Topology based deep learning for biomolecular data

Guowei Wei

Departments of Mathematics

Michigan State University

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

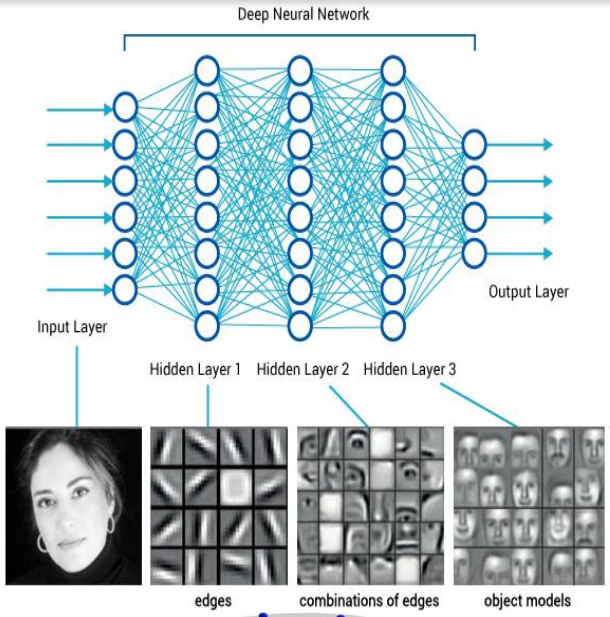
American Institute of Mathematics

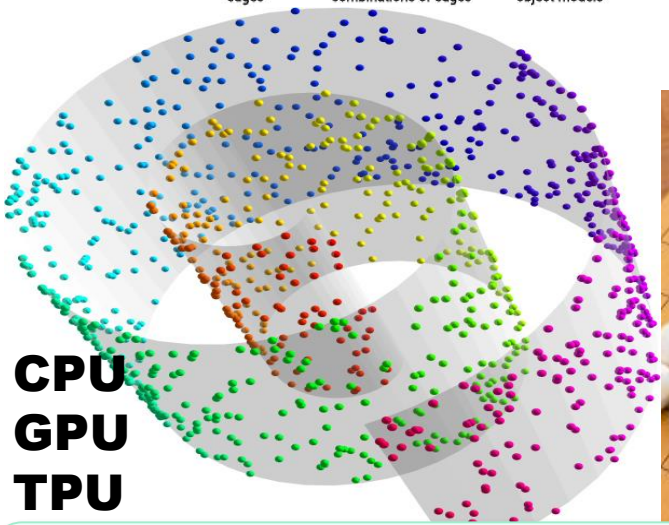
July 23-28, 2017

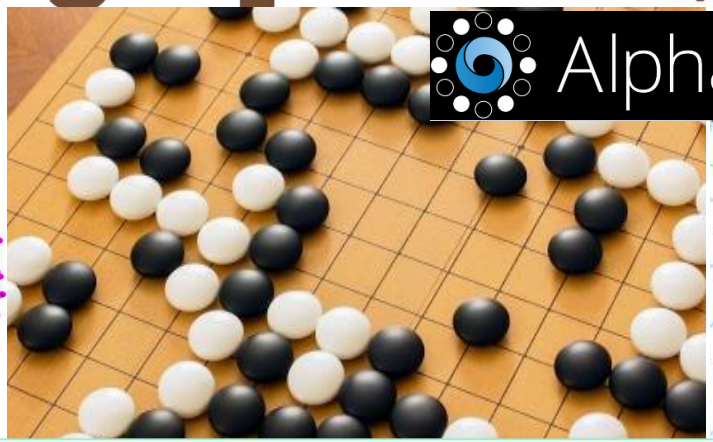
Grant support:

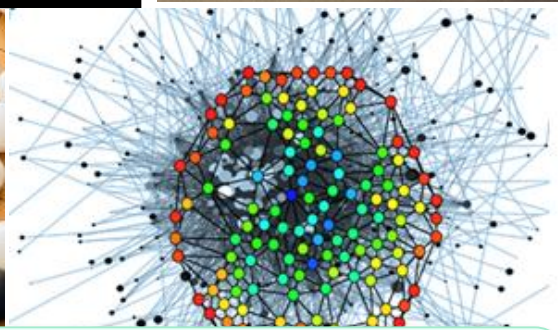
NSF, NIH and Michigan State University

Welcome to big-data era



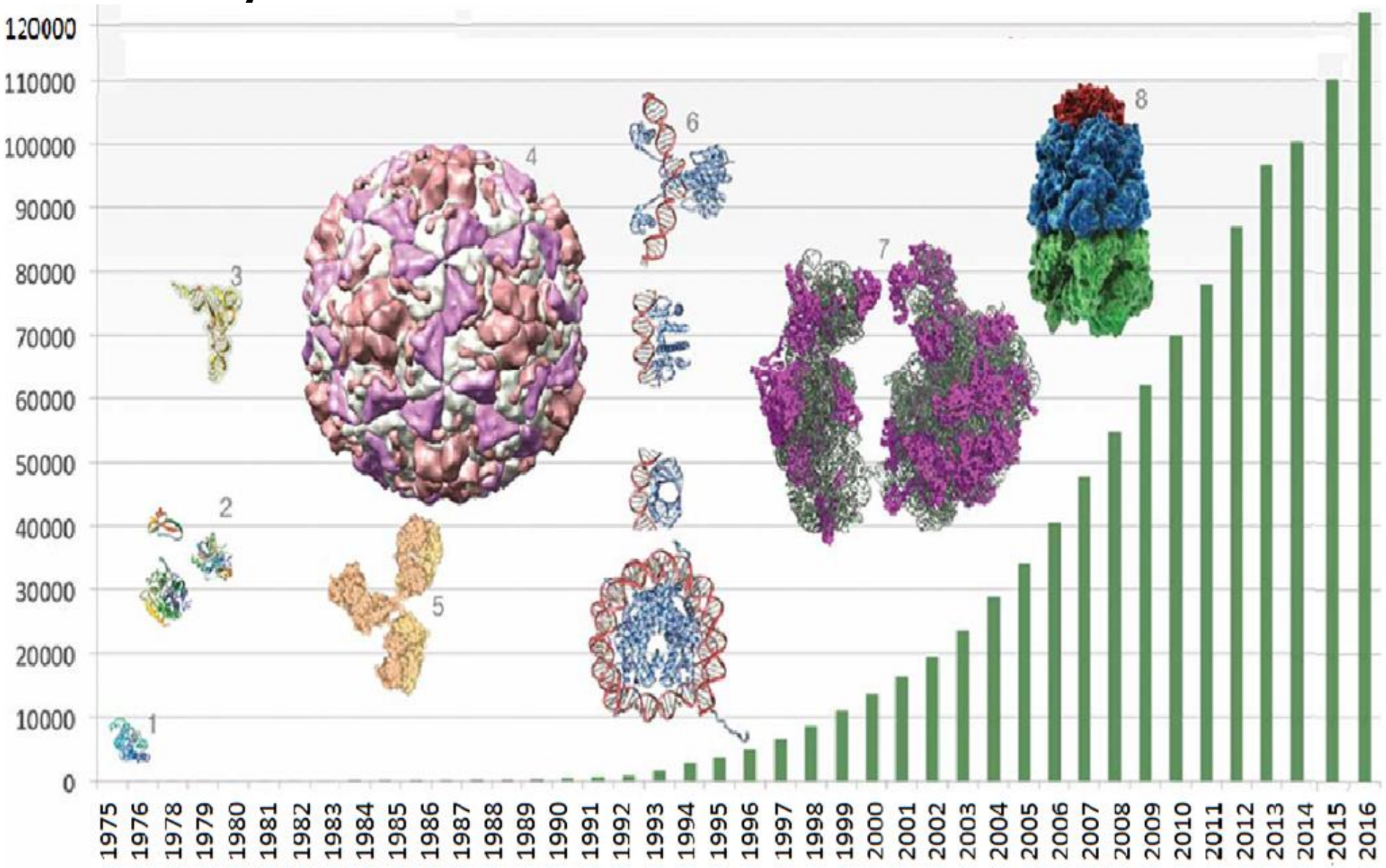






Half of all jobs will be done by robots in the near future

Yearly Growth of Total Structures in the Protein Data Bank



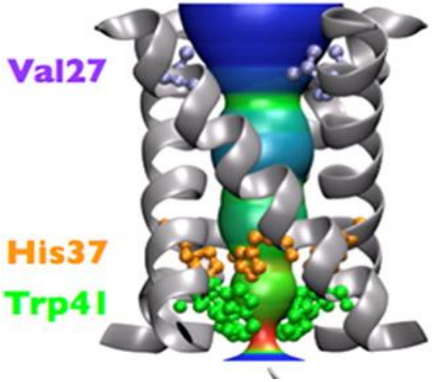
Biological sciences are undergoing a historic transition: From qualitative, phenomenological, and descriptive to quantitative, analytical and predictive, as quantum physics did a century ago

Machine learning for drug design and discovery

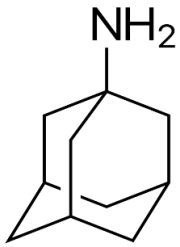
1) Disease identification
2) Target hypothesis
3) Virtual screening
4) Drug structural optimization in the target binding site
5) Preclinical *in vitro* and *in vivo* test
6) Clinical test
7) Optimize drug's efficacy, toxicity, pharmacokinetics, and pharmacodynamics properties (quantitative systems pharmacology)

Influenza -- flu virus

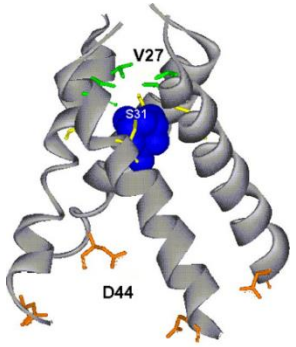
M2 channel



Amantadine

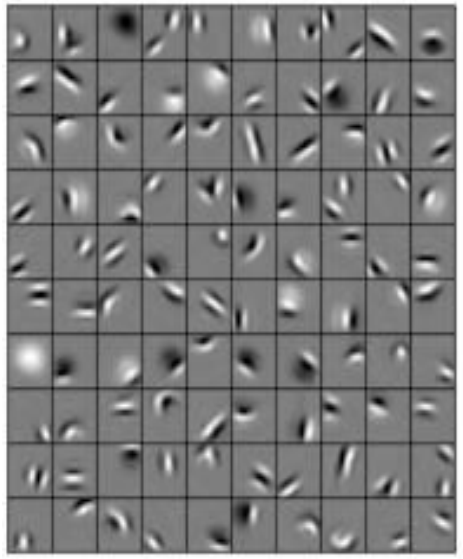


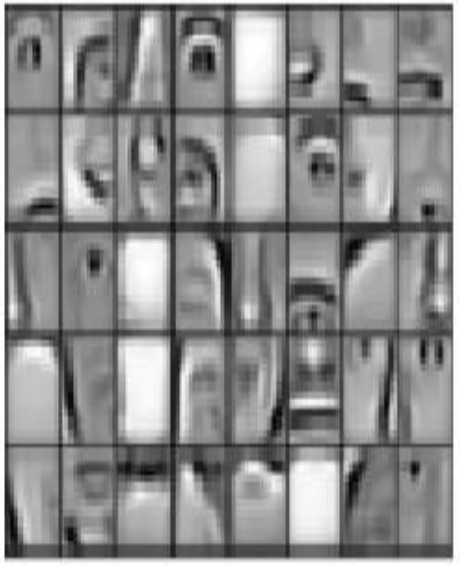
M2-A complex

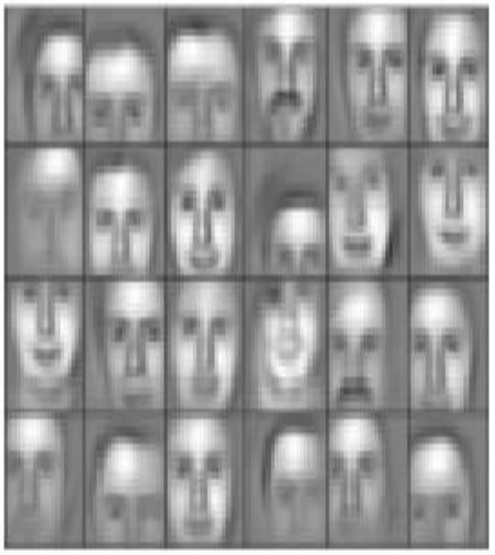


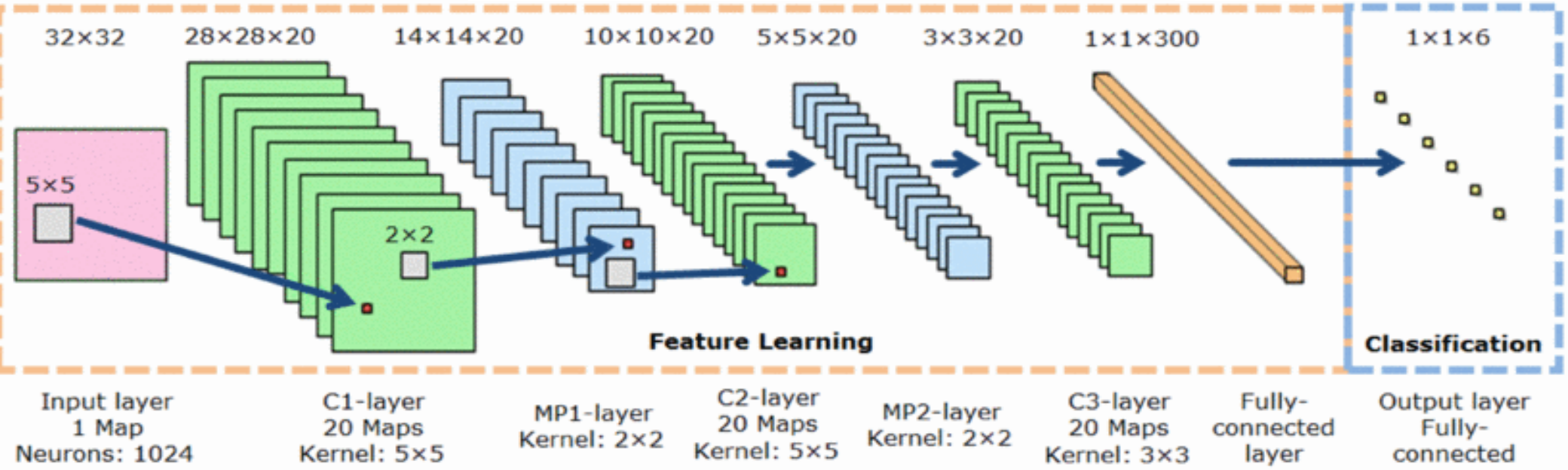
Deep learning

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









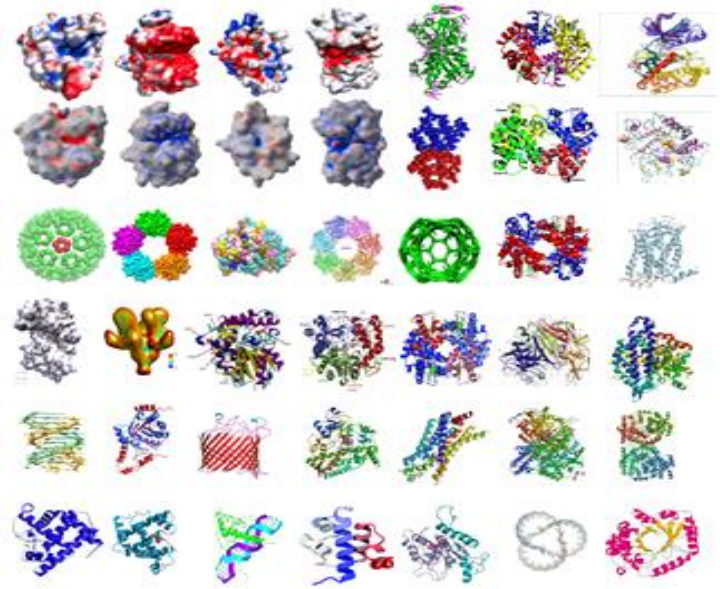
How to do deep learning for 3D biomolecular data?

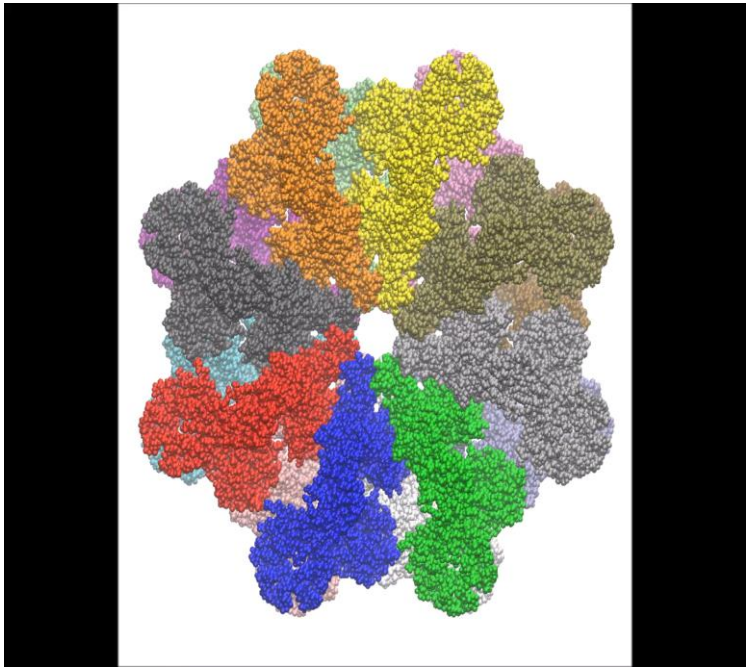
Obstacles for deep learning of 3D biomolecules:

- **Molecules are too small to be visible.**
- **Geometric models for visualization hold partial truth and are non-unique.**
- **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.**
- **Complexity: Atom types, nonstationary & non-Markovian.**

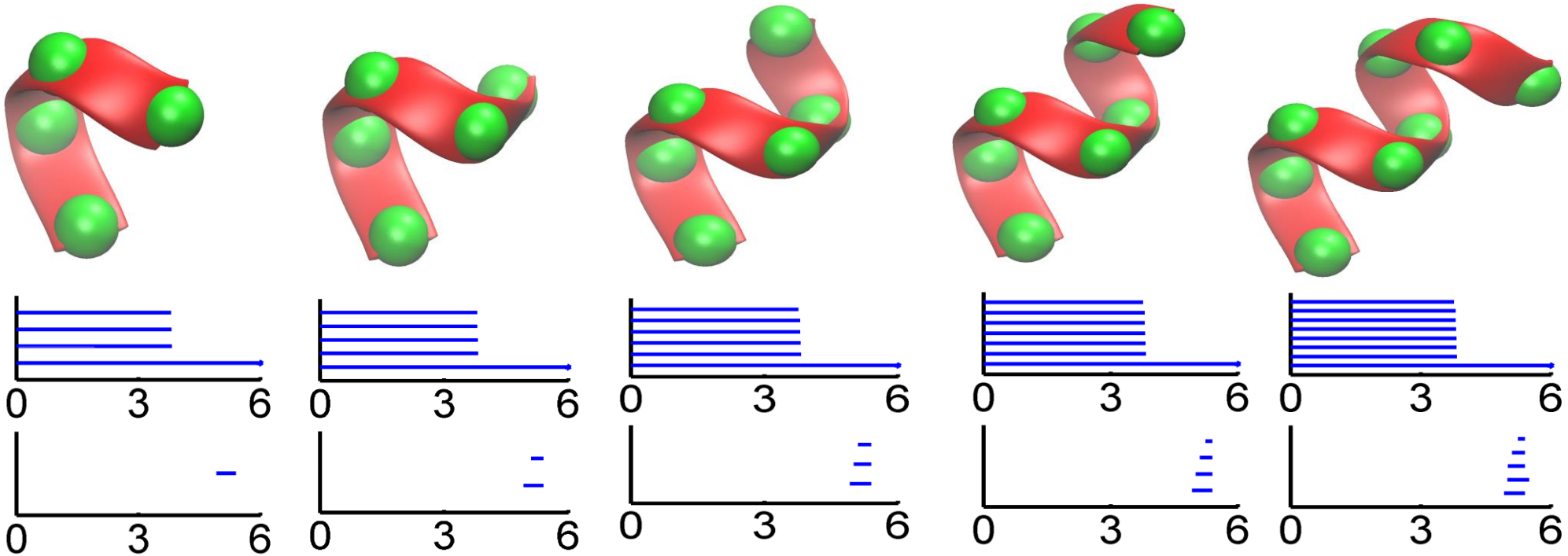
Solution:

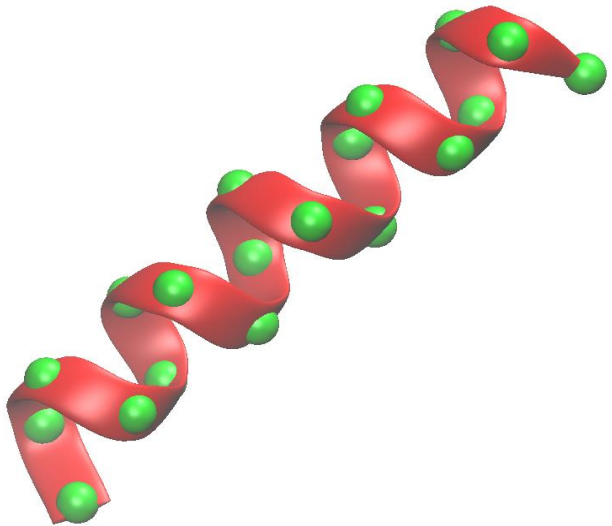
- **Topological simplification**
- **Dimensionality reduction**

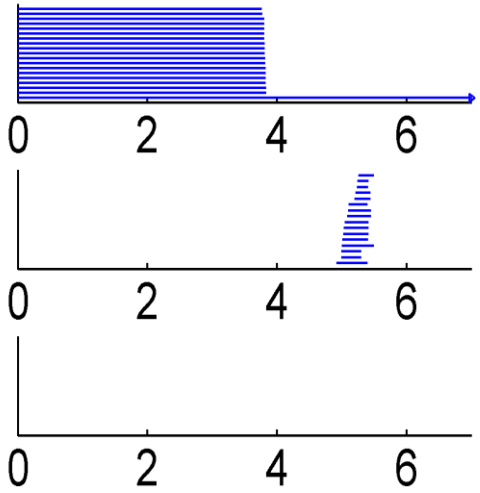




Topological fingerprints of an alpha helix







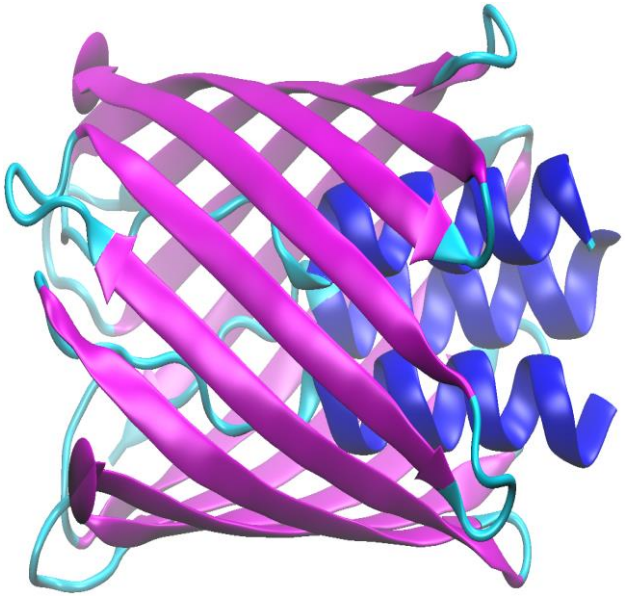
Short bars are NOT noise!

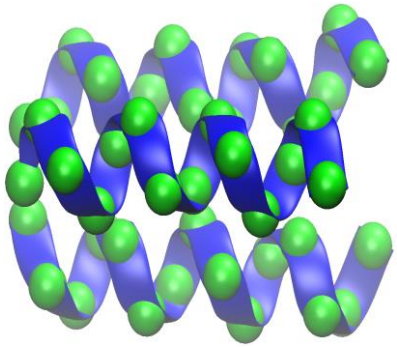
(Xia & Wei, IJNMBE, 2014)

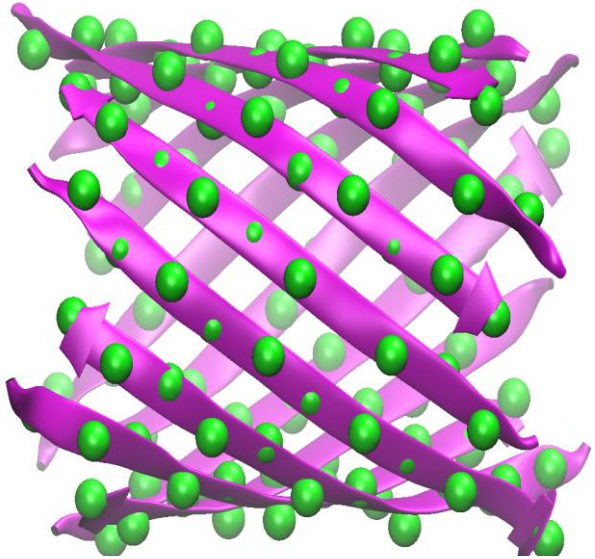
Topological fingerprints of beta barrel

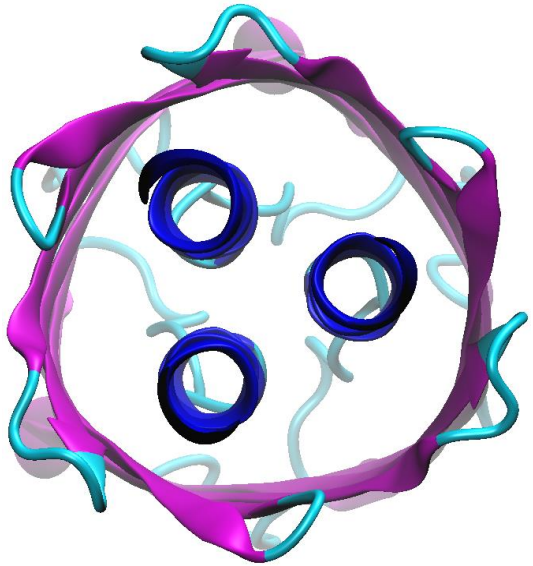
(Xia & Wei, IJNMBE, 2014)

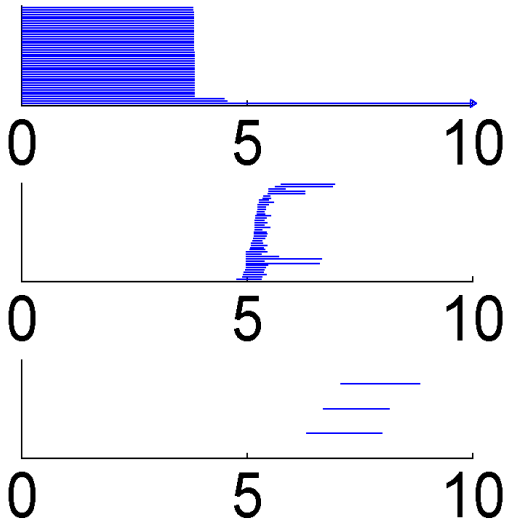
Protein:2GR8

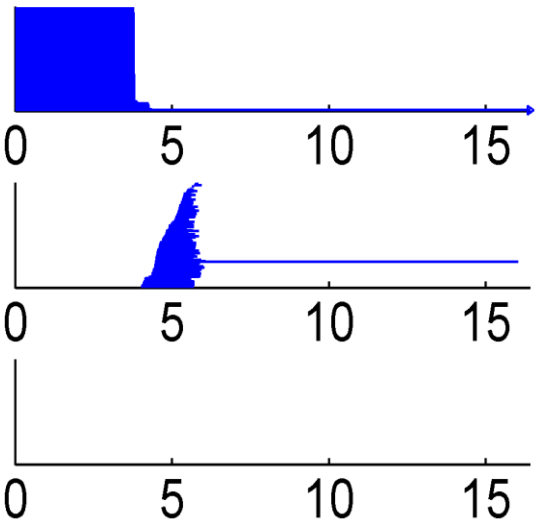






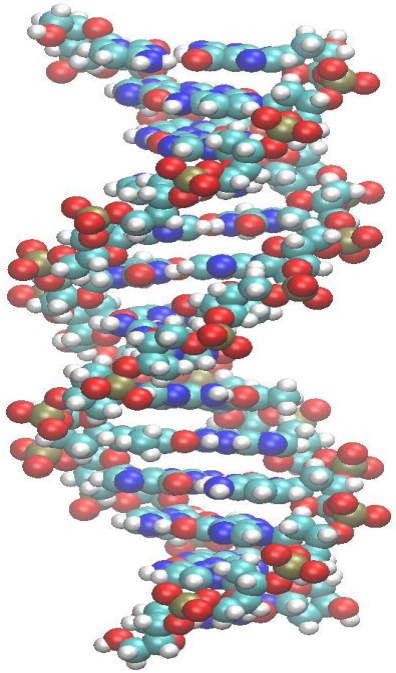


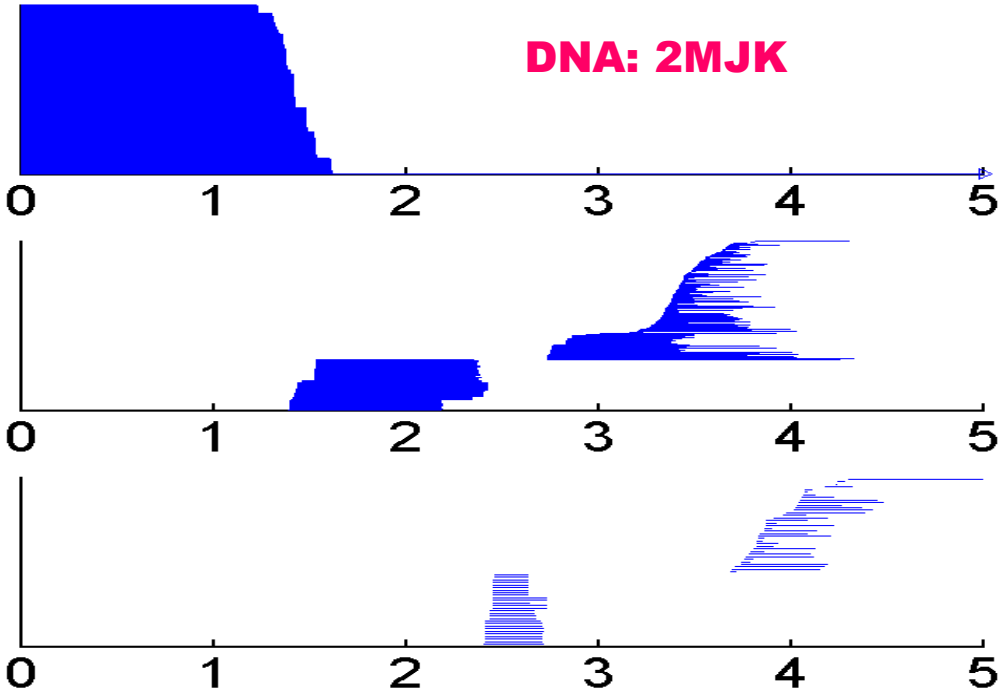


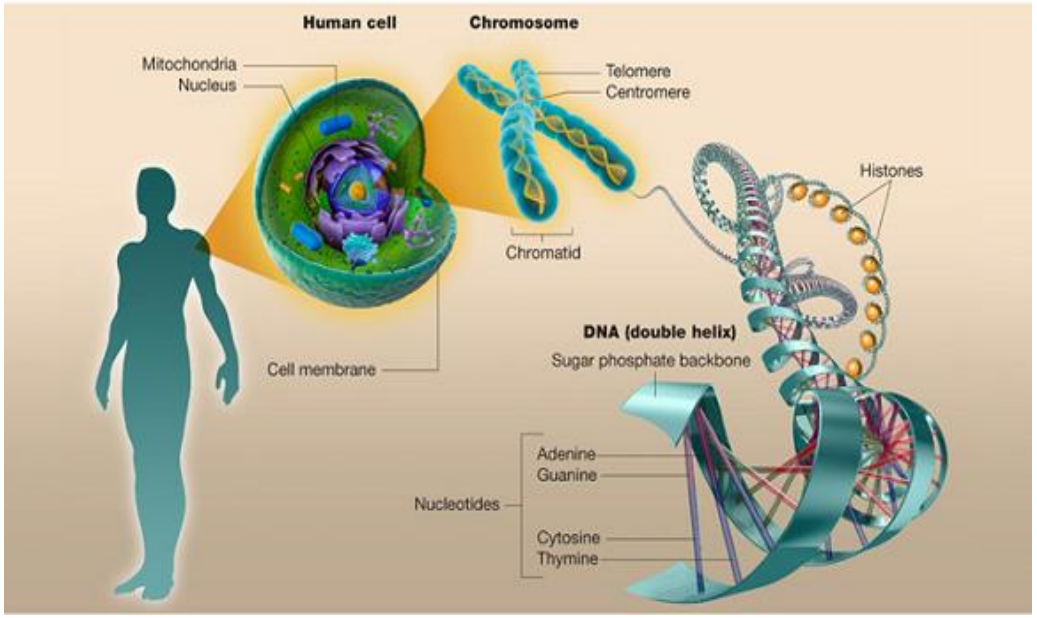


DNA topological fingerprints

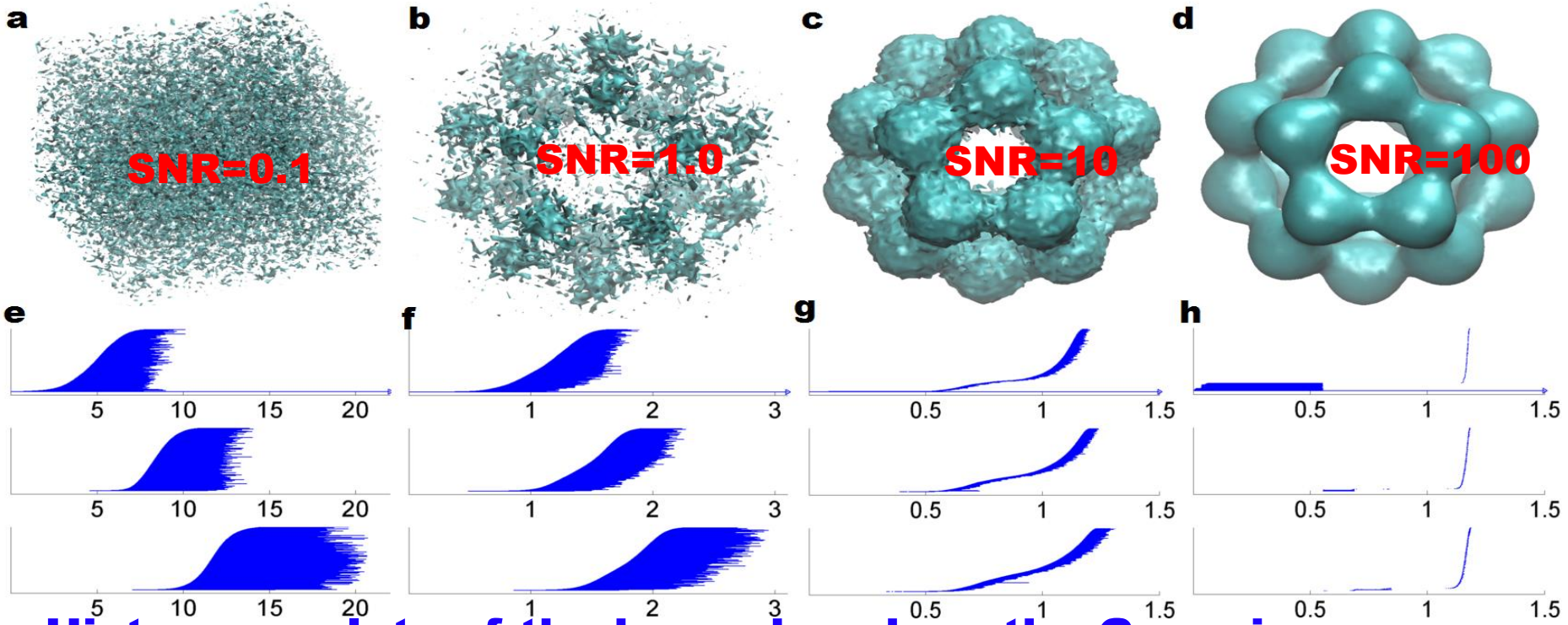
Personalized topological genome library



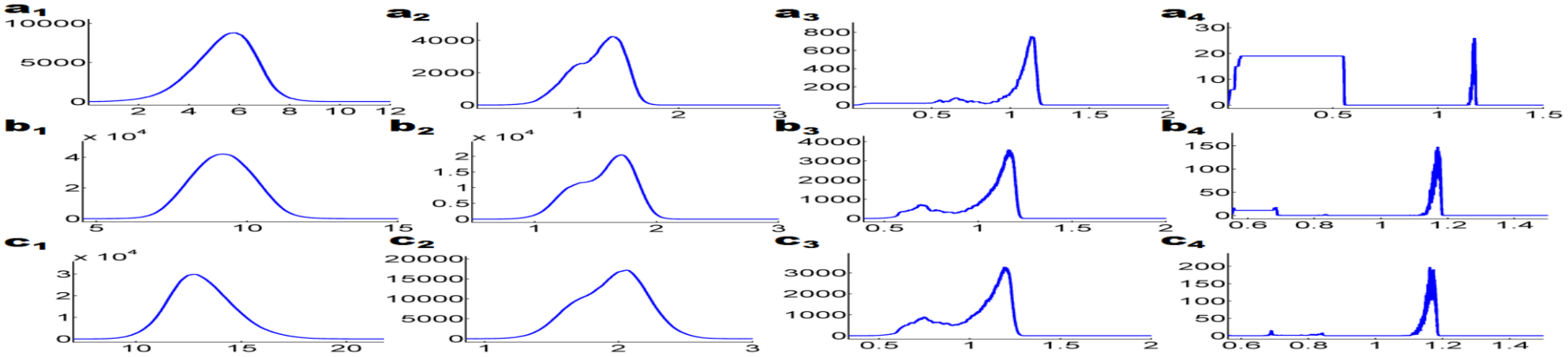




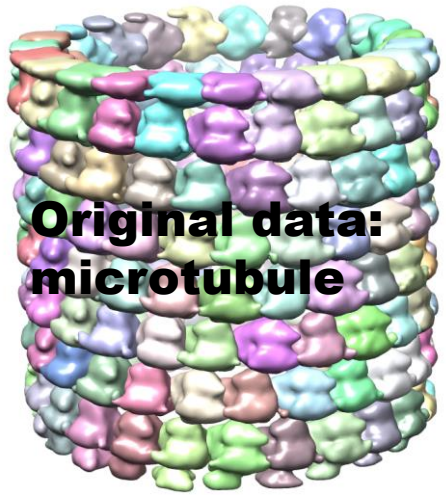
Topological signature of Gaussian noise

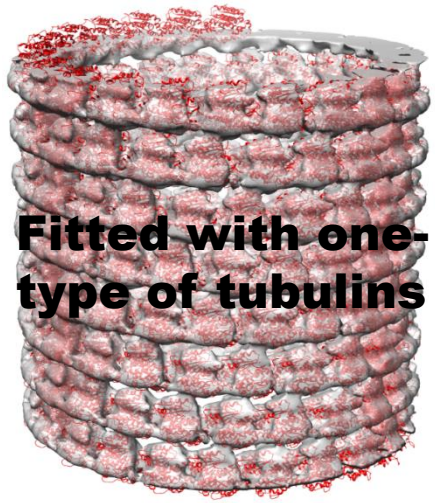


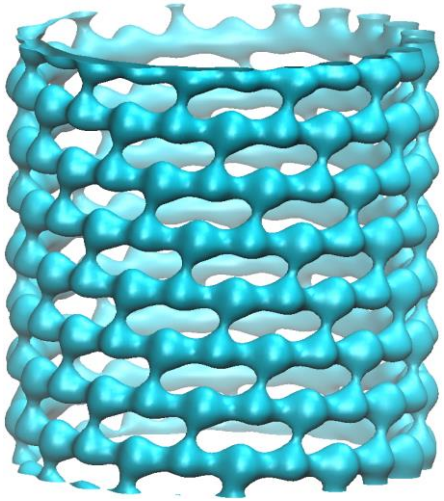
Histogram plots of the barcodes show the Gaussian distribution of the noise

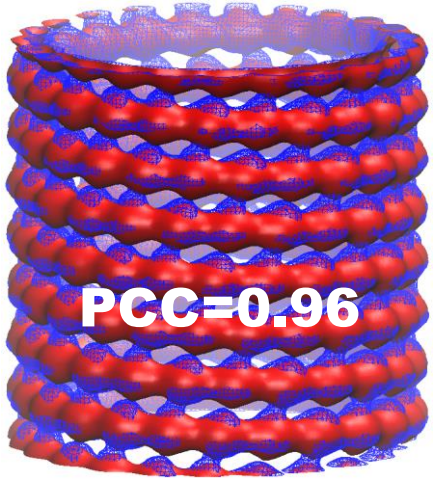


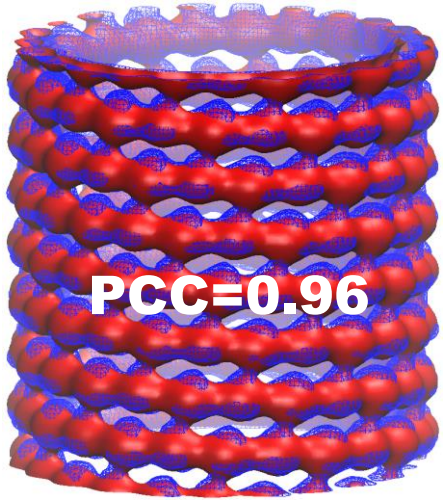
Persistent homology for ill-posed inverse problems



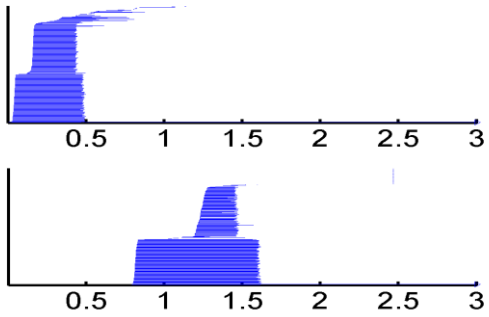


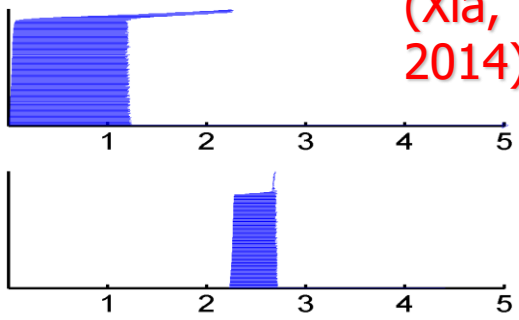


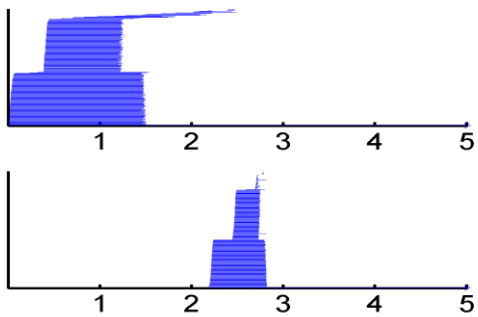




(Xia, Wei, 2014)

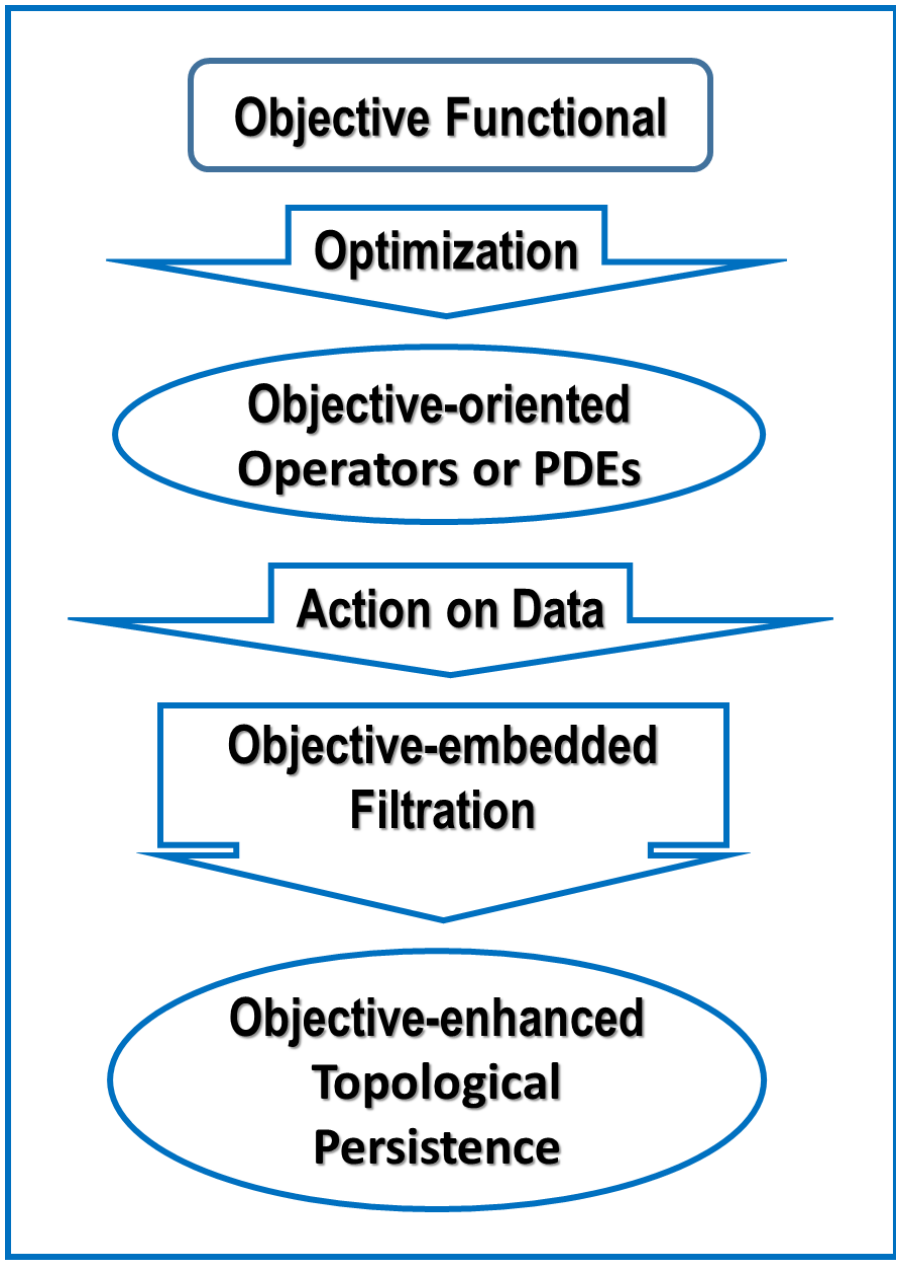






Objective oriented persistent homology

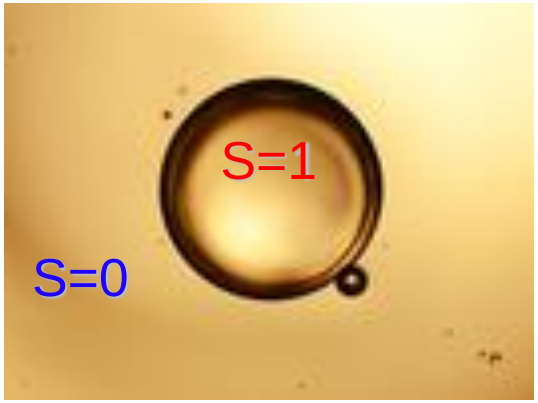
(Wang & Wei, JCP, 2016)



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:



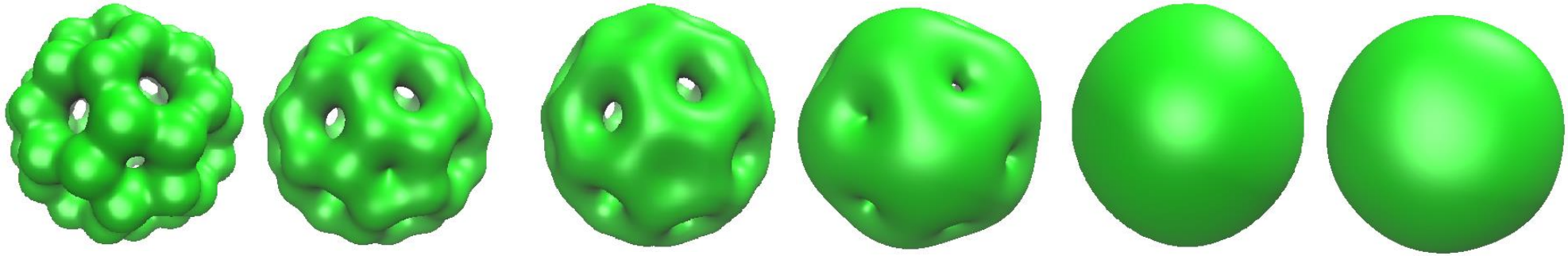
Generalized Laplace-Beltrami flow

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

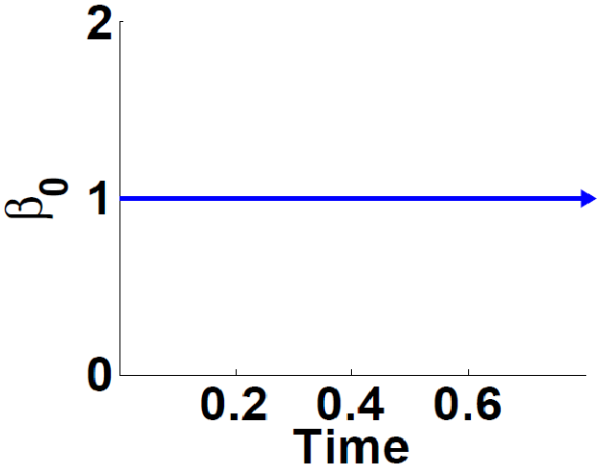
Objective oriented persistent homology

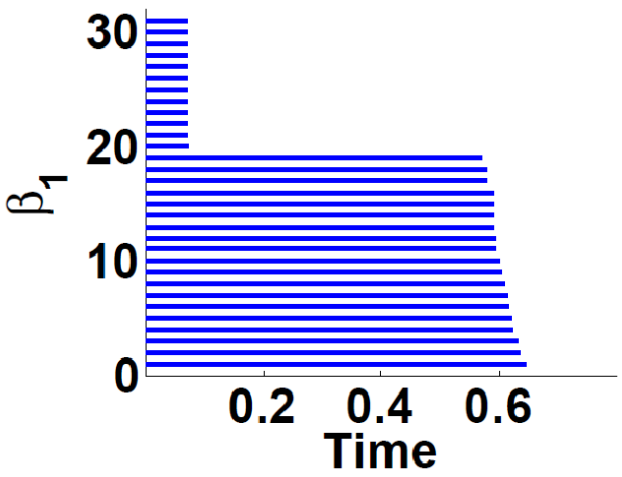
Level sets generated from Laplace-Beltrami flow

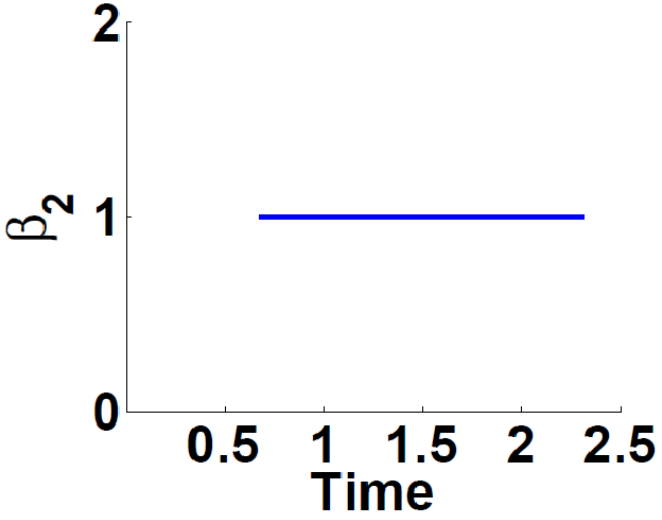
$$\frac{\partial S}{\partial t} = |\nabla S| \left[\nabla \bullet \frac{\gamma \nabla S}{|\nabla S|} \right]$$



(Wang & Wei, JCP, 2016)





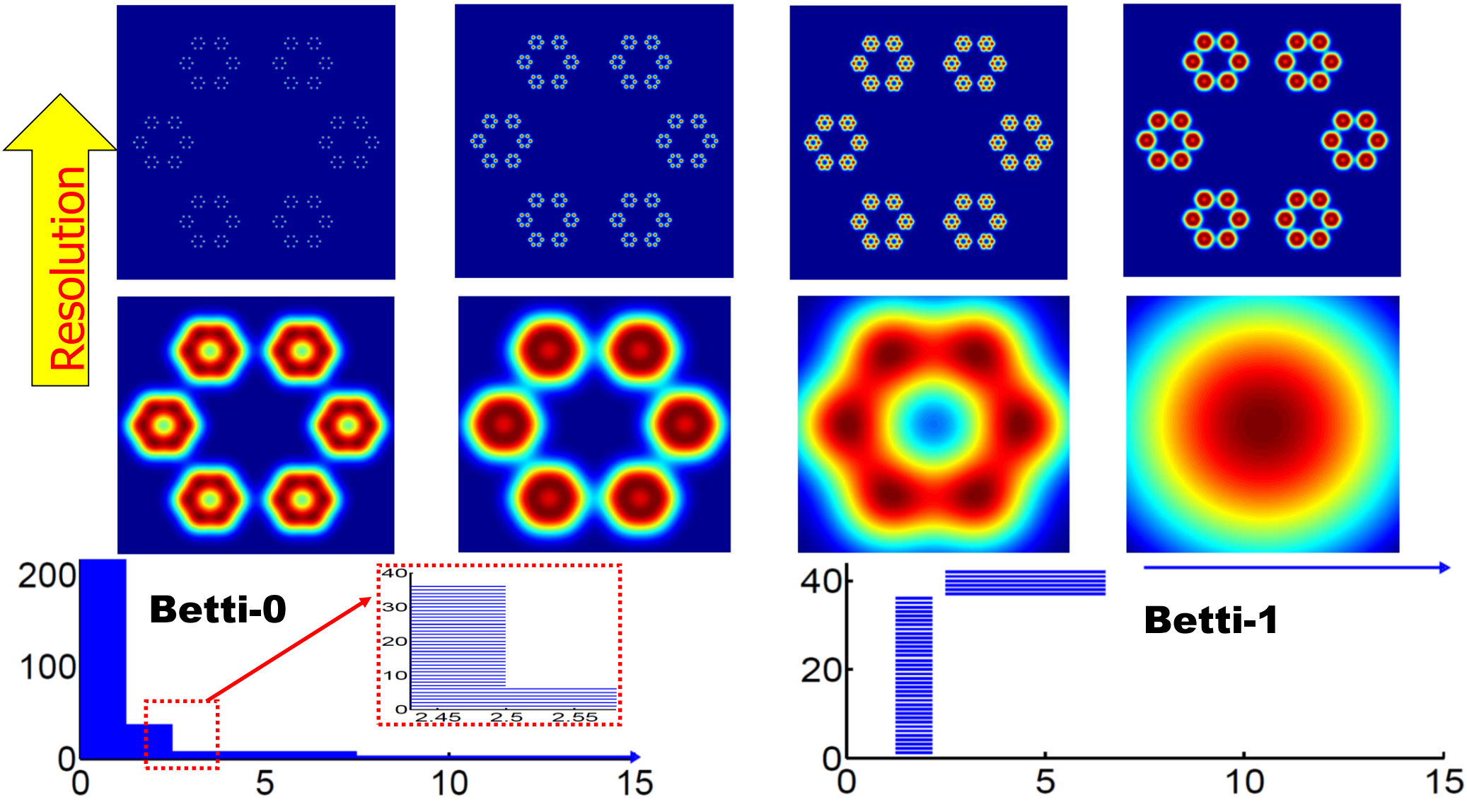


Multiresolution induced multiscale of a fractal

Introducing the resolution:

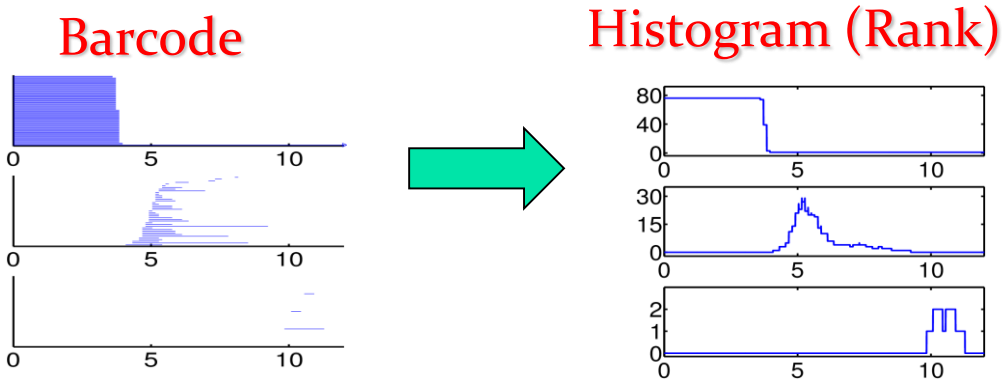
(Xia, Zhao & Wei, JCB, 2015)

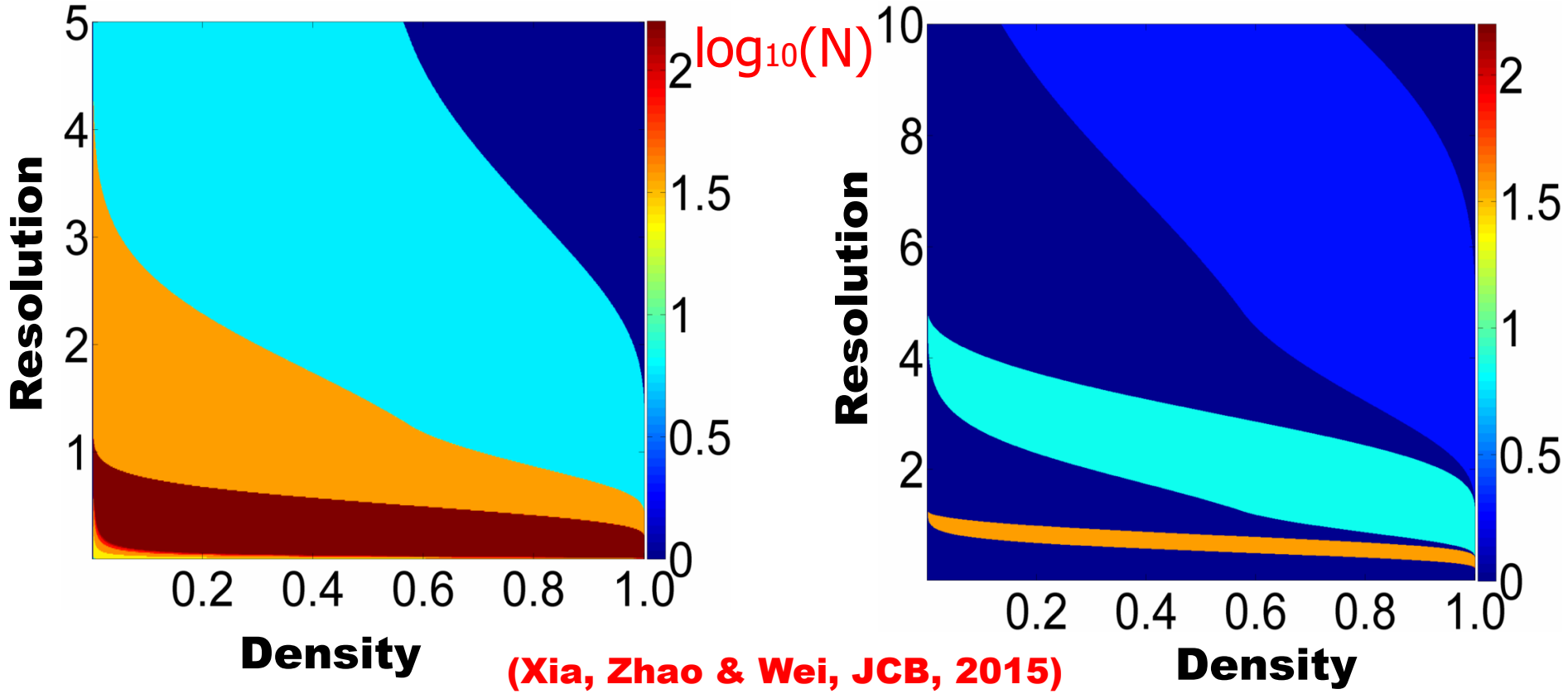
$$\rho(r,\eta)=\sum_j \exp\left(-\frac{(r-r_j)^2}{\eta^2}\right)$$



Multiresolution induced multidimensional topology of a fractal

$$\rho(r,\eta)=\sum_j \exp\left(-\frac{(r-r_j)^2}{\eta^2}\right)$$





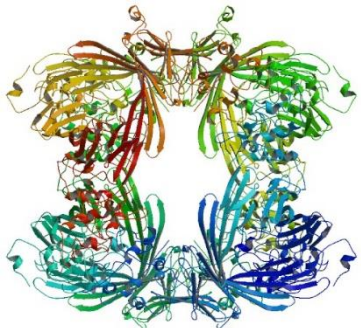
(Xia, Zhao & Wei, JCB, 2015)

High dimensional persistence generated from multi-parameter filtration --- Each *independent* parameter leads to a genuine dimension!

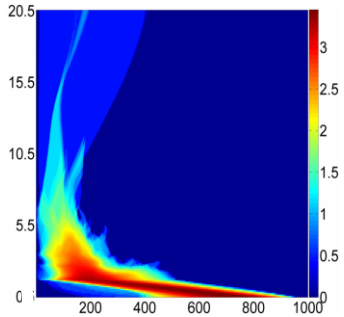
2-dimensional:

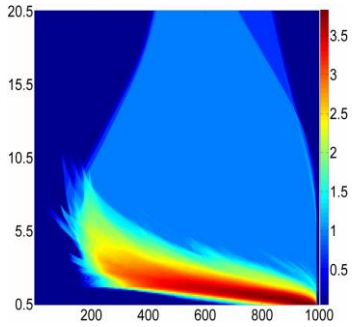
$$\rho(r,\eta)=\sum_j \exp\left(-\frac{(r-r_j)^2}{\eta^2}\right)$$

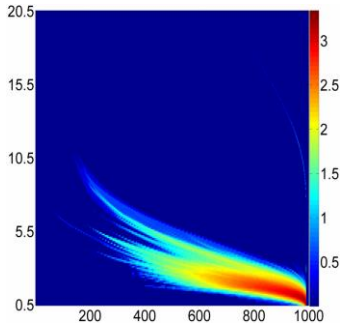
$$\{\rho(r,\eta)\geq c_k\}_{k=1}^{N}$$



2YGD







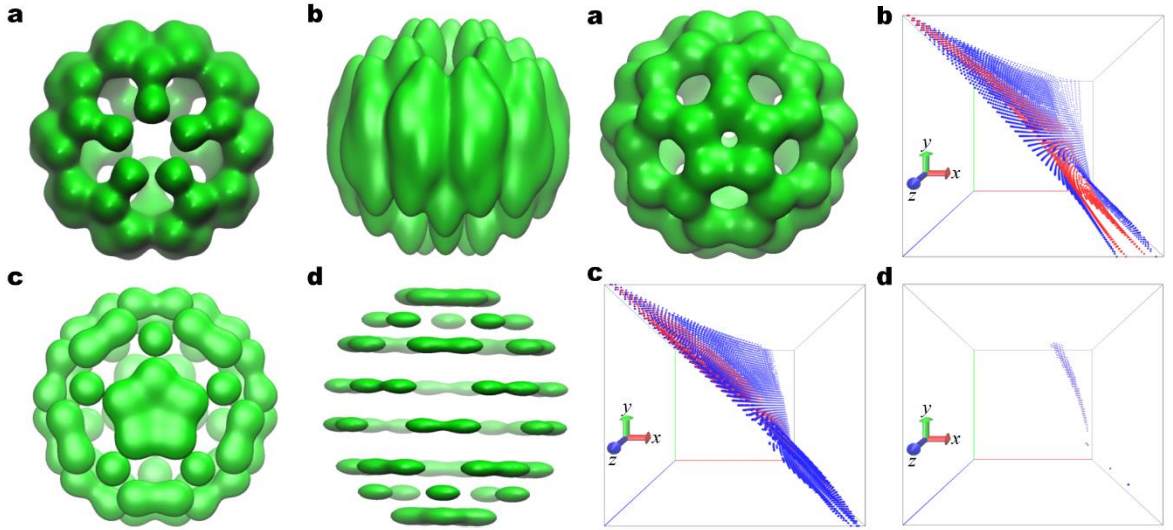
Resolution

Density

3-dimensional:

$$\rho(r,\eta_x,\eta_z)=\sum_j \exp\left(-\frac{(x-x_j)^2+(y-y_j)^2}{\eta_x^{\,2}}+\frac{(z-z_j)^2}{\eta_z^{\,2}}\right)$$

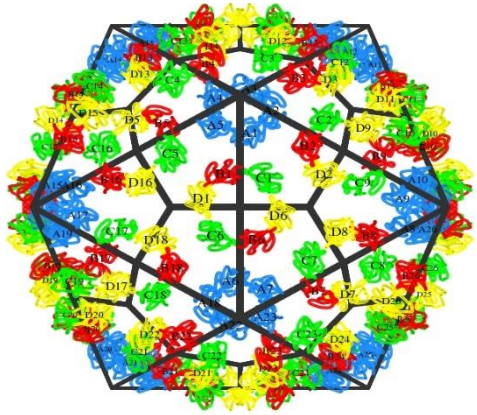
$$\{\rho(r,\eta_x,\eta_z)\geq c_k\}_{k=1}^{N}$$

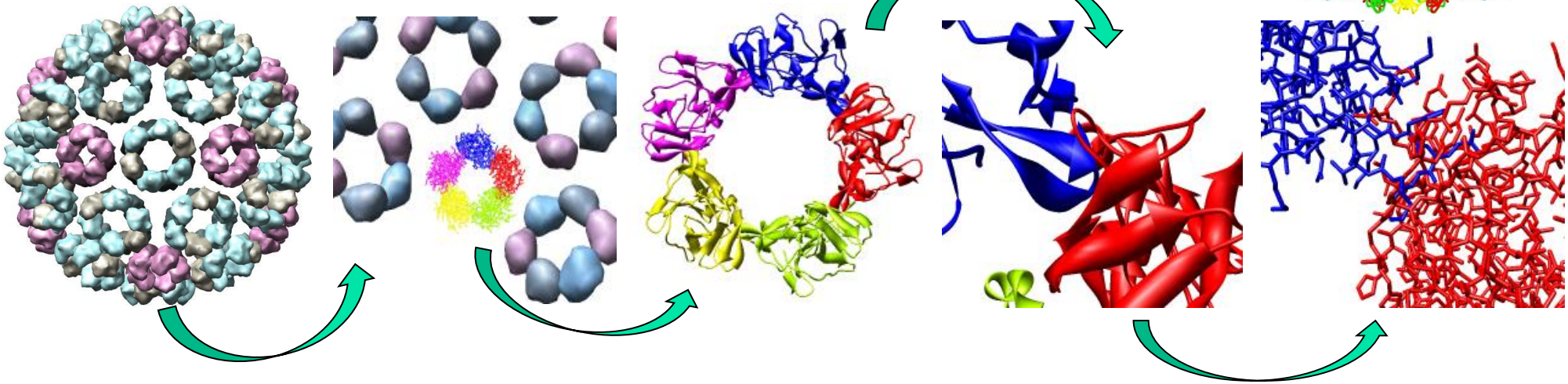


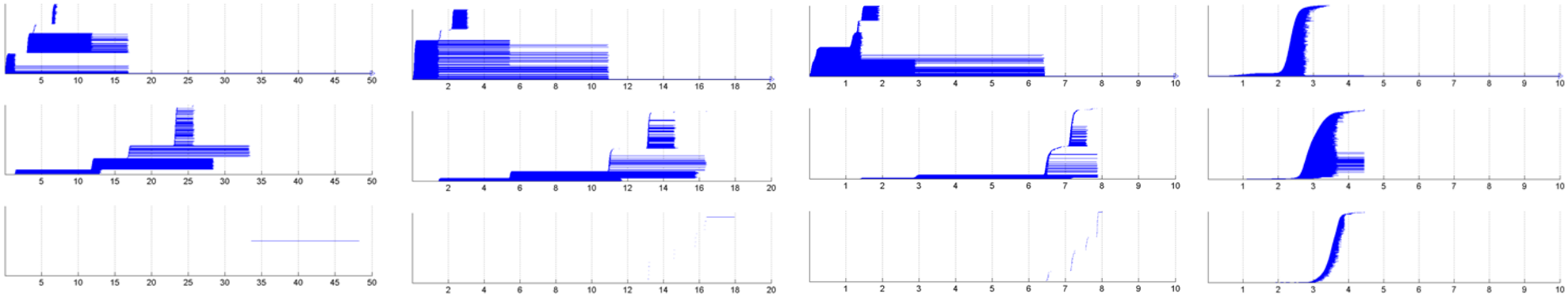
(Xia & Wei, JCC, 2015)

Multiscale topological persistence of a virus

Virus 1DYL has 12 pentagons and 30 hexagons with icosahedral symmetry. The diameter is about 700 Angstrom.



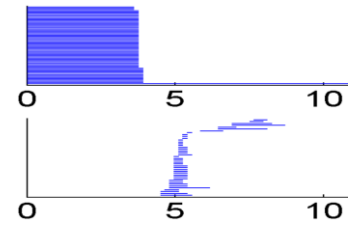


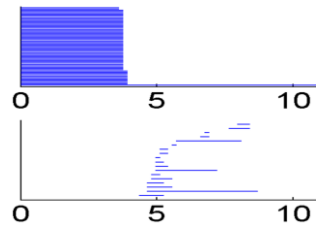


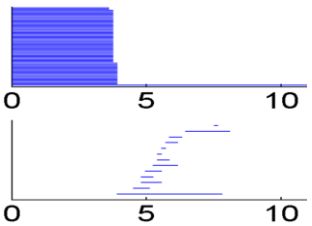
Scale=12A **Scale=8A** **Scale=4A** **Scale=2A**

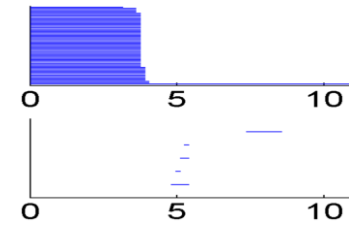
Topological analysis of protein folding

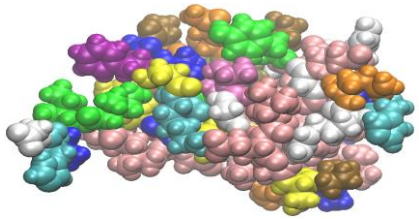
ID: 1I2T

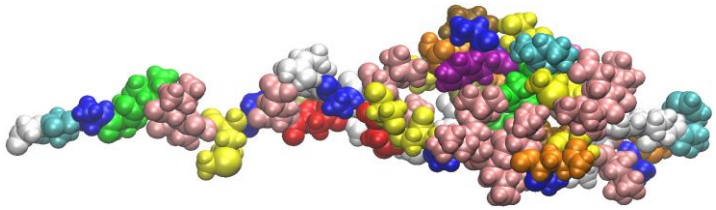


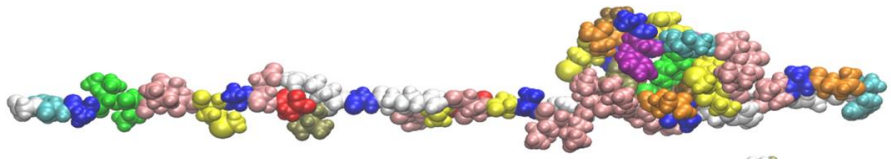


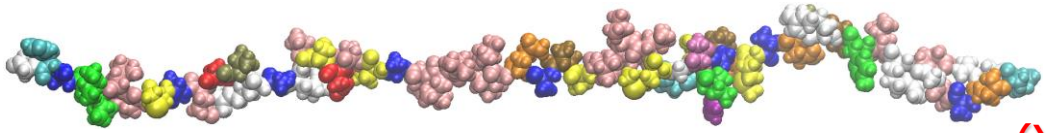


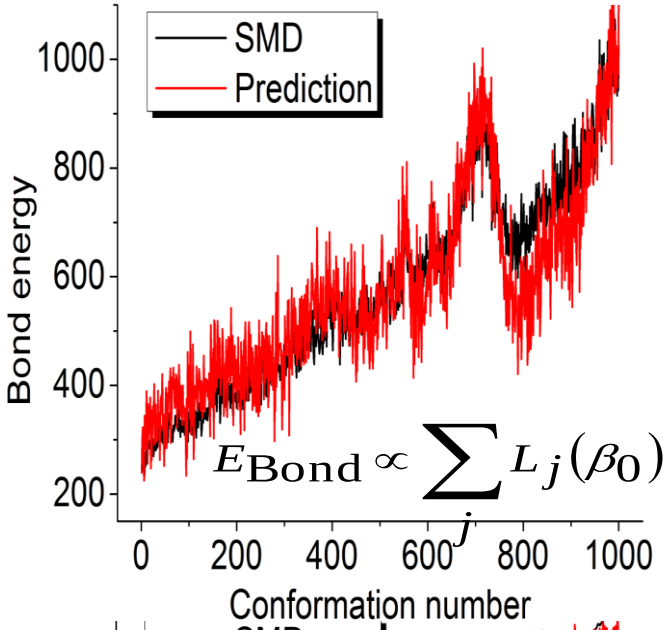


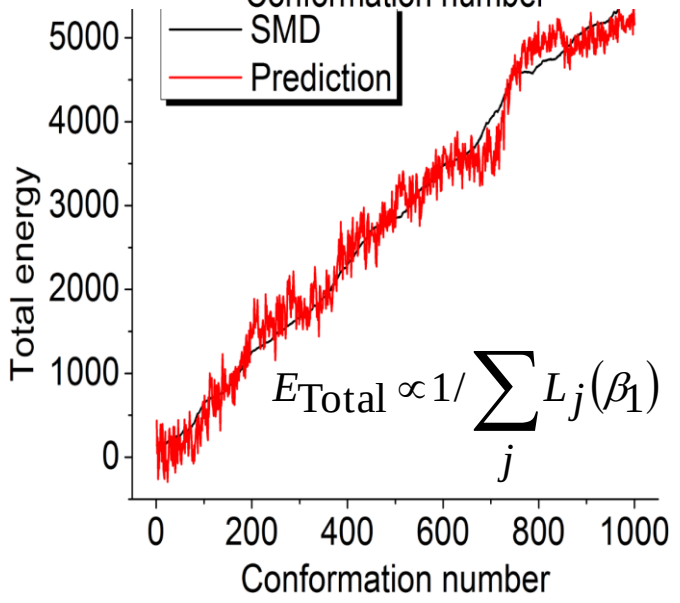








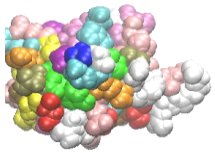


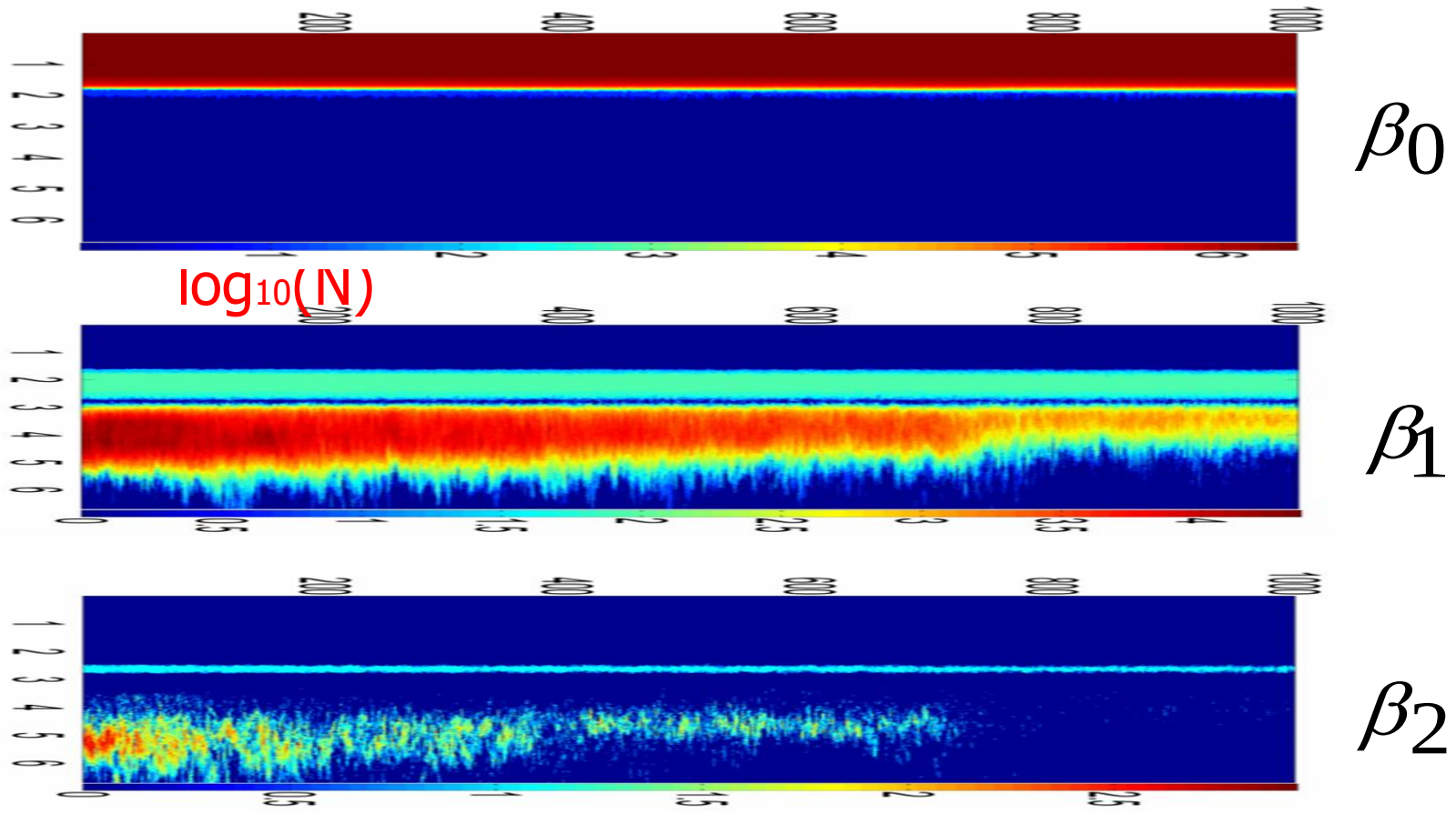


Quantitative!

(Xia, Wei, IJNMBE, 2014)

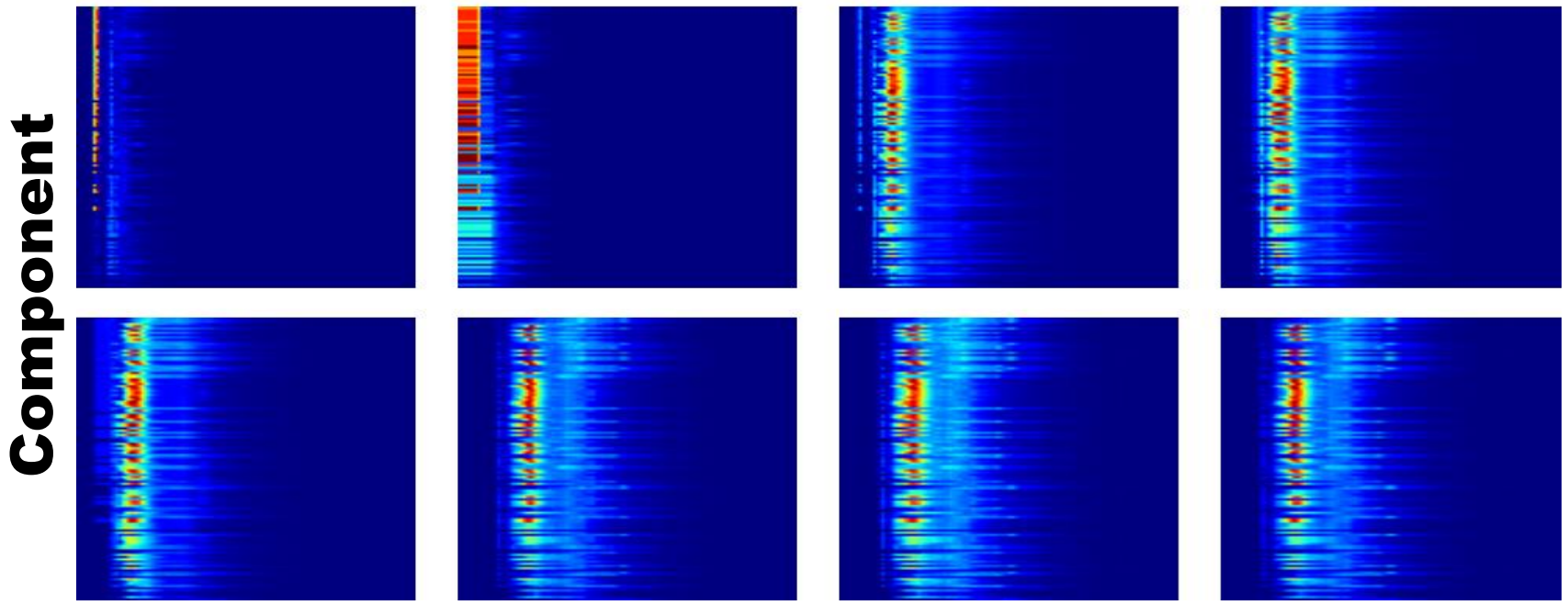
2D persistence in protein 1UBQ unfolding





(Xia & Wei, JCC, 2015)

Multicomponent and multichannel persistent homology for a protein-drug complex

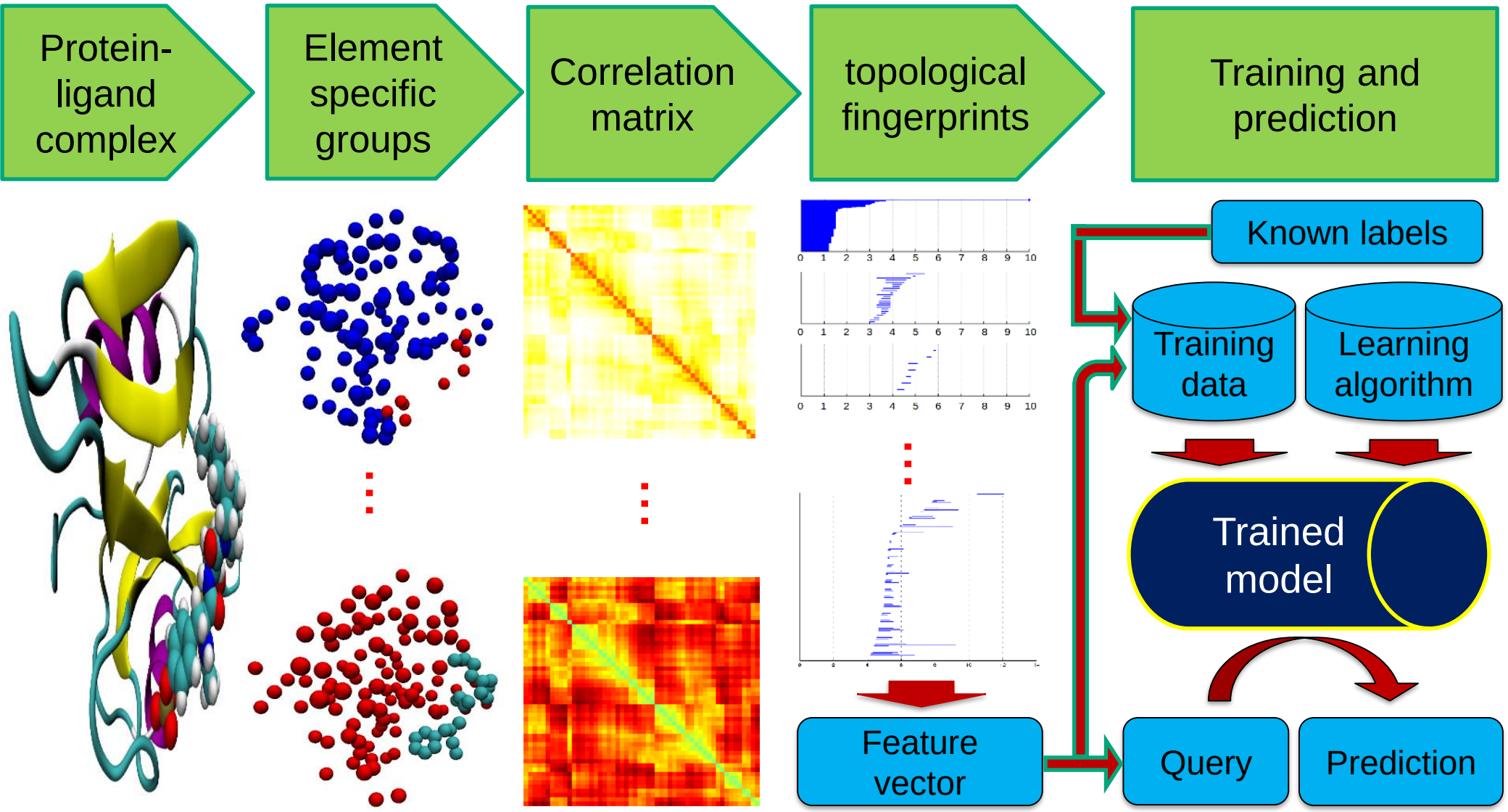


Components are generated from element specific persistent homology. Eight channels are constructed from births, deaths and persistences at Betti-0, Betti-1 and Betti-2.

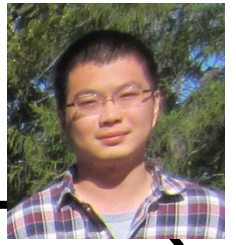
(Cang & Wei, IJNMBE, 2017)

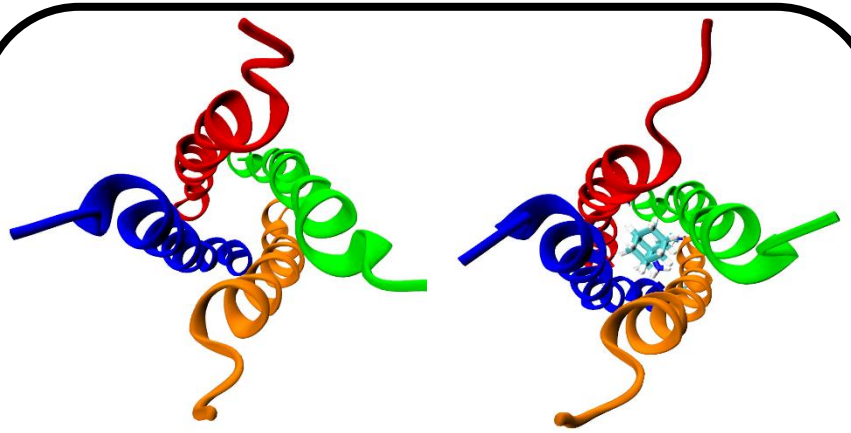
Topology based learning architecture

(Cang & Wei, Bioinfomatics, 2017)

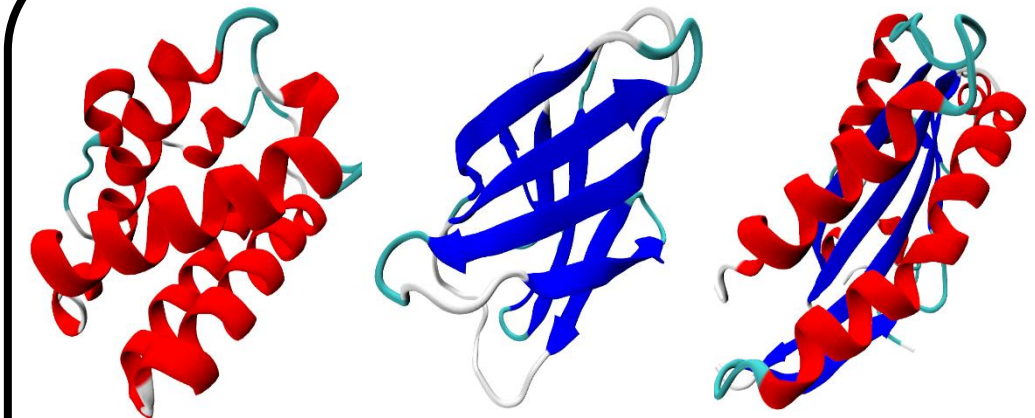


Topological fingerprint based machine learning method for the classification of 2400 proteins



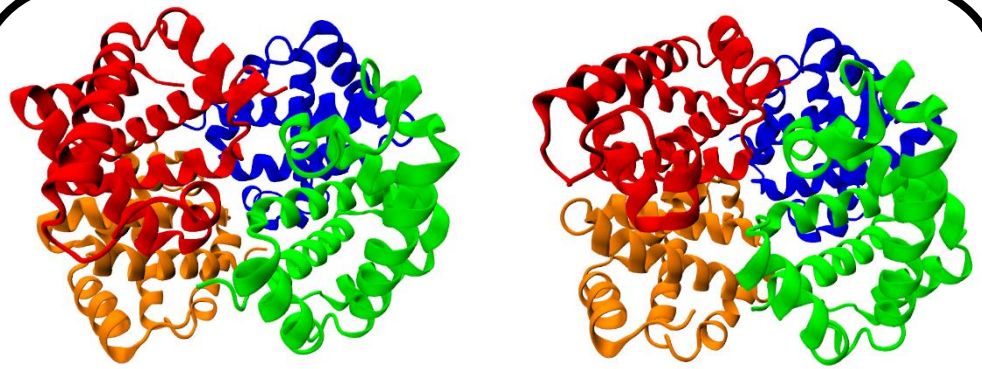


Influenza A virus drug inhibition: 96% Accuracy



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

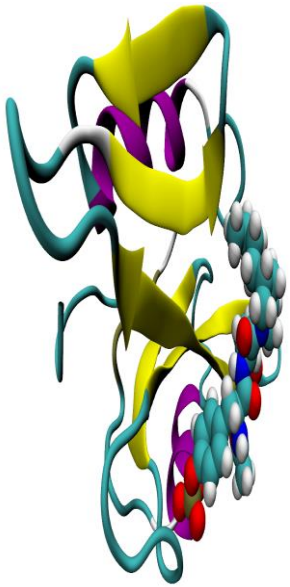
(Cang et al, MBMB, 2015)

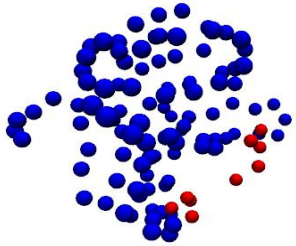


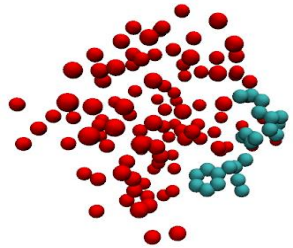
Hemoglobins in their relaxed and taut forms: 80% accuracy

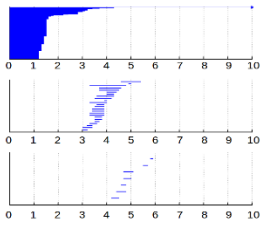
55 classification tasks of protein superfamilies over 1357 proteins from Protein Classification Benchmark Collection: 82% accuracy

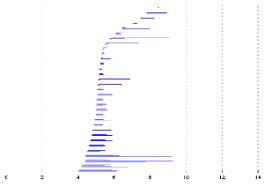
Topology based convolutional deep Learning









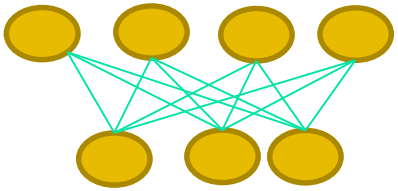


Multichannel images (54x200)

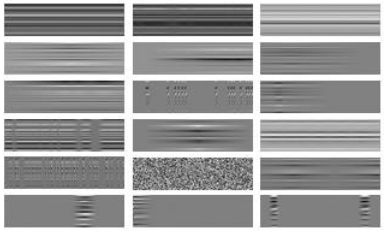
Convolution (128x200)

Pooling (128x100)

Flattening (1xN)



Prediction



Original Complex

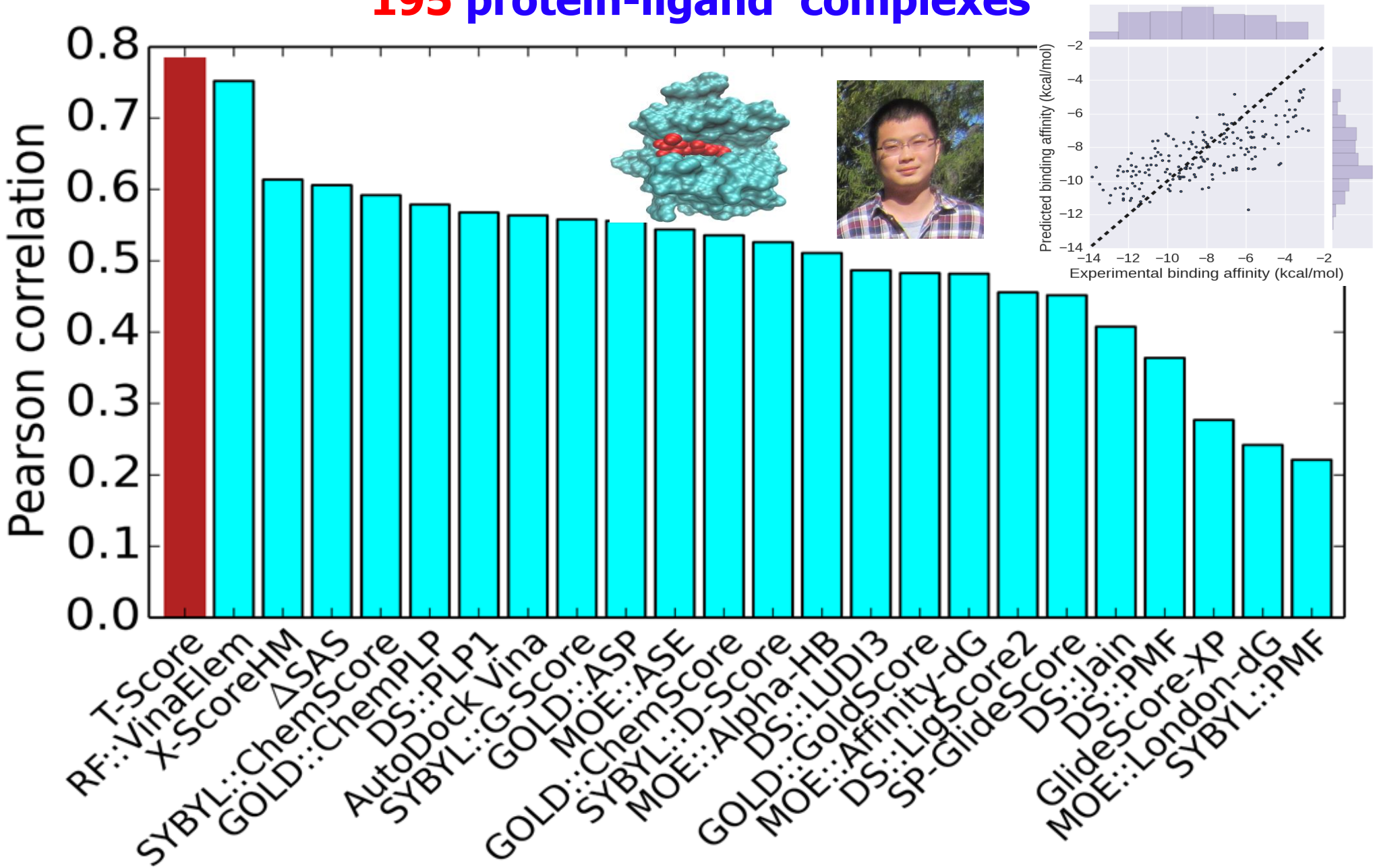
Classify atoms into element specific groups

Generate topological fingerprints

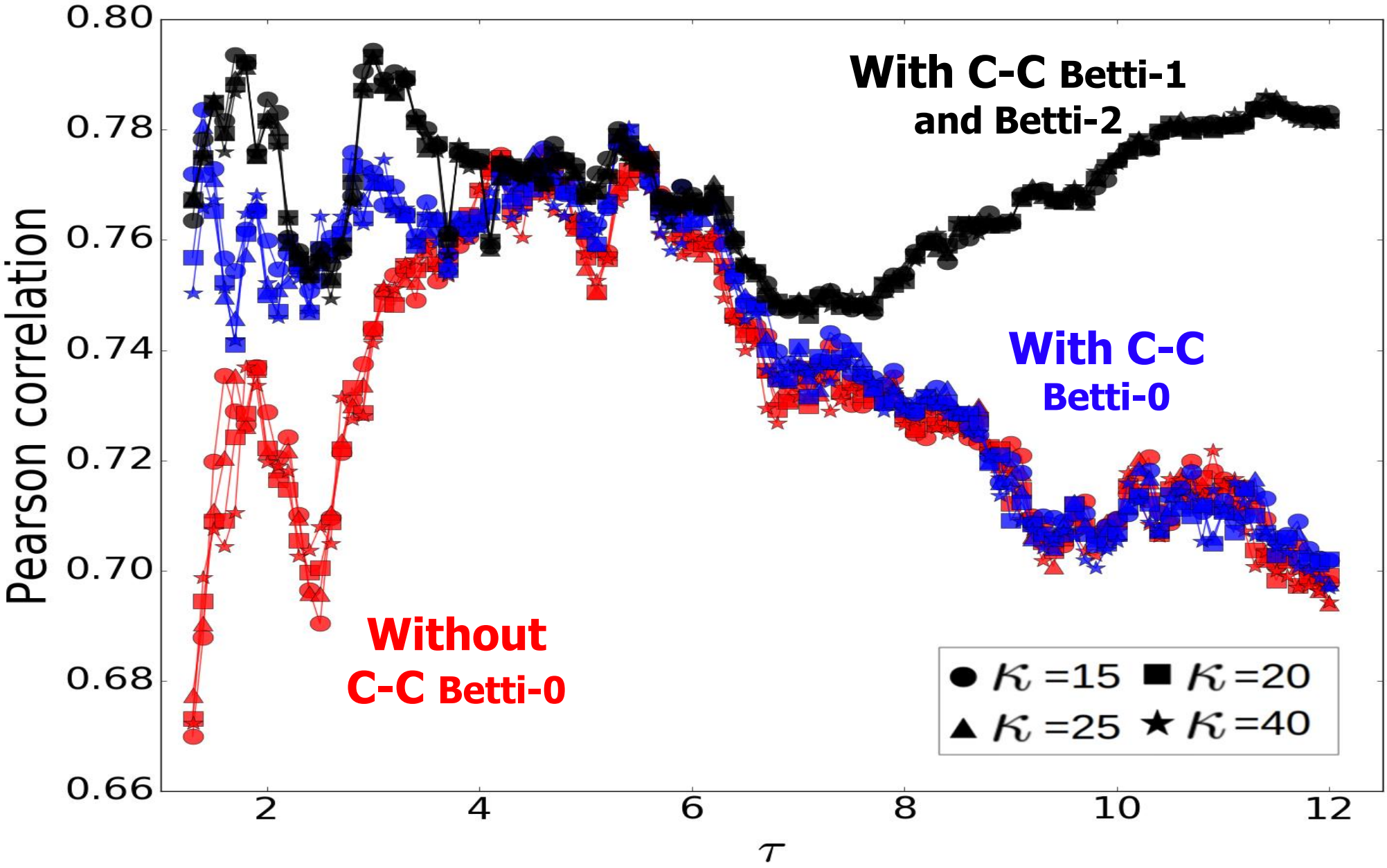
Convolutional deep learning neural network

(Cang & Wei, Plos CB, 2017)

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

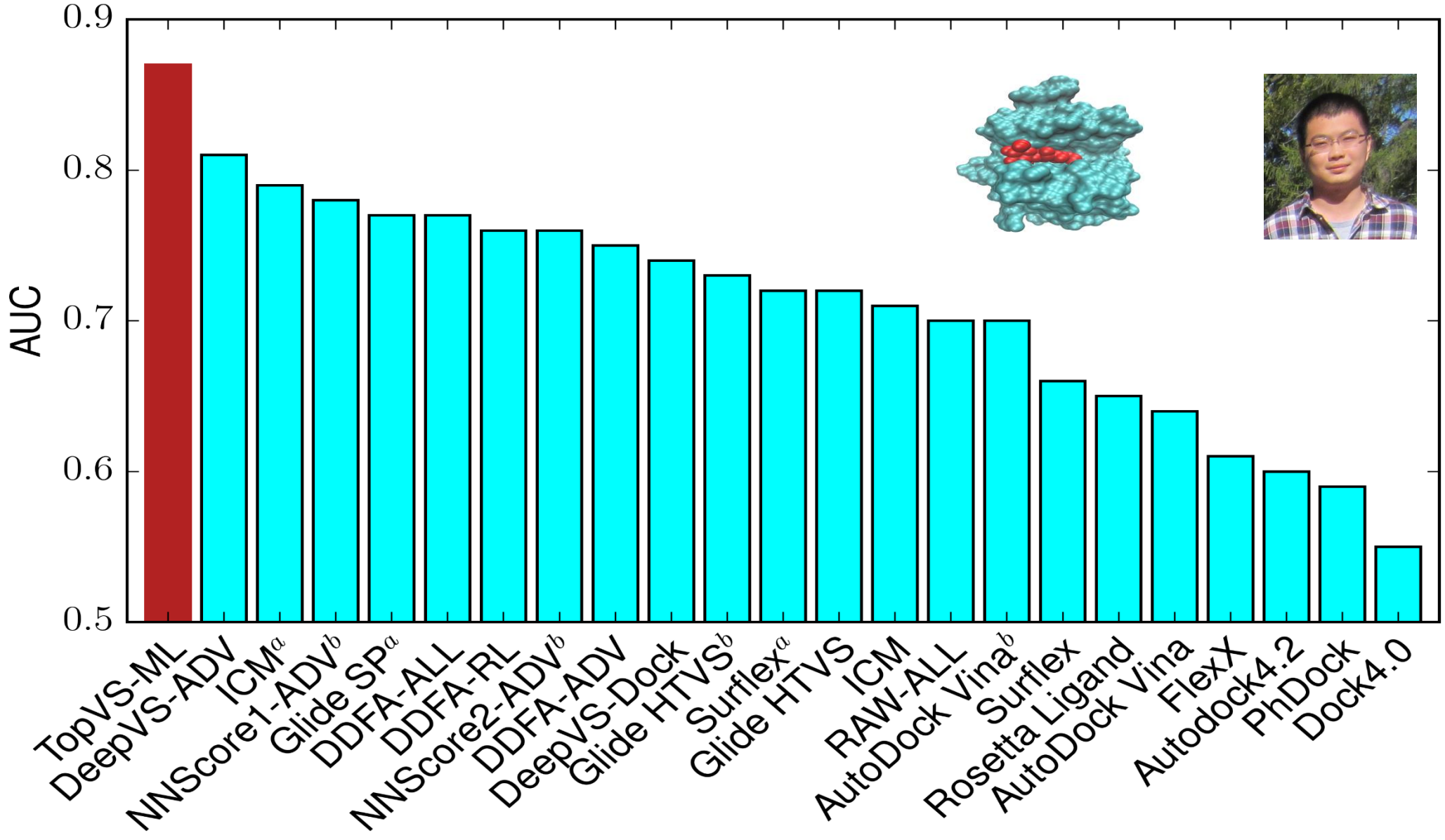


Topology based binding affinity prediction of PDBBind v2007 core set of 195 protein-ligand complexes



Directory of Useful Decoy (DUD)

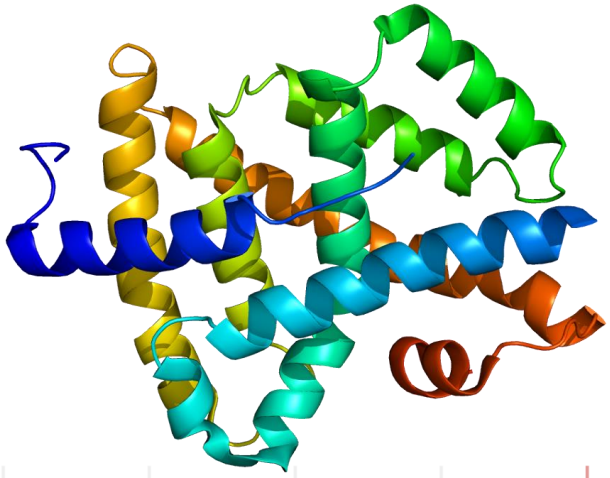
Classification of 98266 compounds containing 95316 decoys and 2950 active ligands binding to 40 targets from six families

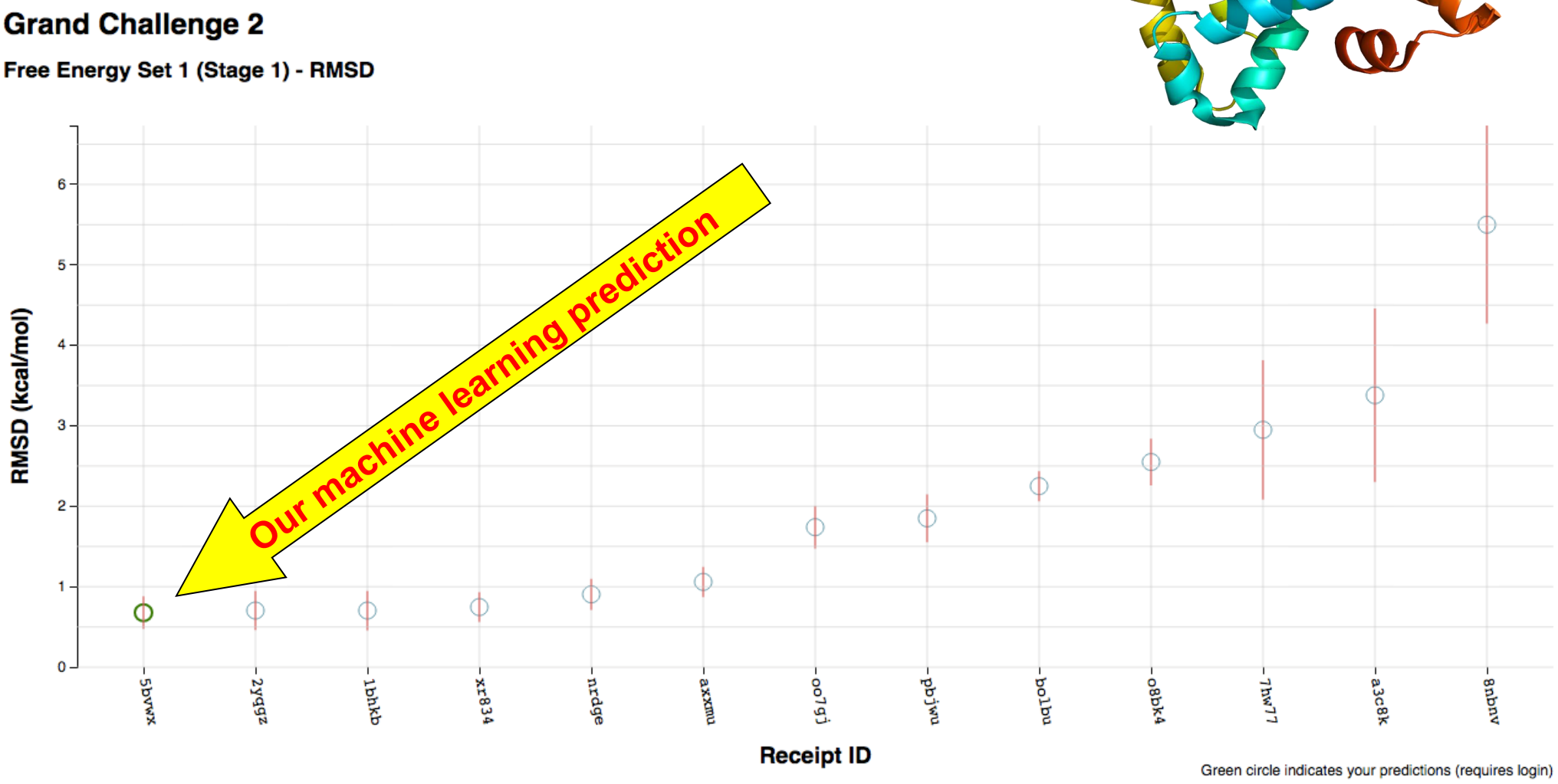


D3R Grand Challenge 2

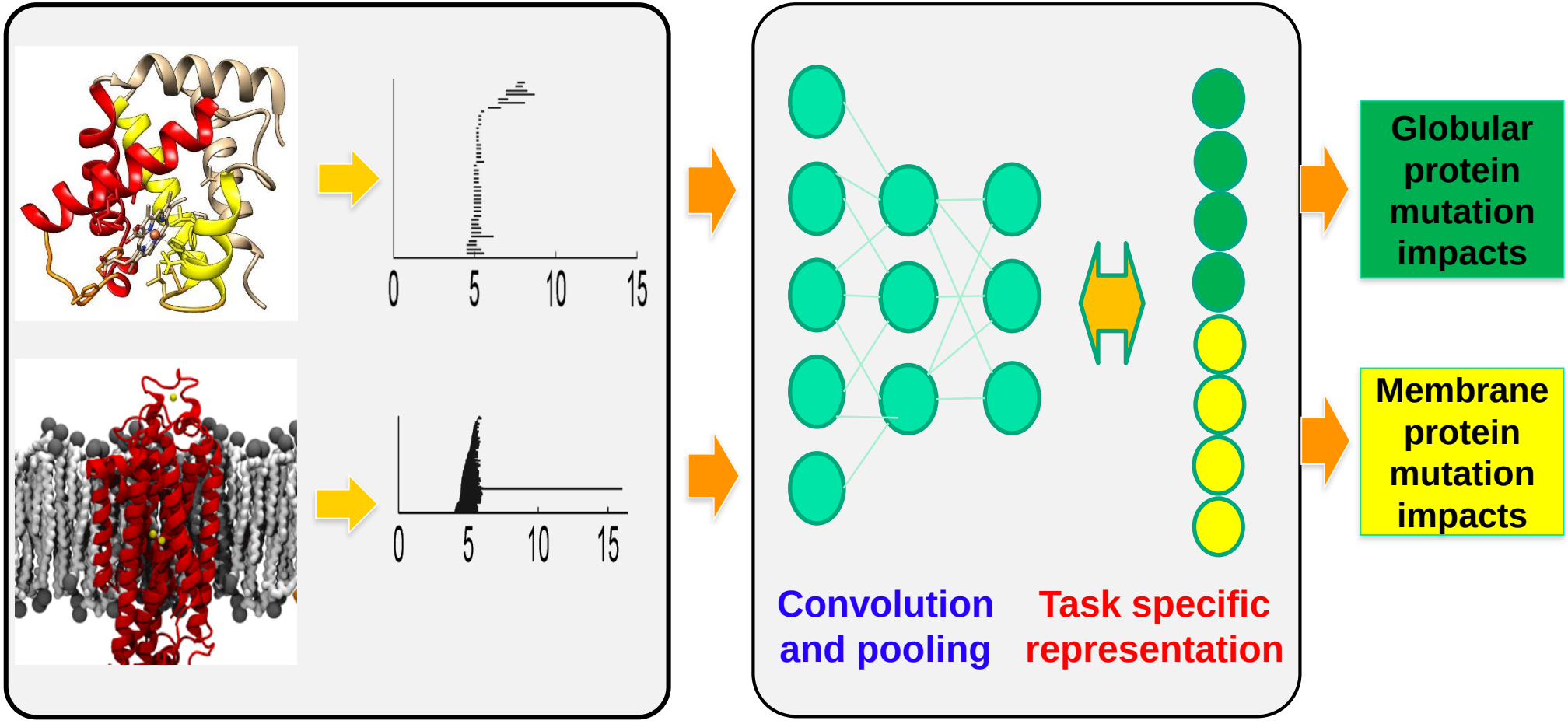
Given: Farnesoid X receptor (FXR) and 102 ligands

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





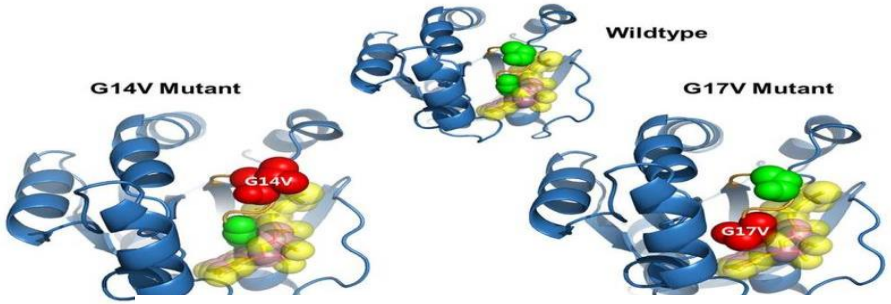
Topological Multi-Task Deep Learning

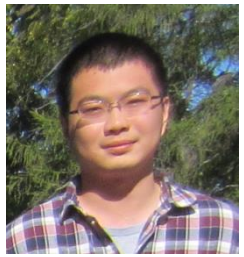


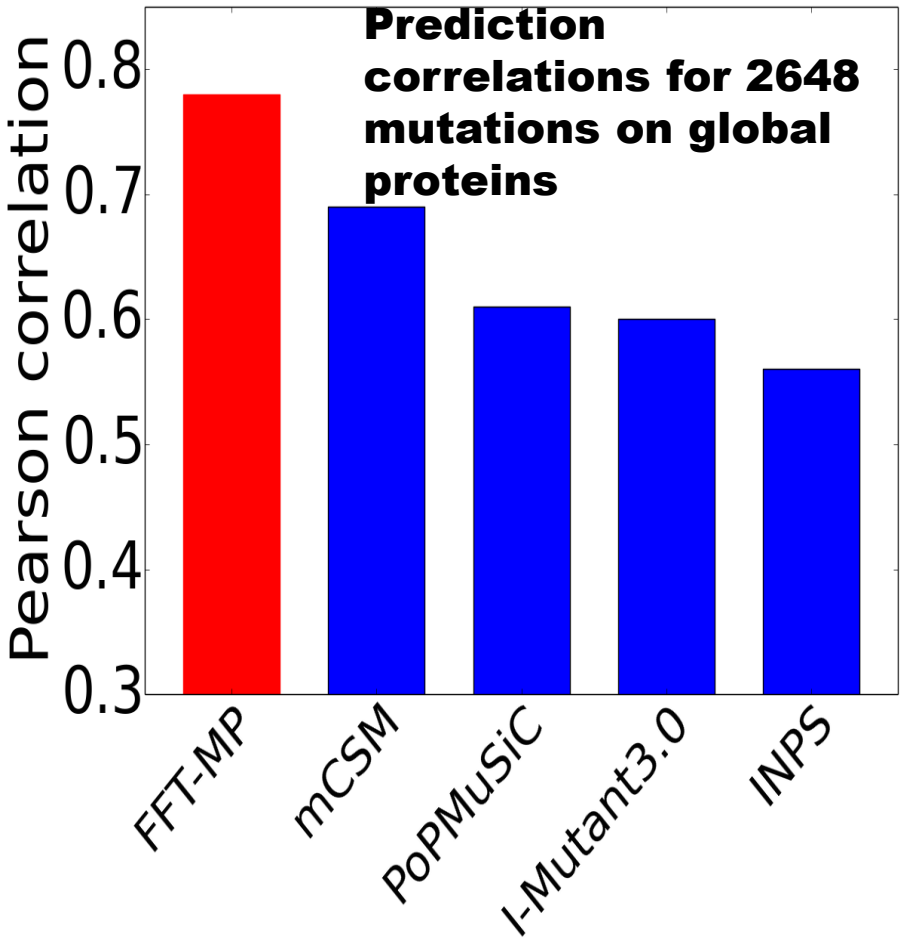
Topological feature extraction

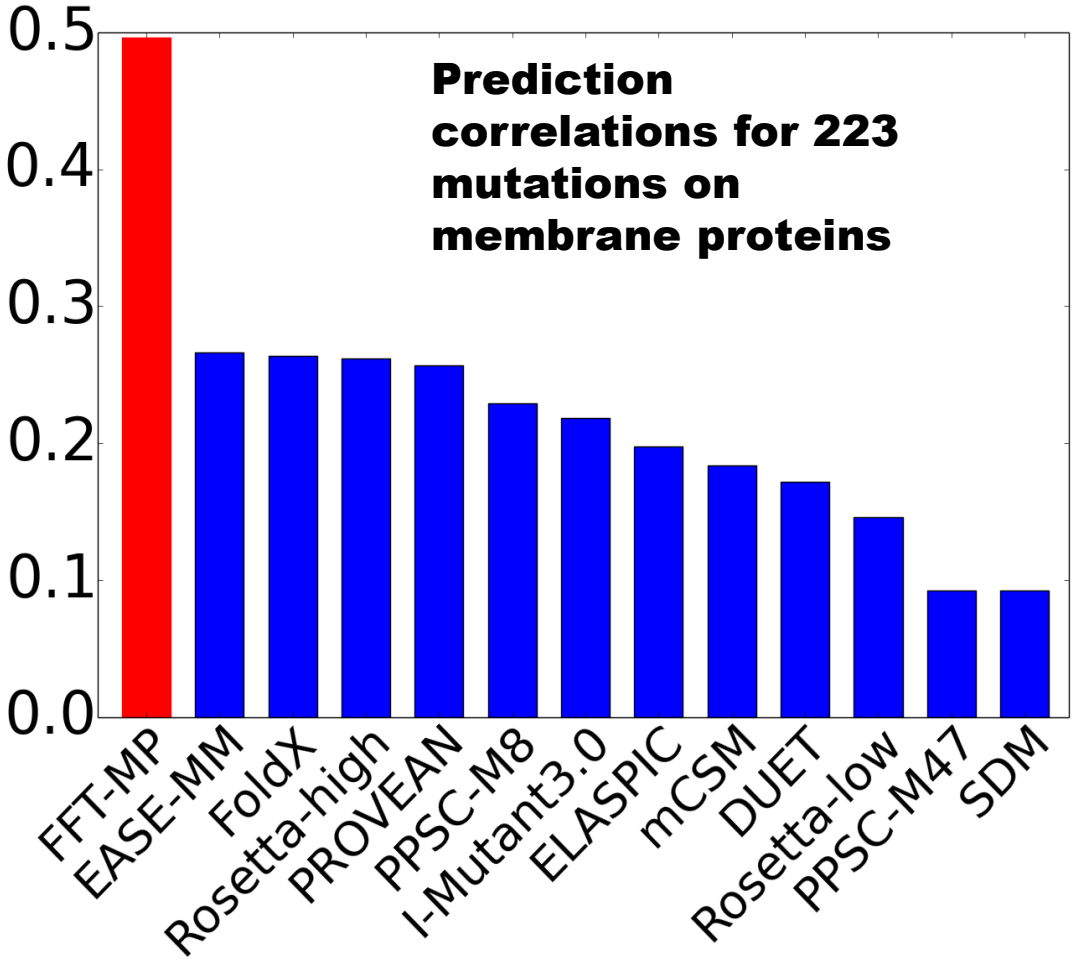
Multi-task topological deep learning

Blind prediction of mutation energies









(Cang & Wei, Bioinformatics, 2017)

Comparison of the predictabilities of mutation predictors (MPs) constructed by Topology (T-MP), Electrostatics (E-MP), Geometry (G-MP), Sequence (S-MP) and High-level feature (H-MP)

Features	S350		S2648		Q3421		M223	
	R_P	RMSE	R_P	RMSE	R_P	RMSE	R_P	RMSE
T-MP-2	0.817(0.002)	0.92(0.004)	0.789(0.005)	0.91(0.009)	0.803(0.008)	1.18(0.020)	0.575(0.019)	1.08(0.018)
T-MP-1	0.765(0.003)	1.02(0.006)	0.746(0.006)	0.98(0.009)	0.767(0.006)	1.27(0.014)	0.543(0.022)	1.11(0.020)
E-MP	0.760(0.003)	1.02(0.005)	0.721(0.005)	1.02(0.008)	0.733(0.009)	1.34(0.018)	0.525(0.026)	1.14(0.026)
G-MP	0.759(0.004)	1.03(0.006)	0.716(0.004)	1.03(0.007)	0.724(0.008)	1.37(0.015)	0.474(0.033)	1.17(0.027)
S-MP	0.609(0.005)	1.26(0.006)	0.616(0.006)	1.16(0.007)	0.581(0.006)	1.61(0.008)	0.379(0.029)	1.27(0.025)
H-MP	0.686(0.004)	1.14(0.006)	0.662(0.009)	1.11(0.013)	0.654(0.009)	1.50(0.016)	0.231(0.048)	1.41(0.043)

(Cang & Wei, Bioinformatics, 2017)

Prediction of partition coefficients: Star Set (223 molecules)

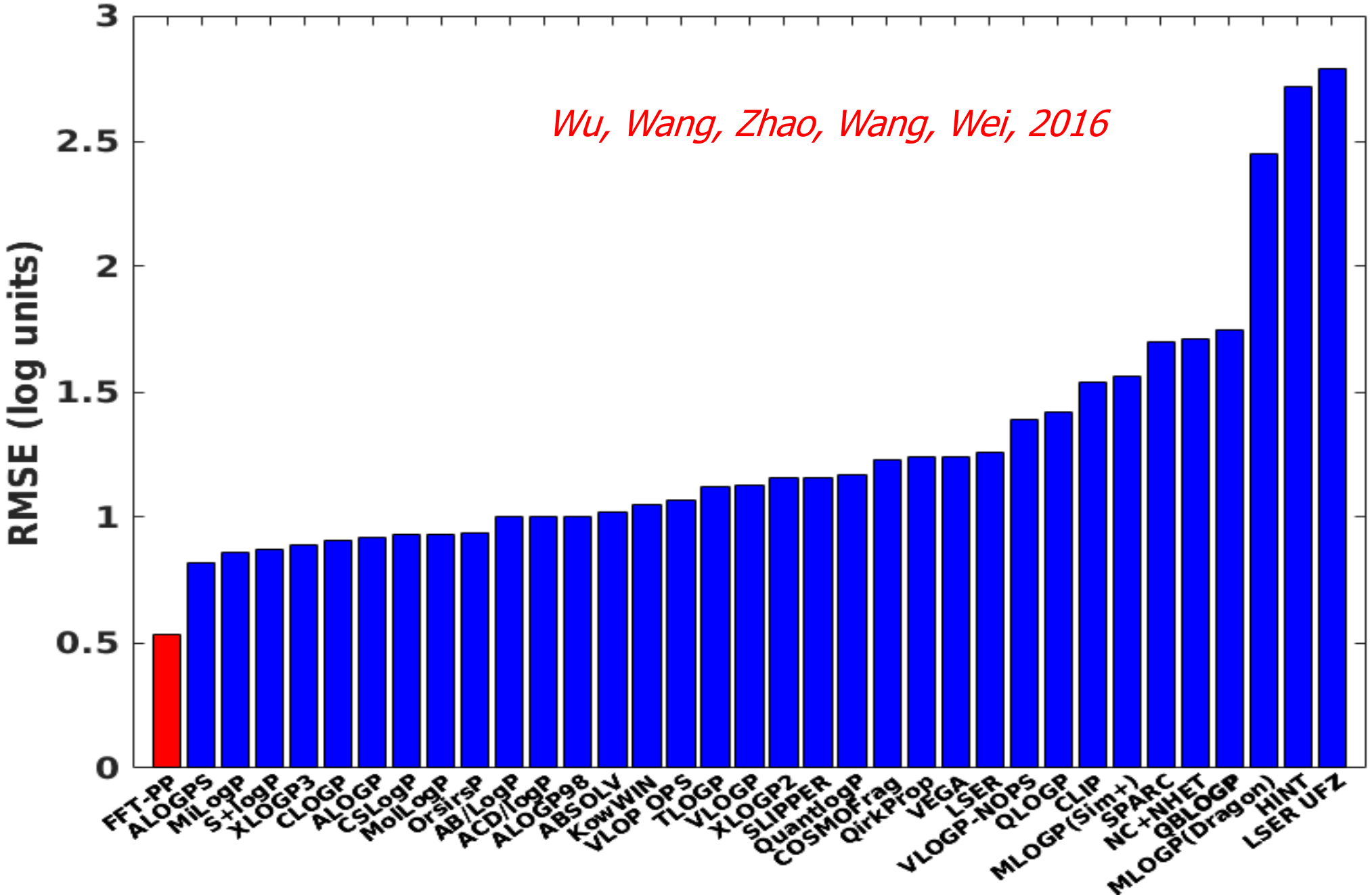
Wu, Wang, Zhao, Wang, Wei, 2016

RMSE (log units)

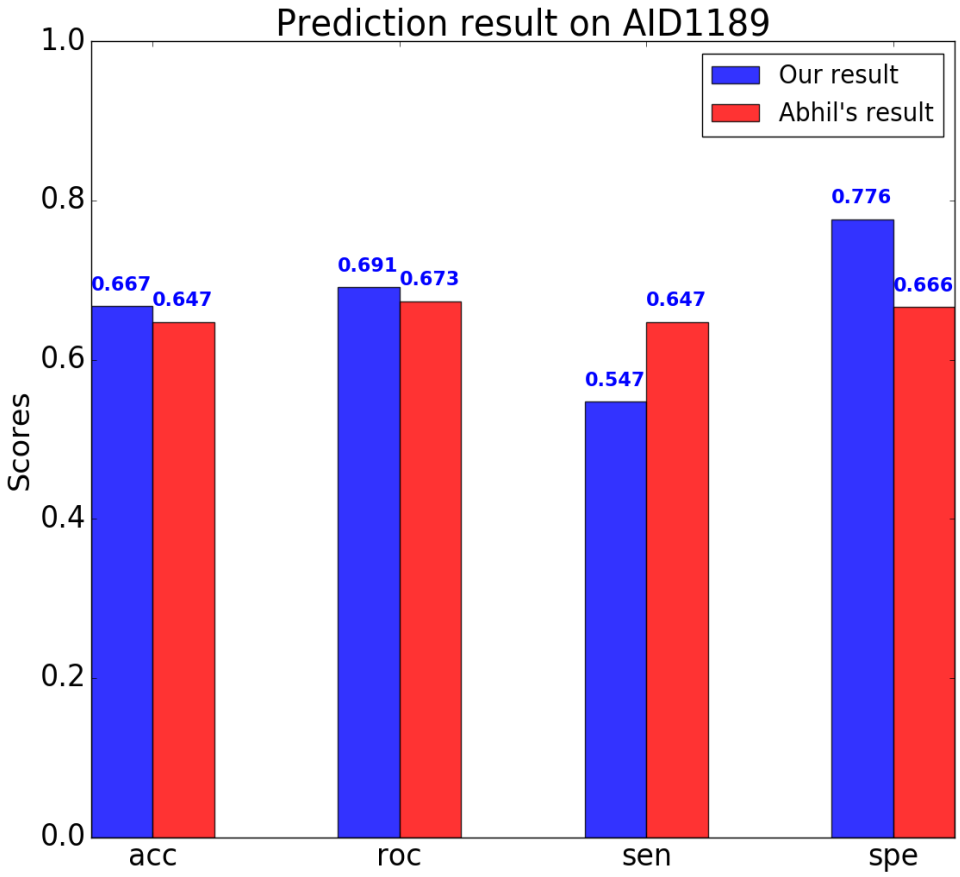
2
1.5
1
0.5
0

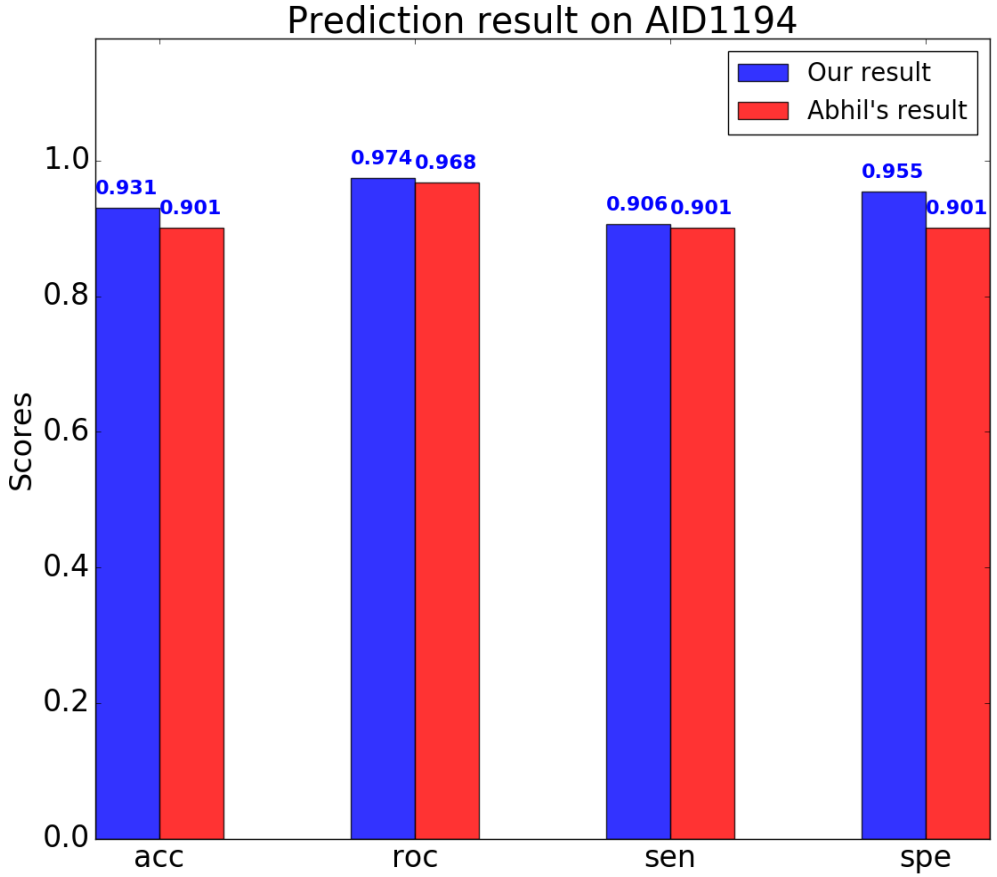
FFT-PP
AB/LogP
S+logP
ACD/logP
CLOGP
VLOGP OPS
ALOGPS
MiLogP
XLOGP3
KowWIN
CSLogP
ALOGP
MolLogP
ALOGP98
OsirisP
VLOGP
TLOGP
ABSOLV
QikProp
QuantlogP
SLIPPER
COSMOFrag
XLOGP2
QLOGP
VEGA
CLIP
LSER
MLOGP(Sim+)
NC+NHET
SPARC
MLOGP(Dragon)
LSER UFZ
VLOGP-NOPS
HINT
GBLOGP

Prediction of drug solubility



Mutagenicity test sets





Cao, Cang, Wu, Wei, 2017

Concluding remarks

- **Multicomponent persistent homology, multidimensional persistent homology, element specific persistent homology, object orientated persistent homology are proposed to retain essential chemical and biological information during the topological simplification of biomolecular geometric complexity.**

- **The abovementioned approaches are integrated with advanced machine learning and deep learning algorithms to achieve the state-of-the-art predictions of protein-ligand binding affinities, mutation induced protein stability changes, drug toxicity, solubility and partition coefficients.**


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