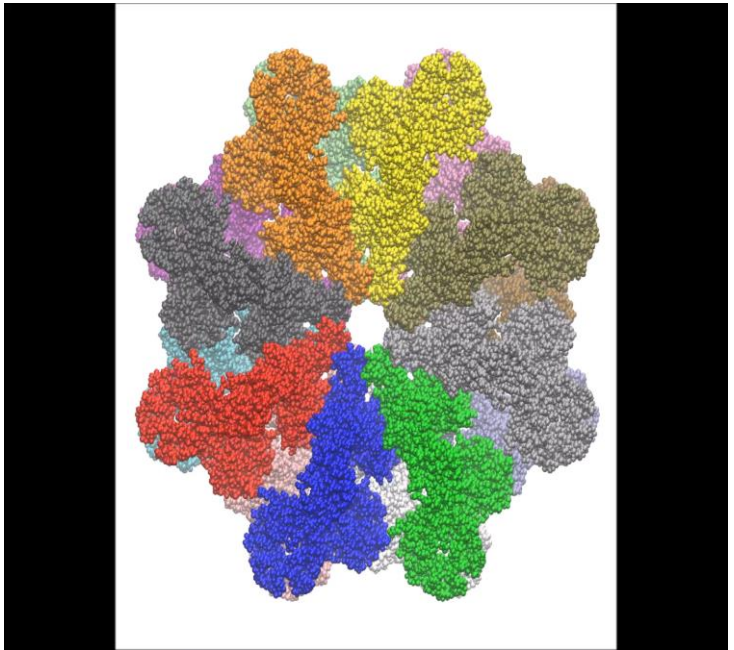
Obstacles for deep learning of 3D biomolecules:

- **Geometric dimensionality: R^{3N}, where N~5500 for a protein.**
- **Machine learning dimensionality: > $m1024^3$, where m is the number of atom types in a protein.**
- **Molecules have different sizes --- non-scalable.**
- **Complexity: biochemistry & biophysics**

Solution:

- **Topological simplification**
- **Dimensionality reduction & unification (scalability)**

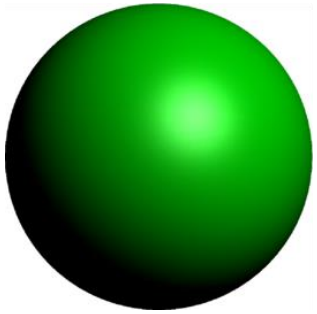


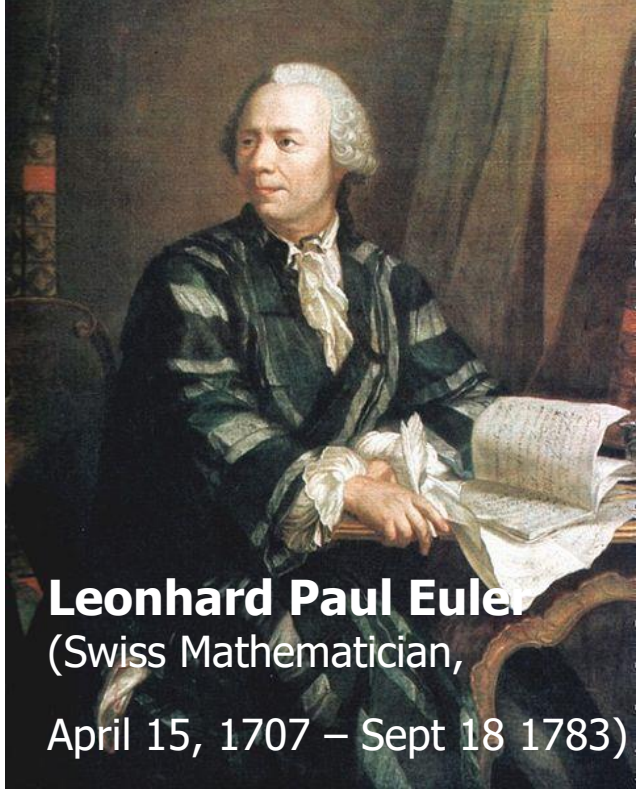


Classical topological objects

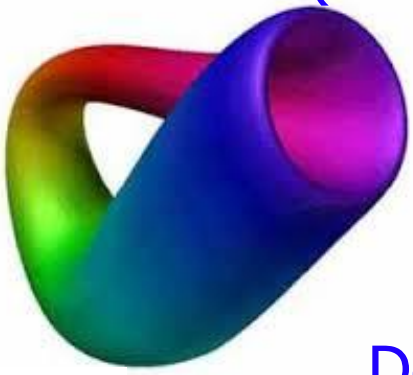
Möbius Strips (1858)

Sphere

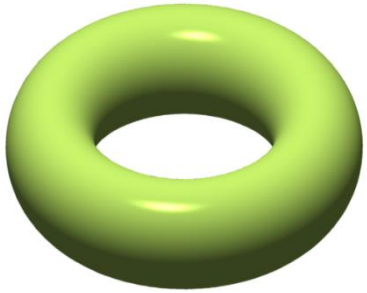




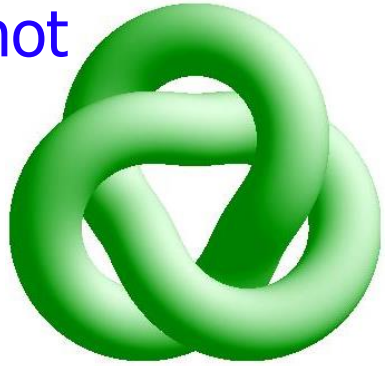
Klein Bottle (1882)



Torus

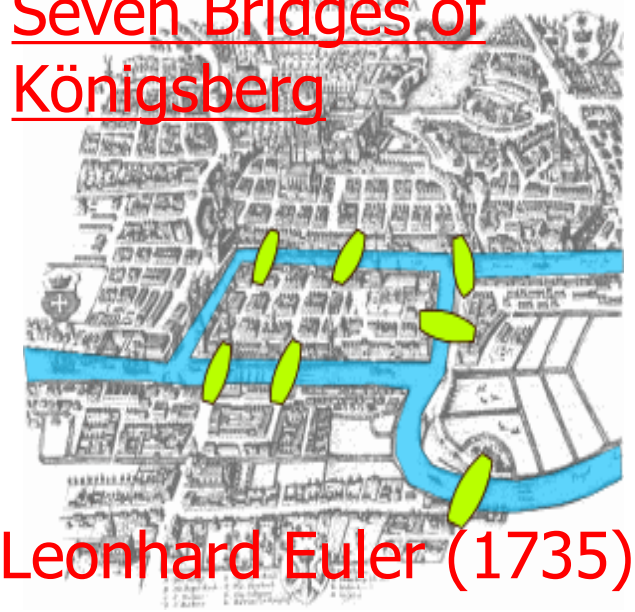


Trefoil Knot



Double Torus

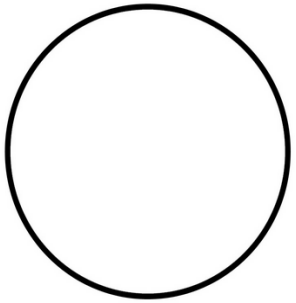
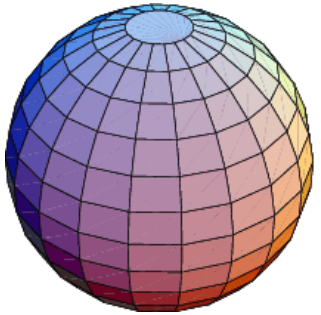
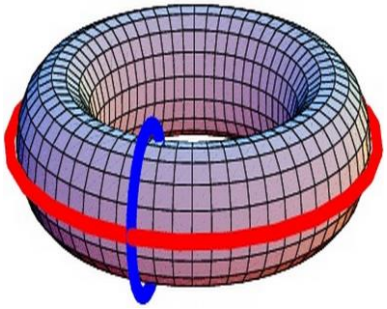
Seven Bridges of Königsberg



Leonhard Euler (1735)

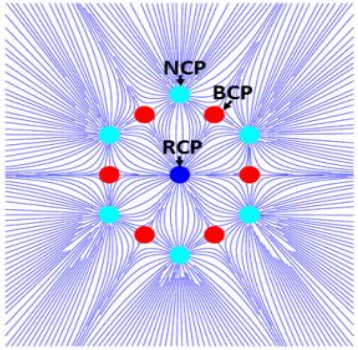
Topological invariants: Betti numbers

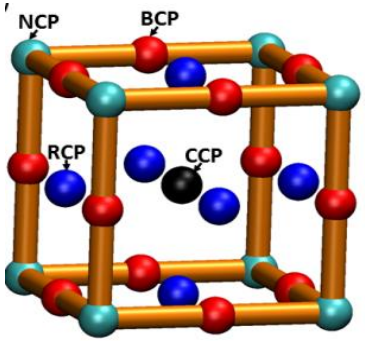
β_0 is the number of connected components.
β_1 is the number of tunnels or circles.
β_2 is the number of cavities or voids.

Point	Circle	Sphere	Torus
			
$\beta_0 = 1$	$\beta_0 = 1$	$\beta_0 = 1$	$\beta_0 = 1$
$\beta_1 = 0$	$\beta_1 = 1$	$\beta_1 = 0$	$\beta_1 = 2$
$\beta_2 = 0$	$\beta_2 = 0$	$\beta_2 = 1$	$\beta_2 = 1$

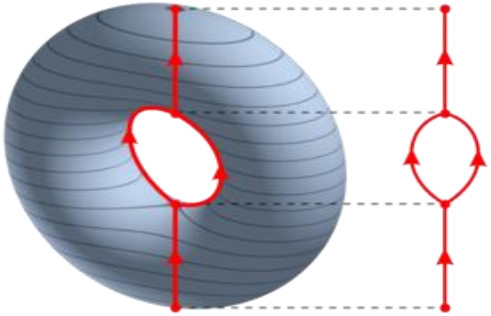
Topological simplification

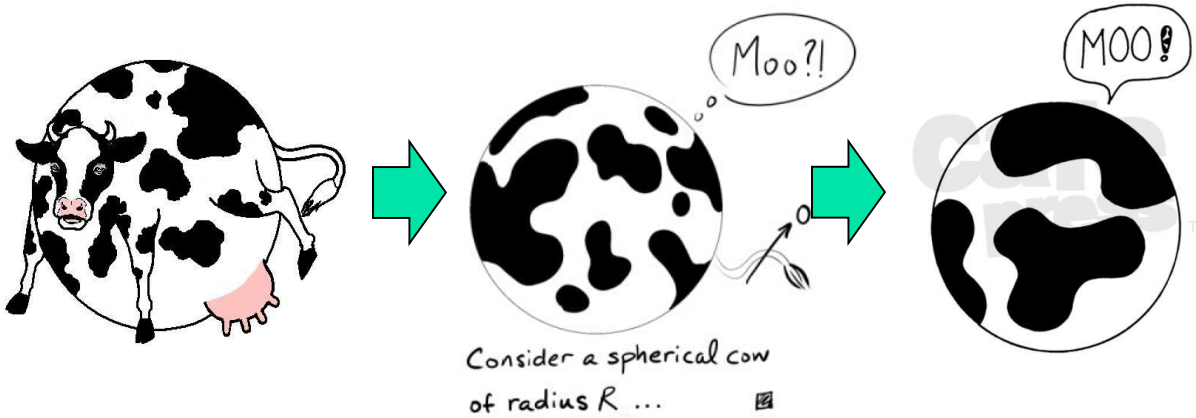
Poincare-Hopf index



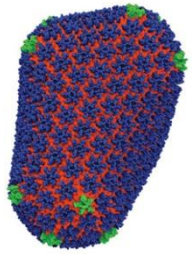


Morse theory

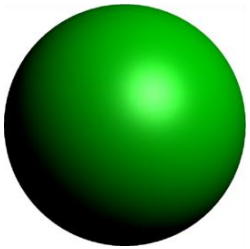




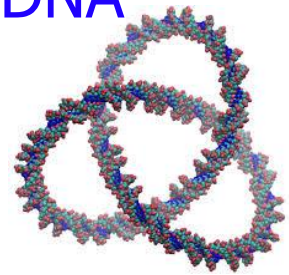
Mug ⟺ Doughnut



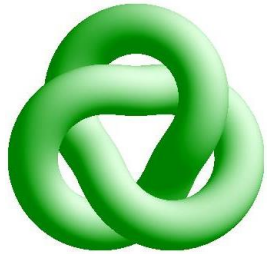
HIV 4.2 million atoms



DNA



Trefoil Knot



Opportunities, challenges and promises

Opportunities from topological methods:

- ❖ **New approach for big data characterization and classification.**
- ❖ **Dramatic reduction of dimensionality and data size.**
- ❖ **Applicable to a variety of fields.**

Challenges with topological methods:

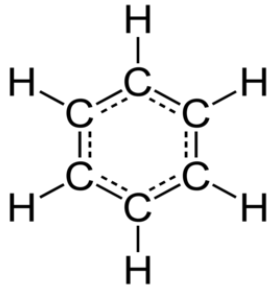
- ➢ **Geometric methods are often inundated with too much structural detail.**
- ➢ **Topological tools incur too much reduction of original geometric information.**
- ➢ **Topology is hardly used for quantitative prediction.**

Promises from persistent homology:

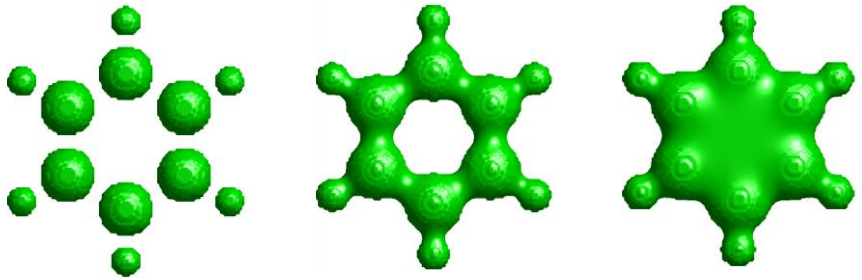
- ✓ **Embeds geometric information in topological invariants.**
- ✓ **Bridges the gap between geometry and topology.**

Persistent homology answers following questions

What is the topology of a benzene?

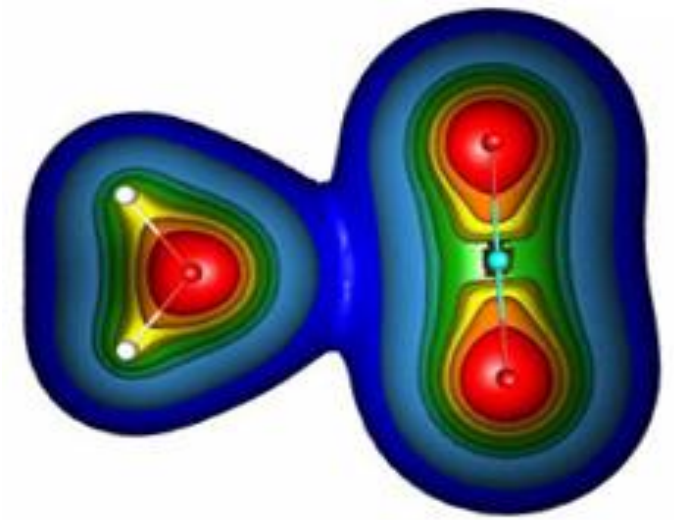


Level sets generated by Laplace-Beltrami flows:



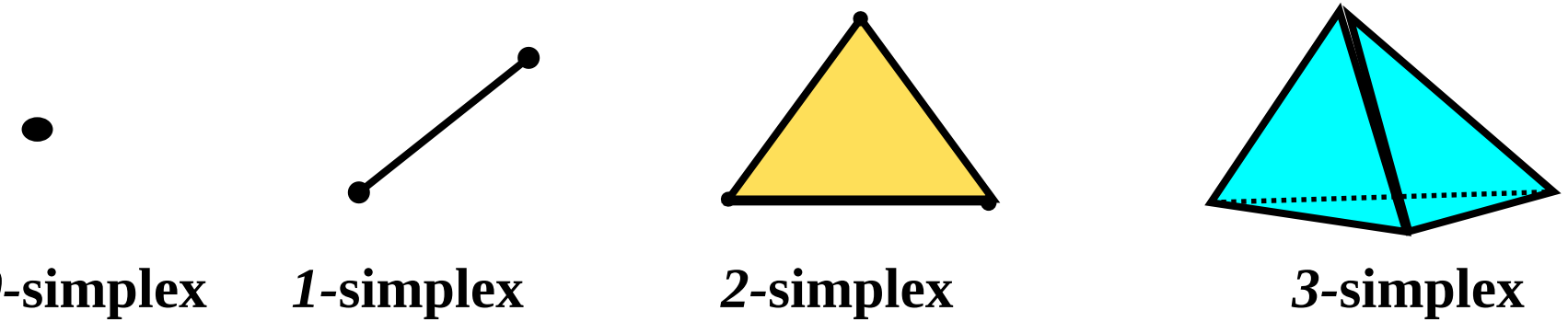
What is the topology of a H_2O-CO_2 complex?

Electron density level sets computed by using quantum mechanics:



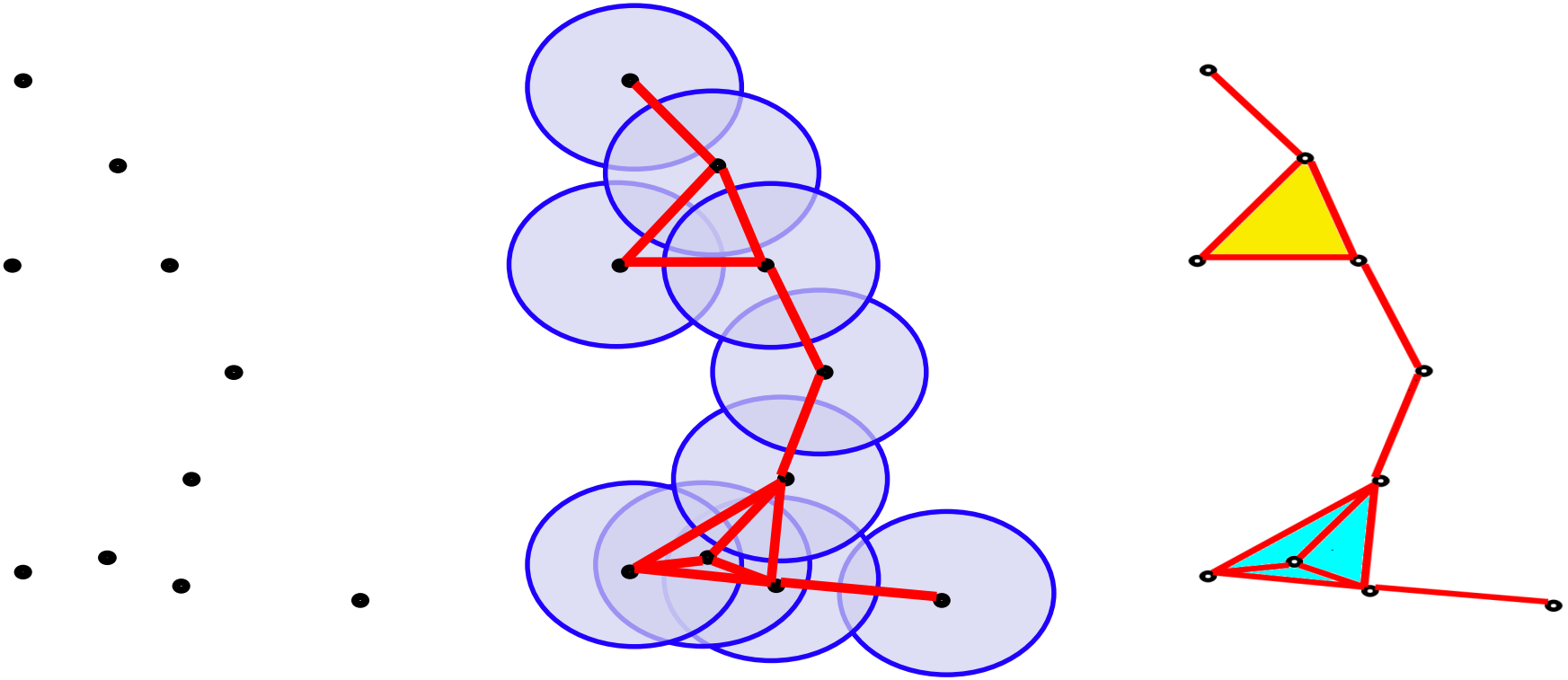
Vietoris-Rips complexes of planar point sets

Simplexes:



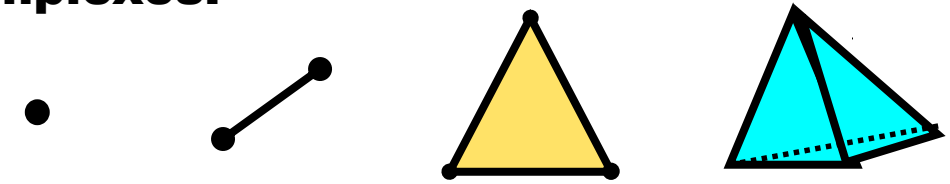
0-simplex *1*-simplex *2*-simplex *3*-simplex

Simplicial complexes of ten points:



Topological modeling - Persistent homology

Simplexes:



0-simplex *1*-simplex *2*-simplex *3*-simplex

k-chain: $\sum_i c_i \sigma_i^k$

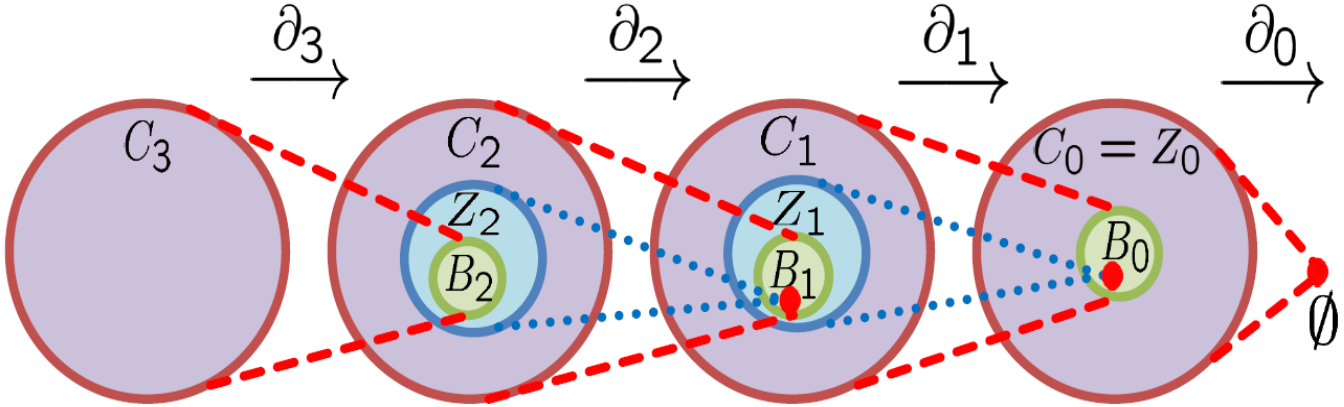
Chain group: $C_k(K, Z_2)$

Frosini and Nandi (1999),
Robins (1999),
Edelsbrunner, Letscher and Zomorodian (2002), Edelsbrunner and Harer, (2007)
Kaczynski, Mischaikow and Mrozek (2004),
Zomorodian and Carlsson (2005),
Ghrist (2008),
.....

Boundary operator:

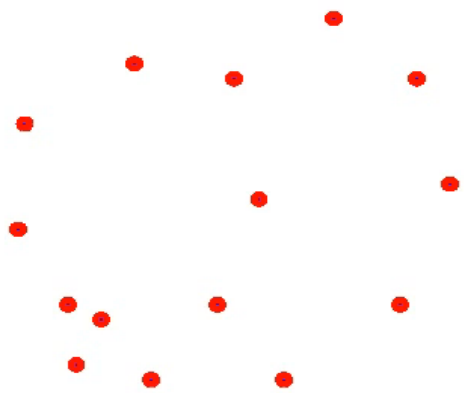
$$\partial_k s^k = \sum_{i=0}^{k} (-1)^i \{v_0, v_1, \ldots, \hat{v}_i, \ldots, v_k\}$$

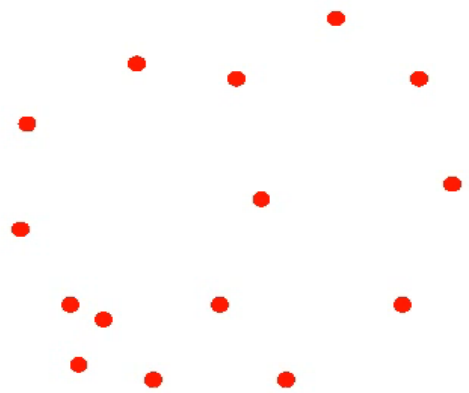
$$Z_k = \mathrm{Ker}\, \partial_k$$
$$B_k = \mathrm{Im}\, \partial_{k+1}$$
$$H_k = Z_k / B_k$$
$$\beta_k = \mathrm{Rank}\,(H_k)$$

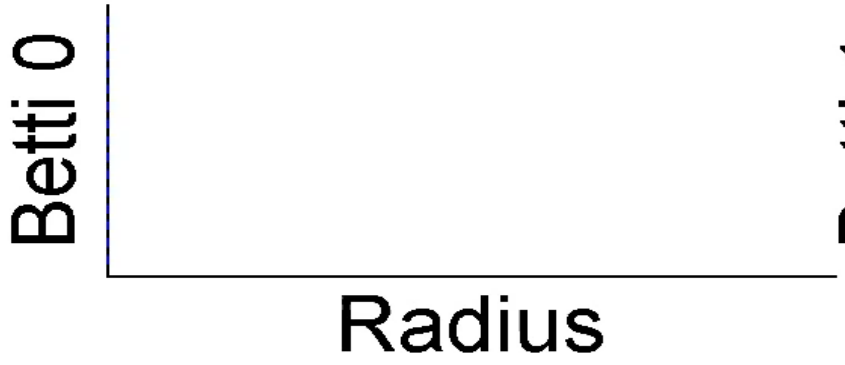


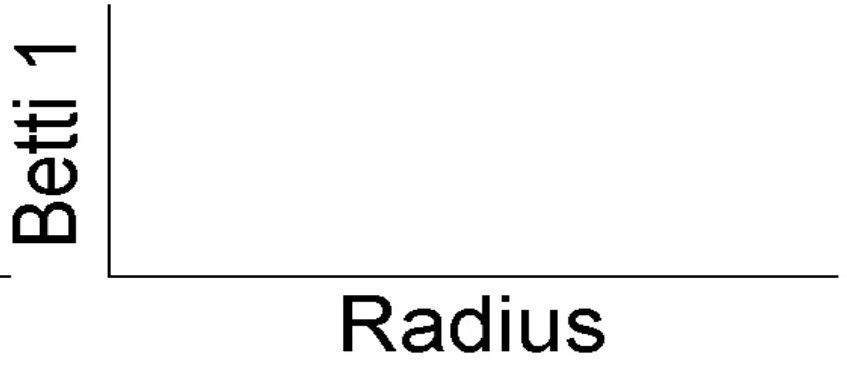
Vietoris-Rips complexes, persistent homology and persistent barcodes

(Xia, Wei, 2014)

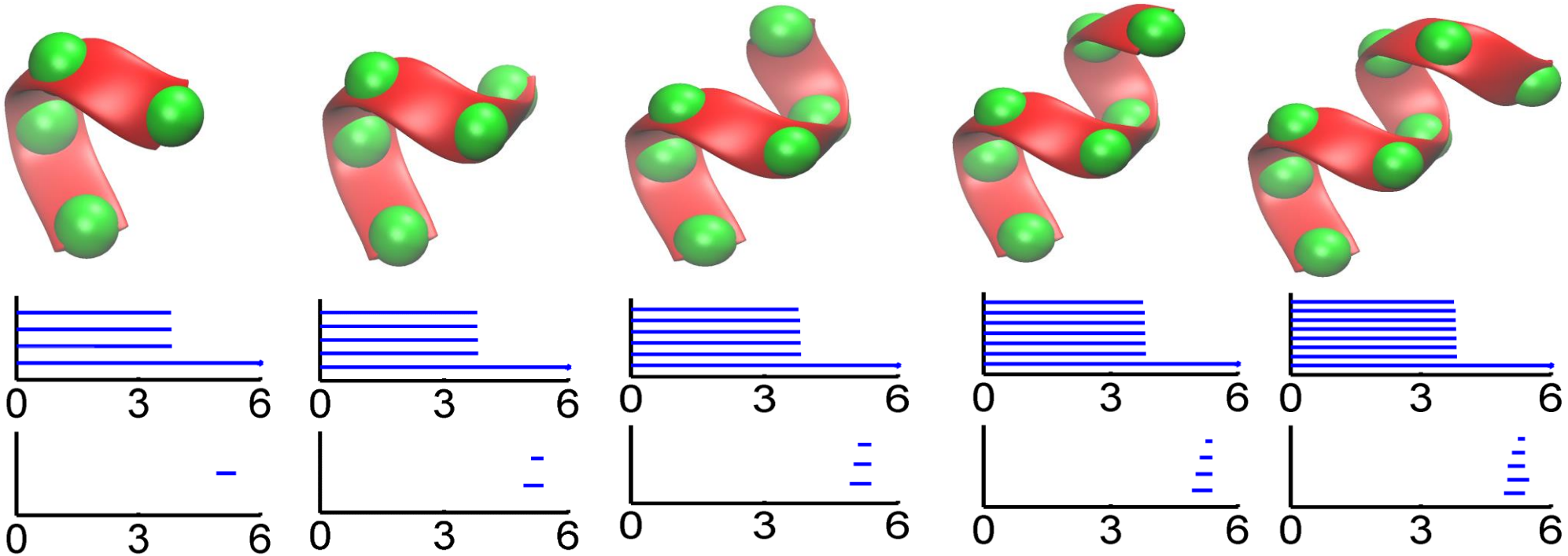


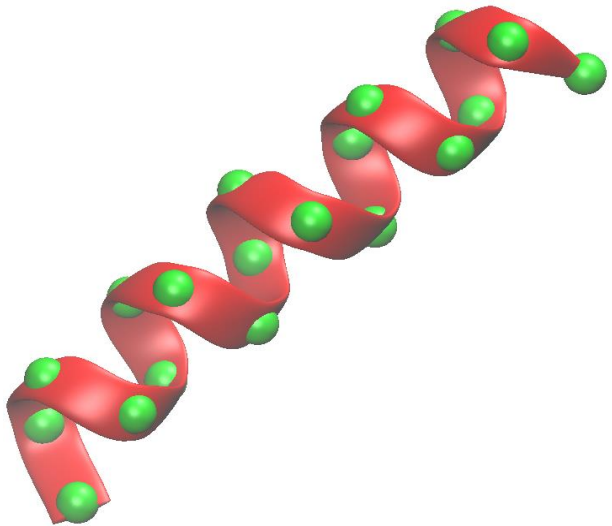


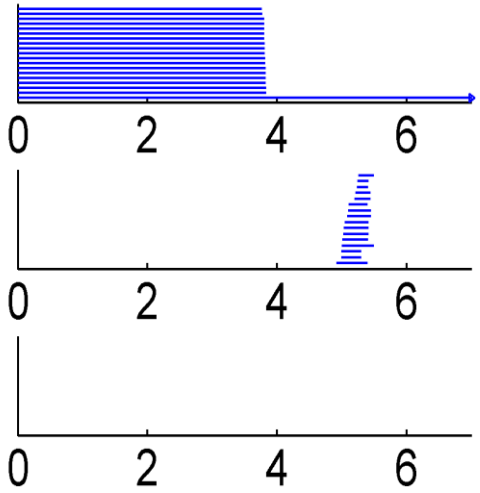




Topological fingerprints of an alpha helix







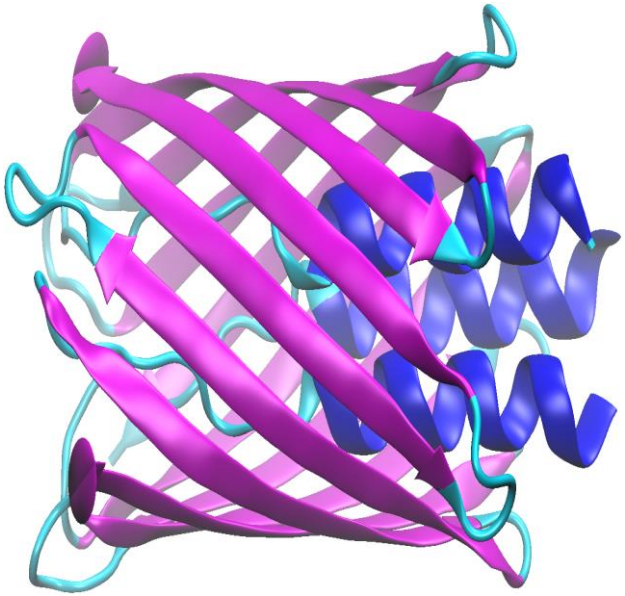
Short bars are NOT noise!

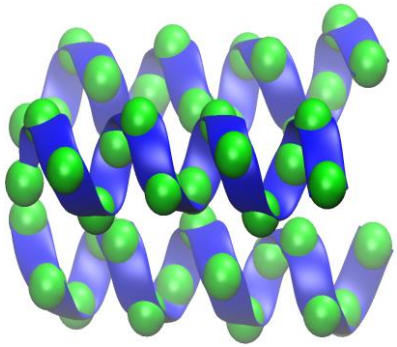
(Xia & Wei, IJNMBE, 2014)

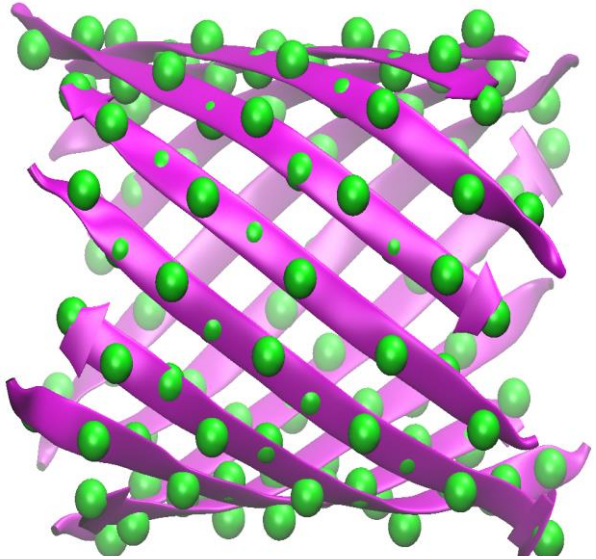
Topological fingerprints of beta barrel

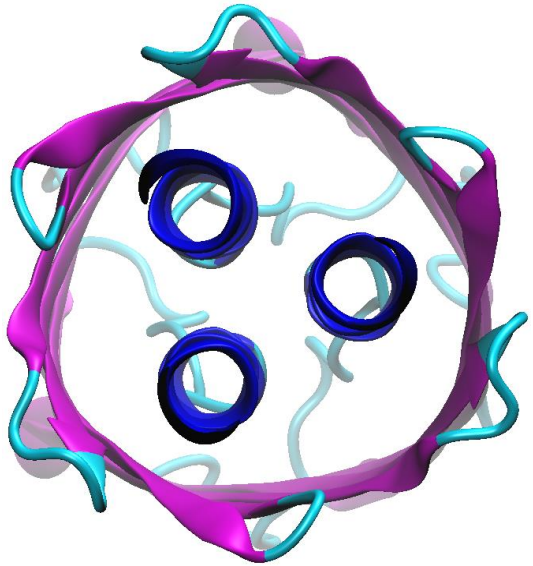
(Xia & Wei, IJNMBE, 2014)

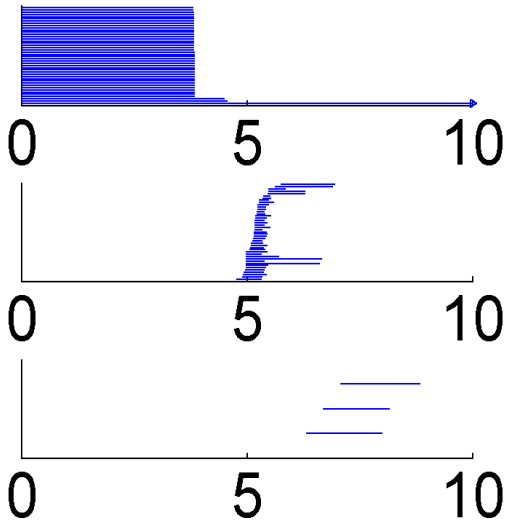
Protein:2GR8

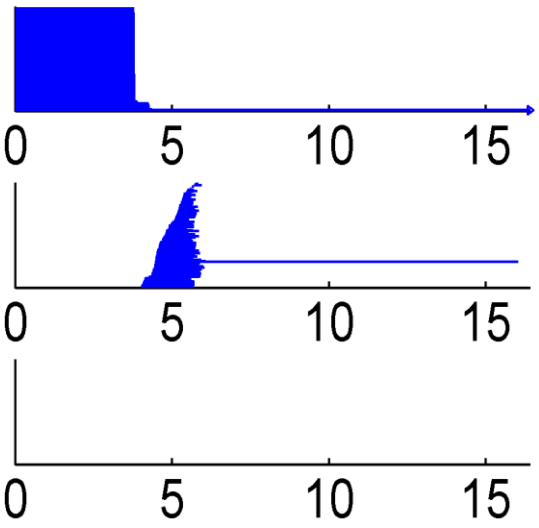






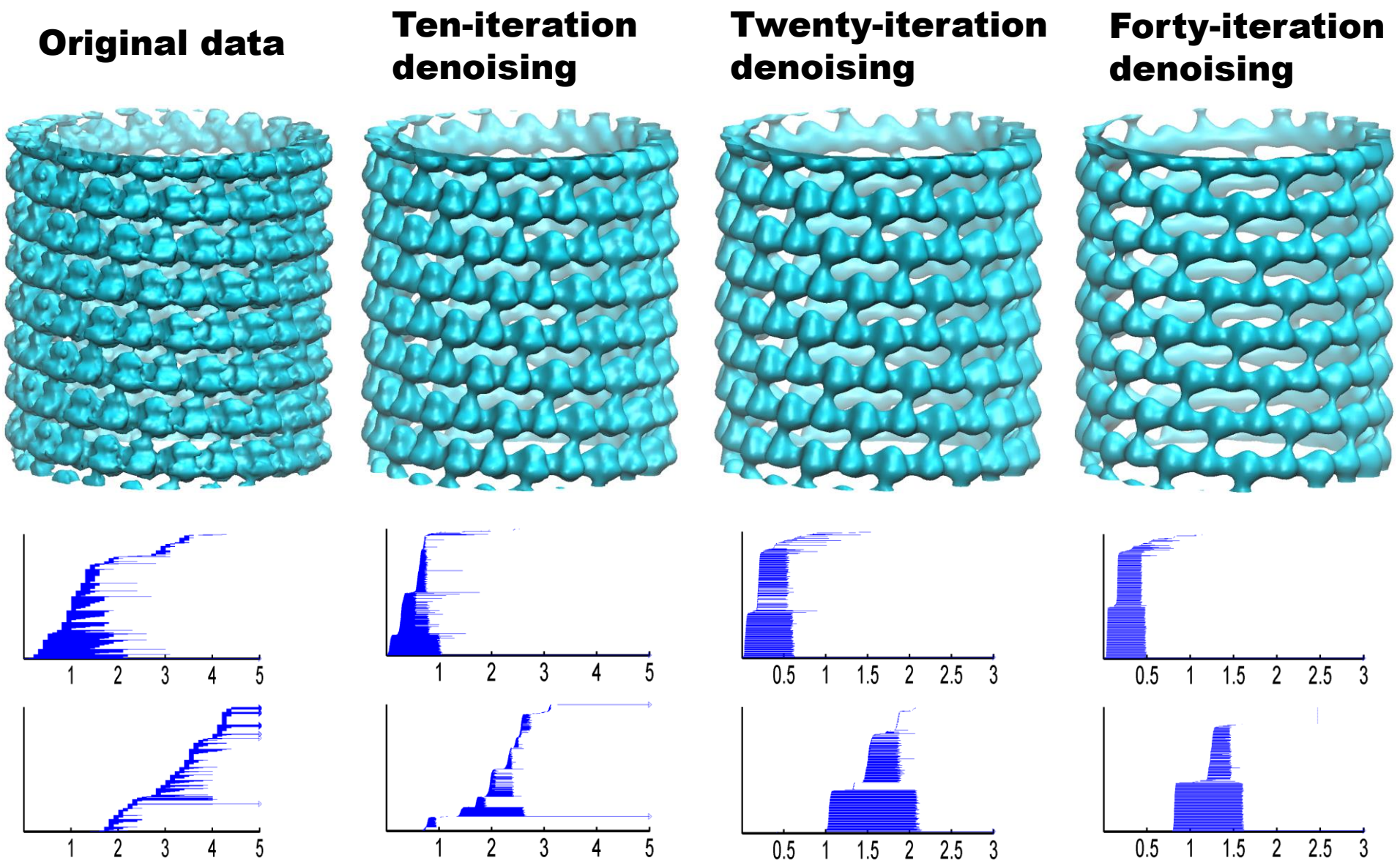




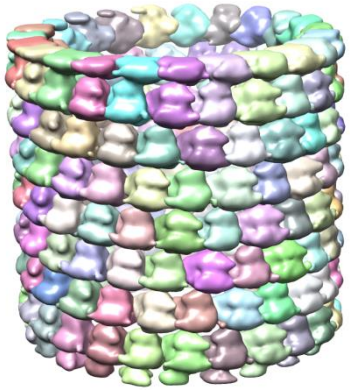


Topological noise reduction

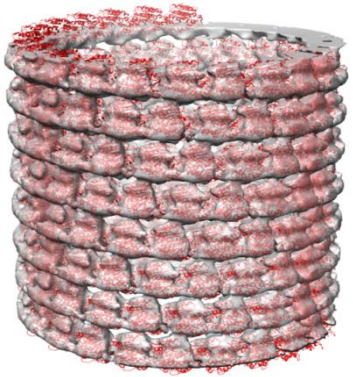
(Xia & Wei, IJNMBE 2015)



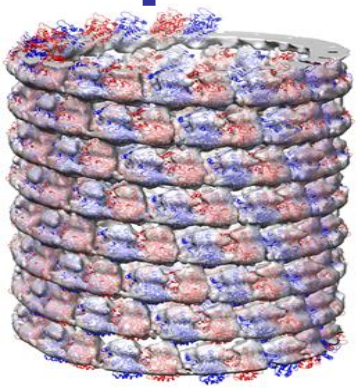
Persistent homology for ill-posed inverse problems



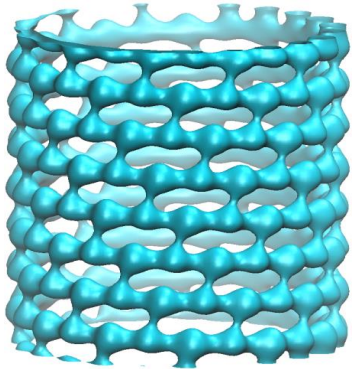
Original data: microtubule

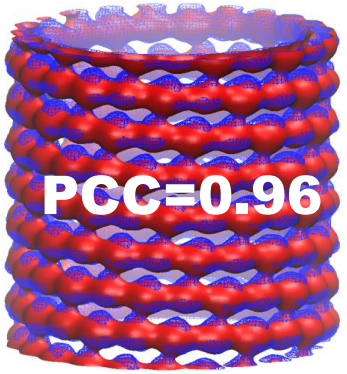


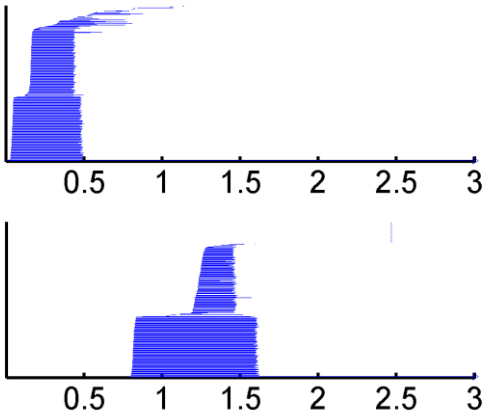
Fitted with one-type of tubulins

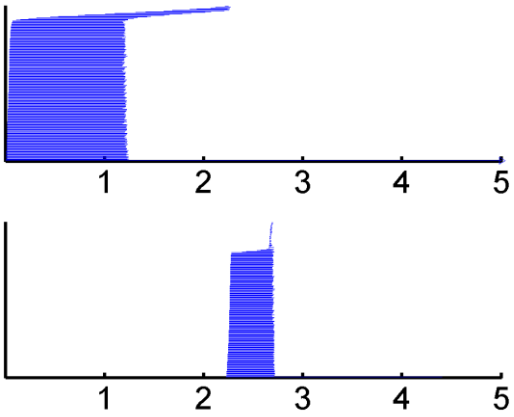


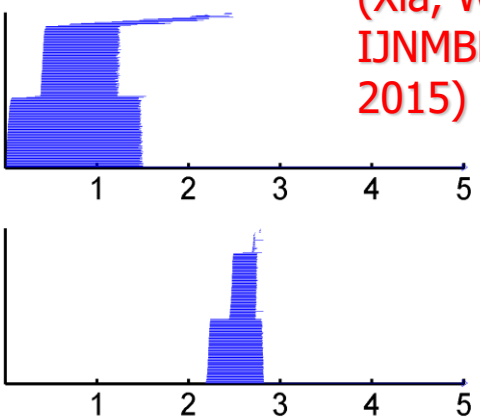
Fitted with two-types of tubulins







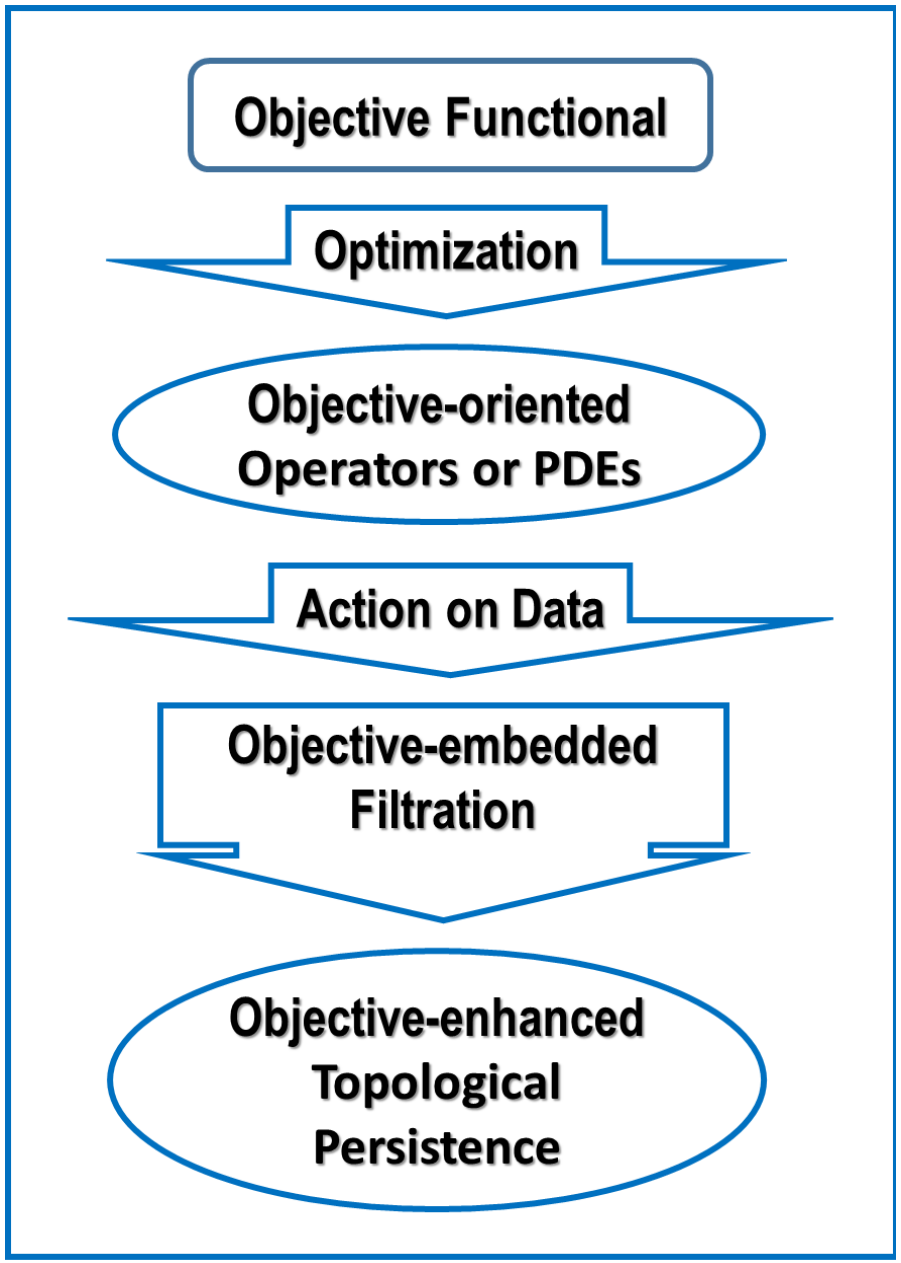




(Xia, Wei, IJNMBE, 2015)

Objective oriented persistent homology

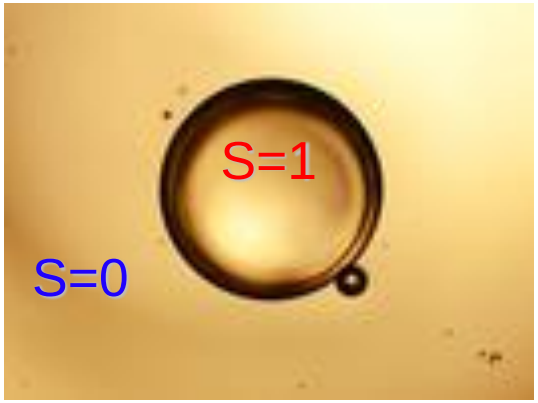
(Wang & Wei, JCP, 2016)



Objective: Minimal surface energy

$$G = \int g[area]\, dr, \quad area = |\nabla S|$$

where gamma (γ) is the surface tension, and S is a surface characteristic function:



Generalized Laplace-Beltrami flow

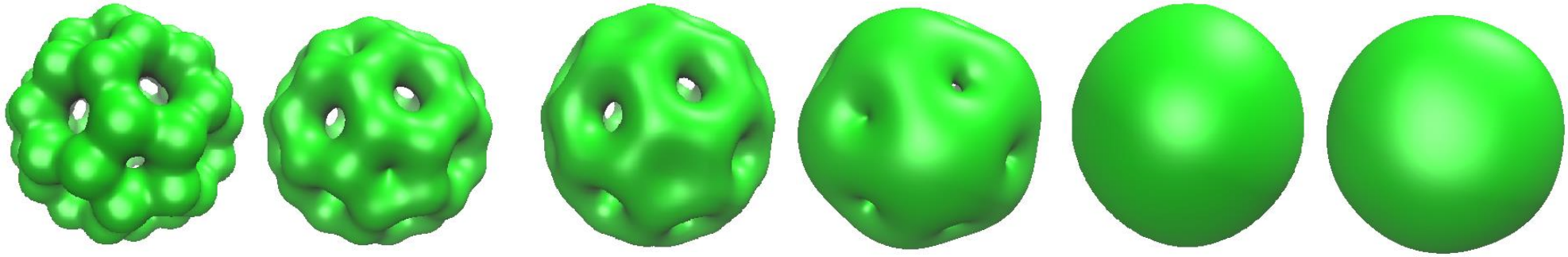
$$\frac{\partial S}{\partial t} = |\nabla S| \left[\nabla \bullet \frac{\gamma \nabla S}{|\nabla S|} \right]$$

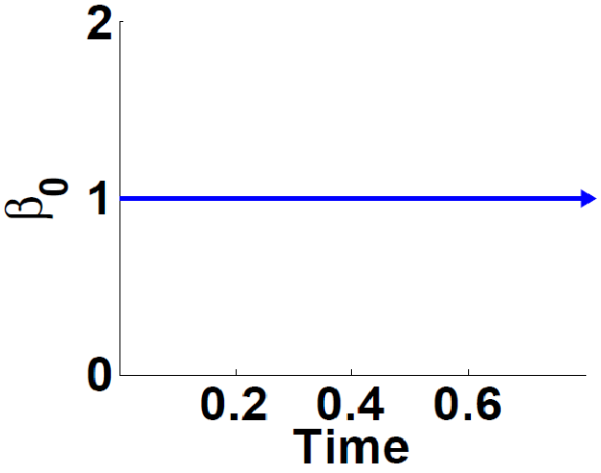
Objective oriented persistent homology

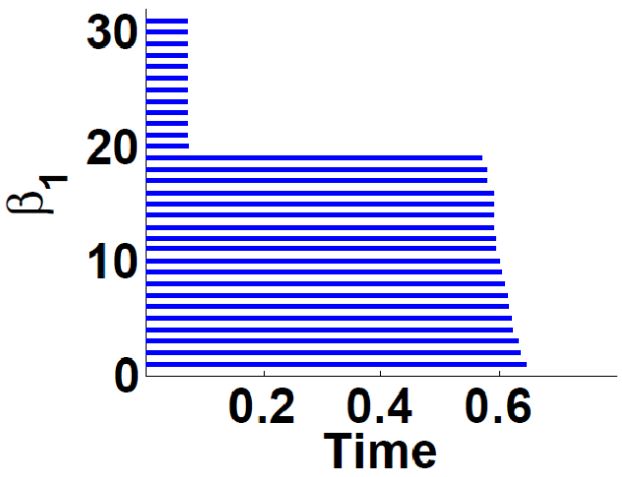
Level sets generated from Laplace-Beltrami flow

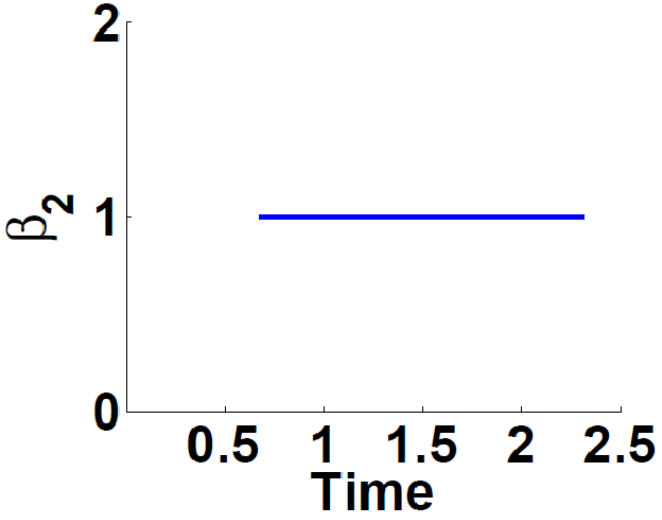
$$\frac{\partial S}{\partial t} = |\nabla S| \left[\nabla \bullet \frac{\gamma \nabla S}{|\nabla S|} \right]$$

(Wang & Wei, JCP, 2016)



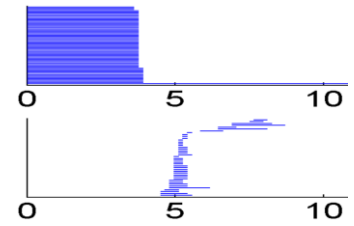


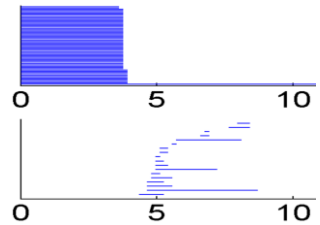


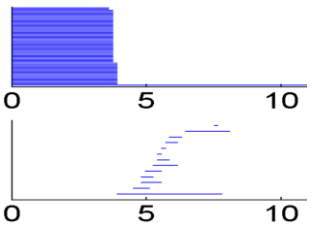


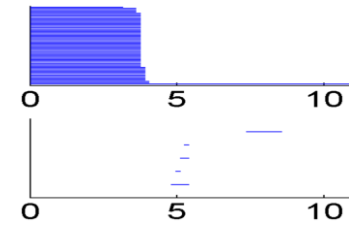
Topological analysis of protein folding

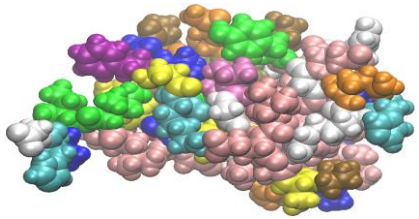
ID: 1I2T

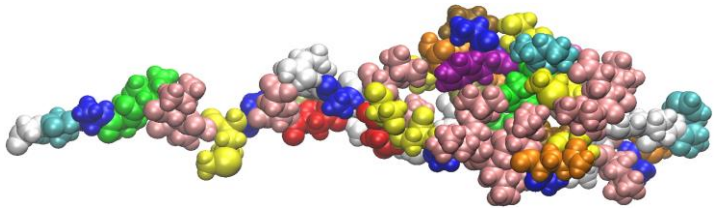


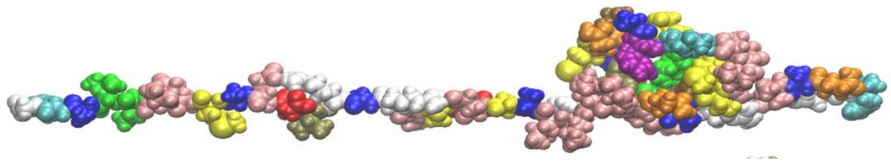


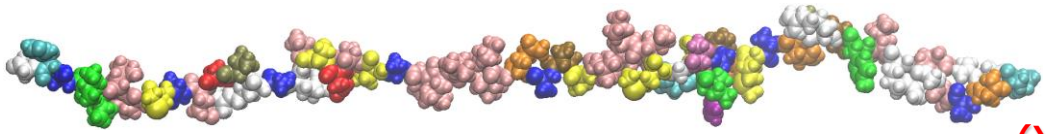


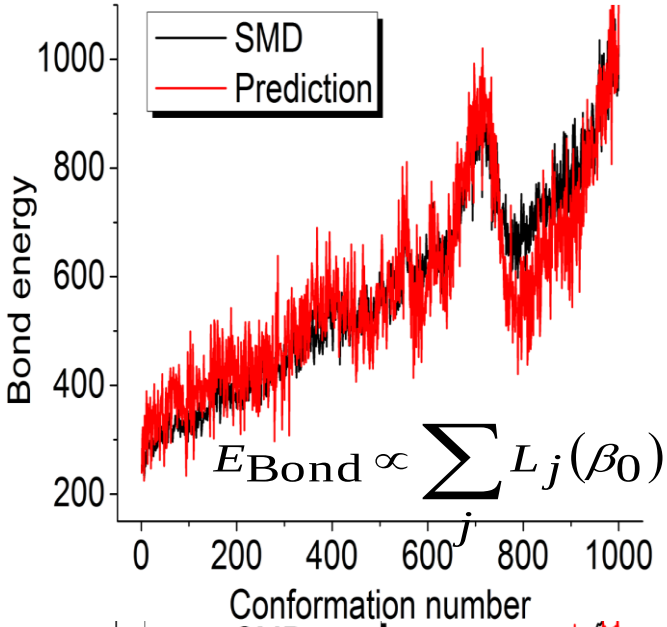


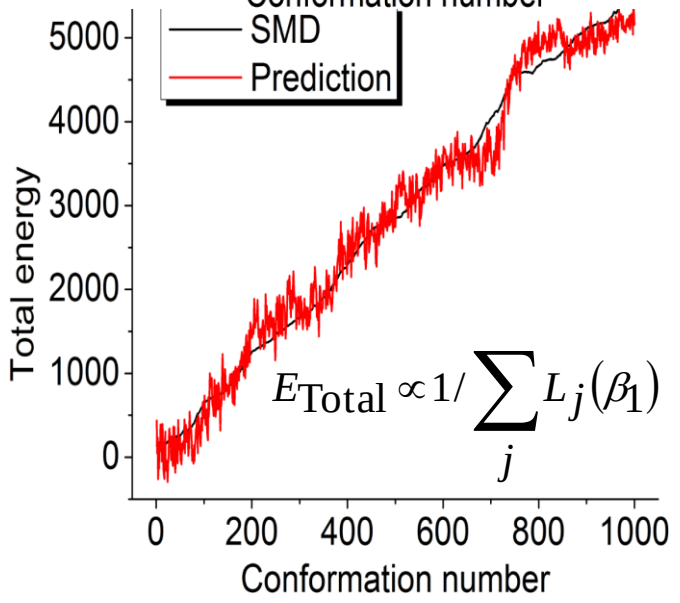








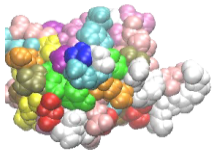


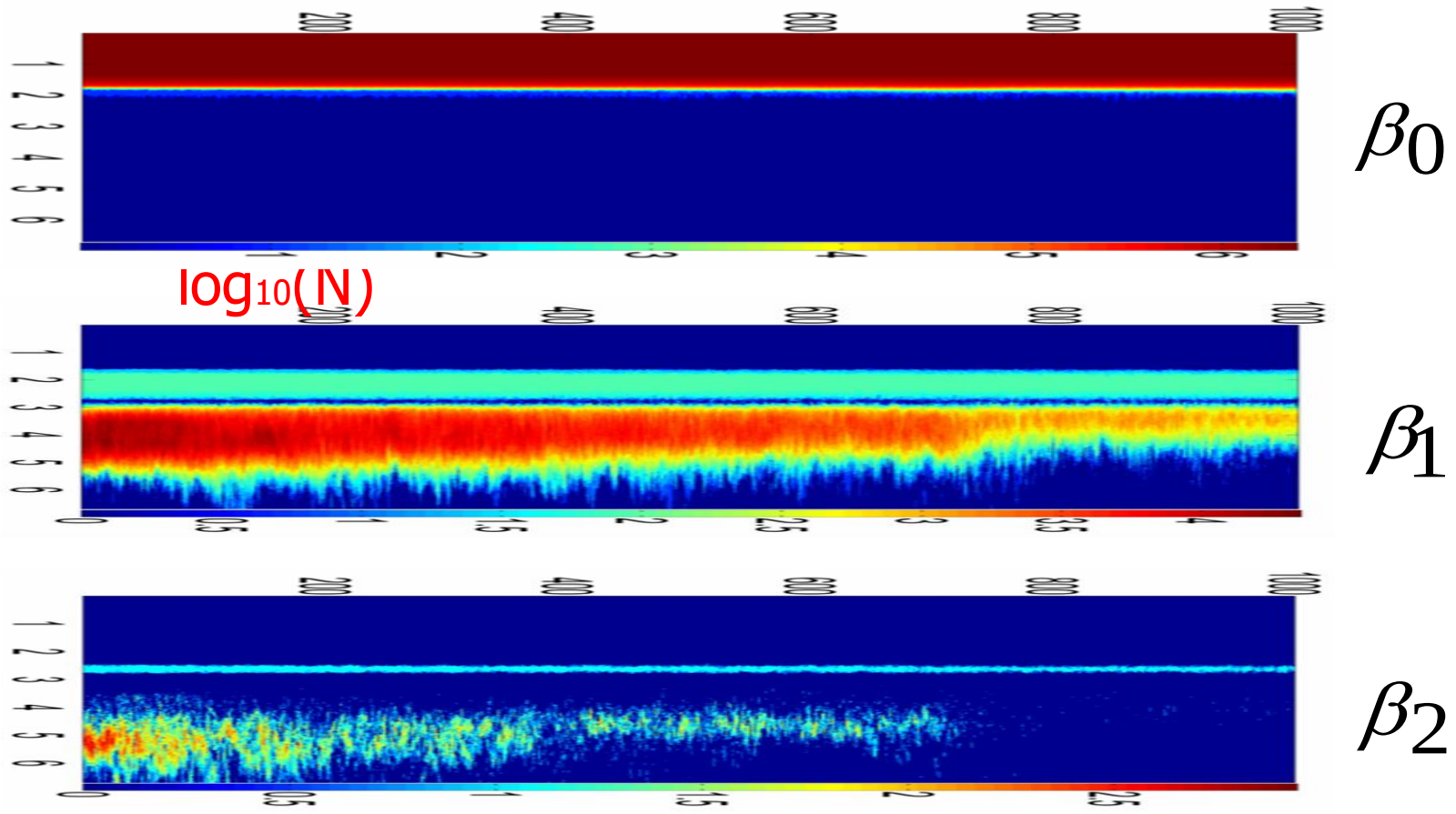


Quantitative!

(Xia, Wei, IJNMBE, 2014)

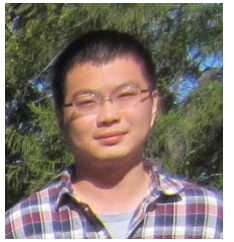
2D persistence in protein 1UBQ unfolding

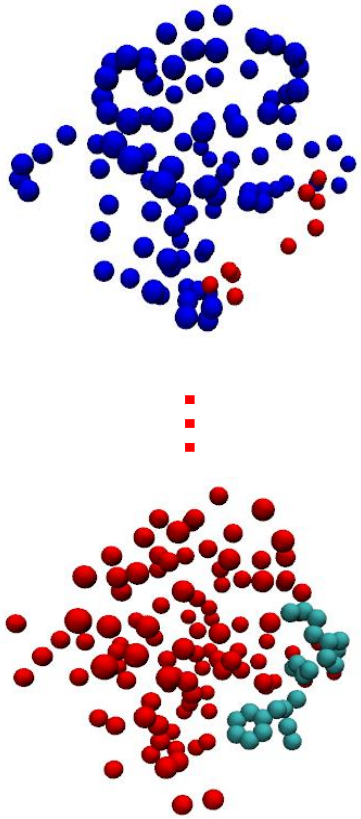


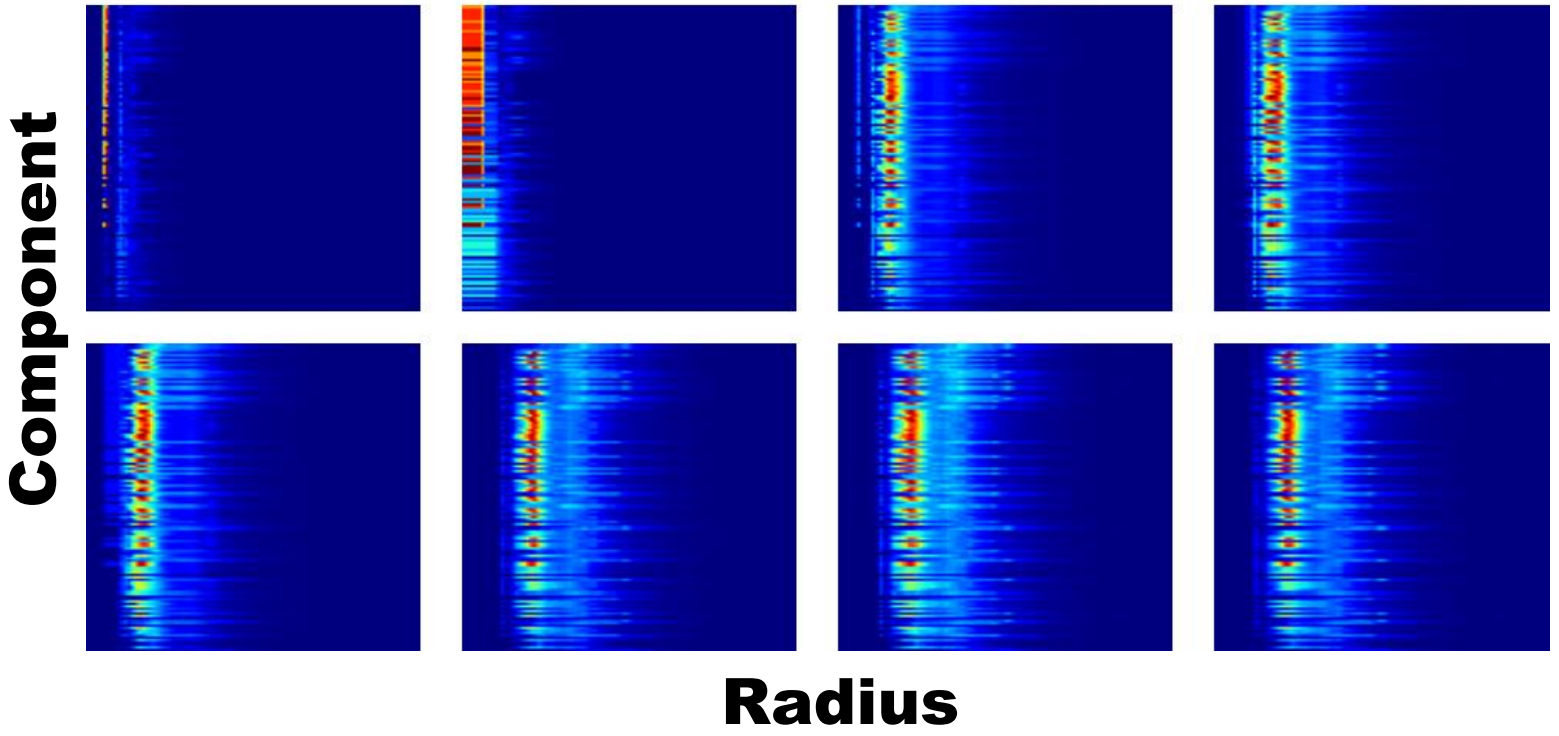


(Xia & Wei, JCC, 2015)

Multicomponent and multichannel persistent homology for a protein-drug complex





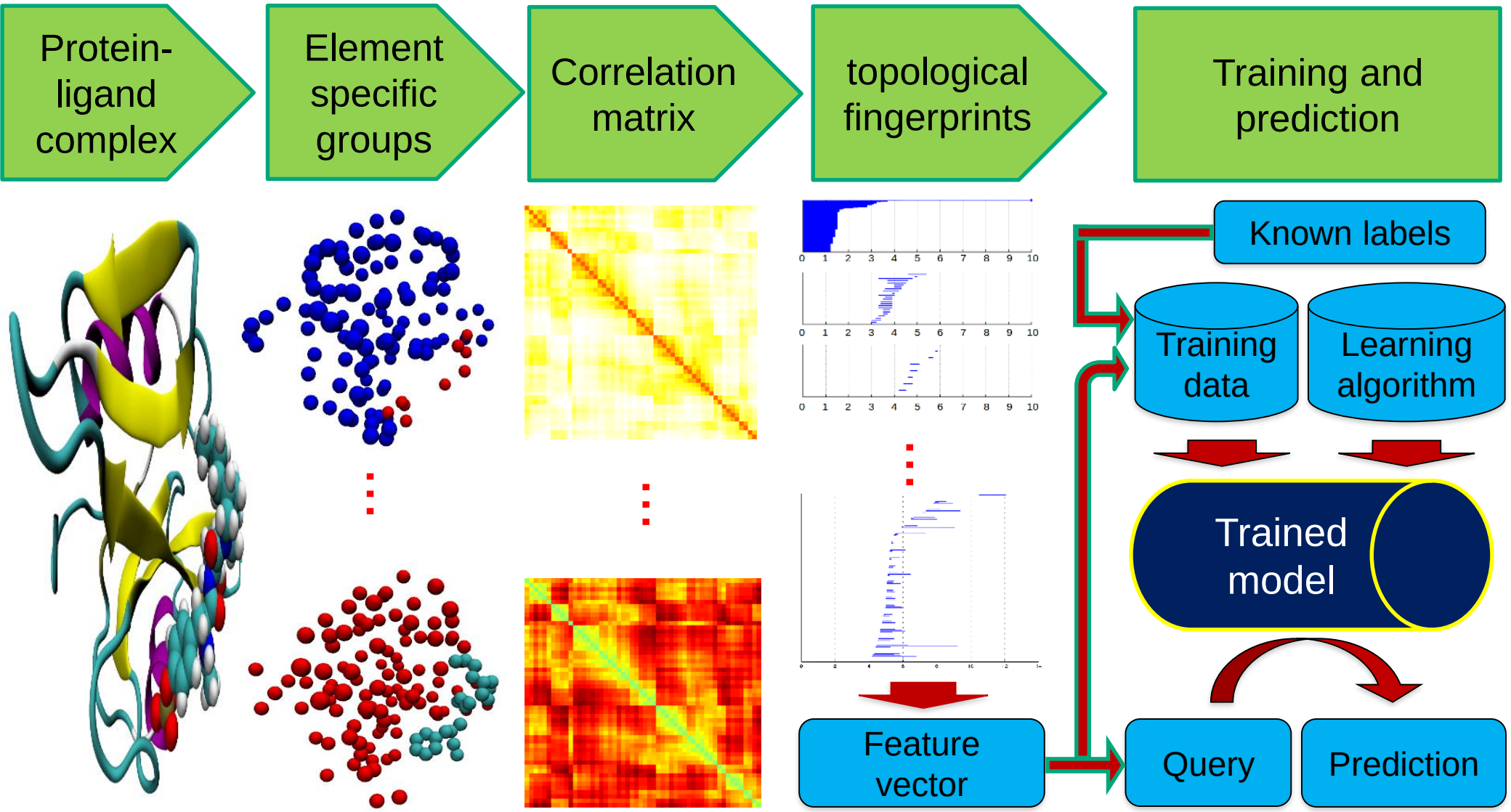


Components are generated from element specific persistent homology. Eight channels are constructed from births, deaths and persistences at Betti-0, Betti-1 and Betti-2.

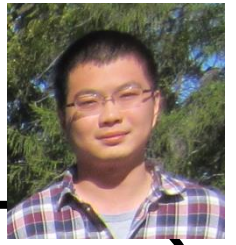
(Cang & Wei, IJNMBE, 2017)

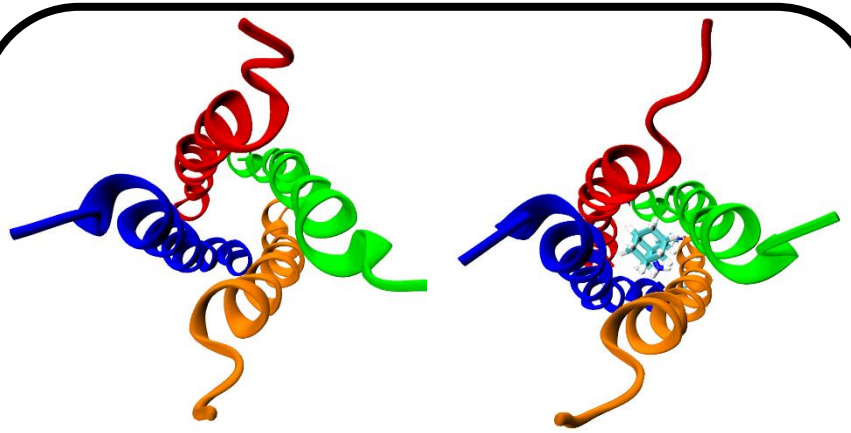
Topology based learning architecture

(Cang & Wei, IJNMBE, 2017)

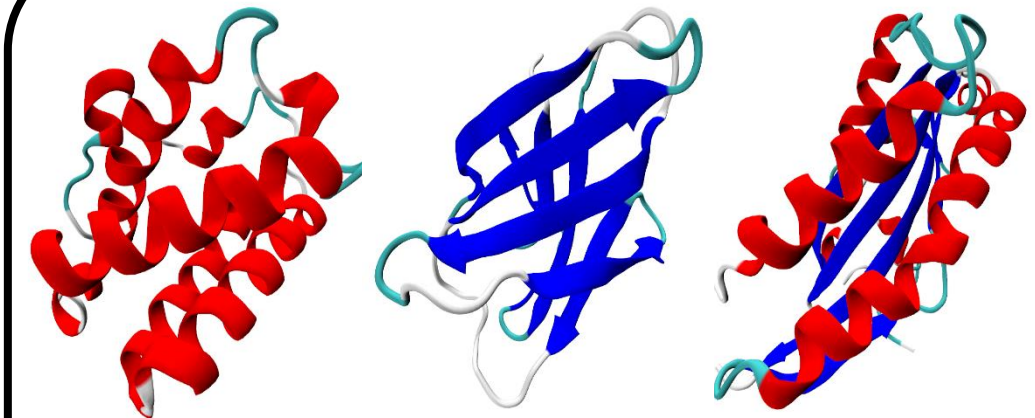


Topological fingerprint based machine learning method for the classification of 2400 proteins



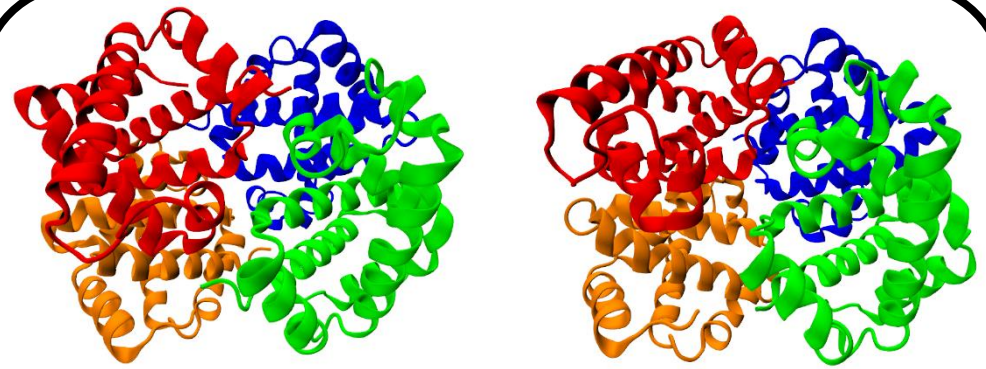


Influenza A virus drug inhibition: 96% Accuracy



Protein domains: 85% Accuracy
(Alzheimer's disease)

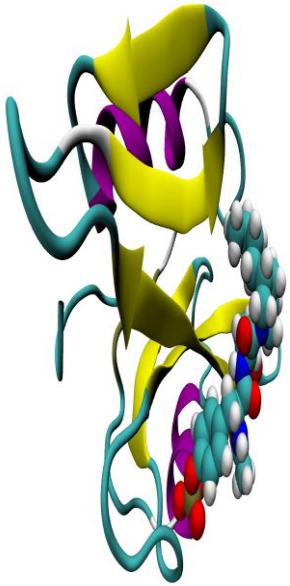
(Cang et al, MBMB, 2015)

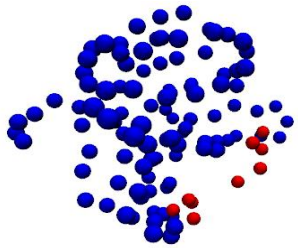


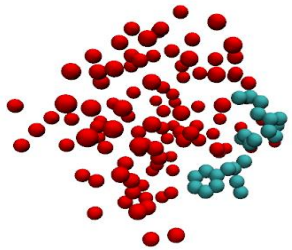
Hemoglobins in their relaxed and taut forms: 80% accuracy

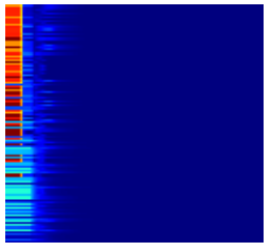
55 classification tasks of protein superfamilies over 1357 proteins from Protein Classification Benchmark Collection: 82% accuracy

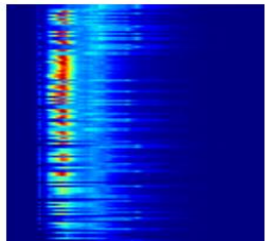
Topological convolutional deep Learning architecture









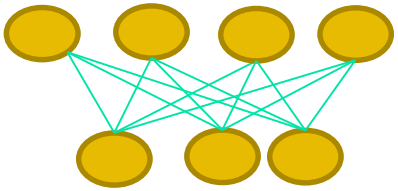


Multichannel images (54x200)

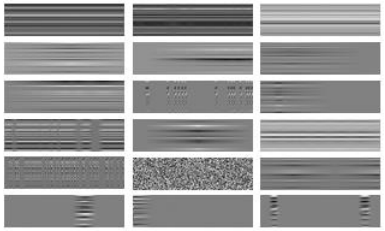
Convolution (128x200)

Pooling (128x100)

Flattening (1xN)



Prediction



Original Complex

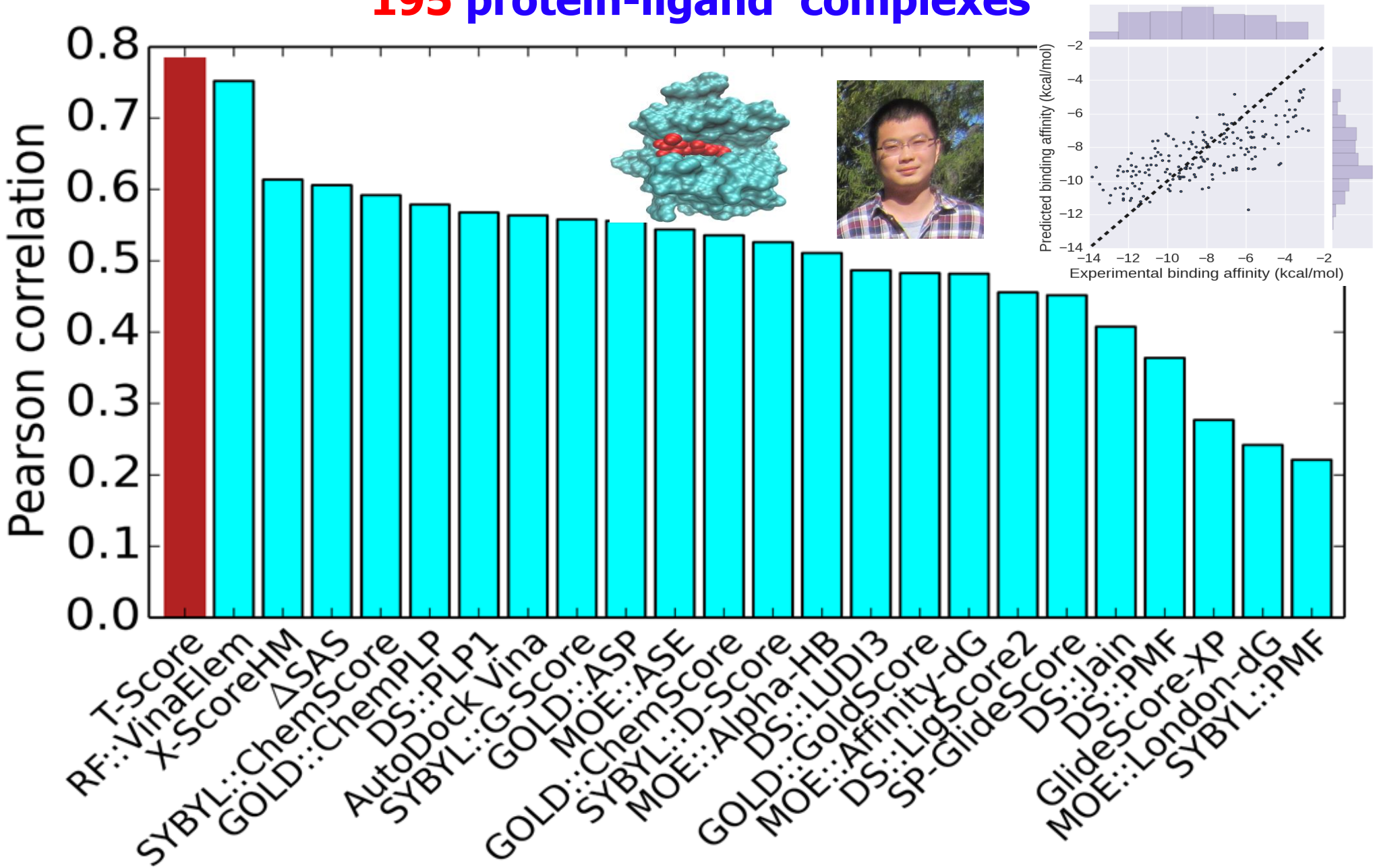
Classify atoms into element specific groups

Generate topological fingerprints

Convolutional deep learning neural network

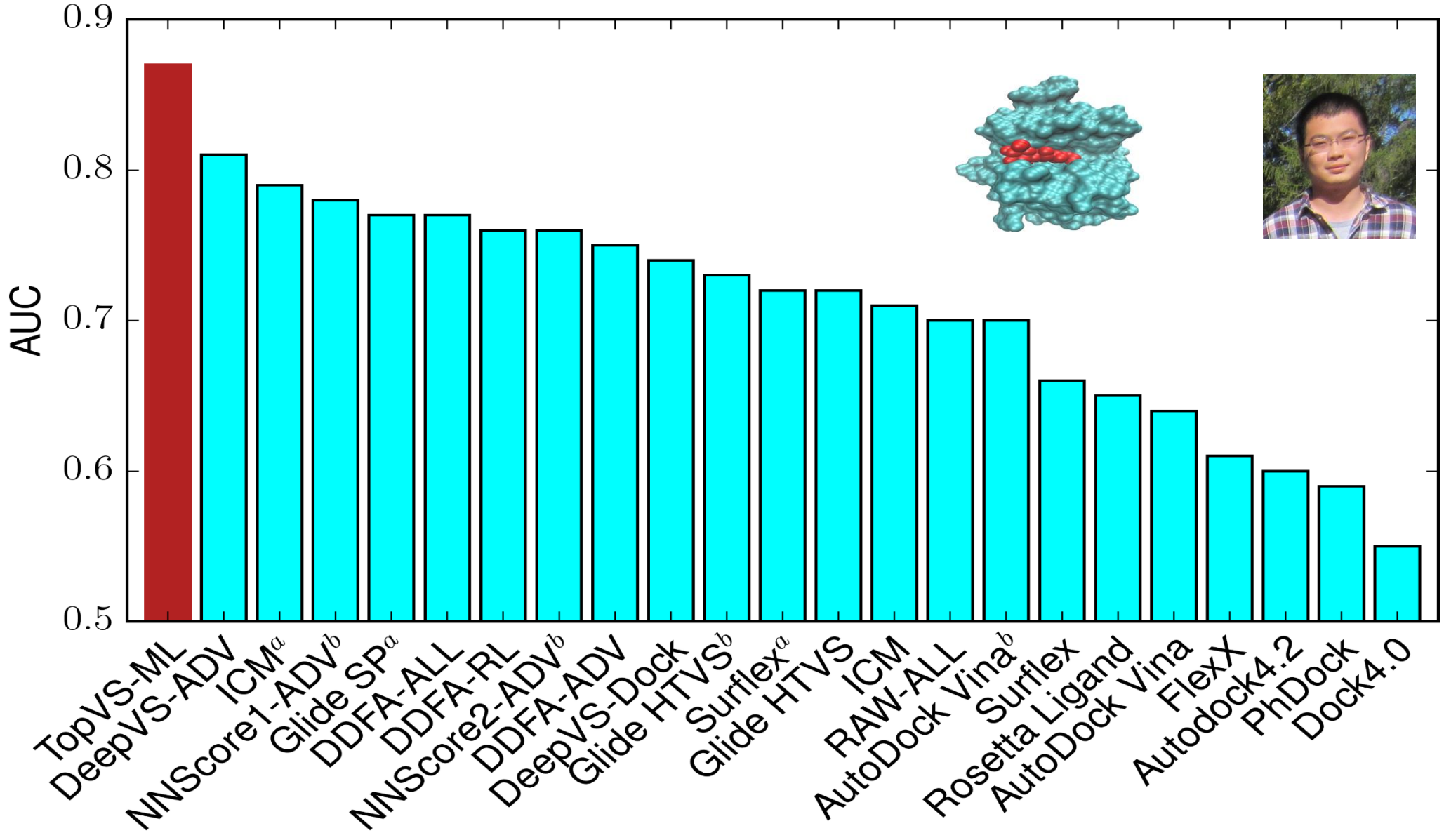
(Cang & Wei, PLOS CB, 2017)

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



Directory of Useful Decoy (DUD)

Classification of 98266 compounds containing 95316 decoys and 2950 active ligands binding to 40 targets from six families



Drug Design & Discovery Resource (D3R) Grand Challenge 2

Given: Farnesoid X receptor (FXR) and 102 ligands

Tasks: Dock 102 ligands to FXR, and compute their poses, binding free energies and energy ranking

Duc Nguyen

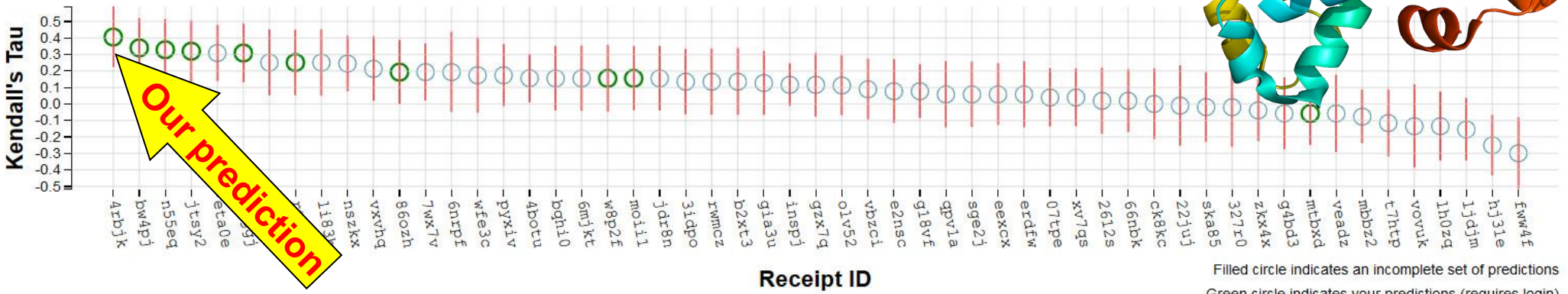
D3R Grand Challenge 2

Given: Farnesoid X receptor (FXR) and 102 ligands

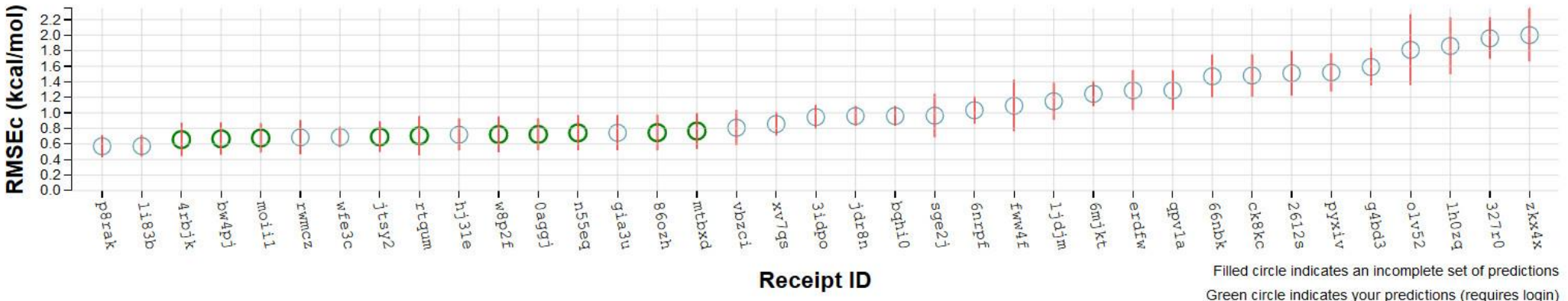
Tasks: Dock 102 ligands to FXR, and compute their poses, binding free energies and energy ranking

Grand Challenge 2

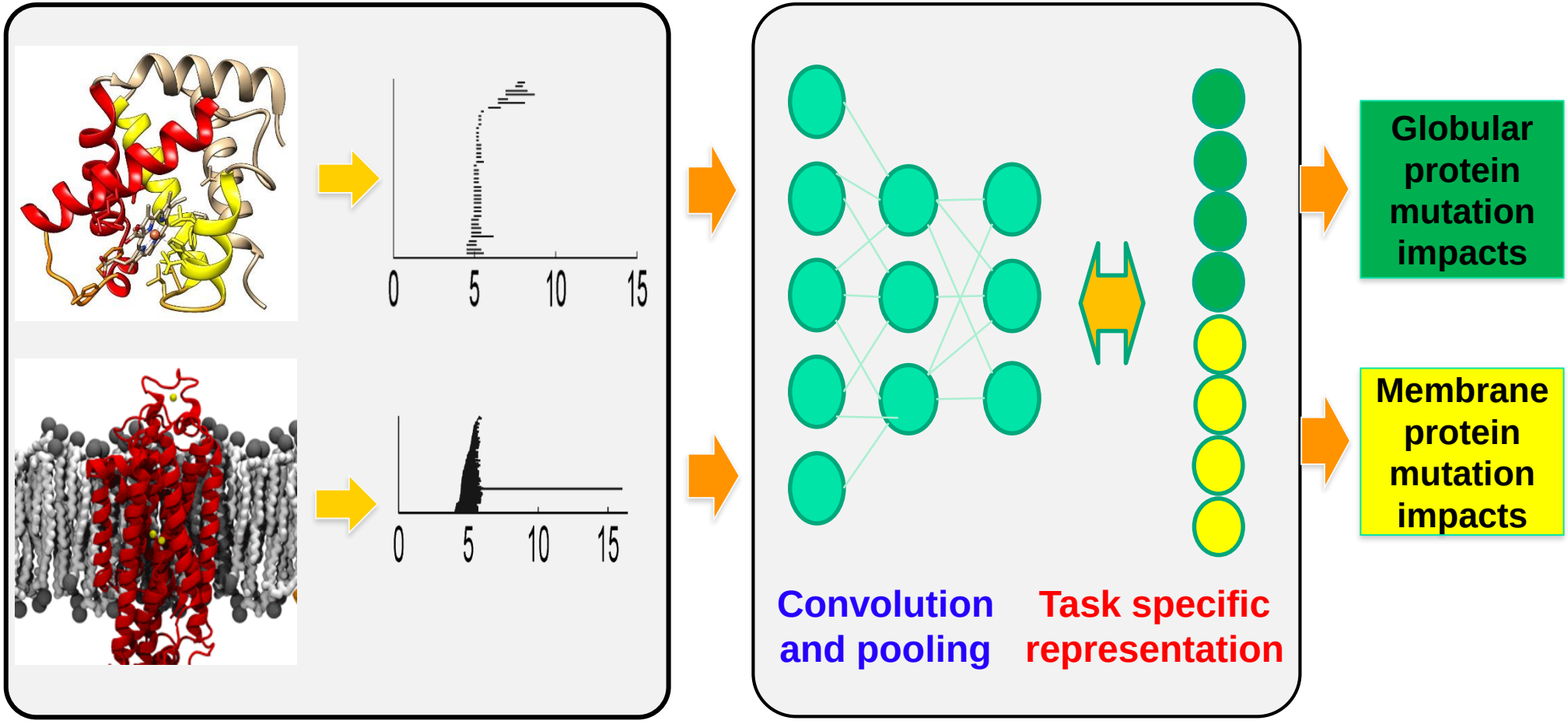
Free Energy Set 1 (Stage 2) - Kendall's Tau



Free Energy Set 1 (Stage 2) - RMSEc



Topological Multi-Task Deep Learning

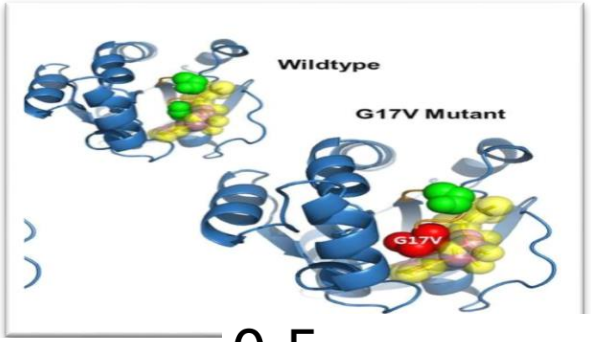


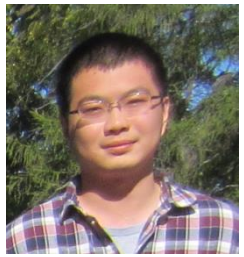
Topological feature extraction

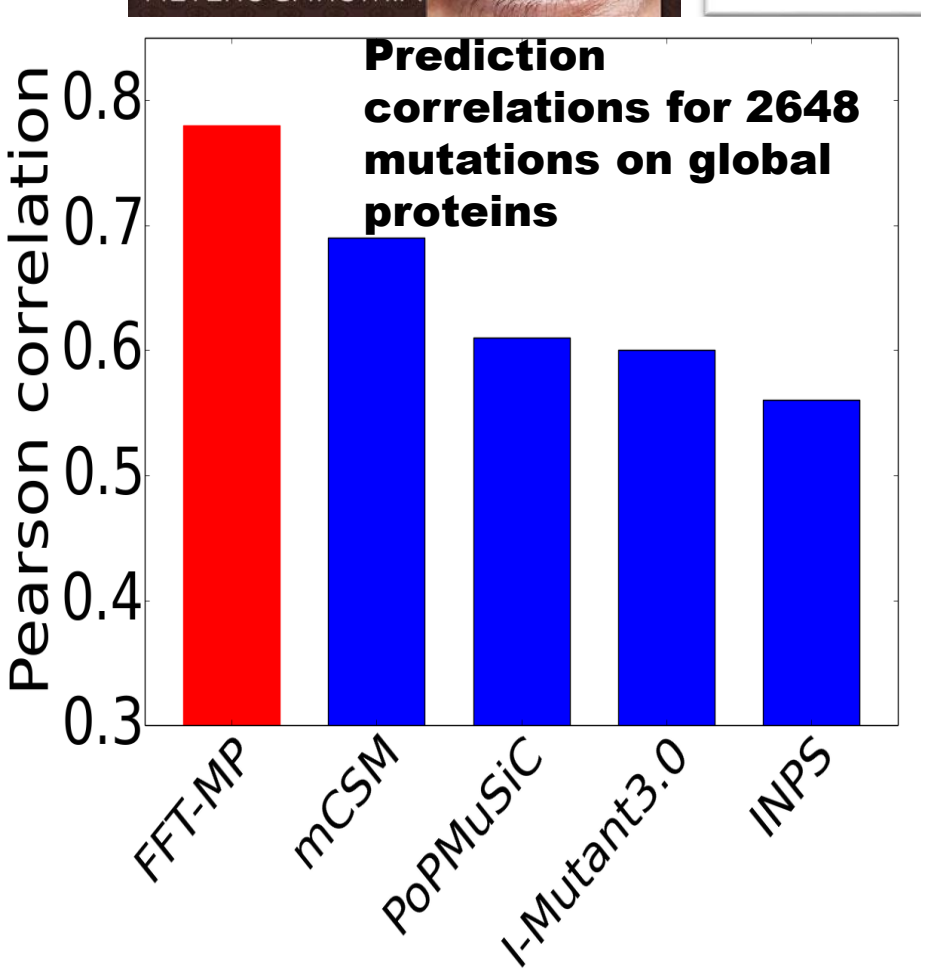
Multi-task topological deep learning

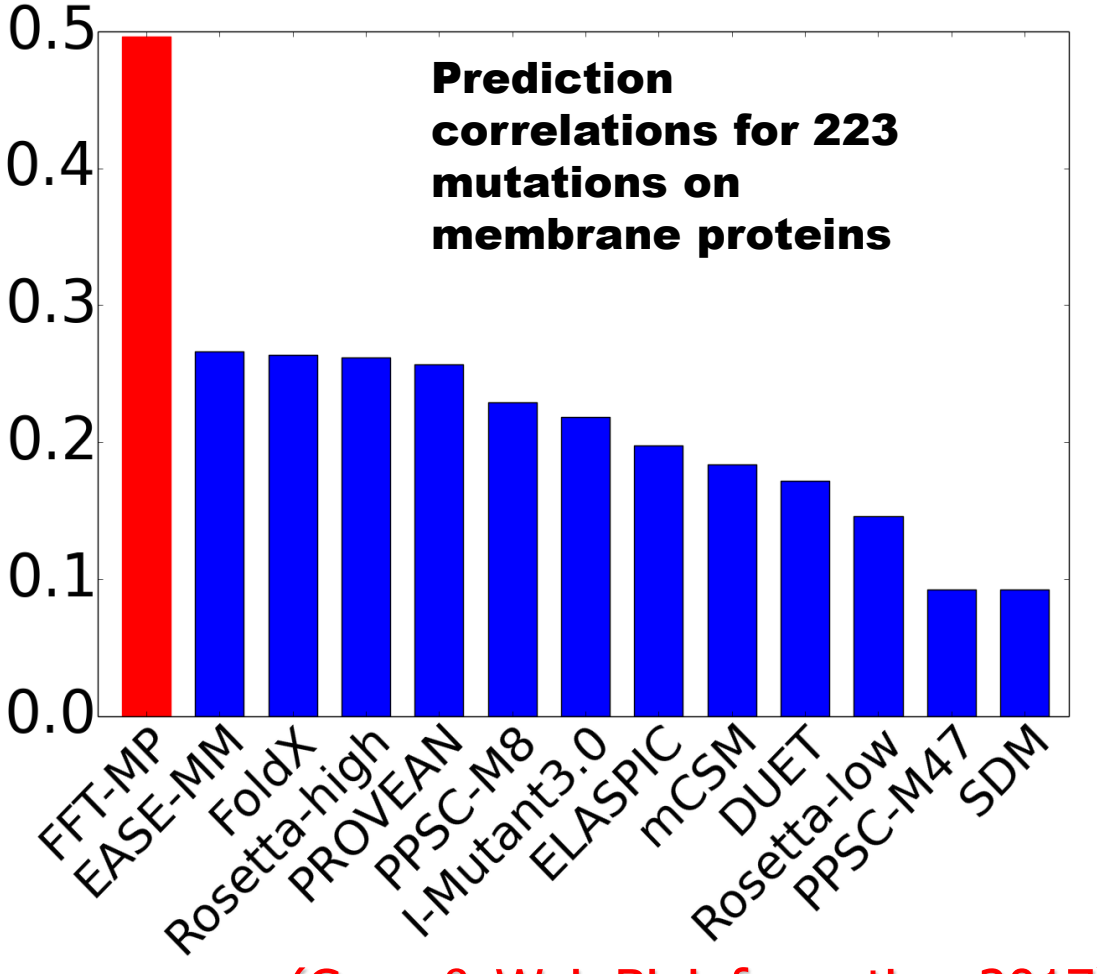
(Cang & Wei, PLOS CB, 2017)

Blind prediction of mutation energies









(Cang & Wei, Bioinformatics, 2017)

Prediction of partition coefficients: Star Set (223 molecules)

Wu, Wang, Zhao, Wang, Wei, 2016

RMSE (log units)

2
1.5
1
0.5
0

FFT-PP
AB/LogP
S+logP
ACD/logP
CLOGP
VLOGP OPS
ALOGPS
MiLogP
XLOGP3
KowWIN
CSLogP
ALOGP
MolLogP
ALOGP98
OsirisP
VLOGP
TLOGP
ABSOLV
QikProp
QuantlogP
SLIPPER
COSMOFrag
XLOGP2
QLOGP
VEGA
CLIP
LSER
MLOGP(Sim+)
NC+NHET
SPARC
MLOGP(Dragon)
LSER UFZ
VLOGP-NOPS
HINT
GBLOGP

Concluding remarks

- Multidimensional, multicomponent, multichannel and objective orientated persistent homologies are introduced to retain essential chemical and biological information during the topological simplification of biomolecular geometric complexity.
- The abovementioned approaches are integrated with advanced machine learning, including deep learning, to achieve the state-of-the-art predictions of protein-ligand binding affinities & ranking, mutation induced protein stability changes, and drug partition coefficients.

Take home messages

- Molecular based mathbio (3 NSF-Simons Centers)
- Topological data analysis
- Machine learning


N Baker
(PNNL)
P Bates
(MSU)
Z Burton
(MSU)
K Dong
(MSU)
M Feig
(MSU)
H Hong
(MSU)
J Hu
(MSU)
Y Tong
(MSU)
X Ye
(UKLR)
Shan Zhao
Alabam
Y Zhou
CSU
W Geng
SMU
Duan Chen
UNCC
Zhan Chen
GSU
Lin
ONL
Mu
S Yang
CU KH
S.N. Yu
Boston
Y.H. Sun
Denver
L. H. Hu
Chicago