Is it time for a great chemistry between mathematics and biology?

Guowei Wei

Departments of Mathematics

Michigan State University

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

Workshop on Geometry, Imaging and Computing
Center of Mathematical Sciences and Applications
Harvard University
March 24-26

Grant support: NIH, NSF, MSU and BMS

Mathematics and Natural Sciences

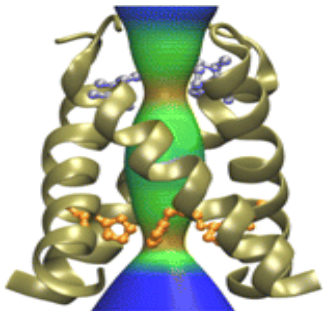
- **Mathematics is the foundation for Newtonian mechanics, Hamiltonian mechanics, Maxwell's electromagnetic theory, Boltzmann theory, statistical mechanics, thermodynamics, Einstein's theory of relativity, and quantum mechanics.**
- **Nobel Prize winner Eugene Wigner: "The Unreasonable Effectiveness of Mathematics in the Natural Sciences".**
- **Mathematics has got more abstract since 1950s while biology became microscopic in 1960s.**
- **Biology assumed an omics dimension (i.e., big data) around 2000.**
- **The power of machine learning and deep learning has burst since 2014.**
- **Biological sciences are undergoing a historic transition from qualitative, phenomenological, and descriptive to quantitative, analytical and predictive, as quantum physics did a century ago.**
- **It is time to invent *biology-inspired math* and discover *math-governed rules of life!***

Drug design and discovery

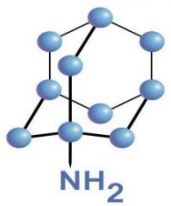
1) Disease identification (physiology)
2) Target hypothesis (biochem./mole. biol.)
3) Virtual screening: binding affinity, solubility, partition coefficient, toxicity, and side-effects (biophysics/bioinformatics)
4) Drug structural optimization in the target binding site (biochemistry/biophysics/synthetic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

Influenza -- flu virus

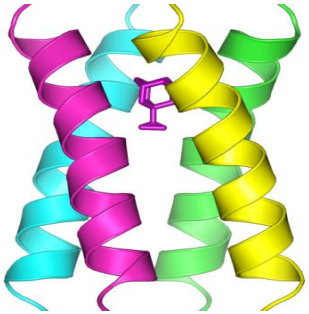
M2 channel

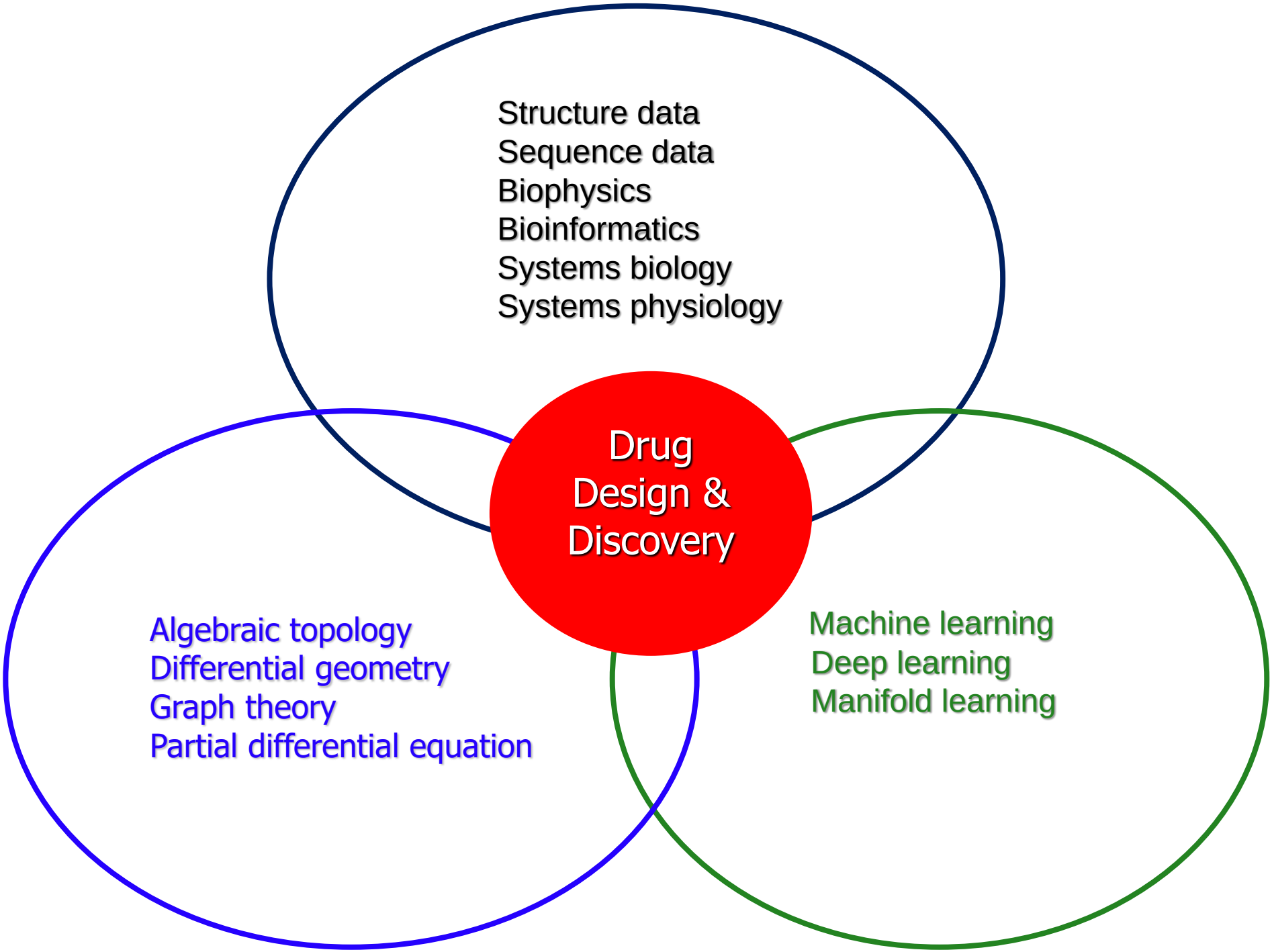


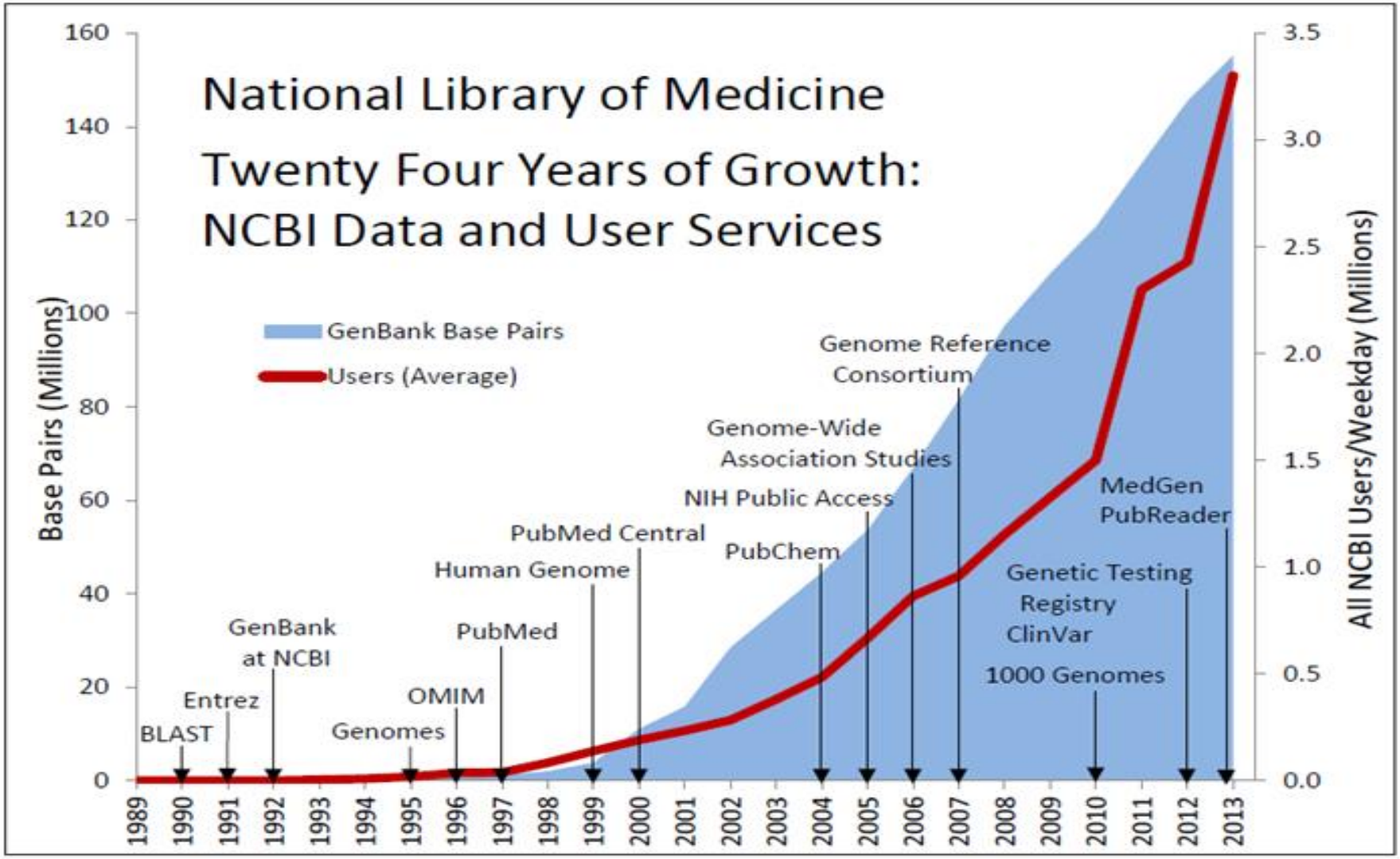
Amantadine



M2-A complex

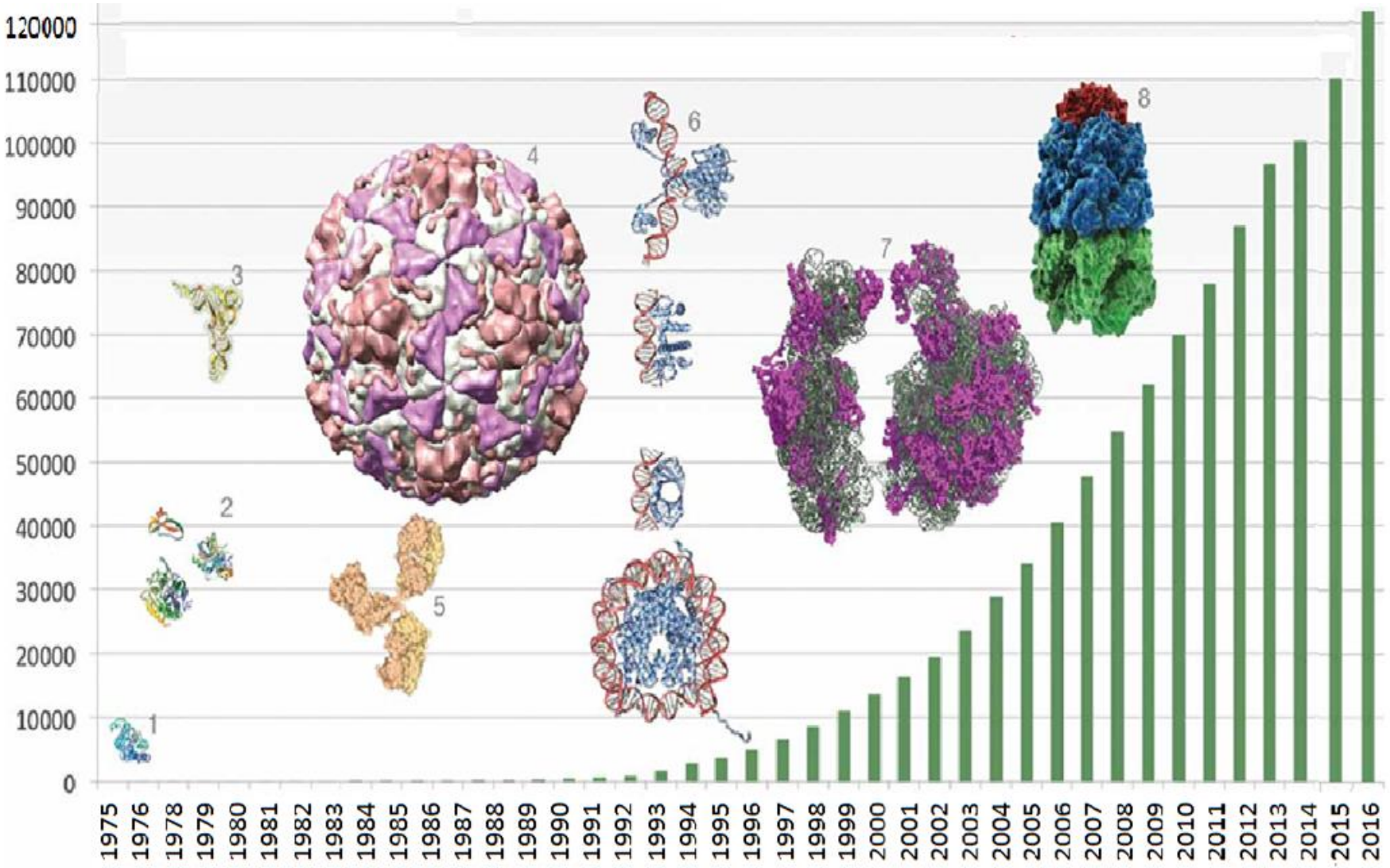



Structure data
Sequence data
Biophysics
Bioinformatics
Systems biology
Systems physiology
Drug
Design &
Discovery
Algebraic topology
Differential geometry
Graph theory
Partial differential equation
Machine learning
Deep learning
Manifold learning



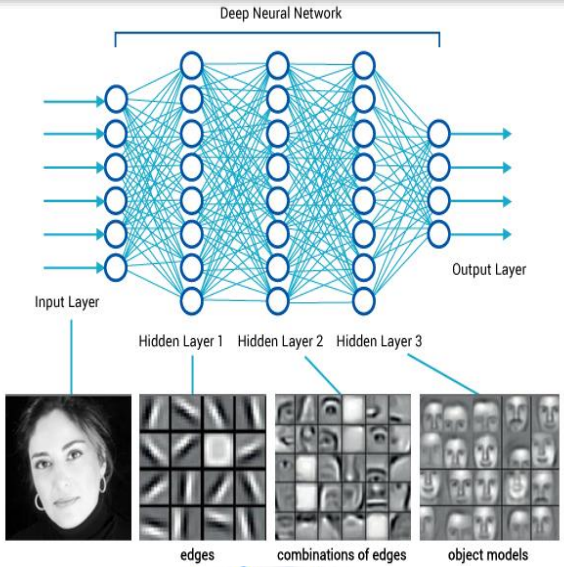
		GenBank		Whole Genome Shotgun	
Release	**Date**	**Bases**	**Sequences**	**Bases**	**Sequences**
224	Feb 2018	253630708098	207040555	2608532210351	564286852

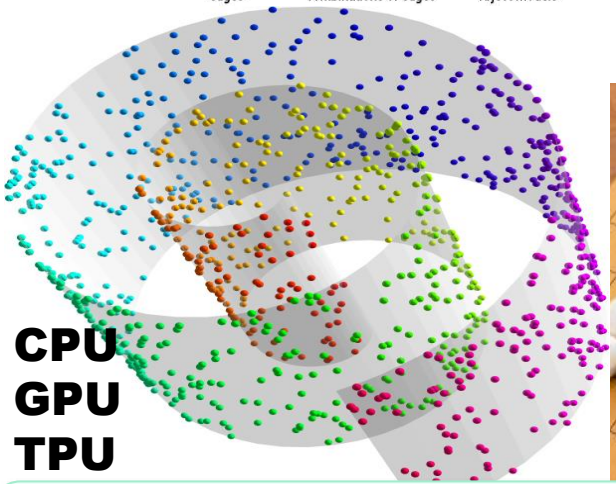
Yearly Growth of Total Structures in the Protein Data Bank

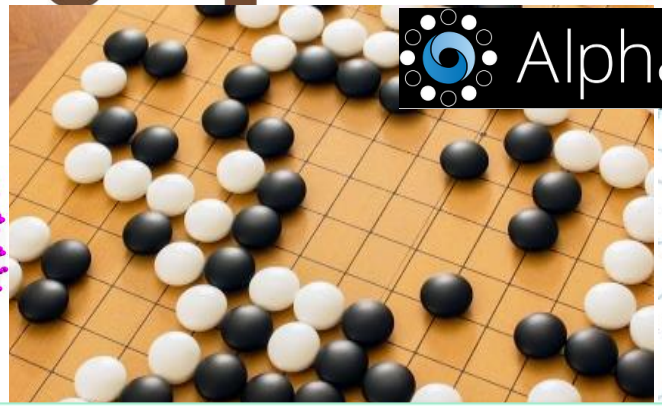


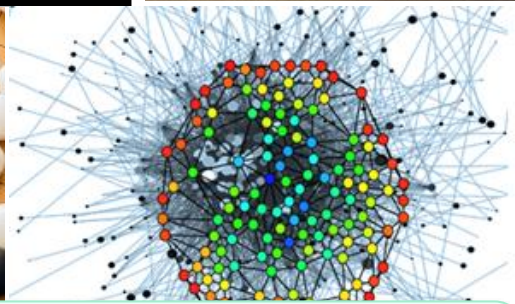
Mach 25, 2018: 138,878

Welcome to big-data era





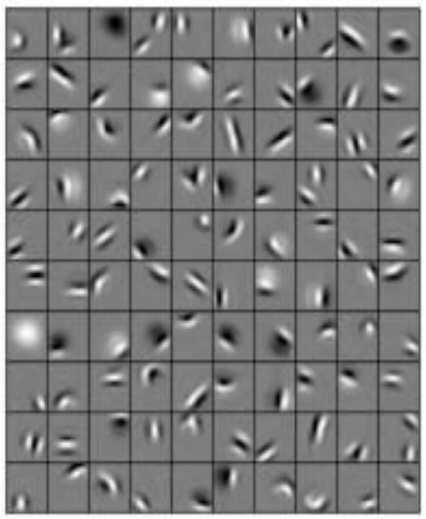


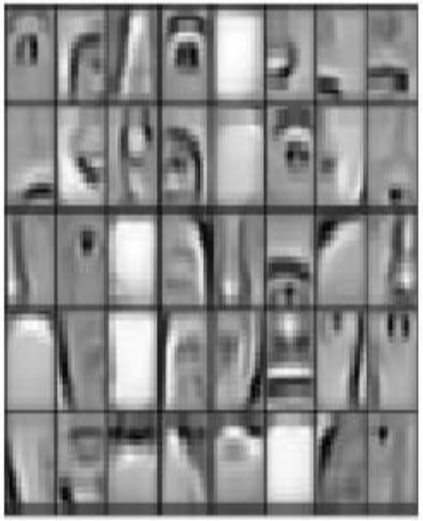


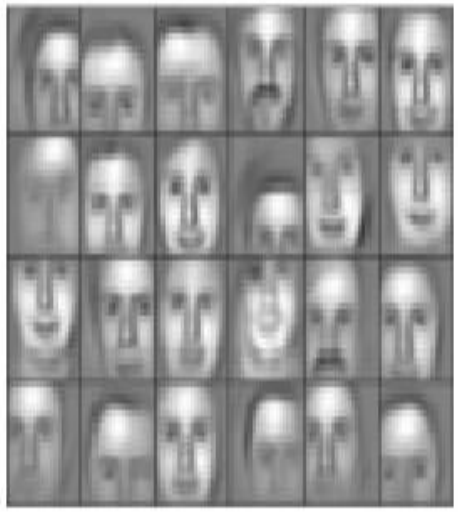
Half of all jobs will be done by robots in the near future

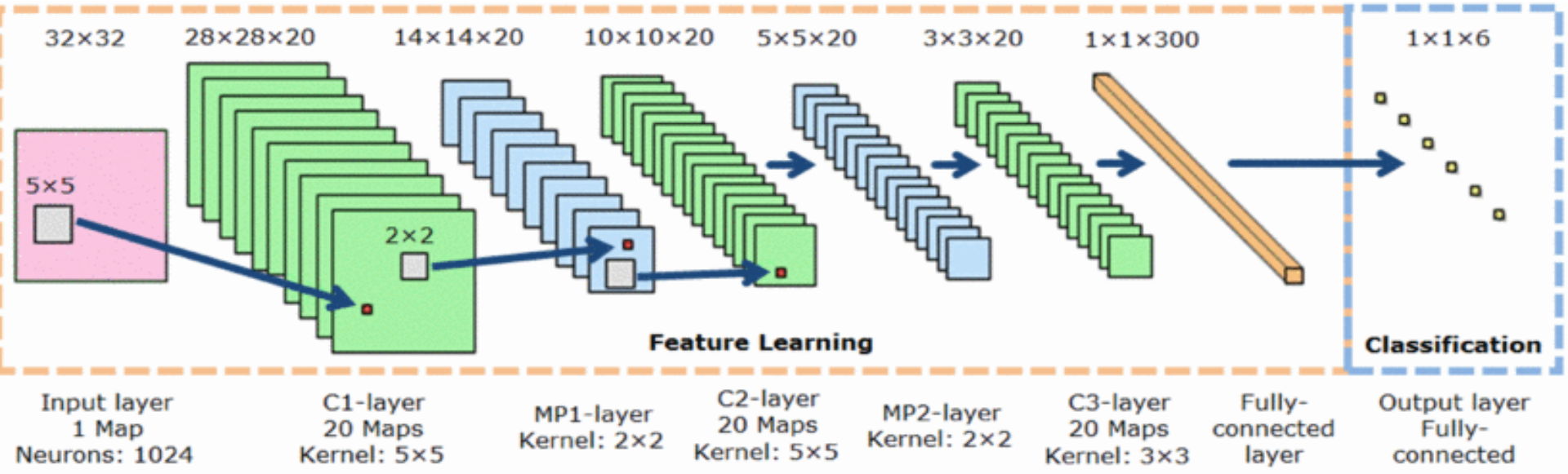
Deep learning

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









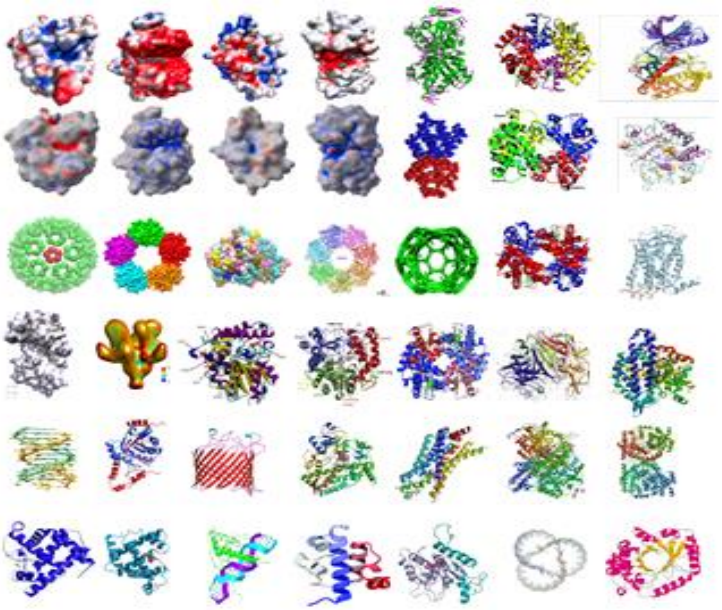
How to do deep learning for 3D biomolecular data?

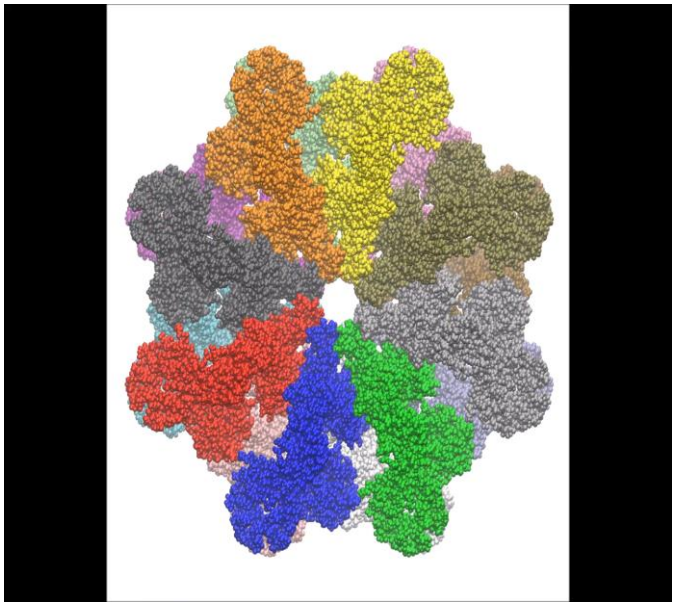
Obstacles for deep learning of 3D biomolecules:

- **Geometric dimensionality: R^{3N}, where N~5500 for a protein.**
- **Machine learning dimensionality: > 1024^{3m}, 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**

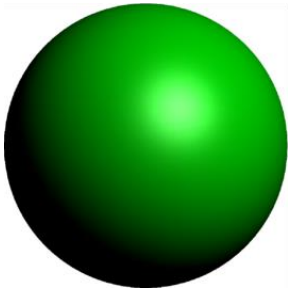


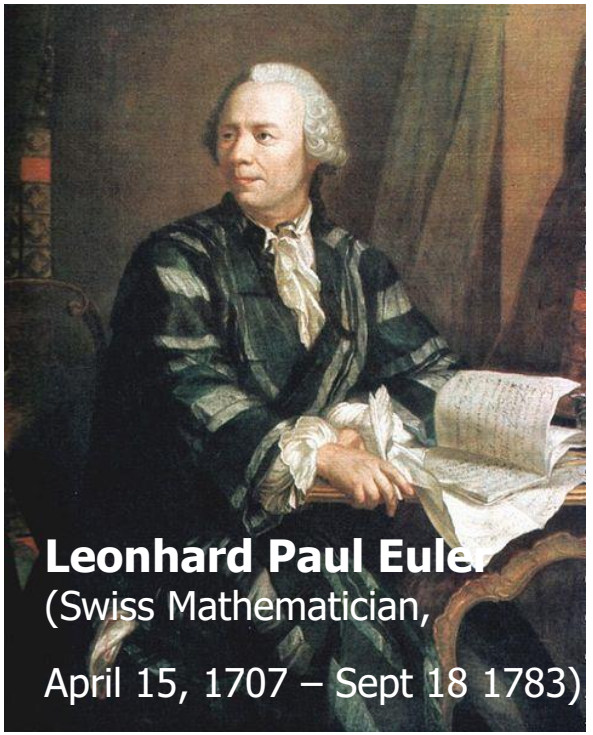


Classical topological objects

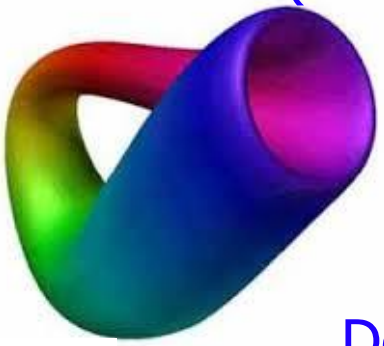
Möbius Strips (1858)

Sphere



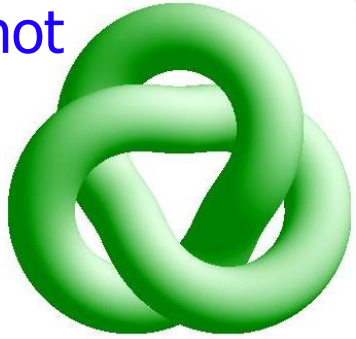


Klein Bottle (1882)



Torus

Trefoil Knot



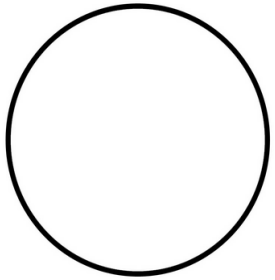
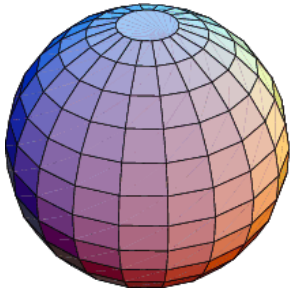
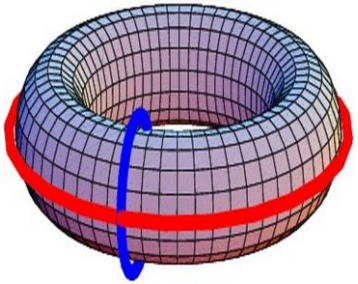
Double Torus

Seven Bridges of Königsberg

Leonhard Euler (1735)

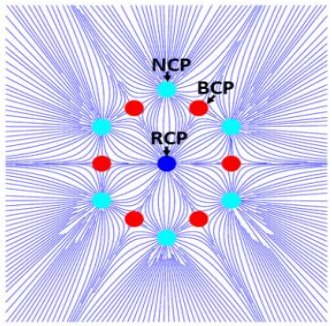
Topological invariants: Betti numbers

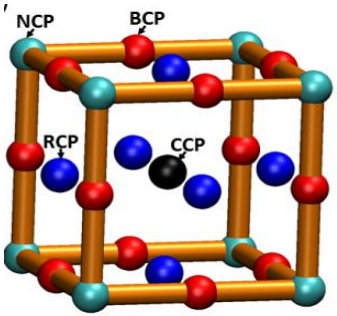
β_0 is the number of connected components.
β_1 is the number of tunnels or circles.
β_2 is the number of cavities or voids.

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

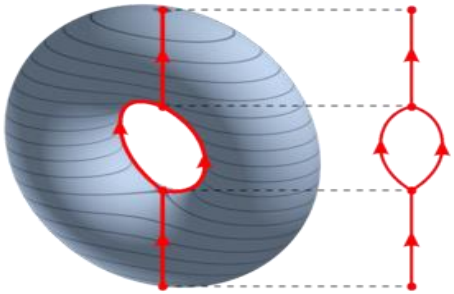
Topological simplification

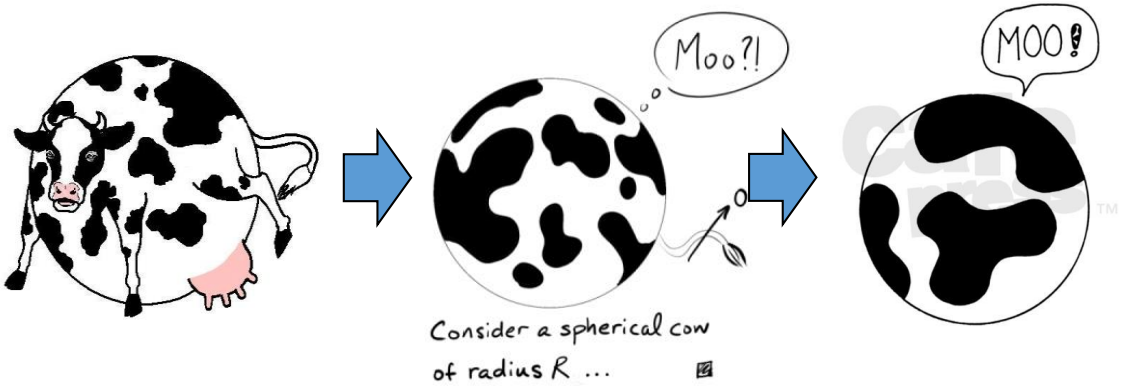
Poincare-Hopf index



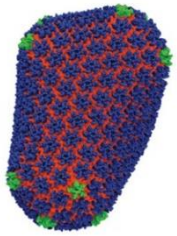


Morse theory

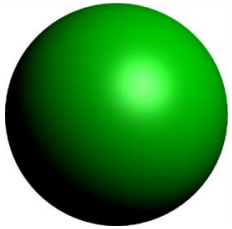




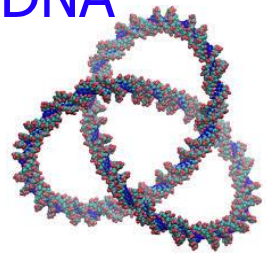
Mug ⇔ **Doughnut**



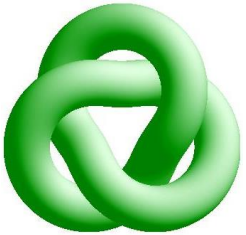
HIV 4.2 million atoms



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:

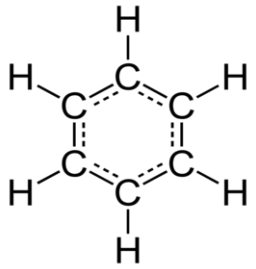
- ➢ **Geometric methods are often inundated with too much structural detail.**
- ➢ **Topological tools incur too much reduction of original geometric information.**
- ➢ **Topology is hardly used for quantitative prediction.**

Promises from persistent homology:

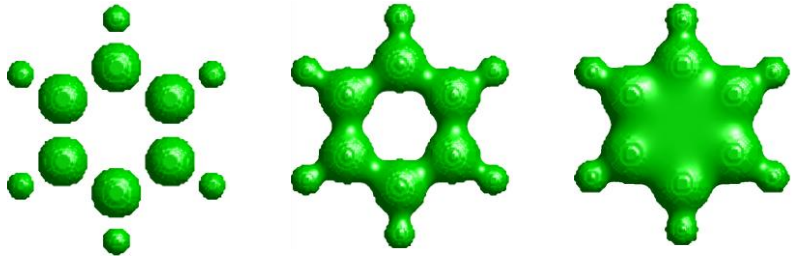
- ✓ **Embeds geometric information in topological invariants.**
- ✓ **Bridges the gap between geometry and topology.**

Persistent homology answers following questions

What is the topology of a benzene?

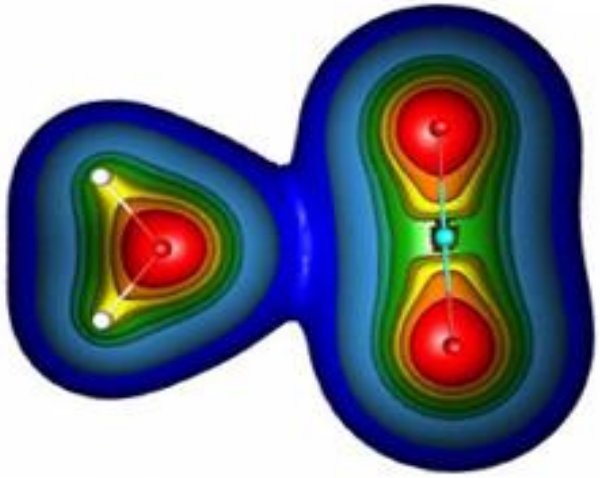


Level sets generated by Laplace-Beltrami flows:



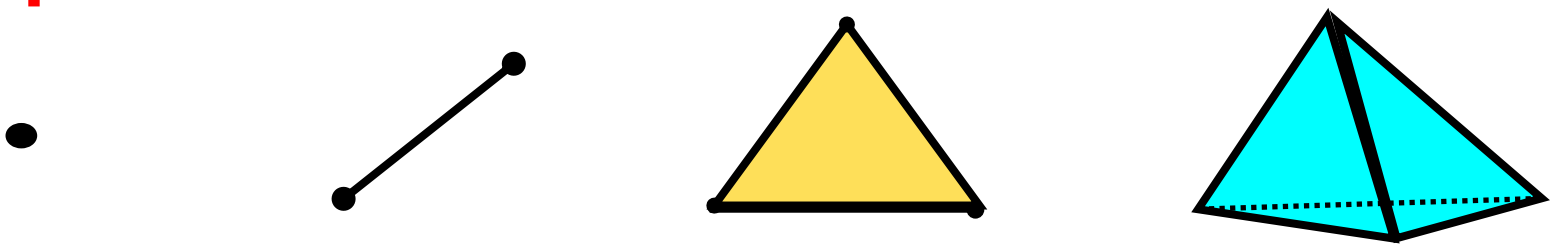
What is the topology of a H_2O-CO_2 complex?

Electron density level sets computed by using quantum mechanics:



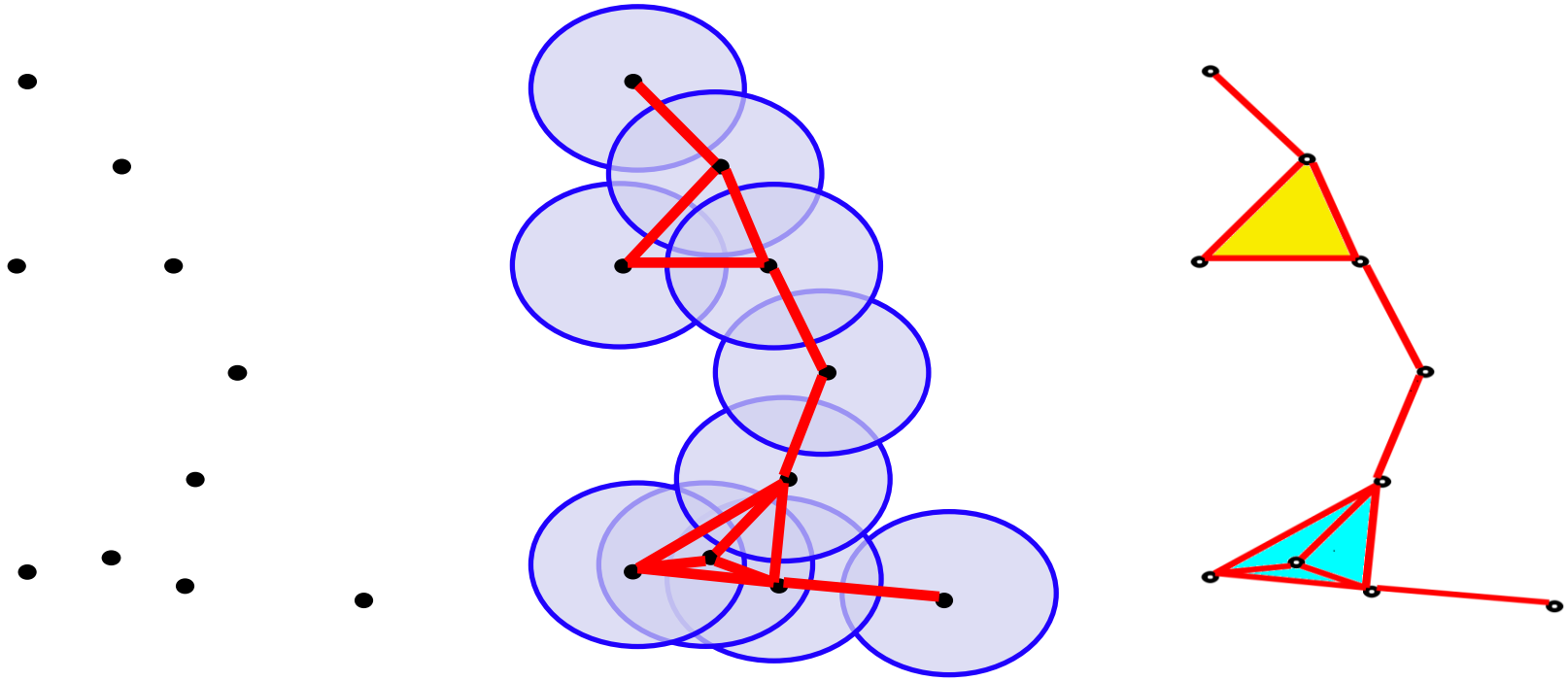
Vietoris-Rips complexes of planar point sets

Simplexes:



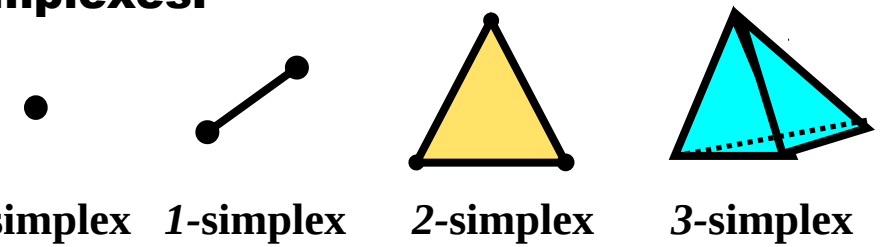
0-simplex *1*-simplex *2*-simplex *3*-simplex

Simplicial complexes of ten points:



Topological modeling - Persistent homology

Simplexes:



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

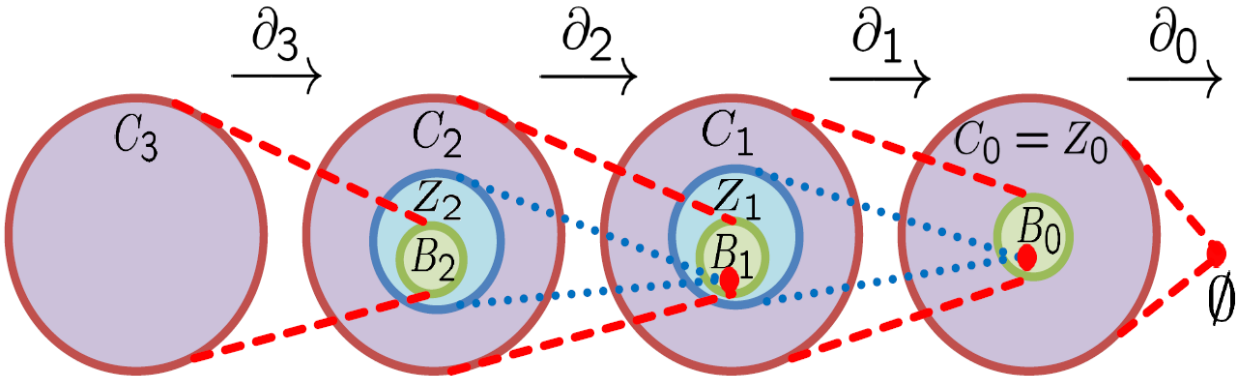
k-chain: $\sum_i c_i \sigma_i^k$

Chain group: $C_k(K, Z_2)$

Boundary operator:

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

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

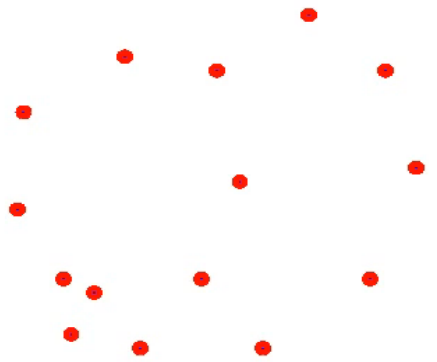


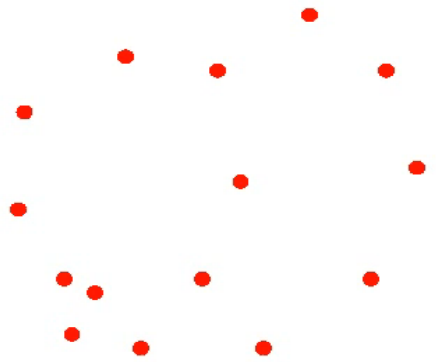
Filtration

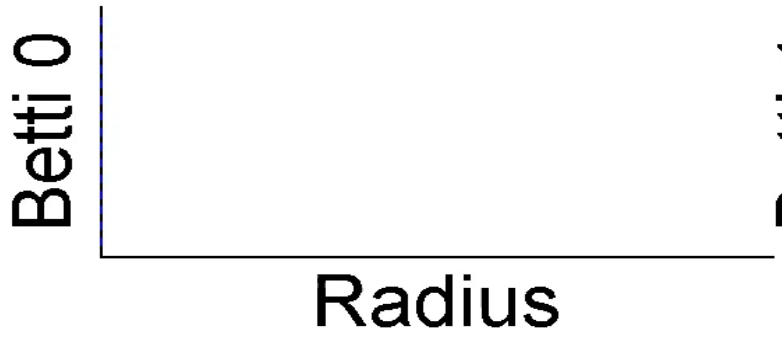
Frosini and Nandi (1999),
Robins (1999),
Edelsbrunner, Letscher and Zomorodian (2002), Edelsbrunner and Harer, (2007)
Kaczynski, Mischaikow and Mrozek (2004),
Zomorodian and Carlsson (2005),
Ghrist (2008),
.....

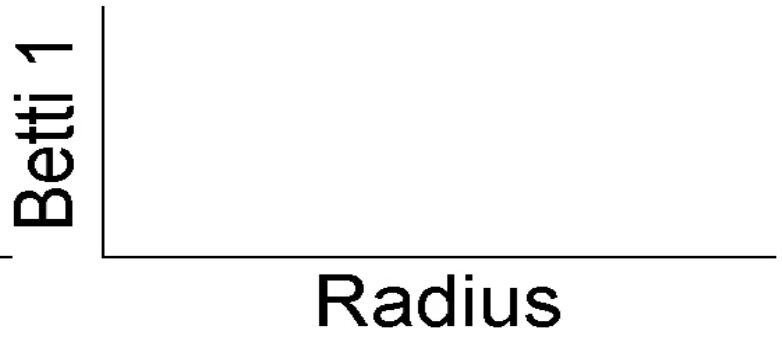
Vietoris-Rips complexes, persistent homology and persistent barcodes

(Xia, Wei, 2014)

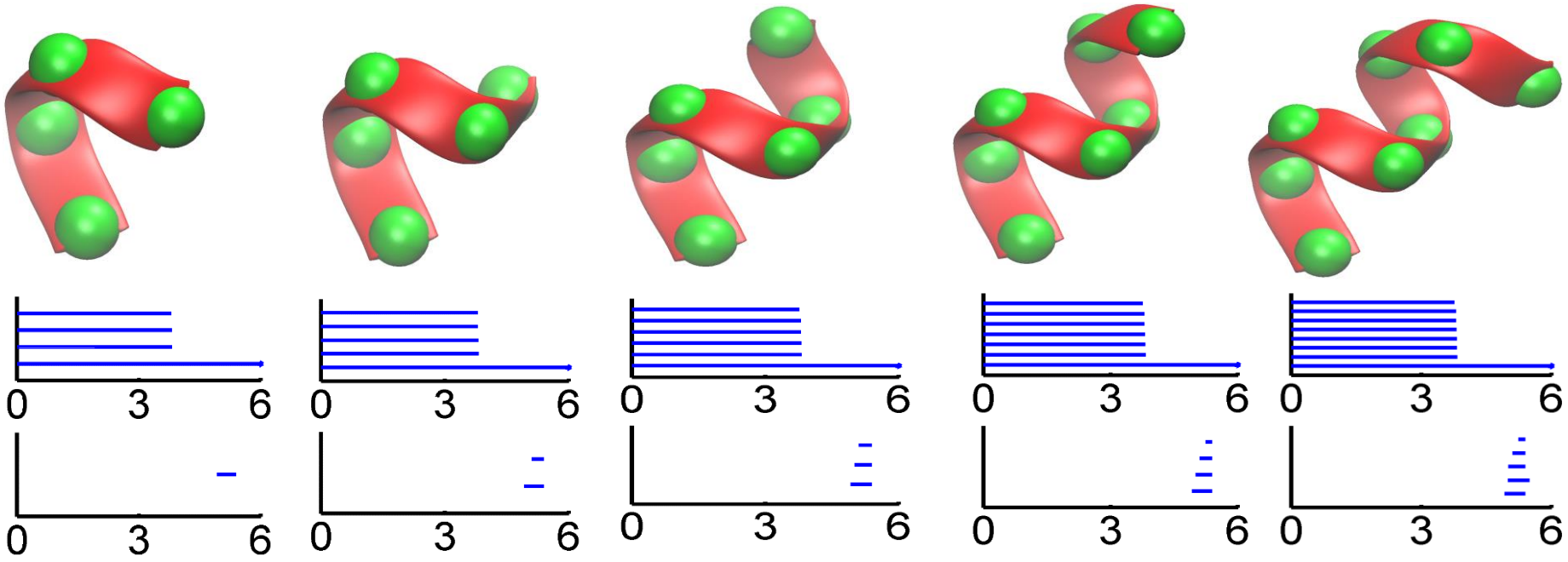


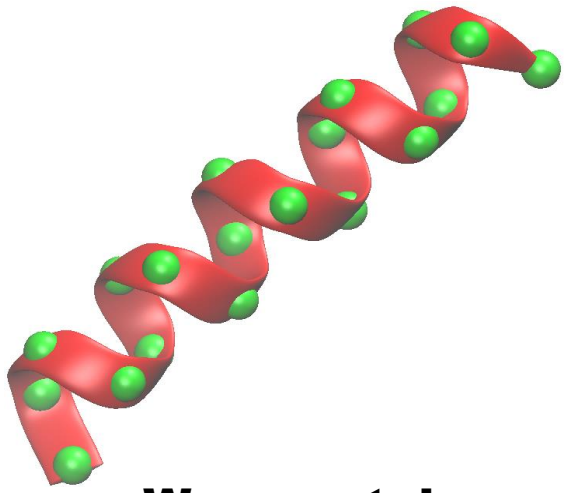


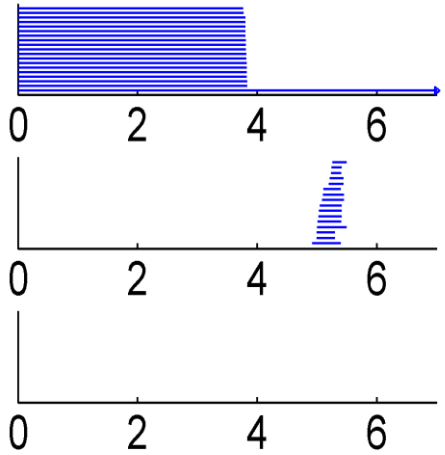




Topological fingerprints of an alpha helix







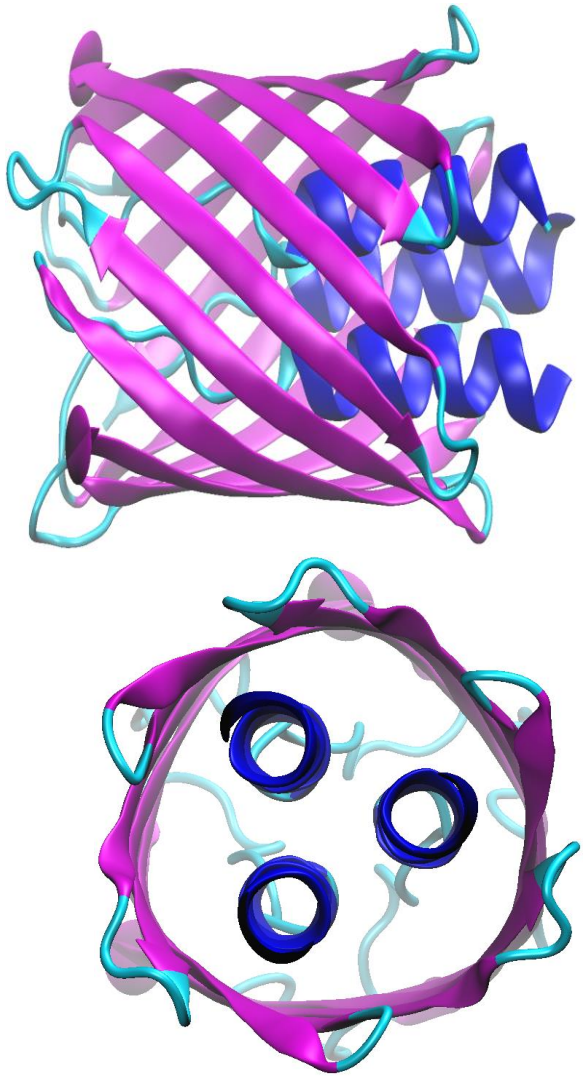
Short bars are NOT noise!

(Xia & Wei, IJNMBE, 2014)

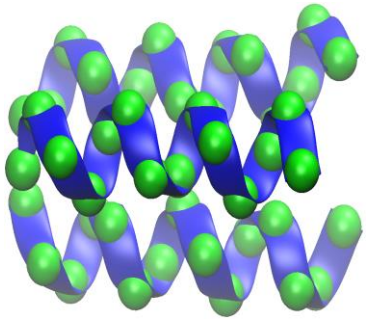
Wasserstein metrics?

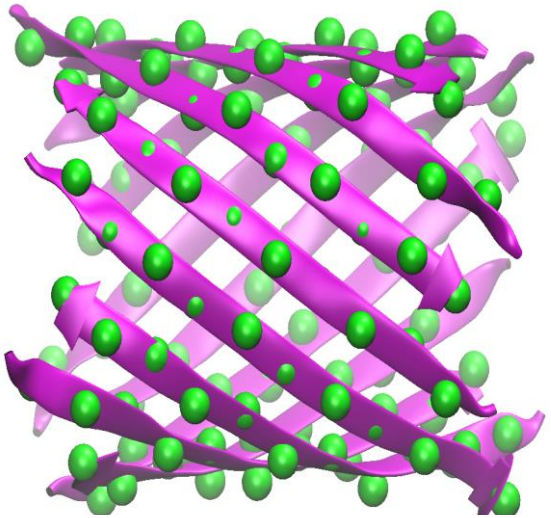
Topological fingerprints of beta barrel

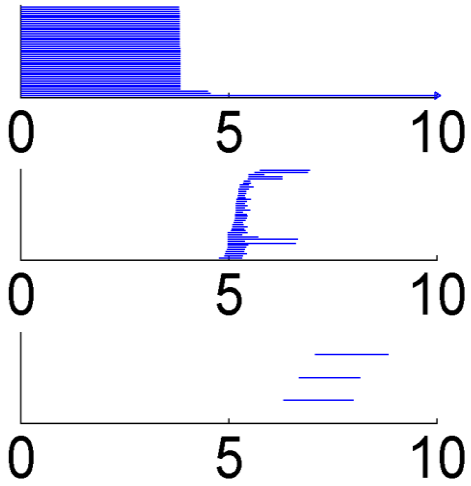
(Xia & Wei, IJNMBE, 2014)

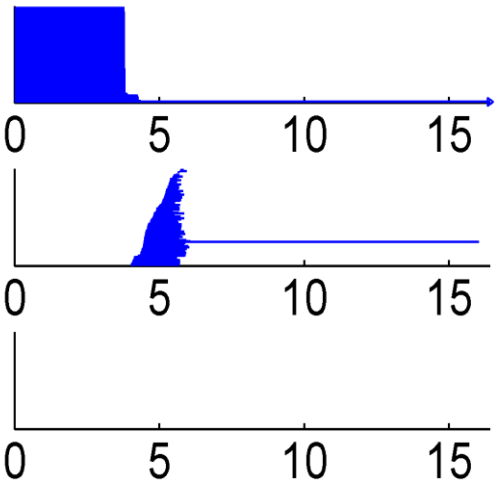


Protein:2GR8

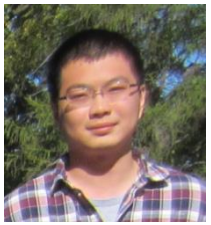


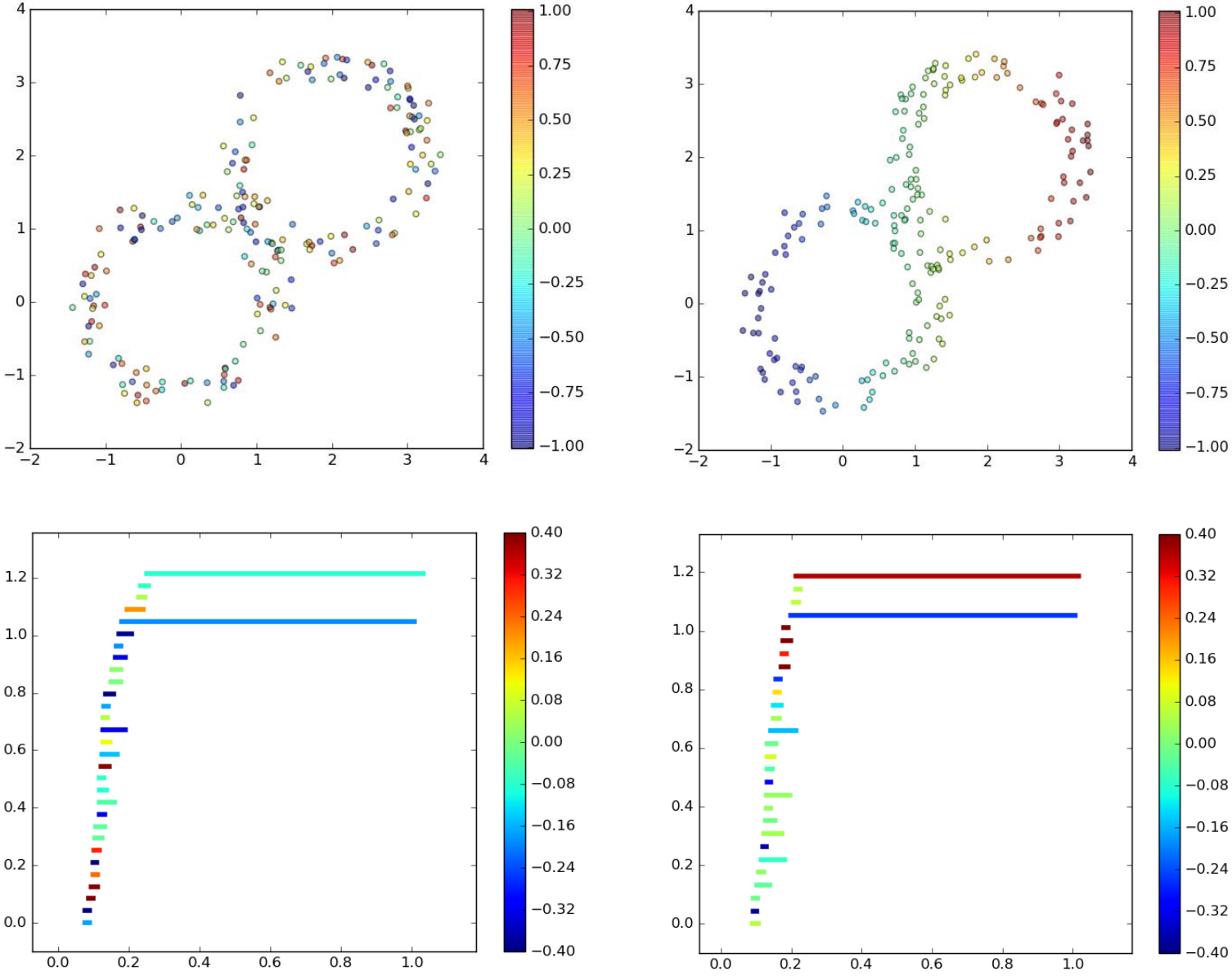






Persistent cohomology, Hodge theory and discrete exterior calculus



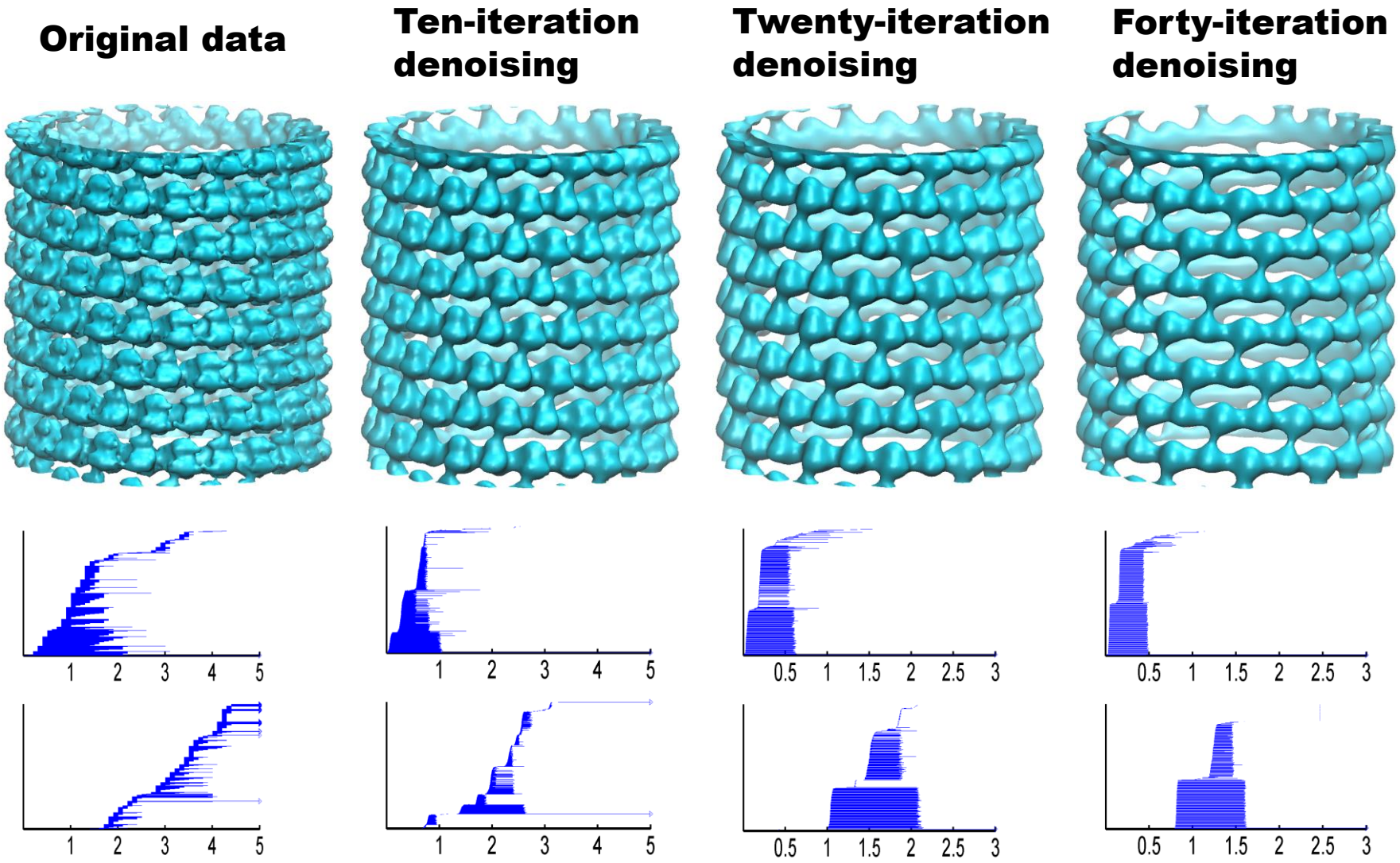


Wasserstein metrics

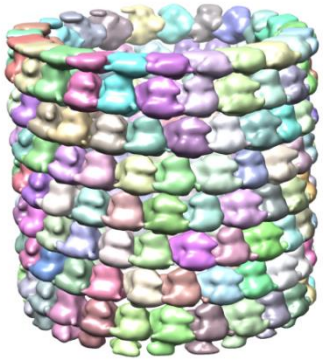
(Cang & Wei, 2018)

Topological noise reduction via geometric PDE

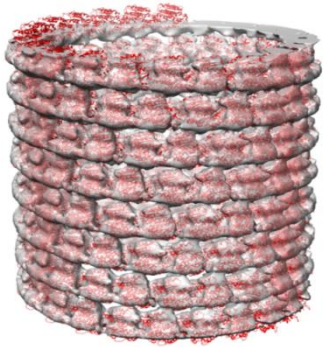
(Xia & Wei, IJNMBE 2015)



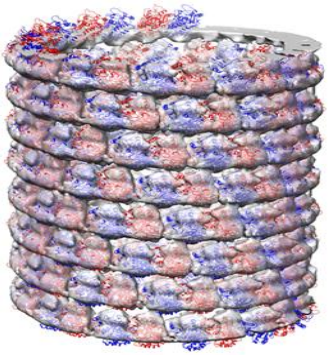
Persistent homology for ill-posed inverse problems



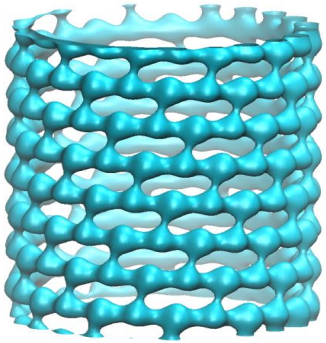
Original data: microtubule

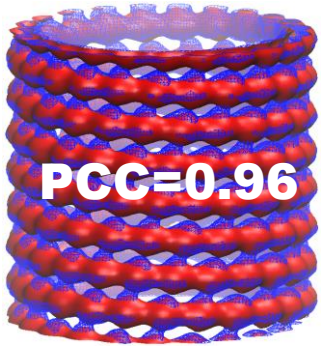


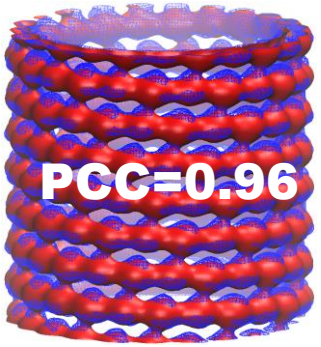
Fitted with one-type of tubulins

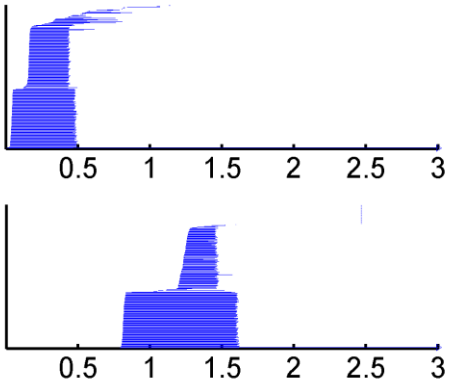


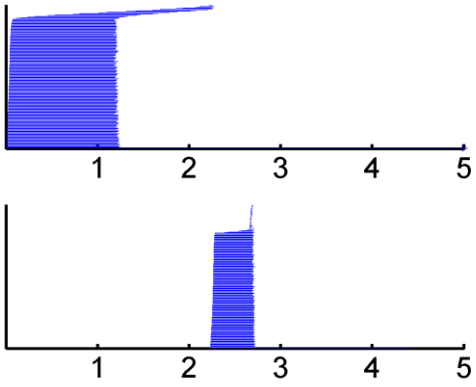
Fitted with two-types of tubulins

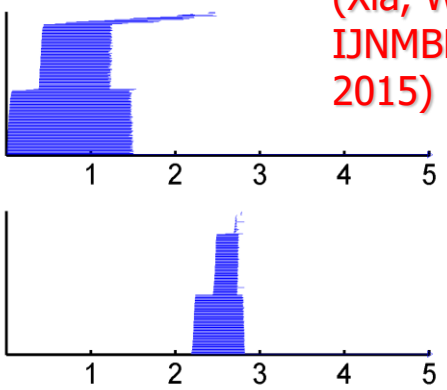








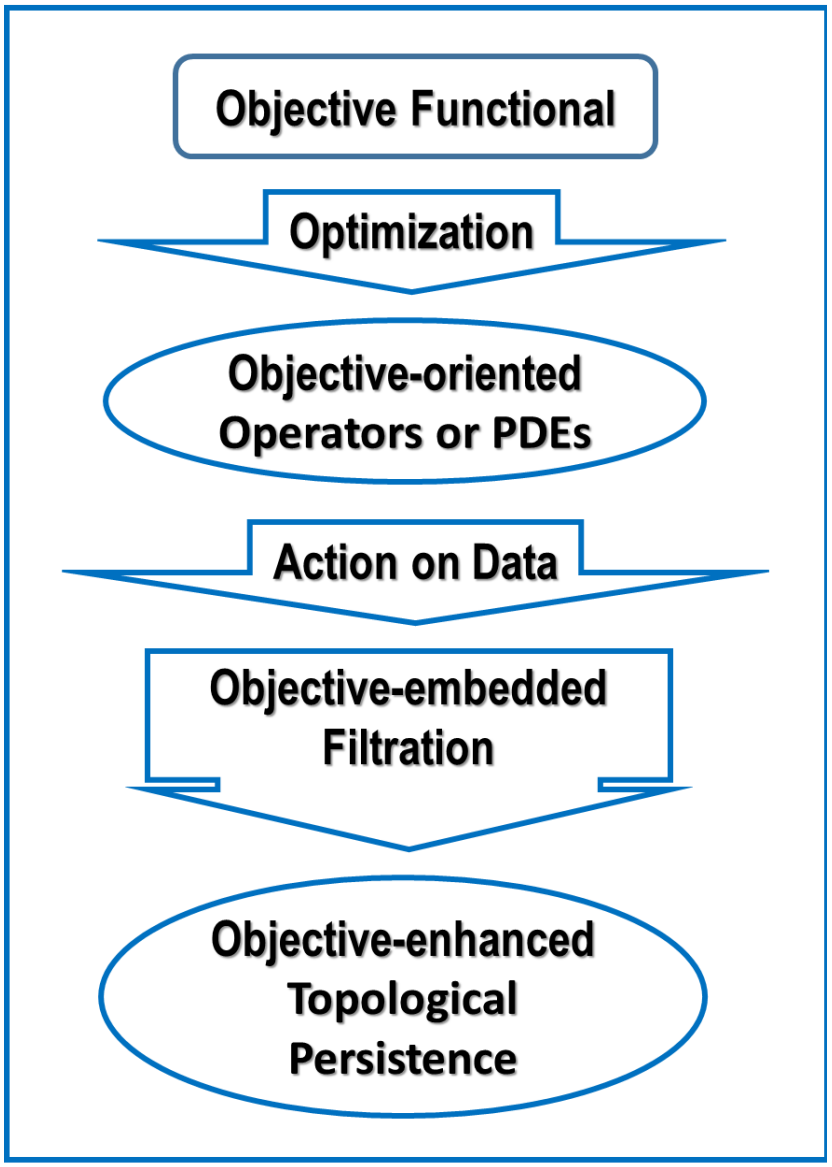




(Xia, Wei, IJNMBE, 2015)

Objective oriented persistent homology

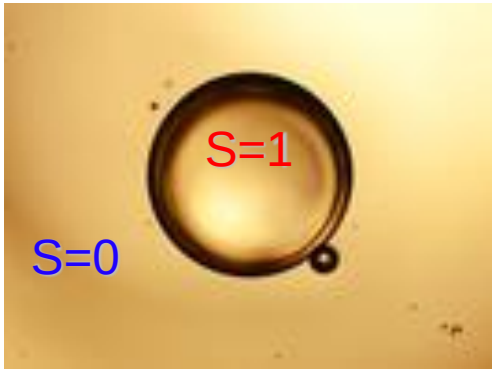
(Wang & Wei, JCP, 2016)



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

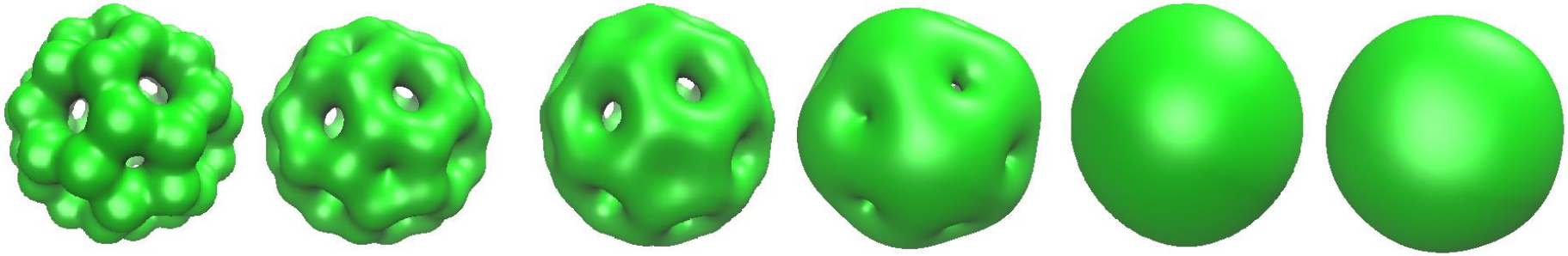
$$\frac{\partial S}{\partial t} = |\nabla S| \left[\nabla \bullet \frac{\gamma \nabla S}{|\nabla S|} \right]$$

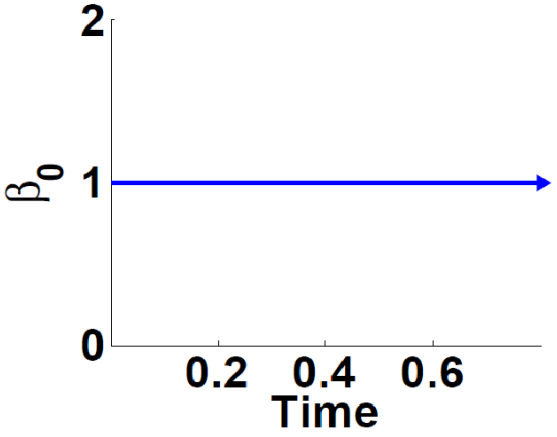
Objective oriented persistent homology

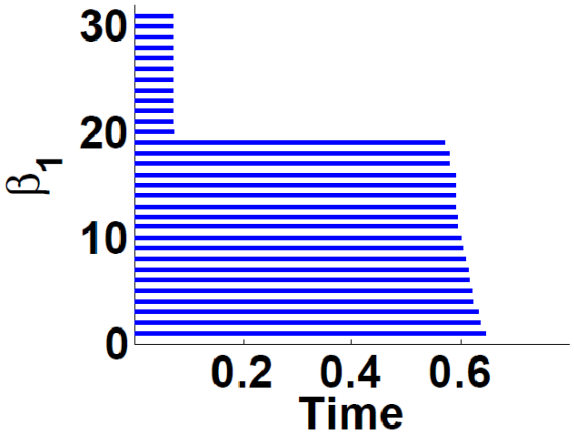
Level sets generated from Laplace-Beltrami flow

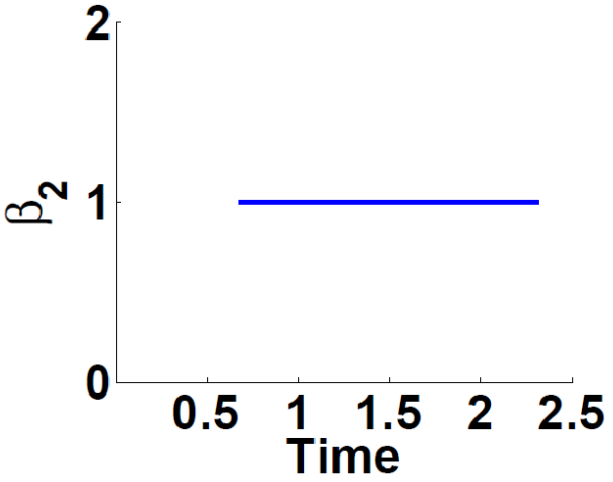
$$\frac{\partial S}{\partial t} = |\nabla S| \left[\nabla \bullet \frac{\gamma \nabla S}{|\nabla S|} \right]$$

(Wang & Wei, JCP, 2016)



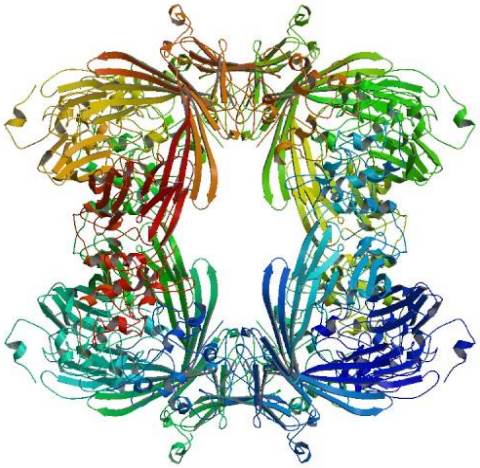


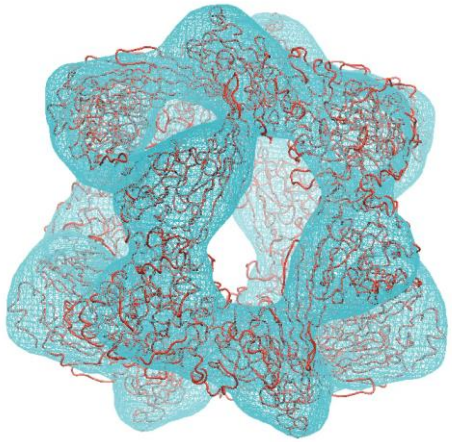


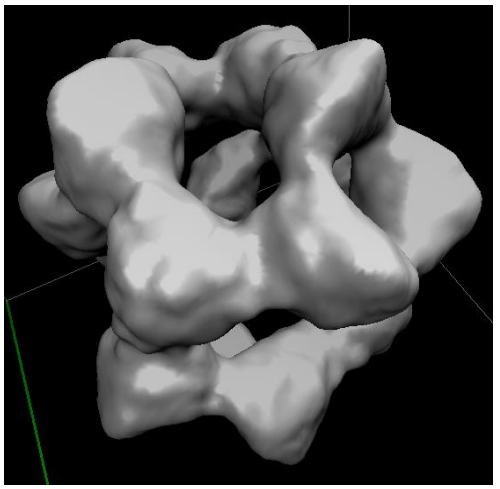


Barcodes are generated by cubical complex and cubical homology

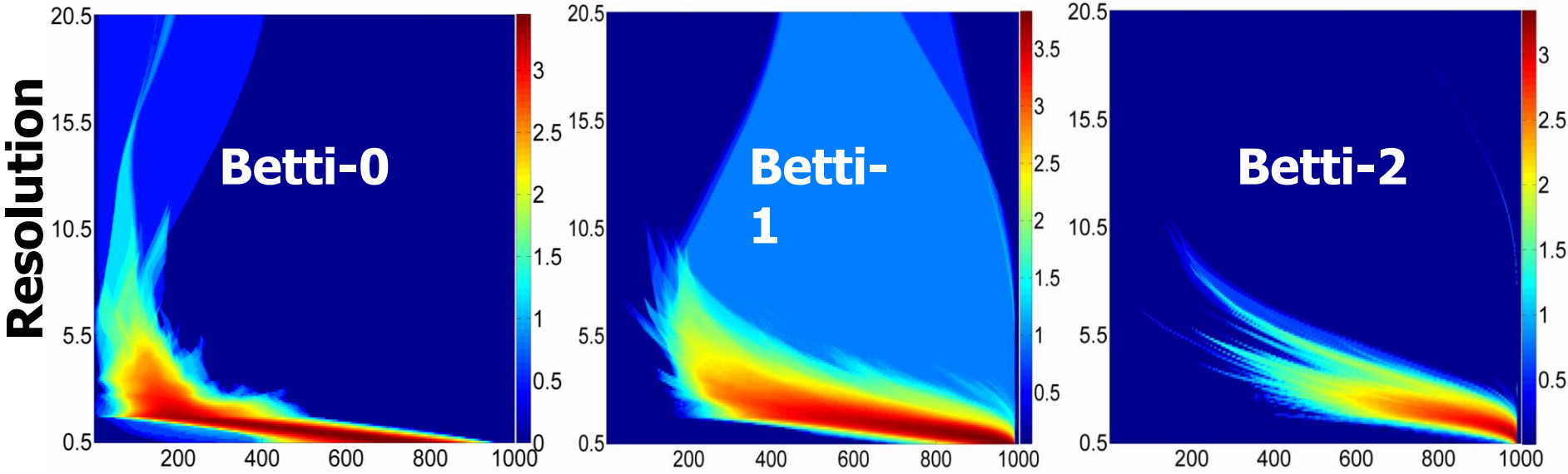
Multiresolution 2D persistence in protein complex 2YGD







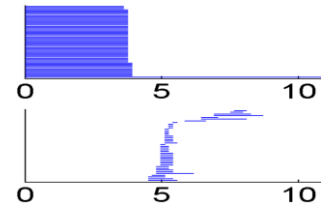
$\log_{10}(N)$

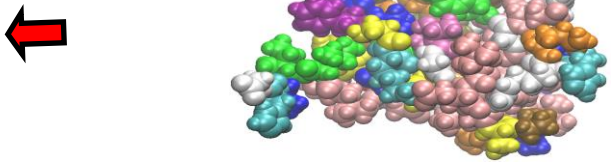


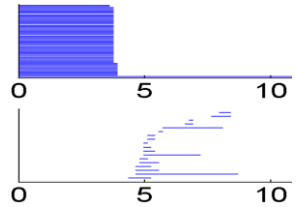
(Xia & Wei, JCC, 2015)

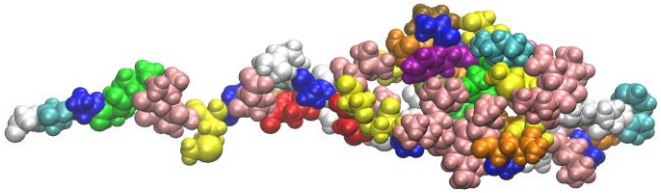
Topological analysis of protein folding

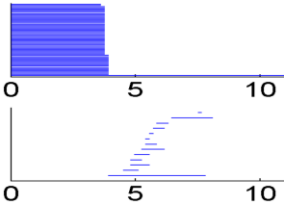
ID: 1I2T

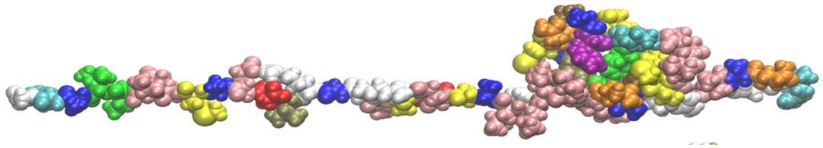


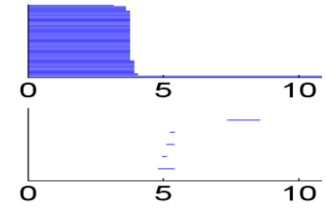


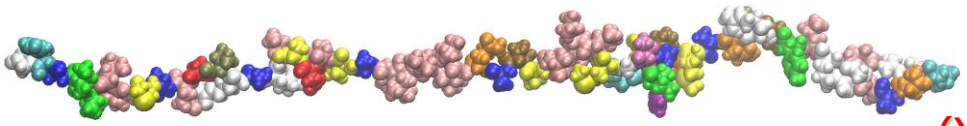


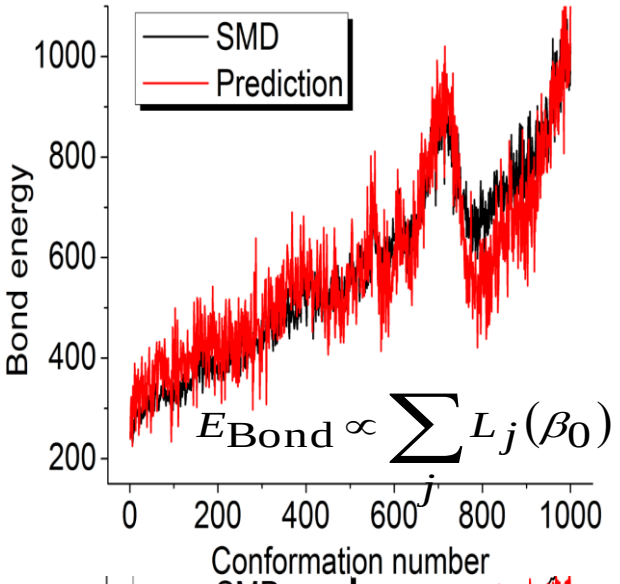


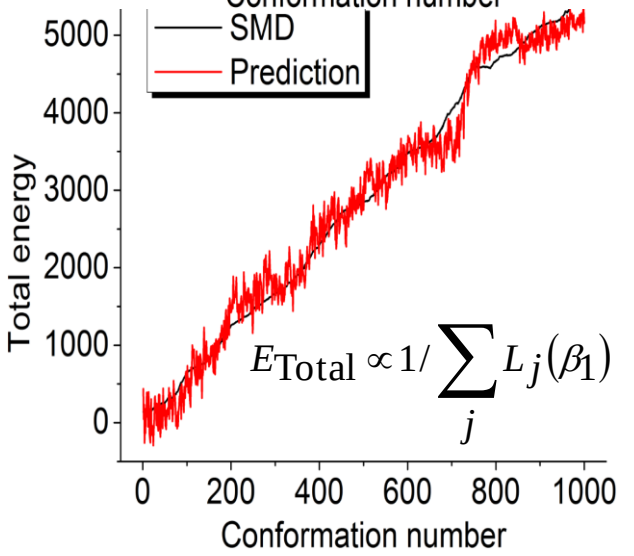








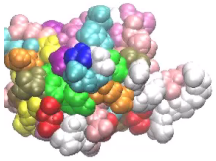


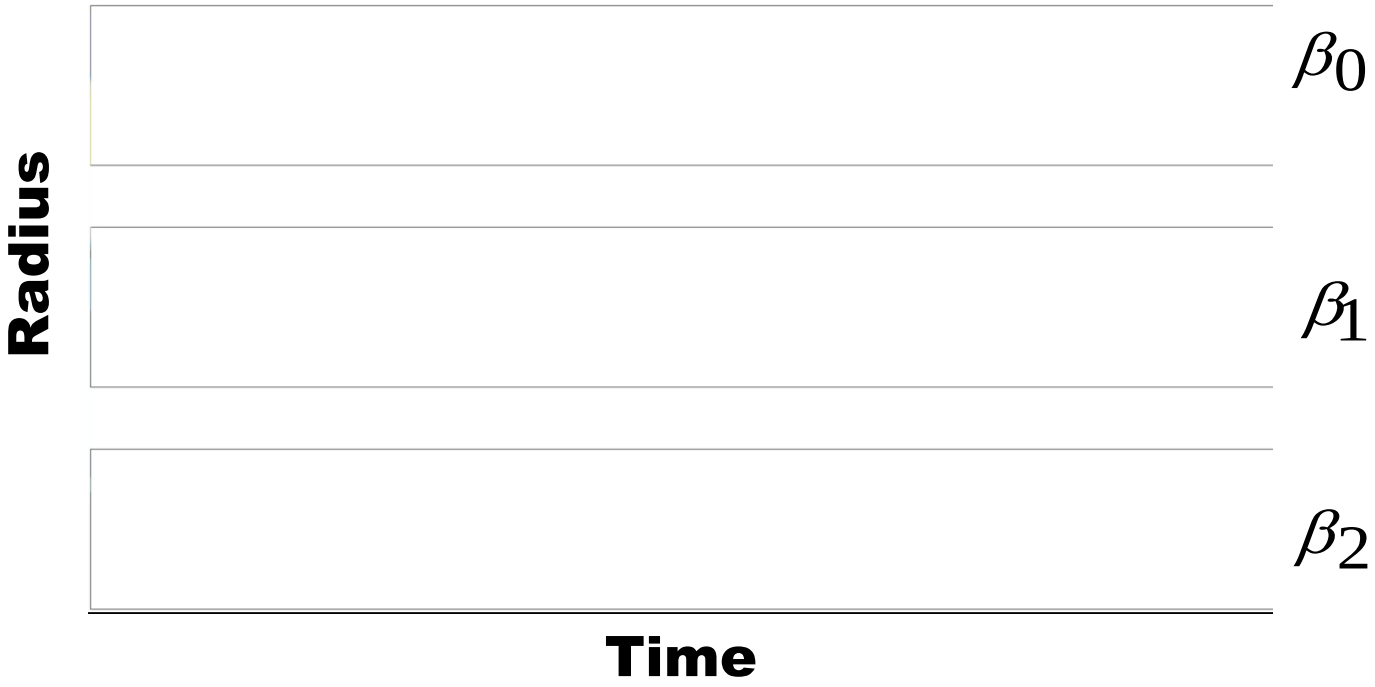


Quantitative!

(Xia, Wei, IJNMBE, 2014)

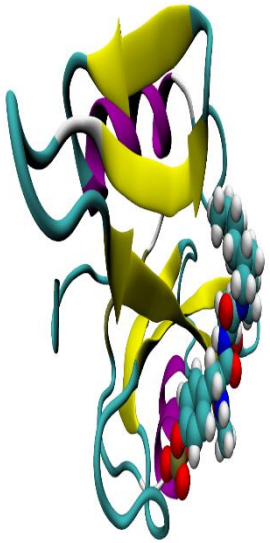
2D persistent homology of protein 1UBQ unfolding

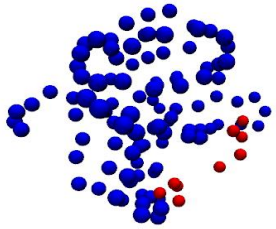


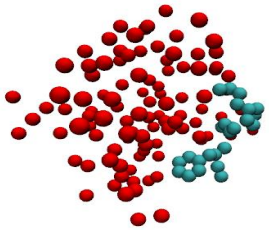


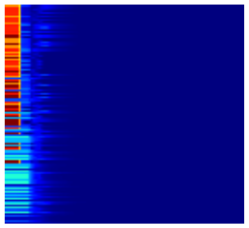
(Xia & Wei, JCC, 2015)

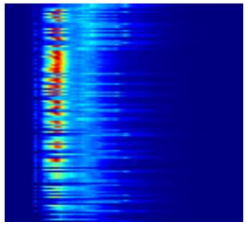
Topological convolutional deep Learning architecture









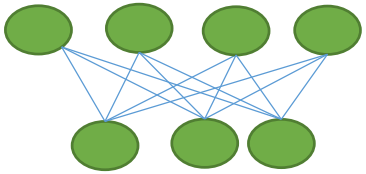


Multichannel images (54x200)

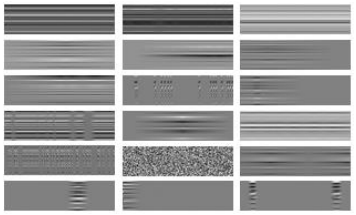
Convolution (128x200)

Pooling (128x100)

Flattening (1xN)



Prediction



Original protein-ligand Complex

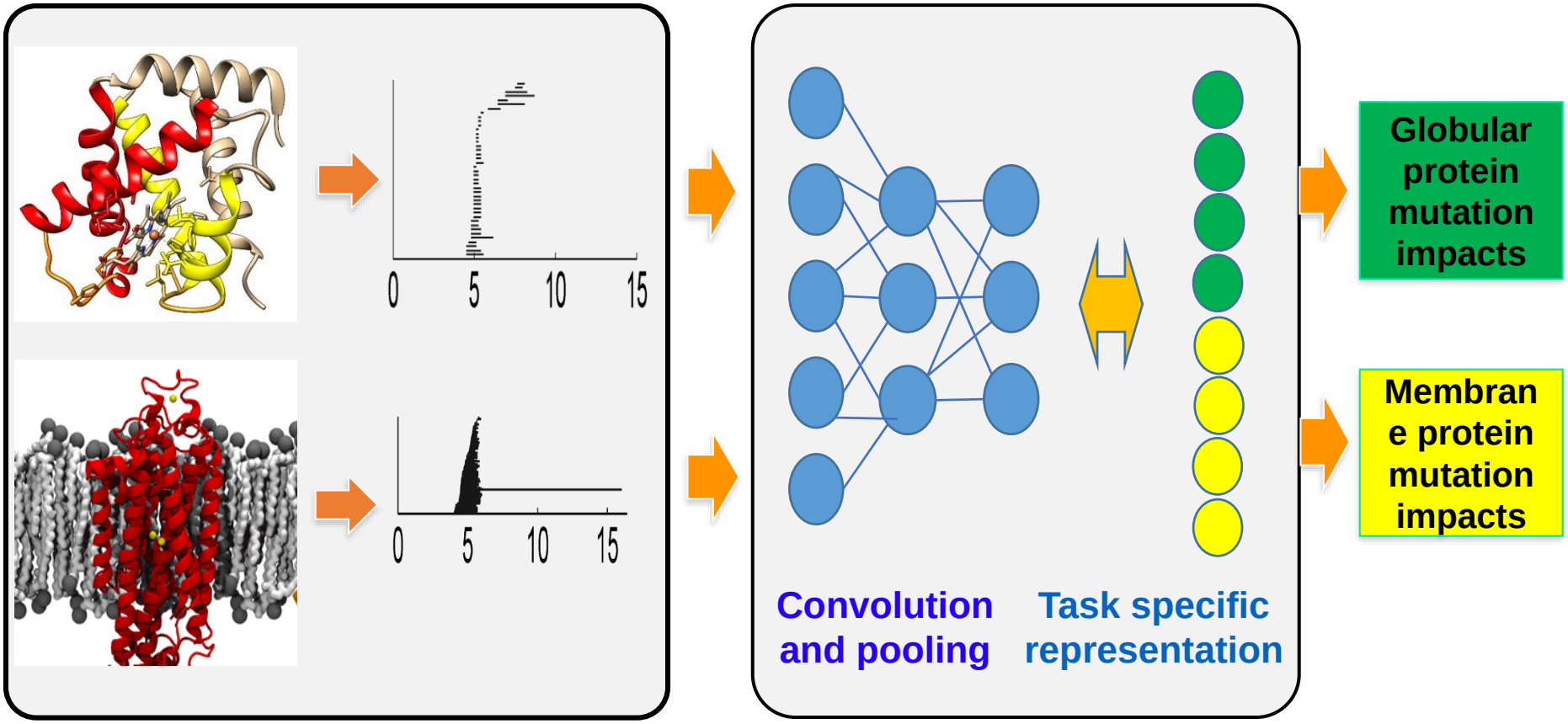
Classify atoms into element specific groups

Generate topological, DG & graph fingerprints

Convolutional deep learning neural network

(Cang & Wei, PLOS CB, 2017)

Topological Multi-Task Deep Learning

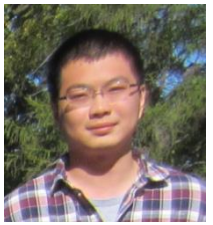


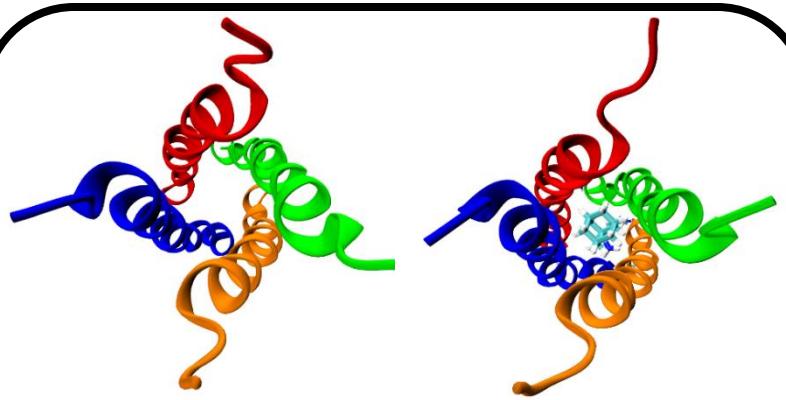
Topological feature extraction

Multi-task topological deep learning

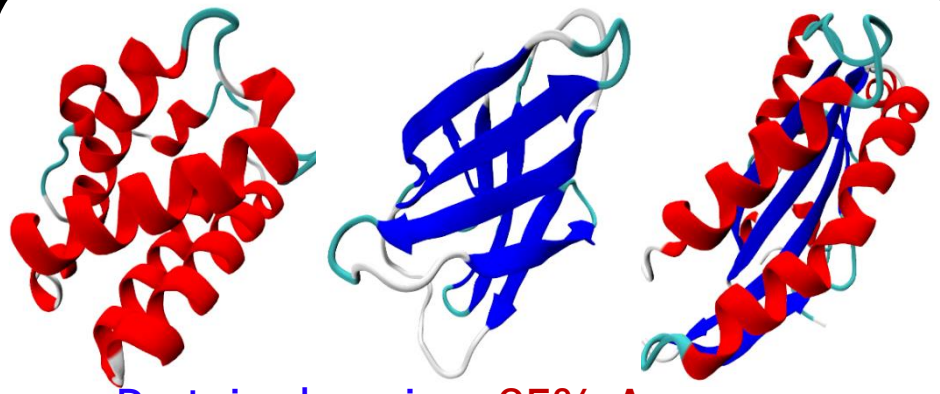
(Cang & Wei, PLOS CB, 2017)

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



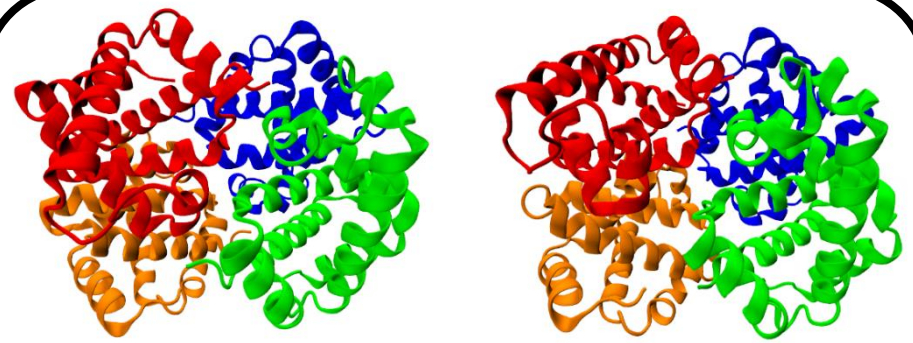


Influenza A virus drug inhibition: 96% Accuracy



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

(Cang et al, MBMB, 2015)



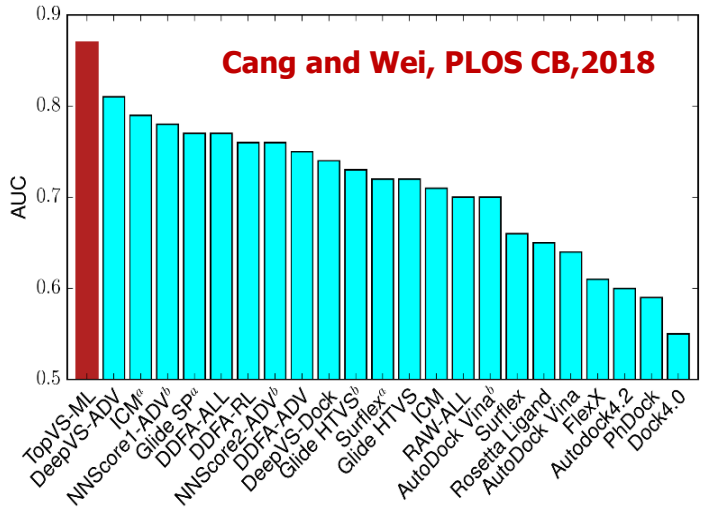
Hemoglobins in their relaxed and taut forms: 80% accuracy

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

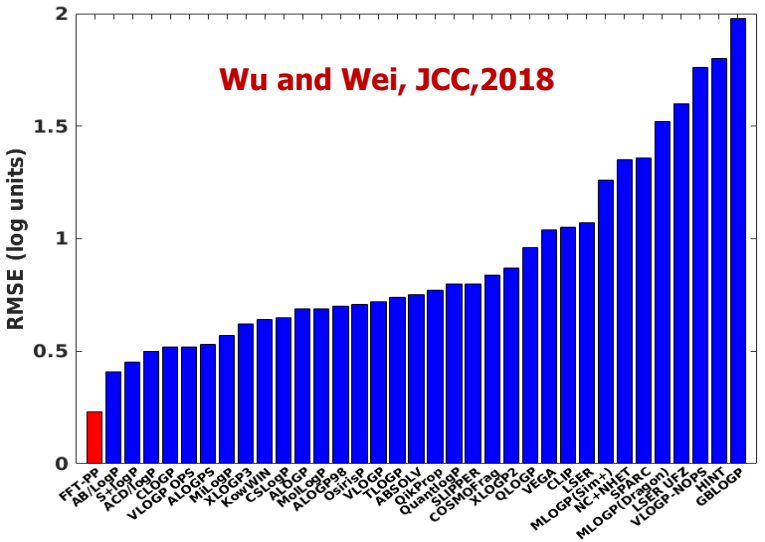
Topological learning based predictions

Classification of ligands & decoys

DUD database 128,374 protein-ligand/decoy pairs



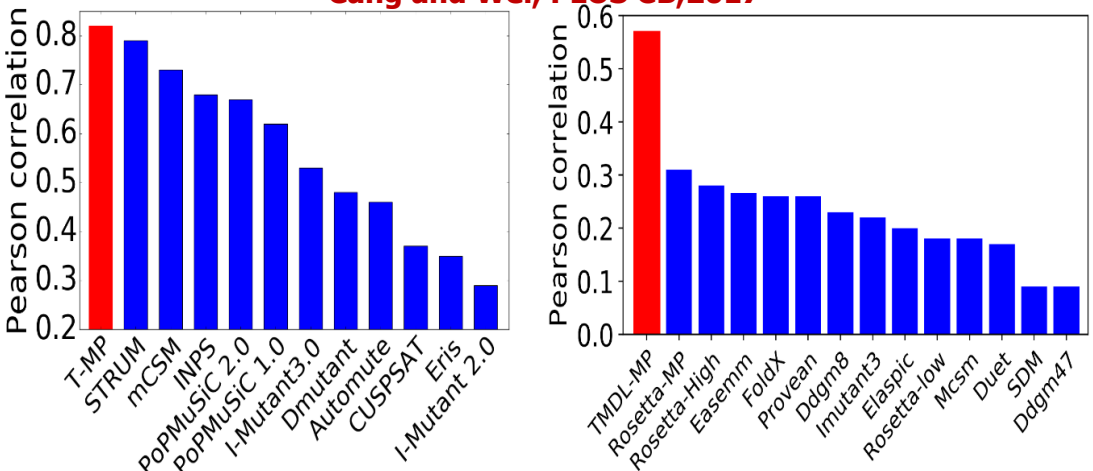
Prediction RMSD of LogP (Star set)



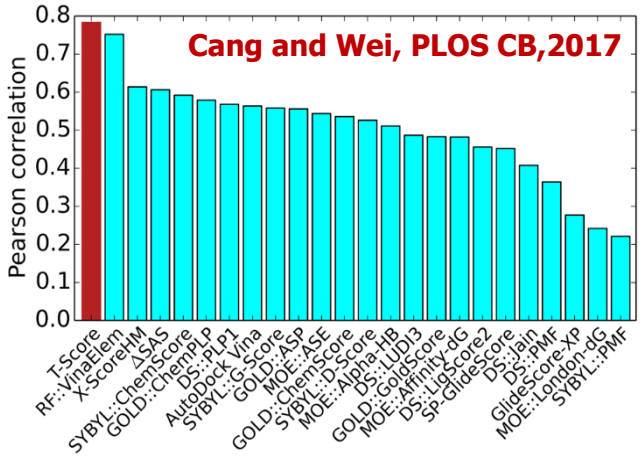
Prediction correlations for 2648 mutations on globular proteins

Prediction correlations for 223 mutations on membrane proteins

Cang and Wei, PLOS CB,2017

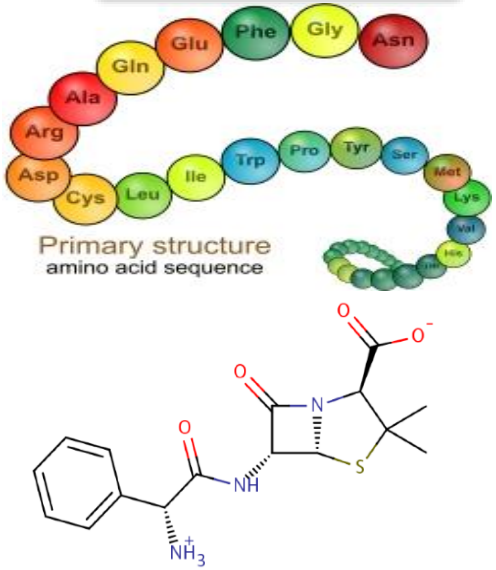


Binding affinity prediction of PDBBind v2013 core set of 195 complexes

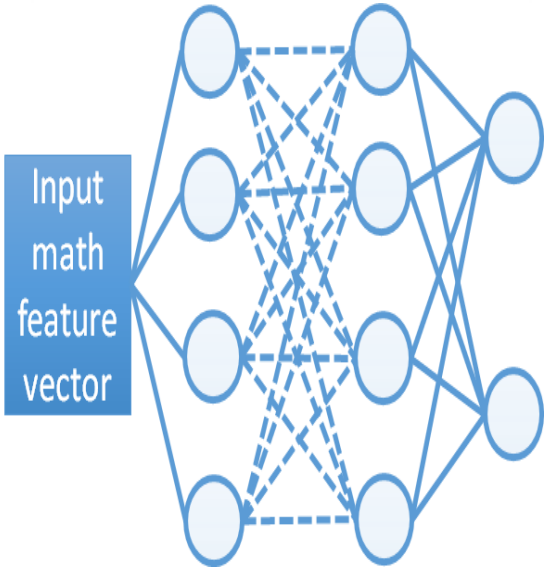


D3R Grand Challenge in drug design

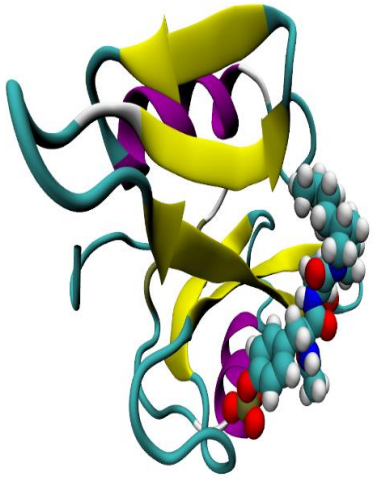
Given data



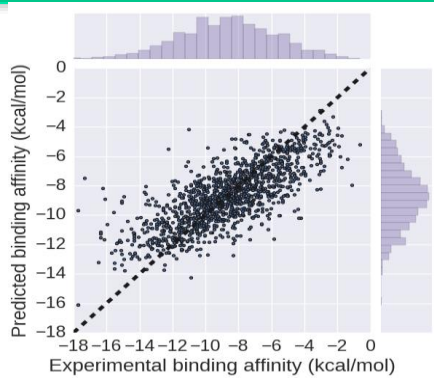
Math based CNN

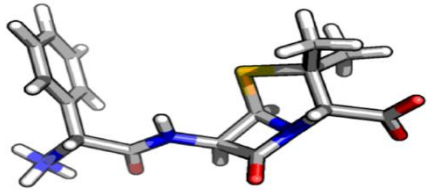


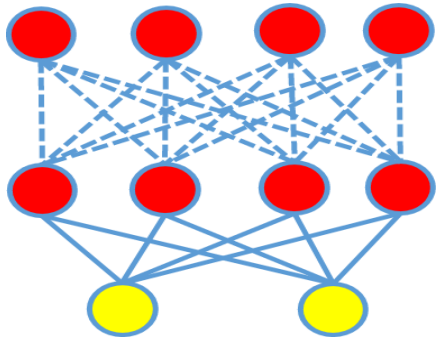
Predicted complex



Final predictions to be compared with experiments



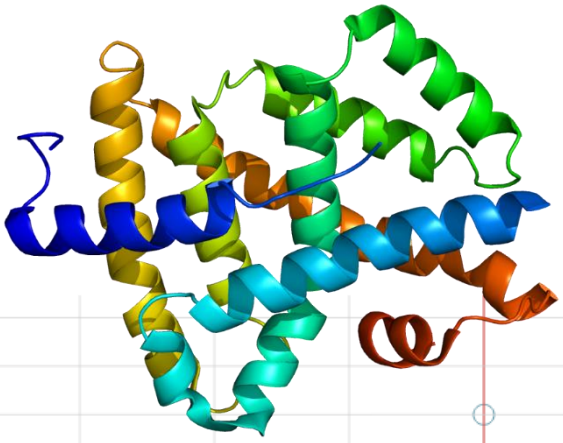




Drug Design & Discovery Resource (D3R) Grand Challenge 2

Given: Farnesoid X receptor (FXR) and 102 ligands

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



Dr Duc Nguyen

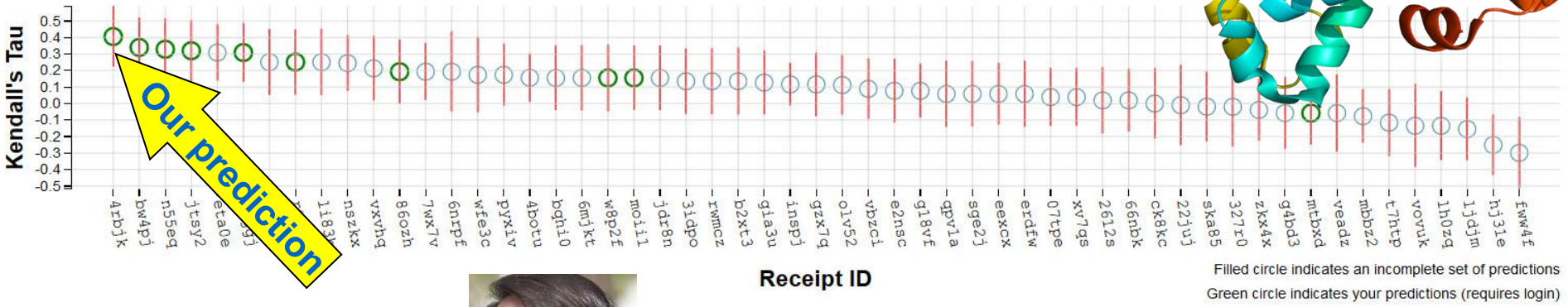
D3R Grand Challenge 2 (2016-2017)

Given: Farnesoid X receptor (FXR) and 102 ligands

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

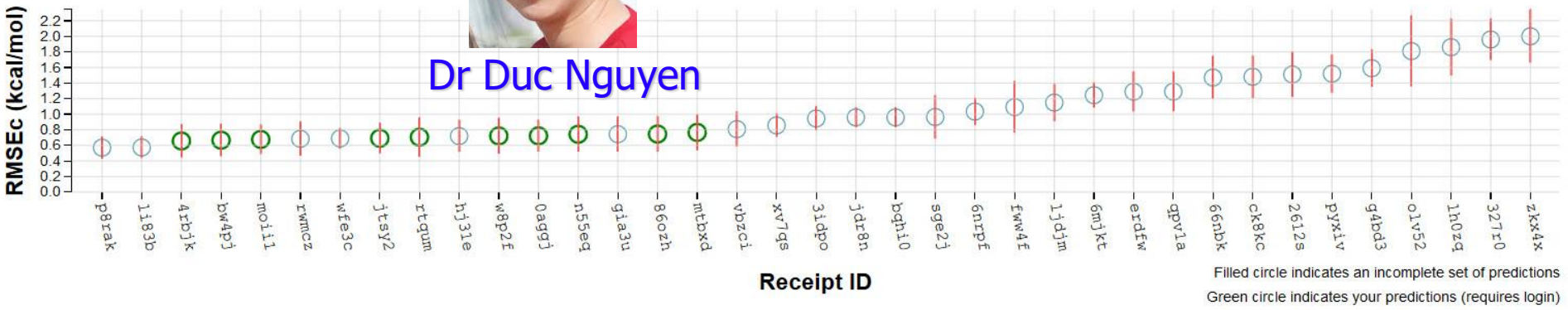
Grand Challenge 2

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



Dr Duc Nguyen

Free Energy Set 1 (Stage 2) - RMSEc



D3R Grand Challenge 3 (2017-2018)

Preliminary Evaluations, Subject to Revision and Refinement

Cathepsin Stage 1 Pose Predictions Scoring
Free Energy Sets
Cathepsin Stage 1B Pose Prediction
Cathepsin Stage 2 Scoring (partials)
Free Energy Sets Affinity ranking of 24 Complexes

1st

Zixuan Cang

Dr Duc Nguyen

VEGFR2 Scoring (partials) **JAK SC2** Scoring (partials) **p38-α** Scoring (partials)

1st

JAK SC3 Free Energy Sets

1st

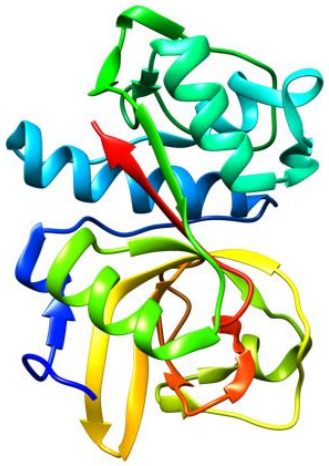
TIE2 Scoring (partials)
Free Energy Set 1

1st

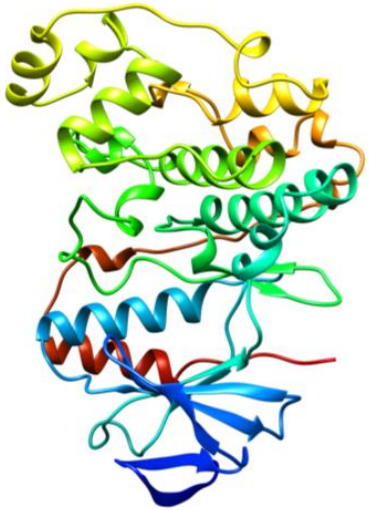
Free Energy Set 2

1st

ABL1 Scoring



Cathepsin S



Kinase: p38-α

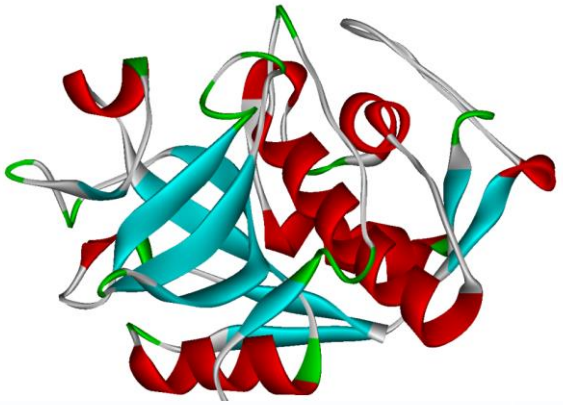
Eight of our predictions were ranked 1st in a total of 21 competitions.

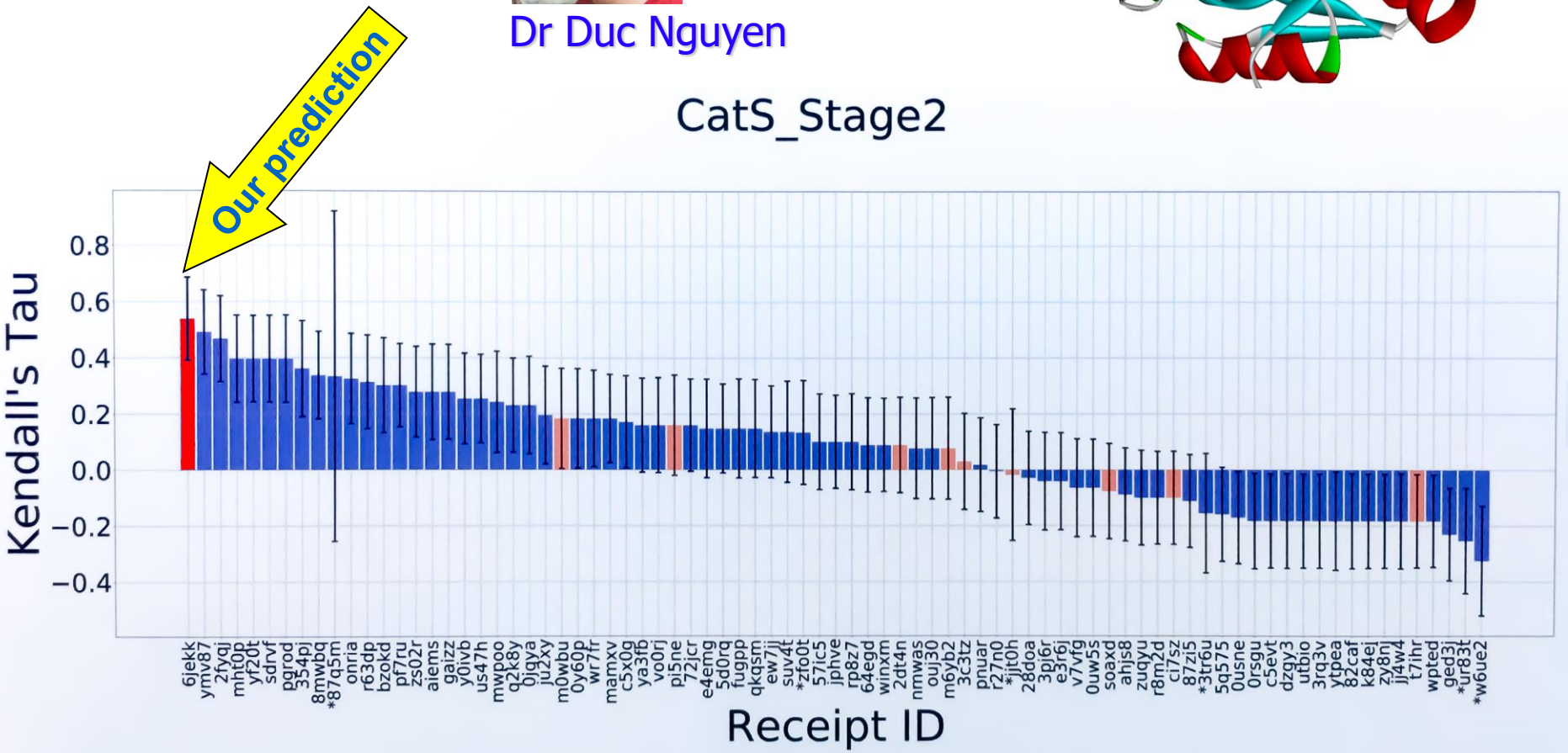
D3R Grand Challenge 3 (2017-2018)

Given: X-ray crystal structures of cathepsin (CatS) and 24 ligands

Tasks: Compute their binding affinity ranking

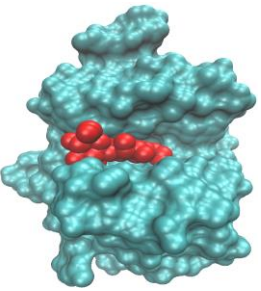
Dr Duc Nguyen

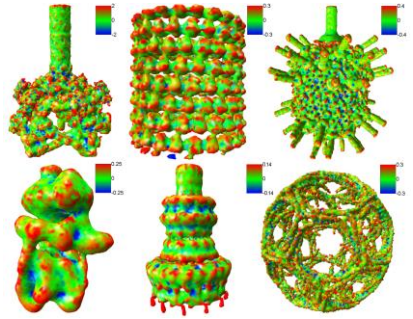




Our other methods for drug design and discovery

- Geometric graph theory, algebraic graph theory and topological graph theory
- Differential geometry: geodesic distance, curvatures and curvature tensors
- Partial differential equation based biophysical models





US patent pending

Quantitative systems pharmacology modeling

Predicting drug pharmacokinetics and pharmacodynamics by integrating

- Systems biology, protein networks, signal transduction pathways
- Cellular biology and cellular mechanics
- Systems physiological modeling
- Clinical data and virtual patient simulation

In collaboration with Bristol-Myers Squibb (BMS)

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.**
- **Our goal is to create mathematical jobs and kill experimental jobs in drug design and biology.**
- **Postdocs are wanted**


N Baker
(PNNL)
P Bates
(MSU)
Z Burton
(MSU)
K Dong
(MSU)
H Hong
(MSU)
J Hu
(MSU)
Munch
(MSU)
Y
Tong
(MSU)
X Ye
(UKLR)
Shan Zhao
Alabam
Y Zhou
CSU
W Geng
SMU
Duan Chen
UNCC
Zhan Chen
GSU
Lin Mu
ONL
S Yang
CU KH
S.N. Yu
Wall Street
Y.H. S
Der
L. H. Hu
Chicago
J. Park
Indiana
Wang
UCLA
Y Cao
P. I.
C Opron
UM
K Xia
NTU